CHAPTER 1. INTRODUCTION.

1.1. Algorithms for diophantine equations.

This thesis deals with certain types of *diophantine equations*. An *equation* is a mathematical formula, expressing equality of two expressions that involve one or more unknowns (variables). *Solving* an equation means finding all *solutions*, i.e. the values that can be substituted for the unknowns such that the equation becomes a true statement. An equation is called a *diophantine* equation if the solutions are restricted to be *integers* in some sense, usually the ordinary rational integers (elements of $\mathbb{Z}$) or some subset of that.

Examples of diophantine equations that will be studied in this thesis are

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

(the Ramanujan-Nagell equation, having only the solutions given by $(\pm x,n) = (1,3), (3,4), (5,5), (11,7), (181,15)$, see Chapter 4);

$$2^x = 3^y + 5^z$$

(a purely exponential equation, having only the solutions $(x,y,z) = (1,0,0), (2,1,0), (3,1,1), (5,3,1), (7,1,3)$, see Chapter 6);

$$y^2 = x^3 - 4 \cdot x + 1$$

(an elliptic equation, having only 22 solutions, of which the largest are $(x,y) = (1274,\pm 45473)$, see Chapter 8). The three examples mentioned here are only examples of classes of equations which we study. We also study (in Chapter 5) a *diophantine inequality* (a formula expressing that one expression is larger than another, where solutions are again restricted to integers). In the following discussion the statements about diophantine equations also hold for this inequality.

What the equations treated in this book have in common is that they can all be solved by the same method. This method consists essentially of three

parts: a transformation step, a application of the Gelfond-Baker theory, and a diophantine approximation step. We explain these steps briefly.

First, one transforms the equation to a purely exponential equation or inequality, i.e. a diophantine equation or inequality where the unknowns are all in the exponents, such as in the second example given above. Each type of diophantine equation needs a particular kind of transformation, so that it is difficult to be more specific at this point. In some instances, such as in the second example above, this transformation is easy, if not trivial. In other instances, as in the first example above, it uses some arguments from algebraic number theory, or, as in the third example above, a lot of them.

In general, such a purely exponential equation has the form

$$\sum_{i=1}^{t} c_i \cdot \prod_{j=1}^{s_i} \alpha_{ij}^{n_{ij}} = c_0 \cdot \prod_{j=1}^{s_0} \alpha_{0j}^{n_{0j}} \quad , \tag{1.1}$$

and a corresponding purely exponential inequality looks like

$$\left| \sum_{i=1}^{t} c_i \cdot \prod_{j=1}^{s_i} \alpha_{ij}^{n_{ij}} \right| < \min_i \left| c_i \cdot \prod_{j=1}^{s_i} \alpha_{ij}^{n_{ij}} \right|^{\delta} \tag{1.2}$$

where t, s_i, c_i, α_{ij}, δ are constants with t, $s_i \in \mathbb{N}$, $0 < \delta < 1$, and c_i, α_{ij} belong to some algebraic extension of $\mathbb{Q}$, and where the n_{ij} are the unknowns in $\mathbb{Z}$. We now suppose that the number of terms t on the left hand side is equal to 2. This restriction is essential for the second step, in which we use results from the so-called theory of linear forms in logarithms, also known as the Gelfond-Baker theory. (Some special exponential equations with more terms can also be treated by the Gelfond-Baker method, by reducing them to exponential inequalities with two terms, cf. Stroeker and Tijdeman [1982], Alex [1985^a], [1985^b], Tijdeman and Wang [1987].)

An exponential equation or inequality such as (1.1) or (1.2) gives rise to a *linear form in logarithms*

$$\Lambda = \log \beta_0 + \sum_{i=1}^{m} n_i \cdot \log \beta_i \quad ,$$

where the β_i are algebraic constants, and the n_i are integral unknowns. Here, the logarithms can be real or complex, or can be p-adic. This relation between equation and linear form in logarithms is such that for a large solution of the equation the linear form is extremely close to zero (in the

real or complex sense, or in the p-adic sense). The Gelfond-Baker theory provides effectively computable lower bounds for the absolute values (respectively p-adic values) of such linear forms in logarithms of algebraic numbers. In many cases these bounds have been explicitly computed. Comparing the so-found upper and lower bounds it is possible to obtain explicit upper bounds for the solutions of the exponential diophantine equation or inequality, leading to upper bounds for the solutions of the original equation. This second step, unlike the first (transformation) step, is of a rather general nature.

We remark that many authors have given effectively computable upper bounds for the solutions of a wide variety of diophantine equations, by applying the method sketched above. For a survey, see Shorey and Tijdeman [1986]. Often these authors were satisfied with the knowledge of the existence of such bounds, and they did not actually compute them. If they computed bounds, they did not always determine all the solutions. In this thesis, solving an equation will always mean: explicitly finding all the solutions.

After the second step, the problem of solving the diophantine equation is reduced to a finite problem, which is the treated in the third part of the method. Namely, since we have found explicit upper bounds for the absolute values of the (integral) unknowns, we have to check only finitely many possibilities for the unknowns. However, the word *finite* does not mean the same as *small* or *trivial*. In fact, the constants appearing in the bounds that the Gelfond-Baker theory provides for linear forms in logarithms are rather large. Therefore, in practice the upper bounds that can be obtained in this way for the solutions of purely exponential equations can be for instance as large as 10^{40}. This is far too large to admit simple enumeration of all the possibilities, even with the fastest of computers today.

Notwithstanding, it is generally assumed that the upper bounds found in this way are far from the actual largest solution. Therefore, it is worthwile to search for methods to reduce these upper bounds to a size that can be more easily handled. Often it is possible to devise such a method using directly certain properties of the original diophantine equation, for example that large solutions must satisfy certain congruences modulo large numbers, or some reciprocity condition (for some examples, see Grinstead [1978], Brown [1985], Pinch [1987], and Pethö [1983]). The disadvantage of such methods is that they work only for that particular type of diophantine equation, so that

in general for each type of equation a new reduction method must be devised. It would therefore be interesting to have methods for reducing upper bounds for the solutions of inequalities for linear forms in logarithms. They would be useful for solving any type of diophantine problem that leads to such inequalities.

Such methods are provided by that part of the theory of *diophantine approximation* that is concerned with studying how close to zero a linear form can be for given values of the variables. Recently important progress has been made in this field, leading to practically efficient *algorithms*, which can be employed for many diophantine equations to show that in a certain interval $[X_1, X_0]$ no solutions exist. Usually X_1 is of the order of magnitude of $\log X_0$. When for X_0 the theoretical upper bound for the solutions is substituted, a new upper bound X_1 is found. For many equations the upper bound X_0 is well within reach of practical application of these algorithms, within only a few minutes of computer time. This thus leads in practice to methods for finding all the solutions of many types of diophantine equations, for which alternative methods have not yet been found or employed with success.

It is mainly in this third part of the method that new developments can be reported. The arguments we use in the first and second parts are mainly classical. They are applied to types of equations that have been studied before, and also to new types of equations.

The method outlined above, and used in this thesis to solve many examples of various diophantine equations, is of an "algorithmic" nature. In a sense it lies between "ad hoc" methods and "theoretical" methods. This we shall explain below. Let a set of diophantine equations with an unspecified parameter in it be given. As an example of such a set, consider the generalized Ramanujan–Nagell equation $x^2 + D = 2^n$, where D is a parameter, and x, n are the unknowns.

An *ad hoc method* is a method for solving the equation for specific values of the parameters only. However, it may not work at all for other than these particular values. The first example of solving an equation of the type $x^2 + D = 2^n$ occurring in the literature is that by Nagell [1948] of $D = 7$. The method he used is of an ad hoc nature, since it depends heavily on the special choice of 7 for the parameter D.

A *theoretical method* is capable of proving results that hold for some large set of values of the parameters. The Gelfond-Baker theory is of a theoretical nature, since it yields upper bounds for the solutions of many equations in terms of their parameters. Other examples are the theory of quadratic reciprocity, that shows that $x^2 + D = 2^n$ has no solutions at all if D is odd, at least 5, and not congruent to $7 \pmod 8$, and the theory of hypergeometric functions, which Beukers [1981] used to show that the solutions (x,n) of $x^2 + D = 2^n$ satisfy $n < 435 + 10 \cdot {}^2\log|D|$, and if $|D| < 2^{96}$ then moreover $n < 18 + 2 \cdot {}^2\log|D|$. Theoretical methods are often too general to be able to produce all the solutions of a given equation.

An *algorithmic method* is a method that is guaranteed to work for any set of values of the parameters, but has to be applied separately to each particular set of parameter values, in order to produce all the solutions. The methods used in this thesis are mainly of such an algorithmic nature. For the equation $x^2 + D = 2^n$ (and actually for a more general equation) we will give an algorithmic method in Chapter 4. In fact, since Beukers' above-mentioned result provides a small upper bound for the solutions, it can be made algorithmic by providing a simple method of enumerating all the solutions below the upper bound. In order to make the Gelfond-Baker theory algorithmic, enumeration of all possibilities is impractical. Therefore more ingenious ways of determining all the solutions below a large upper bound have to be found. We remark that Beukers' method for the more general equation $x^2 + D = p^n$ also has an ad hoc aspect, since it does not work for any value of p. Our method of Chapter 4 does not have this disadvantage.

An ideal towards which one might strive in solving diophantine equations is to devise a computer algorithm, which only has to be fed with the parameters of the equation, and after a short time gives a list of all the solutions as output. The user should have a guarantee (in the strictest, mathematical sense of proof), that no solutions have been missed.

At first sight the method outlined above, and described in this dissertation, seems to be a good candidate to be developed into such a general applicable algorithm. Namely, the second step is of a quite general nature, providing upper bounds for exponential diophantine equations that are explicit in the parameters of the equation. Also the third step, the algorithmic diophantine approximation part, works in principle for any set of values substituted for the parameters. However, the computations have to be performed separately for

each particular set of values.

The main difficulties in devising such a 'diophantine machine' are in the first part of the method outlined above, especially if some algebraic number theory is used. Developments taking place in the theory of algorithmic algebraic number theory on computing fundamental units and on finding factorizations of prime numbers in algebraic extensions, are of importance here. We believe that when suitable algorithms of this kind are available, it will be possible in principle to make such a 'diophantine machine'. The generality of such an algorithm is restricted by the generality of the first step, the transformation to the linear form in logarithms. In this thesis we use computer algorithms only if the magnitude of the computational tasks makes this necessary, and keep to "manual" work otherwise. In this way we also try to keep the presentation of the methods lucid.

The reader should be aware of the fact that the computer programs and their results are part of the proofs of many of our theorems on specific diophantine equations. It is however impossible to publish all details of these programs and computations. The interested reader may obtain the details from the author by request, and is invited to check the computations himself.

The book by Shorey and Tijdeman [1986] gives a good survey of the diophantine equations for which computable upper bounds for the solutions can be found using the Gelfond-Baker method (see also Shorey, van der Poorten, Tijdeman and Schinzel [1977], and Stroeker and Tijdeman [1982]). Some of these equations can be completely solved by the methods described in this thesis, among which there are purely exponential equations, equations involving binary recurrence sequences, and Thue equations and Thue-Mahler equations. Especially the latter two are of importance in various other parts of number theory. For example, they are the key to solving Mordell equations and various equations arising in algebraic number theory. The Gelfond-Baker method was used to actually solve a diophantine equation for the first time in the work of Baker and Davenport [1969], who solved the system of diophantine equations

$$3 \cdot x^2 - 2 = y^2 \ , \quad 8 \cdot x^2 - 7 = z^2 \ .$$

Other equations occurring in the literature for which upper bounds for the solutions can be computed, cannot be treated as easily by our algorithmic

methods, because the application of the theory of linear forms in logarithms is more complicated for these equations, and moreover the upper bounds are essentially too large. An example of this kind is the Catalan equation $a^x - b^y = 1$ in integers a, b, x, y , all ≥ 2 . Catalan conjectured in 1844 that this equation has only the solution $(a,b,x,y) = (3,2,2,3)$. Tijdeman [1976] proved that the solutions of the Catalan equation are bounded by a computable number. This number can be taken to be $\exp(\exp(\exp(\exp(730))))$, according to Langevin [1976]. However, we fail to see how the methods that we describe in the forthcoming chapters can be applied for completely solving the Catalan equation.

Another diophantine equation, that for centuries has attracted the attention of many mathematicians, is the Fermat equation $x^n + y^n = z^n$ in integers x, y, z, n , with $n \geq 3$ and $x \cdot y \cdot z \neq 0$. It is conjectured to have no solutions. Faltings [1983] proved that for fixed n the number of solutions is finite. His proof is ineffective, and not based on the Gelfond-Baker theory. The Gelfond-Baker theory seems not to be strong enough to dela with the Fermat equation in its full generality, not even if n is fixed. For partial results on the Fermat equation that have been obtained using this theory, see Tijdeman [1985] and Chapter 11 of Shorey and Tijdeman [1986].

We remark that for many diophantine equations recently important progress has been made in determining upper bounds for the number of solutions. See Evertse [1983] and Evertse, Győry, Stewart and Tijdeman [1987] for a survey. These results are often remarkably sharp, but ineffective, so that they cannot be used for actually finding the solutions.

To conclude this section we give an overview of the remaining chapters of this thesis. It is divided into three parts: Chapter 1 is introductory, Chapters 2 and 3 give the necessary preliminaries, and Chapters 4 to 8 deal with various types of diophantine equations.

Sections 1.2 to 1.5 give a short introduction to the Gelfond-Baker theory, diophantine approximation theory, the algorithmic aspects of diophantine approximation, and the procedure for reducing upper bounds, respectively, for the non-specialist. Chapter 2 contains the preliminary results that we need from algebraic number theory and from the theory of p-adic numbers and functions, and quotes in full detail the theorems from the Gelfond-Baker theory which we need. It concludes with some remarks on numerical methods.

Chapter 3 gives in detail the algorithms in the field of diophantine approximation theory that we apply in the subsequent chapters.

The remaining Chapters 4 to 8 are each devoted to a certain type of diophantine equation. Let $p_1, \ldots, p_s$ be a fixed set of distinct primes. Let S be the set of positive integers composed of primes $p_1, \ldots, p_s$ only.

Chapter 4 deals with elements of binary recurrence sequences ("generalized Fibonacci sequences") that are in S, and gives their application to mixed quadratic-exponential equations, such as the generalized Ramanujan-Nagell equation $x^2 + k \in S$ (k fixed). The diophantine approximation part of this chapter is interesting for two reasons: the p-adic approximation is very simple, and in the case of the recurrence having negative discriminant, a nice interplay of p-adic and real/complex approximation arguments occurs. The research for Chapter 4 was done partly in cooperation with A. Pethö from Debrecen. The results have been published in Pethö and de Weger [1986] and de Weger [1986^b].

Chapter 5 deals with the diophantine inequality $0 < x - y < y^{\delta}$, where x, y are in S, and $\delta \in (0,1)$ is fixed. Chapter 6 deals with $x + y = z$, where $x, y, z \in S$, which can be considered as the p-adic analogue of the inequality of Chapter 5. These two equations are the simplest examples of diophantine equations that can be treated by our method. Since they are already purely exponential equations, the first (transformation) step is trivial. So the linear forms in logarithms are directly related to the equations themselves. Therefore they serve as good examples to get a clear understanding of the diophantine approximation part of our method. The results of these chapters have been published in de Weger [1987^a].

Chapter 7 studies the equation $x + y = z^2$, where $x, y \in S$, and $z \in \mathbb{Z}$. This equation is a further generalization of the generalized Ramanujan-Nagell equation, studied in Chapter 4.

In Chapter 8 a procedure is given to solve Thue equations, that works in principle for Thue equations of any degree. It is applied to find all integral points on the elliptic curve $y^2 = x^3 - 4 \cdot x + 1$. We also mention briefly how Thue-Mahler equations can be dealt with. This chapter has been written jointly with N. Tzanakis from Iraklion. The results have been

published in Tzanakis and de Weger [1987], and in de Weger [1987^b].

1.2. The Gelfond-Baker method.

In Section 1.1 we have explained that before applying the Gelfond-Baker method to some diophantine equation, the equation should be transformed into a purely exponential diophantine equation or inequality with not too many terms (cf. (1.1), (1.2)). In this section we sketch the arguments from the Gelfond-Baker theory that lead to upper bounds for the variables of this exponential equation/inequality.

Let us first treat the case of the inequality (1.2). It may be assumed to have the form

$$\left| \alpha_0 \cdot \prod_{i=1}^{s} \alpha_i^{n_i} - 1 \right| < C_0 \cdot \exp(-\delta \cdot N) ,$$

where the α_i are fixed algebraic numbers, $N = \max|n_i|$, and C_0, δ are positive constants. In the examples we study, we encounter one of the following two cases: either all α_i are real, or $|\alpha_i| = 1$ for all i. In the real case, if N is large enough, the linear form in logarithms

$$\Lambda = \log|\alpha_0| + \sum_{i=1}^{s} n_i \cdot \log|\alpha_i|$$

must satisfy

$$|\Lambda| < C_0' \cdot \exp(-\delta \cdot N) \tag{1.3}$$

for some C_0'. In the complex case, the same inequality (1.3) follows for the linear form

$$\begin{aligned} \Lambda &= \mathrm{Log}\ \alpha_0 + \sum_{i=1}^{s} n_i \cdot \mathrm{Log}\ \alpha_i + k \cdot \mathrm{Log}(-1) \\ &= i \cdot \left(\mathrm{Arg}\ \alpha_0 + \sum_{i=1}^{s} n_i \cdot \mathrm{Arg}\ \alpha_i + k \cdot \pi \right) , \end{aligned}$$

where the Log and Arg functions take their principal values. Now we can apply one of the many results from the Gelfond-Baker theory, giving an explicit lower bound for $|\Lambda|$ in terms of N, e.g. the following theorem.

THEOREM 1.1. (Baker [1972]). *Let* Λ *be as above. There exist computable constants* C_1, C_2, *depending on the* α_i *only, such that if* $\Lambda \neq 0$ *then*

$$|\Lambda| > \exp\left(-(C_1+C_2\cdot\log N)\right) .$$

We know that $\Lambda \neq 0$. Combining (1.3) and Theorem 1.1 we obtain

$$N < \frac{C_1 + \log C_0'}{\delta} + \frac{C_2}{\delta} \cdot \log N .$$

It follows that N is bounded from above.

Next, consider the exponential equation (1.1). We can write it as

$$\alpha_0 \cdot \prod_{i=1}^{s} \alpha_i^{n_i} - 1 = \beta_0 \cdot \prod_{j=1}^{r} \beta_j^{m_j} ,$$

where the α_i, β_j are fixed algebraic numbers. Let H_p be the maximum of the $|n_i|$, $|m_j|$ where i, j run through the set of indices for which α_i resp. β_j are non-units. Let H be the maximum of the $|n_i|$, $|m_j|$ where i, j run through the set of all indices. Suppose that p is a rational prime lying above β_j for some j. There are constants c_1, c_2 such that

$$\mathrm{ord}_p\left(\alpha_0 \cdot \prod_{i=1}^{s} \alpha_i^{n_i} - 1\right) \leq c_1 + c_2 \cdot m_j .$$

Assuming that $\mathrm{ord}_p(\alpha_i) = 0$ for all i, we may write down a p-adic linear form in logarithms

$$\Lambda = \log_p \alpha_0 + \sum_{i=1}^{s} n_i \cdot \log_p \alpha_i ,$$

for which, if m_j is large enough, it follows that

$$\mathrm{ord}_p(\Lambda) \leq c_1 + c_2 \cdot m_j . \tag{1.4}$$

We are now in a position to apply the following result from the p-adic Gelfond-Baker theory. Here, $N = \max|n_i|$.

THEOREM 1.2. (van der Poorten [1977], Yu [1987[a]]). *Let* Λ, p *be as above. There exist computable constants* C_3, C_4, *depending only on the* α_i *and on* p, *such that if* $\Lambda \neq 0$ *then*

$$\mathrm{ord}_p(\Lambda) < C_3 + C_4 \cdot \log N .$$

Applying (1.4) and Theorem 1.2 for all possible p we obtain constants C_3', C_4' with

$$H_p < C_3' + C_4' \cdot \log H \ .$$

If $H \leq C_5 \cdot H_p$ for some constant C_5, then this immediately yields an upper bound for H. If $H > C_5 \cdot H_p$, then it can be shown that there exists a conjugate of the α_i, β_j, denoted with a prime sign, for which

$$\left| \beta_0' \cdot \prod_{j=1}^{r} \beta_j'^{m_j} \right| < \exp(-C_6 \cdot H)$$

for a constant C_6 (cf. the proof of Theorem 1.4, pp. 45–49, of Shorey and Tijdeman [1986]). Now we can apply Theorem 1.1. This yields

$$\left| \alpha_0' \cdot \prod_{i=1}^{s} \alpha_i'^{n_i} - 1 \right| > \exp\left(-(C_7 + C_8 \cdot \log H)\right) \ .$$

It follows that H is bounded from above.

If it happens that none of the α_i, β_j are units, then of course the application of Theorem 1.2 suffices.

We remark that, in order to be able to completely solve a diophantine equation, it is crucial that all constants can be computed explicitly. Therefore we can only use the bounds from the Gelfond-Baker theory that are completely explicit. We give details of such theorems in Section 2.4.

1.3. Theoretical diophantine approximation.

In this section we briefly mention some results from diophantine approximation theory, thus giving a background to the next section. We refer to Koksma [1937], Cassels [1957] (Chapters I and III) and to Hardy and Wright [1979] (Chapters XI and XXIII), for further details.

The simplest form of diophantine approximation in the real case is that of approximation of a real number ϑ by rational numbers p/q. It is well known that if ϑ is irrational, then there exist infinitely many solutions $(p,q) \in \mathbb{Z} \times \mathbb{N}$ with $(p,q) = 1$ of the diophantine inequality

$$\left| \vartheta - \frac{p}{q} \right| < q^{-2} \ .$$

All convergents from the continued fraction expansion of ϑ are such solutions. The convergents are simple to compute for any particular $\vartheta \in \mathbb{R}$.

One way of generalizing this is to study simultaneous approximations to a set of real numbers $\vartheta_1, \ldots, \vartheta_n$, i.e. rational approximations to ϑ_i all having the same denominator. It is well known that the system of inequalities

$$| \vartheta_i - \frac{p_i}{q} | < q^{-(1+1/n)} \quad \text{for} \quad i = 1, \ldots, n$$

has infinitely many solutions $(p_1,\ldots,p_n,q)$ if at least one of the ϑ_i is irrational. But it is much harder to find solutions of such inequalities than in the case $n = 1$. Some multi-dimensional continued fraction algorithms have been devised (cf. Brentjes [1981] for a survey), but they seem not to have the desired simplicity and generality. We shall see later how we can apply the so-called L^3-algorithm to this problem.

Another way of generalizing the simplest case of diophantine approximation is to study linear forms, such as

$$L = \sum_{j=1}^{m} q_j \cdot \vartheta_j \ ,$$

where $\vartheta_1, \ldots, \vartheta_m$ are given real numbers, and $q_1, \ldots, q_m$ are the unknowns in $\mathbb{Z}$. Put $Q = \max|q_i|$. A classical theorem guarantees the existence of a solution $(p,q_1,\ldots,q_m)$ of the inequality

$$| L - p | < Q^{-m} \ .$$

Note that the case $m = 1$ becomes our first inequality on dividing by $q = q_1$. Also in this case the L^3-algorithm is very useful, as we shall see below.

We can incorporate the two generalizations above in a further generalization, that of simultaneous approximation of linear forms. Let real numbers ϑ_{ij} be given for $i = 1, \ldots, n$, $j = 1, \ldots, m$. Put

$$L_i = \sum_{j=1}^{m} q_j \cdot \vartheta_{ij} \quad \text{for} \quad i = 1, \ldots, n \ .$$

A celebrated theorem of Minkowski states that there exists a solution $(p_1,\ldots,p_n,q_1,\ldots,q_m)$ of the system of inequalities

$$| L_i - p_i | < Q^{-m/n} \quad \text{for} \quad i = 1, \ldots, n .$$

As we shall show in Section 1.4, the L^3-algorithm may be applied to this general form. We actually compute solutions of systems of inequalities that are slightly weaker in the sense that the right hand side is multiplied by a constant larger than 1.

We now consider inhomogeneous approximation. This means that for all i there is an inhomogeneous term β_i in the linear form L_i, viz.

$$L_i = \beta_i + \sum_{j=1}^{m} q_j \cdot \vartheta_{ij} \quad \text{for} \quad i = 1, \ldots, n .$$

Again, there exists a constant c such that the system

$$| L_i - p_i | < c \cdot Q^{-m/n} \quad \text{for} \quad i = 1, \ldots, n ,$$

under some independence condition on the β_i and ϑ_{ij}, has a solution. This is Kronecker's theorem. The simplest case $m = n = 1$ comes down to

$$| q \cdot \vartheta - p + \beta | < c \cdot q^{-1} .$$

To conclude this section, we remark that there is a p-adic analogue of this theory of diophantine approximation, founded by Mahler and Lutz. If we replace in the above considerations $\mathbb{R}$ by $\mathbb{Q}_p$, the absolute value $|\cdot|$ by the p-adic value $|\cdot|_p$, and the measure Q for an approximation $(p_1,\ldots,p_n,q_1,\ldots,q_m)$ by any convex norm $\Phi(p_1,\ldots,p_n,q_1,\ldots,q_m)$, then the p-adic analogues of the theorems of Minkowski and Kronecker are essentially analogous to the above mentioned results in the real case. See Koksma [1937] for references to Mahler's work, and Lutz [1951].

1.4. Computational diophantine approximation.

In this section we give some idea of practically solving the diophantine approximation problems that we encounter in solving diophantine equations. In this section we give no rigorous treatment. We neglect worst cases, and concentrate on how things are expected to work, and appear to work in practice. For a more rigorous theoretical treatment we refer to a forthcoming publication by Tijdeman, Wang and the present author. In the subsequent

chapters of this thesis many examples are given, showing that our methods are indeed useful in practice. Applying the method in practice may be the best way of acquiring the necessary *Fingerspitzengefühl* for the method.

We shall deal with the following computational diophantine approximation problem. Let ϑ_{ij}, $\beta_i \in \mathbb{R}$ be given, and let $p_1, \ldots, p_n, q_1, \ldots, q_m$ be integral unknowns with $Q = \max|q_j|$. Let L_i be as above. Let a positive constant Q_0, assumed to be a rather large number, 10^{50} say, be given. Find a lower bound for the value of

$$\max_i | L_i - p_i | ,$$

where $(p_1,\ldots,p_n,q_1,\ldots,q_m)$ runs through the set of values with $Q \leq Q_0$. From the theory outlined in Section 1.3 it follows that one will be satisfied if this lower bound is of the size $Q_0^{-m/n}$. For the p-adic case an analogous problem may be formulated.

Related problems in diophantine approximation theory are those of actually finding a good or the best solution of $\max|L_i-p_i| < \epsilon$ for a fixed $\epsilon > 0$. As we shall see, the L^3-algorithm is a very useful tool for finding good solutions. The problem of finding the best solution however seems to be essentially more difficult. We note that in most of our applications of solving diophantine equations it suffices to have a suitable lower bound for $\max|L_i-p_i|$.

The computational tool that we use to solve the afore-mentioned problems is the so-called L^3-lattice basis reduction algorithm, described in Lenstra, Lenstra and Lovász [1982]. We shall give details of this algorithm in Chapter 3. Below we briefly indicate how it can be used to solve diophantine approximation problems.

Let Γ be a lattice in $\mathbb{R}^n$. The L^3-algorithm accepts as input an arbitrary basis $\underline{b}_1, \ldots, \underline{b}_n$ of Γ. As output it gives another basis $\underline{c}_1, \ldots, \underline{c}_n$ of the same lattice Γ, that is a so-called *reduced* basis. The concept *reduced* means something like nearly orthogonal. From a reduced basis it is possible to compute lower bounds for the following two quantities: (i), the length of the non-zero lattice point that is nearest to the origin, viz.

$$\ell(\Gamma) = \min_{\underline{0} \neq \underline{x} \in \Gamma} |\underline{x}| ,$$

(cf. Lenstra, Lenstra and Lovász [1982], Proposition (1.11) and our Lemma 3.4), (ii), for any given point $\underline{y} \in \mathbb{R}^n$, the distance from $\underline{y}$ to the nearest lattice point, viz.

$$\ell(\Gamma,\underline{y}) = \min_{\underline{x}\in\Gamma} |\underline{x}-\underline{y}| ,$$

(cf. our Lemmas 3.5 and 3.6). This algorithm enjoys the property that these lower bounds are usually near to the actual minimal solutions. In a typical situation, where the lattice is not too distorted, the vectors $\underline{c}_i$ of the reduced basis all have about the same length, which is of the order of magnitude of

$$\det(\Gamma)^{1/n} ,$$

The value of $\ell(\Gamma)$ as well as the lower bounds computed for it, are about as large as that. If $\underline{y}$ is not too close to a lattice point, the same holds for $\ell(\Gamma,\underline{y})$. Moreover, the running time of the algorithm is good, both in the theoretical sense (it is polynomial-time in the length of the input-parameters), and in the practical sense.

To solve the problem of finding a lower bounds for $\max|L_i-p_i|$ as formulated above, we take the lattice Γ as follows. Let C be an integer, at least as large as $Q_0^{1+m/n}$. The lattice Γ, of dimension $n + m$, is defined by specifying a basis, namely the column vectors $\underline{b}_1, \ldots, \underline{b}_{n+m}$ of the matrix

$$\mathcal{B} = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ \text{Ø} & & 1 & \\ [C\cdot\vartheta_{11}] & \ldots & [C\cdot\vartheta_{1m}] & -C \\ \vdots & & \vdots & \\ [C\cdot\vartheta_{n1}] & \ldots & [C\cdot\vartheta_{nm}] & \end{pmatrix} \begin{matrix} \\ \\ \text{Ø} \\ \\ \\ \\ -C \end{matrix} .$$

(The symbol Ø means that all not explicitly given entries are zero). Applying the L^3-algorithm to this lattice we find a reduced basis, of which the basis vectors will have lengths of about $C^{n/(m+n)}$, which is roughly the size of Q_0. Generally speaking, the larger C is, the larger the lengths of the basis vectors of a reduced basis will be (and the larger the lower bounds for $\ell(\Gamma)$ and $\ell(\Gamma,\underline{y})$ will be).

Let us first treat the homogeneous case, i.e. $\beta_i = 0$ for all i. Consider the lattice point $\underline{x} = \mathcal{B}\cdot(q_1,\ldots,q_m,p_1,\ldots p_n)^T$. It is equal to

$$\underline{x} = (q_1, \ldots, q_m, \tilde{L}_1 - C \cdot p_1, \ldots, \tilde{L}_n - C \cdot p_n)^T ,$$

where

$$\tilde{L}_i = \sum_{j=1}^{m} q_j \cdot [C \cdot \vartheta_{ij}] \quad \text{for} \quad i = 1, \ldots, n .$$

From the application of the L^3-algorithm we find a lower bound for $\ell(\Gamma)$, of size Q_0. We assume it to be large enough (if this is not the case, we try a somewhat larger value for C, and perform the L^3-algorithm again for the lattice defined for this C). So we may assume that there is a small constant c_1 such that

$$\sum_{i=1}^{n} (\tilde{L}_i - C \cdot p_i)^2 \geq \ell(\Gamma)^2 - m \cdot Q_0^2 > c_1 \cdot Q_0^2 .$$

We have $|\tilde{L}_i - C \cdot L_i| \leq m \cdot Q_0$, so we may assume that for small constants c_2, c_3

$$\max |L_i - p_i| > c_2 \cdot C^{-1} \cdot \max |\tilde{L}_i - C \cdot p_i| > c_3 \cdot Q_0 / C .$$

By the choice of C this last bound has the required size.

Next, we study the inhomogeneous case, where not all β_i are zero. We take the same lattice Γ as in the homogeneous case (note that the lattice definition depends only on the ϑ_{ij} and the C). Consider the point

$$\underline{y} = (0, \ldots, 0, -[C \cdot \beta_1], \ldots, -[C \cdot \beta_n])^T .$$

From the reduced basis found by the L^3-algorithm we have a lower bound for $\ell(\Gamma, \underline{y})$. Assume that it is large enough, and of size Q_0. We take the same lattice point $\underline{x} = \mathcal{B} \cdot (q_1, \ldots, q_m, p_1, \ldots p_n)^T$ as in the homogeneous case. Then

$$\underline{x} - \underline{y} = (q_1, \ldots, q_m, \tilde{L}_1 - C \cdot p_1, \ldots, \tilde{L}_n - C \cdot p_n)^T ,$$

where

$$\tilde{L}_i = [C \cdot \beta_i] + \sum_{j=1}^{m} q_j \cdot [C \cdot \vartheta_{ij}] \quad \text{for} \quad i = 1, \ldots, n .$$

The same reasoning as in the homogeneous case now yields the desired result. Note that if we have performed the L^3-algorithm once for given ϑ_{ij}, we may use the result to treat the homogeneous case, and many inhomogeneous cases with different β_i's as well, as long as the ϑ_{ij}'s are the same.

The above process describes how to find lower bounds for systems of diophantine inequalities. It will be clear from the above that it is not difficult to find good solutions, i.e. $(q_1,\ldots,q_m, p_1,\ldots,p_n)$ with $Q \le Q_0$ and $\max|L_i-p_i|$ near to the best possible value. In particular, the basis vectors of a reduced basis are adequate for the homogeneous case, and for the inhomogeneous case the lattice points near to $\underline{y}$ will be such solutions. The lattice points near to $\underline{y}$ are not difficult to find once a reduced basis is available. Specifically, if $s_1, \ldots, s_n \in \mathbb{R}$ are the coordinates of $\underline{y}$ with respect to a reduced basis, then one may take the lattice points with coordinates $t_i \in \mathbb{Z}$ that are near to s_i ($i = 1, \ldots, n$).

In the definition of the matrix above the expressions $[C\cdot\vartheta_{ij}]$ occur. Using these expressions we have constructed a lattice Γ that is completely integral, i.e. $\Gamma \subset \mathbb{Z}^{m+n}$. The L^3-algorithm can be adapted to work exact for those lattices, so that rounding-off errors are avoided (cf. Section 3.5). The "errors" occur in the difference between the $\tilde{L}_i$ and the $C\cdot L_i$, and are thus kept under control by choosing the proper constants c_1, c_2, c_3 . Of course one should take care to have the numerical values of the ϑ_{ij} and the β_i correct to a sufficient precision. We shall discuss such numerical problems briefly in Section 2.5.

A possible variation of the above diophantine approximation problem is to give weights to the linear forms L_i , i.e. to look for a lower bound for

$$\max_i w_i\cdot| L_i - p_i | ,$$

where the w_i are fixed positive numbers. This situation can be dealt with easily by replacing the C 's in the $(n+i)$ th row of the matrix by proper constants depending on w_i .

Another variation is the problem where not all the variables q_j have the same upper bound Q_0 . To illustrate this, assume that $n = 1$, and that

$$L = \sum_{j=1}^{m} q_j\cdot\vartheta_j .$$

Now suppose that for some $Q_1 > Q_2$ (it will be handy to have $Q_2 \mid Q_1$) we are interested in the solutions with

$$|q_j| \le Q_1 \quad \text{for} \quad j = 1, \ldots, m_1 ,$$

$$|q_j| \le Q_2 \quad \text{for} \quad j = m_1+1, \ldots, m ,$$

$$|p| \le Q_2 .$$

Next let C be of the size of $Q_1^{m_1} \cdot Q_2^{m-m_1+1}$, and take the matrix

$$\begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & \varnothing & \\ & \varnothing & & Q_1/Q_2 & & & \\ & & & & \ddots & & \\ & & & & & Q_1/Q_2 & \\ [C\cdot\vartheta_1] & \ldots & [C\cdot\vartheta_{m_1}] & [C\cdot\vartheta_{m_1+1}] & \ldots & [C\cdot\vartheta_m] & -C\cdot Q_1/Q_2 \end{pmatrix} .$$

For a lattice point $(q_1,\ldots,q_m,\tilde{L}-C\cdot p)^T$ we expect that $|\tilde{L}-C\cdot p| > c\cdot Q_1$ for some c. It follows that $|L-p| > c'\cdot Q_1^{-(m_1-1)}\cdot Q_2^{-(m-m_1+1)}/C > c''\cdot Q_1^{-(m_1-1)}\cdot Q_2^{-(m-m_1+1)}$ for some c', c''. This variant is useful when a combination of real and p-adic techniques is used, such as for the Thue-Mahler equation (see Section 8.6).

We conclude this section by giving the analogous method of p-adic diophantine approximation. We assume that the ϑ_{ij}, β_i are in $\mathbb{Q}_p$, and, moreover, that they are p-adic integers. Let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For any p-adic integer γ and any $\mu \in \mathbb{N}_0$ we denote by $\gamma^{(\mu)}$ the unique rational integer such that

$$\gamma \equiv \gamma^{(\mu)} \pmod{p^\mu} , \quad 0 \le \gamma^{(\mu)} < p^\mu .$$

Let $\mu \in \mathbb{N}$ be such that p^μ is roughly the same size as $Q_0^{1+m/n}$, and assume that μ is large enough (it is the analogue of the constant C in the real case above). Take for Γ the lattice of which a basis is given by the column vectors of the matrix

$$\mathcal{B} = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & \varnothing & \\ \varnothing & & 1 & & & \\ \vartheta_{11}^{(\mu)} & \ldots & \vartheta_{1m}^{(\mu)} & p^\mu & & \\ \vdots & & \vdots & & \ddots & \\ \vartheta_{n1}^{(\mu)} & \ldots & \vartheta_{nm}^{(\mu)} & \varnothing & & p^\mu \end{pmatrix} .$$

Consider the lattice point

$$\mathcal{B}\cdot(q_1,\ldots,q_m,z_1,\ldots,z_n)^T = (q_1,\ldots,q_m,p_1,\ldots,p_n)^T \ .$$

Then it is obvious that

$$p_i = \sum_{j=1}^{m} q_j\cdot\vartheta_{ij}^{(\mu)} + z_i\cdot p^{\mu} \ .$$

Hence the lattice Γ can be described as the set

$$\Gamma = \{\ (q_1,\ldots,q_m,p_1,\ldots,p_n)^T \in \mathbb{Z}^{m+n} \ \mid \ \sum_{j=1}^{m} q_j\cdot\vartheta_{ij} \equiv p_i \pmod{p^{\mu}} \quad \text{for} \quad i = 1, \ldots, n \ \} \ .$$

The L^3-algorithm provides a lower bound for the length of the nonzero vectors in this set, which is of the same size as $p^{\mu\cdot n/(n+m)}$, and that of Q_0. This yields the desired result, if μ is taken large enough.

For the inhomogeneous case, put

$$\underline{y} = (0,\ldots,0,-\beta_1^{(\mu)},\ldots,-\beta_n^{(\mu)})^T \ ,$$

and consider the set

$$\Gamma^* = \{\ (q_1,\ldots,q_m,p_1,\ldots,p_n)^T \in \mathbb{Z}^{m+n} \ \mid \ \beta_i + \sum_{j=1}^{m} q_j\cdot\vartheta_{ij} \equiv p_i \pmod{p^{\mu}} \quad \text{for} \quad i = 1, \ldots, n \ \} \ .$$

Then $\underline{x} \in \Gamma^*$ if and only if $\underline{x} + \underline{y} \in \Gamma$, so Γ^* is a translated lattice. A lower bound for $\ell(\Gamma,\underline{y})$ now yields the desired result.

Again variations are possible, as in the real case, e.g. by replacing on the (n+i) th row the μ by different μ_i. It is even possible in this way to treat more than one prime p at the same time.

We conclude this section with three remarks. Firstly, in the case that the dimension of the lattice under consideration is only 2, the L^3-algorithm is essentially the continued fraction algorithm, and so yields nothing new. For the p-adic continued fraction algorithm, see de Weger [1986[a]]. Secondly, the inhomogeneous case of diophantine approximation of one linear form of real numbers can also be treated by what is known as Davenport's lemma, cf. Baker and Davenport [1969] (and its multi-dimensional generalization, cf. Ellison

[1971[a]]). We will return to this in Chapter 3, and explain there why we prefer our method.

Finally, one of the nice features of the above method of practical diophantine approximation is that if an extreme solution exists, then in the homogeneous case the lattice (with proper constant C or μ) will be distorted. This means that the reduced basis will not be as nice as expected, for example there might be a basis vector in it that is substantially shorter than the other ones. In the inhomogeneous case the existence of an extreme solution means that there is a lattice point extremely near to $\underline{y}$. The algorithm detects such an extraordinary situation at once, and in most cases the extremal solution is presented explicitly (e.g. in the homogeneous case as one of the vectors of the reduced basis). One can check whether this extremal solution actually satisfies the original equation, and then proceed by replacing in the above reasoning $\ell(\Gamma)$ or $\ell(\Gamma,\underline{y})$ by lower bounds for all vectors in the lattice except the extremal one. These new lower bounds will in general be of the expected size. However, when we solved diophantine equations in practice, we have never met such an extraordinary situation.

1.5. The procedure for reducing upper bounds.

We have seen in Section 1.2 how upper bounds for the solutions of the exponential inequalities and equations occurring there can be found. In Section 1.4 we have studied some diophantine approximation theory from a practical point of view. Now these two things come together.

From the application of the Gelfond-Baker theory we are left with the following problem. We have a linear form

$$\Lambda = \beta + \sum_{j=1}^{m} n_j \cdot \vartheta_j \ ,$$

where the β and ϑ_j are constants (that they are logarithms of algebraic numbers is now of no importance anymore), and the n_j are integral unknowns. We know that Λ is extremely close to 0, namely

$$|\Lambda| < c \cdot \exp(-\delta \cdot N) \ ,$$

where c, δ are (small) constants, and $N = \max|n_j|$. Finally, we have an explicit upper bound N_0 for N. This N_0 is very large, 10^{50} say.

It will be clear from Section 1.4 that the methods outlined there are of use for solving this problem. For Q_0 we take N_0. We have $n = 1$. In the real case we expect, by choosing C at least of size N_0^{m+1}, that

$$|\Lambda| > c' \cdot N_0^{-m} ,$$

for a small constant c'. It follows by combining the two inequalities for $|\Lambda|$ that

$$N < \log(c/c')/\delta + (m/\delta) \cdot \log N_0 .$$

So the upper bound N_0 for N is reduced to an upper bound N_1 of the size of $\log N_0$, which is a considerable improvement indeed. We now may apply the procedure with N_1 instead of N_0, and repeat until no further improvement is obtained. In practice it appears almost always to be the case that in that situation the reduced upper bound is near to the actual largest solution, anyway so small that simple methods of finding all the solutions below that bound suffice.

In the p-adic case an analogous reduction of upper bounds can be reached, following a similar argument. We have for the linear form Λ (cf. (1.4)),

$$\mathrm{ord}_p(\Lambda) \leq c_1 + c_2 \cdot m_j ,$$

where c_1, c_2 are small constants, and m_j is one of the variables. Moreover, the variables are bounded by a large constant N_0, that is explicitly known. We take μ such that p^μ is at least of size N_0^{m+1}, so that the lower bound for the shortest nonzero vector in Γ (or Γ^*) is larger than $\sqrt{m} \cdot N_0$. Then it follows that the elements of the lattice Γ (or of the translated lattice Γ^*) cannot be solutions of (1.2). Therefore,

$$c_1 + c_2 \cdot m_j < \mu ,$$

so that we find a new upper bound for m_j, that is of the size of μ, which is about $\log N_0 / \log p$. We repeat this procedure for all the m_j, in order to obtain a reduced upper bound for H_p. If this is not yet sufficient to derive at once a reduced upper bound for H, then we can do so by applying a reduction step for real linear forms, where we may take advantage of the fact that for some of the variables a much better upper bound has just been found (cf. the second variation in Section 1.4). Again we repeat the whole procedure as far as possible.