CAMBRIDGE

Cambridge University Press
978-0-521-64633-8 - The Algorithmic Resolution of Diophantine Equations
Nigel P. Smart
Excerpt
More information

CHAPTER I

Introduction

This book shall concern itself with the study of modern methods for solving diophantine equations. The study of diophantine equations goes back to the ancient Greeks. The most famous example from that time, $X^2 + Y^2 = Z^2$, is still being taught in schools today. Many of the ideas in this book can be traced back to earlier times, so I shall start by giving a brief outline of the history of the subject. This will be to set the scene and raise the problems that will hopefully be answered in the following chapters.

By a diophantine equation we mean, intuitively, an equation where we are interested only in integer or rational solutions. For example, Fermat's famous 'last theorem', now Wiles' theorem, says that the only integer solutions to the equation

$$X^n + Y^n = Z^n \tag{I.1}$$

with $n \geq 3$ are given by $XYZ = 0$. Another important class of examples is elliptic curves, which are curves of the form

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6.$$

When studying equation (I.1) it clearly makes no difference if we study the rational solutions (X, Y, Z) or the integral solutions. However, when looking at elliptic curves it makes a great deal of difference whether we want to determine the rational or integral solutions. An elliptic curve can (and often does) possess an infinite number of rational solutions, but it will only ever possess a finite number of integral solutions, as we shall see in a later chapter.

Factoring an integer can be considered as solving a diophantine equation. Given an integer N the problem of factoring can be presented as finding the integral solutions to the equation

$$N = pq$$

where $p, q \in \mathbb{N}$.

Diophantine equations have over the centuries provided a fertile ground for mathematical investigation. This is at first glance surprising, as finding solutions to an equation in the real numbers appears easy. We can, for instance, just draw a graph, and the integers are considered a much simpler mathematical object than the reals.

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-64633-8 - The Algorithmic Resolution of Diophantine Equations
Nigel P. Smart
Excerpt
More information

I.1. A brief history

The study of diophantine equations dates back to at least 1600BC. The earliest work of importance seems to have been on the problem of determining 'Pythagorean' triples, that is, non-trivial solutions (X, Y, Z) to the equation

$$X^2 + Y^2 = Z^2.$$

Any school child knows about the triple $(3, 4, 5)$, while any undergraduate would understand the proof that all triples are given by (up to an interchange of x and y)

$$\begin{aligned} x &= \pm d(a^2 - b^2), \\ y &= \pm 2abd, \\ z &= \pm d(a^2 + b^2), \end{aligned}$$

where $a, b, d \in \mathbb{Z}$ with $\gcd(a, b) = 1$.

The name diophantine equations is in honour of the mathematician Diophantos, who lived in Alexandria around 300AD. Diophantos' work *Arithmetica* was one of the ancient texts that went 'missing' in Europe in the Dark and Middle Ages. The *Arithmetica* originally consisted of 13 books, of which only 6 have survived into the modern era. Two translations of the remaining books were made in the sixteenth and seventeenth centuries. It was in the margin of Fermat's copy of Bachet's translation of *Arithmetica* that Fermat made his famous marginal note that equation (I.1) has no non-trivial solutions.

Pierre de Fermat (1601–65) gave a large number of legacies to mathematics, and in particular number theory, the most famous of these being the above–mentioned last theorem. More important was his introduction of the so called 'method of descent'. In this method one supposes one has a solution which is as 'small' as possible, and then one produces by some means an even 'smaller' solution. This contradiction tells us that our original solution could not have existed in the first place. Fermat applied his method of descent to show

THEOREM I.1 (Fermat). *The equation*

$$z^2 = x^4 + y^4$$

has no non-trivial integer solutions.

PROOF. Suppose that there is a non–trivial solution. We can clearly assume that it satisfies $(x, y) = 1$ and without loss of generality we can assume that x is odd, y is even and both are positive. By applying the formulae for pythagorean triples given above, we can then write

$$x^2 = a^2 - b^2 \ , \ y^2 = 2ab \ , \ z = a^2 + b^2,$$

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-64633-8 - The Algorithmic Resolution of Diophantine Equations
Nigel P. Smart
Excerpt
More information

for two coprime integers a and b. We then apply the formulae again to the equation $x^2 + b^2 = a^2$ to obtain

$$a = p^2 + q^2 \ , \ x = p^2 - q^2 \ , \ b = 2pq,$$

where p and q are two coprime integers. Since y is even we obtain

$$\left(\frac{y}{2}\right)^2 = \frac{ab}{2} = pq(p^2 + q^2),$$

which leads us deduce, as p, q and p^2+q^2 are coprime, that there exist positive integers X, Y and Z such that

$$p = X^2 \ , \ q = Y^2 \ , \ p^2 + q^2 = Z^2,$$

and so (X, Y, Z) is another solution to

$$Z^2 = X^4 + Y^4.$$

To sum up, we have from one solution to our equation deduced another solution to our equation. The trick of the proof is to show that this new solution is 'smaller' than the original one. If we can do this then this method of descending to a 'smaller' solution cannot be carried on indefinitely, and so the original solution could not have existed in the first place.

Note that

$$y = 2XY\sqrt{X^4 + Y^4},$$

so if either X or Y is zero then y is also zero, which would mean that (x, y, z) was a trivial solution. Hence neither X nor Y can be zero. It is then clear that $X < y$ and $Y < y$. So the new solution must be 'smaller' than the old solution. □

As a corollary we easily deduce that Fermat's last theorem holds for the exponent $n = 4$. The method of descent has since been adapted and now the name 'descent' is often given to any process whereby the existence or non-existence of solutions to some equation is proved by means of considering other, in most cases smaller, solutions to either the same equation or a related set of equations. Using the method of descent it is believed that Fermat managed to show that if p is a prime congruent to 1 modulo 4 then the equation

$$x^2 + y^2 = p$$

always has an integer solution. However, no proof of this result by Fermat survives; the earliest known proof dates back as far as Euler. This descent method uses a known solution to one equation to deduce a solution to a similar equation with smaller coefficients:

THEOREM I.2. *Let p denote a prime congruent to one modulo four. Then there exists a solution in integers to the equation*

$$p = x^2 + y^2.$$

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-64633-8 - The Algorithmic Resolution of Diophantine Equations
Nigel P. Smart
Excerpt
More information

PROOF. Clearly there is a solution (x_1, y_1) to the equation

$$mp = x_1^2 + y_1^2$$

for some positive integer value of m. For example we can take x_1 to denote a square root of -1 modulo p and y_1 to be 1. With this choice we can assume that $m < p$. Now choose two integers (u, v) such that $u \equiv x_1 \pmod{m}$, $v \equiv y_1 \pmod{m}$ and $u, v \in [-m/2, m/2]$. We then have that

$$u^2 + v^2 \equiv x_1^2 + y_1^2 \pmod{m},$$

so that

$$u^2 + v^2 = rm$$

for some positive integer value of $r < m$. Now set

$$x_1' = x_1 u + y_1 v \ , \ y_1' = x_1 v - y_1 u.$$

Then we notice that both x_1' and y_1' are multiplies of m. So we set $x_1' = mx_2$ and $y_1' = my_2$. But then

$$\begin{aligned} m^2(x_2^2 + y_2^2) &= x_1'^2 + y_1'^2 = (x_1 u + y_1 v)^2 + (x_1 v - y_1 u)^2 \\ &= (x_1^2 + y_1^2)(u^2 + v^2) \\ &= rpm^2. \end{aligned}$$

Hence (x_2, y_2) is a solution to the equation

$$x_2^2 + y_2^2 = rp,$$

where $r < m$. We can continue carrying out this process, but not indefinitely, as the values of r are positive and get successively smaller. Hence at some point we will reach $r = 1$ and we will have a solution to our equation. □

Euler and Lagrange also gave a proof of a result which had been asserted by Fermat, namely that every integer could be written as the sum of four rational squares. Fermat had claimed he had a proof of this result which also used his method of descent.

Our story now jumps forward a century to the time of Gauss, who was born in Braunschweig in 1777. Gauss has obtained the reputation of being one of the most original mathematicians in history. Up to his death, in 1855, he worked in various areas of mathematics and physics such as algebra, magnetism and probability. However, it is his work in number theory which interests us. Gauss studied the integer solutions to quadratic equations such as

$$ax^2 + bxy + cy^2 + dx + ey + f = 0,$$

where a, b, c, d, e, f are given integers. This includes the case of Pell's equation

$$y^2 - Dx^2 = 1,$$

which provides a useful motivating example in undergraduate courses for the topics of quadratic fields and continued fractions. Some examples of Pell's equation had even been solved in Diophantos's *Arithmetica*, and some had

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-64633-8 - The Algorithmic Resolution of Diophantine Equations
Nigel P. Smart
Excerpt
More information

been studied by the Hindu mathematicians Brahmagupta and Bhaskara in the seventh and twelfth centuries respectively. It was Euler who first spotted the link between solutions of Pell's equation and the continued fraction expansion of the quadratic irrational $\sqrt{D}$, a link which is sometimes used today to compute fundamental units of real quadratic number fields. Gauss developed the arithmetic of quadratic number fields, in particular their class groups. However, he expressed everything in terms of the theory of binary quadratic forms. The development of the notion of ideals was to come after Gauss.

The challenge of Fermat's last theorem led, amongst other problems, to the development in the nineteenth century of the subject known as algebraic number theory. This subject provided the testing ground for much of modern algebra. The development of the notions of rings, ideals, modules, unique factorization domains and other basic notions can be traced back to the investigation of algebraic number fields. The list of mathematicians involved in this theory, Dedekind, Dirichlet, Galois, Kummer, Minkowski, etc., can be recounted by any undergraduate number theorist.

Another interest of Fermat was the study of elliptic curves. That the rational points form a group has been known for over a century, as the group law can be deduced from the classical chord–tangent process, a process linked to the addition formulae for elliptic integrals. Weierstrass (1815–97) had studied such elliptic functions and had expanded on the work of the Norwegian mathematician Abel (1802–29). Abel had discovered the class of transcendental functions which we now call Abelian functions, of which elliptic functions are an example. Earlier Gauss had seen the link between elliptic integrals and his arithmetic–geometric mean, but his work in this area was not published until after his death (and after the publication of Abel's work).

A similar process to the chord–tangent process had been used by Fermat to deduce more solutions to an elliptic curve given a known solution, this being known as Fermat's method of ascent. The name ascent comes from the fact that the process, in general, produces larger and larger rational solutions from known smaller ones.

Jules Henri Poincaré (1854–1912), although more famous for his works in topology and mathematical physics, in the late 1800s conjectured that the set of rational points on any given elliptic curve, including an additional point 'at infinity', formed a finitely generated abelian group. In 1922 Mordell proved Poincaré's conjecture using a technique based on Fermat's method of descent.

The first half of the twentieth century also saw the introduction of 'local methods' into number theory by Hasse, Hensel, Skolem and others. These were used, especially by Skolem, to find all the solutions to a large number of individual equations. Indeed Skolem's method was still the main one in use up to the late 1970s, and it is still of relevance today. Later, in Chapter III, we shall give some examples of the use of Skolem's method to solve equations.

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-64633-8 - The Algorithmic Resolution of Diophantine Equations
Nigel P. Smart
Excerpt
More information

In 1913, Ramanujan asked what are the positive integral solutions to the diophantine equation

$$x^2 + 7 = 2^n.$$

In 1948 Nagell showed that there are only five positive integral solutions given by

$$(n, x) = (3, 1), (4, 3), (5, 5), (7, 11), (15, 181).$$

This equation is now called the Ramanujan–Nagell equation. A proof that these are the only five positive solutions of the Ramanujan–Nagell equation can be found in many standard textbooks, such as [**187**]. Various authors have considered generalizations of the Ramanujan–Nagell equation, the most general form being that considered by Pethő and de Weger,

$$x^2 + k = p_1^{z_1} \cdots p_t^{z_t}.$$

In [**151**] a general method is given to deal with this generalized Ramanujan–Nagell equation. This shows a marked change of emphasis, from the study of individual equations to the study of generalized classes of equations.

Prior to the Second World War, if the solutions to an equation were able to be determined explicitly then usually some special trick was used which only applied to that equation. If you wanted to find the solutions to a similar equation, you had to find a similar trick which worked in this new case. Very few methods, Skolem's being the notable exception, could be applied to a large number of equations without alteration.

Given a diophantine equation, or even a class of equations such as 'all elliptic curves', there are some natural questions that come to mind:

1. Is the number of rational (or integral) solutions finite or infinite?
2. If the number of rational (or integral) solutions is finite, can we give a procedure which in a finite amount of time will determine all the solutions?
3. If the number of rational (or integral) solutions is infinite, is it possible to express all solutions in terms of some 'basic' ones? For instance, given Mordell's result that the rational points on an elliptic curve form a finitely generated group, can we construct explicit generators for this group in any given example?

Answering either of the final two questions clearly is a harder task than the first one.

If we can prove that an equation has a finite number of solutions in a way which gives an algorithm to determine all the solutions, then one is said to have an **effective** proof. If on the other hand one can only answer the question of the finiteness of the number of solutions without giving an algorithm, then one is said to have an **ineffective** proof. Hilbert, in the tenth of his famous problems at the turn of the century, asked whether there existed an algorithm which could determine whether any diophantine equation had finitely or infinitely many solutions.

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-64633-8 - The Algorithmic Resolution of Diophantine Equations
Nigel P. Smart
Excerpt
More information

In 1970 Matijasevič proved the non-existence of such an algorithm. See [**45**] for a general discussion on Hilbert's tenth problem and Matijasevič's solution. However, this still left the question open as to whether large classes of equations could be tackled using algorithms.

Three years before Matijasevič's proof Baker [**4**] had given an effective proof of a result of Thue. Thue [**196**] had shown in an ineffective way that there are only finitely many integer solutions to equations of the form

$$F(X,Y) = m,$$

where $F(X,Y)$ is a binary form of degree greater than 2 and m is some fixed integer, for example, equations such as

$$X^3 + 2XY^2 + Y^3 = 2$$

or

$$X^{16} - 33Y^{16} = 42775.$$

Such equations are now called Thue equations. Baker showed how one could bound the size of the solutions in terms of the coefficients of F and the integer m. Hence there is a finite search region within which all the solutions must lie. However, the bounds given by Baker are huge and certainly not meant for a practical solution of Thue equations.

As Baker's method results in a finite search region, it was not long before people started thinking of ways of reducing the size of this region. This was began in the early 1970s with work of Baker, Davenport and Ellison, see [**5**] and [**49**] . However, in the mid–1980s after the development of the LLL–algorithm by Lenstra, Lenstra and Lovász [**117**] a new technique was available. This technique, which was pioneered in de Weger's thesis [**208**] allowed the computational resolution of equations which only a few years previously had been the preserve of theoreticians.

I.2. Algorithms

In this book we shall mainly be concerned with effective proofs. But we shall also be interested in techniques which not only tell us how we could compute all the solutions to an equation in principle but how we could do so in **practice**. Consider the following example. Suppose one could show that all integral solutions to the Thue equation

$$X^7 + 2Y^7 = 1$$

satisfied

$$|X|, |Y| \leq 10^{40}.$$

We would then know that there were finitely many solutions. Not only that but we could give a computer the task of computing all the solutions. If the computer could check whether a pair (X,Y) was a solution in, say, one billionth of a second, then it would take the computer around 10^{50} years to

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-64633-8 - The Algorithmic Resolution of Diophantine Equations
Nigel P. Smart
Excerpt
More information

go through all the possible solutions to the equation. Unfortunately this is much longer than the estimated age of the Universe.

Therefore we cannot be content only with effective methods. What will interest us is methods which can be applied in practice to actually determine all the solutions to a problem. Even here one has problems, as an algorithm which will determine all solutions in one example may not work in another. It could fail either because the algorithm does not apply or because the algorithm would take too long to be of any practical benefit.

We shall be interested in practical algorithms which apply to wide classes of equations and we will be satisfied if the algorithm works for a wide number of examples in the class. We shall not discuss the problems of growth of expected running time and other complexity theoretic issues. Indeed the complexity theoretic study of algorithms for diophantine equations is not anywhere near as well developed as it is for other areas in computational number theory, but we shall need an intuitive concept.

Upper bounds on run times of algorithms depend, of course, on the definition one is using for time. The standard definition is to measure time in terms of bit operations; this is often referred to as 'bit complexity'. We shall not really worry about the exact definitions from complexity theory, but we will have a need to measure and compare one algorithm's expected run time against another. For this comparison we will be content with just an intuitive understanding; for a more complete discussion of such matters you should consult [**3**].

We shall sometimes refer to an algorithm as running in polynomial or exponential time in some parameter. We shall now review this concept briefly, as it may be new to some readers. An algorithm is said to run in polynomial time, with respect to some parameter B, if its run time is bounded by a polynomial function of B; in other words its run time is $O(B^n)$ for some fixed number n. An algorithm is said to run in exponential time if it runs in time $O(n^B)$. An algorithm is said to run in subexponential time if its run time is certainly faster than exponential in behavior but not necessarily as good as polynomial in growth.

Normally the run time is measured in terms of the length of the input data. So an algorithm which has as input a single number N is said to run in polynomial time if we can bound the run time by $O((\log N)^n)$, for some fixed number n, while an algorithm runs in exponential time if we can bound the run time by $O(N^n)$, for some fixed n. This is because the length of the input to the algorithm is of size $O(\log N)$.

It is convenient to introduce the estimate

$$L_N(\alpha, \beta) = O\left(\left(e^{(\log N)^\alpha (\log\log N)^{1-\alpha}}\right)^{\beta+o(1)}\right).$$

This interpolates between polynomial time, $\alpha = 0$, and exponential time, $\alpha = 1$. This estimate occurs quite a lot in number theoretic algorithms. For example the best–known factoring algorithm has complexity $L_N(1/3, c)$ for

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-64633-8 - The Algorithmic Resolution of Diophantine Equations
Nigel P. Smart
Excerpt
More information

some constant c, while the best–known algorithms for computing the class group and unit group of a number field have complexity $L_D(1/2, c)$, where D is the discriminant of the field and one assumes the generalized Riemann hypothesis (GRH).

In this book three main techniques will be discussed for solving diophantine equations, being:

- The application of 'local' considerations. This includes Skolem's method and sieving.
- Reducing the problem of finding the solutions to one equation to the problem of finding the solutions to another set of hopefully easier equations. We shall meet this, for instance, when we discuss the method of descent for elliptic curves.
- The application of Baker's theory of linear forms in logarithms and the use of a method to reduce the huge bounds resulting from this theory.

Baker's technique itself divides into four main steps:

1. Reduce the solution to a problem for which we know Baker's techniques will apply.
2. Produce effective bounds on solutions using deep theoretical results of Baker, Yu, Waldschmidt, Wüstholz and others.
3. Reduce these large bounds to something more manageable, using a computational technique first developed by de Weger.
4. Find clever search techniques to find all solutions under the reduced bounds.

In recent years other techniques from areas such as arithmetic geometry have been developed. These allow one, for instance, to find all rational points on some curves of genus greater than one. That there are finitely many such rational points follows from the deep, but ineffective, work of Faltings. In later sections we shall consider these new methods and ideas and current open problems.

Many algorithms to find all the solutions to a particular equation require fast and efficient techniques for finding one (or a few) solutions to another set of equations. Sometimes we need to find all solutions below a certain upper bound. To solve this problem we will use a 'sieving' technique which can be found in various guises throughout this book. Often it is the sieve, which locates all small solutions or exhibits a single solution, which is the slowest part of the entire solution process.

I.3. What is a diophantine equation?

Often the above definition of a diophantine equation is too restrictive. We would like to consider generalizations of the intuitive notion of diophantine equation considered above. We shall now define the exact notion which shall be used in this book. What was important in our intuitive definition was that a diophantine equation was not only the equation but also the set for

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-64633-8 - The Algorithmic Resolution of Diophantine Equations
Nigel P. Smart
Excerpt
More information

which were trying to find solutions in. In the examples above this set was always $\mathbb{Z}$ or $\mathbb{Q}$. It makes sense to admit other sets as candidates for containing solutions.

Let K denote an algebraic number field and let S_i denote some well–defined subsets of K, for $i = 1, \ldots, n$. If $F(X_1, \ldots, X_n)$ denotes some function which maps $S_1 \times \cdots \times S_n$ to K, then we define a diophantine equation to be the equation

$$F(X_1, \ldots, X_n) = 0,$$

where we are interested in determining the structure of all the solutions in $S_1 \times \cdots \times S_n$.

There is a common equivalence we use for inhomogenous equations in two variables or homogenous equations in three variables. Such equations are usually thought of as curves, so we often speak of 'points on the curve' as meaning 'solutions to the equation'. Clearly our solutions (and hence points) can come from various sets, the term 'integral point' being reserved for points which have coordinates in the ring $\mathbb{Z}$.

For instance, we can use this to define what it means for an elliptic curve defined over a number field to have integral points. Let

$$Y^2 = X^3 + AX + B$$

denote an elliptic curve defined over K, by which we mean $A, B \in K$. Then determining all the solutions in $\mathcal{O}_K \times \mathcal{O}_K$ to the equation

$$F(X, Y) = Y^2 - X^3 - AX - B = 0$$

is what we mean by finding all the integral points on the elliptic curve, in this example $S_1 = S_2 = \mathcal{O}_K$. If we replace S_1 and S_2 by K, we are then faced with the task of determining the structure of the K-rational points on the elliptic curve.

In this vein one of the most important types of equation we shall meet is the two–term S-unit equations. In these equations we let S_1 and S_2 denote two finitely generated subgroups of K^* and let α_1, α_2 denote two fixed elements of K^*. By solving a two–term S-unit equation we mean determining all the solutions to the following equation, with $X_1 \in S_1$ and $X_2 \in S_2$:

$$F(X_1, X_2) = \alpha_1 X_1 + \alpha_2 X_2 + 1 = 0.$$

I.4. An elliptic curve

We end this introduction with an example of a diophantine equation which can be solved using standard techniques from undergraduate number theory. We shall then discuss just how lucky we are in this situation and the type of problems which can occur when one carries out the following ideas for other examples.

We look at the problem of finding all the integer solutions to a specific example of an elliptic curve. The example we have chosen is a standard

© Cambridge University Press www.cambridge.org