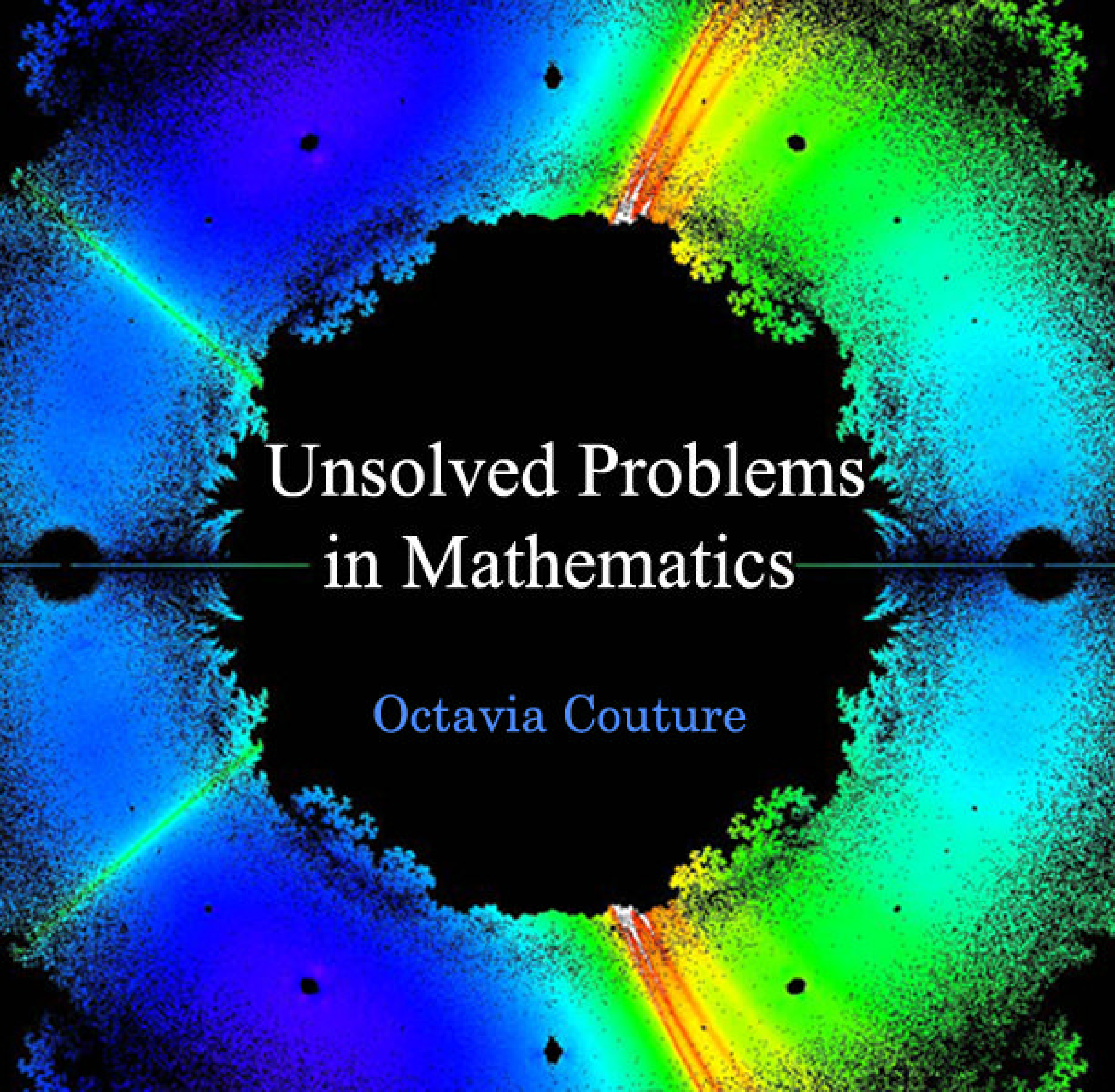


Unsolved Problems in Mathematics

Octavia Couture

First Edition, 2012

ISBN 978-81-323-4253-3

Published by:
White Word Publications
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter 1

Goldbach's Conjecture

Goldbach's conjecture is one of the oldest unsolved problems in number theory and in all of mathematics. It states:

> Every even integer greater than 2 can be expressed as the sum of two primes.

. . .

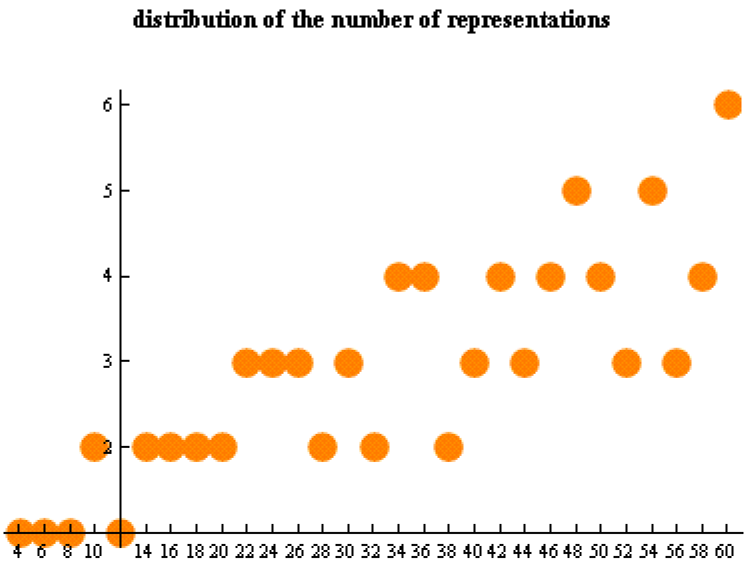
$\{52 = 5 + 47, 52 = 11 + 41, 52 = 23 + 29\}$

$\{54 = 7 + 47, 54 = 11 + 43, 54 = 13 + 41, 54 = 17 + 37, 54 = 23 + 31\}$

$\{56 = 3 + 53, 56 = 13 + 43, 56 = 19 + 37\}$

$\{58 = 5 + 53, 58 = 11 + 47, 58 = 17 + 41, 58 = 29 + 29\}$

$\{60 = 7 + 53, 60 = 13 + 47, 60 = 17 + 43, 60 = 19 + 41, 60 = 23 + 37,$



distribution of the number of representations

The number of ways an even number can be represented as the sum of two primes

Such a number is called a **Goldbach number**. Expressing a given even number as a sum of two primes is called a **Goldbach partition** of the number. For example,

4 = 2 + 2
6 = 3 + 3
8 = 3 + 5
10 = 7 + 3 or 5 + 5
12 = 5 + 7
14 = 3 + 11 or 7 + 7

Origins

On 7 June 1742, the German mathematician Christian Goldbach originally of Brandenburg-Prussia wrote a letter to Leonhard Euler (letter XLIII) in which he proposed the following conjecture:

> Every integer which can be written as the sum of two primes, can also be written as the sum of as many primes as one wishes, until all terms are units.

He then proposed a second conjecture in the margin of his letter:

> Every integer greater than 2 can be written as the sum of three primes.

He considered 1 to be a prime number, a convention subsequently abandoned. The two conjectures are now known to be equivalent, but this did not seem to be an issue at the time. A modern version of Goldbach's marginal conjecture is:

> **Every integer greater than 5 can be written as the sum of three primes**.

Euler replied in a letter dated 30 June 1742, and reminded Goldbach of an earlier conversation they had ("...so Ew vormals mit mir communicirt haben.."), in which Goldbach remarked his original (and not marginal) conjecture followed from the following statement

> **Every even integer greater than 2 can be written as the sum of two primes**,

which is thus also a conjecture of Goldbach. In the letter dated 30 June 1742, Euler stated:

“Dass ... ein jeder numerus par eine summa duorum primorum sey, halte ich für ein ganz gewisses theorema, ungeachtet ich dasselbe necht demonstriren kann.” ("every even integer is a sum of two primes. I regard this as a completely certain theorem, although I cannot prove it.")

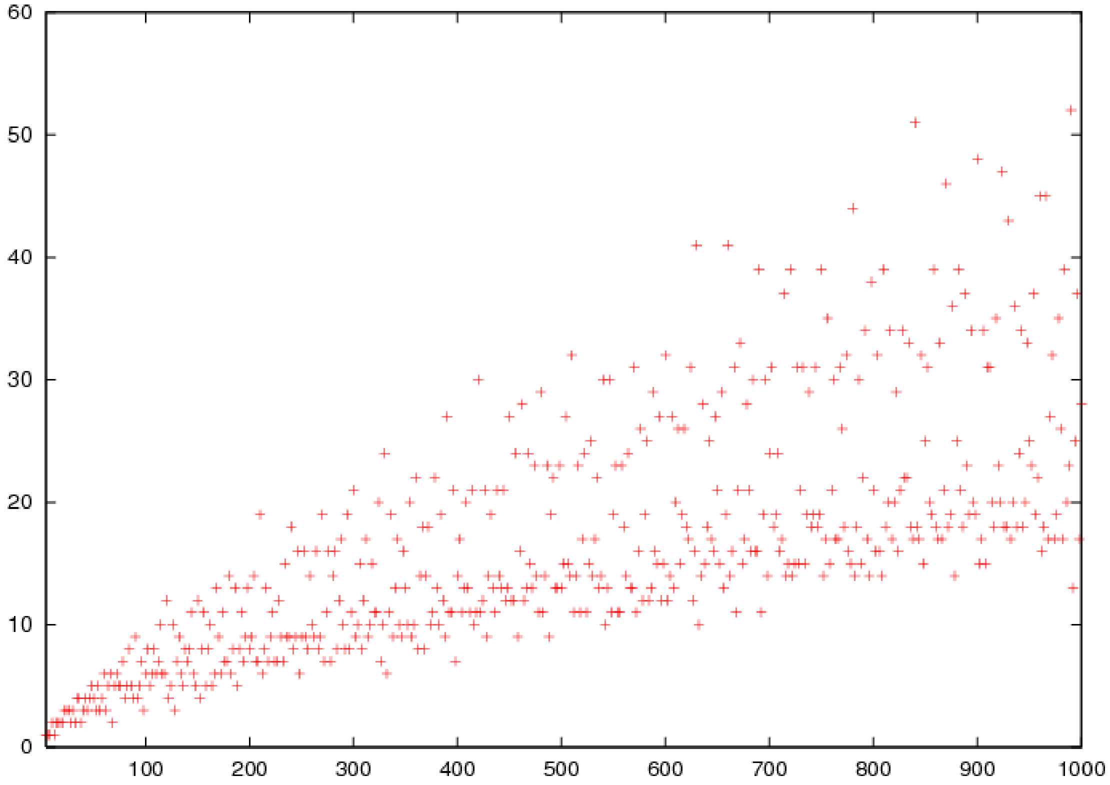
Goldbach's third version (equivalent to the two other versions) is the form in which the conjecture is usually expressed today. It is also known as the "strong", "even", or "binary" Goldbach conjecture, to distinguish it from a weaker corollary. The strong Goldbach conjecture implies the conjecture that **all odd numbers greater than 7 are the sum of three odd primes**, which is known today variously as the "weak" Goldbach conjecture, the "odd" Goldbach conjecture, or the "ternary" Goldbach conjecture. Both questions have remained unsolved ever since, although the weak form of the conjecture appears to be much closer to resolution than the strong one. If the strong Goldbach conjecture is true, the weak Goldbach conjecture will be true by implication.

Verified results

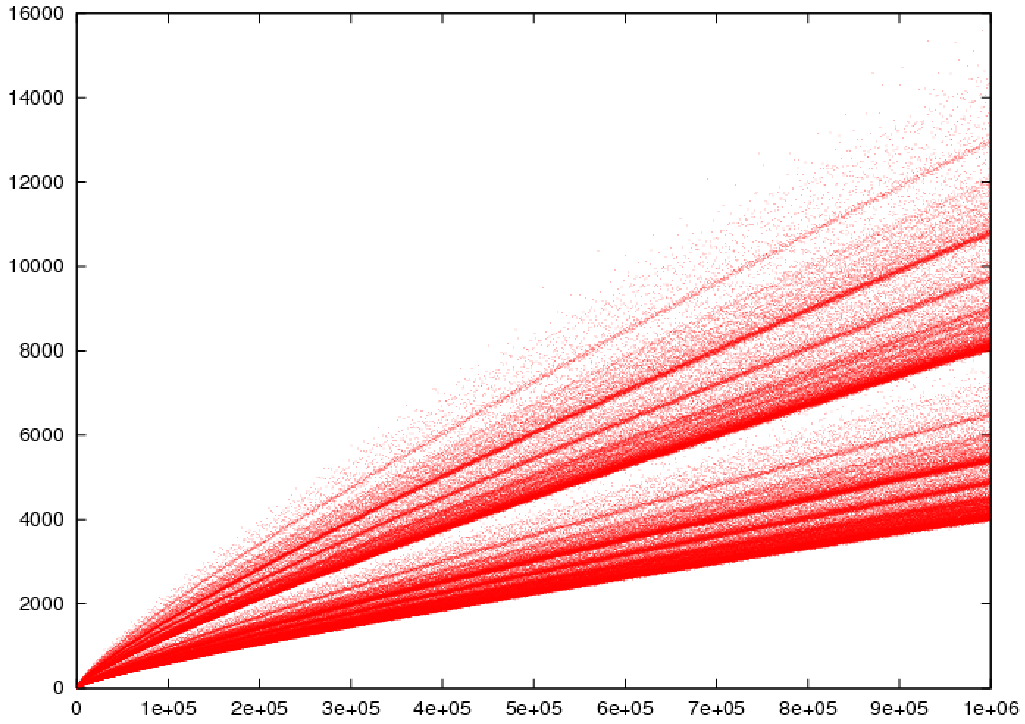
For small values of *n*, the strong Goldbach conjecture (and hence the weak Goldbach conjecture) can be verified directly. For instance, N. Pipping in 1938 laboriously verified the conjecture up to $n \leq 10^5$. With the advent of computers, many more small values of *n* have been checked; T. Oliveira e Silva is running a distributed computer search that has verified the conjecture for $n \leq 1.609 \times 10^{18}$ and some higher small ranges up to 4×10^{18} (double-checked up to 1×10^{17}).

Heuristic justification

Statistical considerations which focus on the probabilistic distribution of prime numbers present informal evidence in favour of the conjecture (in both the weak and strong forms) for sufficiently large integers: the greater the integer, the more ways there are available for that number to be represented as the sum of two or three other numbers, and the more "likely" it becomes that at least one of these representations consists entirely of primes.



Number of ways to write an even number *n* as the sum of two primes ($4 \leq n \leq 1{,}000$)



Number of ways to write an even number n as the sum of two primes ($4 \le n \le 1{,}000{,}000$)

A very crude version of the heuristic probabilistic argument (for the strong form of the Goldbach conjecture) is as follows. The prime number theorem asserts that an integer *m* selected at random has roughly a $1/\ln m$chance of being prime. Thus if *n* is a large even integer and *m* is a number between 3 and *n*/2, then one might expect the probability of *m* and $n - m$ simultaneously being prime to be $1/[\ln m \ln(n-m)]$. This heuristic is non-rigorous for a number of reasons; for instance, it assumes that the events that *m* and $n - m$ are prime are statistically independent of each other. Nevertheless, if one pursues this heuristic, one might expect the total number of ways to write a large even integer *n* as the sum of two odd primes to be roughly

$$\sum_{m=3}^{n/2} \frac{1}{\ln m} \frac{1}{\ln(n-m)} \approx \frac{n}{2\ln^2 n}.$$

Since this quantity goes to infinity as *n* increases, we expect that every large even integer has not just one representation as the sum of two primes, but in fact has very many such representations.

The above heuristic argument is actually somewhat inaccurate, because it ignores some dependence between the events of *m* and $n - m$ being prime. For instance, if *m* is odd then $n - m$ is also odd, and if *m* is even, then $n - m$ is even, a non-trivial relation because (besides 2) only odd numbers can be prime. Similarly, if *n* is divisible by 3, and *m* was already a prime distinct from 3, then $n - m$ would also be coprime to 3 and thus be slightly more likely to be prime than a general number. Pursuing this type of analysis more carefully, Hardy and Littlewood in 1923 conjectured (as part of their famous *Hardy–Littlewood prime tuple conjecture*) that for any fixed $c \geq 2$, the number of representations of a large integer n as the sum of c primes $n = p_1 + \cdots + p_c$ with $p_1 \leq \cdots \leq p_c$ should be asymptotically equal to

$$\left(\prod_p \frac{p\gamma_{c,p}(n)}{(p-1)^c}\right) \int_{2\leq x_1\leq\cdots\leq x_c: x_1+\cdots+x_c=n} \frac{dx_1\cdots dx_{c-1}}{\ln x_1\cdots\ln x_c}$$

where the product is over all primes p, and $\gamma_{c,p}(n)$ is the number of solutions to the equation $n - q_1 + \cdots + q_c \mod p$ in modular arithmetic, subject to the constraints $q_1, \ldots, q_c \neq 0 \mod p$. This formula has been rigorously proven to be asymptotically valid for $c \geq 3$ from the work of Vinogradov, but is still only a conjecture when $c = 2$. In the latter case, the above formula simplifies to 0 when n is odd, and to

$$2\Pi_2 \left(\prod_{p|n;p\geq 3} \frac{p-1}{p-2}\right) \int_2^n \frac{dx}{\ln^2 x} \approx 2\Pi_2 \left(\prod_{p|n;p\geq 3} \frac{p-1}{p-2}\right) \frac{n}{\ln^2 n}$$

when n is even, where Π_2 is the twin prime constant

$$\Pi_2 := \prod_{p\geq 3} \left(1 - \frac{1}{(p-1)^2}\right) = 0.6601618158\ldots.$$

This asymptotic is sometimes known as the *extended Goldbach conjecture*. The strong Goldbach conjecture is in fact very similar to the twin prime conjecture, and the two conjectures are believed to be of roughly comparable difficulty.

The Goldbach partition functions shown here can be displayed as histograms which informatively illustrate the above equations.

Rigorous results

Considerable work has been done on the weak Goldbach conjecture.

The strong Goldbach conjecture is much more difficult. Using the method of Vinogradov, Chudakov, van der Corput, and Estermann showed that almost all even numbers can be written as the sum of two primes (in the sense that the fraction of even numbers which can be so written tends towards 1). In 1930, Lev Schnirelmann proved that every even number $n \geq 4$ can be written as the sum of at most 20 primes. This result was subsequently improved by many authors; currently, the best known result is due to Olivier Ramaré, who in 1995 showed that every even number $n \geq 4$ is in fact the sum of at most six primes. In fact, resolving the weak Goldbach conjecture will also directly imply that every even number $n \geq 4$ is the sum of at most four primes. Leszek Kaniecki showed every odd integer is a sum of at most five primes, under Riemann Hypothesis.

Chen Jingrun showed in 1973 using the methods of sieve theory that every sufficiently large even number can be written as the sum of either two primes, or a prime and a semiprime (the product of two primes)—e.g., 100 = 23 + 7·11.

In 1975, Hugh Montgomery and Robert Charles Vaughan showed that "most" even numbers were expressible as the sum of two primes. More precisely, they showed that there existed positive constants c and C such that for all sufficiently large numbers N, every even number less than N is the sum of two primes, with at most CN^{1-c} exceptions. In particular, the set of even integers which are not the sum of two primes has density zero.

Linnik proved in 1951 the existence of a constant K such that every sufficiently large even number is the sum of two primes and at most K powers of 2. Roger Heath-Brown and Jan-Christoph Schlage-Puchta in 2002 found that $K = 13$ works. This was improved to K=8 by Pintz and Ruzsa.

One can pose similar questions when primes are replaced by other special sets of numbers, such as the squares. For instance, it was proven by Lagrange that every positive integer is the sum of four squares.

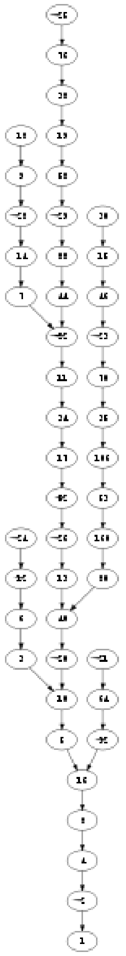
As with many famous conjectures in mathematics, there are a number of purported proofs of the Goldbach conjecture, none accepted by the mathematical community.

Similar conjectures

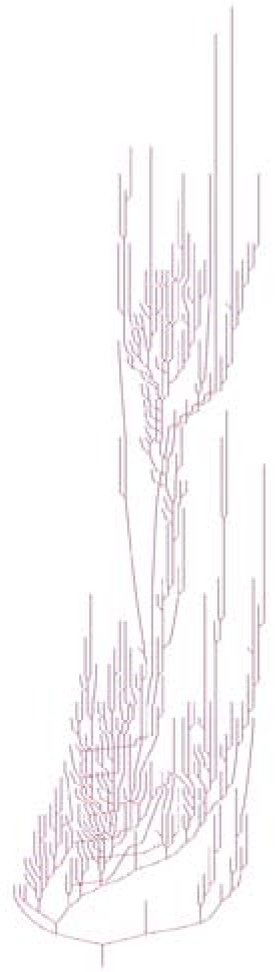
- Lemoine's conjecture (also called *Levy's conjecture*) – states that all odd integers greater than 5 can be represented as the sum of an odd prime number and an even semiprime.
- Waring–Goldbach problem – asks whether large numbers can be expressed as a sum, with at most a constant number of terms, of like powers of primes.

Chapter 2

Collatz Conjecture



Directed graph showing the orbits of small numbers under the Collatz map. The Collatz conjecture is equivalent to the statement that all paths eventually lead to 1.



Directed graph showing the orbits of the first 1000 numbers.

The **Collatz conjecture** is an unsolved conjecture in mathematics named after Lothar Collatz, who first proposed it in 1937. The conjecture is also known as the **3*n* + 1 conjecture**, the **Ulam conjecture** (after Stanisław Ulam), **Kakutani's problem** (after Shizuo Kakutani), the **Thwaites conjecture** (after Sir Bryan Thwaites), **Hasse's algorithm** (after Helmut Hasse), or the **Syracuse problem**; the sequence of numbers involved is referred to as the **hailstone sequence** or **hailstone numbers**, or as **wondrous numbers**.

Take any natural number *n*. If *n* is even, divide it by 2 to get *n* / 2, if *n* is odd multiply it by 3 and add 1 to obtain 3*n* + 1. Repeat the process (which has been called "Half Or Triple Plus One", or **HOTPO**) indefinitely. The conjecture is that no matter what number you start with, you will always eventually reach 1. The property has also been called **oneness**.

Paul Erdős said about the Collatz conjecture: "Mathematics is not yet ready for such problems." He offered $500 for its solution.

In 2006, researchers Kurtz and Simon, building on earlier work by J.H. Conway in the 1970s, proved that a natural generalization of the Collatz problem is undecidable. However, as this proof depends upon the generalization, it cannot be applied to the original Collatz problem.

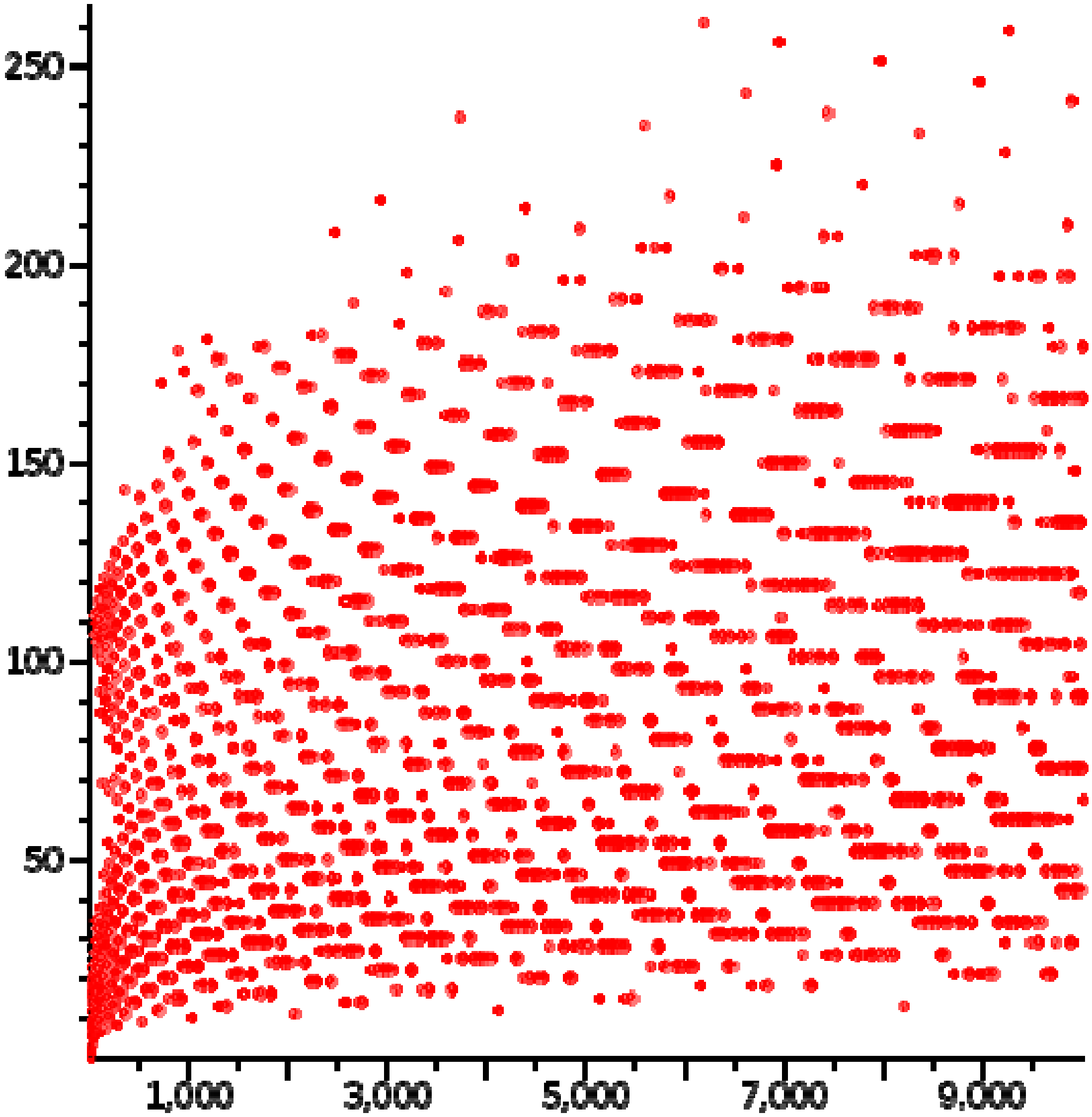
Statement of the problem

Consider the following operation on an arbitrary positive integer:

- If the number is even, divide it by two.
- If the number is odd, triple it and add one.

In modular arithmetic notation, define the function *f* as follows:

$$f(n) = \begin{cases} n/2 & \text{if } n \equiv 0 \pmod{2} \\ 3n+1 & \text{if } n \equiv 1 \pmod{2} \end{cases}$$



Numbers from 1 to 9999 and their corresponding total stopping time.

Now, form a sequence by performing this operation repeatedly, beginning with any positive integer, and taking the result at each step as the input at the next.

In notation:

$$a_i = \begin{cases} n & \text{for } i = 0 \\ f(a_{i-1}) & \text{for } i > 0. \end{cases}$$

or

$$a_i = \frac{1}{2}a_{i-1} - \frac{1}{4}(5a_{i-1}+2)((-1)^{a_{i-1}}-1)$$

The Collatz conjecture is: *This process will eventually reach the number 1, regardless of which positive integer is chosen initially.*

That smallest *i* such that a_i=1 is called the **total stopping time** of *n*. The conjecture asserts that every *n* has a well-defined stopping time. If, for some *n*, such an *i* doesn't exist, we say that *n* has infinite total stopping time and the conjecture is false.

If the conjecture is false, it can only be because there is some starting number which gives rise to a sequence which does not contain 1. Such a sequence might enter a repeating cycle that excludes 1, or increase without bound. No such sequence has been found.

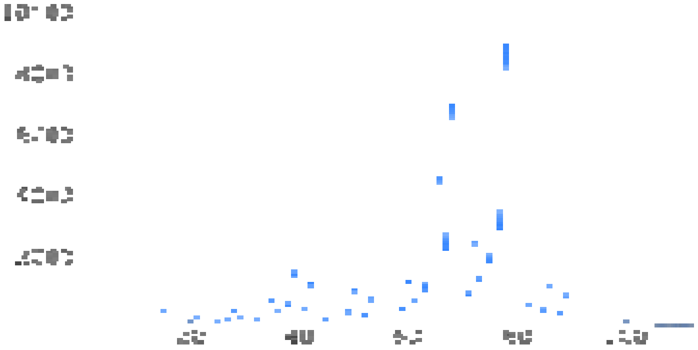
Examples

For instance, starting with *n* = 6, one gets the sequence 6, 3, 10, 5, 16, 8, 4, 2, 1.

n = 11, for example, takes longer to reach 1: 11, 34, 17, 52, 26, 13, 40, 20, 10, 5, 16, 8, 4, 2, 1.

The sequence for *n* = 27, listed and graphed below, takes 111 steps, climbing to over 9000 before descending to 1.

> { 27, 82, 41, 124, 62, 31, 94, 47, 142, 71, 214, 107, 322, 161, 484, 242, 121, 364, 182, 91, 274, 137, 412, 206, 103, 310, 155, 466, 233, 700, 350, 175, 526, 263, 790, 395, 1186, 593, 1780, 890, 445, 1336, 668, 334, 167, 502, 251, 754, 377, 1132, 566, 283, 850, 425, 1276, 638, 319, 958, 479, 1438, 719, 2158, 1079, 3238, 1619, 4858, 2429, **7288**, 3644, 1822, 911, 2734, 1367, 4102, 2051, 6154, 3077, ***9232***, 4616, 2308, 1154, 577, 1732, 866, 433, 1300, 650, 325, 976, 488, 244, 122, 61, 184, 92, 46, 23, 70, 35, 106, 53, 160, 80, 40, 20, 10, 5, 16, 8, 4, 2, 1 }



Starting values with longer stopping time than any smaller starting value are known as "high water marks". The high water mark for numbers less than 100 million is 63,728,127, with 949 steps. The high water mark for numbers less than 1 billion is

670,617,279, with 986 steps. The high water mark for numbers less than 10 billion is 9,780,657,630, with 1132 steps. High water marks are given by sequence A006877 in On-Line Encyclopedia of Integer Sequences, and the number of steps for each starting value are given by A006878.

The powers of two converge to one in the fewest steps, because 2^n is halved *n* times to reach one, and is never increased.

Program to calculate Collatz sequences

A specific Collatz sequence can be easily computed, as is shown by this pseudocode example:

```
function collatz(n)
    while n > 1
        show n
        if n is odd then
            set n = 3n + 1
        else
            set n = n / 2
        endif
    endwhile
    show n
```

This program halts when the sequence reaches 1, in order to avoid printing an endless cycle of 4, 2, 1. If the Collatz conjecture is true, the program will always **halt** (stop) no matter what positive starting integer is given to it.

m-cycles

The proof of the conjecture can indirectly be done by proving the following:

- no infinite divergent trajectory occurs
- no cycle occurs

thus all numbers have a trajectory down to 1.

In 1977, R. Steiner, and in 2000 and 2002, J. Simons and B. de Weger (based on Steiner's work), proved the nonexistence of certain types of cycles.

Notation

To explain this we refer to the definition as given in the section Syracuse function below:

Define the transformation for odd positive integer numbers *a* and *b*, positive integer numbers *A*:

b = T(a;A) *meaning* b = (3*a + 1)/2^A

where *A* has the maximum value which leaves *b* an integer.

Example:
a=7. *Then* b=T(7,A) and 3*a + 1 = 22,
so A = 1
b = (3*7 + 1)/2^1 = 11
11 = T(7;1)

Define the concatenation (extensible up to arbitrary length):

b = T(T(a;A);B) = T(a;A,B)
Example:

b = T(7;A,B) = T(7;1,1) = ((3*7 + 1)/2^1*3 + 1)/2^B = (3*11 + 1)/2^B = 34/2^1 = 17

17 = T(7;1,1)

Define a "one-peak-transformation" of *L* ascending and *1* descending exponents/steps:

b = T(a;1,1,1,1,1,....,1,A) = $T(a;(1)_L,A)$

with *L* (arbitrary) many exponents *1* and exactly one exponent *A>1*

Example:
b = T(7;1,1,A) = $T(7;(1)_2,A)$ = (17*3 + 1)/2^A = 52/2^2 = 13
13 = $T(7;(1)_2,2)$

then call the construction

a = $T(a;(1)_L,A)$ // *arbitrary positive value for number of increasing steps L*

a "***1-cycle***" of length *N = L + 1* (steps).

Theorems

- Steiner proved in 1977 there is no *1-cycle*. However many *L* steps may be chosen, a number *x* satisfying the loop-condition is never an integer.
- Simons proved in 2000 (based on Steiner's method) there is no *2-cycle* $a = T(a;(1)_L,A,(1)_M,B)$ however many *L* and *M* steps may be chosen.
- Simons/deWeger in 2003 extended their own proof up to "68-cycles": there is no *m-cycle* up to $m=68$ $a=T(a;(1)_{L1},A_1,(1)_{L2},A_2,...,(1)_{L68},A_{68})$. Whatever number of steps in L_1 to L_{68} and whatever exponents A_1 to A_{68} one may choose, there is no positive odd integer number *a* satisfying the cycle condition. Steiner claimed in a usenet discussion he could extend this up to *m=71*.

Supporting arguments

Although the conjecture has not been proven, most mathematicians who have looked into the problem think the conjecture is true because experimental evidence and heuristic arguments support it.

Experimental evidence

The conjecture has been checked by computer for all starting values up to 20 × $2^{58} \approx 5.764\times10^{18}$. All initial values tested so far eventually end in the repeating cycle {4,2,1}, which has only three terms. It is also known that {4,2,1} is the only repeating cycle possible with fewer than 35400 terms.

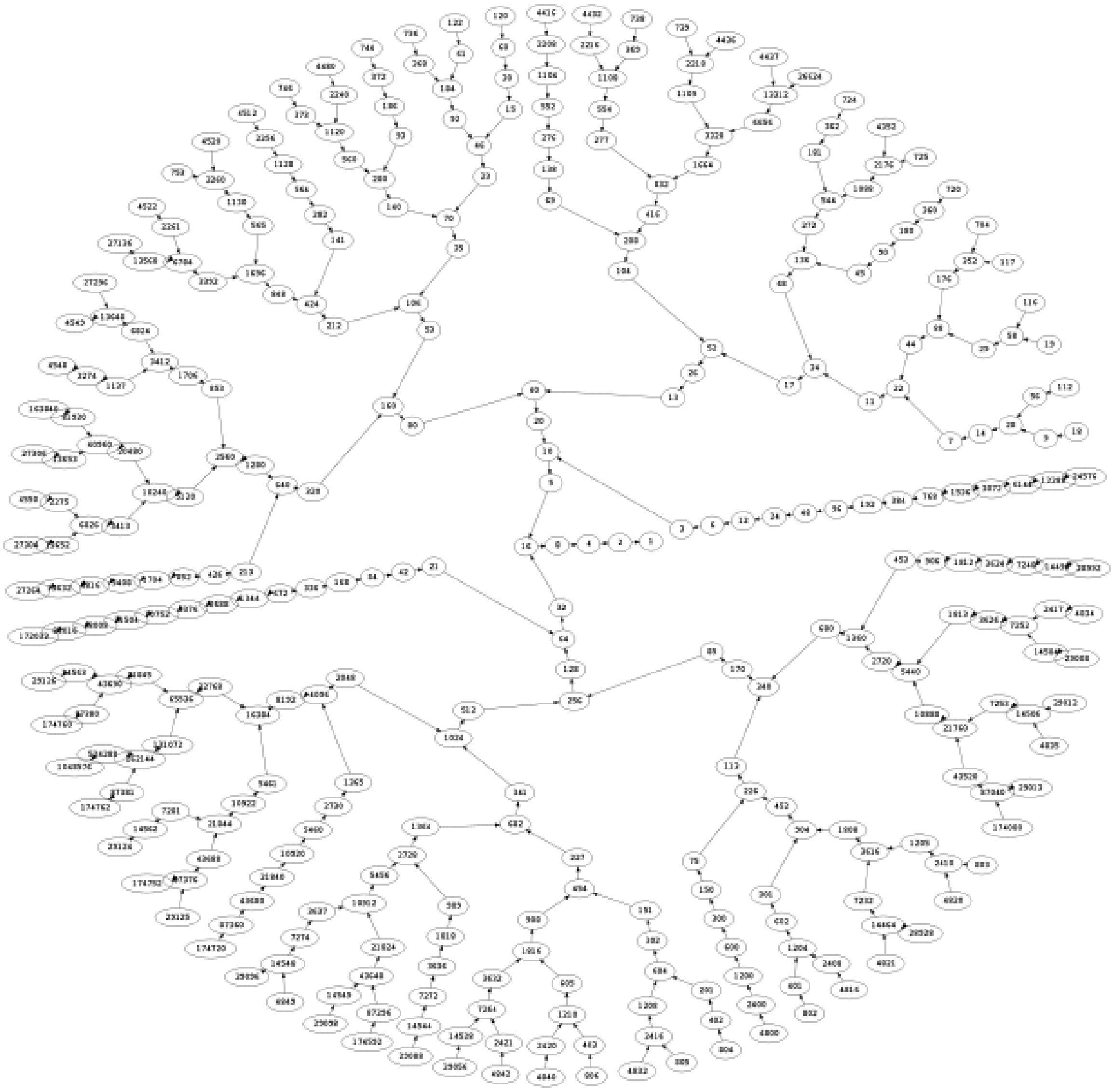
Such computer evidence is not a proof that the conjecture is true. As shown in the cases of the Pólya conjecture, the Mertens conjecture and the Skewes' number, sometimes a conjecture's only counterexamples are found when using very large numbers. Since sequentially examining all natural numbers is a process which can never be completed, such an approach can never demonstrate that the conjecture is true, merely that no counterexamples have yet been discovered.

A probabilistic heuristic

If one considers only the *odd* numbers in the sequence generated by the Collatz process, then each odd number is on average 3/4 of the previous one. (More precisely, the geometric mean of the ratios of outcomes is 3/4.) This yields a heuristic argument that every Collatz sequence should decrease in the long run, although this is not evidence against other cycles, only against divergence. The argument is not a proof because it assumes that Collatz sequences are assembled from uncorrelated probabilistic events. (It does rigorously establish that the 2-adic extension of the Collatz process has 2 division steps for every multiplication step for almost all 2-adic starting values.)

Other formulations of the conjecture

In reverse



The first 20 levels of the *Collatz graph* generated in bottom-up fashion. The graph includes all numbers with an orbit length of 20 or less.

There is another approach to prove the conjecture, which considers the bottom-up method of growing the so called *Collatz graph*. The *Collatz graph* is a graph defined by the inverse relation

$$R(n) = \begin{cases} 2n & \text{if } n \equiv 0,1,2,3,5 \\ (n-1)/3 & \text{if } n \equiv 4 \end{cases} \pmod{6}.$$

So, instead of proving that all natural numbers eventually lead to 1, we can prove that 1 leads to all natural numbers. For any integer n, $n \equiv 1 \pmod 2$ iff $3n + 1 \equiv 4 \pmod 6$. Equivalently, $(n - 1)/3 \equiv 1 \pmod 2$ iff $n \equiv 4 \pmod 6$. Conjecturally, this inverse relation forms a tree except for the 1-2-4 loop (the inverse of the 1-4-2 loop of the unaltered function f defined in the statement of the problem above). When the relation $3n + 1$ of the function f is replaced by the common substitute "shortcut" relation $(3n + 1)/2$, the Collatz graph is defined by the inverse relation,

$$R(n) = \begin{cases} 2n & \text{if } n \equiv 0, 1 \\ (2n-1)/3 & \text{if } n \equiv 2 \end{cases} \pmod 3.$$

Conjecturally, this inverse relation forms a tree except for a 1-2 loop (the inverse of the 1-2 loop of the function $f(n)$ revised as indicated above).

As rational numbers

The natural numbers can be converted to rational numbers in a certain way. To get the rational version, find the highest power of two less than or equal to the number, use it as the denominator, and subtract it from the original number for the numerator (527 → 15/512). To get the natural version, add the numerator and denominator (255/256 → 511).

The Collatz conjecture then says that the numerator will eventually equal zero. The Collatz function changes to:

$$f(n,d) = \begin{cases} (3n+d+1)/2d & \text{if } 3n+d+1 < 2d \\ (3n-d+1)/4d & \text{if } 3n+d+1 \geq 2d \end{cases}$$

(n = numerator; d = denominator).

This works because $3x + 1 = 3(d + n) + 1 = (2d) + (3n + d + 1) = (4d) + (3n - d + 1)$. Reducing a rational before every operation is required to get x as an odd.

As an abstract machine that computes in base two

Repeated applications of the Collatz function can be represented as an abstract machine that handles strings of bits. The machine will perform the following three steps on any odd number until only one "1" remains:

1. Append 1 to the (right) end of the number in binary (giving $2n+1$);
2. Add this to the original number by binary addition (giving $2n+1 + n = 3n+1$);
3. Remove all trailing "0"s (i.e. repeatedly divide by two until the result is odd).

This prescription is plainly equivalent to computing a Collatz sequence in base two.

Example

The starting number 7 is written in base two as 111. The resulting Collatz sequence is:

```
     111
    1111
   1011~~0~~
   10111
  10001~~0~~
 100011
 1101~~00~~
 11011
101~~000~~
 1011
1~~0000~~
```

As a parity sequence

For this section, consider the Collatz function in the slightly modified form

$$f(n) = \begin{cases} n/2 & \text{if } n \equiv 0 \\ (3n+1)/2 & \text{if } n \equiv 1. \end{cases} \pmod{2}$$

This can be done because when n is odd, $3n + 1$ is always even.

If P(…) is the parity of a number, that is $P(2n) = 0$ and $P(2n + 1) = 1$, then we can define the Collatz parity sequence for a number n as $p_i = P(a_i)$, where $a_0 = n$, and $a_{i+1} = f(a_i)$.

Using this form for $f(n)$, it can be shown that the parity sequences for two numbers m and n will agree in the first k terms if and only if m and n are equivalent modulo 2^k. This implies that every number is uniquely identified by its parity sequence, and moreover that if there are multiple Collatz cycles, then their corresponding parity cycles must be different.

The proof is simple: it is easy to verify by hand that applying the f function k times to the number $a\ 2^k+b$ will give the result $a\ 3^c+d$, where d is the result of applying the f function k times to b, and c is how many odd numbers were encountered during that sequence. So the parity of the first k numbers is determined purely by b, and the parity of the (k+1)th number will change if the least significant bit of a is changed.

The Collatz Conjecture can be rephrased as stating that the Collatz parity sequence for every number eventually enters the cycle 0 → 1 → 0.

As a tag system

For the Collatz function in the form

$$f(n) = \begin{cases} n/2 & \text{if } n \equiv 0 \\ (3n+1)/2 & \text{if } n \equiv 1. \end{cases} \pmod 2$$

Collatz sequences can be computed by the extremely simple 2-tag system with production rules *a* → *bc*, *b* → *a*, *c* → *aaa*. In this system, the positive integer *n* is represented by a string of *n a*'s, and iteration of the tag operation halts on any word of length less than 2. (Adapted from De Mol.)

The Collatz conjecture equivalently states that this tag system, with an arbitrary finite string of *a*'s as the initial word, eventually halts.

Extensions to larger domains

Iterating on all integers

An obvious extension is to include all integers, not just positive integers. In this case there are a total of 5 known cycles, which all integers seem to eventually fall into under iteration of *f*. These cycles are listed here, starting with the well-known cycle for positive n.

Odd values are listed in bold. Each cycle is listed with its member of least absolute value (which is always odd or zero) first.

Cycle	Odd-value cycle length	Full cycle length
1 → 4 → 2 → **1 …**	1	3
0 → 0 **…**	0	1
−1 → −2 → **−1 …**	1	2
−5 → −14 → **−7** → −20 → −10 → **−5 …**	2	5
−17 → −50 → **−25** → −74 → **−37** → −110 → **−55** → −164 → −82 → **−41** → −122 → **−61** → −182 → **−91** → −272 → −136 → −68 → −34 → **−17 …**	7	18

The Generalized Collatz Conjecture is the assertion that every integer, under iteration by *f*, eventually falls into one of these five cycles.

Iterating with odd denominators or 2-adic integers

The standard Collatz map can be extended to (positive or negative) rational numbers which have odd denominators when written in lowest terms. The number is taken to be odd or even according to whether its numerator is odd or even. A closely related fact is that the Collatz map extends to the ring of 2-adic integers, which contains the ring of rationals with odd denominators as a subring.

The parity sequences as defined above are no longer unique for fractions. However, it can be shown that any possible parity cycle is the parity sequence for exactly one fraction: if a cycle has length n and includes odd numbers exactly m times at indices k_0, …, k_{m-1}, then the unique fraction which generates that parity cycle is

$$\frac{3^{m-1}2^{k_0} + ... + 3^0 2^{k_{m-1}}}{2^n - 3^m}.$$

For example, the parity cycle (1 0 1 1 0 0 1) has length 7 and has 4 odd numbers at indices 0, 2, 3, and 6. The unique fraction which generates that parity cycle is

$$\frac{3^3 2^0 + 3^2 2^2 + 3^1 2^3 + 3^0 2^6}{2^7 - 3^4} = \frac{151}{47}.$$

The complete cycle being: 151/47 → 250/47 → 125/47 → 211/47 → 340/47 → 170/47 → 85/47 → 151/47

Although the cyclic permutations of the original parity sequence are unique fractions, the cycle is not unique, each permutation's fraction being the next number in the loop cycle:

(0 1 1 0 0 1 1) → $\dfrac{3^3 2^1 + 3^2 2^2 + 3^1 2^5 + 3^0 2^6}{2^7 - 3^4} = \dfrac{250}{47}$

(1 1 0 0 1 1 0) → $\dfrac{3^3 2^0 + 3^2 2^1 + 3^1 2^4 + 3^0 2^5}{2^7 - 3^4} = \dfrac{125}{47}$

(1 0 0 1 1 0 1) → $\dfrac{3^3 2^0 + 3^2 2^3 + 3^1 2^4 + 3^0 2^6}{2^7 - 3^4} = \dfrac{211}{47}$

(0 0 1 1 0 1 1) → $\dfrac{3^3 2^2 + 3^2 2^3 + 3^1 2^5 + 3^0 2^6}{2^7 - 3^4} = \dfrac{340}{47}$

(0 1 1 0 1 1 0) → $\dfrac{3^3 2^1 + 3^2 2^2 + 3^1 2^4 + 3^0 2^5}{2^7 - 3^4} = \dfrac{170}{47}$

$$(1\ 1\ 0\ 1\ 1\ 0\ 0) \rightarrow \frac{3^3 2^0 + 3^2 2^1 + 3^1 2^3 + 3^0 2^4}{2^7 - 3^4} = \frac{85}{47}$$

Also, for uniqueness, the parity sequence should be "prime", i.e., not partitionable into identical sub-sequences. For example, parity sequence (1 1 0 0 1 1 0 0) can be partitioned into two identical sub-sequences (1 1 0 0)(1 1 0 0). Calculating the 8-element sequence fraction gives

$$(1\ 1\ 0\ 0\ 1\ 1\ 0\ 0) \rightarrow \frac{3^3 2^0 + 3^2 2^1 + 3^1 2^4 + 3^0 2^5}{2^8 - 3^4} = \frac{125}{175}$$

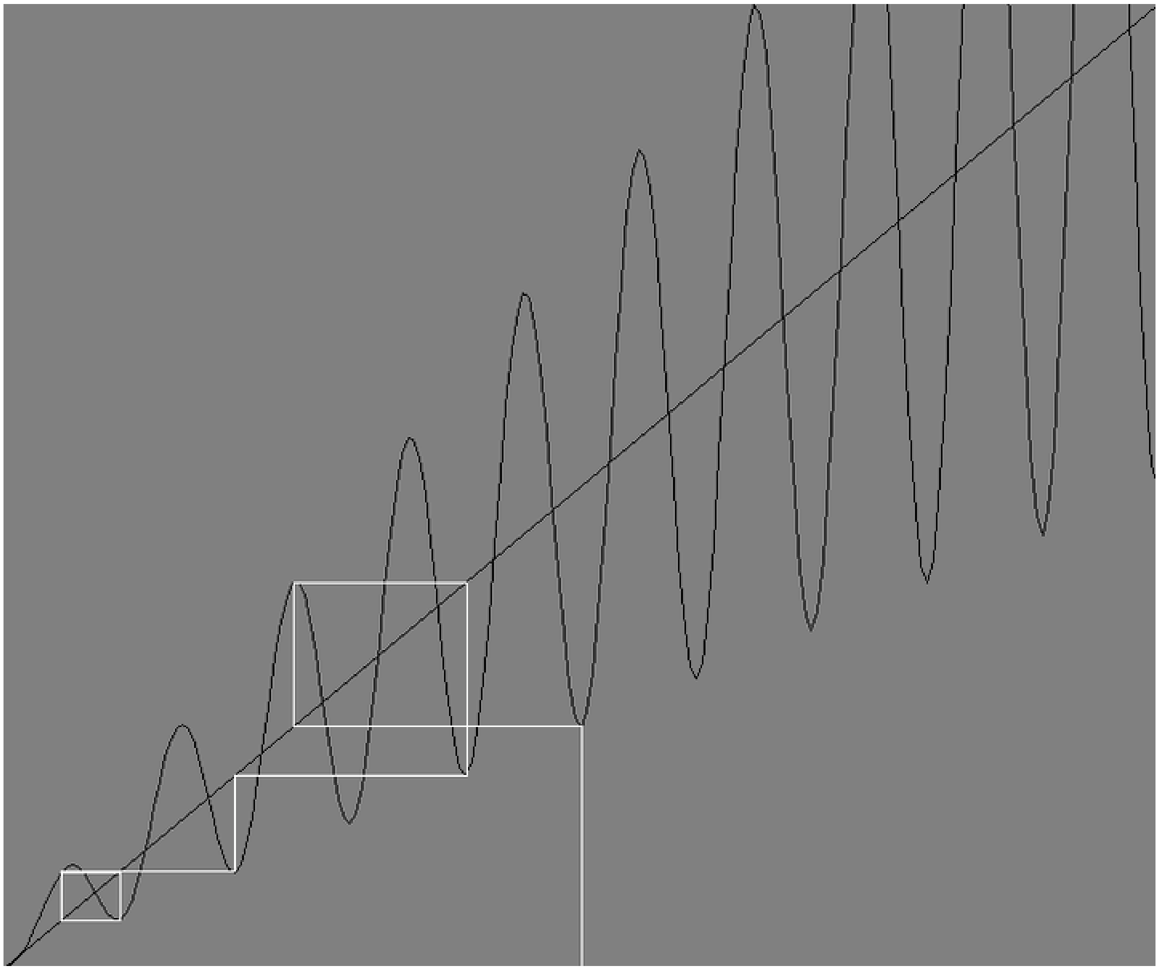
But when reduced to lowest terms {5/7}, it is the same as that of the 4-element subsequence

$$(1\ 1\ 0\ 0) \rightarrow \frac{3^1 2^0 + 3^0 2^1}{2^4 - 3^2} = \frac{5}{7}$$

And this is because the 8-element parity sequence actually represents two circuits of the loop cycle defined by the 4-element parity sequence.

In this context, the Collatz conjecture is equivalent to saying that (0 1) is the only cycle which is generated by positive whole numbers (i.e. 1 and 2).

Iterating on real or complex numbers



Cobweb plot of the orbit 10-5-8-4-2-1-2-1-2-1-etc. in the real extension of the Collatz map (optimized by replacing "3*n* + 1" with "(3*n* + 1)/2")

The Collatz map can be viewed as the restriction to the integers of the smooth real and complex map

$$f(z)=\frac{1}{2}z\cos^2\left(\frac{\pi}{2}z\right)+(3z+1)\sin^2\left(\frac{\pi}{2}z\right),$$

which simplifies to $\frac{1}{4}(2+7z-(2+5z)\cos(\pi z))$.

If the standard Collatz map defined above is optimized by replacing the relation 3*n* + 1 with the common substitute "shortcut" relation (3*n* + 1)/2, it can be viewed as the restriction to the integers of the smooth real and complex map

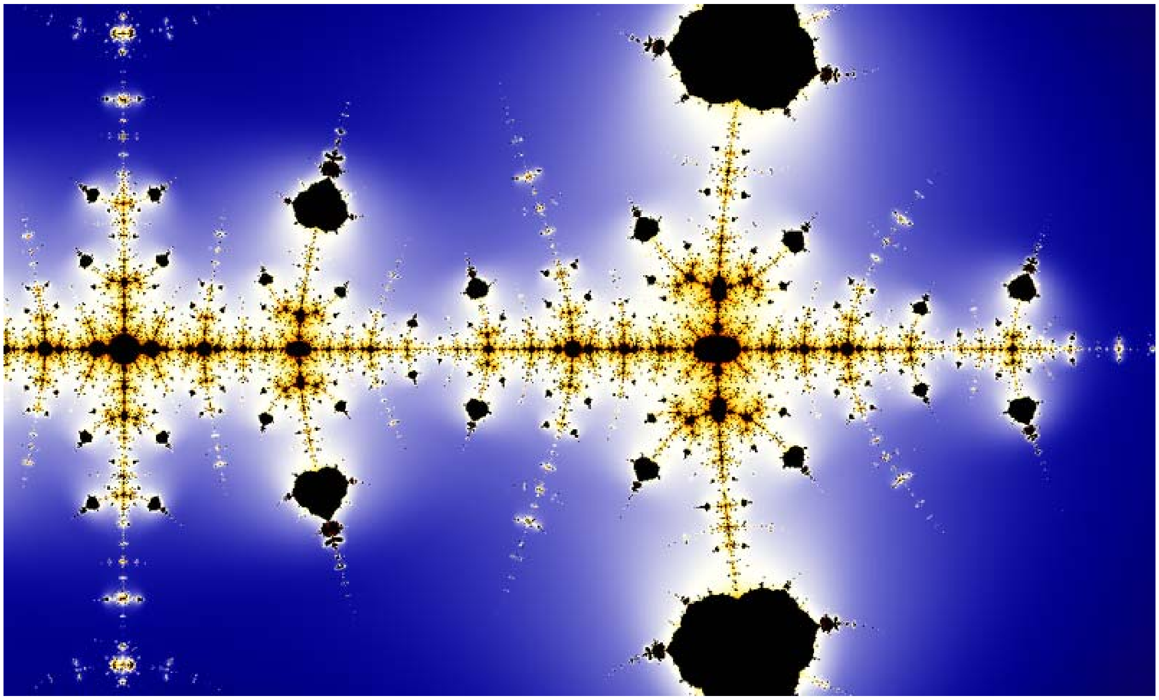
$$f(z) = \frac{1}{2} z \cos^2\left(\frac{\pi}{2} z\right) + \frac{1}{2}(3z+1)\sin^2\left(\frac{\pi}{2} z\right),$$

which simplifies to $\frac{1}{4}(1 + 4z - (1+2z)\cos(\pi z))$.

Collatz fractal

Iterating the above optimized map in the complex plane produces the Collatz fractal.

The point of view of iteration on the real line was investigated by Chamberland (1996), and on the complex plane by Letherman, Schleicher, and Wood (1999).



Collatz map fractal in a neighbourhood of the real line

Optimizations

The As a parity sequence section above gives a way to speed up simulation of the sequence. To jump ahead *k* steps on each iteration (using the *f* function from that section), break up the current number into two parts, *b* (the *k* least significant bits, interpreted as an integer), and *a* (the rest of the bits as an integer). The result of jumping ahead *k* steps can be found as:

$f^k(a\ 2^k+b) = a\ 3^{c[\mathrm{b}]}+d[\mathrm{b}]$.

The *c* and *d* arrays are precalculated for all possible *k*-bit numbers *b*, where *d* [b] is the result of applying the *f* function *k* times to *b*, and *c* [b] is the number of odd numbers encountered on the way. For example, if k=5, you can jump ahead 5 steps on each iteration by separating out the 5 least significant bits of a number and using:

> *c* [0...31] = {0,3,2,2,2,2,2,4,1,4,1,3,2,2,3,4,1,2,3,3,1,1,3,3,2,3,2,4,3,3,4,5}
> *d* [0...31] =
> {0,2,1,1,2,2,2,20,1,26,1,10,4,4,13,40,2,5,17,17,2,2,20,20,8,22,8,71,26,26,80,242}.

This requires 2^k precomputation and storage to speed up the resulting calculation by a factor of *k*.

For the special purpose of searching for a counterexample to the Collatz conjecture, this precomputation leads to an even more important acceleration which is due to Tomás Oliveira e Silva and is used in the record confirmation of the Collatz conjecture. If, for some given *b* and *k*, the inequality

$$f^k(a\ 2^k+b) = a\ 3^{c[\mathrm{b}]}+d[\mathrm{b}] < a\ 2^k+b$$

holds for all *a*, then the first counterexample, if it exists, cannot be *b* modulo 2^k. For instance, the first counterexample must be odd because $f(2n) = n$; and it must be 3 mod 4 because $f^3(4n+1) = 3n+1$. For each starting value *a* which is not a counterexample to the Collatz conjecture, there is a *k* for which such an inequality holds, so checking the Collatz conjecture for one starting value is as good as checking an entire congruence class. As *k* increases, the search only needs to check those residues *b* that are not eliminated by lower values of *k*. On the order of $3^{k/2}$ residues survive. For example, the only surviving residues mod 32 are 7, 15, 27, and 31; only 573,162 residues survive mod 2^{25} = 33,554,432.

Syracuse function

If *k* is an odd integer, then $3k + 1$ is even, so we can write $3k + 1 = 2^a k'$, with k' odd and $a \geq 1$. We define a function *f* from the set *I* of odd integers into itself, called the *Syracuse Function,* by taking $f(k) = k'$ (sequence A075677 in OEIS).

Some properties of the Syracuse function are:

- $f(4k + 1) = f(k)$ for all *k* in *I*.
- For all $p \geq 2$ and *h* odd, $f^{p-1}(2^p h - 1) = 2\ 3^{p-1}h - 1$ (here, f^{p-1} is function iteration notation).
- For all odd *h*, $f(2h - 1) \leq (3h - 1)/2$

The Syracuse Conjecture is that for all *k* in *I*, there exists an integer $n \geq 1$ such that $f^n(k) = 1$. Equivalently, let *E* be the set of odd integers *k* for which there exists an integer $n \geq 1$ such that $f^n(k) = 1$. The problem is to show that $E = I$. The following is the beginning of an attempt at a proof by induction:

1, 3, 5, 7, and 9 are known to be elements of E. Let k be an odd integer greater than 9. Suppose that the odd numbers up to and including $k - 2$ are in E and let us try to prove that k is in E. As k is odd, $k + 1$ is even, so we can write $k + 1 = 2^p h$ for $p \geq 1$, h odd, and $k = 2^p h - 1$. Now we have:

- If $p = 1$, then $k = 2h - 1$. It is easy to check that $f(k) < k$, so $f(k) \in E$; hence $k \in E$.
- If $p \geq 2$ and h is a multiple of 3, we can write $h = 3h'$. Let $k' = 2^{p+1}h' - 1$; then $f(k') = k$, and as $k' < k$, k' is in E; therefore $k = f(k') \in E$.
- If $p \geq 2$ and h is not a multiple of 3 but $h \equiv (-1)^p \bmod 4$, we can still show that $k \in E$.

The problematic case is that where $p \geq 2$, h not multiple of 3 and $h \equiv (-1)^{p+1} \bmod 4$. Here, if we manage to show that for every odd integer k', $1 \leq k' \leq k-2$; $3k' \in E$ we are done.

Chapter 3

Union-Closed Sets Conjecture and Barnette's Conjecture

Union-closed sets conjecture

In combinatorial mathematics, the **union-closed sets conjecture** is an elementary problem, posed by Péter Frankl in 1979 and still open. A family of sets is said to be *union-closed* if the union of any two sets from the family remains in the family. The conjecture states that for any finite union-closed family of finite sets, other than the family consisting only of the empty set, there exists an element that belongs to at least half of the sets in the family.

Equivalent forms

If *F* is a union-closed family of sets, the family of complement sets to sets in *F* is closed under intersection, and an element that belongs to at least half of the sets of *F* belongs to at most half of the complement sets. Thus, an equivalent form of the conjecture (the form in which it was originally stated) is that, for any intersection-closed family of sets that contains more than one set, there exists an element that belongs to at most half of the sets in the family.

Although stated above in terms of families of sets, Frankl's conjecture has also been formulated and studied as a question in lattice theory. A lattice is a partially ordered set in which for two elements *x* and *y* there is a unique greatest element less than or equal to both of them (the *meet* of *x* and *y*) and a unique least element greater than or equal to both of them (the *join* of *x* and *y*). The family of all subsets of a set *S*, ordered by set inclusion, forms a lattice in which the meet is represented by the set-theoretic intersection and the join is represented by the set-theoretic union; a lattice formed in this way is called a Boolean lattice. The lattice-theoretic version of Frankl's conjecture is that in any finite lattice there exists an element *x* that is not the join of any two smaller elements, and such that the number of elements greater than or equal to *x* totals at most half the lattice, with equality only if the lattice is a Boolean lattice. As Abe (2000) shows, this statement about lattices is equivalent to the Frankl conjecture for union-closed sets: each lattice can be translated into a union-closed set family, and each union-closed set family can be translated into a lattice, such that the truth of the Frankl conjecture for the translated object implies the truth of the conjecture for the original object. This lattice-theoretic version of the conjecture is known to be true for several natural subclasses of lattices (Abe 2000; Poonen 1992; Reinhold 2000) but remains open in the general case.

Families known to satisfy the conjecture

The conjecture has been proven for many special cases of union-closed set families. In particular, it is known to be true for

- families of at most 36 sets,
- families of sets such that their union has at most 11 elements, and
- families of sets in which the smallest set has one or two elements.

History

Péter Frankl stated the conjecture, in terms of intersection-closed set families, in 1979, and so the conjecture is usually credited to him and sometimes called the **Frankl conjecture**. The earliest publication of the union-closed version of the conjecture appears to be by Duffus (1985).

Barnette's conjecture

Barnette's conjecture is an unsolved problem in graph theory, a branch of mathematics, concerning Hamiltonian cycles in graphs. It is named after David W. Barnette, a professor emeritus at the University of California, Davis; it states that every bipartite polyhedral graph with three edges per vertex has a Hamiltonian cycle.

Definitions

A planar graph is an undirected graph that can be embedded into the Euclidean plane without any crossings. A planar graph is called polyhedral if and only if it is 3-vertex-connected, that is, if there do not exist two vertices the removal of which would disconnect the rest of the graph. A graph is bipartite if its vertices can be colored with two different colors such that each edge has one endpoint of each color. A graph is cubic (or 3-regular) if each vertex is the endpoint of exactly three edges. And, a graph is Hamiltonian if there exists a cycle that passes exactly once through each of its vertices. Barnette's conjecture states that every cubic bipartite polyhedral graph is Hamiltonian.

By Steinitz's theorem, a planar graph represents the edges and vertices of a convex polyhedron if and only if it is polyhedral. And, a planar graph is bipartite if and only if, in a planar embedding of the graph, all face cycles have even length. Therefore, Barnette's conjecture may be stated in an equivalent form: suppose that a three-dimensional convex polyhedron has an even number of edges on each of its faces. Then, according to the conjecture, the graph of the polyhedron has a Hamiltonian cycle.

History

In P. G. Tait (1884) conjectured that every cubic polyhedral graph is Hamiltonian; this came to be known as Tait's conjecture. It was disproven by W. T. Tutte (1946), who constructed a counterexample with 46 vertices; other researchers later found even smaller counterexamples. However, none of these known counterexamples is bipartite. Tutte himself conjectured that every cubic 3-connected bipartite graph is Hamiltonian, but this was shown to be false by the discovery of a counterexample, the Horton graph. David W. Barnette (1969) proposed a weakened combination of Tait's and Tutte's conjectures, stating that every bipartite cubic polyhedron is Hamiltonian, or, equivalently, that every counterexample to Tait's conjecture is non-bipartite.

Equivalent forms

Kelmans (1994) showed that Barnette's conjecture is equivalent to a stronger statement, that for every two edges *e* and *f* on the same face of a bipartite cubic polyhedron, there exists a Hamiltonian cycle that contains *e* but does not contain *f*. Clearly, if this stronger statement is true, then every bipartite cubic polyhedron contains a Hamiltonian cycle: just choose *e* and *f* arbitrarily. In the other directions, Kelman showed that a counterexample to this stronger conjecture could be transformed into a counterexample to the original Barnette conjecture.

Barnette's conjecture is also equivalent to the statement that the vertices of every cubic bipartite polyhedral graph can be partitioned into two subsets in such a way that every cycle of the graph passes through both subsets; that is, the graph can be covered by two induced forests.

Partial results

Although the truth of Barnette's conjecture remains unknown, computational experiments have shown that there is no counterexample with fewer than 86 vertices.

If Barnette's conjecture turns out to be false, then it can be shown to be NP-complete to test whether a bipartite cubic polyhedron is Hamiltonian. If a planar graph is bipartite and cubic but only 2-connected, then it may be non-Hamiltonian, and it is NP-complete to test Hamiltonicity for these graphs.

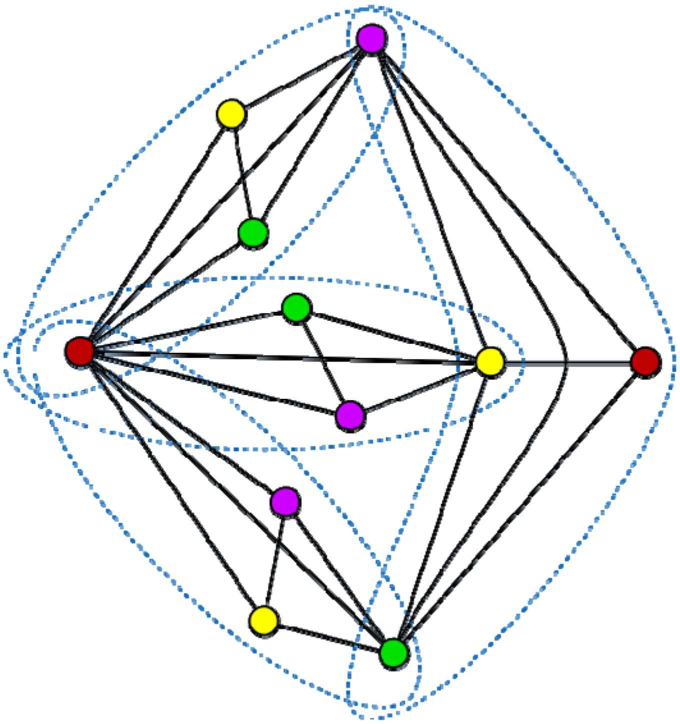
Related problems

A related conjecture of Barnette states that every cubic polyhedral graph in which all faces have six or fewer edges is Hamiltonian. Computational experiments have shown that, if a counterexample exists, it would have to have more than 177 vertices.

Chapter 4

Erdős–Faber–Lovász Conjecture and Inverse Galois Problem

Erdős–Faber–Lovász conjecture



An instance of the Erdős–Faber–Lovász conjecture: a graph formed from four cliques of four vertices each, any two of which intersect in a single vertex, can be four-colored.

In graph theory, the **Erdős–Faber–Lovász conjecture** is an unsolved problem about graph coloring, named after Paul Erdős, Vance Faber, and László Lovász, who formulated it in 1972. It says:

If *k* complete graphs, each having exactly *k* vertices, have the property that every pair of complete graphs has at most one shared vertex, then the union of the graphs can be colored with *k* colors.

Equivalent formulations

Haddad & Tardif (2004) introduced the problem with a story about seating assignment in committees: suppose that, in a university department, there are *k* committees, each consisting of *k* faculty members, and that all committees meet in the same room, which has *k* chairs. Suppose also that at most one person belongs to the intersection of any two committees. Is it possible to assign the committee members to chairs in such a way that each member sits in the same chair for all the different committees to which he or she belongs? In this model of the problem, the faculty members correspond to graph vertices, committees correspond to complete graphs, and chairs correspond to vertex colors.

A *linear hypergraph* is a hypergraph with the property that every two hyperedges have at most one vertex in common. A hypergraph is said to be uniform if all of its hyperedges have the same number of vertices as each other. The *n* cliques of size *n* in the Erdős–Faber–Lovász conjecture may be interpreted as the hyperedges of an *n*-uniform linear hypergraph that has the same vertices as the underlying graph. In this language, the Erdős–Faber–Lovász conjecture states that, given any *n*-uniform linear hypergraph with *n* hyperedges, one may *n*-color the vertices such that each hyperedge has one vertex of each color.

A *simple hypergraph* is a hypergraph in which at most one hyperedge connects any pair of vertices and there are no hyperedges of size at most one. In the graph coloring formulation of the Erdős–Faber–Lovász conjecture, it is safe to remove vertices that belong to a single clique, as their coloring presents no difficulty; once this is done, the hypergraph that has a vertex for each clique, and a hyperedge for each graph vertex, forms a simple hypergraph. And, the hypergraph dual of vertex coloring is edge coloring. Thus, the Erdős–Faber–Lovász conjecture is equivalent to the statement that any simple hypergraph with *n* vertices has chromatic index (edge coloring number) at most *n*.

The graph of the Erdős–Faber–Lovász conjecture may be represented as an intersection graph of sets: to each vertex of the graph, correspond the set of the cliques containing that vertex, and connect any two vertices by an edge whenever their corresponding sets have a nonempty intersection. Using this description of the graph, the conjecture may be restated as follows: if some family of sets has *n* total elements, and any two sets intersect in at most one element, then the intersection graph of the sets may be *n*-colored.

The intersection number of a graph *G* is the minimum number of elements in a family of sets whose intersection graph is *G*, or equivalently the minimum number of vertices in a hypergraph whose line graph is *G*. Klein & Margraf (2003) define the linear intersection number of a graph, similarly, to be the minimum number of vertices in a linear hypergraph whose line graph is *G*. As they observe, the Erdős–Faber–Lovász conjecture is equivalent to the statement that the chromatic number of any graph is at most equal to its linear intersection number.

Haddad & Tardif (2004) present another yet equivalent formulation, in terms of the theory of clones.

History and partial results

Paul Erdős, Vance Faber, and László Lovász formulated the conjecture in 1972. Paul Erdős originally offered US$50 for proving the conjecture in the affirmative, and later raised the reward to US$500.

Chiang & Lawler (1988) proved that the chromatic number of the graphs in the conjecture is at most $3k/2 - 2$, and Kahn (1992) improved this to $k + o(k)$.

Related problems

It is also of interest to consider the chromatic number of graphs formed as the union of *k* cliques of *k* vertices each, without restricting how big the intersections of pairs of cliques can be. In this case, the chromatic number of their union is at most $1 + k\sqrt{k - 1}$, and some graphs formed in this way require this many colors.

A version of the conjecture that uses the fractional chromatic number in place of the chromatic number is known to be true. That is, if a graph *G* is formed as the union of *k* *k*-cliques that intersect pairwise in at most one vertex, then *G* can be *k*-colored.

In the framework of edge coloring simple hypergraphs, Hindman (1981) defines a number *L* from a simple hypergraph as the number of hypergraph vertices that belong to a hyperedge of three or more vertices. He shows that, for any fixed value of *L*, a finite calculation suffices to verify that the conjecture is true for all simple hypergraphs with that value of *L*. Based on this idea, he shows that the conjecture is indeed true for all simple hypergraphs with $L \leq 10$. In the formulation of coloring graphs formed by unions of cliques, Hindman's result shows that the conjecture is true whenever at most ten of the cliques contain a vertex that belongs to three or more cliques. In particular, it is true for $n \leq 10$.

Inverse Galois problem

In Galois theory, the **inverse Galois problem** concerns whether or not every finite group appears as the Galois group of some Galois extension of the rational numbers **Q**. This problem, first posed in the 19th century, is unsolved.

More generally, let *G* be a given finite group, and let *K* be a field. Then the question is this: is there a Galois extension field *L*/*K* such that the Galois group of the extension is isomorphic to *G*? One says that **G is realizable over K** if such a field *L* exists.

Partial results

There is a great deal of detailed information in particular cases. It is known that every finite group is realizable over any function field in one variable over the complex numbers **C**, and more generally over function fields in one variable over any algebraically closed field of characteristic zero. Shafarevich showed that every finite solvable group is realizable over **Q**. It also known that every sporadic group, except possibly the Mathieu group M_{23}, is realizable over **Q**.

Hilbert had shown that this question is related to a rationality question for *G*: if *K* is any extension of **Q**, on which *G* acts as an automorphism group and the invariant field K^G is rational over **Q**, then *G* is realizable over **Q**. Here *rational* means that it is a purely transcendental extension of **Q**, generated by an algebraically independent set. This criterion can for example be used to show that all the symmetric groups are realizable.

Much detailed work has been carried out on the question, which is in no sense solved in general. Some of this is based on constructing *G* geometrically as a Galois covering of the projective line: in algebraic terms, starting with an extension of the field **Q**(*t*) of rational functions in an indeterminate *t*. After that, one applies Hilbert's irreducibility theorem to specialise *t*, in such a way as to preserve the Galois group.

A simple example: cyclic groups

It is possible, using classical results, to construct explicitly a polynomial whose Galois group over **Q** is the cyclic group **Z**/*n***Z** for any positive integer *n*. To do this, choose a prime *p* such that $p \equiv 1 \pmod{n}$; this is possible by Dirichlet's theorem. Let **Q**(μ) be the cyclotomic extension of **Q** generated by μ, where μ is a primitive p^{th} root of unity; the Galois group of **Q**(μ)/**Q** is cyclic of order $p - 1$.

Since *n* divides $p - 1$, the Galois group has a cyclic subgroup *H* of order $(p - 1)/n$. The fundamental theorem of Galois theory implies that the corresponding fixed field

$$F = \mathbf{Q}(\mu)^H$$

has Galois group **Z**/n**Z** over **Q**. By taking appropriate sums of conjugates of μ, following the construction of Gaussian periods, one can find an element α of F that generates F over **Q**, and compute its minimal polynomial.

This method can be extended to cover all finite abelian groups, since every such group appears in fact as a quotient of the Galois group of some cyclotomic extension of **Q**. (This statement should not though be confused with the Kronecker–Weber theorem, which lies significantly deeper.)

Worked example: the cyclic group of order three

For $n = 3$, we may take $p = 7$. Then Gal(**Q**(μ)/**Q**) is cyclic of order six. Let us take the generator η of this group which sends μ to μ^3. We are interested in the subgroup $H = \{1, \eta^3\}$ of order two. Consider the element $\alpha = \mu + \eta^3(\mu)$. By construction, α is fixed by H, and only has three conjugates over **Q**, given by

$$\alpha = \mu + \mu^6, \quad \beta = \eta(\alpha) = \mu^3 + \mu^4, \quad \gamma = \eta^2(\alpha) = \mu^2 + \mu^5.$$

Using the identity $1 + \mu + \mu^2 + ... + \mu^6 = 0$, one finds that

$$\alpha + \beta + \gamma = -1,$$
$$\alpha\beta + \beta\gamma + \gamma\alpha = -2, \text{ and}$$
$$\alpha\beta\gamma = 1.$$

Therefore α is a root of the polynomial

$$(x - \alpha)(x - \beta)(x - \gamma) = x^3 + x^2 - 2x - 1,$$

which consequently has Galois group **Z**/3**Z** over **Q**.

Symmetric and alternating groups

Hilbert showed that all symmetric and alternating groups are represented as Galois groups of polynomials with rational coefficients.

The polynomial $x^n + ax + b$ has discriminant

$$(-1)^{n(n-1)/2}[n^n b^{n-1} + (-1)^{1-n}(n-1)^{n-1}a^n].$$

We take the special case

$$f(x,s) = x^n - sx - s.$$

Substituting a prime integer for s in $f(x,s)$ gives a polynomial (called a **specialization** of $f(x,s)$) that by Eisenstein's criterion is irreducible. Then $f(x,s)$ must be irreducible over **Q**(s). Furthermore, $f(x,s)$ can be written

$$x^n - x/2 - 1/2 - (s - 1/2)(x + 1)$$

and $f(x,1/2)$ can be factored to:

$$(x - 1)(1 + 2x + 2x^2 + ... + 2x^{n-1})/2$$

whose second factor is irreducible by Eisenstein's criterion. We have now shown that the group Gal($f(x,s)/\mathbf{Q}(s)$) is doubly transitive.

We can then find that this Galois group has a transposition. Use the scaling $(1 - n)x = ny$ to get

$$y^n - s((1 - n)/n)^{n-1}y - s((1 - n)/n)^n$$

and with $t = s(1 - n)^{n-1}/n^n$ get

$$g(y,t) = y^n - nty + (n - 1)t$$

which can be arranged to

$$y^n - y - (n - 1)(y - 1) + (t - 1)(-ny + n - 1).$$

Then $g(y,1)$ has 1 as a double zero and its other $n - 2$ zeros are simple, and a transposition in Gal($f(x,s)/\mathbf{Q}(s)$) is implied. Any finite doubly transitive permutation group containing a transposition is a full symmetric group.

Hilbert's irreducibility theorem then implies that an infinite set of rational numbers give specializations of $f(x,t)$ whose Galois groups are S_n over the rational field $\mathbf{Q}$. In fact this set of rational numbers is dense in $\mathbf{Q}$.

The discriminant of $g(y,t)$ equals

$$(-1)^{n(n-1)/2}n^n(n - 1)^{n-1}t^{n-1}(1 - t)$$

and this is not in general a perfect square.

Alternating groups

Solutions for alternating groups must be handled differently for odd and even degrees. Let

$$t = 1 - (-1)^{n(n-1)/2}nu^2$$

Under this substitution the discriminant of $g(y,t)$ equals

$$n^{n+1}(n - 1)^{n-1}t^{n-1}u^2$$

which is a perfect square when n is odd.

In the even case let t be the reciprocal of

$$1 + (-1)^{n(n-1)/2}(n - 1)u^2$$

and $1 - t$ becomes

$$t(-1)^{n(n-1)/2}(n - 1)u^2$$

and the discriminant becomes

$$n^n(n - 1)^n t^n u^2$$

which is a perfect square when n is even.

Again, Hilbert's irreducibility theorem implies the existence of infinitely many specializations whose Galois groups are alternating groups.

Rigid groups

Suppose that $C_1,...,C_n$ are conjugacy classes of a finite group G, and A be the set of n-tuples $(g_1,...g_n)$ of G such that g_i is in C_i and the product $g_1...g_n$ is trivial. Then A is called **rigid** if it is nonempty, G acts transitively on it by conjugation, and each element of A generates G.

Thompson (1984) showed that if a finite group G has a rigid set then it can often be realized as a Galois group over a cyclotomic extension of the rationals. (More precisely, over the cyclotomic extension of the rationals generated by the values of the irreducible characters of G on the conjugacy classes C_i.)

This can be used to show that many finite simple groups, including the monster simple group, are Galois groups of extensions of the rationals.

The prototype for rigidity is the symmetric group S_n, which is generated by an n-cycle and a transposition whose product is an (n-1)-cycle. The construction in the preceding section used these generators to establish a polynomial's Galois group.

A construction with an elliptic modular function

Let n be any integer greater than 1. A lattice Λ in the complex plane with period ratio τ has a sublattice Λ' with period ratio $n\tau$. The latter lattice is one of a finite set of sublattices permuted by the modular group PSL(2,Z), which is based on changes of basis for Λ. Let j denote the elliptic modular function of Klein. Define the polynomial φ_n as the product of the differences $(X\text{-}j(\Lambda_i))$ over the conjugate sublattices. As a polynomial in X, φ_n has coefficients that are polynomials over Q in $j(\tau)$.

On the conjugate lattices, the modular group acts as PGL(2,Z_n). It follows that φ_n has Galois group isomorphic to PGL(2,Z_n) over Q(J(τ)).

Use of Hilbert's irreducibility theorem gives an infinite (and dense) set of rational numbers specializing φ_n to polynomials with Galois group PGL(2,Z_n) over Q. The groups PGL(2,Z_n) include infinitely many non-solvable groups.

Chapter 5

Problems in Loop Theory and Quasigroup Theory

In mathematics, especially abstract algebra, loop theory and quasigroup theory are active research areas with many open problems. As in other areas of mathematics, such problems are often made public at professional conferences and meetings. Many of the problems posed here first appeared in the *Loops (Prague)* conferences and the *Mile High (Denver)* conferences.

Open problems (Moufang loops)

Abelian by cyclic groups resulting in Moufang loops

Let L be a Moufang loop with normal abelian subgroup (associative subloop) M of odd order such that L/M is a cyclic group of order bigger than 3. (i) Is L a group? (ii) If the orders of M and L/M are relatively prime, is L a group?

- *Proposed:* by Michael Kinyon, based on (Chein and Rajah, 2000)
- *Comments:* The assumption that L/M has order bigger than 3 is important, as there is a (commutative) Moufang loop L of order 81 with normal commutative subgroup of order 27.

(Doro's Conjecture) Does a Moufang loop with trivial nucleus necessarily have normal commutant?

Doro conjectured that a Moufang loop with trivial nucleus has normal commutant. Is it true?

- *Proposed:* at Milehigh conference on quasigroups, loops, and nonassocitaive systems, Denver 2005

Embedding CMLs of period 3 into alternative algebras

Conjecture: Any finite commutative Moufang loop of period 3 can be embedded into a commutative alternative algebra.

- *Proposed:* by Alexander Grishkov at Loops '03, Prague 2003

Minimal presentations for loops M(G,2)

For a group G, define $M(G,2)$ on G x C_2 by $(g,0)(h,0) = (gh,0)$, $(g,0)(h,1) = (hg,1)$, $(g,1)(h,0) = (gh^{-1},1)$, $(g,1)(h,1) = (h^{-1}g,0)$. Find a minimal presentation for the Moufang loop $M(G,2)$ with respect to a presentation for G.

- *Proposed:* by Petr Vojtěchovský at Loops '03, Prague 2003
- *Comments:* Chein showed in (Chein, 1974) that $M(G,2)$ is a Moufang loop that is nonassociative if and only if G is nonabelian. Vojtěchovský (Vojtěchovský, 2003) found a minimal presentation for $M(G,2)$ when G is a 2-generated group.

Moufang loops of order p^2q^3 and pq^4

Let p and q be distinct odd primes. If q is not congruent to 1 modulo p, are all Moufang loops of order p^2q^3 groups? What about pq^4?

- *Proposed:* by Andrew Rajah at Loops '99, Prague 1999
- *Comments:* If q is not congruent to 1 modulo p, then all Moufang loops of order pq^3 are groups. It is also known that for an odd prime q all Moufang loops of order q^4 are groups iff $q > 3$.

Moufang loops with non-normal commutant

Is there a Moufang loop whose commutant is not normal?

- *Proposed:* by Andrew Rajah at Loops '03, Prague 2003

(Phillips' problem) Odd order Moufang loop with trivial nucleus

Is there a Moufang loop of odd order with trivial nucleus?

- *Proposed:* by Andrew Rajah at Loops '03, Prague 2003

Presentations for finite simple Moufang loops

Find presentations for all nonassociative finite simple Moufang loops in the variety of Moufang loops.

- *Proposed:* by Petr Vojtěchovský at Loops '03, Prague 2003

- *Comments:* It is shown in (Vojtěchovský, 2003) that every nonassociative finite simple Moufang loop is generated by 3 elements, with explicit formulas for the generators.

Torsion in free Moufang loops

Let MF_n be the free Moufang loop with n generators.

Conjecture: MF_3 is torsion free but MF_n with n>4 is not.

- *Proposed:* by Alexander Grishkov at Loops '03, Prague 2003

Open problems (Bol loops)

Nilpotency degree of the left multiplication group of a left Bol loop

For a left Bol loop Q, find some relation between the nilpotency degree of the left multiplication group of Q and the structure of Q.

- *Proposed:* at Milehigh conference on quasigroups, loops, and nonassociative systems, Denver 2005

Are two Bol loops with similar multiplication tables isomorphic?

Let (Q, *), (Q, +) be two quasigroups defined on the same underlying set Q. The distance d(* , +) is the number of pairs (a,b) in Q x Q such that $a * b \neq a + b$. Call a class of finite quasigroups *quadratic* if there is a positive real number α such that any two quasigroups (Q, *), (Q, +) of order n from the class satisfying $d(* , +) < \alpha n^2$ are isomorphic. Are Moufang loops quadratic? Are Bol loops quadratic?

- *Proposed:* by Aleš Drápal at Loops '99, Prague 1999
- *Comments:* Drápal proved in (Drápal, 1992) that groups are quadratic with α = 1 / 9, and in (Drápal, 2000) that 2-groups are quadratic with α = 1 / 4.

Campbell-Hausdorff series for analytic Bol loops

Determine the Campbell-Hausdorff series for analytic Bol loops.

- *Proposed:* by M. A. Akivis and V. V. Goldberg at Loops '99, Prague 1999
- *Comments:* The problem has been partially solved for local analytic Bruck loops in (Nagy, 2002).

Universally flexible loop that is not middle Bol

A loop is *universally flexible* if every one of its loop isotopes is flexible, that is, satisfies (xy)x=x(yx). A loop is *middle Bol* if every one of its loop isotopes has the antiautomorphic inverse property, that is, satisfies $(xy)^{-1}=y^{-1}x^{-1}$. Is there a finite, universally flexible loop that is not middle Bol?

- *Proposed:* by Michael Kinyon at Loops '03, Prague 2003

Finite simple Bol loop with nontrivial conjugacy classes

Is there a finite simple nonassociative Bol loop with nontrivial conjugacy classes?

- *Proposed:* by Kenneth W. Johnson and Jonathan D. H. Smith at the 2nd Mile High Conference on Nonassociative Mathematics, Denver 2009

Open problems (Nilpotency and solvability)

Niemenmaa's conjecture and related problems

Let Q be a loop whose inner mapping group is nilpotent. Is Q nilpotent? Is Q solvable? If Q is also finite, is the multiplication group of Q solvable?

- *Proposed:* at Loops '03 and '07, Prague 2003 and 2007
- *Comments:* Niemenmaa conjectures that every loop with nilpotent inner mapping group is nilpotent.

Loops with abelian inner mapping group

Let Q be a loop with abelian inner mapping group. Is Q nilpotent? If so, is there a bound on the nilpotency class of Q? In particular, can the nilpotency class of Q be higher than 3?

- *Proposed:* at Loops '07, Prague 2007
- *Comments:* When the inner mapping group Inn(Q) is finite and abelian, then Q is nilpotent (Niemenaa and Kepka). The first question is therefore open only in the infinite case. Call loop Q of *Csörgõ type* if it is nilpotent of class at least 3, and Inn(Q) is abelian. No loop of Csörgõ type of nilpotency class higher than 3 is known. Loops of Csörgõ type exist (Csörgõ, 2004), Buchsteiner loops of Csörgõ type exist (Csörgõ, Drápal and Kinyon, 2007), and Moufang loops of Csörgõ type exist (Nagy and Vojtěchovský, 2007). On the other hand, there are no groups of Csörgõ type (folklore), there are no commutative Moufang loops of Csörgõ type (Bruck), and there are no Moufang p-loops of Csörgõ type for p>3 (Nagy and Vojtěchovský, 2007).

Number of nilpotent loops up to isomorphism

Determine the number of nilpotent loops of order 24 up to isomorphism.

- *Proposed:* by Petr Vojtěchovský at the 2nd Mile High Conference on Nonassociative Mathematics, Denver 2009

Open problems (Quasigroups)

Classification of finite simple paramedial quasigroups

Classify the finite simple paramedial quasigroups.

- *Proposed:* by Jaroslav Ježek and Tomáš Kepka at Loops '03, Prague 2003

Existence of infinite simple paramedial quasigroups

Are there infinite simple paramedial quasigroups?

- *Proposed:* by Jaroslav Ježek and Tomáš Kepka at Loops '03, Prague 2003

Minimal isotopically universal varieties of quasigroups

A variety V of quasigroups is *isotopically universal* if every quasigroup is isotopic to a member of V. Is the variety of loops a minimal isotopically universal variety? Does every isotopically universal variety contain the variety of loops or its parastrophes?

- *Proposed:* by Tomáš Kepka and Petr Němec at Loops '03, Prague 2003
- *Comments:* Every quasigroup is isotopic to a loop, hence the variety of loops is isotopically universal.

Small quasigroups with quasigroup core

Does there exist a quasigroup Q of order q=14, 18, 26 or 42 such that the operation * defined on Q by x * y = y - xy is a quasigroup operation?

- *Proposed:* by Parascovia Syrbu at Loops '03, Prague 2003

Parity of the number of quasigroups up to isomorphism

Let Q(n) be the number of isomorphism classes of quasigroups of order n. Is Q(n) odd for every n?

- *Proposed:* by Douglas Stones at 2nd Mile High Conference on Nonassociative Mathematics, Denver 2009
- *Comments:* The numbers Q(n) are known for 0 < n < 11, and all these are odd.

Uniform construction of latin squares?

Construct a latin square L of order n as follows: Let G = $K_{n,n}$ be the complete bipartite graph with distinct weights on its n^2 edges. Let M_1 be the cheapest matching in G, M_2 the cheapest matching in G with M_1 removed, and so on. Each matching M_i determines a permutation p_i of 1, ..., n. Let L be obtained from G by placing the permutation p_i into row i of L. Does this procedure result in a uniform distribution on the space of latin squares of order n?

- *Proposed:* by Gábor Nagy at the 2nd Mile High Conference on Nonassociative Mathematics, Denver 2009

Open problems (Miscellaneous)

Bound on the size of multiplication groups

For a loop Q, let Mlt(Q) denote the multiplication group of Q, that is, the group generated by all left and right translations. Is | Mlt(Q) | < f(| Q |) for some variety of loops and for some polynomial f?

- *Proposed:* at the Milehigh conference on quasigroups, loops, and nonassociative systems, Denver 2005

Does every alternative loop have 2-sided inverses?

Does every alternative loop, that is, every loop satisfying x(xy)=(xx)y and x(yy)=(xy)y, have 2-sided inverses?

- *Proposed:* by Warren D. Smith
- *Comments:* There are infinite alternative loops without 2-sided inverses, cf. (Ormes and Vojtěchovský, 2007)

Finite simple A-loop

Find a nonassociative finite simple A-loop, if such a loop exists.

- *Proposed:* by Michael Kinyon at Loops '03, Prague 2003
- *Comments:* It is known that if the order of such a loop is odd, the loop must have exponent p for some prime p.

Universality of Osborn loops

A loop is *Osborn* if it satisfies the identity $x((yz)x) = (x^{\lambda}\backslash y)(zx)$. Is every Osborn loop universal, that is, is every isotope of an Osborn loop Osborn? If not, is there a nice identity characterizing universal Osborn loops?

- *Proposed:* by Michael Kinyon at Milehigh conference on quasigroups, loops, and nonassociative systems, Denver 2005
- *Comments:* Moufang and conjugacy closed loops are Osborn.

Solved problems

The following problems were posed as open at various conferences and have since been solved.

Buchsteiner loop that is not conjugacy closed

Is there a Buchsteiner loop that is not conjugacy closed? Is there a finite simple Buchsteiner loop that is not conjugacy closed?

- *Proposed:* at Milehigh conference on quasigroups, loops, and nonassociative systems, Denver 2005
- *Solved by:* Piroska Csörgõ, Aleš Drápal, and Michael Kinyon
- *Solution:* The quotient of a Buchsteiner loop by its nucleus is an abelian group of exponent 4. In particular, no nonassociative Buchsteiner loop can be simple. There exists a Buchsteiner loop of order 128 which is not conjugacy closed.

Classification of Moufang loops of order 64

Classify nonassociative Moufang loops of order 64.

- *Proposed:* at Milehigh conference on quasigroups, loops, and nonassociative systems, Denver 2005
- *Solved by:* Gábor P. Nagy and Petr Vojtěchovský
- *Solution:* There are 4262 nonassociative Moufang loops of order 64. They were found by the method of group modifications in (Vojtěchovský, 2006), and it was shown in (Nagy and Vojtěchovský, 2007) that the list is complete. The latter paper uses a linear-algebraic approach to Moufang loop extensions.

Conjugacy closed loop with nonisomorphic one-sided multiplication groups

Construct a conjugacy closed loop whose left multiplication group is not isomorphic to its right multiplication group.

- *Proposed:* by Aleš Drápal at Loops '03, Prague 2003
- *Solved by:* Aleš Drápal
- *Solution:* There is such a loop of order 9. In can be obtained in the LOOPS package by the command `CCLoop(9,1)`.

Existence of a finite simple Bol loop

Is there a finite simple Bol loop that is not Moufang?

- *Proposed at:* Loops '99, Prague 1999
- *Solved by:* Gábor P. Nagy, 2007.
- *Solution:* A simple Bol loop that is not Moufang will be called *proper*. There are several families of proper simple Bol loops. A smallest proper simple Bol loop is of order 24. There is also a proper simple Bol loop of exponent 2, and a proper simple Bol looop of odd order.
- *Comments:* The above constructions solved additional two open problems:
    - Is there a finite simple Bruck loop that is not Moufang? Yes, since any proper simple Bol loop of exponent 2 is Bruck.
    - Is every Bol loop of odd order solvable? No, as witnessed by any proper simple Bol loop of odd order.

Left Bol loop with trivial right nucleus

Is there a finite non-Moufang left Bol loop with trivial right nucleus?

- *Proposed:* at Milehigh conference on quasigroups, loops, and nonassociative systems, Denver 2005
- *Solved by:* Gábor P. Nagy, 2007
- *Solution:* There is a finite simple left Bol loop of exponent 2 of order 96 with trivial right nucleus. Also, using an exact factorization of the Mathieu group M_{24}, it is possible to construct a non-Moufang simple Bol loop which is a G-loop.

Lagrange property for Moufang loops

Does every finite Moufang loop have the strong Lagrange property?

- *Proposed:* by Orin Chein at Loops '99, Prague 1999
- *Solved by:* Alexander Grishkov and Andrei Zavarnitsine, 2003
- *Solution:* Every finite Moufang loop has the strong Lagrange property (SLP). Here is an outline of the proof:
    - According to (Chein et al. 2003), it suffices to show SLP for nonassociative finite simple Moufang loops (NFSML).
    - It thus suffices to show that the order of a maximal subloop of an NFSML L divides the order of L.
    - A countable class of NFSMLs $M(q)$ was discovered in (Paige 1956), and no other NSFMLs exist by (Liebeck 1987).
    - Grishkov and Zavarnitsine matched maximal subloops of loops $M(q)$ with certain subgroups of groups with triality in (Grishkov and Zavarnitsine, 2003).

Quasivariety of cores of Bol loops

Is the class of cores of Bol loops a quasivariety?

- *Proposed:* by Jonathan D. H. Smith and Alena Vanžurová at Loops '03, Prague 2003
- *Solved by:* Alena Vanžurová, 2004.
- *Solution:* No, the class of cores of Bol loops is not closed under subalgebras. Furthermore, the class of cores of groups is not closed under subalgebras. Here is an outline of the proof:
    - Cores of abelian groups are medial, by (Romanowska and Smith, 1985), (Rozskowska-Lech, 1999).
    - The smallest nonabelian group S_3 has core containing a submagma G of order 4 that is not medial.
    - If G is a core of a Bol loop, it is a core of a Bol loop of order 4, hence a core of an abelian group, a contradiction.

Chapter 6

Hilbert's Problems

Hilbert's problems are a list of twenty-three problems in mathematics published by German mathematician David Hilbert in 1900. The problems were all unsolved at the time, and several of them were very influential for 20th century mathematics. Hilbert presented ten of the problems (1, 2, 6, 7, 8, 13, 16, 19, 21 and 22) at the Paris conference of the International Congress of Mathematicians, speaking on 8 August in the Sorbonne. The complete list of 23 problems was later published, most notably in English translation in 1902 by Mary Frances Winston Newson in the *Bulletin of the American Mathematical Society*.

Nature and influence of the problems

Hilbert's problems ranged greatly in topic and precision. Some of them are propounded precisely enough to enable a clear affirmative/negative answer, like the 3rd problem (probably the easiest for a nonspecialist to understand and also the first to be solved) or the notorious 8th problem (the Riemann hypothesis). There are other problems (notably the 5th) for which experts have traditionally agreed on a single interpretation and a solution to the accepted interpretation has been given, but for which there remain unsolved problems which are so closely related as to be, perhaps, part of what Hilbert intended. Sometimes Hilbert's statements were not precise enough to specify a particular problem but were suggestive enough so that certain problems of more contemporary origin seem to apply, e.g. most modern number theorists would probably see the 9th problem as referring to the (conjectural) Langlands correspondence on representations of the absolute Galois group of a number field. Still other problems (e.g. the 11th and the 16th) concern what are now flourishing mathematical subdisciplines, like the theories of quadratic forms and real algebraic curves.

There are two problems which are not only unresolved but may in fact be unresolvable by modern standards. The 6th problem concerns the axiomatization of physics, a goal that twentieth century developments of physics (including its recognition as a discipline independent from mathematics) seem to render both more remote and less important than in Hilbert's time. Also, the 4th problem concerns the foundations of geometry, in a manner which is now generally judged to be too vague to enable a definitive answer.

Remarkably, the other twenty-one problems have all received significant attention, and late into the twentieth century work on these problems was still considered to be of the greatest importance. Notably, Paul Cohen received the Fields Medal during 1966 for his work on the first problem, and the negative solution of the tenth problem during 1970 by Matiyasevich (completing work of Davis, Putnam and Robinson) generated similar acclaim. Aspects of these problems are still of great interest today.

Ignorabimus

Several of the Hilbert problems have been resolved (or arguably resolved) in ways that would have been profoundly surprising, and even disturbing, to Hilbert himself. Following Frege and Russell, Hilbert sought to define mathematics logically using the method of formal systems, i.e., finitistic proofs from an agreed-upon set of axioms. One of the main goals of Hilbert's program was a finitistic proof of the consistency of the axioms of arithmetic: that is his second problem.

However, Gödel's second incompleteness theorem gives a precise sense in which such a finitistic proof of the consistency of arithmetic is provably impossible. Hilbert lived for 12 years after Gödel's theorem, but he does not seem to have written any formal response to Gödel's work. But doubtless the significance of Gödel's work to mathematics as a whole (and not just to formal logic) was amply and dramatically illustrated by its applicability to one of Hilbert's problems.

Hilbert's tenth problem does not ask whether there exists an algorithm for deciding the solvability of Diophantine equations, but rather asks for the *construction* of such an algorithm: "to devise a process according to which it can be determined in a finite number of operations whether the equation is solvable in rational integers." That this problem was solved by showing that there cannot be any such algorithm would presumably have been very surprising to him.

In discussing his opinion that every mathematical problem should have a solution, Hilbert allows for the possibility that the solution could be a proof that the original problem is impossible. Famously, he stated that the point is to know one way or the other what the solution is, and he believed that we always can know this, that in mathematics there is not any "ignorabimus" (statement that the truth can never be known). It seems unclear whether he would have regarded the solution of the tenth problem as an instance of ignorabimus: what we are proving not to exist is not the integer solution, but (in a certain sense) our own ability to discern whether a solution exists.

On the other hand, the status of the first and second problems is even more complicated: there is not any clear mathematical consensus as to whether the results of Gödel (in the case of the second problem), or Gödel and Cohen (in the case of the first problem) give definitive negative solutions or not, since these solutions apply to a certain formalization of the problems, a formalization which is quite reasonable but is not necessarily the only possible one.

The 24th Problem

Hilbert originally included 24 problems on his list, but decided against including one of them in the published list. The "24th problem" (in proof theory, on a criterion for simplicity and general methods) was rediscovered in Hilbert's original manuscript notes by German historian Rüdiger Thiele in 2000.

Sequels

Since 1900, other mathematicians and mathematical organizations have announced problem lists, but, with few exceptions, these collections have not had nearly as much influence nor generated as much work as Hilbert's problems.

One of the exceptions is furnished by three conjectures made by André Weil during the late 1940s (the Weil conjectures). In the fields of algebraic geometry, number theory and the links between the two, the Weil conjectures were very important. The first of the Weil conjectures was proved by Bernard Dwork, and a completely different proof of the first two conjectures via l-adic cohomology was given by Alexander Grothendieck. The last and deepest of the Weil conjectures (an analogue of the Riemann hypothesis) was proven by Pierre Deligne in what some argue as one of the greatest mathematical achievements of all time. Both Grothendieck and Deligne were awarded the Fields medal. However, the Weil conjectures in their scope are more like a single Hilbert problem, and Weil never intended them as a programme for all mathematics. This is somewhat ironic, since arguably Weil was the mathematician of the 1940s and 1950s who best played the Hilbert role, being conversant with nearly all areas of (theoretical) mathematics and having been important in the development of many of them.

Paul Erdős is legendary for having posed hundreds, if not thousands, of mathematical problems, many of them profound. Erdős often offered monetary rewards; the size of the reward depended on the perceived difficulty of the problem.

The end of the millennium, being also the centennial of Hilbert's announcement of his problems, was a natural occasion to propose "a new set of Hilbert problems." Several mathematicians accepted the challenge, notably Fields Medalist Steve Smale, who responded to a request of Vladimir Arnold by proposing a list of 18 problems. Smale's problems have thus far not received much attention from the media, and it is unclear how much serious attention they are getting from the mathematical community.

At least in the mainstream media, the *de facto* 21st century analogue of Hilbert's problems is the list of seven Millennium Prize Problems chosen during 2000 by the Clay Mathematics Institute. Unlike the Hilbert problems, where the primary award was the admiration of Hilbert in particular and mathematicians in general, each prize problem includes a million dollar bounty. As with the Hilbert problems, one of the prize problems (the Poincaré conjecture) was solved relatively soon after the problems were announced.

Noteworthy for its appearance on the list of Hilbert problems, Smale's list and the list of Millennium Prize Problems — and even, in its geometric guise, in the Weil Conjectures — is the Riemann hypothesis. Notwithstanding some famous recent assaults from major mathematicians of our day, many experts believe that the Riemann hypothesis will be included in problem lists for centuries yet. Hilbert himself declared: "If I were to awaken after having slept for a thousand years, my first question would be: has the Riemann hypothesis been proven?"

During 2008, DARPA announced its own list of 23 problems which it hoped could cause major mathematical breakthroughs, "thereby strengthening the scientific and technological capabilities of DoD".

Summary

Of the cleanly-formulated Hilbert problems, problems 3, 7, 10, 11, 13, 14, 17, 19, 20, and 21 have a resolution that is accepted by consensus. On the other hand, problems 1, 2, 5, 9, 15, 18^{+}, and 22 have solutions that have partial acceptance, but there exists some controversy as to whether it resolves the problem.

The + on 18 denotes that the Kepler conjecture solution is a computer-assisted proof, a notion anachronistic for a Hilbert problem and to some extent controversial because of its lack of verifiability by a human reader in a reasonable time.

That leaves 8 (the Riemann hypothesis) and 12 unresolved, both being in the field of number theory. On this classification 4, 6, 16, and 23 are too vague to ever be described as solved. The withdrawn 24 would also be in this class.

Table of problems

Hilbert's twenty-three problems are:

Problem	Brief explanation	Status	Year Solved
1st	The continuum hypothesis (that is, there is no set whose cardinality is strictly between that of the integers and that of the real numbers)	Proven to be impossible to prove or disprove within the Zermelo–Fraenkel set theory with or without the Axiom of Choice. There is no consensus on whether this is a solution to the problem.	1963
2nd	Prove that the axioms of arithmetic are consistent.	There is no consensus on whether results of Gödel and Gentzen give a solution to the problem as stated by Hilbert. Gödel's second incompleteness theorem, proved in 1931, shows that no proof of its consistency can be carried out within arithmetic itself. Gentzen proved in 1936	1936?

		that the consistency of arithmetic follows from the well-foundedness of the ordinal ε_0.	
3rd	Given any two polyhedra of equal volume, is it always possible to cut the first into finitely many polyhedral pieces which can be reassembled to yield the second?	Resolved. Result: no, proved using Dehn invariants.	1900
4th	Construct all metrics where lines are geodesics.	Too vague to be stated resolved or not.[n 1]	–
5th	Are continuous groups automatically differential groups?	Resolved by Andrew Gleason, depending on how the original statement is interpreted. If, however, it is understood as an equivalent of the Hilbert–Smith conjecture, it is still unsolved.	1953?
6th	Axiomatize all of physics	Unresolved. [n 2]	–
7th	Is a^{b} transcendental, for algebraic $a \neq 0,1$ and irrational algebraic b ?	Resolved. Result: yes, illustrated by Gelfond's theorem or the Gelfond–Schneider theorem.	1935
8th	The Riemann hypothesis ("the real part of any non-trivial zero of the Riemann zeta function is ½") and other prime number problems, among them Goldbach's conjecture and the twin prime conjecture	Unresolved.	–
9th	Find most general law of the reciprocity theorem in any algebraic number field	Partially resolved.[n 3]	–
10th	Find an algorithm to determine whether a given polynomial Diophantine equation with integer coefficients has an integer solution.	Resolved. Result: impossible, Matiyasevich's theorem implies that there is no such algorithm.	1970
11th	Solving quadratic forms with algebraic numerical coefficients.	Partially resolved.	–
12th	Extend the Kronecker–Weber theorem on abelian extensions of the rational numbers to any base number field.	Unresolved.	–
13th	Solve all 7-th degree equations using continuous functions of two parameters.	Resolved. The problem was solved affirmatively by Vladimir Arnold based on work by Andrei Kolmogorov. [n 5]	1957
14th	Is the ring of invariants of an algebraic group acting on a polynomial ring always finitely generated?	Resolved. Result: no, counterexample was constructed by Masayoshi Nagata.	1959
15th	Rigorous foundation of Schubert's enumerative calculus.	Partially resolved.	–
16th	Describe relative positions of ovals originating from a real algebraic curve and as limit cycles of a polynomial vector field on the plane.	Unresolved.	–

17th	Expression of definite rational function as quotient of sums of squares	Resolved. Result: An upper limit was established for the number of square terms necessary.	1927
18th	(a) Is there a polyhedron which admits only an anisohedral tiling in three dimensions? (b) What is the densest sphere packing?	(a) Resolved. Result: yes (by Karl Reinhardt). (b) Resolved by computer-assisted proof. Result: cubic close packing and hexagonal close packing, both of which have a density of approximately 74%.[n 6]	(a) 1928 (b) 1998
19th	Are the solutions of Lagrangians always analytic?	Resolved. Result: yes, proven by Ennio de Giorgi and, independently and using different methods, by John Forbes Nash.	1957
20th	Do all variational problems with certain boundary conditions have solutions?	Resolved. A significant topic of research throughout the 20th century, culminating in solutions for the non-linear case.	–
21st	Proof of the existence of linear differential equations having a prescribed monodromic group	Resolved. Result: Yes or no, depending on more exact formulations of the problem.	–
22nd	Uniformization of analytic relations by means of automorphic functions	Resolved.	–
23rd	Further development of the calculus of variations	Unresolved.	–

Chapter 7

Hadamard's Maximal Determinant Problem

Hadamard's maximal determinant problem, named after Jacques Hadamard, asks for the largest possible determinant of a matrix with elements restricted to the set {1,−1}. The question for matrices with elements restricted to the set {0,1} is equivalent: the maximal determinant of a {1,−1} matrix of size n is 2^{n-1} times the maximal determinant of a {0,1} matrix of size $n-1$. The problem was first posed by Hadamard in the 1893 paper in which he presented his famous determinant bound, but the problem remains unsolved for matrices of general size. Hadamard's bound implies that {1, −1}-matrices of size n have determinant at most $n^{n/2}$. Matrices whose determinants attain the bound are now known as Hadamard matrices, although examples had earlier been found by Sylvester. Any two rows of an $n \times n$ Hadamard matrix are orthogonal, which is impossible for a {1, −1} matrix when n is an odd number greater than 1. When $n \equiv 2 \pmod 4$, two rows that are both orthogonal to a third row cannot be orthogonal to each other. Together, these statements imply that an $n \times n$ Hadamard matrix can exist only if $n = 1$, 2, or a multiple of 4. Hadamard matrices have been well studied, but it is not known whether a Hadamard matrix of size $4k$ exists for every $k \geq 1$. The smallest n for which an $n \times n$ Hadamard matrix is not known to exist is 668.

Matrix sizes n for which $n \equiv 1$, 2, or 3 (mod 4) have received less attention. Nevertheless, some results are known, including:

- tighter bounds for $n \equiv 1$, 2, or 3 (mod 4) (These bounds, due to Barba, Ehlich, and Wojtas, are known not to be always attainable.),
- a few infinite sequences of matrices attaining the bounds for $n \equiv 1$ or 2 (mod 4),
- a number of matrices attaining the bounds for specific $n \equiv 1$ or 2 (mod 4),
- a number of matrices not attaining the bounds for specific $n \equiv 1$ or 3 (mod 4), but that have been proved by exhaustive computation to have maximal determinant.

The design of experiments in statistics makes use of {1, −1} matrices X (not necessarily square) for which the information matrix $X^{\mathrm{T}}X$ has maximal determinant. (The notation X^{T}

denotes the transpose of X.) Such matrices are known as D-optimal designs. If X is a square matrix, it is known as a saturated D-optimal design.

Equivalence and normalization of {1, −1} matrices

Any of the following operations, when performed on a $\{1, -1\}$ matrix R, changes the determinant of R only by a minus sign:

- Negation of a row.
- Negation of a column.
- Interchange of two rows.
- Interchange of two columns.

Two $\{1,-1\}$ matrices, R_1 and R_2, are considered **equivalent** if R_1 can be converted to R_2 by some sequence of the above operations. The determinants of equivalent matrices are equal, except possibly for a sign change, and it is often convenient to standardize R by means of negations and permutations of rows and columns. A $\{1, -1\}$ matrix is **normalized** if all elements in its first row and column equal 1. When the size of a matrix is odd, it is sometimes useful to use a different normalization in which every row and column contains an even number of elements 1 and an odd number of elements −1. Either of these normalizations can be accomplished using the first two operations.

Connection of the maximal determinant problems for {1, −1} and {0, 1} matrices

There is a one-to-one map from the set of normalized $n\times n$ $\{1, -1\}$ matrices to the set of $(n-1)\times(n\text{-}1)$ $\{0, 1\}$ matrices under which the magnitude of the determinant is reduced by a factor of 2^{1-n}. This map consists of the following steps.

1. Subtract row 1 of the $\{1, -1\}$ matrix from rows 2 through n. (This does not change the determinant.)
2. Extract the $(n-1)\times(n-1)$ submatrix consisting of rows 2 through n and columns 2 through n. This matrix has elements 0 and −2. (The determinant of this submatrix is the same as that of the original matrix, as can be seen by performing a cofactor expansion on column 1 of the matrix obtained in Step 1.)
3. Divide the submatrix by −2 to obtain a $\{0, 1\}$ matrix. (This multiplies the determinant by $(-2)^{1\text{-}n}$.)

Example:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \rightarrow \left[\begin{array}{c|ccc} 1 & 1 & 1 & 1 \\ \hline 0 & -2 & -2 & 0 \\ 0 & 0 & -2 & -2 \\ 0 & -2 & 0 & -2 \end{array}\right] \rightarrow \begin{bmatrix} -2 & -2 & 0 \\ 0 & -2 & -2 \\ -2 & 0 & -2 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

In this example, the original matrix has determinant −16 and its image has determinant $2 = -16 \cdot (-2)^{-3}$.

Since the determinant of a {0, 1} matrix is an integer, the determinant of an $n \times n$ {1, −1} matrix is an integer multiple of 2^{n-1}.

Upper bounds on the maximal determinant

Gram matrix

Let R be an n by n {1, −1} matrix. The **Gram matrix** of R is defined to be the matrix $G = RR^{\mathrm{T}}$. From this definition it follows that G

1. is an integer matrix,
2. is symmetric,
3. is positive-semidefinite,
4. has constant diagonal whose value equals n.

Negating rows of R or applying a permutation to them results in the the same negations and permutation being applied both to the rows, and to the corresponding columns, of G. We may also define the matrix $G'=R^{\mathrm{T}}R$. The matrix G is the usual Gram matrix of a set of vectors, derived from the set of rows of R, while G' is the Gram matrix derived from the set of columns of R. A matrix R for which $G = G'$ is a normal matrix. Every known maximal-determinant matrix is equivalent to a normal matrix, but it is not known whether this is always the case.

Hadamard's bound (for all *n*)

Hadamard's bound can be derived by noting that $|\det R| = (\det G)^{1/2} \leq (\det nI)^{1/2} = n^{n/2}$, which is a consequence of the observation that nI, where I is the n by n identity matrix, is the unique matrix of maximal determinant among matrices satisfying properties 1–4. That det R must be an integer multiple of 2^{n-1} can be used to provide another demonstration that Hadamard's bound is not always attainable. When n is odd, the bound $n^{n/2}$ is either non-integer or odd, and is therefore unattainable except when $n = 1$. When $n = 2k$ with k odd, the highest power of 2 dividing Hadamard's bound is 2^k which is less than 2^{n-1} unless $n = 2$. Therefore Hadamard's bound is unattainable unless $n = 1$, 2, or a multiple of 4.

Barba's bound for *n* odd

When n is odd, property 1 for Gram matrices can be strengthened to

1. G is an odd-integer matrix.

This allows a sharper upper bound to be derived: $|\det R| = (\det G)^{1/2} \leq (\det (n\text{-}1)I+J)^{1/2} = (2n-1)^{1/2}(n-1)^{(n-1)/2}$, where J is the all-one matrix. Here $(n\text{-}1)I+J$ is the maximal-determinant matrix satisfying the modified property 1 and properties 2–4. It is unique up to multiplication of any set of rows and the corresponding set of columns by −1. The bound is not attainable unless $2n-1$ is a perfect square, and is therefore never attainable when $n \equiv 3 \pmod 4$.

The Ehlich–Wojtas bound for $n \equiv 2 \pmod 4$

When n is even, the set of rows of R can be partitioned into two subsets.

- Rows of **even type** contain an even number of elements 1 and an even number of elements −1.
- Rows of **odd type** contain an odd number of elements 1 and an odd number of elements −1.

The dot product of two rows of the same type is congruent to $n \pmod 4$; the dot product of two rows of opposite type is congruent to $n+2 \pmod 4$. When $n \equiv 2 \pmod 4$, this implies that, by permuting rows of R, we may assume the **standard form**,

$$G = \begin{bmatrix} A & B \\ B^{\mathrm{T}} & D \end{bmatrix},$$

where A and D are symmetric integer matrices whose elements are congruent to 2 (mod 4) and B is a matrix whose elements are congruent to 0 (mod 4). In 1964, Ehlich and Wojtas independently showed that in the maximal determinant matrix of this form, A and D are both of size $n/2$ and equal to $(n-2)I+2J$ while B is the zero matrix. This optimal form is unique up to multiplication of any set of rows and the corresponding set of columns by −1 and to simultaneous application of a permutation to rows and columns. This implies the bound $\det R \leq (2n-2)(n-2)^{(n-2)/2}$. Ehlich showed that if R attains the bound, and if the rows and columns of R are permuted so that both $G = RR^{\mathrm{T}}$ and $G' = R^{\mathrm{T}}R$ have the standard form and are suitably normalized, then we may write

$$R = \begin{bmatrix} W & X \\ Y & Z \end{bmatrix}$$

where W, X, Y, and Z are $(n/2)\times(n/2)$ matrices with constant row and column sums w, x, y, and z that satisfy $z = -w$, $y = x$, and $w^2+x^2 = 2n-2$. Hence the Ehlich–Wojtas bound is not attainable unless $2n-2$ is expressible as the sum of two squares.

Ehlich's bound for $n \equiv 3 \pmod 4$

When n is odd, then by using the freedom to multiply rows by −1, one may impose the condition that each row of R contain an even number of elements 1 and an odd number of elements −1. It can be shown that, if this normalization is assumed, then property 1 of G may be strengthened to

1. *G* is a matrix with integer elements congruent to *n* (mod 4).

When $n \equiv 1 \pmod 4$, the optimal form of Barba satisfies this stronger property, but when $n \equiv 3 \pmod 4$, it does not. This means that the bound can be sharpened in the latter case. Ehlich showed that when $n \equiv 3 \pmod 4$, the strengthened property 1 implies that the maximal-determinant form of *G* can be written as $B-J$ where *J* is the all-one matrix and *B* is a block-diagonal matrix whose diagonal blocks are of the form $(n\text{-}3)I+4J$. Moreover, he showed that in the optimal form, the number of blocks, *s*, depends on *n* as shown in the table below, and that each block either has size *r* or size *r+1* where $r = \lfloor n/s \rfloor$.

n	*s*
3	3
7	5
11	5 or 6
15 − 59	6
≥ 63	7

Except for *n*=11 where there are two possibilities, the optimal form is unique up to multiplication of any set of rows and the corresponding set of columns by −1 and to simultaneous application of a permutation to rows and columns. This optimal form leads to the bound

$$\det R \leq (n-3)^{(n-s)/2}(n-3+4r)^{u/2}(n+1+4r)^{v/2}\left[1-\frac{ur}{n-3+4r}-\frac{v(r+1)}{n+1+4r}\right]^{1/2},$$

where $v = n-rs$ is the number of blocks of size r+1 and $u = s-v$ is the number of blocks of size *r*. Cohn analyzed the bound and determined that, apart from $n = 3$, it is an integer only for $n = 112t^2 \pm 28t+7$ for some positive integer *t*. Tamura derived additional restrictions on the attainability of the bound using the Hasse-Minkowski theorem on the rational equivalence of quadratic forms, and showed that the smallest $n > 3$ for which Ehlich's bound is conceivably attainable is 511.

Maximal determinants up to size 18

The maximal determinants of {1, −1} matrices up to size $n = 18$ are given in the following table. Size 19 is the smallest open case. In the table, $D(n)$ represents the maximal determinant divided by 2^{n-1}. Equivalently, $D(n)$ represents the maximal determinant of a {0, 1} matrix of size $n-1$.

n	*D*(*n*)	Notes
1	1	Hadamard matrix
2	1	Hadamard matrix
3	1	Attains Ehlich bound

4	2	Hadamard matrix
5	3	Attains Barba bound; circulant matrix
6	5	Attains Ehlich–Wojtas bound
7	9	98.20% of Ehlich bound
8	32	Hadamard matrix
9	56	84.89% of Barba bound
10	144	Attains Ehlich–Wojtas bound
11	320	94.49% of Ehlich bound; three non-equivalent matrices
12	1458	Hadamard matrix
13	3645	Attains Barba bound; maximal-determinant matrix is {1,−1} incidence matrix of projective plane of order 3
14	9477	Attains Ehlich–Wojtas bound
15	25515	97.07% of Ehlich bound
16	131072	Hadamard matrix; five non-equivalent matrices
17	327680	87.04% of Barba bound; three non-equivalent matrices
18	1114112	Attains Ehlich–Wojtas bound; three non-equivalent matrices

Chapter 8

Gauss Circle Problem and Inscribed Square Problem

Gauss circle problem

In mathematics, the **Gauss circle problem** is the problem of determining how many lattice points there are in a circle centred at the origin and with radius *r*. The first progress on a solution was made by Carl Friedrich Gauss, hence its name.

The problem

Consider a circle in $\mathbf{R}^2$ with centre at the origin and radius $r \geq 0$. Gauss' circle problem asks how many points there are inside this circle of the form (m,n) where *m* and *n* are both integers. Since the equation of this circle is given in Cartesian coordinates by $x^2 + y^2 = r^2$, the question is equivalently asking how many pairs of integers *m* and *n* there are such that

$$m^2 + n^2 \leq r^2.$$

If the answer for a given *r* is denoted by $N(r)$ then the following list shows the first few values of $N(r)$ for *r* an integer between 0 and 10:

1, 5, 13, 29, 49, 81, 113, 149, 197, 253, 317 (sequence A000328 in OEIS).

Bounds on a solution and conjecture

The area inside a circle of radius *r* is given by πr^2, and since a square of area 1 in $\mathbf{R}^2$ contains one integer point, the expected answer to the problem could be about πr^2. In fact it should be slightly higher than this, since circles are more efficient at enclosing space than squares. So in fact it should be expected that

$$N(r) = \pi r^2 + E(r)$$

for some error term *E*(*r*). Finding a correct upper bound for *E*(*r*) is thus the form the problem has taken.

Gauss managed to prove that

$$E(r) \leq 2\sqrt{2}\pi r.$$

Hardy and, independently, Landau found a lower bound by showing that

$$E(r) \neq o\left(r^{1/2}(\log r)^{1/4}\right),$$

using the little o-notation. It is conjectured that the correct bound is

$$E(r) = O\left(r^{1/2+\varepsilon}\right).$$

Writing $|E(r)| \leq Cr^t$, the current bounds on *t* are

$$\frac{1}{2} < t \leq \frac{131}{208} = 0.6298\ldots,$$

with the lower bound from Hardy and Landau in 1915, and the upper bound proved by Huxley in 2000.

In 2007, Sylvain Cappell and Julius Shaneson deposited a paper in the arXiv claiming to prove the bound of $O(r^{1/2+\varepsilon})$. Kannan Soundararajan is reported to have found a mistake in the proof.

Exact forms

The value of *N*(*r*) can be given by several series. In terms of a sum involving the floor function it can be expressed as:

$$N(r) = 1 + 4\sum_{i=0}^{\infty}\left(\left\lfloor\frac{r^2}{4i+1}\right\rfloor - \left\lfloor\frac{r^2}{4i+3}\right\rfloor\right).$$

A much simpler sum appears if the sum of squares function $r_2(n)$ is defined as the number of ways of writing the number *n* as the sum of two squares. Then

$$N(r) = \sum_{n=0}^{r^2} r_2(n).$$

Generalisations

Although the original problem asks for integer lattice points in a circle, there is no reason not to consider other shapes or conics, indeed Dirichlet's divisor problem is the equivalent problem where the circle is replaced by the rectangular hyperbola. Similarly one could extend the question from two dimensions to higher dimensions (A00605 in 3, A055410 in 4, A055411 in 5, A005412 etc. in 6 and higher), and ask for integer points within a sphere or other objects. If one ignores the geometry and merely considers the problem an algebraic one of Diophantine inequalities then there one could increase the exponents appearing in the problem from squares to cubes, or higher.

The primitive circle problem

Another generalisation is to calculate the number of coprime integer solutions *m*, *n* to the equation

$$m^2 + n^2 \leq r^2.$$

This problem is known as the **primitive circle problem**, as it involves searching for primitive solutions to the original circle problem. If the number of such solutions is denoted *V*(*r*) then the values of *V*(*r*) for *r* taking integer values from 0 on are (A175341)

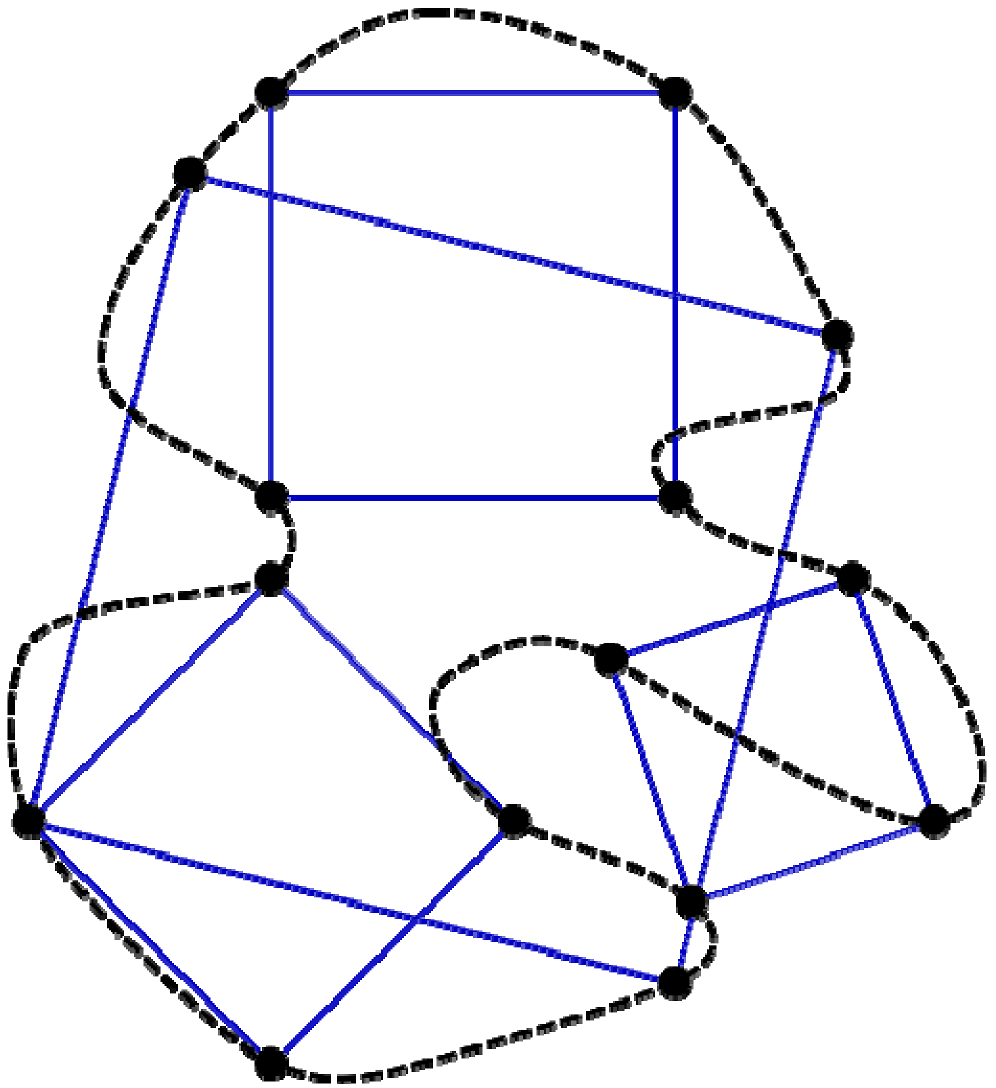
$$0, 4, 8, 16, 32, 48, 72, 88, 120, 152, \ldots$$

Using the same ideas as the usual Gauss circle problem and the fact that the probability that two integers are coprime is $6/\pi^2$, it is relatively straightforward to show that

$$V(r) = \frac{6}{\pi}r^2 + O(r^{1+\varepsilon}).$$

As with the usual circle problem, the problematic part of the primitive circle problem is reducing the exponent in the error term. At present the best known exponent is $221/304 + \varepsilon$ if one assumes the Riemann hypothesis. On the other hand no exponent less than one has been proved unconditionally.

Inscribed square problem



Example black dashed curve goes through corners of several blue squares.

The **inscribed square problem** is an unsolved question in geometry: *Does every plane simple curve contain all four vertices of some square?* This is known to be true if the curve is convex or piecewise smooth and in other special cases. The problem was proposed by Otto Toeplitz in 1911. Some early positive results were obtained by Arnold Emch and Lev Shnirelman. As of 2007, the general case remains open.

Overview

Let C be a Jordan curve. A polygon P is **inscribed in *C*** if all vertices of P belong to C. The **inscribed square problem** asks:

> *Does every Jordan curve admit an inscribed square?*

It is *not* required that the vertices of the square appear along the curve in any particular order.

Some figures, such circles and squares, admit infinitely many inscribed squares. If C is an obtuse triangle then it admits exactly one inscribed square.

The most encompassing result to date is due to Stromquist, who proved that every *local monotone* plane simple curve admits an inscribed square. The condition is that for any point p, the curve C can be locally represented as a graph of a function $y = f(x)$. More precisely, for any point p on C there is a neighborhood $U(p)$ such that no chord of C in this neighborhood is parallel to a fixed direction $n(p)$ (the direction of the "y-axis"). Locally monotone curves includes all closed convex curves and all piecewise-C^1 curves without cusps.

The affirmative answer is also known for centrally symmetric curves.

Variants and generalizations

One may ask whether other shapes can be inscribed into an arbitrary Jordan curve. It is known that for any triangle T and Jordan curve C, there is a triangle similar to T and inscribed in C. Moreover, the set of the vertices of such triangles is dense in C. In particular, there is always an inscribed equilateral triangle. It is also known that any Jordan curve admits an inscribed rectangle.

Some generalizations of the inscribed square problem consider inscribed polygons for curves and even more general continua in higher dimensional Euclidean spaces. For example, Stromquist proved that every continuous closed curve C in $\mathbf{R}^n$ satisfying "Condition A" that no two chords of C in a suitable neighborhood of any point are perpendicular admits an inscribed quadrilateral with equal sides and equal diagonals. This class of curves includes all C^2 curves. Nielsen and Wright proved that any symmetric continuum K in $\mathbf{R}^n$ contains many inscribed rectangles. H.W. Guggenheimer proved that every hypersurface C^3-diffeomorphic to the sphere S^{n-1} contains 2^n vertices of a regular Euclidean n-cube.

Chapter 9

Burnside's Problem

The **Burnside problem**, posed by William Burnside in 1902 and one of the oldest and most influential questions in group theory, asks whether a finitely generated group in which every element has finite order must necessarily be a finite group. In plain language, if by looking at individual elements of a group we suspect that the whole group is finite, must it indeed be true? The problem has many variants that differ in the additional conditions imposed on the orders of the group elements.

Brief history

Initial work pointed towards the affirmative answer. For example, if a group *G* is generated by *m* elements and the order of each element of *G* is a divisor of 4, then *G* is finite. Moreover, A. I. Kostrikin (for the case of a prime exponent) and Efim Zelmanov (in general) proved that, among the finite groups with given number of generators and exponent, there exists a largest one. Issai Schur showed that any finitely generated periodic group that was a subgroup of the group of invertible n x n complex matrices was finite; he used this theorem to prove the Jordan–Schur theorem.

Nevertheless, the general answer to Burnside's problem turned out to be negative. In 1964, Golod and Shafarevich constructed an infinite group of Burnside type without assuming that all elements have uniformly bounded order. In 1968, Pyotr Novikov and Sergei Adian's supplied a negative solution to the bounded exponent problem for all odd exponents larger than 4381. In 1982, A. Yu. Ol'shanskii found some striking counterexamples for sufficiently large odd exponents (greater than 10^{10}), and supplied a considerably simpler proof based on geometric ideas.

The case of even exponents turned out to be much harder to settle. In 1992 S. V. Ivanov announced the negative solution for sufficiently large even exponents divisible by a large power of 2 (detailed proofs were published in 1994 and occupied some 300 pages). Later joint work of Ol'shanskii and Ivanov established a negative solution to an analogue of Burnside's problem for hyperbolic groups, provided the exponent is sufficiently large. By contrast, when the exponent is small and different from 2,3,4 and 6, very little is known.

General Burnside problem

A group G is called periodic if every element has finite order; in other words, for each g in G, there exists some positive integer n such that $g^n = 1$. Clearly, every finite group is periodic. There exist easily defined groups such as the p^∞-group which are infinite periodic groups; but the latter group cannot be finitely generated.

The **general Burnside problem** can be posed as follows:

> If G is a periodic group, and G is finitely generated, then must G necessarily be a finite group?

This question was answered in the negative in 1964 by Evgeny Golod and Igor Shafarevich, who gave an example of an infinite p-group that is finitely generated. However, the orders of the elements of this group are not *a priori* bounded by a single constant.

Bounded Burnside problem

Part of the difficulty with the general Burnside problem is that the requirements of being finitely generated and periodic give very little information about the possible structure of a group. Consider a periodic group G with the additional property that there exists a single integer n such that for all g in G, $g^n = 1$. A group with this property is said to be *periodic with bounded exponent n*, or just a *group with exponent n*. Burnside problem for groups with bounded exponent asks:

> If G is a finitely generated group with exponent n, is G necessarily finite?

It turns out that this problem can be restated as a question about the finiteness of groups in a particular family. The **free Burnside group** of rank m and exponent n, denoted $\mathrm{B}(m, n)$, is a group with m distinguished generators $x_1,\ldots,x_m$ in which the identity $x^n = 1$ holds for all elements x, and which is the "largest" group satisfying these requirements. More precisely, the characteristic property of $\mathrm{B}(m, n)$ is that, given any group G with m generators $g_1,\ldots,g_m$ and of exponent n, there is a unique homomorphism from $\mathrm{B}(m, n)$ to G that maps the ith generator x_i of $\mathrm{B}(m, n)$ into the ith generator g_i of G. In the language of group presentations, free Burnside group $\mathrm{B}(m, n)$ has m generators $x_1,\ldots,x_m$ and the relations $x^n = 1$ for each word x in $x_1,\ldots,x_m$, and any group G with m generators of exponent n is obtained from it by imposing additional relations. The existence of the free Burnside group and its uniqueness up to an isomorphism are established by standard techniques of group theory. Thus if G is any finitely generated group of exponent n, then G a homomorphic image of $\mathrm{B}(m, n)$, where m is the number of generators of G. Burnside's problem can now be restated as follows:

> For which positive integers m, n is the free Burnside group $\mathrm{B}(m,n)$ finite?

The full solution to Burnside's problem in this form is not known. Burnside considered some easy cases in his original paper:

- For $m = 1$ and any positive n, B(1, n) is the cyclic group of order n.
- B(m, 2) is the direct product of m copies of the cyclic group of order 2. The key step is to observe that the identities $a^2 = b^2 = (ab)^2 = 1$ together imply that $ab = ba$, so that a free Burnside group of exponent two is necessarily abelian.

The following additional results are known (Burnside, Sanov, M. Hall):

- B(m,3), B(m,4), and B(m,6) are finite for all m.

The particular case of B(2, 5) remains open: as of 2005, it was not known whether this group is finite.

The breakthrough in Burnside's problem was achieved by Pyotr Novikov and Sergei Adian in 1968. Using a complicated combinatorial argument, they demonstrated that for every odd number n with $n > 4381$, there exist infinite, finitely generated groups of exponent n. Adian later improved the bound on the odd exponent to 665. The case of even exponent turned out to be considerably more difficult. It was only in 1992 that Sergei Vasilievich Ivanov was able to prove an analogue of Novikov–Adian theorem: for any $m > 1$ and an even $n \geq 2^{48}$, n divisible by 2^9, the group B(m, n) is infinite. Both Novikov–Adian and Ivanov established considerably more precise results on the structure of the free Burnside groups. In the case of the odd exponent, all finite subgroups of the free Burnside groups were shown to be cyclic groups. In the even exponent case, each finite subgroup is contained in a product of two dihedral groups, and there exist non-cyclic finite subgroups. Moreover, the word and conjugacy problems were shown to be effectively solvable in B(m, n) both for the cases of odd and even exponents n.

A famous class of counterexamples to Burnside's problem is formed by finitely generated non-cyclic infinite groups in which every nontrivial proper subgroup is a finite cyclic group, the so-called Tarski Monsters. First examples of such groups were constructed by A. Yu. Ol'shanskii in 1979 using geometric methods, thus affirmatively solving O. Yu. Schmidt's problem. In 1982 Ol'shanskii was able to strengthen his results to establish existence, for any sufficiently large prime number p (one can take $p > 10^{75}$) of a finitely generated infinite group in which every nontrivial proper subgroup is a cyclic group of order p. In a paper published in 1996, Ivanov and Ol'shanskii solved an analogue of Burnside's problem in an arbitrary hyperbolic group for sufficiently large exponents.

Restricted Burnside problem

The **restricted Burnside problem**, formulated in the 1930s, asks another, related, question:

> If it is known that a group G with m generators and exponent n is finite, can one conclude that the order of G is bounded by some constant depending only on m

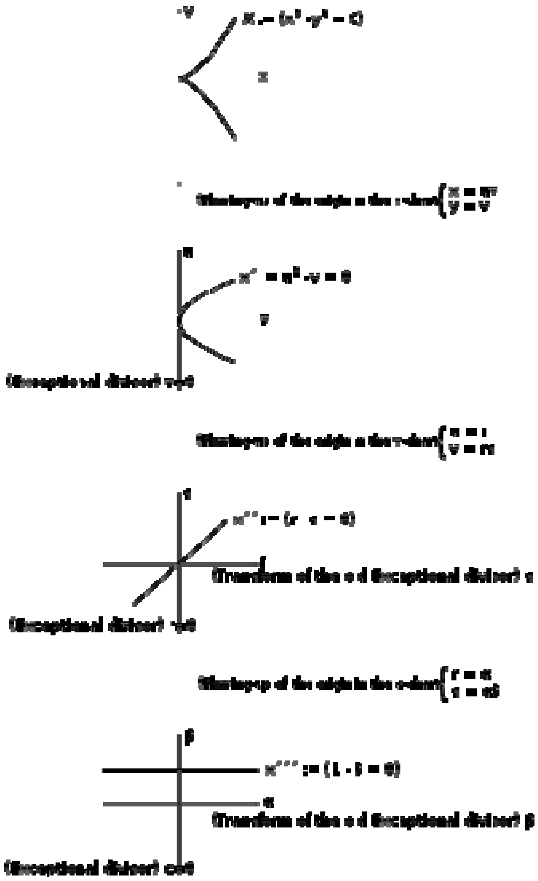
> and n? Equivalently, are there only finitely many *finite* groups with m generators of exponent n, up to isomorphism?

This variant of the Burnside problem can also be stated in terms of certain universal groups with m generators and exponent n. By basic results of group theory, the intersection of two subgroups of finite index in any group is itself a subgroup of finite index. Let M be the intersection of all subgroups of the free Burnside group B(m, n) which have finite index, then M is a normal subgroup of B(m, n) (otherwise, there exists a subgroup $g^{-1}Mg$ with finite index containing elements not in M). One can therefore define a group $B_0(m,n)$ to be the factor group B(m,n)/M. Every finite group of exponent n with m generators is a homomorphic image of $B_0(m,n)$. The restricted Burnside problem then asks whether $B_0(m,n)$ is a finite group.

In the case of the prime exponent p, this problem was extensively studied by A. I. Kostrikin during the 1950s, prior to the negative solution of the general Burnside problem. His solution, establishing the finiteness of $B_0(m,p)$, used a relation with deep questions about identities in Lie algebras in finite characteristic. The case of arbitrary exponent has been completely settled in the affirmative by Efim Zelmanov, who was awarded the Fields Medal in 1994 for his work.

Chapter 10

Resolution of Singularities


y
x
(Exceptional divisor)
(Transform of the old Exceptional divisor)
(Exceptional divisor)
β
α
(Transform of the old Exceptional divisor)
(Exceptional divisor)

Strong desingularization of $X := (x^2 - y^3 = 0) \subset W := \mathbf{R}^2$.Observe that the resolution does not stop after the first blowing-up, when the strict transform is smooth, but when it is simple normal crossings with the exceptional divisors.

In algebraic geometry, the problem of **resolution of singularities** asks whether every algebraic variety *V* has a resolution, a non-singular variety *W* with a proper birational map $W \to V$. For varieties over fields of characteristic 0 this was proved in Hironaka (1964), while for varieties over fields of characteristic p it is an open problem in dimensions at least 4.

Definitions

Originally the problem of resolution of singularities was to find a nonsingular model for the function field of a variety X, in other words a complete non-singular variety X' with the same function field. In practice it is more convenient to ask for a different condition as follows: a variety X has a **resolution of singularities** if we can find a non-singular variety X' and a proper birational map from X' to X. The condition that the map is proper is needed to exclude trivial solutions, such as taking X' to be the subvariety of non-singular points of X.

More generally, it is often useful to resolve the singularities of a variety X embedded into a larger variety W. Suppose we have a closed embedding of X into a regular variety W. A **strong desingularization** of X is given by a proper birational morphism from a regular variety W' to W subject to some of the following conditions (the exact choice of conditions depends on the author):

1. The strict transform X' of X is regular, and transverse to the exceptional locus of the resolution morphism (so in particular it resolves the singularities of X).
2. The map from the strict transform of X to X is an isomorphism away from the singular points of X.
3. W' is constructed by repeatedly blowing up regular closed subvarieties, transverse to the exceptional locus of the previous blowings up.
4. The construction of W' is functorial for *smooth* morphisms to W and embeddings of W into a larger variety. (It cannot be made functorial for all (not necessarily smooth) morphisms in any reasonable way.)
5. The morphism from X' to X does not depend on the embedding of X in W. Or in general, the sequence of blowings up is functorial with respect to smooth morphisms.

Hironaka showed that there is a strong desingularization satisfying the first three conditions above whenever X is defined over a field of characteristic 0, and his construction was improved by several authors (see below) so that it satisfies all conditions above.

Resolution of singularities of curves

Every algebraic curve has a unique nonsingular projective model, which means that all resolution methods are essentially the same because they all construct this model. In higher dimensions this is no longer true: varieties can have many different nonsingular projective models.

Kollar (2007) lists about 20 ways of proving resolution of singularities of curves.

Newton's method

Resolution of singularities of curves was essentially first proved by Newton (1676), who showed the existence of Puiseux series for a curve from which resolution follows easily.

Riemann's method

Riemann constructed a smooth Riemann surface from the function field of a complex algebraic curve, which gives a resolution of its singularities. This can be done over more general fields by using the set of discrete valuation rings of the field as a substitute for the Riemann surface.

Albanese's method

Albanese's method consists of taking a curve that spans a projective space of sufficiently large dimension (more than twice the degree of the curve) and repeatedly projecting down from singular points to projective spaces of smaller dimension. This method extends to higher dimensional varieties, and shows that any *n*-dimensional variety has a projective model with singularities of multiplicity at most *n*!, which when *n* is one means that there are no singular points.

Normalization

A one step method of resolving singularities of a curve is to take the normalization of the curve. Normalization removes all singularities in codimension 1, so it works for curves but not in higher dimensions.

Valuation rings

Another one-step method of resolving singularities of a curve is to take a space of valuation rings of the function field of the curve. This space can be made into a nonsingular projective curve birational to the original curve. This only gives a weak resolution, because there is in general no morphism from this nonsingular projective curve to the original curve.

Blowing up

Repeatedly blowing up the singular points of a curve will eventually resolve the singularities. The main task with this method is to find a way to measure the complexity of a singularity and to show that blowing up improves this measure. There are many ways to do this. For example, one can use the arithmetic genus of the curve.

Noether's method

Noether's method takes a plane curve and repeatedly applies quadratic transformations (determined by a singular points and two points in general position). Eventually this produces a plane curve whose only singularities are ordinary multiple points (all tangent lines have multiplicity 1).

Bertini's method

Bertini's method is similar to Noether's method. It starts with a plane curve, and repeatedly applies birational transformations to the plane to improve the curve. The birational transformations are more complicated than the quadratic transformations used in Noether's method, but produce the better result that the only singularities are ordinary double points.

Resolution of singularities of surfaces

Surfaces have many different nonsingular projective models (unlike the case of curves where the nonsingular projective model is unique). However a surface still has a unique minimal resolution, that all others factor through (all others are resolutions of it). In higher dimensions there need not be a minimal resolution.

Resolution for surfaces over the complex numbers was given informal proofs by Levi (1899), Chisini (1921) and Albanese (1924). A rigorous proof was first given by Walker (1935), and an algebraic proof for all fields of characteristic 0 was given by Zariski (1939). Abhyankar (1956) gave a proof for surfaces of non-zero characteristic. Resolution of singularities has also been shown for all excellent 2-dimensional schemes (including all arithmetic surfaces) by Lipman (1978).

Normalization and blowup

The usual method of resolution of singularities for surfaces is to repeatedly alternate normalizing the surface (which kills codimension 1 singularities) with blowing up points (which makes codimension 2 singularities better, but may introduce new codimension 1 singularities).

Jung's method

By applying strong embedded resolution for curves, Jung (1908) reduces to a surface with only rather special singularities (abelian quotient singularities) which are then dealt with explicitly. The higher-dimensional version of this method is de Jong's method.

Albanese method

In general the analogue of Albanese's method for curves shows that for any variety one can reduce to singularities of order at most $n!$, where n is the dimension. For surfaces this reduces to the case of singularities of order 2, which are easy enough to do explicitly.

Hironaka's method

Hironaka's method for arbitrary characteristic 0 varieties gives a resolution method for surfaces, which involves repeatedly blowing up points or smooth curves in the singular set.

Lipman's method

Lipman (1978) showed that a surface *Y* (a 2-dimensional reduced Noetherian scheme) has a desingularization if and only if its normalization is finite over *Y* and analytically normal (the completions of its singular points are normal) and has only finitely many singular points. In particular if *Y* is excellent then it has a desingularization.

His method was to consider normal surfaces *Z* with a birational proper map to *Y* and show that there is a minimal one with minimal possible arithmetic genus. He then shows that all singularities of this minimal *Z* are pseudo rational, and shows that pseudo rational singularities can be resolved by repeatedly blowing up points.

Resolution of singularities in higher dimensions

The problem of resolution of singularities in higher dimensions is notorious for many incorrect published proofs and announcements of proofs that never appeared.

Zariski's method

For 3-folds the resolution of singularities was proved in characteristic 0 by Zariski (1944).

Abhyankar's method

Abhyankar (1966) proved resolution of singularities in characteristic greater than 6. The restriction on the characteristic arises because Abhyankar shows that it is possible to resolve any singularity of a 3-fold of multiplicity less than the characteristic, and then uses Albanese's method to show that singularities can be reduced to those of multiplicity at most (dimension)! = 3! = 6.

Cossart and Piltant (2008, 2009) proved resolution of singularities of 3-folds in all characteristics.

Hironaka's method

Resolution of singularities in characteristic 0 in all dimensions was first proved by Hironaka (1964). He proved that it was possible to resolve singularities of varieties over fields of characteristic 0 by repeatedly blowing up along non-singular subvarieties, using a very complicated argument by induction on the dimension. Simplified versions of his formidable proof were given by several people, including Bierstone & Milman (1997), Encinas & Villamayor (1998), Encinas & Hauser (2002), Cutkosky (2004), Wlodarczyk (2005), Kollár (2007). Some of the recent proofs are about a tenth of the length of Hironaka's original proof, and are easy enough to give in an introductory graduate course.

De Jong's method

de Jong (1996) found a different approach to resolution of singularities, generalizing Jung's method for surfaces, which was used by Bogomolov & Pantev (1996) and by Abramovich & de Jong (1997) to prove resolution of singularities in characteristic 0. De Jong's method gave a weaker result for varieties of all dimensions in characteristic *p*, which was strong enough to act as a substitute for resolution for many purposes. De Jong proved that for any variety *X* over a field there is a dominant proper morphism which preserves the dimension from a regular variety onto *X*. This need not be a birational map, so is not a resolution of singularities, as it may be generically finite to one and so involves a finite extension of the function field of *X*. De Jong's idea was to try to represent *X* as a fibration over a smaller space *Y* with fibers that are curves (this may involve modifying *X*), then eliminate the singularities of *Y* by induction on the dimension, then eliminate the singularities in the fibers.

Resolution for schemes and status of the problem

It is easy to extend the definition of resolution to all schemes. Not all schemes have resolutions of their singularities: Grothendieck (1965, section 7.9) showed that if a locally Noetherian scheme *X* has the property that one can resolve the singularities of any finite integral scheme over *X*, then *X* must be quasi-excellent. Grothendieck also suggested that the converse might hold: in other words, if a locally Noetherian scheme *X* is reduced and quasi excellent, then it is possible to resolve its singularities. When *X* is defined over a field of characteristic 0, this follows from Hironaka's theorem, and when *X* has dimension at most 2 it was prove by Lipman. In general it would follow if it is possible to resolve the singularities of all integral complete local rings.

Hauser (2010) gave a survey of work on the unsolved characteristic *p* resolution problem.

Method of proof in characteristic zero

There are many constructions of strong desingularization but all of them give essentially the same result. In every case the global object (the variety to be desingularized) is replaced by local data (the ideal sheaf of the variety and those of the exceptional divisors and some *orders* that represents how much should be resolved the ideal in that step). With this local data the centers of blowing-up are defined. The centers will be defined locally and therefore it is a problem to guarantee that they will match up into a global center. This can be done by defining what blowings-up are allowed to resolve each ideal. Done this appropriately will make the centers match automatically. Another way is to define a local invariant depending on the variety and the history of the resolution (the previous local centers) so that the centers consist of the maximum locus of the invariant.

The definition of this is made such that making this choice is meaningful, giving smooth centers transversal to the exceptional divisors.

In either case the problem is reduced to resolve singularities of the tuple formed by the ideal sheaf and the extra data (the exceptional divisors and the order, d, to which the resolution should go for that ideal). This tuple is called a *marked ideal* and the set of points in which the order of the ideal is larger than d is called its co-support. The proof that there is a resolution for the marked ideals is done by induction on dimension. The induction breaks in two steps:

1. Functorial desingularization of marked ideal of dimension $n - 1$ implies functorial desingularization of marked ideals of maximal order of dimension n.
2. Functorial desingularization of marked ideals of maximal order of dimension n implies functorial desingularization of (a general) marked ideal of dimension n.

Here we say that a marked ideal is of *maximal order* if at some point of its co-support the order of the ideal is equal to d. A key ingredient in the strong resolution is the use of the Hilbert–Samuel function of the local rings of the points in the variety. This is one of the components of the resolution invariant.

Examples

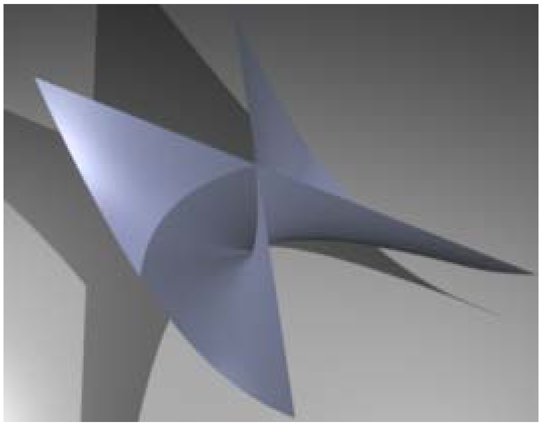
Multiplicity need not decrease under blowup

The most obvious invariant of a singularity is its multiplicity. However this need not decrease under blowup, so it is necessary to use more subtle invariants to measure the improvement.

For example, the rhamphoid cusp $y^2 = x^5$ has a singularity of order 2 at the origin. After blowing up at its singular point it becomes the ordinary cusp $y^2 = x^3$, which still has multiplicity 2.

In the previous example it was fairly clear that the singularity improved since the degree of one of the monimials defining it got smaller. This does not happen in general. An example where it does not is given by the isolated singularity of $x^2 + y^3z + z^3 = 0$ at the origin. Blowing it up gives the singularity $x^2 + y^2z + yz^3 = 0$. It is not immediately obvious that this new singularity is better, as both singularities have multiplicity 2 and are given by the sum of monomials of degrees 2, 3, and 4.

Blowing up the most singular points does not work



Whitney umbrella

A natural idea for improving singularities is to blow up the locus of the "worst" singular points. The Whitney umbrella $x^2 = y^2z$ has singular set the z axis, most of whose point are ordinary double points, but there is a more complicated pinch point singularity at the origin, so blowing up the worst singular points suggests that one should start by blowing up the origin. However blowing up the origin reproduces the same singularity on one of the coordinate charts. So blowing up the (apparently) "worst" singular points does not improve the singularity. Instead the singualrity can be resolved by blowing up along the z-axis.

There are algorithms that work by blowing up the "worst" singular points in some sense, such as (Bierstone & Milman 1997), but this example shows that the definition of the "worst" points needs to be quite subtle.

For more complicated singularities, such as $x^2 = y^mz^n$ which is singular along $x = yz = 0$, blowing up the worst singularity at the origin produces the singularities $x^2 = y^{m+n-2}z^n$ and $x^2 = y^mz^{m+n-2}$ which are worse than the original singularity if m and n are both at least 3.

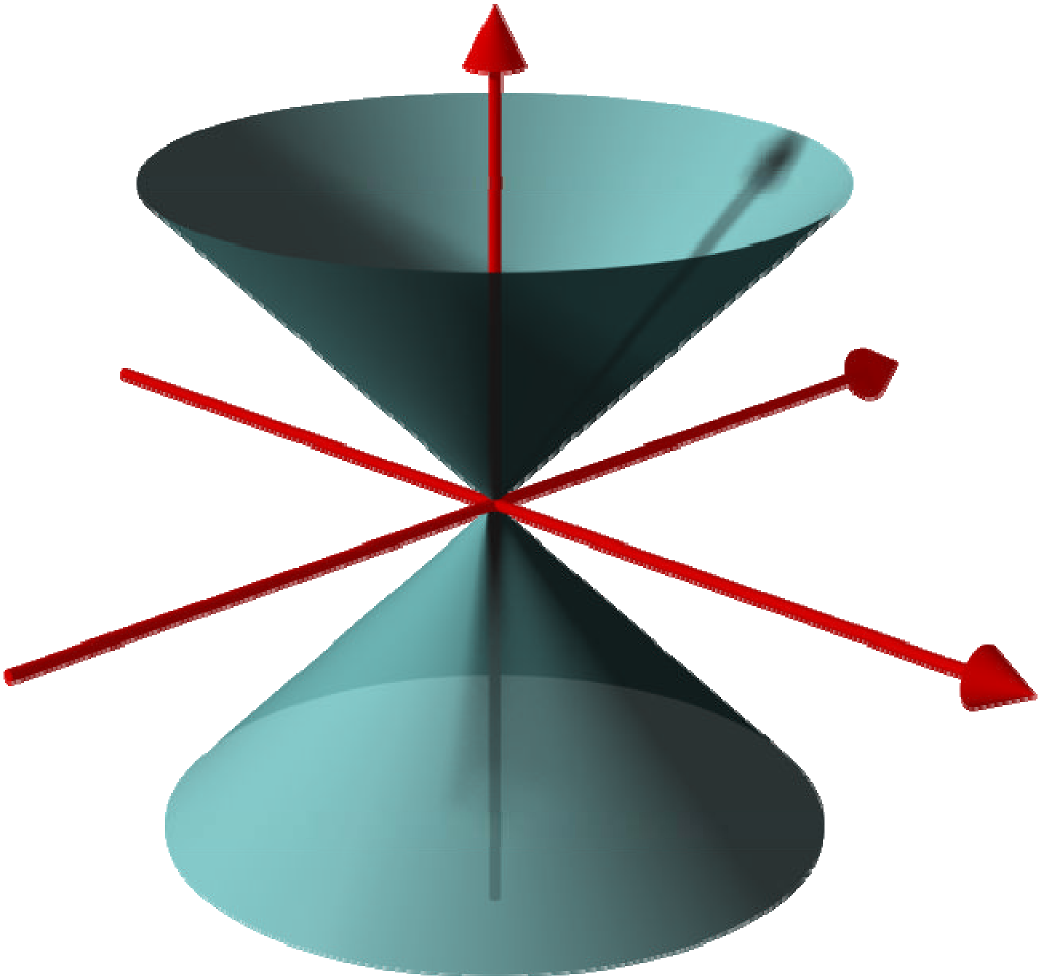
After the resolution the total transform, the union of the strict transform, X, and the exceptional divisors, is a variety with singularities of the simple normal crossings type. Then it is natural to consider the possibility of resolving singularities without resolving this type of singularities, this is finding a resolution that is an isomorphism over the set of smooth and simple normal crossing points. When X is a divisor, i.e. it can be embedded as a codimension one subvariety in a smooth variety it is known to be true the existence of the strong resolution avoiding simple normal crossing points. Whitney's umbrella shows that it is not possible to resolve singularities avoiding blowing-up the normal crossings singularities.

Incremental resolution procedures need memory

A natural way to resolve singularities is to repeatedly blow up some canonically chosen smooth subvariety. This runs into the following problem. The singular set of $x^2 = y^2z^2$ is the pair of lines given by the *y* and *z* axes. The only reasonable varieties to blow up are the origin, one of these two axes, or the whole singular set (both axes). However the whole singular set cannot be used since it is not smooth, and choosing one of the two axes breaks the symmetry between them so is not canonical. This means we have to start by blowing up the origin, but this reproduces the original singularity, so we seem to be going round in circles.

The solution to this problem is that although blowing up the origin does not change the type of the singularity, it does give a subtle improvement: it breaks the symmetry between the two singular axes because one of them is an exceptional divisor for a previous blowup, so it is now permissible to blow up just one of these. However in order to exploit this the resolution procedure needs to treat these 2 singularities differently, even though they are locally the same. This is sometimes done by giving the resolution procedure some memory, so the center of the blowup at each step depends not only on the singularity, but on the previous blowups used to produce it.

Resolutions are not functorial



Conical singularity $x^2 + y^2 = z^2$

Some resolution methods (in characteristic 0) are functorial for all smooth morphisms. However it is not possible to find a strong resolution functorial for all (possibly non-smooth) morphisms. An example is given by the map from the affine plane A^2 to the conical singularity $x^2 + y^2 = z^2$ taking (X,Y) to $(2XY, X^2 - Y^2, X^2 + Y^2)$. The *XY*-plane is already nonsingular so should not be changed by resolution, and any resolution of the conical singularity factorizes through the minimal resolution given by blowing up the singular point. However the rational map from the *XY*-plane to this blowup does not extend to a regular map.

Minimal resolutions need not exist

Minimal resolutions (resolutions such that every resolution factors through them) exist in dimensions 1 and 2, but not always in higher dimensions. The Atiyah flop gives an example in 3 dimensions of a singularity with no minimal resolution. Let Y be the zeros of $xy = zw$ in $\mathbf{A}^4$, and let V be the blowup of Y at the origin. The exceptional locus of this blowup is isomorphic to $\mathbf{P}^1 \times \mathbf{P}^1$, and can be blown down to $\mathbf{P}^1$ in 2 different ways, giving two small resolutions X_1 and X_2 of Y, neither of which can be blown down any further.

Resolutions should not commute with products

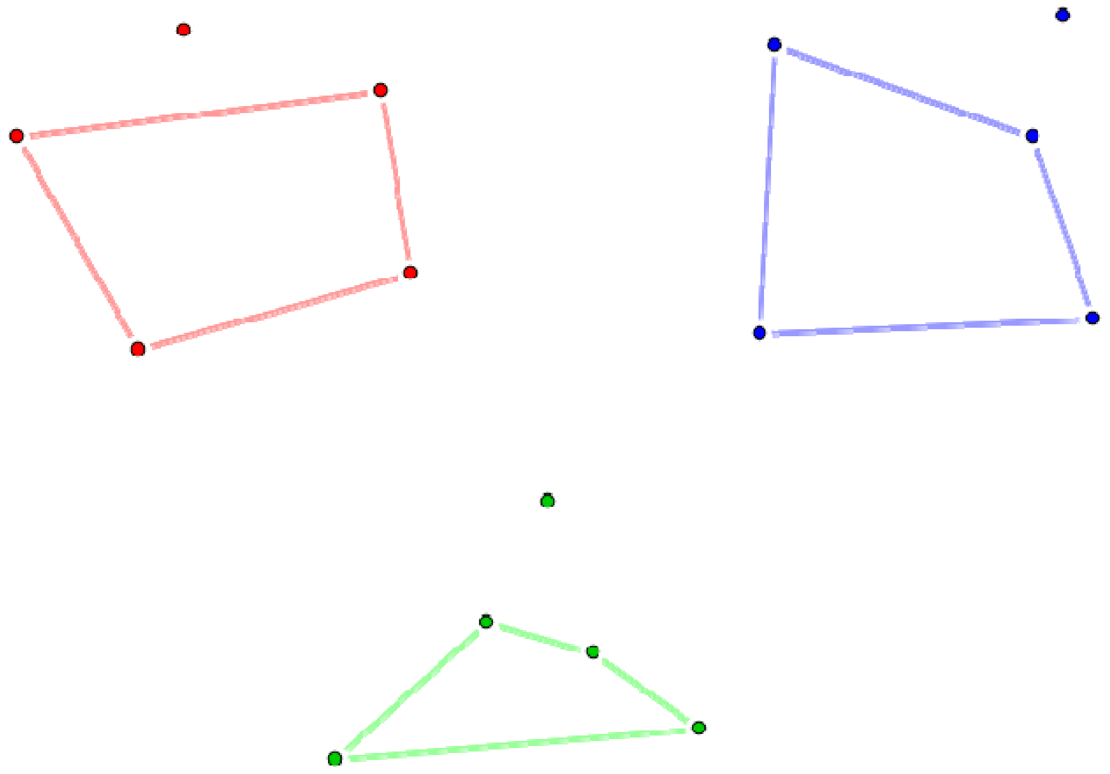
Kollar (2007, example 3.4.4, page 121) gives the following example showing that one cannot expect a sufficiently good resolution procedure to commute with products. If $f{:}A \longrightarrow B$ is the blowup of the origin of a quadric cone B in affine 3-space, then $f \times f{:}A \times A \longrightarrow B \times B$ cannot be produced by an etale local resolution procedure, essentially because the exceptional locus has 2 components that intersect.

Singularities of toric varieties

Singularities of toric varieties give examples of high dimensional singularities that are easy to resolve explicitly. A toric variety is defined by a fan, a collection of cones in a lattice. The singularities can be resolved by subdividing each cone into a union of cones each of which is generated by a basis for the lattice, and taking the corresponding toric variety.

Chapter 11

Happy Ending Problem



The Happy Ending problem: every set of five points in general position contains the vertices of a convex quadrilateral.

The **Happy Ending problem** (so named by Paul Erdős because it led to the marriage of George Szekeres and Esther Klein) is the following statement:

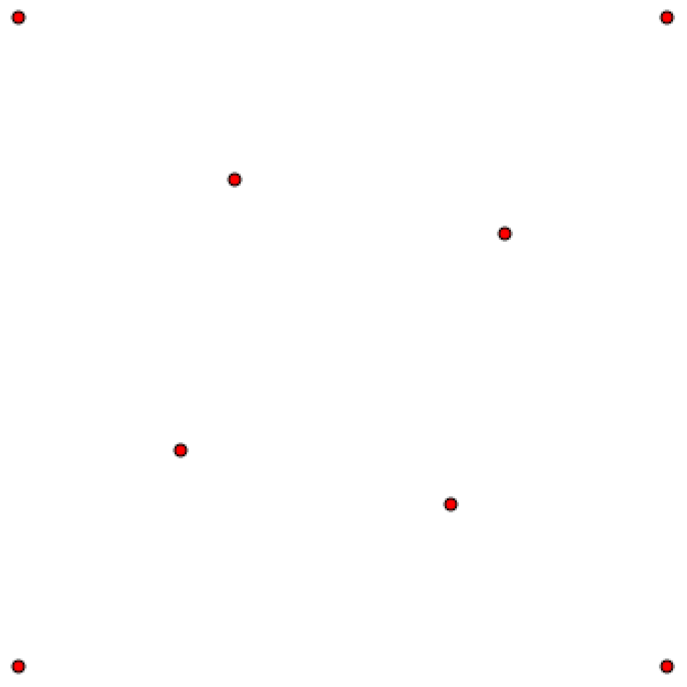
Theorem. Any set of five points in the plane in general position has a subset of four points that form the vertices of a convex quadrilateral.

This was one of the original results that led to the development of Ramsey theory.

The Happy Ending theorem can be proven by a simple case analysis: If four or more points are vertices of the convex hull, any four such points can be chosen. If on the other hand the point set has the form of a triangle with two points inside it, the two inner points and one of the triangle sides can be chosen.

The **Erdős–Szekeres conjecture** states precisely a more general relationship between the number of points in a general-position point set and its largest convex polygon. It remains unproven, but less precise bounds are known.

Larger polygons



A set of eight points in general position with no convex pentagon.

Erdős & Szekeres (1935) proved the following generalisation:

Theorem. For any positive integer *N*, any sufficiently large finite set of points in the plane in general position has a subset of *N* points that form the vertices of a convex polygon.

The proof appeared in the same paper that proves the Erdős–Szekeres theorem on monotonic subsequences in sequences of numbers.

Denoting by *f*(*N*) the minimal possible *M* for a set of *M* points in general position must contain a convex *N*-gon, it is known that

- *f*(3) = 3, trivially.
- *f*(4) = 5.
- *f*(5) = 9. A set of eight points with no convex pentagon is shown in the illustration, demonstrating that *f*(5) > 8; the more difficult part of the proof is to show that every set of nine points in general position contains the vertices of a convex pentagon.
- *f*(6) = 17.
- The value of *f*(*N*) is unknown for all *N* > 6; by the result of Erdős & Szekeres (1935) it is known to be finite.

On the basis of the known values of *f*(*N*) for *N* = 3, 4 and 5, Erdős and Szekeres conjectured in their original paper that

$$f(N) = 1 + 2^{N-2} \quad \text{for all } N \geq 3.$$

They proved later, by constructing explicit examples, that

$$f(N) \geq 1 + 2^{N-2},$$

but the best known upper bound when *N* ≥ 7 is

$$f(N) \leq \binom{2N-5}{N-2} + 1 = O\left(\frac{4^N}{\sqrt{N}}\right).$$

Empty polygons

One may also consider the question of whether any sufficiently large set of points in general position has an *empty* quadrilateral, pentagon, etc., that is, one that contains no other input point. The original solution to the Happy Ending problem can be adapted to show that any five points in general position have an empty convex quadrilateral, as shown in the illustration, and any ten points in general position have an empty convex pentagon. However, there exist arbitrarily large sets of points in general position that contain no empty heptagon.

For a long time the question of the existence of empty hexagons remained open, but Nicolás (2007) and Gerken (2008) proved that every sufficiently large point set in general position contains a convex empty hexagon. More specifically, Gerken showed that the number of points needed is no more than *f*(9) for the same function *f* defined above, while Nicolás showed that the number of points needed is no more than *f*(25). Valtr (2006) supplies a simplification of Gerken's proof that however requires more points, *f*(15) instead of *f*(9). At least 30 points are needed: there exists a set of 29 points in general position with no empty convex hexagon.

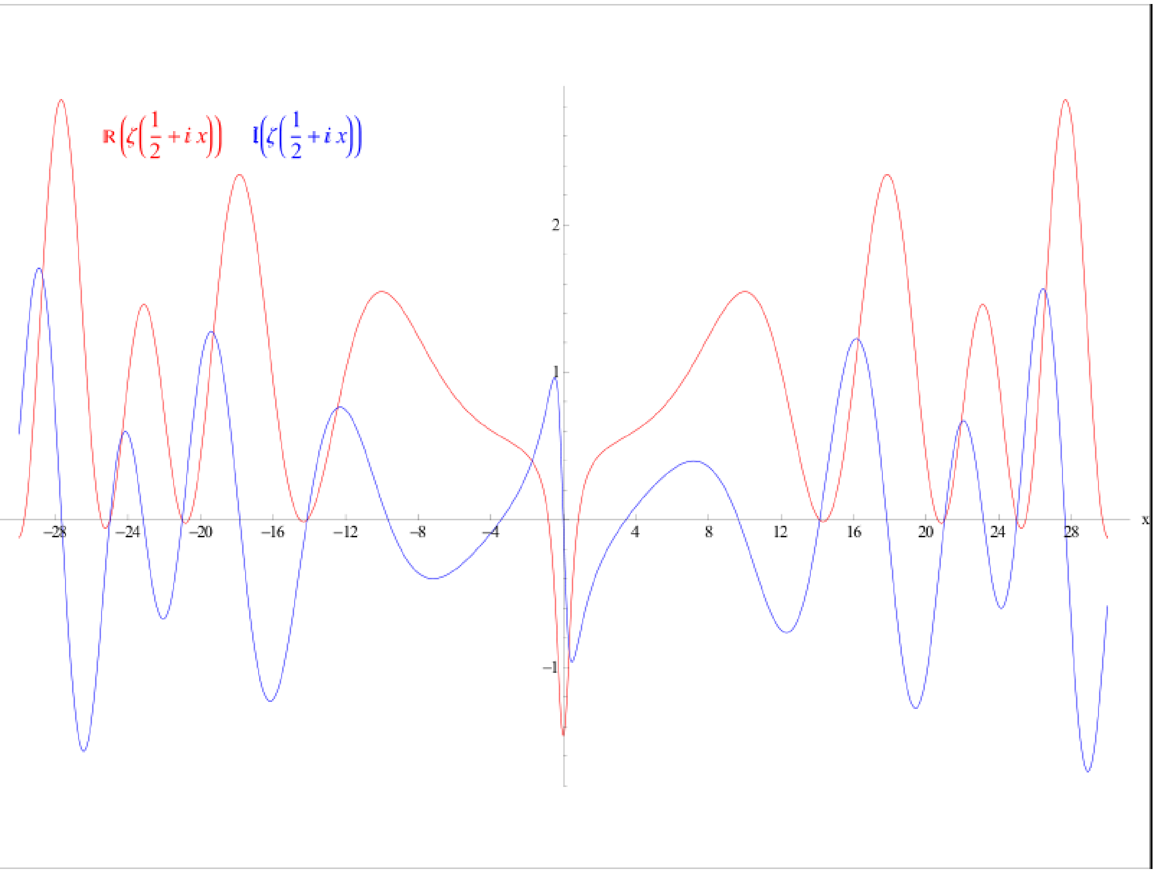
Related problems

The problem of finding sets of *n* points minimizing the number of convex quadrilaterals is equivalent to minimizing the crossing number in a straight-line drawing of a complete graph. The number of quadrilaterals must be proportional to the fourth power of *n*, but the precise constant is not known.

It is straightforward to show that, in higher dimensional Euclidean spaces, sufficiently large sets of points will have a subset of *k* points that forms the vertices of a convex polytope, for any *k* greater than the dimension: this follows immediately from existence of convex *k*-gons in sufficiently large planar point sets, by projecting the higher dimensional point set into an arbitrary two-dimensional subspace. However, the number of points necessary to find *k* points in convex position may be smaller in higher dimensions than it is in the plane, and it is possible to find subsets that are more highly constrained. In particular, in *d* dimensions, every $d + 3$ points in general position have a subset of $d + 2$ points that form the vertices of a cyclic polytope. More generally, for every *d* and $k > d$ there exists a number $m(d,k)$ such that every set of $m(d,k)$ points in general position has a subset of *k* points that form the vertices of a neighborly polytope.

Chapter 12

Riemann Hypothesis



The real part (red) and imaginary part (blue) of the Riemann zeta function along the critical line Re(*s*) = 1/2. The first non-trivial zeros can be seen at Im(*s*) = ±14.135, ±21.022 and ±25.011.

In mathematics, the **Riemann hypothesis**, proposed by Bernhard Riemann (1859), is a conjecture about the distribution of the zeros of the Riemann zeta function which states that all non-trivial zeros have real part 1/2. The name is also used for some closely related analogues, such as the Riemann hypothesis for curves over finite fields.

The Riemann hypothesis implies results about the distribution of prime numbers that are in some ways as good as possible. Along with suitable generalizations, it is considered by some mathematicians to be the most important unresolved problem in pure mathematics (Bombieri 2000). The Riemann hypothesis is part of Problem 8, along with the Goldbach conjecture, in Hilbert's list of 23 unsolved problems, and is also one of the Clay Mathematics Institute Millennium Prize Problems. Since it was formulated, it has withstood concentrated efforts from many outstanding mathematicians. In 1973, Pierre Deligne proved an analogue of the Riemann Hypothesis for zeta functions of varieties defined over finite fields. The full version of the hypothesis remains unsolved, although modern computer calculations have shown that the first 10 trillion zeros lie on the critical line.

The Riemann zeta function $\zeta(s)$ is defined for all complex numbers $s \neq 1$. It has zeros at the negative even integers (i.e. at $s = -2, -4, -6, \ldots$). These are called the **trivial zeros**. The Riemann hypothesis is concerned with the non-trivial zeros, and states that:

> The real part of any non-trivial zero of the Riemann zeta function is 1/2.

Thus the non-trivial zeros should lie on the **critical line**, $1/2 + it$, where t is a real number and i is the imaginary unit.

There are several popular books on the Riemann hypothesis, such as Derbyshire (2003), Rockmore (2005), Sabbagh (2003), du Sautoy (2003). The books Edwards (1974), Patterson (1988) and Borwein et al. (2008) give mathematical introductions, while Titchmarsh (1986), Ivić (1985) and Karatsuba & Voronin (1992) are advanced monographs.

The Riemann zeta function

The Riemann zeta function is given for complex s with real part greater than 1 by

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \frac{1}{1^s} + \frac{1}{2^s} + \frac{1}{3^s} + \cdots.$$

Leonhard Euler showed that it is given by the Euler product

$$\zeta(s) = \prod_{p \text{ prime}} \frac{1}{1-p^{-s}} = \frac{1}{1-2^{-s}} \cdot \frac{1}{1-3^{-s}} \cdot \frac{1}{1-5^{-s}} \cdot \frac{1}{1-7^{-s}} \cdots \frac{1}{1-p^{-s}} \cdots$$

where the infinite product extends over all prime numbers *p*, and again converges for complex *s* with real part greater than 1. The convergence of the Euler product shows that ζ(*s*) has no zeros in this region, as none of the factors have zeros.

The Riemann hypothesis discusses zeros outside the region of convergence of this series, so it needs to be analytically continued to all complex *s*. This can be done by expressing it in terms of the Dirichlet eta function as follows. If *s* has positive real part, then the zeta function satisfies

$$\left(1-\frac{2}{2^s}\right)\zeta(s)=\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^s}=\frac{1}{1^s}-\frac{1}{2^s}+\frac{1}{3^s}-\cdots$$

where the series on the right converges whenever *s* has positive real part. Thus, this alternative series extends the zeta function from Re(*s*) > 1 to the larger domain Re(*s*) > 0.

In the strip 0 < Re(*s*) < 1 the zeta function satisfies the functional equation

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

One may then define ζ(*s*) for all remaining nonzero complex numbers *s* by assuming that this equation holds outside the strip as well, and letting ζ(*s*) equal the right-hand side of the equation whenever *s* has non-positive real part. If *s* is a negative even integer then ζ(*s*) = 0 because the factor sin(π*s*/2) vanishes; these are the **trivial zeros** of the zeta function. (If *s* is a positive even integer this argument does not apply because the zeros of sin are cancelled by the poles of the gamma function.) The value ζ(0) = −1/2 is not determined by the functional equation, but is the limiting value of ζ(*s*) as *s* approaches zero. The functional equation also implies that the zeta function has no zeros with negative real part other than the trivial zeros, so all non-trivial zeros lie in the **critical strip** where *s* has real part between 0 and 1.

History

"…es ist sehr wahrscheinlich, dass alle Wurzeln reell sind. Hiervon wäre allerdings ein strenger Beweis zu wünschen; ich habe indess die Aufsuchung desselben nach einigen flüchtigen vergeblichen Versuchen vorläufig bei Seite gelassen, da er für den nächsten Zweck meiner Untersuchung entbehrlich schien."

"…it is very probable that all roots are real. Of course one would wish for a rigorous proof here; I have for the time being, after some fleeting vain attempts, provisionally put aside the search for this, as it appears dispensable for the next objective of my investigation."

Riemann's statement of the Riemann hypothesis, from (Riemann 1859). (He was discussing a version of the zeta function, modified so that its roots are real rather than on the critical line.)

In his 1859 paper *On the Number of Primes Less Than a Given Magnitude* Riemann found an explicit formula for the number of primes π(x) less than a given number x. His formula was given in terms of the related function

$$\Pi(x) = \pi(x) + \frac{1}{2}\pi(x^{1/2}) + \frac{1}{3}\pi(x^{1/3}) + \cdots$$

which counts primes where a prime power p^n counts as $1/n$ of a prime. The number of primes can be recovered from this function by

$$\pi(x) = \sum_{n=1}^{\infty} \frac{\mu(n)}{n}\Pi(x^{1/n}) = \Pi(x) - \frac{1}{2}\Pi(x^{1/2}) - \frac{1}{3}\Pi(x^{1/3}) - \cdots,$$

where μ is the Möbius function. Riemann's formula is then

$$\Pi_0(x) = \mathrm{Li}(x) - \sum_{\rho} \mathrm{Li}(x^{\rho}) - \log(2) + \int_x^{\infty} \frac{dt}{t(t^2-1)\log(t)}$$

where the sum is over the nontrivial zeros of the zeta function and where Π_0 is a slightly modified version of Π that replaces its value at its points of discontinuity by the average of its upper and lower limits:

$$\Pi_0(x) = \lim_{\varepsilon \to 0} \frac{\Pi(x-\varepsilon) + \Pi(x+\varepsilon)}{2}.$$

The summation in Riemann's formula is not absolutely convergent, but may be evaluated by taking the zeros ρ in order of the absolute value of their imaginary part. The function Li occurring in the first term is the (unoffset) logarithmic integral function given by the Cauchy principal value of the divergent integral

$$\mathrm{Li}(x) = \int_0^x \frac{dt}{\log(t)}.$$

The terms Li(x^{ρ}) involving the zeros of the zeta function need some care in their definition as Li has branch points at 0 and 1, and are defined (for $x > 1$) by analytic continuation in the complex variable ρ in the region Re(ρ) > 0, i.e. they should be considered as Ei(ρ ln x). The other terms also correspond to zeros: the dominant term Li(x) comes from the pole at $s = 1$, considered as a zero of multiplicity −1, and the remaining small terms come from the trivial zeros.

This formula says that the zeros of the Riemann zeta function control the oscillations of primes around their "expected" positions. Riemann knew that the non-trivial zeros of the zeta function were symmetrically distributed about the line $s = 1/2 + it$, and he knew that all of its non-trivial zeros must lie in the range $0 \leq \mathrm{Re}(s) \leq 1$. He checked that a few of the zeros lay on the critical line with real part 1/2 and suggested that they all do; this is the Riemann hypothesis.

Consequences of the Riemann hypothesis

The practical uses of the Riemann hypothesis include many propositions which are known to be true under the Riemann hypothesis, and some which can be shown to be equivalent to the Riemann hypothesis.

Distribution of prime numbers

Riemann's explicit formula for the number of primes less than a given number in terms of a sum over the zeros of the Riemann zeta function says that the magnitude of the oscillations of primes around their expected position is controlled by the real parts of the zeros of the zeta function. In particular the error term in the prime number theorem is closely related to the position of the zeros: for example, the supremum of real parts of the zeros is the infimum of numbers β such that the error is O(x^β) (Ingham 1932).

Von Koch (1901) proved that the Riemann hypothesis is equivalent to the "best possible" bound for the error of the prime number theorem.

A precise version of Koch's result, due to Schoenfeld (1976), says that the Riemann hypothesis is equivalent to

$$|\pi(x) - \mathrm{Li}(x)| < \frac{1}{8\pi}\sqrt{x}\,\log(x), \qquad \text{for all } x \ge 2657.$$

Growth of arithmetic functions

The Riemann hypothesis implies strong bounds on the growth of many other arithmetic functions, in addition to the primes counting function above.

One example involves the Möbius function μ. The statement that the equation

$$\frac{1}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{\mu(n)}{n^s}$$

is valid for every *s* with real part greater than 1/2, with the sum on the right hand side converging, is equivalent to the Riemann hypothesis. From this we can also conclude that if the Mertens function is defined by

$$M(x) = \sum_{n \le x} \mu(n)$$

then the claim that

$$M(x) = O(x^{1/2+\varepsilon})$$

for every positive ε is equivalent to the Riemann hypothesis (Titchmarsh 1986). The determinant of the order *n* Redheffer matrix is equal to *M*(*n*), so the Riemann hypothesis can also be stated as a condition on the growth of these determinants. The Riemann hypothesis puts a rather tight bound on the growth of *M*, since Odlyzko & te Riele (1985) disproved the slightly stronger Mertens conjecture

$$|M(x)| \le \sqrt{x}.$$

The Riemann hypothesis is equivalent to many other conjectures about the rate of growth of other arithmetic functions aside from μ(*n*). A typical example is Robin's theorem (Robin 1984), which states that if σ(*n*) is the divisor function, given by

$$\sigma(n) = \sum_{d|n} d$$

then

$$\sigma(n) < e^\gamma n \log \log n$$

for all *n* > 5040 if and only if the Riemann hypothesis is true, where γ is the Euler–Mascheroni constant.

Another example was found by Franel & Landau (1924) showing that the Riemann hypothesis is equivalent to a statement that the terms of the Farey sequence are fairly regular. More precisely, if F_n is the Farey sequence of order *n*, beginning with 1/*n* and up to 1/1, then the claim that for all ε > 0

$$\sum_{i=1}^{m} |F_n(i) - i/m| = O(n^{1/2+\epsilon})$$

is equivalent to the Riemann hypothesis. Here $m = \sum_{i=1}^{n} \phi(i)$ is the number of terms in the Farey sequence of order *n*.

For an example from group theory, if *g*(*n*) is Landau's function given by the maximal order of elements of the symmetric group S_n of degree *n*, then Massias, Nicolas & Robin (1988) showed that the Riemann hypothesis is equivalent to the bound

$\log g(n) < \sqrt{\operatorname{Li}^{-1}(n)}$ for all sufficiently large *n*.

Lindelöf hypothesis and growth of the zeta function

The Riemann hypothesis has various weaker consequences as well; one is the **Lindelöf hypothesis** on the rate of growth of the zeta function on the critical line, which says that, for any $\varepsilon > 0$,

$$\zeta\left(\frac{1}{2}+it\right) = O(t^{\varepsilon}),$$

as t tends to infinity.

The Riemann hypothesis also implies quite sharp bounds for the growth rate of the zeta function in other regions of the critical strip. For example, it implies that

$$e^{\gamma} \leq \limsup_{t\rightarrow+\infty} \frac{|\zeta(1+it)|}{\log\log t} \leq 2e^{\gamma}$$
$$\frac{6}{\pi^2}e^{\gamma} \leq \limsup_{t\rightarrow+\infty} \frac{1/|\zeta(1+it)|}{\log\log t} \leq \frac{12}{\pi^2}e^{\gamma}$$

so the growth rate of $\zeta(1+it)$ and its inverse would be known up to a factor of 2 (Titchmarsh 1986).

Large prime gap conjecture

The prime number theorem implies that on average, the gap between the prime p and its successor is $\log p$. However, some gaps between primes may be much larger than the average. Cramér proved that, assuming the Riemann hypothesis, every gap is $O(\sqrt{p}\log p)$. This is a case when even the best bound that can currently be proved using the Riemann hypothesis is far weaker than what seems to be true: Cramér's conjecture implies that every gap is $O((\log p)^2)$ which, while larger than the average gap, is far smaller than the bound implied by the Riemann hypothesis. Numerical evidence supports Cramér's conjecture (Nicely 1999).

Criteria equivalent to the Riemann hypothesis

Many statements equivalent to the Riemann hypothesis have been found, though so far none of them have led to much progress in solving it. Some typical examples are as follows.

The Riesz criterion was given by Riesz (1916), to the effect that the bound

$$-\sum_{k=1}^{\infty} \frac{(-x)^k}{(k-1)!\,\zeta(2k)} = O\left(x^{1/4+\epsilon}\right)$$

holds for all ε > 0 if and only if the Riemann hypothesis holds.

Nyman (1950) proved that the Riemann Hypothesis is true if and only if the space of functions of the form

$$f(x) = \sum_{\nu=1}^{n} c_\nu \rho(\theta_\nu / x)$$

where ρ(*z*) is the fractional part of *z*, $0 \le \theta_\nu \le 1$, and

$$\sum_{\nu=1}^{n} c_\nu \theta_\nu = 0 \quad ,$$

is dense in the Hilbert space $L^2(0,1)$ of square-integrable functions on the unit interval. Beurling (1955) extended this by showing that the zeta function has no zeros with real part greater than 1/*p* if and only if this function space is dense in $L^p(0,1)$

Salem (1953) showed that the Riemann hypothesis is true if and only if the integral equation

$$\int_0^\infty \frac{z^{-\sigma-1}\phi(z)\, dz}{e^{x/z}+1} = 0$$

has no non-trivial bounded solutions φ for 1/2<σ<1.

Weil's criterion is the statement that the positivity of a certain function is equivalent to the Riemann hypothesis. Related is Li's criterion, a statement that the positivity of a certain sequence of numbers is equivalent to the Riemann hypothesis.

Speiser (1934) proved that the Riemann hypothesis is equivalent to the statement that ζ'(*s*), the derivative of ζ(*s*), has no zeros in the strip

$$0 < \Re(s) < \frac{1}{2}.$$

That ζ has only simple zeros on the critical line is equivalent to its derivative having no zeros on the critical line.

Consequences of the generalized Riemann hypothesis

Several applications use the generalized Riemann hypothesis for Dirichlet L-series or zeta functions of number fields rather than just the Riemann hypothesis. Many basic properties of the Riemann zeta function can easily be generalized to all Dirichlet L-series, so it is plausible that a method that proves the Riemann hypothesis for the Riemann zeta function would also work for the generalized Riemann hypothesis for Dirichlet L-functions. Several results first proved using the generalized Riemann hypothesis were later given unconditional proofs without using it, though these were usually much harder. Many of the consequences on the following list are taken from Conrad (2010).

- In 1913, Gronwall showed that the generalized Riemann hypothesis implies that Gauss's list of imaginary quadratic fields with class number 1 is complete, though Baker, Stark and Heegner later gave unconditional proofs of this without using the generalized Riemann hypothesis.
- In 1917, Hardy and Littlewood showed that the generalized Riemann hypothesis implies a conjecture of Chebyshev that

  $$\lim_{x\to 1^-}\sum_{p>2}(-1)^{(p+1)/2}x^p=+\infty$$

  which says that in some sense primes 3 mod 4 are more common than primes 1 mod 4.

- In 1923 Hardy and Littlewood showed that the generalized Riemann hypothesis implies a weak form of the Goldbach conjecture for odd numbers: that every sufficiently large odd number is the sum of 3 primes, though in 1937 Vinogradov gave an unconditional proof. In 1997 Deshouillers, Effinger, te Riele, and Zinoviev showed that the generalized Riemann hypothesis implies that every odd number greater than 5 is the sum of 3 primes.
- In 1934, Chowla showed that the generalized Riemann hypothesis implies that the first prime in the arithmetic progression *a* mod *m* is at most $Km^2\log(m)^2$ for some fixed constant *K*.
- In 1967, Hooley showed that the generalized Riemann hypothesis implies Artin's conjecture on primitive roots.
- In 1973, Weinberger showed that the generalized Riemann hypothesis implies that Euler's list of idoneal numbers is complete.
- Weinberger (1973) showed that the generalized Riemann hypothesis for the zeta functions of all algebraic number fields implies that any number field with class number 1 is either Euclidean or an imaginary quadratic number field of discriminant −19, −43, −67, or −163.
- In 1976, G. Miller showed that the generalized Riemann hypothesis implies that one can test if a number is prime in polynomial times. In 2002, Manindra Agrawal, Neeraj Kayal and Nitin Saxena proved this result unconditionally using the AKS primality test.
- Odlyzko (1990) discussed how the generalized Riemann hypothesis can be used to give sharper estimates for discriminants and class numbers of number fields.
- In 1997 Ono and Soundararajan showed that the generalized Riemann hypothesis implies that Ramanujan's integral quadratic form $x^2 + y^2 + 10z^2$ represents all integers that it represents locally, with exactly 18 exceptions.

Generalizations and analogues of the Riemann hypothesis

Dirichlet L-series and other number fields

The Riemann hypothesis can be generalized by replacing the Riemann zeta function by the formally similar, but much more general, global L-functions. In this broader setting, one expects the non-trivial zeros of the global *L*-functions to have real part 1/2. It is these conjectures, rather than the classical Riemann hypothesis only for the single Riemann zeta function, which accounts for the true importance of the Riemann hypothesis in mathematics.

The generalized Riemann hypothesis extends the Riemann hypothesis to all Dirichlet L-functions. In particular it implies the conjecture that Siegel zeros (zeros of *L* functions between 1/2 and 1) do not exist.

The extended Riemann hypothesis extends the Riemann hypothesis to all Dedekind zeta functions of algebraic number fields. The extended Riemann hypothesis for abelian extension of the rationals is equivalent to the generalized Riemann hypothesis. The Riemann hypothesis can also be extended to the L-functions of Hecke characters of number fields.

The grand Riemann hypothesis extends it to all automorphic zeta functions, such as Mellin transforms of Hecke eigenforms.

Function fields and zeta functions of varieties over finite fields

Artin (1924) introduced global zeta functions of (quadratic) function fields and conjectured an analogue of the Riemann hypothesis for them, which has been proven by Hasse in the genus 1 case and by Weil (1948) in general. For instance, the fact that the Gauss sum, of the quadratic character of a finite field of size *q* (with *q* odd), has absolute value

$$\sqrt{q}$$

is actually an instance of the Riemann hypothesis in the function field setting. This led Weil (1949) to conjecture a similar statement for all algebraic varieties; the resulting Weil conjectures were proven by Pierre Deligne (1974, 1980).

Selberg zeta functions

Selberg (1956) introduced the Selberg zeta function of a Riemann surface. These are similar to the Riemann zeta function: they have a functional equation, and an infinite product similar to the Euler product but taken over closed geodesics rather than primes. The Selberg trace formula is the analogue for these functions of the explicit formulas in prime number theory. Selberg proved that the Selberg zeta functions satisfy the analogue of the Riemann hypothesis, with the imaginary parts of their zeros related to the eigenvalues of the Laplacian operator of the Riemann surface.

Ihara zeta functions

The Ihara zeta function of a finite graph is an analogue of the Selberg zeta function introduced by Yasutaka Ihara. A regular finite graph is a Ramanujan graph, a mathematical model of efficient communication networks, if and only if its Ihara zeta function satisfies the analogue of the Riemann hypothesis as was pointed out by T. Sunada.

Montgomery's pair correlation conjecture

Montgomery (1973) suggested the pair correlation conjecture that the correlation functions of the (suitably normalized) zeros of the zeta function should be the same as those of the eigenvalues of a random hermitian matrix. Odlyzko (1987) showed that this is supported by large scale numerical calculations of these correlation functions.

Montgomery showed that (assuming the Riemann hypothesis) at least 2/3 of all zeros are simple, and a related conjecture is that all zeros of the zeta function are simple (or more generally have no non-trivial integer linear relations between their imaginary parts). Dedekind zeta functions of algebraic number fields, which generalize the Riemann zeta function, often do have multiple complex zeros. This is because the Dedekind zeta functions factorize as a product of powers of Artin L-functions, so zeros of Artin L-functions sometimes give rise to multiple zeros of Dedekind zeta functions. Other examples of zeta functions with multiple zeros are the L-functions of some elliptic curves: these can have multiple zeros at the real point of their critical line; the Birch-Swinnerton-Dyer conjecture predicts that the multiplicity of this zero is the rank of the elliptic curve.

Other zeta functions

There are many other examples of zeta functions with analogues of the Riemann hypothesis, some of which have been proved. Goss zeta functions of function fields have a Riemann hypothesis, proved by Sheats (1998). The main conjecture of Iwasawa theory, proved by Barry Mazur and Andrew Wiles for cyclotomic fields, and Wiles for totally real fields, identifies the zeros of a p-adic L-function with the eigenvalues of an operator, so can be thought of as an analogue of the Hilbert–Pólya conjecture for p-adic L-functions (Wiles 2000).

Attempts to prove the Riemann hypothesis

Several mathematicians have addressed the Riemann hypothesis, but none of their attempts have yet been accepted as correct solutions. Watkins (2007) lists some incorrect solutions, and more are frequently announced.

Operator theory

Hilbert and Polya suggested that one way to derive the Riemann hypothesis would be to find a self-adjoint operator, from the existence of which the statement on the real parts of the zeros of $\zeta(s)$ would follow when one applies the criterion on real eigenvalues. Some support for this idea comes from several analogues of the Riemann zeta functions whose zeros correspond to eigenvalues of some operator: the zeros of a zeta function of a variety over a finite field correspond to eigenvalues of a Frobenius element on an etale cohomology group, the zeros of a Selberg zeta function are eigenvalues of a Laplacian operator of a Riemann surface, and the zeros of a p-adic zeta function correspond to eigenvectors of a Galois action on ideal class groups.

Odlyzko (1987) showed that the distribution of the zeros of the Riemann zeta function shares some statistical properties with the eigenvalues of random matrices drawn from the Gaussian unitary ensemble. This gives some support to the Hilbert–Pólya conjecture.

In 1999, Michael Berry and Jon Keating conjectured that there is some unknown quantization $\hat{H}$ of the classical Hamiltonian $H = xp$ so that

$$\zeta(1/2 + i\hat{H}) = 0$$

and even more strongly, that the Riemann zeros coincide with the spectrum of the operator $1/2 + i\hat{H}$. This is to be contrasted to canonical quantization which leads to the Heisenberg uncertainty principle $[x,p] = 1 / 2$ and the natural numbers as spectrum of the quantum harmonic oscillator. The crucial point is that the Hamiltonian should be a self-adjoint operator so that the quantization would be a realization of the Hilbert–Pólya program. In a connection with this Quantum mechanical problem Berry and Connes had proposed that the inverse of the potential of the Hamiltonian is connected to the half-derivative of the function $N(s) = \frac{1}{\pi} Arg\xi(1/2 + i\sqrt{s})$ then, in Berry-Connes approach $V^{-1}(x) = \sqrt{(4\pi)}\frac{d^{1/2}N(x)}{dx^{1/2}}$ (Connes 1999). This yields to a Hamiltonian whose eigenvalues are the square of the imaginary part of the Riemann zeros, also the functional determinant of this Hamiltonian operator is just the Riemann Xi-function

The analogy with the Riemann hypothesis over finite fields suggests that the Hilbert space containing eigenvectors corresponding to the zeros might be some sort of first cohomology group of the spectrum Spec(**Z**) of the integers. Deninger (1998) described some of the attempts to find such a cohomology theory.

Zagier (1983) constructed a natural space of invariant functions on the upper half plane which has eigenvalues under the Laplacian operator corresponding to zeros of the Riemann zeta function, and remarked that in the unlikely event that one could show the existence of a suitable positive definite inner product on this space the Riemann hypothesis would follow. Cartier (1982) discussed a related example, where due to a bizarre bug a computer program listed zeros of the Riemann zeta function as eigenvalues of the same Laplacian operator.

Schumayer & Hutchinson (2011) surveyed some of the attempts to construct a suitable physical model related to the Riemann zeta function.

Lee–Yang theorem

The Lee–Yang theorem states that the zeros of certain partition functions in statistical mechanics all lie on a "critical line" with real part 0, and this has led to some speculation about a relationship with the Riemann hypothesis (Knauf 1999).

Turán's result

Pál Turán (1948) showed that if the functions

$$\sum_{n=1}^{N} n^{-s}$$

have no zeros when the real part of s is greater than one then

$$T(x) = \sum_{n \le x} \frac{\lambda(n)}{n} \ge 0 \quad \text{for all } x > 0,$$

where $\lambda(n)$ is the Liouville function given by $(-1)^r$ if n has r prime factors. He showed that this in turn would imply that the Riemann hypothesis is true. However Haselgrove (1958) proved that $T(x)$ is negative for infinitely many x (and also disproved the closely related Polya conjecture), and Borwein, Ferguson & Mossinghoff (2008) showed that the smallest such x is 72185376951205. Spira (1968) showed by numerical calculation that the finite Dirichlet series above for N=19 has a zero with real part greater than 1. Turán also showed that a somewhat weaker assumption, the nonexistence of zeros with real part greater than $1+N^{-1/2+\varepsilon}$ for large N in the finite Dirichlet series above, would also imply the Riemann hypothesis, but Montgomery (1983) showed that for all sufficiently large N these series have zeros with real part greater than $1 + (\log \log N)/(4 \log N)$. Therefore, Turán's result is vacuously true and cannot be used to help prove the Riemann hypothesis.

Noncommutative geometry

Connes (1999, 2000) has described a relationship between the Riemann hypothesis and noncommutative geometry, and shows that a suitable analogue of the Selberg trace formula for the action of the idèle class group on the adèle class space would imply the Riemann hypothesis. Some of these ideas are elaborated in Lapidus (2008).

Hilbert spaces of entire functions

Louis de Branges (1992) showed that the Riemann hypothesis would follow from a positivity condition on a certain Hilbert space of entire functions. However Conrey & Li (2000) showed that the necessary positivity conditions are not satisfied.

Quasicrystals

The Riemann hypothesis implies that the zeros of the zeta function form a quasicrystal, meaning a distribution with discrete support whose Fourier transform also has discrete support. Dyson (2009) suggested trying to prove the Riemann hypothesis by classifying, or at least studying, 1-dimensional quasicrystals.

Multiple zeta functions

Deligne's proof of the Riemann hypothesis over finite fields used the zeta functions of product varieties, whose zeros and poles correspond to sums of zeros and poles of the original zeta function, in order to bound the real parts of the zeros of the original zeta function. By analogy, Kurokawa (1992) introduced multiple zeta functions whose zeros and poles correspond to sums of zeros and poles of the Riemann zeta function. To make the series converge he restricted to sums of zeros or poles all with non-negative imaginary part. So far, the known bounds on the zeros and poles of the multiple zeta functions are not strong enough to give useful estimates for the zeros of the Riemann zeta function.

Location of the zeros

Number of zeros

The functional equation combined with the argument principle implies that the number of zeros of the zeta function with imaginary part between 0 and *T* is given by

$$N(T)=\frac{1}{\pi}\mathrm{Arg}(\xi(s))=\frac{1}{\pi}\mathrm{Arg}(\Gamma(s/2)\pi^{-s/2}\zeta(s)s(s-1)/2)$$

for *s*=1/2+i*T*, where the argument is defined by varying it continuously along the line with Im(*s*)=*T*, starting with argument 0 at ∞+i*T*. This is the sum of a large but well understood term

$$\frac{1}{\pi}\mathrm{Arg}(\Gamma(s/2)\pi^{-s/2}s(s-1)/2)=\frac{T}{2\pi}\log\frac{T}{2\pi}-\frac{T}{2\pi}+7/8+O(1/T)$$

and a small but rather mysterious term

$$S(T) = \frac{1}{\pi}\mathrm{Arg}(\zeta(1/2 + iT)) = O(\log(T)).$$

So the density of zeros with imaginary part near T is about $\log(T)/2\pi$, and the function S describes the small deviations from this. The function $S(t)$ jumps by 1 at each zero of the zeta function, and for $t \geq 8$ it decreases monotonically between zeros with derivative close to $-\log t$.

Karatsuba (1996) proved that every interval $(T,T + H]$ for $H \geq T^{27/82+\varepsilon}$contains at least

$$H(\ln T)^{1/3}e^{-c\sqrt{\ln\ln T}}$$

points where the function $S(t)$ changes sign.

Selberg (1946) showed that the average moments of even powers of S are given by

$$\int_0^T |S(t)|^{2k}dt = \frac{(2k)!}{k!(2\pi)^{2k}}T(\log\log T)^k + O(T(\log\log T)^{k-1/2}).$$

This suggests that $S(T)/(\log\log T)^{1/2}$ resembles a Gaussian random variable with mean 0 and variance $2\pi^2$ (Ghosh (1983) proved this fact). In particular $|S(T)|$ is usually somewhere around $(\log\log T)^{1/2}$, but occasionally much larger. The exact order of growth of $S(T)$ is not known. There has been no unconditional improvement to Riemann's original bound $S(T)=O(\log T)$, though the Riemann hypothesis implies the slightly smaller bound $S(T)=O(\log T/\log\log T)$ (Titchmarsh 1985). The true order of magnitude may be somewhat less than this, as random functions with the same distribution as $S(T)$ tend to have growth of order about $\log(T)^{1/2}$. In the other direction it cannot be too small: Selberg (1946) showed that $S(T) \neq o((\log T)^{1/3}/(\log\log T)^{7/3})$, and assuming the Riemann hypothesis Montgomery showed that $S(T) \neq o((\log T)^{1/2}/(\log\log T)^{1/2})$.

Numerical calculations confirm that S grows very slowly: $|S(T)| < 1$ for $T < 280$, $|S(T)| < 2$ for $T < 6800000$, and the largest value of $|S(T)|$ found so far is not much larger than 3 (Odlyzko 2002).

Riemann's estimate $S(T) = O(\log T)$ implies that the gaps between zeros are bounded, and Littlewood improved this slightly, showing that the gaps between their imaginary parts tends to 0.

The theorem of Hadamard and de la Vallée-Poussin

Hadamard (1896) and de la Vallée-Poussin (1896) independently proved that no zeros could lie on the line $\mathrm{Re}(s) = 1$. Together with the functional equation and the fact that there are no zeros with real part greater than 1, this showed that all non-trivial zeros must lie in the interior of the critical strip 0 < Re(*s*) < 1. This was a key step in their first proofs of the prime number theorem.

Both the original proofs that the zeta function has no zeros with real part 1 are similar, and depend on showing that if ζ(1+*it*) vanishes, then ζ(1+2*it*) is singular, which is not possible. One way of doing this is by using the inequality

$$|\zeta(\sigma)^3\zeta(\sigma+it)^4\zeta(\sigma+2it)| \geq 1$$ for σ>1, *t* real,

and looking at the limit as σ tends to 1. This inequality follows by taking the real part of the log of the Euler product to see that

$$|\zeta(\sigma+it)| = \exp \Re \sum_{p^n} \frac{p^{-n(\sigma+it)}}{n} = \exp \sum_{p^n} \frac{p^{-n\sigma}\cos(t\log p^n)}{n}$$

(where the sum is over all prime powers p^n) so that

$$|\zeta(\sigma)^3\zeta(\sigma+it)^4\zeta(\sigma+2it)| = \exp \sum_{p^n} p^{-n\sigma}\frac{3+4\cos(t\log p^n)+\cos(2t\log p^n)}{n}$$

which is at least 1 because all the terms in the sum are positive, due to the inequality

$$3+4\cos(\theta)+\cos(2\theta) = 2(1+\cos(\theta))^2 \geq 0.$$

Zero-free regions

De la Vallée-Poussin (1899-1900) proved that if σ+*it* is a zero of the Riemann zeta function, then 1-σ ≥ C/log(*t*) for some positive constant *C*. In other words zeros cannot be too close to the line σ=1: there is a zero-free region close to this line. This zero-free region has been enlarged by several authors. Ford (2002) gave a version with explicit numerical constants: ζ(σ + i*t*) ≠ 0 whenever |*t*| ≥ 3 and

$$\sigma \geq 1 - \frac{1}{57.54(\log|t|)^{2/3}(\log\log|t|)^{1/3}}.$$

Zeros on the critical line

Hardy (1914) and Hardy & Littlewood (1921) showed there are infinitely many zeros on the critical line, by considering moments of certain functions related to the zeta function. Selberg (1942) proved that at least a (small) positive proportion of zeros lie on the line. Levinson (1974) improved this to one-third of the zeros by relating the zeros of the zeta function to those of its derivative, and Conrey (1989) improved this further to two-fifths.

Most zeros lie close to the critical line. More precisely, Bohr & Landau (1914) showed that for any positive ε, all but an infinitely small proportion of zeros lie within a distance ε of the critical line. Ivić (1985) gives several more precise versions of this result, called **zero density estimates**, which bound the number of zeros in regions with imaginary part at most T and real part at least 1/2+ε.

The Hardy-Littlewood conjectures

In 1914 Godfrey Harold Hardy proved that $\zeta\left(\tfrac{1}{2}+it\right)$has infinitely many real zeros.

Let $N(T)$ be the total number of real zeros, $N_0(T)$ be the total number of zeros of odd order of the function $\zeta\left(\tfrac{1}{2}+it\right)$, lying on the interval $(0,T]$.

The next two conjectures of Hardy and John Edensor Littlewood on the distance between real zeros of $\zeta\left(\tfrac{1}{2}+it\right)$and on the density of zeros of $\zeta\left(\tfrac{1}{2}+it\right)$on intervals $(T,T+H]$ for sufficiently great $T>0$, $H=T^{a+\varepsilon}$and with as less as possible value of $a>0$, where $\varepsilon>0$is an arbitrarily small number, open two new directions in the investigation of the Riemann zeta function:

1. for any $\varepsilon>0$there exists $T_0=T_0(\varepsilon)>0$such that for $T\geq T_0$and $H=T^{0.25+\varepsilon}$the interval $(T,T+H]$ contains a zero of odd order of the function $\zeta\left(\tfrac{1}{2}+it\right)$.

2. for any $\varepsilon>0$there exist $T_0=T_0(\varepsilon)>0$and $c=c(\varepsilon)>0$, such that for $T\geq T_0$and $H=T^{0.5+\varepsilon}$the inequality $N_0(T+H)-N_0(T)\geq cH$is true.

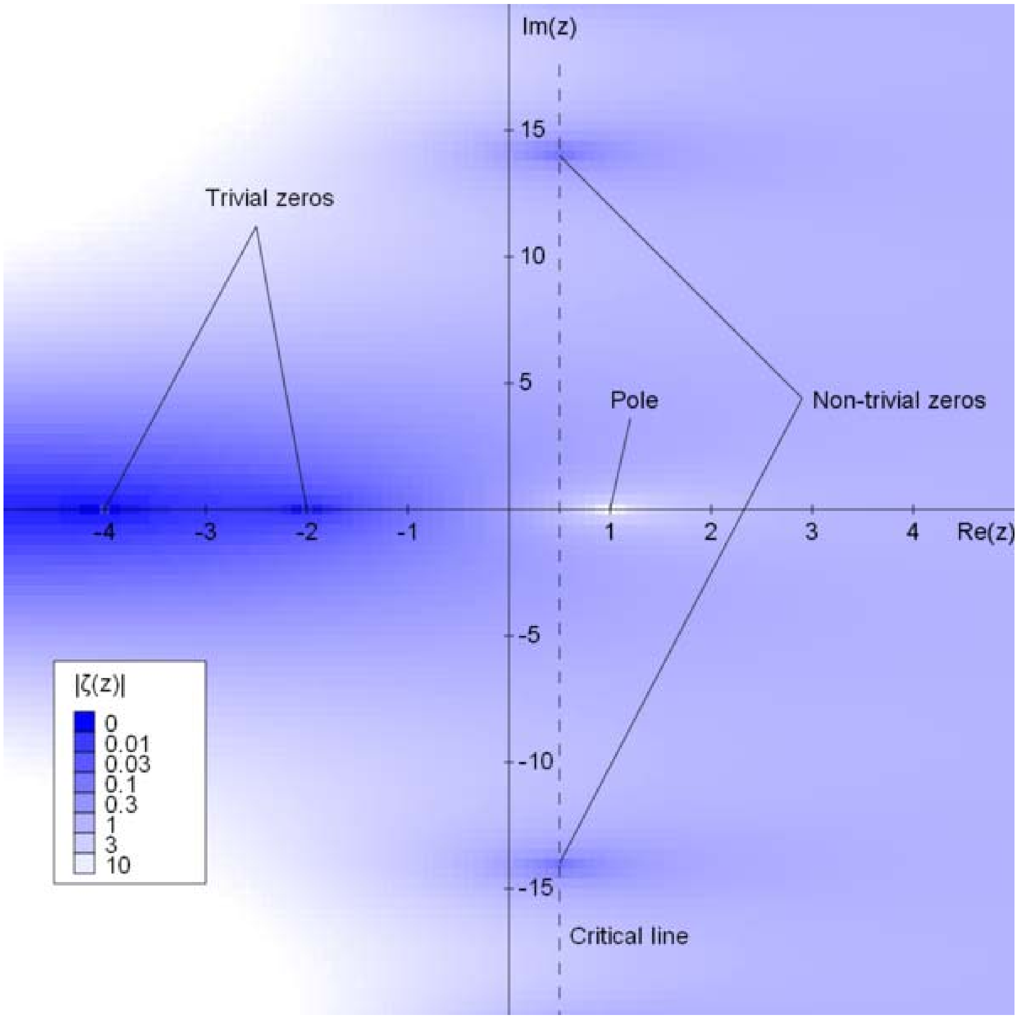
The Selberg conjecture

Atle Selberg (1942) investigated the problem of Hardy-Littlewood **2** and proved that for any $\varepsilon>0$there exists such $T_0=T_0(\varepsilon)>0$and $c=c(\varepsilon)>0$, such that for $T\geq T_0$and $H=T^{0.5+\varepsilon}$the inequality $N(T+H)-N(T)\geq cH\log T$is true. Selberg conjectured that this could be tightened to $H=T^{0.5}$. A. A. Karatsuba (1984a, 1984b, 1985) proved that for a fixed εsatisfying the condition $0<\varepsilon<0.001$, a sufficiently large T and $H=T^{a+\varepsilon}$, $a=\frac{27}{82}=\frac{1}{3}-\frac{1}{246}$, the interval $(T,T+H)$ contains at least $cH\ln T$ real zeros of the Riemann zeta function $\zeta\left(\tfrac{1}{2}+it\right)$and therefore confirmed the Selberg conjecture. The estimates of Selberg and Karatsuba can not be improved in respect of the order of growth as $T\to+\infty$.

Karatsuba (1992) proved that an analog of the Selberg conjecture holds for almost all intervals $(T,T+H]$, $H=T^{\varepsilon}$, where $\varepsilon$is an arbitrarily small fixed positive number. The

Karatsuba method permits to investigate zeros of the Riemann zeta-function on "supershort" intervals of the critical line, that is, on the intervals (T,$T + H$], the length H of which grows slower than any, even arbitrarily small degree T. In particular, he proved that for any given numbers ε, $\varepsilon_1$satisfying the conditions $0 < \varepsilon, \varepsilon_1 < 1$almost all intervals ($T$,$T + H$] for $H \geq \exp\{(\ln T)^{\varepsilon}\}$contain at least $H(\ln T)^{1-\varepsilon_1}$zeros of the function $\zeta(\frac{1}{2} + it)$. This estimate is quite close to the one that follows from the Riemann hypothesis.

Numerical calculations



Absolute value of the ζ-function

The function

$$\pi^{-s/2}\Gamma(s/2)\zeta(s)$$

has the same zeros as the zeta function in the critical strip, and is real on the critical line because of the functional equation, so one can prove the existence of zeros exactly on the real line between two points by checking numerically that the function has opposite signs at these points. Usually one writes

$$\zeta(1/2+it)=Z(t)e^{-i\pi\theta(t)}$$

where Hardy's function *Z* and the Riemann-Siegel theta function θ are uniquely defined by this and the condition that they are smooth real functions with θ(0)=0. By finding many intervals where the function *Z* changes sign one can show that there are many zeros on the critical line. To verify the Riemann hypothesis up to a given imaginary part *T* of the zeros, one also has to check that there are no further zeros off the line in this region. This can be done by calculating the total number of zeros in the region and checking that it is the same as the number of zeros found on the line. This allows one to verify the Riemann hypothesis computationally up to any desired value of *T* (provided all the zeros of the zeta function in this region are simple and on the critical line).

Some calculations of zeros of the zeta function are listed below. So far all zeros that have been checked are on the critical line and are simple. (A multiple zero would cause problems for the zero finding algorithms, which depend on finding sign changes between zeros.)

Year	Number of zeros	Author
1859?	3	B. Riemann used the Riemann-Siegel formula (unpublished, but reported in Siegel 1932).
1903	15	J. P. Gram (1903) used Euler–Maclaurin summation and discovered Gram's law. He showed that all 10 zeros with imaginary part at most 50 range lie on the critical line with real part 1/2 by computing the sum of the inverse 10th powers of the roots he found.
1914	79 ($\gamma_n \leq 200$)	R. J. Backlund (1914) introduced a better method of checking all the zeros up to that point are on the line, by studying the argument *S*(*T*) of the zeta function.
1925	138 ($\gamma_n \leq 300$)	J. I. Hutchinson (1925) found the first failure of Gram's law, at the Gram point g_{126}.
1935	195	E. C. Titchmarsh (1935) used the recently rediscovered Riemann-Siegel formula, which is much faster than Euler–Maclaurin summation.It takes about $O(T^{3/2+\varepsilon})$ steps to check zeros with imaginary part less than *T*, while the Euler–Maclaurin method takes about $O(T^{2+\varepsilon})$ steps.
1936	1041	E. C. Titchmarsh (1936) and L. J. Comrie were the

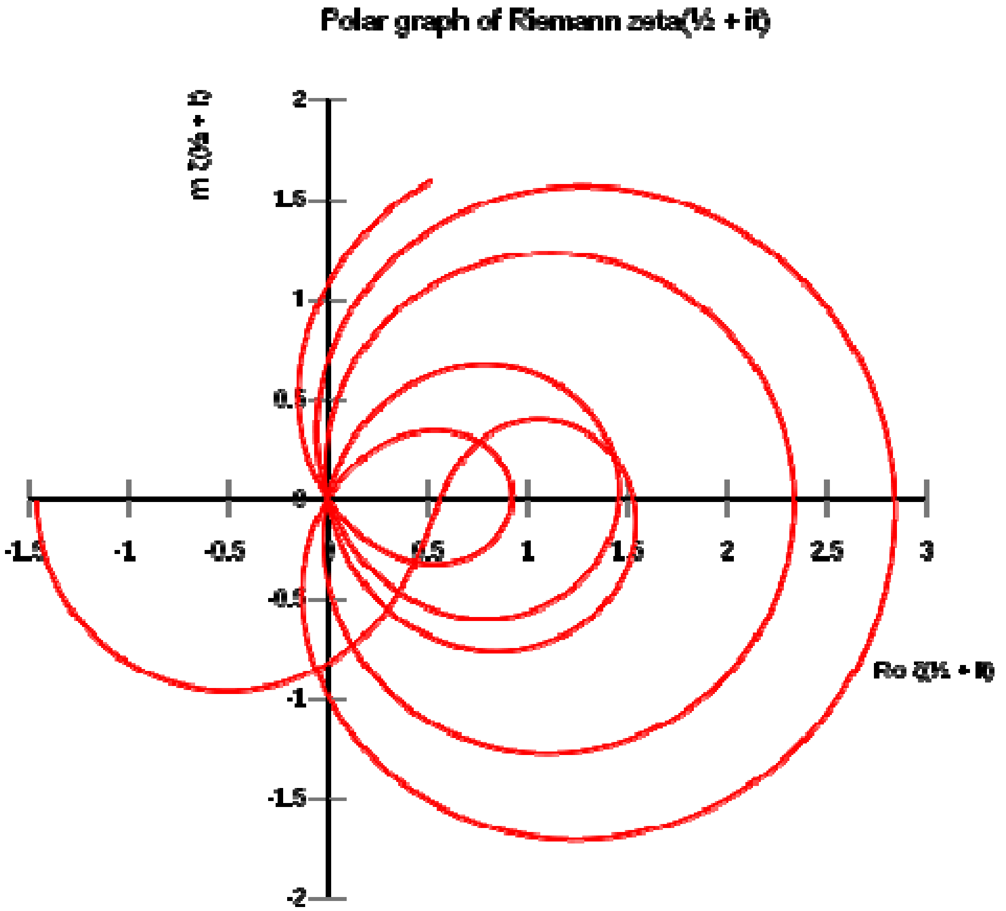
		last to find zeros by hand.
1953	1104	A. M. Turing (1953) found a more efficient way to check that all zeros up to some point are accounted for by the zeros on the line, by checking that *Z* has the correct sign at several consecutive Gram points and using the fact that *S*(*T*) has average value 0. This requires almost no extra work because the sign of *Z* at Gram points is already known from finding the zeros, and is still the usual method used. This was the first use of a digital computer to calculate the zeros.
1956	15000	D. H. Lehmer (1956) discovered a few cases where the zeta function has zeros that are "only just" on the line: two zeros of the zeta function are so close together that it is unusually difficult to find a sign change between them. This is called "Lehmer's phenomenon", and first occurs at the zeros with imaginary parts 7005.063 and 7005.101, which differ by only .04 while the average gap between other zeros near this point is about 1.
1956	25000	D. H. Lehmer
1958	35337	N. A. Meller
1966	250000	R. S. Lehman
1968	3500000	Rosser, Yohe & Schoenfeld (1969) stated Rosser's rule (described below).
1977	40000000	R. P. Brent
1979	81000001	R. P. Brent
1982	200000001	R. P. Brent, J. van de Lune, H. J. J. te Riele, D. T. Winter
1983	300000001	J. van de Lune, H. J. J. te Riele
1986	1500000001	van de Lune, te Riele & Winter (1986) gave some statistical data about the zeros and give several graphs of *Z* at places where it has unusual behavior.
1987	A few of large (~10^{12}) height	A. M. Odlyzko (1987) computed smaller numbers of zeros of much larger height, around 10^{12}, to high precision to check Montgomery's pair correlation conjecture.
1992	A few of large (~10^{20}) height	A. M. Odlyzko (1992) computed a 175 million zeroes of heights around 10^{20} and a few more of heights around 2×10^{20}, and gave an extensive discussion of the results.

1998	10000 of large (~10^{21}) height	A. M. Odlyzko (1998) computed some zeros of height about 10^{21}
2001	10000000000	J. van de Lune (unpublished)
2004	900000000000	S. Wedeniwski (ZetaGrid distributed computing)
2004	10000000000000 and a few of large (up to ~10^{24}) heights	X. Gourdon (2004) and Patrick Demichel used the Odlyzko–Schönhage algorithm. They also checked two billion zeros around heights 10^{13}, 10^{14}, ... , 10^{24}.

Gram points

A Gram point is a value of t such that $\zeta(1/2 + it) = Z(t)e^{-i\theta(t)}$ is a non-zero real; these are easy to find because they are the points where the Euler factor at infinity $\pi^{-s/2}\Gamma(s/2)$ is real at $s = 1/2 + it$, or equivalently $\theta(t)$ is a multiple $n\pi$ of π. They are usually numbered as g_n for $n = -1, 0, 1, ...$, where g_n is the unique solution of $\theta(t) = n\pi$ with $t \geq 8$ (θ is increasing beyond this point; there is a second point with $\theta(t) = -\pi$ near 3.4, and $\theta(0) = 0$). Gram observed that there was often exactly one zero of the zeta function between any two Gram points; Hutchinson called this observation **Gram's law**. There are several other closely related statements that are also sometimes called Gram's law: for example, $(-1)^n Z(g_n)$ is usually positive, or $Z(t)$ usually has opposite sign at consecutive Gram points. The imaginary parts γ_n of the first few zeros (in blue) and the first few Gram points g_n are given in the following table

		g_{-1}	γ_1	g_0	γ_2	g_1	γ_3	g_2	γ_4	g_3	γ_5	g_4	γ_6	g_5
0	3.4	9.667	14.135	17.846	21.022	23.170	25.011	27.670	30.425	31.718	32.935	35.467	37.586	38.999



This shows the values of ζ(1/2+it) in the complex plane for 0 ≤ t ≤ 34. (For t=0, ζ(1/2) ≈ -1.460 corresponds to the leftmost point of the red curve.) Gram's law states that the curve usually crosses the real axis once between zeros.

The first failure of Gram's law occurs at the 127'th zero and the Gram point g_{126}, which are in the "wrong" order.

g_{124}	γ_{126}	g_{125}	g_{126}	$\boldsymbol{\gamma_{127}}$	γ_{128}	g_{127}	γ_{129}	g_{128}
279.148	279.229	280.802	282.455	**282.465**	283.211	284.104	284.836	285.752

A Gram point *t* is called good if the zeta function is positive at 1/2 + *it*. The indices of the "bad" Gram points where *Z* has the "wrong" sign are 126, 134, 195, 211,... (sequence A114856 in OEIS). A **Gram block** is an interval bounded by two good Gram points such that all the Gram points between them are bad. A refinement of Gram's law called Rosser's rule due to Rosser, Yohe & Schoenfeld (1969) says that Gram blocks often have the expected number of zeros in them (the same as the number of Gram intervals), even though some of the individual Gram intervals in the block may not have exactly one zero in them. For example, the interval bounded by g_{125} and g_{127} is a Gram block containing a unique bad Gram point g_{126}, and contains the expected number 2 of zeros although neither of its two Gram intervals contains a unique zero. Rosser et al. checked that there were no exceptions to Rosser's rule in the first 3 million zeros, although there are infinitely many exceptions to Rosser's rule over the entire zeta function.

Gram's rule and Rosser's rule both say that in some sense zeros do not stray too far from their expected positions. The distance of a zero from its expected position is controlled by the function *S* defined above, which grows extremely slowly: its average value is of the order of $(\log \log T)^{1/2}$, which only reaches 2 for T around 10^{24}. This means that both rules hold most of the time for small *T* but eventually break down often.

Arguments for and against the Riemann hypothesis

Mathematical papers about the Riemann hypothesis tend to be cautiously noncommittal about its truth. Of authors who express an opinion, most of them, such as Riemann (1859) or Bombieri (2000), imply that they expect (or at least hope) that it is true. The few authors who express serious doubt about it include Ivić (2008) who lists some reasons for being skeptical, and Littlewood (1962) who flatly states that he believes it to be false, and that there is no evidence whatever for it and no imaginable reason for it to be true. The consensus of the survey articles (Bombieri 2000, Conrey 2003, and Sarnak 2008) is that the evidence for it is strong but not overwhelming, so that while it is probably true there is some reasonable doubt about it.

Some of the arguments for (or against) the Riemann hypothesis are listed by Sarnak (2008), Conrey (2003), and Ivić (2008), and include the following reasons.

- Several analogues of the Riemann hypothesis have already been proved. The proof of the Riemann hypothesis for varieties over finite fields by Deligne (1974) is possibly the single strongest theoretical reason in favor of the Riemann hypothesis. This provides some evidence for the more general conjecture that all zeta functions associated with automorphic forms satisfy a Riemann hypothesis, which includes the classical Riemann hypothesis as a special case. Similarly Selberg zeta functions satisfy the analogue of the Riemann hypothesis, and are in some ways similar to the Riemann zeta function, having a functional equation and an infinite product expansion analogous to the Euler product expansion. However there are also some major differences; for example they are not given by Dirichlet series. The Riemann hypothesis for the Goss zeta function was proved by Sheats (1998). In contrast to these positive examples, however, some Epstein zeta functions do not satisfy the Riemann hypothesis, even though they have an infinite number of zeros on the critical line (Titchmarsh 1986). These functions are quite similar to the Riemann zeta function, and have a Dirichlet series expansion and a functional equation, but the ones known to fail the Riemann hypothesis do not have an Euler product and are not directly related to automorphic representations.
- The numerical verification that many zeros lie on the line seems at first sight to be strong evidence for it. However analytic number theory has had many conjectures supported by large amounts of numerical evidence that turn out to be false. See

Skewes number for a notorious example, where the first exception to a plausible conjecture related to the Riemann hypothesis probably occurs around 10^{316}; a counterexample to the Riemann hypothesis with imaginary part this size would be far beyond anything that can currently be computed. The problem is that the behavior is often influenced by very slowly increasing functions such as log log *T*, that tend to infinity, but do so so slowly that this cannot be detected by computation. Such functions occur in the theory of the zeta function controlling the behavior of its zeros; for example the function *S*(*T*) above has average size around $(\log \log T)^{1/2}$. As *S*(*T*) jumps by at least 2 at any counterexample to the Riemann hypothesis, one might expect any counterexamples to the Riemann hypothesis to start appearing only when *S*(*T*) becomes large. It is never much more than 3 as far as it has been calculated, but is known to be unbounded, suggesting that calculations may not have yet reached the region of typical behavior of the zeta function.

- Denjoy's probabilistic argument for the Riemann hypothesis (Edwards 1974): If μ(*x*) is a random sequence of "1"s and "−1"s then, for every ε > 0, the function

  $$M(x) = \sum_{n \le x} \mu(n)$$

  (the values of which are positions in a simple random walk) satisfies the bound

  $$M(x) = O(x^{1/2+\varepsilon})$$

  with probability 1. The Riemann hypothesis is equivalent to this bound for the Möbius function μ and the Mertens function *M* derived in the same way from it. In other words, the Riemann hypothesis is in some sense equivalent to saying that μ(*x*) behaves like a random sequence of coin tosses. When μ(*x*) is non-zero its sign gives the parity of the number of prime factors of *x*, so informally the Riemann hypothesis says that the parity of the number of prime factors of an integer behaves randomly. Such probabilistic arguments in number theory often give the right answer, but tend to be very hard to make rigorous, and occasionally give the wrong answer for some results, such as Maier's theorem.

- The calculations in Odlyzko (1987) show that the zeros of the zeta function behave very much like the eigenvalues of a random Hermitian matrix, suggesting that they are the eigenvalues of some self-adjoint operator, which would imply the Riemann hypothesis. However all attempts to find such an operator have failed.
- There are several theorems, such as Goldbach's conjecture for sufficiently large odd numbers, that were first proved using the generalized Riemann hypothesis, and later shown to be true unconditionally. This could be considered as weak evidence for the generalized Riemann hypothesis, as several of its "predictions" turned out to be true.
- Lehmer's phenomenon (Lehmer 1956) where two zeros are sometimes very close is sometimes given as a reason to disbelieve in the Riemann hypothesis. However one would expect this to happen occasionally just by chance even if the Riemann hypothesis were true, and Odlyzko's calculations suggest that nearby pairs of zeros occur just as often as predicted by Montgomery's conjecture.

- Patterson (1988) suggests that the most compelling reason for the Riemann hypothesis for most mathematicians is the hope that primes are distributed as regularly as possible.