

Lecture Notes on  
Decision Problems in Algebra

Joseph Flenner  
University of Notre Dame

September 1, 2011



---

# Chapter 1

## Introduction

### 1.1 The decision problem

We begin with some general remarks on what is meant by decision problems. Let  $\mathcal{L}$  be a first-order language, and  $S$  a set of  $\mathcal{L}$ -sentences. By the *decision problem* for  $S$  we mean the question, given a sentence  $\varphi$  of  $\mathcal{L}$ , whether  $\varphi \in S$ ? The main issue is of course whether there is an algorithm which, given input  $\varphi$ , answers this question (correctly!) for  $\varphi$ .<sup>1</sup>

This is most commonly seen in one of the following specific forms:

- (a) If  $T$  is an  $\mathcal{L}$ -theory, say  $T$  is *decidable* if there is an algorithm deciding membership in  $T$ .
- (b) If  $\mathfrak{A}$  is an  $\mathcal{L}$ -structure, and  $T$  is the  $\mathcal{L}(\mathfrak{A})$ -theory of  $\mathfrak{A}$  (i.e. the theory of  $\mathfrak{A}$  in the language  $\mathcal{L} \cup \{c_a \mid a \in \mathfrak{A}\}$  expanding  $\mathcal{L}$  by a new constant symbol for each element of  $\mathfrak{A}$ ), say  $\mathfrak{A}$  is *decidable* if there is an algorithm deciding membership in  $T$ .

In the second case, it would generally be assumed that  $\mathfrak{A}$  is a *computable structure*, in the sense that there is some encoding of  $\mathfrak{A}$  whereby the relations and functions of  $\mathcal{L}$ , as interpreted in  $\mathfrak{A}$ , are computable. But in the first case, this is not necessary, so for example one speaks of the decidability of the theories of  $\mathbb{R}$  and  $\mathbb{C}$ . Both questions, however, are largely only meaningful if  $\mathcal{L}$  is countable.

In fact, in these notes we will most frequently be working in the language of rings  $\mathcal{L}_R = \{+, \times, 0, 1\}$ , so if not stated otherwise this should be assumed to be the language. The constants 0 and 1 are included for convenience, though since both

---

<sup>1</sup>Though I am not convinced that this is really a very useful distinction to make, we are distinguishing between the concepts of decidability (for sets of sentences) and computability (for sets of natural numbers). This is not a universal convention.

I do think, however, that it is worthwhile to clearly distinguish between two common informal uses of the term *undecidable problem*: in the sense that we are using it, but also in the sense e.g. of the Continuum Hypothesis relative to ZFC. Though thematically linked, these are categorically different kinds of questions, the former concerned with the overall question of membership in a set, the latter with the status of a specific sentence.

elements are definable in any ring (or semiring, in the case of  $\mathbb{N}$ ), this can often be done without if there's any reason to be stingy. However, defining them costs a couple quantifiers (“ $\exists x (\forall y (x + y = x) \wedge \varphi(x, \dots))$ ” vs. “ $\varphi(0, \dots)$ ”), which is one pretty good reason to keep them around as constants.

Okay. So, what are the main strategies for showing that a theory or a structure is decidable? First of all, if a theory  $T$  is complete and has a computable axiomatization (namely, a decidable subset  $S \subseteq T$  such that  $T$  is the deductive closure of  $S$ ), then  $T$  is decidable. Indeed, if  $\varphi$  is an  $\mathcal{L}$ -sentence, then by completeness of  $T$  either  $\varphi \in T$  or  $\neg\varphi \in T$ . Since  $T$  is the deductive closure of  $S$ , there is a proof of either  $S \vdash \varphi$  or  $S \vdash \neg\varphi$ . Now, by enumerating all proofs from  $S$  until a proof of either  $\varphi$  or  $\neg\varphi$  is found, we obtain a decision procedure for membership in  $T$ . Recall moreover that by the Completeness Theorem of Gödel,  $T$  is identical with the set of sentences true in all models of  $S$ .

Let us also note that the above criterion is also necessary. If  $T$  is decidable, then  $T$  is a computable axiomatization of itself.

The question for incomplete (but deductively closed) theories is subtler. Here, given again a computable axiomatization  $S$  of  $T$ , we may attempt a search for a proof  $S \vdash \varphi$ . If  $\varphi \in T$ , then eventually the search will be fruitful and the algorithm confirms  $\varphi \in T$ . But if neither  $\varphi$  nor  $\neg\varphi$  is in  $T$ , we may be stuck eternally waiting for a confirmation that will never come. This, in practice, would suck. In such a case we speak of the theory  $T$  being *computably enumerable*.

Another drawback to this approach is that the algorithm, even in case  $T$  is complete, is not a great one. Performing an all out search for a proof of  $\varphi$  or  $\neg\varphi$  tends to be far from tractable. In cases where we are only concerned with the existence or nonexistence of an algorithm, this is no concern. But often we learn more about the theories or structures in question if we can find a more insightful algorithm.

In fact, most of the classical decidability results rely on a computable version of quantifier elimination:

**Definition 1.1.1.** *A theory  $T$  computably eliminates quantifiers if there is an algorithm which, given a formula  $\varphi(\bar{x})$ , produces a quantifier free formula  $\psi(\bar{x})$  which is equivalent in  $T$  to  $\varphi$ , i.e.  $T \models \forall \bar{x} (\varphi(\bar{x}) \leftrightarrow \psi(\bar{x}))$ .*

**Proposition 1.1.2.** *If  $\mathfrak{A}$  is an  $\mathcal{L}$ -structure such that the substructure  $\mathfrak{C}$  of  $\mathfrak{A}$  generated by the constants of  $\mathcal{L}$  is computable, and  $\text{Th}(\mathfrak{A})$  computably eliminates quantifiers, then  $\text{Th}(\mathfrak{A})$  is decidable.*

*Proof.* Let  $\varphi$  be a sentence. Constructing a quantifier free sentence  $\psi$  equivalent to  $\varphi$  in  $\text{Th}(\mathfrak{A})$ , since  $\varphi \in \text{Th}(\mathfrak{A})$  iff  $\psi \in \text{Th}(\mathfrak{A})$  it suffices to check the latter condition. But the truth of  $\psi$  can be evaluated by computability of  $\mathfrak{C}$ .  $\square$

This method can be weakened somewhat. Recall that a theory  $T$  is called *model complete* if, for every pair of models  $\mathfrak{A}, \mathfrak{B} \models T$  with  $\mathfrak{A} \subseteq \mathfrak{B}$ ,  $\mathfrak{A} \preceq \mathfrak{B}$ . A computable version of model completeness can be devised via the following fact.

**Proposition 1.1.3.** *An  $\mathcal{L}$ -theory  $T$  is model complete if and only if for every  $\mathcal{L}$ -formula  $\varphi(\bar{x})$ , there is a quantifier free formula  $\psi(\bar{x}, \bar{y})$  such that*

$$T \models \forall \bar{x} (\varphi(\bar{x}) \leftrightarrow \exists \bar{y} \psi(\bar{x}, \bar{y})). \quad (1.1)$$

So, let us call  $T$  *computably model complete* if there is an algorithm which for each  $\varphi(\bar{x})$  produces a  $\psi(\bar{x}, \bar{y})$  as in (1.1).

**Corollary 1.1.4.** *Suppose  $\mathfrak{A}$  is a (countable) computable structure such that  $\text{Th}(\mathfrak{A})$  is computably model complete. Then  $\mathfrak{A}$  is decidable.*

*Proof.* We seek an algorithm to decide whether an  $\mathcal{L}(\mathfrak{A})$ -formula  $\varphi(\bar{a})$  is true in  $\mathfrak{A}$ . Suppose  $\exists \bar{y} \psi_0(\bar{x}, \bar{y})$  and  $\exists \bar{y} \psi_1(\bar{x}, \bar{y})$  are logically equivalent in  $T$  to  $\varphi(\bar{x})$  and  $\neg\varphi(\bar{x})$ , respectively. We may assume that the lengths  $\ell(\bar{y})$  of the tuples  $\bar{y}$  are in both cases the same, by appending dummy variables to the shorter one if needed.

So  $\mathfrak{A} \models \varphi(\bar{a})$  iff  $\mathfrak{A} \models \exists \bar{y} \psi_0(\bar{a}, \bar{y})$ , and likewise  $\neg\varphi(\bar{a})$  and  $\exists \bar{y} \psi_1(\bar{a}, \bar{y})$ . Now, enumerate all  $\ell(\bar{y})$ -tuples  $\bar{b}$  of  $\mathfrak{A}$  and check successively whether  $\mathfrak{A} \models \psi_0(\bar{a}, \bar{b})$  or  $\mathfrak{A} \models \psi_1(\bar{a}, \bar{b})$ . Since  $\mathfrak{A}$  must satisfy either  $\exists \bar{y} \psi_0(\bar{a}, \bar{y})$  or  $\exists \bar{y} \psi_1(\bar{a}, \bar{y})$  (and not both), eventually this process will halt with the conclusion  $\mathfrak{A} \models \varphi(\bar{a})$  or  $\mathfrak{A} \models \neg\varphi(\bar{a})$ , respectively.  $\square$

Some decidable theories:

1. The theory ACF of algebraically closed fields (Tarski, ??).
2. The theory RCF of real closed fields (Tarski [27], 1930).
3. The theory of abelian groups (Szmielew [26], 1955).
4. The theory of the  $p$ -adic numbers  $\mathbb{Q}_p$  (Ax & Kochen [1], 1966; Cohen [3], 1969).

Most of you have already seen the first two of these. We'll get to the 4th and, time permitting, the 3rd later on. For now, we instead turn to the question of undecidable theories and structures. Here the situation gets a bit murky. While it is clear enough that one proves decidability by explicitly exhibiting a decision procedure, proving the nonexistence of such a procedure is rather problematic.

It seems to basically boil down to two options: either be incredibly industrious and inventive and prove this from scratch, or be a regular person and find a way to reduce the problem in question to another one already known to be undecidable. These are not meant to be mutually exclusive. The reduction of one decision problem to another often also requires a great deal of ingenuity. But undecidability results proved from scratch are rare and remarkable specimens indeed.

Let us point out these two particular examples of problems proved directly to be undecidable<sup>2</sup>:

1. Number theory, the complete theory of the natural numbers  $\mathbb{N}$ . This is Gödel's Incompleteness Theorem (1931).
2. Hilbert's Tenth Problem, the set of sentences

$$\exists x_1, \dots, x_n P(x_1, \dots, x_n) = 0$$

<sup>2</sup>Admittedly, these proofs still do operate by interpretation of known unsolvable problems or paradoxes such as the Halting Problem or the Liar's Paradox.

true in the integers, where  $P$  is a polynomial over  $\mathbb{Z}$ . This was proved over the course of a couple decades, beginning around 1950 with key contributions from Martin Davis, Hilary Putnam, and Julia Robinson, and finishing in 1970 with the final ingredient by Yuri Matiyasevich.

Both of these have spawned large families of undecidable children. But how does this reduction process work? A first idea:

**Proposition 1.1.5.** *Suppose  $\mathfrak{A}$  and  $\mathfrak{B}$  are  $\mathcal{L}$ -structures with  $\mathfrak{B} \subseteq \mathfrak{A}$  such that  $\mathfrak{B}$  as a set is definable (without parameters) in  $\mathfrak{A}$ .*

- (i) *If  $\text{Th}(\mathfrak{B})$  is undecidable, then so is  $\text{Th}(\mathfrak{A})$ .*
- (ii) *If  $\mathfrak{B}$  is undecidable, then so is  $\mathfrak{A}$ .*

*Proof.* We give only a sketch of (i), as the details would be a bit tedious. Let  $\varphi(x)$  define  $\mathfrak{B}$  in  $\mathfrak{A}$ . Supposing  $\text{Th}(\mathfrak{A})$  were decidable, obtain a decision procedure for  $\text{Th}(\mathfrak{B})$  as follows. For an  $\mathcal{L}$ -sentence  $\psi$ , construct a new sentence  $\psi^*$  by relativizing each subformula  $\exists x \sigma(x, \bar{y})$  in  $\psi$  to  $\exists x (\sigma(x, \bar{y}) \wedge \varphi(x))$ . One then verifies that  $\psi^* \in \text{Th}(\mathfrak{A})$  iff  $\psi \in \text{Th}(\mathfrak{B})$ .  $\square$

This too is somewhat stronger than what is really needed. Ultimately we will not only want to work with undecidable substructures of a larger structure, but structures, possibly in a different language, that can in a sense be simulated within an undecidable structure. The tool we are after here is an *interpretation*.

**Definition 1.1.6.** *Suppose that  $\mathfrak{A}$  is an  $\mathcal{L}_1$ -structure,  $C \subseteq \mathfrak{A}$ , and  $\mathfrak{B}$  an  $\mathcal{L}_2$ -structure. Then  $\mathfrak{A}$  interprets  $\mathfrak{B}$  over  $C$  if there are a  $C$ -definable set  $\mathcal{B} \subseteq \mathfrak{A}^k$  and a  $C$ -definable equivalence relation  $\varepsilon$  on  $\mathcal{B}$  such that:*

- (a) *The  $\varepsilon$ -equivalence classes of  $\mathcal{B}$  are in bijection  $i : \mathcal{B}/\varepsilon \xrightarrow{\sim} \mathfrak{B}$  such that*
- (b) *The relations and functions induced on  $\mathcal{B}$  by the ( $\mathcal{L}_2$ ) relations and functions of  $\mathfrak{B}$  via  $i$  are all  $C$ -definable.*

So, the idea is that the  $C$ -definable quotient  $\mathcal{B}/\varepsilon$  in  $\mathfrak{A}$  can  $C$ -definably mimic  $\mathfrak{B}$ . Here's an example.

**Proposition 1.1.7.** *The rational numbers  $\mathbb{Q}$  is interpretable in the integers  $\mathbb{Z}$ .*

*Proof.* Define  $\mathcal{Q} := \{\langle x, y \rangle \in \mathbb{Z}^2 \mid y \neq 0\}$  and the equivalence relation

$$\langle x_1, y_1 \rangle \varepsilon \langle x_2, y_2 \rangle \Leftrightarrow x_1 y_2 = x_2 y_1.$$

Clearly  $\mathbb{Q}$  is in bijection with  $\mathcal{Q}/\varepsilon$  by  $i : \langle x, y \rangle \mapsto x/y$ .

It remains only to show that the graphs of  $+$  and  $\times$  induced on  $\mathcal{Q}/\varepsilon$  are definable. But it is immediate that the former is defined by

$$\langle x_1, y_1 \rangle + \langle x_2, y_2 \rangle = \langle x_3, y_3 \rangle \Leftrightarrow \langle x_1 y_2 + x_2 y_1, y_1 y_2 \rangle \varepsilon \langle x_3, y_3 \rangle$$

and the latter by

$$\langle x_1, y_1 \rangle \times \langle x_2, y_2 \rangle = \langle x_3, y_3 \rangle \Leftrightarrow \langle x_1 x_2, y_1 y_2 \rangle \varepsilon \langle x_3, y_3 \rangle.$$

$\square$

Returning to the decidability question, the analog to Proposition 1.1.5 states:

**Proposition 1.1.8.** *Suppose that  $\mathfrak{A}$  and  $\mathfrak{B}$  are as in Definition 1.1.6. Then*

(i) *If  $\text{Th}(\mathfrak{B})$  is undecidable, then so is the  $\mathcal{L}_1(C)$ -theory of  $\mathfrak{A}$ .*

(ii) *If  $\mathfrak{B}$  is undecidable, then so is  $\mathfrak{A}$ .*

*Proof.* Again, just a sketch. The idea is to proceed as in 1.1.5, but in constructing  $\psi^*$  we must not only relativize a subformula  $\exists x \sigma(x, \bar{y})$  to  $S$ , but also replace each instance of  $=$  in  $\sigma$  with  $\varepsilon$  and each instance of a relation or function symbol with its corresponding  $\mathcal{L}_1(C)$ -definition.  $\square$

In the next section we give two examples of Propositions 1.1.5 and 1.1.8 in action.

## 1.2 Two examples

Let us state once again for the record Gödel's theorem on undecidability of arithmetic (which is not the Incompleteness Theorem exactly but is an immediate consequence of it).

**Theorem 1.2.1** (Gödel [10], 1931). *The complete theory of  $\mathbb{N}$  in the language of rings is undecidable.*

What follows are two relatively simple examples of using 1.1.5 and 1.1.8 to transfer undecidability from  $\mathbb{N}$ .

### 1.2.1 $\mathbb{N}$ in the language of successor and divisibility

This example is from Robinson [24]. Robinson proved that multiplication and addition are definable on the positive integers using only the successor function and the divisibility relation. Her proof appears to use the famous theorem of Dirichlet on primes in arithmetic progressions, though this is not mentioned so perhaps she had something more elementary in mind. The proof of Dirichlet's theorem can be found in most introductory number theory texts, for instance [22].

**Theorem 1.2.2** (Dirichlet, 1837). *If  $a, m \in \mathbb{N}$  are relatively prime, then the arithmetic sequence  $a, a + m, a + 2m, \dots$  contains infinitely many primes  $p$ , i.e.  $p \equiv a \pmod{m}$ .*

**Corollary 1.2.3.** *Given  $a, b, n \in \mathbb{N}$ , there are distinct primes  $x, y, z$ , all relatively prime to  $a, b$ , and  $n$  and with  $z > n$ , such that  $z$  divides both  $ax + 1$  and  $by + 1$ .*

In other words,  $x$  and  $y$  can be chosen so that  $ax + 1$  and  $by + 1$  have arbitrarily large common factors.

*Proof.* Take  $z$  to be any prime larger than  $n$  which is also relatively prime to  $a$  and  $b$ . Now  $a$  and  $b$  are invertible modulo  $z$ , so we can find solutions  $x_0, y_0$  to the congruences  $ax_0 \equiv by_0 \equiv -1 \pmod{z}$ . Such  $x_0$  and  $y_0$  must also be relatively prime to

$z$ , so setting  $x_n = x_0 + nz$  and  $y_n = y_0 + nz$ , by Dirichlet's Theorem there are infinitely many primes  $x$  and  $y$  from the sequences  $\{x_n\}_{n \geq 0}$  and  $\{y_n\}_{n \geq 0}$ , respectively. In particular,  $x$  and  $y$  can be chosen distinct from each other as well as from all divisors of  $a$ ,  $b$ , and  $n$ . Since  $ax \equiv by \equiv -1 \pmod{z}$ , these  $x, y, z$  fulfill the requirements.  $\square$

We are now ready to prove Robinson's result.

**Theorem 1.2.4.** *The structure  $\langle \mathbb{N}, +, \times, 0, 1 \rangle$  is interpretable over  $\emptyset$  in the structure of the natural numbers in the language with the successor function and divisibility relation  $\mathcal{L} = \{s, |, 0, 1\}$ .*

*Proof.* In fact we follow Robinson's lead in excluding 0, since it complicates matters by demanding constant special consideration and repetitive exceptional cases. Thus while we usually follow the convention  $0 \in \mathbb{N}$ , for this proof only let's say that  $\mathbb{N}$  starts at 1. The proof including 0 in  $\mathbb{N}$  can be adapted with suitable easy, but tedious, modifications.

The set  $\mathcal{B}$  and equivalence relation  $\varepsilon$  from Definition 1.1.6 are simply  $\mathbb{N}$  and  $=$ . The interesting part is defining  $+$  and  $\times$  in terms of  $s$  and  $|$ .

Note that the least common multiple function  $[a, b]$  is definable in terms of  $|$ . Indeed,  $[a, b] = c$  iff  $\forall x (c | x \leftrightarrow (a | x \wedge b | x))$ . Similarly, the relation  $\perp$  of being relatively prime is definable from  $|$ ,  $a \perp b$  iff  $\forall x ((x | a \wedge x | b) \leftrightarrow x = 1)$ .

(Incidentally, 1 is still definable in this language, since  $x = 1$  iff  $\forall y (s(y) \neq x)$  iff  $\forall y (x | y)$ .)

Now, we need to show that  $+$  and  $\times$  are definable in  $\langle \mathbb{N}, s, |, 1 \rangle$ . First, multiplication. We claim that  $a \times b = c$  iff

$$\begin{aligned} & a = b = c = 1 \vee \\ & \forall x, y, z \left( (a \perp x \wedge b \perp y \wedge c \perp x \wedge c \perp y \wedge x \perp y \wedge z | s([a, x]) \wedge z | s([b, y])) \right. \\ & \left. \rightarrow \exists w (z | w \wedge s(w) = [c, [x, y]]) \right). \end{aligned}$$

First, if  $a \times b = c$  then take  $x$  prime to  $a$ ,  $c$ , and  $y$ ;  $y$  prime to  $b$ ,  $c$ , and  $x$ ; and  $z$  dividing both  $s([a, x]) = ax + 1$  and  $s([b, y]) = by + 1$ , as in the antecedent of the long conditional in the above formula. Then  $ax \equiv by \equiv -1 \pmod{z}$  implies  $abxy \equiv 1 \pmod{z}$ , that is, there exists  $w$  such that  $z | w$  and  $w = abxy - 1$ , giving  $s(w) = cxy = [c, [x, y]]$  as required.

Conversely, suppose that the formula holds for  $a, b, c$ . If  $c = 1$ , then also  $a = b = 1$ , since otherwise, by taking  $x = y = z = 1$  in the formula we would get a  $w$  such that  $s(w) = 1$ , which is impossible.

So assume  $c \neq 1$ . Then let  $x, y, z$  be any positive integers with  $z > abc$  satisfying everything in the second line of the formula. The existence of such integers is guaranteed by Corollary 1.2.3. Notice in particular that the least common multiples reduce to multiplication by relative primeness, so  $z | s([a, x]) = ax + 1$  and  $z | s([b, y]) = by + 1$ , and we thus obtain  $w$  satisfying  $z | w$  and  $w + 1 = cxy$ . Hence

$$\begin{aligned} ax &\equiv by \equiv -1 \pmod{z} \\ cxy &\equiv 1 \pmod{z} \end{aligned}$$

implies (by multiplying both sides of the second congruence by  $ab$ ) that

$$c \equiv ab \pmod{z}.$$

Since  $z$  is larger than both  $ab$  and  $c$ , we conclude that  $c = ab$ .

Next, in giving a definition of the graph of addition, we may now freely use multiplication as well as divisibility and the successor. Accordingly, we note that

$$\begin{aligned} \mathfrak{s}(ac)\mathfrak{s}(bc) = \mathfrak{s}(c^2\mathfrak{s}(ab)) &\Leftrightarrow \mathfrak{s}(ac)\mathfrak{s}(bc) = \mathfrak{s}(c^2\mathfrak{s}(ab)) \\ &\Leftrightarrow (ac + 1)(bc + 1) = c^2(ab + 1) + 1 \\ &\Leftrightarrow abc^2 + ac + bc + 1 = abc^2 + c^2 + 1 \\ &\Leftrightarrow c(a + b) = c^2 \\ &\Leftrightarrow a + b = c \end{aligned}$$

(using in the last step  $c \neq 0$ ). □

In view of Proposition 1.1.8, we conclude

**Proposition 1.2.5.** *The structure  $\langle \mathbb{N}, \mathfrak{s}, |, 0, 1 \rangle$  is undecidable.*

### 1.2.2 Undecidability of $\mathbb{Z}$

As a second example, we show (again in the language of rings) that the natural numbers are a definable subset of the integers. Here we again need to invoke a classical theorem of number theory, but in this case we take the opportunity to give a complete proof of this theorem. This is both because it's a charming proof, and also because unlike the previous example this is one that will be used repeatedly in what follows.

The theorem is the Four Square Theorem of Lagrange, which states that every natural number can be expressed as the sum of four squares. The proof we give, however, is due to Hurwitz [14] and is based on factorization in the quaternions.

**Definition 1.2.6.** *The ring  $\mathbb{H}$  of quaternions consists of elements of the form  $\alpha = a_1 + a_2\mathbf{i} + a_3\mathbf{j} + a_4\mathbf{k}$  with  $a_i \in \mathbb{R}$  and such that  $\mathbf{i}, \mathbf{j}, \mathbf{k}$  commute with elements of  $\mathbb{R}$  and satisfy the identities  $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$ . The coordinates of  $\alpha$  are the values  $a_1, a_2, a_3, a_4$ .*

It is easy to check that  $\mathbf{ij} = -\mathbf{ji} = \mathbf{k}$ ,  $\mathbf{jk} = -\mathbf{kj} = \mathbf{i}$ , and  $\mathbf{ki} = -\mathbf{ik} = \mathbf{j}$ , so that  $\mathbb{H}$  is noncommutative. Though this will not be of use for us, it is interesting to note that  $\mathbb{H}$  is one of only three division rings which are finite-dimensional as algebras over  $\mathbb{R}$  (the other two being  $\mathbb{R}$  and  $\mathbb{C}$ ). Most basic algebra books mention the quaternions, but if your appetite is such to demand a whole book about them, try [4].

Much like in  $\mathbb{C}$ , if we define the conjugate of  $\alpha = a_1 + a_2\mathbf{i} + a_3\mathbf{j} + a_4\mathbf{k}$  by  $\bar{\alpha} = a_1 - a_2\mathbf{i} - a_3\mathbf{j} - a_4\mathbf{k}$ , then  $\mathbb{H}$  admits a norm  $N(\alpha) = \alpha\bar{\alpha} = a_1^2 + a_2^2 + a_3^2 + a_4^2$ . That the norm is multiplicative,  $N(\alpha\beta) = N(\alpha)N(\beta)$  for all  $\alpha, \beta \in \mathbb{H}$ , can be checked explicitly

or using Euler's four-square identity

$$\begin{aligned} (a_1^2 + a_2^2 + a_3^2 + a_4^2)(b_1^2 + b_2^2 + b_3^2 + b_4^2) = \\ (a_1b_1 - a_2b_2 - a_3b_3 - a_4b_4)^2 + (a_1b_2 + a_2b_1 + a_3b_4 - a_4b_3)^2 + \\ (a_1b_3 - a_2b_4 + a_3b_1 + a_4b_2)^2 + (a_1b_4 + a_2b_3 - a_3b_2 + a_4b_1)^2 \end{aligned}$$

(the quantities being squared on the right being precisely the coordinates obtained on multiplying  $\alpha = a_1 + a_2\mathbf{i} + a_3\mathbf{j} + a_4\mathbf{k}$  and  $\beta = b_1 + b_2\mathbf{i} + b_3\mathbf{j} + b_4\mathbf{k}$ ).

The proof proceeds much like Dedekind's proof of Fermat's Theorem on primes representable as the sum of two squares: an odd prime  $p$  is the sum of two squares iff  $p \equiv 1 \pmod{4}$ . Dedekind's idea was to study which primes factor in the *Gaussian integers*, the subring of  $\mathbb{C}$  consisting of complex numbers  $a + b\mathbf{i}$  with  $a, b \in \mathbb{Z}$ .

We are thus led to consider the so-called *Lipschitz integers*,

$$L = \{a_1 + a_2\mathbf{i} + a_3\mathbf{j} + a_4\mathbf{k} \in \mathbb{H} \mid a_i \in \mathbb{Z} \text{ for each } i\}.$$

Accordingly, Lagrange's Theorem can be restated to say that *every natural number is the norm of a Lipschitz integer*. But weirdly, the Lipschitz integers somehow fail to be the right analogy to the integers inside of  $\mathbb{H}$ , because there is no Euclidean algorithm in  $L$ .

**Definition 1.2.7.** *The Hurwitz integers is the subring  $H$  of the quaternions consisting of the elements*

$$\frac{a_1}{2} + \frac{a_2}{2}\mathbf{i} + \frac{a_3}{2}\mathbf{j} + \frac{a_4}{2}\mathbf{k}$$

*with the integers  $a_1, a_2, a_3, a_4$  either all even or all odd. Equivalently, Hurwitz integers are quaternions whose coordinates are either all in  $\mathbb{Z}$  or all in  $\mathbb{Z} + \frac{1}{2}$ .*

In fact it is not an entirely trivial exercise to prove that  $H$  is closed under multiplication, but the reward is a version of the Euclidean algorithm:

**Lemma 1.2.8.** *Given two Hurwitz integers  $\alpha, \beta \in H$  with  $\beta \neq 0$ , there are  $\delta, \rho \in H$  with  $N(\rho) < N(\beta)$  such that  $\alpha = \delta\beta + \rho$ .*

*Proof.* Consider  $\alpha\beta^{-1} \in \mathbb{H}$  as an element of  $\mathbb{R}^4$ . The nearest Lipschitz integer  $\delta$  (not necessarily unique) has distance from  $\alpha\beta^{-1}$  at most

$$\sqrt{(1/2)^2 + (1/2)^2 + (1/2)^2 + (1/2)^2} = 1.$$

Furthermore, this distance is in fact  $< 1$  unless  $\alpha\beta^{-1} \in H \setminus L$ , in which case set  $\delta = \alpha\beta^{-1}$  instead. Thus in either case,  $\delta \in H$  and  $N(\alpha\beta^{-1} - \delta) < 1$ .

Now, letting  $\rho = (\alpha\beta^{-1} - \delta)\beta$ , we have  $\alpha = \delta\beta + \rho$  and  $N(\rho) = N(\alpha\beta^{-1} - \delta)N(\beta) < N(\beta)$ .  $\square$

**Lemma 1.2.9.** *Every  $\alpha \in H$  can be written as the product of a unit  $v \in H$  and a Lipschitz integer  $\lambda \in L$ .*

*Proof.* Any  $\alpha \in H \setminus L$  can be written in the form

$$(b_1 + b_2\mathbf{i} + b_3\mathbf{j} + b_4\mathbf{k}) + \left( \pm \frac{1}{2} \pm \frac{1}{2}\mathbf{i} \pm \frac{1}{2}\mathbf{j} \pm \frac{1}{2}\mathbf{k} \right) \quad (1.2)$$

with  $b_1, b_2, b_3, b_4$  all *even* integers. Let  $\beta = b_1 + b_2\mathbf{i} + b_3\mathbf{j} + b_4\mathbf{k}$  and

$$v = \left( \pm \frac{1}{2} \pm \frac{1}{2}\mathbf{i} \pm \frac{1}{2}\mathbf{j} \pm \frac{1}{2}\mathbf{k} \right)$$

as in (1.2). It is clear that  $v$  is a unit with inverse  $\bar{v}$ . Then  $\lambda = \bar{v}(\beta + v) = \bar{v}\beta + 1$  works, since

$$v\lambda = v\bar{v}\beta + v = \beta + v = \alpha.$$

In particular,  $\lambda \in L$  because  $\bar{v}\beta \in L$ , since the coordinates of  $\beta$  are all even.  $\square$

Therefore, the norm of a Hurwitz integer  $\alpha$ ,  $N(\alpha) = N(v\lambda) = N(v)N(\lambda) = N(\lambda)$ , is equal to the norm of some  $\lambda \in L$ . It follows that norms of Hurwitz integers are sums of four integer squares (and in particular are themselves integers). Incidentally, this gives an easy way of describing the units in  $H$ .

**Lemma 1.2.10.** *The units of  $H$  are precisely those elements of  $H$  of the form  $\pm 1$ ,  $\pm \mathbf{i}$ ,  $\pm \mathbf{j}$ ,  $\pm \mathbf{k}$ , and  $\frac{1}{2}(\pm 1 \pm \mathbf{i} \pm \mathbf{j} \pm \mathbf{k})$  (so, 24 of them altogether).*

*Proof.* Since a unit  $v \in H$  must have norm  $N(v) = 1$ , one verifies easily that the listed elements are the only possibilities.  $\square$

We need one last number theoretic lemma before proving the main theorem.

**Lemma 1.2.11.** *For any odd prime  $p$ , there are  $m, n \in \mathbb{N}$  such that*

$$p \mid 1 + m^2 + n^2.$$

*Proof.* Note that the  $\frac{p+1}{2}$  integers  $0, 1, 2, \dots, \frac{p-1}{2}$  all have distinct squares modulo  $p$ . Thus  $1 + m^2$  takes on  $\frac{p+1}{2}$  distinct residues modulo  $p$ , and likewise  $-n^2$ . Hence by the pigeonhole principle, there must be  $m, n \in \mathbb{N}$  such that  $1 + m^2 \equiv -n^2 \pmod{p}$ , i.e.  $p \mid 1 + m^2 + n^2$ .  $\square$

**Theorem 1.2.12** (Lagrange, 1770). *Every natural number  $n \in \mathbb{N}$  can be written as the sum of four integer squares  $n = a^2 + b^2 + c^2 + d^2$ .*

*Proof.* We show that  $n$  can be expressed as the norm of a Lipschitz integer. By Lemma 1.2.9, it suffices to prove that  $n = N(\alpha)$  for some  $\alpha \in H$ . Also, by multiplicativity of  $N$ , it suffices to prove this for primes  $p$ . As the case  $p = 2 = N(1 + \mathbf{i})$  is trivial, we assume  $p$  is odd.

Now let  $m$  and  $n$  be as in Lemma 1.2.11, say  $pq = 1 + m^2 + n^2$ . So

$$p \mid 1 + m^2 + n^2 = (1 + m\mathbf{i} + n\mathbf{j})(1 - m\mathbf{i} - n\mathbf{j}).$$

But  $p$  divides neither  $1 + m\mathbf{i} + n\mathbf{j}$  nor  $1 - m\mathbf{i} - n\mathbf{j}$  in  $H$ , since for example

$$\frac{1}{p} + \frac{m}{p}\mathbf{i} + \frac{n}{p}\mathbf{j} \notin H.$$

Let  $\alpha \in H$  be a nonzero Hurwitz integer of minimal norm which can be expressed in the form  $\alpha = \lambda(1 - m\mathbf{i} - n\mathbf{j}) + \mu p$  for  $\lambda, \mu \in H$ . Using Lemma 1.2.8 to write  $p = \delta\alpha + \rho$  with  $N(\rho) < N(\alpha)$ , since

$$\rho = p - \delta\alpha = (-\delta\lambda)(1 - m\mathbf{i} - n\mathbf{j}) + (1 - \delta\mu)p$$

the minimality of  $N(\alpha)$  gives  $\rho = 0$ . Therefore  $p = \delta\alpha$ . By the same reasoning,  $1 - m\mathbf{i} - n\mathbf{j} = \gamma\alpha$  for some  $\gamma \in H$ .

Since  $N(\delta)N(\alpha) = N(p) = p^2$  in  $\mathbb{N}$ , we must have one of  $N(\alpha) = 1$ ,  $N(\alpha) = p$ , or  $N(\alpha) = p^2$ .

But  $N(\alpha) \neq 1$ , because in this case we could assume  $\alpha = 1$ , and multiplying  $1 = \lambda(1 - m\mathbf{i} - n\mathbf{j}) + \mu p$  on the right by  $1 + m\mathbf{i} + n\mathbf{j}$  gives

$$\begin{aligned} 1 + m\mathbf{i} + n\mathbf{j} &= \lambda(1 + m^2 + n^2) + \mu(1 + m\mathbf{i} + n\mathbf{j})p \\ &= (\lambda q + \mu(1 + m\mathbf{i} + n\mathbf{j}))p, \end{aligned}$$

which violates  $p \nmid 1 + m\mathbf{i} + n\mathbf{j}$ .

Likewise  $N(\alpha) \neq p^2$ , because in this case  $N(\delta) = 1$  means that  $\delta$  is a unit in  $H$ . Now  $\gamma\alpha = (\gamma\delta^{-1})p = 1 - m\mathbf{i} - n\mathbf{j}$  contradicts  $p \nmid 1 - m\mathbf{i} - n\mathbf{j}$ .

Therefore we must have  $N(\alpha) = p$ , completing the proof.  $\square$

The main idea here, once  $p$  has been shown to divide  $(1 + m\mathbf{i} + n\mathbf{j})(1 - m\mathbf{i} - n\mathbf{j})$  but not  $1 + m\mathbf{i} + n\mathbf{j}$  or  $1 - m\mathbf{i} - n\mathbf{j}$ , is to use unique factorization to show that  $p$  factors nontrivially in  $H$ . Though a form of unique factorization does hold in  $H$ , it is rather delicate to handle this in a noncommutative domain. We dodge this complication by using the Euclidean algorithm from Lemma 1.2.8 to get a concept of greatest common divisor in  $H$ .

Finally, we conclude

**Corollary 1.2.13.**  $\mathbb{N}$  is defined in  $\mathbb{Z}$  by the formula

$$v(x) \equiv \exists x_1, x_2, x_3, x_4 (x = x_1^2 + x_2^2 + x_3^2 + x_4^2).$$

Consequently,  $\mathbb{Z}$  is undecidable as a ring.

---

## Chapter 2

# The integers and the rational numbers

In this section, absolute values and completions of  $\mathbb{Q}$  are introduced and then used to study rational quadratic forms. The crowning achievement is Hasse's theorem [11] that the representation of a rational number by a rational quadratic form is equivalent to its representation in every completion of  $\mathbb{Q}$ . Invoking the terminology that  $\mathbb{Q}$  is a global field while its completions are local fields (as in, localized to a specific metric), this translates to the “local-global principle” that understanding rational quadratic forms *globally* equates to understanding them *locally everywhere*.

Unfortunately, based on time constraints it is necessarily to omit some of the key number theoretic details, as Hasse's theorem is by no means elementary. The main goal is to develop the ideas at least far enough to understand the principle, if not the ideas of the proof. This may seem anticlimactic as our intended application, Robinson's result that  $\mathbb{Z}$  is definable in  $\mathbb{Q}$  (and hence that  $\mathbb{Q}$  is undecidable), will end up using only lemmas stated without proof. However, many of the results will again be relevant in later sections.

### 2.1 Absolute values and completions of $\mathbb{Q}$

Recall from undergraduate analysis courses how  $\mathbb{R}$  is constructed from  $\mathbb{Q}$ . One notices that not all Cauchy sequences converge in  $\mathbb{Q}$ , exposing some missing holes in the rational number line. This problem is addressed by defining a new field  $\mathbb{R}$  on equivalence classes of Cauchy sequences.  $\mathbb{R}$  is *complete* in the sense that every sequence which should converge does converge. This notion of convergence is based fundamentally on the absolute value, which gives the metric  $d(x, y) = |x - y|$  on  $\mathbb{Q}$ . So, let's generalize this.

I have used the texts by Engler and Prestel [9] and Koblitz [17] as references for this section, and both would be a good starting point for further reading on valuation theory and the  $p$ -adics.

### 2.1.1 Absolute values

**Definition 2.1.1.** An absolute value on a field  $K$  is a map  $|\cdot|: K \rightarrow \mathbb{R}$  such that:

- (a)  $|x| \geq 0$  for all  $x$  and  $|x| = 0$  iff  $x = 0$ ,
- (b)  $|xy| = |x||y|$ , and
- (c)  $|x + y| \leq |x| + |y|$  (the triangle inequality).

Note that it follows easily from the definition that for any absolute value,  $|1| = |-1| = 1$ .

An important dividing line between absolute values depends on what  $|n|$  looks like as  $n$  ranges over  $\mathbb{N}$ .

**Proposition 2.1.2.** The set  $\{|n| \mid n \in \mathbb{N}\}$  is bounded in  $\mathbb{R}$  iff  $|\cdot|$  satisfies for all  $x, y$

$$|x + y| \leq \max\{|x|, |y|\}. \quad (2.1)$$

*Proof.* First of all, if (2.1) is satisfied then it is easily seen by induction that  $|n| = |1 + \dots + 1| \leq |1| = 1$ . So  $|n|$  is not only bounded in  $\mathbb{R}$ , but bounded by 1.

Conversely, suppose  $|n| \leq C$  for all  $n \in \mathbb{N}$ . Then for all  $x, y$ , and  $n$

$$|x + y|^n = \left| \sum_{i=0}^n \binom{n}{i} x^i y^{n-i} \right| \leq \sum_{i=0}^n \binom{n}{i} |x^i y^{n-i}| \quad (2.2)$$

$$\leq (n+1)C (\max\{|x|, |y|\})^n \quad (2.3)$$

with the inequality in (2.2) a consequence of the triangle inequality, while the one in (2.3) is due to the absolute value of the binomial coefficient being bounded by  $C$  and  $|x^i y^{n-i}| \leq (\max\{|x|, |y|\})^n$ .

Now, from  $|x + y|^n \leq (\max\{|x|, |y|\})^n$  we get

$$|x + y| \leq \sqrt[n]{n+1} \sqrt[n]{C} \max\{|x|, |y|\}.$$

Since this is true for all  $n$ , the inequality still holds in the limit as  $n \rightarrow \infty$ . Using

$$\lim_{n \rightarrow \infty} \sqrt[n]{n+1} = \lim_{n \rightarrow \infty} \sqrt[n]{C} = 1$$

we thus obtain  $|x + y| \leq \max\{|x|, |y|\}$ . □

The result is that the absolute values on  $K$  divide precisely into two classes:

1. The *archimedean* absolute values, for which there are natural numbers of arbitrarily large absolute value.
2. The *nonarchimedean* absolute values, which in place of the triangle inequality satisfy the stronger *ultrametric inequality* of (2.1).

Just like the usual absolute value in  $\mathbb{Q}$ ,  $\mathbb{R}$ , or  $\mathbb{C}$ , an absolute value defines a metric  $d$  on its field by  $d(x, y) = |x - y|$ . A topology on  $K$  is then devised in the customary way by taking as basic open sets the open balls

$$B_r(\alpha) = \{x \in K \mid |x - \alpha| < r\}$$

(with center  $\alpha$  and radius  $r$ ).

In the ultrametric case, however, this topology does not always align with our intuition about balls. For example, one can show without too much trouble for a nonarchimedean absolute value that if  $|x| \neq |y|$  then in fact  $|x + y| = \max\{|x|, |y|\}$ , and consequently that if  $\beta \in B_r(\alpha)$  then  $B_r(\alpha) = B_r(\beta)$ . So in fact every element in an open ball is at the center.

One more note on this topology before we return to the rational numbers.

**Definition 2.1.3.** *Two absolute values  $|\cdot|_1, |\cdot|_2$  on a field  $K$  are equivalent if they induce the same topology.*

**Lemma 2.1.4.** *If there is a real number  $c > 0$  such that  $|x|_1 = (|x|_2)^c$  for all  $x \in K$ , then  $|\cdot|_1$  and  $|\cdot|_2$  are equivalent.*

*Proof.* The open ball  $B_r^2(\alpha)$  of radius  $r$  in the sense of  $|\cdot|_2$  is the open ball  $B_{r^c}^1(\alpha)$  of radius  $r^c$  in the sense of  $|\cdot|_1$ , and vice versa with  $1/c$  in place of  $c$ .  $\square$

### 2.1.2 Absolute values on $\mathbb{Q}$

As we will be studying arbitrary absolute values on  $\mathbb{Q}$ , let us introduce the convention that the ordinary absolute value will be denoted by  $|\cdot|_\infty$ ,

$$|x|_\infty = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0. \end{cases}$$

A second absolute value, the *trivial absolute value*  $|\cdot|_0$ , is obtained for  $\mathbb{Q}$  (and indeed for any field) by setting  $|x|_0 = 1$  for all nonzero  $x$ . The associated topology is the discrete topology: every set is open. This is obviously a degenerate case deserving our utmost scorn.

A third example comes from fixing a prime  $p$  and measuring how divisible by  $p$  a rational number is. Accordingly, define for nonzero  $x \in \mathbb{Q}$

$$\text{ord}_p(x) = n \text{ iff } x = p^n \frac{a}{b} \text{ with } p \nmid a, b.$$

**Definition 2.1.5.** *The  $p$ -adic absolute value  $|\cdot|_p$  on  $\mathbb{Q}$  is defined by*

$$|x|_p = \begin{cases} 0 & \text{for } x = 0 \\ p^{-\text{ord}_p(x)} & \text{for } x \neq 0. \end{cases}$$

It is not very hard to check that this truly is an absolute value, and in fact a nonarchimedean one since  $\text{ord}_p(n) \geq 0$  implies  $|n|_p \leq 1$  for all nonzero  $n \in \mathbb{N}$ . A remarkable fact is that up to equivalence, these examples constitute a complete list of absolute values on  $\mathbb{Q}$ .

**Theorem 2.1.6** (Ostrowski [23], 1916). *Every nontrivial absolute value  $|\cdot|$  on  $\mathbb{Q}$  is equivalent to  $|\cdot|_p$  for some prime  $p$  or  $p = \infty$ .*

*Proof.* We distinguish between two cases depending on whether or not  $|\cdot|$  is archimedean.

*The nonarchimedean case:* If  $|\cdot|$  is nonarchimedean, then obviously by (2.1)  $|n| \leq 1$  for every  $n \in \mathbb{N}$ . Since nontriviality is assumed, there must be some nonzero  $n$  with  $|n| < 1$ . Taking  $p$  to be the least such number, since  $|\cdot|$  respects multiplication it follows that  $p$  must be prime.

We claim that for any prime  $q \neq p$ ,  $|q| = 1$ . Taking  $a, b \in \mathbb{Z}$  such that  $ap + bq = 1$ ,

$$1 = |ap + bq| \leq \max\{|ap|, |bq|\} \leq 1.$$

Because  $|ap| < 1$ , it must be that  $|bq| = |q| = 1$ .

This implies that for any integer  $n$ ,  $|n| = |p|^{\text{ord}_p(n)}$ . Since by multiplicativity  $|n/m| = |n|/|m|$ , we thus have

$$|x| = |n|/|m| = |p|^{\text{ord}_p(n) - \text{ord}_p(m)} = |p|^{\text{ord}_p(x)}$$

for every nonzero  $x = n/m \in \mathbb{Q}$ .

Now let  $c$  be a positive real number such that  $|p| = p^{-c}$ , namely  $c = -\log_p |p|$ . Then from the calculation above, for any nonzero  $x \in \mathbb{Q}$ ,

$$|x| = |p|^{\text{ord}_p(x)} = p^{-c \text{ord}_p(x)} = \left(p^{-\text{ord}_p(x)}\right)^c = (|x|_p)^c.$$

Hence by Lemma 2.1.4,  $|\cdot|$  is equivalent to the  $p$ -adic absolute value  $|\cdot|_p$ .

*The archimedean case:* In this case the triangle inequality (and  $|1| = 1$ ) implies that  $|n| \leq n$  for all  $n \in \mathbb{N}$ . Assuming  $|\cdot|$  to be archimedean, first fix a natural number  $n > 1$  such that  $|n| > 1$ .<sup>1</sup>

Consider another  $m \in \mathbb{N}$  with  $|m| > 1$ . For any  $r \in \mathbb{N}$ , we can write  $m^r = a_0 + a_1 n + \dots + a_k n^k$  in base  $n$ , with  $0 \leq a_i < n$  for each  $i$ ,  $a_k \neq 0$ , and

$$k \leq \log_n(m^r) = r \frac{\ln(m)}{\ln(n)}.$$

Now, using  $|a_i| \leq a_i \leq n - 1$  and  $|n|^i \leq |n|^k$ ,

$$|m|^r \leq \sum_{i=0}^k |a_i| |n|^i \leq (n-1)(k+1) |n|^k \leq (n-1) \left( r \frac{\ln(m)}{\ln(n)} + 1 \right) |n|^{r \frac{\ln(m)}{\ln(n)}}.$$

If we take  $r$ th roots in the above inequality, as in the proof of Proposition 2.1.2, we obtain

$$|m| \leq \sqrt[r]{n-1} \sqrt[r]{r \frac{\ln(m)}{\ln(n)} + 1} |n|^{\frac{\ln(m)}{\ln(n)}}.$$

<sup>1</sup>This proof will imply that for any archimedean absolute value on  $\mathbb{Q}$ ,  $|n| \geq 1$  for all natural numbers  $n \geq 1$ , and can readily be adapted to show this fact directly: one leaves out the assumption  $|n| > 1$  and obtains in case  $|n| \leq 1$  that also  $|m| \leq 1$ .

As this too is true for any  $r \in \mathbb{N}$ , taking the limit as  $r \rightarrow \infty$  gives the result

$$|m|^{\ln(n)} \leq |n|^{\ln(m)}.$$

Finally, we may swap the roles of  $m$  and  $n$  to get the reverse inequality, so we conclude that in fact

$$|m| = |n|^{\frac{\ln(m)}{\ln(n)}} = |n|^{\log_n(m)}.$$

This holds for any  $m \in \mathbb{N}$  with  $|m| > 1$ .

Let  $c \in \mathbb{R}$  be such that  $|n| = n^c$  (noting that  $0 < c \leq 1$ ). We have thus shown that if  $m$  is an integer and  $|m| > 1$ , then

$$|m| = |\pm m| = \|m\|_\infty = n^{c \log_n(|m|_\infty)} = (|m|_\infty)^c.$$

But for any  $x \in \mathbb{Q}$ ,  $x$  can be written as the quotient of integers  $m_1/m_2$  with both  $|m_1|, |m_2| > 1$ . This is because if, say,  $|m_1| < 1$ , then finding a suitable integer  $s$  such that  $|s| > 1/|m_1|$  courtesy of Proposition 2.1.2, then  $|sm_1| > 1$  and we could instead write  $x = sm_1/sm_2$ .

It follows that for all  $x \in \mathbb{Q}$ ,  $|x| = (|x|_\infty)^c$ , and  $|\cdot|$  is equivalent to the usual absolute value  $|\cdot|_\infty$  by Lemma 2.1.4.  $\square$

### 2.1.3 The $p$ -adic numbers

Now let's look sketchily at the construction of the  $p$ -adic numbers  $\mathbb{Q}_p$ , the completion of  $\mathbb{Q}$  with respect to the  $p$ -adic absolute value. These fields were first described by Hensel at the end of the 19th century. Much greater detail can be found, again, in [17].

A sequence  $\{x_n\}_{n \in \mathbb{N}}$  of rational numbers is *Cauchy* with respect to an absolute value  $|\cdot|$  iff for every  $\varepsilon > 0$  there is an  $N$  such that  $|x_n - x_m| < \varepsilon$  whenever  $m, n > N$ . It is *convergent* if there is an  $x$  with the property that, for every  $\varepsilon > 0$ , there is  $N$  such that  $|x - x_n| < \varepsilon$  whenever  $n > N$ .

In the  $p$ -adic absolute value, smallness of  $|x_n - x_m|_p$  corresponds to largeness of  $\text{ord}_p(x_n - x_m)$ . In other words,  $|x_n - x_m|_p$  is small if  $x_n - x_m$  is divisible by a large power of  $p$ . For example, 3-adically speaking the integer 7625597484987 is peculiarly much smaller than 2 (which for its part is as 3-adically large as an integer can be), and hence nowhere near 7625597484985.

As with the real absolute value, not all Cauchy sequences in  $\mathbb{Q}$  relative to  $|\cdot|_p$  converge. For example, the sequence given by

$$x_n = 1 + p + \dots + p^n$$

is Cauchy but fails to converge.

Now proceed as in completing  $\mathbb{Q}$  to  $\mathbb{R}$ . Define a notion of equivalence for two Cauchy sequences, show that it respects coordinatewise multiplication and addition, that  $\text{ord}_p$  and  $|\cdot|_p$  can be extended sensibly to these sequences, and so on. The  $p$ -adic numbers  $\mathbb{Q}_p$  are this field of equivalence classes of Cauchy sequences, and  $\mathbb{Q}_p$  is complete in the sense that every Cauchy sequence in  $\mathbb{Q}_p$  converges in  $\mathbb{Q}_p$ .

In practice this is a somewhat unwieldy way to think about  $\mathbb{Q}_p$ . It is nicer to picture a  $p$ -adic number as a series

$$a_n p^n + a_{n+1} p^{n+1} + a_{n+2} p^{n+2} + \dots \quad (2.4)$$

with  $n \in \mathbb{Z}$  ( $n$  may be negative!) and for each  $i \geq n$ ,  $0 \leq a_i < p$ . Addition and multiplication on such series goes as expected. Define also  $|a_n p^n + \dots|_p = p^{-n}$  if  $a_n \neq 0$ . This is a field complete with respect to  $|\cdot|_p$  containing  $\mathbb{Q}$ . Now by uniqueness of the completion we see that this must be the same  $\mathbb{Q}_p$  constructed earlier.

Ostrowski's Theorem, therefore, gives us a short yet exhaustive list of the completions of  $\mathbb{Q}$  with respect to a nontrivial absolute value:

1. Relative to the real absolute value  $|\cdot|_\infty$ , the real numbers  $\mathbb{Q}_\infty = \mathbb{R}$ .
2. For each prime  $p$ , relative to  $|\cdot|_p$ , the  $p$ -adic numbers  $\mathbb{Q}_p$ .

By analogy to  $\mathbb{Z}$ ,  $\mathbb{Q}_p$  contains a subring  $\mathbb{Z}_p$ , the *p-adic integers*, consisting of those  $x \in \mathbb{Q}_p$  for which  $|x|_p \leq 1$  (i.e.,  $n \geq 0$  in (2.4)).  $\mathbb{Z} \subseteq \mathbb{Z}_p$  and in fact the localized ring  $\mathbb{Z}_{(p)} = \{a/b \in \mathbb{Q} \mid p \nmid b\}$  is also a subring of  $\mathbb{Z}_p$ .

By contrast with  $\mathbb{R}$ , this absolute value gives what is called a *discrete valuation*. The 'valuation' is the (integer-valued) function  $\text{ord}_p$  on which  $|\cdot|_p$  is based. Also, from the above definition of  $|\cdot|_p$  on  $\mathbb{Q}_p$  it is clear that for every  $x \in \mathbb{Q}_p$  there is a  $y \in \mathbb{Q}$  such that  $|x|_p = |y|_p$ .

It is convenient in working with  $\mathbb{Q}_p$  to extend the congruence notation  $x \equiv y \pmod{p^n}$  in the natural way to mean  $\text{ord}_p(x - y) \geq n$ , or  $|x - y|_p \leq p^{-n}$ .

A fundamental tool for the  $p$ -adic fields is the following fact known as *Hensel's Lemma*. This is based on a version of Newton's Method for approximating roots of polynomials, but unlike in  $\mathbb{R}$ , in  $\mathbb{Q}_p$  this method always converges to a root.

**Theorem 2.1.7** (Hensel). *Suppose that  $f(X) \in \mathbb{Z}_p[X]$  is a polynomial over the p-adic integers, and  $a \in \mathbb{Z}_p$  such that  $f(a) \equiv 0 \pmod{p}$  but  $f'(a) \not\equiv 0 \pmod{p}$ . Then there is  $\tilde{a} \in \mathbb{Z}_p$  such that  $f(\tilde{a}) = 0$  and  $a \equiv \tilde{a} \pmod{p}$ .*

*Proof.* We inductively find a sequence  $a = a_0, a_1, \dots$  in  $\mathbb{Z}_p$  satisfying

- (i)  $a_n \equiv a_{n-1} \pmod{p^n}$ ,
- (ii)  $f(a_n) \equiv 0 \pmod{p^{n+1}}$ , and
- (iii)  $f'(a_n) \equiv f'(a) \not\equiv 0 \pmod{p}$ .

Clearly this will suffice as then  $\{a_n\}_{n \geq 0}$  would form a Cauchy sequence converging to the root  $\tilde{a}$  of  $f(X)$ .

So, supposing  $a_0, \dots, a_{n-1}$  already constructed, we seek  $a_n = a_{n-1} + p^n b$  (some  $b \in \mathbb{Z}$ ) such that  $f(a_n) \equiv 0 \pmod{p^{n+1}}$ . In fact if  $b$  satisfies

$$p^n b \equiv -\frac{f(a_{n-1})}{f'(a_{n-1})} \pmod{p^{n+1}} \quad (2.5)$$

then the desired congruences hold by the computations

$$\begin{aligned} f(a_{n-1} + p^n b) &= \sum_{i \geq 0} \frac{f^{(i)}(a_{n-1})}{i!} (p^n b)^i \\ &= f(a_{n-1}) + f'(a_{n-1}) p^n b + \frac{f''(a_{n-1})}{2!} p^{2n} b^2 + \dots \\ &\equiv f(a_{n-1}) + f'(a_{n-1}) p^n b \pmod{p^{n+1}} \end{aligned}$$

and  $f(a_{n-1} + p^n b) \equiv f(a) \equiv f(a_0) \not\equiv 0 \pmod{p}$ .

That such a  $b$  as in (2.5) can always be found follows from the fact that  $f(a_{n-1}) \equiv 0 \pmod{p^n}$  and  $f'(a_{n-1}) \not\equiv 0 \pmod{p}$ . In particular note that  $f'(a_{n-1})$  is invertible modulo  $p$  by the condition (iii).  $\square$

With a small amount of extra care in the above proof one can also get uniqueness of the root  $\tilde{a}$ , but we won't need this. As a sample application we prove:

**Corollary 2.1.8.** *Let  $p$  be an odd prime, and suppose  $c \in \mathbb{Q}_p$ ,  $\text{ord}_p(c) = n$ . Then  $c$  is a square in  $\mathbb{Q}_p$  iff  $n$  is even and  $c/p^n$  is a square modulo  $p$  (that is, there exists  $a \in \mathbb{Z}_p$  such that  $a^2 \equiv c/p^n \pmod{p}$ ).*

*Proof.* Consider the polynomial  $f(X) = X^2 - c/p^n \in \mathbb{Z}_p[x]$ . If  $a^2 \equiv c/p^n \pmod{p}$ , then  $\text{ord}_p(c/p^n) = 0$  implies  $\text{ord}_p(a) = 0$ . Thus we have  $f(a) \equiv 0 \pmod{p}$  and

$$f'(a) = 2a \not\equiv 0 \pmod{p}.$$

Now Hensel's Lemma gives  $\tilde{a} \in \mathbb{Z}_p$  such that  $\tilde{a}^2 = c/p^n$ . Setting  $n = 2k$ , we get that  $(p^k \tilde{a})^2 = c$ .

Conversely, suppose that  $c = \alpha^2$  for  $\alpha \in \mathbb{Q}_p$ . Then  $n = \text{ord}_p(c) = 2 \text{ord}_p(\alpha)$  implies that  $n$  is even, and  $(\alpha/p^{\text{ord}_p(\alpha)})^2 \equiv c/p^n \pmod{p}$ .  $\square$

Finally, this last result together with Ostrowski's Theorem gives a nice characterization of the squares in  $\mathbb{Q}$  which conveniently foreshadows the next section. In fact something weaker than 2.1.8 will suffice. We need only the fact that  $\text{ord}_p(c)$  is even if  $c$  is a square in  $\mathbb{Q}_p$ . This also holds in  $\mathbb{Q}_2$ .

**Proposition 2.1.9.** *For any  $c \in \mathbb{Q}$ ,  $c$  is a square in  $\mathbb{Q}$  iff  $c$  is a square in  $\mathbb{Q}_p$  for every  $p$  (including  $p = \infty$ ).*

*Proof.* One direction is obvious. If  $b^2 = c$  and  $b \in \mathbb{Q}$ , then since  $\mathbb{Q} \subseteq \mathbb{Q}_p$  for every  $p$ ,  $c$  is also a square in each  $\mathbb{Q}_p$ .

For the converse, suppose that  $c \in \mathbb{Q}$  and  $c$  is a square in every  $\mathbb{Q}_p$ . Writing  $c = (-1)^{e_0} p_1^{e_1} \cdots p_n^{e_n}$  with  $p_1, \dots, p_n$  distinct primes,  $e_0 \in \{0, 1\}$  and  $e_i \in \mathbb{Z}$  for  $i \geq 1$ , observe that

- (a)  $e_0 = 0$  since  $c$  is a square in  $\mathbb{R}$ , and
- (b)  $e_i = \text{ord}_{p_i}(c)$  is even for every  $i \geq 1$  since  $c$  is a square in  $\mathbb{Q}_{p_i}$ .

$$\text{Now } b = p_1^{e_1/2} \cdots p_n^{e_n/2} \in \mathbb{Q} \text{ and } b^2 = c. \quad \square$$

This sort of result is commonly referred to as a ‘local-global principle’. It says that something is true globally ( $c$  is square in the global field  $\mathbb{Q}$ ) if and only if it is true everywhere locally ( $c$  is square in every completion of  $\mathbb{Q}$ , the local fields  $\mathbb{Q}_p$ ).

## 2.2 Quadratic forms over $\mathbb{Q}$

Here we collect some facts about quadratic forms that will be needed. For further information I highly recommend the books by Cassels [2] (the rational case) and Lam [18] (over general fields of characteristic  $\neq 2$ ).

The motivation for studying quadratic forms, in terms of Robinson’s definition of  $\mathbb{Z}$  in  $\mathbb{Q}$ , is the sort of local-global principle hinted at in Proposition 2.1.9. By looking at the completion at a particular prime  $p$ , Robinson is able to find a set of criteria that imply that the denominator of a rational number is not divisible by  $p$ . Then by quantifying over all primes, we can require that the denominator of a rational number  $a \in \mathbb{Q}$  is not divisible by any prime, so  $a \in \mathbb{Z}$ .

### 2.2.1 Basic facts

**Definition 2.2.1.** An  $n$ -ary quadratic form over a field  $K$  is a homogeneous polynomial  $f(X_1, \dots, X_n)$  of degree two,

$$f(X_1, \dots, X_n) = \sum_{i,j \leq n} a_{ij} X_i X_j \quad (2.6)$$

(not all  $a_{ij} = 0$ ).

As long as  $K$  has characteristic  $\neq 2$ , by replacing each  $a_{ij}$  with  $\frac{a_{ij} + a_{ji}}{2}$  it is possible to assume in (2.6) that the coefficients  $a_{ij} = a_{ji}$ .

A quadratic form can be thought of as a matrix equation. If  $A_f = (a_{ij})$  and the variables are written as a column vector

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

then (2.6) becomes  $f(X) = X^T A_f X$ . Now the previous remark means that  $A_f$  may be chosen to be a *symmetric* matrix.

**Definition 2.2.2.** Two  $n$ -ary quadratic forms  $f(X)$  and  $g(X)$  are equivalent if there exists a nonsingular linear change of variables  $S$  (or if you prefer, an invertible  $n \times n$  matrix  $S$ ) such that  $f(X) = g(S(X))$ .

Also,  $f(X)$  is diagonal if  $A_f$  is diagonal, i.e.

$$f(X) = a_1 X_1^2 + \dots + a_n X_n^2.$$

**Proposition 2.2.3.** Every quadratic form over a field of characteristic  $\neq 2$  is equivalent to a diagonal quadratic form.

*Proof.* This is actually a general fact about symmetric matrices. Replacing  $X$  by  $S(X)$  in  $f(X) = X^T A_f X$  gives

$$f(S(X)) = (SX)^T A_f(SX) = X^T (S^T A_f S) X.$$

Two matrices  $A, B$  are called *congruent* if there exists  $S$  such that  $B = S^T A S$ . Every symmetric matrix  $A$  is congruent to a diagonal matrix  $B$ . To see this, consider multiplying  $A$  by a sequence of row operations to make  $A$  upper triangular. The corresponding column operations are the transpose of the row operations, and multiplying  $A$  by these on the other side will simultaneously make  $A$  lower triangular.  $\square$

**Definition 2.2.4.** A quadratic form  $f(X_1, \dots, X_n)$  over  $K$  represents  $\alpha \in K$  if there exist  $\beta_1, \dots, \beta_n \in K$  (not all 0) such that  $f(\beta_1, \dots, \beta_n) = \alpha$ .

A quadratic form is isotropic if it represents 0.

We wish to study the representation of various rational numbers by certain quadratic forms. Proposition 2.2.3 shows that attention can be restricted to diagonal quadratic forms. The next proposition will show that attention can be restricted further to representation of 0 at the cost of adding one additional variable. This observation is due to Hensel.

**Proposition 2.2.5.** If the (diagonal) quadratic form

$$f(X) = a_1 X_1^2 + \dots + a_n X_n^2$$

(with  $a_1 \cdots a_n \neq 0$ ) represents  $\alpha$ , say  $f(\beta_1, \dots, \beta_n) = \alpha$ , then in fact  $\beta_1, \dots, \beta_n$  may be chosen all nonzero.

*Proof.* Without loss of generality we suppose that  $\beta_1 \neq 0$  but  $\beta_2 = 0$ , and find  $\tilde{\beta}_1, \tilde{\beta}_2$  both nonzero such that

$$a_1 \beta_1^2 = a_1 \beta_1^2 + a_2 \beta_2^2 = a_1 \tilde{\beta}_1^2 + a_2 \tilde{\beta}_2^2.$$

Proceeding thusly we eliminate one by one every  $\beta_i = 0$ . There are two cases.

Case 1:  $a_1 a_2 \neq \pm 1$ .

Then let

$$\tilde{\beta}_1 = \beta_1 \frac{1 - a_1 a_2}{1 + a_1 a_2}, \quad \tilde{\beta}_2 = \beta_1 \frac{2 a_1}{1 + a_1 a_2}$$

noting that these are both well-defined and nonzero by  $a_1 a_2 \neq \pm 1$ . Then compute directly

$$\begin{aligned} a_1 \tilde{\beta}_1^2 + a_2 \tilde{\beta}_2^2 &= a_1 \beta_1^2 \left( \frac{1 - 2 a_1 a_2 + a_1^2 a_2^2}{(1 + a_1 a_2)^2} \right) + a_2 \beta_1^2 \left( \frac{4 a_1^2}{(1 + a_1 a_2)^2} \right) \\ &= a_1 \beta_1^2 \left( \frac{1 - 2 a_1 a_2 + a_1^2 a_2^2}{(1 + a_1 a_2)^2} \right) + a_1 \beta_1^2 \left( \frac{4 a_1 a_2}{(1 + a_1 a_2)^2} \right) \\ &= a_1 \beta_1^2 \left( \frac{1 + 2 a_1 a_2 + a_1^2 a_2^2}{(1 + a_1 a_2)^2} \right) \\ &= a_1 \beta_1^2 \end{aligned}$$

as required.

*Case 2:*  $a_1 a_2 = \pm 1$ .

In this case, set instead

$$\tilde{\beta}_1 = \beta_1 \frac{1 - 4a_1 a_2}{1 + 4a_1 a_2}, \quad \tilde{\beta}_2 = \beta_1 \frac{4a_1}{1 + 4a_1 a_2}.$$

Essentially the same computation as above then shows  $a_1 \tilde{\beta}_1^2 + a_2 \tilde{\beta}_2^2 = a_1 \beta_1^2$ .  $\square$

**Corollary 2.2.6.** *The quadratic form  $f(X_1, \dots, X_n) = a_1 X_1^2 + \dots + a_n X_n^2$  represents  $\alpha \in K$  iff*

$$g(X_1, \dots, X_n, Y) = a_1 X_1^2 + \dots + a_n X_n^2 - \alpha Y^2$$

*represents 0.*

*Proof.* If  $f(\beta_1, \dots, \beta_n) = \alpha$ , then  $g(\beta_1, \dots, \beta_n, 1) = 0$  shows that  $g$  is isotropic.

Conversely, if  $g(\beta_1, \dots, \beta_n, \gamma) = 0$ , then by the Proposition we may assume  $\gamma \neq 0$ .

Now

$$f(\beta_1/\gamma, \dots, \beta_n/\gamma) = \alpha$$

gives a representation of  $\alpha$  by  $f$ .  $\square$

In other words, the problem of representation of an arbitrary rational number (or any element of a field of characteristic  $\neq 2$ ) by an  $n$ -ary quadratic form reduces to the problem of representation of 0 by an  $(n+1)$ -ary quadratic form.

There is a righteous corollary to this corollary, although we won't find a use for it.

**Corollary 2.2.7.** *Isotropic quadratic forms are universal: they represent every  $\alpha \in K$ .*

*Proof.* If  $f(X)$  represents 0, then clearly for any  $\alpha \in K$ ,  $g(X, Y) = f(X) - \alpha Y^2$  represents 0. By 2.2.6,  $f$  represents  $\alpha$ .  $\square$

### 2.2.2 The Hasse Principle

The rest of this chapter will be completed later.

---

## Chapter 3

# Hilbert's Tenth Problem

We now devote ourselves to a proof of the unsolvability of Hilbert's Tenth Problem, one of the landmark theorems in logic of the last century. Hilbert's question, from his famous 1900 list of twenty-three major open problems for the new century (in English translation: [13]), asks for a process by which it can be determined, in a finite number of steps, whether a polynomial with integer coefficients has an integer root. It is interesting to note that Hilbert did not explicitly allow for the possibility that such an algorithm could not exist.

In the formulation of Chapter 1, this translates into the decidability question for the subset  $D \subseteq \text{Th}(\mathbb{Z})$  consisting of (true) sentences of the form

$$\exists x_1, \dots, x_n P(x_1, \dots, x_n) = 0$$

with  $P \in \mathbb{Z}[x_1, \dots, x_n]$ .

The standard resource for this subject is Matiyasevich's book [21], which gives a very clear presentation of the proof as well as a great deal of related material. However, we take a slightly different approach in presenting the proof more or less as it developed historically over two decades of work between Martin Davis, Hilary Putnam, Julia Robinson, and Yuri Matiyasevich.

In particular, it seems somewhat simpler (though probably less intuitive) to appeal to the theory of recursive functions as a foundation for algorithms rather than Turing machines. Those who insist on a machine-based model of computation can look up a proof of their equivalence, such as in the classic books of Davis [6] or Hermes [12].

The theorem was proved in three major stages. Each step gave an incrementally simpler presentation of the computably enumerable sets: first as sets defined arithmetically by formulas in the so-called Davis Normal Form, secondly as sets of solutions to exponential polynomial equations, and finally as sets of solutions<sup>1</sup> to ordinary polynomial equations. Since there are uncomputable computably enumerable sets, this last step implies the unsolvability of Hilbert's Tenth Problem.

---

<sup>1</sup>More accurately, the projection of these sets to some subset of the variables.

But in fact, this tells us a great deal more. To prove unsolvability of the problem, it would suffice to find just one uncomputable set characterized by a polynomial equation (note that it is clear that every such set is at least computably enumerable). The apparent extreme difficulty of even some very harmless-looking diophantine equations (for example,  $x^3 + y^3 = z^3$ ), coupled with the logician's historical benefit of having grown comfortable with unsolvable problems, makes it seem in hindsight relatively unsurprising that this should be the case.

The theorem of Davis, Putnam, Robinson, and Matiyasevich, however, goes considerably further, stating that *every* computably enumerable set is captured by a polynomial equation! At least to me, even making full use of hindsight, this still seems pretty incredible. A few ramifications will be discussed at the end of the chapter after all the key ideas have been made precise.

### 3.1 Preliminaries

So as not to interrupt the groove by repeatedly pausing to develop auxiliary material, we start with some preliminaries.

#### 3.1.1 Diophantine sets

**Definition 3.1.1.** A set  $S \subseteq \mathbb{Z}^n$  is called *diophantine over  $\mathbb{Z}$*  if there is a polynomial  $P(x_1, \dots, x_n; y_1, \dots, y_m) \in \mathbb{Z}[\bar{x}, \bar{y}]$  such that  $S$  is defined by the formula

$$\exists y_1, \dots, y_m P(x_1, \dots, x_n; y_1, \dots, y_m) = 0. \quad (3.1)$$

*Diophantine sets over  $\mathbb{N}$  are defined analogously except  $P$  is still permitted to be a polynomial over  $\mathbb{Z}$  (or equivalently, allow instead formulas  $\exists \bar{y} P(\bar{x}, \bar{y}) = Q(\bar{x}, \bar{y})$  with  $P, Q \in \mathbb{N}[\bar{x}, \bar{y}]$ ).*

*Finally, a diophantine formula or diophantine definition is one of the given form.*

As mentioned above, every diophantine set is certainly computably enumerable. Hilbert's Tenth Problem then becomes the question whether in fact every diophantine set is computable.

A further observation is that Hilbert's Tenth Problem as stated for the integers is equivalent to the same question over  $\mathbb{N}$ . Namely, if  $P(x_1, \dots, x_n)$  is a polynomial with integer coefficients, then

1.  $P(x_1, \dots, x_n) = 0$  has a solution in natural numbers iff

$$P(x_{11}^2 + x_{12}^2 + x_{13}^2 + x_{14}^2, \dots, x_{n1}^2 + x_{n2}^2 + x_{n3}^2 + x_{n4}^2) = 0$$

has a solution in integers (using Theorem 1.2.12).

2.  $P(x_1, \dots, x_n) = 0$  has a solution in integers iff

$$\prod_{\bar{e} \in \{0,1\}^n} P((-1)^{e_1} x_1, \dots, (-1)^{e_n} x_n) = 0$$

has a solution in natural numbers.

In the coming sections, it will mostly be more convenient to work over  $\mathbb{N}$  since  $\mathbb{N}$  is the foundation for the theory of computation.

It is also equivalent to replace the single polynomial equation in (3.1) by a system of polynomial equations.

**Proposition 3.1.2.** *The set  $\{\bar{x} \mid \exists \bar{y} P_1(\bar{x}, \bar{y}) = \dots = P_k(\bar{x}, \bar{y}) = 0\}$  is diophantine.*

*Proof.* Clearly  $\bar{x}$  is in the set iff  $\exists \bar{y} \left( \sum_{i=1}^k P_i(\bar{x}, \bar{y})^2 = 0 \right)$ . □

Similarly, the diophantine sets are closed under finite unions and intersections.

**Proposition 3.1.3.** *If  $S$  and  $T$  are both diophantine subsets of  $\mathbb{N}^n$ , then so are  $S \cup T$  and  $S \cap T$ .*

*Proof.* Suppose  $S$  and  $T$  are defined by  $\exists \bar{y} P(\bar{x}, \bar{y}) = 0$  and  $\exists \bar{z} Q(\bar{x}, \bar{z}) = 0$  (assuming, as we may, that  $\bar{y}$  and  $\bar{z}$  are entirely distinct sets of variables). Then

$$\bar{x} \in S \cup T \Leftrightarrow \exists \bar{y}, \bar{z} (P(\bar{x}, \bar{y}) = 0 \vee Q(\bar{x}, \bar{z}) = 0) \Leftrightarrow \exists \bar{y}, \bar{z} (P(\bar{x}, \bar{y})Q(\bar{x}, \bar{z}) = 0),$$

$$\bar{x} \in S \cap T \Leftrightarrow \exists \bar{y}, \bar{z} (P(\bar{x}, \bar{y}) = 0 \wedge Q(\bar{x}, \bar{z}) = 0) \Leftrightarrow \exists \bar{y}, \bar{z} (P(\bar{x}, \bar{y})^2 + Q(\bar{x}, \bar{z})^2 = 0)$$

give the desired diophantine formulas. □

However, negation is more problematic.

**Proposition 3.1.4.** *There is a diophantine set  $S \subseteq \mathbb{N}^n$  whose complement  $\mathbb{N}^n \setminus S$  is not diophantine.*

*Proof.* The class of diophantine sets is closed under conjunction, disjunction, and existential quantification. If it were also closed under negation, then all arithmetically definable sets would be diophantine. But not all definable sets in  $\mathbb{N}$  are computably enumerable. □

Curiously, it can be shown that every diophantine formula is equivalent to one of degree at most 4, at the cost of adding many more variables. Since we won't be using this fact, rather than a proof we simply give an illustrative example:

$$\begin{aligned} \exists y \ x^2 + xy^2 = 0 &\Leftrightarrow \exists y, a, b, c (a = x^2 \wedge b = xy \wedge c = by \wedge a + c = 0) \\ &\Leftrightarrow \exists y, a, b, c ((x^2 - a)^2 + (xy - b)^2 + (by - c)^2 + (a + c)^2 = 0). \end{aligned}$$

For later reference, we close the section with a list of a few easy examples of diophantine relations. The *remainder function*  $\text{rem}$  is defined by  $\text{rem}(a, n) = b$  iff  $a \equiv b \pmod{n}$  and  $0 \leq b < n$ .

**Proposition 3.1.5.** *The relations  $x \leq y$ ,  $x \neq y^2$ ,  $x \mid y$ ,  $\text{gcd}(x, y) = z$ ,  $x \equiv y \pmod{z}$ ,  $x \not\equiv y \pmod{z}$ , and  $\text{rem}(x, z) = y$  are all diophantine.*

<sup>2</sup>The relation  $x \neq y$  being diophantine means we can always negate equations inside the scope of an existential quantifier, but by the previous proposition, not outside.

*Proof.* All of these are entirely routine. For example,  $\gcd(x, y) = z$  iff

$$\exists a, b((ax = by + z \vee by = ax + z) \wedge z \mid x \wedge z \mid y)$$

and  $x \not\equiv y \pmod{z}$  iff

$$\exists u, v(u < z \wedge v < z \wedge u \neq v \wedge x \equiv u \pmod{z} \wedge y \equiv v \pmod{z}).$$

□

### 3.1.2 Two coding schemes

It will be necessary to encode sequences of natural numbers as a single number. There are two primary ways of doing this, each with its particular advantages.

The first option is the *Cantor code* which employs the usual method of enumerating the cartesian product of two countable sets as in Figure 3.1, which depicts a bijection  $C : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$  in which for example  $C(0, 2) = 3$ . A convenient fact is that this bijection is described by a polynomial equation.

**Proposition 3.1.6.** For all  $a, b \in \mathbb{N}$ ,  $2C(a, b) = (a + b)^2 + 3a + b$ .

*Proof.* By induction on  $C(a, b)$ . This is clear for  $C(0, 0) = 0$ . For other  $(a, b)$ , we have two cases to consider. First, if  $b > 0$ , then  $C(a, b) + 1 = C(a + 1, b - 1)$ , so we confirm that

$$\begin{aligned} 2C(a + 1, b - 1) &= 2C(a, b) + 2 \\ &= (a + b)^2 + 3a + b + 2 \\ &= 2((a + 1) + (b - 1))^2 + 3(a + 1) + (b - 1) \end{aligned}$$

as required.

If on the other hand  $b = 0$ , then  $C(a, 0) + 1 = C(0, a + 1)$  and

$$2C(0, a + 1) = 2C(a, 0) + 2 = a^2 + 3a + 2 = (a + 1)^2 + (a + 1).$$

So in either case, if  $2C(a, b) = (a + b)^2 + 3a + b$  and  $C(\tilde{a}, \tilde{b}) = C(a, b) + 1$ , then we have shown that  $2C(\tilde{a}, \tilde{b}) = (\tilde{a} + \tilde{b})^2 + 3\tilde{a} + \tilde{b}$ , completing the induction. □

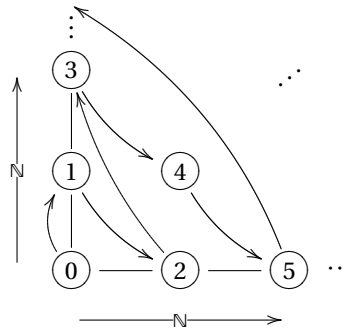


Figure 3.1: The Cantor numbering on  $\mathbb{N} \times \mathbb{N}$

Due to this handy equation, the Cantor code is a very efficient way of coding pairs of natural numbers. It can easily be extended to encode sequences of natural numbers of any fixed length as well. To this end, define the  $n$ th Cantor coding function  $C_n : \mathbb{N}^n \rightarrow \mathbb{N}$  recursively by  $C_2 = C$  and for  $n > 2$

$$C_n(a_1, \dots, a_n) = C(a_1, C_{n-1}(a_2, \dots, a_n)).$$

This can be defined inductively by a system of polynomial equations as in Proposition 3.1.6.

The following simple fact will be needed in Section 3.2.

**Proposition 3.1.7.** *If  $a_i \leq b_i$  for each  $i = 1, \dots, n$ , then  $C_n(a_1, \dots, a_n) \leq C_n(b_1, \dots, b_n)$ .*

*Proof.* The case  $n = 2$  follows from the formula in Proposition 3.1.6, and the general case is handled by an easy induction on  $n$ .  $\square$

For later reference, let us also fix the notation for the decoding functions

$$D_{i,n} = \pi_{i,n} \circ C_n^{-1}$$

giving  $a_i$  from the  $n$ th Cantor code of the  $n$ -tuple  $\langle a_1, \dots, a_n \rangle$ . Like  $C_n$ , each  $D_{i,n}$  is defined by a diophantine formula.

The drawback, however, is that a new system of equations is required any time we wish to lengthen (or shorten) the sequence. It would be much better to have a system that can uniformly allow for sequences of natural numbers of any finite length. This can be accomplished with the Chinese Remainder Theorem:

**Theorem 3.1.8.** *If  $m_1, \dots, m_n$  are (pairwise) relatively prime natural numbers and  $a_1, \dots, a_n \in \mathbb{N}$  are arbitrary, the system of congruences*

$$x \equiv a_i \pmod{m_i} \tag{3.2}$$

*always has a solution. (Furthermore this solution is unique modulo  $\prod_i m_i$ .)*

Consequently, provided we can find a sequence of  $n$  relatively prime integers greater than some  $N$ , then as long as  $N > \max\{a_1, \dots, a_n\}$ , the solution  $x$  to (3.2) uniquely determines the sequence  $a_1, \dots, a_n$ . But getting your hands on such a sequence can be tricky, which makes the Cantor coding preferable if the length of the sequence to be coded is fixed. However, the following proposition will help.

**Proposition 3.1.9.** *If  $N$  is divisible by  $i$  for all  $1 \leq i < n$ , then the numbers*

$$1 + N, 1 + 2N, \dots, 1 + nN$$

*are pairwise relatively prime.*

*Proof.* Suppose there were a prime  $p$  dividing both  $1 + iN$  and  $1 + jN$ , say with  $1 \leq i < j \leq n$ . Then  $p \mid (j - i)N$ , and  $j - i < n$  implies  $j - i \mid N$ . Thus  $p \mid N$ , which contradicts  $p \equiv 1 \pmod{N}$ .  $\square$

So if, for example, we can require that  $N$  is divisible by  $n!$ , then the Chinese Remainder Theorem gives us a useful way of coding sequences of up to  $n$  natural numbers, each less than  $N$ . The main difficulty, which will be dealt with in ??, is giving a *diophantine* definition of  $n!$  (or at least divisibility by all  $i < n$ ) in a language of rings which does not explicitly include the factorial.

### 3.1.3 Recursive functions

As mentioned earlier, we adopt as the basic model of computation the (partial) recursive functions. These are functions on (some subset of)  $\mathbb{N}^n$  built in a finite number of steps by combining some basic functions in some prescribed ways. Let us start by defining an important subclass, the primitive recursive functions.

**Definition 3.1.10.** *The class of primitive recursive functions is the smallest subclass of functions  $f : \mathbb{N}^n \rightarrow \mathbb{N}$  (some  $n \in \mathbb{N}$ ) containing the basic functions*

(a) *The constant function on  $\mathbb{N}^n$ ,  $c_n : \langle x_1, \dots, x_n \rangle \mapsto 0$*

(b) *For each  $i \leq n$  the projection  $\pi_{i,n} : \langle x_1, \dots, x_n \rangle \mapsto x_i$*

(c) *The successor function  $s : x \mapsto x + 1$*

*and closed under the operations of*

(i) *Composition: If  $f : \mathbb{N}^n \rightarrow \mathbb{N}$  and  $g_1, \dots, g_n : \mathbb{N}^m \rightarrow \mathbb{N}$  are primitive recursive then  $f(g_1, \dots, g_n) : \mathbb{N}^m \rightarrow \mathbb{N}$  is as well.*

(ii) *Primitive recursion: If  $h(x_1, \dots, x_n)$  and  $g(y, z, x_1, \dots, x_n)$  are primitive recursive then so is the function  $f : \mathbb{N}^{n+1} \rightarrow \mathbb{N}$  defined by*

- $f(0, x_1, \dots, x_n) = h(x_1, \dots, x_n)$
- $f(y + 1, x_1, \dots, x_n) = g(y, f(y, x_1, \dots, x_n), x_1, \dots, x_n)$ .

As an example, let us show that addition is primitive recursive. Indeed, if  $h(x) = x$  is the projection  $\pi_{1,1}$  and  $g(y, z, x) = z + 1$  the composition  $s(\pi_{2,3})$ , then  $f(y, x)$  defined from  $g$  and  $h$  by primitive recursion satisfies

- $f(0, x) = h(x) = x$
- $f(y + 1, x) = g(y, f(y, x), x) = f(y, x) + 1$ .

So it follows by induction on  $y$  that  $f(y, x) = y + x$ .

Obtaining multiplication similarly, one then sees for example that every polynomial function is primitive recursive. It is also clear that every primitive recursive function is computable. But not every computable function is primitive recursive. For this an additional operation is needed.

**Definition 3.1.11.** *The class of partial recursive functions is the smallest subclass of functions  $f : S \rightarrow \mathbb{N}$ ,  $S \subseteq \mathbb{N}^n$  for some  $n$ , containing the primitive recursive functions and also closed under the operation*

(iii) *Search: If  $g(y, x_1, \dots, x_n)$  is partial recursive then so is the function  $f$  such that  $f(x_1, \dots, x_n)$  is the least  $y$  such that  $g(y, x_1, \dots, x_n) = 0$ , if such  $y$  exists with  $\langle i, x_1, \dots, x_n \rangle \in \text{dom}(g)$  for every  $i \leq y$ .*

*A recursive function is a partial recursive function whose domain is all of  $\mathbb{N}^n$ .*

Again, it should be clear that every recursive function is computable. In particular, as long as the domain of  $f$  is all of  $\mathbb{N}^n$ , we can conduct a search for the least  $y$  as in (iii) being assured that the search will eventually halt. Partial recursive functions, on the other hand, are such that if  $\langle y, x_1, \dots, x_n \rangle \in \text{dom}(f)$  then  $f(y, x_1, \dots, x_n)$  can be computed; but we may not be able to determine in finite time whether  $\langle y, x_1, \dots, x_n \rangle \in \text{dom}(f)$ . The next fact can be taken either as a definition or a theorem. Again, we simply refer to [6] or [12] for details.

**Theorem 3.1.12.** *A set  $S \subseteq \mathbb{N}^n$  is computably enumerable iff  $S$  is the domain of a partial recursive function.*

### 3.1.4 The Pell equation

The *Pell equation*<sup>3</sup> refers to any equation of the form

$$X^2 - aY^2 = 1 \tag{3.3}$$

with  $a > 1$  a nonsquare integer (it can be readily ascertained that (3.3) has no non-trivial integer solutions for  $a$  square or  $\leq 1$ ). A solution is sought in nonzero natural numbers.

The solutions to the Pell equations are quite well understood, and a topic worth exploring. In particular, (3.3) always has a solution in positive integers, in fact infinitely many solutions. Furthermore, once the smallest positive solution has been found (the *fundamental solution*), from this it is fairly easy to generate the complete sequence of solutions.

In the following sections, we will always be choosing  $a$  such that the fundamental solution is immediately apparent. Nevertheless, we prove here the full existence theorem, although only Theorem 3.1.16 will be needed for the applications to Hilbert's Tenth Problem. The theorems are due to Lagrange, though algorithms for solving the Pell equation were known much earlier.<sup>4</sup> Many basic number theory books cover the topic at this level of detail (in particular, I learned the proof of Theorem 3.1.14 from [15]), but there is also an exhaustive but accessible reference in [16].

**Lemma 3.1.13.** *If  $a \in \mathbb{N}$ , then for some  $n \in \mathbb{N}$  the equation*

$$X^2 - aY^2 = n$$

*has infinitely many (integer) solutions.*

<sup>3</sup>It would fly in the face of tradition to mention this equation without also noting that John Pell (1611-1685), though an accomplished number theorist, had little or nothing to do with it. The name seems to be due to a mistaken attribution by Euler.

See [19] for an entertaining discussion of the history of the equation. This includes a reasonably natural combinatorial problem posed by Archimedes, on counting the cattle owned by the god Helios, which leads to an instance of the Pell equation. While Archimedes' cattle problem has a solution, amazingly, the smallest solution is a number of 206,545 digits! That is *a lot* of cattle.

<sup>4</sup>At least in the 7th century in India, and Diophantus in the 3rd century, for example, knew 3.1.15 given an initial solution.

*Proof.* We first show that the inequality

$$|X - \sqrt{a}Y| < 1/Y \quad (3.4)$$

has infinitely many solutions. Take  $k \in \mathbb{N}$  and for  $i \in \{0, 1, \dots, k\}$  consider the  $k+1$  'fractional parts'  $r_i = i\sqrt{a} - \lfloor i\sqrt{a} \rfloor$ . Because  $0 \leq r_i < 1$  for each  $i$ , two of them, say  $r_i$  and  $r_j$ , must have distance less than  $1/k$  from each other,  $|r_j - r_i| < 1/k$ . Assuming  $i < j$ , let  $x = \lfloor i\sqrt{a} \rfloor - \lfloor j\sqrt{a} \rfloor$  and  $y = j - i$ , so

$$|x - \sqrt{a}y| = |j\sqrt{a} - \lfloor j\sqrt{a} \rfloor - i\sqrt{a} + \lfloor i\sqrt{a} \rfloor| = |r_j - r_i| < \frac{1}{k} \leq \frac{1}{y}.$$

This gives one solution  $(x, y)$  to (3.4), but to produce a second one, we need only repeat the process with  $k$  replaced by a larger  $k'$  such that  $1/k' < |x - \sqrt{a}y|$ . As this process can be repeated arbitrarily often, the claim is proven.

Now, if  $(x, y)$  is a solution to (3.4) with  $y > 0$ , then

$$|x^2 - ay^2| = |x - \sqrt{a}y| |x + \sqrt{a}y| \leq \frac{|x + \sqrt{a}y|}{y} \leq \frac{|x - \sqrt{a}y| + 2\sqrt{a}y}{y} < 1 + 2\sqrt{a}$$

(since  $1/y^2 < 1$ ). Thus there are infinitely many pairs of integers  $(x, y)$  such that  $|x^2 - ay^2| < 1 + 2\sqrt{a}$ . Because there are only finitely many positive integers less than  $1 + 2\sqrt{a}$ , there is some such  $n < 1 + 2\sqrt{a}$  for which  $|X^2 - aY^2| = n$  has infinitely many solutions, as required.  $\square$

**Theorem 3.1.14.** *The Pell equation  $X^2 - aY^2 = 1$  with  $a > 1$  nonsquare has a non-trivial solution  $(x, y)$  (that is, with  $x$  and  $y$  both nonzero).*

*Proof.* Take  $n \in \mathbb{N}$  and an infinite family  $\{(x_i, y_i)\}_{i \in \mathbb{N}}$  of solutions to  $|X^2 - aY^2| = n$  as in the lemma. We may assume, further, that  $x_i, y_i > 0$  for each  $i$  and  $x_i \neq x_j$  for  $i \neq j$ . Note that  $n$  cannot be 0 if  $a$  is not square (since neither  $x_i - \sqrt{a}y_i$  nor  $x_i + \sqrt{a}y_i$  can = 0). If  $n = 1$ , we are already done. Thus assume  $n > 1$ .

By the pigeonhole principle, there are  $i, j \in \mathbb{N}$  such that  $x_i \equiv x_j \pmod{n}$  and  $y_i \equiv y_j \pmod{n}$ . Now consider  $(x_i - \sqrt{a}y_i)(x_j + \sqrt{a}y_j) = (x_i x_j - ay_i y_j) + \sqrt{a}(x_i y_j - x_j y_i)$ . We have  $x_i x_j - ay_i y_j \equiv x_i^2 - ay_i^2 \equiv 0 \pmod{n}$  and  $x_i y_j - x_j y_i \equiv x_i y_i - x_j y_i \equiv 0 \pmod{n}$ , so that

$$(x_i - \sqrt{a}y_i)(x_j + \sqrt{a}y_j) = n(x + \sqrt{a}y) \quad (3.5)$$

for some integers  $x, y$ .

We first note that  $y \neq 0$ . Otherwise,  $ny = x_i y_j - x_j y_i = 0$  and  $x_i^2 - ay_i^2 = x_j^2 - ay_j^2 = n$  imply

$$n = x_i^2 - a \left( \frac{x_i y_j}{x_j} \right)^2 = x_i^2 \left( \frac{x_j^2 - ay_j^2}{x_j^2} \right) = n \left( \frac{x_i^2}{x_j^2} \right).$$

Since  $x_i$  and  $x_j$  are both positive, however,  $x_i^2/x_j^2 = 1$  contradicts  $x_i \neq x_j$ .

In fact,  $(x, y)$  gives the desired solution to the Pell equation. It is easily verified that  $n(x - \sqrt{a}y) = (x_i + \sqrt{a}y_i)(x_j - \sqrt{a}y_j)$ , so multiplying (3.5) by  $n(x - \sqrt{a}y)$  gives

$$(x_i^2 - ay_i^2)(x_j^2 - ay_j^2) = n^2 = n^2(x^2 - ay^2).$$

Hence, dividing by  $n^2$  shows that  $(x, y)$  is a solution to the Pell equation, and  $y \neq 0$  guarantees that it is a nontrivial solution.  $\square$

The next step shows how, from a single nontrivial solution to the Pell equation, an infinite number of solutions can be generated.

**Corollary 3.1.15.** *The equation (3.3) has infinitely many solutions.*

*Proof.* Let  $(x_1, y_1)$  be a nontrivial solution to (3.3). We may assume that  $x_1, y_1 \in \mathbb{N}$ . Working in  $\mathbb{Z}[\sqrt{a}]$ , for each  $n \in \mathbb{N}$  define the pair  $(x_n, y_n) \in \mathbb{N}^2$  by

$$x_n + \sqrt{a}y_n = (x_1 + \sqrt{a}y_1)^n. \quad (3.6)$$

Because it follows that  $x_n - \sqrt{a}y_n = (x_1 - \sqrt{a}y_1)^n$  (by induction on  $n$ , for example),

$$\begin{aligned} x_n^2 - ay_n^2 &= (x_n + \sqrt{a}y_n)(x_n - \sqrt{a}y_n) \\ &= (x_1 + \sqrt{a}y_1)^n (x_1 - \sqrt{a}y_1)^n = (x_1^2 - ay_1^2)^n = 1 \end{aligned}$$

shows that  $(x_n, y_n)$  is also a solution to (3.3). Furthermore since  $x_1 + \sqrt{a}y_1 > 1$  by nontriviality of  $(x_1, y_1)$ , the sequence  $\{x_n + \sqrt{a}y_n\}_{i \geq 1}$  is strictly increasing, which implies that the solutions  $(x_n, y_n)$  are all distinct.  $\square$

Now let us fix in (3.6) the fundamental solution  $(x_1, y_1)$ , namely the solution in positive integers which is *minimal* in the sense of  $x_1 + \sqrt{a}y_1$ . Using the notation from Corollary 3.1.15 we show that the sequence  $(x_n, y_n)$  is in fact an exhaustive list of the positive solutions to the Pell equation. In general, the complete list of solutions is given by  $(x_0, y_0) = (1, 0)$  and for each  $n \in \mathbb{N}$

$$(\pm x_n, \pm y_n).$$

**Theorem 3.1.16.** *If  $(x, y)$  satisfies (3.3), then  $x = \pm x_n, y = \pm y_n$  for some  $n \in \mathbb{N}$ .*

*Proof.* Again it will suffice to assume that  $x$  and  $y$  are both positive integers. As above let  $(x_1, y_1)$  be the fundamental solution and  $(x_n, y_n)$  defined as in (3.6).

Since  $x_1 + \sqrt{a}y_1 > 1$ , it follows that  $\{x_n + \sqrt{a}y_n\}$  is unbounded and in particular  $n$  may be fixed so that

$$x_n + \sqrt{a}y_n \leq x + \sqrt{a}y < x_{n+1} + \sqrt{a}y_{n+1} = (x_n + \sqrt{a}y_n)(x_1 + \sqrt{a}y_1). \quad (3.7)$$

We wish to show that in fact  $x_n + \sqrt{a}y_n = x + \sqrt{a}y$ .

Using the fact that  $x_{n+1} + \sqrt{a}y_{n+1} = (x_n + \sqrt{a}y_n)(x_1 + \sqrt{a}y_1)$  and  $x_n - \sqrt{a}y_n = (x_n + \sqrt{a}y_n)^{-1}$ , multiplying (3.7) through by  $x_n - \sqrt{a}y_n$  gives

$$1 \leq (xx_n - ayy_n) + \sqrt{a}(x_ny - xy_n) < x_1 + \sqrt{a}y_1.$$

It can be directly verified that  $(\tilde{x}, \tilde{y}) = (xx_n - ayy_n, x_ny - xy_n)$  is yet another solution to (3.3). To finish it is enough to show that  $\tilde{x}, \tilde{y} \geq 0$ , since then minimality of  $(x_1, y_1)$  gives equality on the left, whence  $x + \sqrt{a}y = x_n + \sqrt{a}y_n$ .

To show  $\tilde{x}, \tilde{y} \geq 0$ , note first that  $\tilde{x}$  and  $\tilde{y}$  cannot both be negative since  $\tilde{x} + \sqrt{a}\tilde{y}$  is positive. If  $\tilde{x} < 0$ ,  $\tilde{y} \geq 0$ , then  $(\tilde{x} + \sqrt{a}\tilde{y})^{-1} = \tilde{x} - \sqrt{a}\tilde{y} < 0$  contradicts  $\tilde{x} + \sqrt{a}\tilde{y} \geq 1$ . If on the other hand  $\tilde{x} \geq 0$ ,  $\tilde{y} < 0$ , then

$$(\tilde{x} + \sqrt{a}\tilde{y})^{-1} = \tilde{x} - \sqrt{a}\tilde{y} > \tilde{x} + \sqrt{a}\tilde{y} \geq 1$$

which is also impossible.  $\square$

## 3.2 The Davis Normal Form

The next step is the characterization of the computably enumerable sets as those defined (in  $\mathbb{N}$ ) by formulas of the form

$$\exists z \forall w \leq z \exists y_1, \dots, y_m P(x, \tilde{y}, w, z) = Q(x, \tilde{y}, w, z)$$

with  $P, Q$  polynomials over  $\mathbb{N}$ . This result was part of Davis' 1950 dissertation and is the main focus of the article [5]. The main idea is the careful use of coding to do many things at once: to encode a computation by primitive recursion, for example, or to abbreviate a string of quantifiers into just one.

### 3.2.1 Bounded formulas

**Definition 3.2.1.** A bounded formula is a first-order formula containing no negations (that is, using only the connectives  $\wedge$  and  $\vee$ ) and in which all universal quantifiers appear in the bounded form  $\forall w \leq z$ .

We work in the language of rings, in  $\mathbb{N}$ . Recall also the function  $\text{rem}$  from Proposition 3.1.5. As diophantine formulas are (trivially) bounded, diophantine functions like  $\text{rem}$  may be used freely in bounded formulas.

**Proposition 3.2.2.** If  $f$  is a primitive recursive function, then (the graph of)  $f$  is defined by a bounded formula.

*Proof.* The proof is by induction on the construction of  $f$ . It is obvious that the constant functions, the projections, and the successor function are defined by bounded formulas (in fact by atomic formulas). The composition of functions defined by bounded formulas can clearly also be defined by a bounded formula.

The only interesting thing here is to show that if  $f(y, x_1, \dots, x_n)$  is constructed from  $h(x_1, \dots, x_n)$  and  $g(y, z, x_1, \dots, x_n)$  by primitive recursion, and  $h$  and  $g$  are defined by bounded formulas, then so is  $f$ . The strategy is to use the Chinese Remainder Theorem to encode the sequence  $f(0, \bar{x}), f(1, \bar{x}), \dots, f(y, \bar{x})$ . A bounded universal quantifier will work to ensure that each successive  $f(i+1, \bar{x})$  is properly obtained from  $f(i, \bar{x})$  and  $g$ .

*Claim:*  $f(y, \bar{x}) = z$  iff

$$\exists N \exists a \forall i \leq y \left\{ \begin{array}{l} \text{rem}(a, N+1) = h(\bar{x}) \wedge \quad (3.8) \\ \left( \begin{array}{l} i = y \vee \\ \text{rem}(a, (i+2)N+1) = g(i, \text{rem}(a, (i+1)N+1), \bar{x}) \end{array} \right) \wedge \quad (3.9) \\ \text{rem}(a, (y+1)N+1) = z \quad (3.10) \end{array} \right.$$

There is a bit of a technical annoyance in the above formula because we want to code the sequence  $f(0, \bar{x}), \dots, f(y, \bar{x})$  by  $a$  modulo the sequence  $N + 1, \dots, (y + 1)N + 1$ , hence all the adjustments to the index  $i$ .

So, first suppose that  $f(y, \bar{x}) = z$ , and let  $N$  be some multiple of  $y!$  bigger than  $f(i, \bar{x})$  for every  $i \leq y$ . By Proposition 3.1.9, the numbers  $N + 1, \dots, (y + 1)N + 1$  are pairwise relatively prime. By the Chinese Remainder Theorem, there is  $a$  such that

$$a \equiv f(i, \bar{x}) \pmod{(i + 1)N + 1}$$

for every  $i \leq y$ , and since  $f(i, \bar{x}) < (i + 1)N + 1$ ,  $\text{rem}(a, (i + 1)N + 1) = f(i, \bar{x})$ . Now

- the case  $i = 0$  gives us (3.8) (using  $f(0, \bar{x}) = h(\bar{x})$ );
- the case  $i = y$  gives (3.10);
- and an induction on  $i \leq y$  gives (3.9) (using  $f(i + 1, \bar{x}) = g(i, f(i, \bar{x}), \bar{x})$ ).

For the converse, suppose that the above formula holds. We show by induction on  $i \leq y$  that  $f(i, \bar{x}) = \text{rem}(a, (i + 1)N + 1)$ . In light of (3.10), this will suffice to prove  $f(y, \bar{x}) = z$  as required. By the primitive recursion construction, (3.8) implies that  $f(0, \bar{x}) = \text{rem}(a, N + 1)$ . Now supposing that  $f(i, \bar{x}) = \text{rem}((i + 1)N + 1)$  and  $i < y$ , from (3.9) we obtain

$$f(i + 1, \bar{x}) = g(i, f(i, \bar{x}), \bar{x}) = \text{rem}(a, (i + 2)N + 1)$$

to complete the induction. Now (3.10) implies  $f(y, \bar{x}) = \text{rem}(a, (y + 1)N + 1) = z$ , as required.

With this claim established, after substituting in the bounded formulas defining  $g$  and  $h$ , the observation that the given formula is itself a bounded formula completes the proof.  $\square$

It is not hard to see that the above result can be extended to all partial recursive functions.

**Corollary 3.2.3.** *Every partial recursive function is defined by a bounded formula.*

*Proof.* It only remains to show that if  $f(x_1, \dots, x_n)$  is obtained from  $g(y, x_1, \dots, x_n)$  by the search operation (i.e.  $f(\bar{x})$  is the least  $y$  such that  $g(y, \bar{x}) = 0$  and  $g(i, \bar{x})$  is defined for all  $i \leq y$ , if such  $y$  exists) and  $g$  is defined by a bounded formula, then so is  $f$ . But this is clear, since  $f(\bar{x}) = z$  iff

$$g(z, \bar{x}) = 0 \wedge \forall y \leq z \exists w (g(y, \bar{x}) = w \wedge w > 0).$$

Recalling again from 3.1.5 that the relation  $w > 0$  is diophantine, this is equivalent to a bounded formula.  $\square$

The main point of this section, the equivalence of computable enumerability and definability by bounded formulas, is now readily accessible. We use 3.1.12 to identify computably enumerable sets with domains of partial recursive functions.

**Proposition 3.2.4.** *The computably enumerable sets are precisely those sets defined by a bounded formula.*

*Proof.* To see that every bounded formula defines a computably enumerable set, we simply invoke Church's Thesis. As an algorithm, in the informal sense, to enumerate every  $n$ -tuple  $\langle a_1, \dots, a_n \rangle$  satisfying a bounded formula  $\varphi(x_1, \dots, x_n)$  is evident, this suffices. The bounding of the universal quantifiers is the key which allows us to get a confirmation of  $\varphi(a_1, \dots, a_n)$ , if true, in a finite number of steps.

On the other hand, if  $S \subseteq \mathbb{N}^n$  is the domain of a partial recursive function  $f(\bar{x})$ , and  $f(\bar{x}) = y$  is defined by the bounded formula  $\varphi(\bar{x}, y)$ , then it is clear that  $S$  is defined by the formula  $\exists y \varphi(\bar{x}, y)$ . Since the existential quantification of a bounded formula is still bounded,  $S$  is indeed defined by a bounded formula.  $\square$

### 3.2.2 Pruning quantifiers in bounded formulas

Now that we are happy with the status of bounded formulas, the attention turns to rearranging and culling quantifiers from bounded formulas to produce a formula in Davis Normal Form (DNF),

$$\exists z \forall w \leq z \varphi(x, w, z)$$

with  $\varphi$  diophantine. (Note that this is the same as the definition given at the beginning of the section.) The objective is to show that every bounded formula is equivalent in  $\mathbb{N}$  to a DNF formula.

**Lemma 3.2.5.** *The formula*

$$\exists z_1 \dots \exists z_n \forall w_1 \leq z_1 \dots \forall w_n \leq z_n \varphi(\bar{x}, \bar{w}, \bar{z}) \quad (3.11)$$

*with  $\varphi$  diophantine is equivalent to one in Davis Normal Form.*

*Proof.* Using the Cantor coding of 3.1.2 to code the sequences  $\bar{w}$  and  $\bar{z}$ , we show that the formula

$$\exists z \forall w \leq z \exists w_1, \dots, w_n, z_1, \dots, z_n \begin{cases} C_n(w_1, \dots, w_n) = w \wedge C_n(z_1, \dots, z_n) = z \wedge \\ (\varphi(\bar{x}, \bar{w}, \bar{z}) \vee \bigvee_{i \leq n} w_i > z_i) \end{cases} \quad (3.12)$$

is equivalent to (3.11).

If (3.11) holds, then let  $z = C_n(z_1, \dots, z_n)$ . For all  $w \leq z$ , as long as  $D_{i,n}(w) = w_i \leq z_i$  for each  $i$ , (3.11) gives  $\varphi(\bar{x}, \bar{w}, \bar{z})$ . Thus (3.12) holds as well.

Conversely, assume (3.12). If  $w_i \leq z_i$  for each  $i$ , then by Proposition 3.1.7,

$$w = C_n(w_1, \dots, w_n) \leq C_n(z_1, \dots, z_n) = z.$$

Now the failure of  $\bigvee_{i \leq n} w_i > z_i$  in (3.12) implies  $\varphi(\bar{x}, \bar{w}, \bar{z})$ , as required.

Finally, we simply note that since  $C_n$  is defined by a diophantine formula,  $\varphi$  and  $x < y$  are diophantine, and conjunctions and disjunctions of diophantine formulas remain diophantine, the formula (3.12) is in Davis Normal Form.  $\square$

The previous lemma shows that we can collapse long sequences of bounded universal quantifiers. The next lemma shows how to rearrange existential and bounded universal quantifiers to eliminate alternations.

**Lemma 3.2.6.** *The formula*

$$\exists z \forall w \leq z \exists s \forall r \leq s \varphi(\bar{x}, r, s, w, z) \quad (3.13)$$

with  $\varphi$  diophantine is equivalent to one in Davis Normal Form.

*Proof.* While the Cantor coding worked in 3.2.5 since the lengths of the sequences are fixed, this time the idea is to use the Chinese Remainder Theorem to code the sequence  $s_0, \dots, s_z$  of  $s$ 's corresponding to each  $w \leq z$ . The formula

$$\exists N \exists a \exists z \forall r \leq a \forall w \leq z \left\{ \begin{array}{l} w \mid N \wedge a > (z+1)N + 1 \wedge \\ (\varphi(\bar{x}, r, s_w, w, z) \vee r > s_w) \end{array} \right. \quad (3.14)$$

(here  $s_w$  is merely an abbreviation for  $\text{rem}(a, (w+1)N + 1)$ ) is equivalent to (3.13).

To see this, assuming first (3.13), for each  $w \leq z$  there is some  $s_w$  such that  $\forall r \leq s_w \varphi(\bar{x}, r, s_w, w, z)$  holds. Let  $N = z!$ , so that  $N+1, \dots, (z+1)N+1$  are (pairwise) relatively prime by 3.1.9. Then the Chinese Remainder Theorem gives us an integer  $a > (z+1)N + 1$  satisfying for each  $w \leq z$

$$\text{rem}(a, (w+1)N + 1) = s_w.$$

For this choice of  $N$  and  $a$ , we thus obtain (3.14).

If, on the other hand, (3.14) holds, then for each  $w \leq z$ , observe that

$$r \leq s_w = \text{rem}(a, (w+1)N + 1) < a$$

(the latter inequality following by definition of  $\text{rem}$ ) implies  $\varphi(\bar{x}, r, s_w, w, z)$ . This gives (3.13).

It remains to show that (3.14) is in DNF. But this follows from Lemma 3.2.5. Again, since the function  $\text{rem}$  is defined by a diophantine formula, and the diophantine formulas are closed under conjunctions and disjunctions, the formula following the bounded universal quantifiers in (3.14) is diophantine. In order to get (3.14) in the exact form of the formula (3.11), one may simply insert a dummy quantifier  $\forall M \leq N$  before the  $\forall r \leq a$ . Since  $M$  does not appear in the ensuing formula, this does not affect the truth of the formula.  $\square$

These lemmas show that bounded universal quantifiers are surprisingly pliable. Unlike the case with unbounded quantifiers, alternations can be eliminated. The two lemmas combine to give

**Proposition 3.2.7.** *Every bounded formula is equivalent to one in Davis Normal Form.*

*Proof.* Let  $\varphi(\bar{x})$  be a bounded formula. By basic syntactic manipulations, we may assume that  $\varphi$  is in prenex normal form

$$Q_1 y_1 Q_2 y_2 \dots Q_n y_n \psi(\bar{x}, \bar{y})$$

with  $\psi$  quantifier free and each  $Q_i$  either an existential quantifier or a bounded universal quantifier. By again inserting dummy quantifiers where needed, in fact it may be assumed that  $\varphi$  is in the form

$$\exists z_1 \forall w_1 \leq z_1 \exists z_2 \forall w_2 \leq z_2 \dots \exists z_n \forall w_n \leq z_n \psi(\bar{x}, \bar{w}, \bar{z}) \quad (3.15)$$

with  $\psi$  diophantine.

We now see by induction on  $n$  that  $\varphi$  is equivalent to a DNF formula. If  $n = 0$  or  $n = 1$ , there is nothing to show. If  $n \geq 2$ , then the subformula

$$\exists z_{n-1} \forall w_{n-1} \leq z_{n-1} \exists w_n \forall w_n \leq z_n \psi(\bar{x}, \bar{w}, \bar{z})$$

is equivalent to a new formula

$$\exists z_{n-1} \forall w_{n-1} \leq z_{n-1} \tilde{\psi}(\bar{x}, \bar{w}, \bar{z})$$

with  $\tilde{\psi}$  diophantine, by 3.2.6. Thus (3.15) is equivalent to

$$\exists z_1 \forall w_1 \leq z_1 \dots \exists z_{n-1} \forall w_{n-1} \leq z_{n-1} \tilde{\psi}(\bar{x}, \bar{w}, \bar{z})$$

and the induction gives that this in turn is equivalent to a DNF formula.  $\square$

Finally, by combining Proposition 3.2.7 with 3.2.4, we immediately obtain Davis' main result:

**Theorem 3.2.8.** *The computably enumerable sets are precisely those sets defined by a formula in Davis Normal Form.*

---

# Bibliography

- [1] J. Ax and S. Kochen, *Diophantine problems over local fields. III. Decidable fields*, Ann. of Math. (2) **83** (1966), 437–456.
- [2] J. W. S. Cassels, *Rational quadratic forms*, London Mathematical Society Monographs, vol. 13, Academic Press Inc., London, 1978.
- [3] P. J. Cohen, *Decision procedures for real and  $p$ -adic fields*, Comm. Pure Appl. Math. **22** (1969), 131–151.
- [4] J. H. Conway and D. A. Smith, *On quaternions and octonions: their geometry, arithmetic, and symmetry*, A K Peters Ltd., Natick, MA, 2003.
- [5] M. Davis, *Arithmetical problems and recursively enumerable predicates*, J. Symbolic Logic **18** (1953), 33–41.
- [6] ———, *Computability and unsolvability*, McGraw-Hill Series in Information Processing and Computers, McGraw-Hill Book Co., Inc., New York, 1958.
- [7] ———, *An explicit diophantine definition of the exponential function*, Comm. Pure Appl. Math. **24** (1971), 137–145.
- [8] M. Davis, H. Putnam, and J. Robinson, *The decision problem for exponential diophantine equations*, Ann. of Math. (2) **74** (1961), 425–436.
- [9] A. J. Engler and A. Prestel, *Valued fields*, Springer Monographs in Mathematics, Springer-Verlag, Berlin, 2005.
- [10] K. Gödel, *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme. I*, Monatsh. Math. **38** (1931), 173–198.
- [11] H. Hasse, *Über die Darstellbarkeit von Zahlen durch quadratische Formen im Körper der rationalen Zahlen*, J. Reine Angew. Math. **152** (1923), 129–148.
- [12] H. Hermes, *Enumerability, decidability, computability. An introduction to the theory of recursive functions*, translated from the German by G. T. Hermann and O. Plassmann. Second revised edition. Die Grundlehren der mathematischen Wissenschaften, Band 127, Springer-Verlag New York, Inc., New York, 1969.
- [13] D. Hilbert, *Mathematical problems*, Bull. Amer. Math. Soc. **8** (1902), no. 10, 437–479.
- [14] A. Hurwitz, *Über die Zahlentheorie der Quaternionen*, Nachr. Königl. Ges. Göttingen Math.-Phys. Kl. **4** (1896), 313–340.

- 
- [15] K. Ireland and M. Rosen, *A classical introduction to modern number theory*, 2nd ed., Graduate Texts in Mathematics, vol. 84, Springer-Verlag, New York, 1990.
- [16] M. J. Jacobson Jr. and H. C. Williams, *Solving the Pell equation*, CMS Books in Mathematics, Springer, New York, 2009.
- [17] N. Koblitz,  *$p$ -adic numbers,  $p$ -adic analysis, and zeta-functions*, 2nd ed., Graduate Texts in Mathematics, vol. 58, Springer-Verlag, New York, 1984.
- [18] T. Y. Lam, *Introduction to quadratic forms over fields*, Graduate Studies in Mathematics, vol. 67, American Mathematical Society, Providence, RI, 2005.
- [19] H. W. Lenstra Jr., *Solving the Pell equation*, Notices Amer. Math. Soc. **49** (2002), no. 2, 182–192.
- [20] Y. V. Matiyasevich, *Enumerable sets are diophantine*, Soviet Math. Dokl. **11** (1970), no. 2, 354–357 (Translation from Russian).
- [21] ———, *Hilbert's Tenth Problem*, Foundations of Computing Series, MIT Press, Cambridge, MA, 1993. Translated from the 1993 Russian original by the author; with a foreword by Martin Davis.
- [22] I. Niven, H. S. Zuckerman, and H. L. Montgomery, *An introduction to the theory of numbers*, 5th ed., John Wiley & Sons Inc., New York, 1991.
- [23] A. Ostrowski, *Über einige Lösungen der Funktionalgleichung  $\varphi(x) \cdot \varphi(y) = \varphi(xy)$* , Acta Math. **41** (1918), no. 1, 271–284.
- [24] J. Robinson, *Definability and decision problems in arithmetic*, J. Symbolic Logic **14** (1949), 98–114.
- [25] ———, *Existential definability in arithmetic*, Trans. Amer. Math. Soc. **72** (1952), 437–449.
- [26] W. Szmielew, *Elementary properties of Abelian groups*, Fund. Math. **41** (1955), 203–271.
- [27] A. Tarski, *A Decision Method for Elementary Algebra and Geometry*, RAND Corporation, Santa Monica, Calif., 1948.