The Algorithmic Solution of Diophantine Equations



Author:
Mahadi Ddamulira

Supervisor:
Prof. Fernando Rodriguez Villegas

The Abdus Salam International Centre for Theoretical Physics
Trieste, Italy

A thesis submitted in partial fulfilment of the requirements for the award of the Postgraduate Diploma in Mathematics

August 2016

Dedication

To my beloved wife, Nagawa Nusha.

Abstract

In this research project, we study some of the local methods which allow us to either completely solve a diophantine equation, aid us in locating the solutions or give us information about the solutions which can be used in more advanced methods.

Key words: p-adic numbers, p-adic numerical analysis, p-adic power series, Hensel's lemma, diophantine equations, Thue equations, Strassmann's theorem, Skolem's method.

Acknowledgements

I would like to thank the Almighty Allah for guiding and giving me natural endowment, skills, healthy mind and a strong body to work on this research project. I would also like to thank my supervisor, Prof. Fernando Rodriguez-Villegas for his help and encouragement throughout this project. Heart felt gratitude to all the members of the Mathematics Section of ICTP for their support during the Postgraduate Diploma programme. Many thanks to Patrizia and Sandra for the continued administrative assistance.

I express my deep gratitude to my family members my beloved wife, Nagawa Nusha, my parents, siblings most especially: Abbey Buyondo and Twahiru Muwoomya for their unconditional love, patience and support in every step of my life. Special thanks to Prof. Florian Luca for his continued encouragement and support through introducing me to independent research in Mathematics at AIMS - Ghana. I am also thankful to Kenneth Muhumuza my Ugandan colleague at ICTP and room-mate for his assistance which he rendered whenever I needed him.

Last but not the least, I acknowledge the financial support from UNESCO, IAEA and the Italian Government which enabled me to come to ICTP and spend full year learning interesting topics and results in mathematics.

Contents

Dedication **ii**

Abstract **iii**

Acknowledgements **iv**

1 Introduction to Local Methods **1**

1.1 The p-adic norm and p-adic numbers 1

1.1.1 Arithmetic in $\mathbb{Q}_p$ 6

1.2 p-adic numerical analysis 10

1.2.1 Newton-Raphson method 10

1.2.2 Power series in one variable 13

1.2.3 Power series in many variables 18

1.2.4 The Iwasawa logarithm 20

1.2.5 p-adic exponential function 24

2 Applications of Local Methods to Diophantine Equations **25**

2.1 Some useful preliminary results 25

2.2 Applications of Strassmann's theorem 27

2.3 Skolem's method 30

2.4 Hasse's principle 34

2.5 Finding small solutions 35

Bibliography **39**

CHAPTER 1

Introduction to Local Methods

In this chapter we give a brief overview of p-adic numbers and various methods from what could be called 'p-adic numerical analysis'. We shall discuss, in particular, the use of p-adic numbers in an elementary way (that is, congruences modulo powers of p) and in a less elementary way (that is, Hensel's lemma, Strassmann's theorem, p-adic power series and Skolem's method). We follow the ideas from [1], [6] and [8].

1.1 The p-adic norm and p-adic numbers

In this section we construct the field of p-adic numbers, $\mathbb{Q}_p$ and state and prove the necessary results needed in order to solve certain Diophantine equations. To do this, we introduce the notion of the p-adic ordinal, ord_p and define the p-adic norm, $|\cdot|_p$ from the p-adic ordinal. This thus allows us to define the field $\mathbb{Q}_p$ as the completion of the field of rational numbers, $\mathbb{Q}$ with respect to the norm $|\cdot|_p$.

Definition 1.1.1. Let R be a ring with unity $1 = 1_R$. A function

$$||\cdot|| : R \longrightarrow \mathbb{R}^+ = \{r \in \mathbb{R} : r \geq 0\}$$

is called a *norm* on R if the following are true:

(1) $||x|| = 0$ if and only if $x = 0$.

(2) $||x \cdot y|| = ||x|| \cdot ||y||$ for all $x, y \in R$.

(3) $||x + y|| \leq ||x|| + ||y||$ for all $x, y \in R$.

Condition (3) is called *the triangle inequality*. A norm is called *non-Archimedean* if condition (3) is replaced by the stronger statement, *the ultrametric inequality*:

(3′) $||x+y|| \leq \max\{||x||, ||y||\}$ for all $x, y \in R$.

otherwise the norm is *Archimedean*.

Example 1.1.1. Let $R \subseteq \mathbb{C}$ be a subring of the complex numbers, $\mathbb{C}$. Then setting $||x|| = |x|$, the usual absolute value, gives a norm on R. In particular, this applies to the cases $R = \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$. This norm is Archimedean because of the inequality:

$$|1+3| = 4 > 3 = |3| = \max\{|1|, |3|\}.$$

We now consider the case of $R = \mathbb{Q}$, the ring of rational numbers a/b, where $a, b \in \mathbb{Z}$ and $b \neq 0$. Let $p \in \{2, 3, 5, 7, 11, 13, \ldots\}$ be a prime number. Every nonzero rational number, x, can be written uniquely in the form

$$x = p^r \frac{a}{b},$$

where $a, b \in \mathbb{Z}$ with $p \nmid a, b$ and $r \in \mathbb{Z}$.

Definition 1.1.2. If x is a nonzero integer, the *p-adic ordinal* of x is

$$\mathrm{ord}_p x = \max\{r : p^r | x\} \geq 0. \tag{1.1}$$

For $a/b \in \mathbb{Q}$, the *p-adic ordinal* of a/b is given by:

$$\mathrm{ord}_p \frac{a}{b} = \mathrm{ord}_p a - \mathrm{ord}_p b. \tag{1.2}$$

For $x = 0$, we agree as a convention to write $\mathrm{ord}_p 0 = \infty$. We notice that in all cases, ord_p gives an integer and that for a rational number a/b, $\mathrm{ord}_p a/b$ is well defined, that is, if $a/b = c/d$ then

$$\mathrm{ord}_p \frac{a}{b} = \mathrm{ord}_p a - \mathrm{ord}_p b = \mathrm{ord}_p c - \mathrm{ord}_p d = \mathrm{ord}_p \frac{c}{d}.$$

Proposition 1.1.1. [6] If $x, y \in \mathbb{Q}$, the ord_p satisfies the following properties:

(1) $\mathrm{ord}_p x = \infty$ if and only if $x = 0$.

(2) $\mathrm{ord}_p(xy) = \mathrm{ord}_p x + \mathrm{ord}_p y$.

(3) $\mathrm{ord}_p(x+y) \geq \min\{\mathrm{ord}_p x, \mathrm{ord}_p y\}$ with equality if $\mathrm{ord}_p x \neq \mathrm{ord}_p y$.

Proof. (1) is trivial as it follows directly from our convention; we, therefore, prove (2) and (3). Let x, y be nonzero rational numbers. Write $x = p^r \frac{a}{b}$ and $y = p^s \frac{c}{d}$, where p is a prime number, $a, b, c, d \in \mathbb{Z}$ with $p \nmid a, b, c, d$ and $r, s \in \mathbb{Z}$. Then:
For (2), $xy = p^r \frac{a}{b} p^s \frac{c}{d} = p^{r+s} \frac{ac}{bd}$, which gives $\mathrm{ord}_p(xy) = r + s = \mathrm{ord}_p x + \mathrm{ord}_p y$, since $p \nmid ac, bd$.
For (3), suppose that $r > s$. Then we have,

$$x + y = p^r \frac{a}{b} + p^s \frac{c}{d} = p^s \left(p^{r-s} \frac{a}{b} + \frac{c}{d} \right) = p^s \left(\frac{p^{r-s} ad + bc}{bd} \right).$$

Similary, if $r < s$,

$$x + y = p^r \frac{a}{b} + p^s \frac{c}{d} = p^r \left(\frac{a}{b} + p^{s-r} \frac{c}{d} \right) = p^r \left(\frac{ad + p^{s-r} bc}{bd} \right).$$

If $r = s$, then we get

$$x + y = p^r \frac{a}{b} + p^r \frac{c}{d} = p^r \left(\frac{a}{b} + \frac{c}{d} \right) = p^r \left(\frac{ad + bc}{bd} \right).$$

Combining these results, since $p \nmid bd$, we get $\mathrm{ord}_p(x + y) \geq \min\{\mathrm{ord}_p x, \mathrm{ord}_p y\}$. □

Definition 1.1.3. For $x \in \mathbb{Q}$, *the p-adic norm* of x is given by

$$|x|_p = \begin{cases} p^{-\mathrm{ord}_p x} & \text{if } x \neq 0 \\ p^{-\infty} = 0 & \text{if } x = 0. \end{cases} \tag{1.3}$$

Proposition 1.1.2. [6] The function $|\cdot|_p : \mathbb{Q} \longrightarrow \mathbb{R}^+$ satisfies the following properties:

(1) $|x|_p = 0$ if and only if $x = 0$;

(2) $|x \cdot y|_p = |x|_p \cdot |y|_p$;

(3) $|x + y|_p \leq \max\{|x|_p, |y|_p\}$, with equality if $|x|_p \neq |y|_p$.

Hence, $|\cdot|_p$ is a *non-Archimedean* norm on $\mathbb{Q}$.

Proof. The proof follows easily from Proposition 1.1.1, as follows. (1) is trivial as it follows directly by definition of $|\cdot|_p$. Thus we prove (2) and (3). Suppose $x, y \in \mathbb{Q}$. Then we may write $x = p^r \frac{a}{b}$ and $y = p^s \frac{c}{d}$, where $a, b, c, d, r, s \in \mathbb{Z}$ with p a prime such that $p \nmid a, b, c, d$.
For (2), we have

$$xy = p^{r+s} \frac{ac}{bd}, \quad \text{with } p \nmid ac, bd \ \ (p\text{-prime}).$$

Therefore, we have

$$|xy|_p \;=\; p^{-(r+s)} \;=\; p^{-r}p^{-s} \;=\; |x|_p|y|_p.$$

For (3), if $x = 0$ or $y = 0$, or if $x + y = 0$, then the property is trivial. Therefore we assume that x, y and $x+y$ are all non-zero. We also assume without loss of generality that $r \leq s$, then

$$|x+y|_p \;=\; \left|\frac{p^r(ad+p^{s-r}bc)}{bd}\right|_p,$$

and this must be less than $|x|_p$ because $(ad + p^{s-r}bc)$ factors as $p^j i$ for some $j \in \mathbb{N}$ and $i \in \mathbb{Z}$. Then $p^{j+r} \parallel \frac{p^r(ad+p^{s-r}bc)}{bd}$, which implies that

$$|x|_p \;=\; p^{-r} \;\geq\; p^{-(j+r)} \;=\; |x+y|_p,$$

so that

$$|x+y|_p \;\leq\; \max\{|x|_p, |y|_p\} \;\leq\; |x|_p + |y|_p.$$

This completes the proof. □

Let p be a prime number and $n \geq 1$. Then from the p-adic expansion

$$n \;=\; n_0 + n_1 p + n_2 p^2 + \cdots + n_\ell p^\ell, \tag{1.4}$$

with $0 \leq n_i \leq p-1$, we define the number

$$\alpha_p(n) \;=\; n_0 + n_1 + n_2 + \cdots + n_\ell. \tag{1.5}$$

In the following example we prove a result that will be useful in determining the radius of convergence of p-adic exponential function, which we shall discuss in the last section of this chapter.

Example 1.1.2. [1] The p-adic ordinal of $n!$ is given by

$$\mathrm{ord}_p(n!) \;=\; \frac{n - \alpha_p(n)}{p-1}. \tag{1.6}$$

Therefore, the p-adic norm of $n!$ is given by

$$|n!|_p \;=\; p^{-(n-\alpha_p(n))/(p-1)}. \tag{1.7}$$

Proof. We prove this result by induction. The claim is true for $n = 1$. Now let $n > 1$ be an integer and assume that

$$\mathrm{ord}_p((n-1)!) \;=\; \frac{n-1-\alpha_p(n-1)}{p-1}.$$

We write n as $n = n_0 + n_1 p + n_2 p^2 + \cdots + n_\ell p^\ell$. If $p \nmid n$, then $n_0 > 0$, hence $n-1 = (n_0-1)+n_1 p+n_2 p^2+\cdots+n_\ell p^\ell$. Therefore, using the fact that $n! = n(n-1)!$ and that $p \nmid n$, we have

$$\mathrm{ord}_p(n!) \;=\; \mathrm{ord}_p(n) + \mathrm{ord}_p((n-1)!) \;=\; \mathrm{ord}_p((n-1)!).$$

Hence, we have

$$\begin{aligned}\mathrm{ord}_p(n!) \;&=\; \frac{n-1-\alpha_p(n-1)}{p-1} \;=\; \frac{n-1-((n_0-1)+n_1 p+n_2 p^2+\cdots+n_\ell p^\ell)}{p-1}\\ &=\; \frac{n-(n_0+n_1 p+n_2 p^2+\cdots+n_\ell p^\ell)}{p-1} \;=\; \frac{n-\alpha_p(n)}{p-1}.\end{aligned}$$

If $p^k \mid n$ for some $k \geq 1$, then $n = n_k p^k + \cdots + n_\ell p^\ell$, where $n_k \geq 1$. It follows that

$$\begin{aligned}n-1 \;&=\; (n_k p^k+\cdots+n_\ell p^\ell)-1 \;=\; p^k-1+(n_k-1)p^k+\cdots+n_\ell p^\ell\\ &=\; (p-1)(1+p+p^2+\cdots+p^{k-1})+(n_k-1)p^k+\cdots+n_\ell p^\ell.\end{aligned}$$

Hence, we have

$$\mathrm{ord}_p((n-1)!) = \frac{n-1-\alpha_p(n-1)}{p-1} = \frac{n-1-(k(p-1)+n_k-1+n_{k+1}+\cdots+n_\ell)}{p-1}$$

Since $\mathrm{ord}_p(n!) = \mathrm{ord}_p(n) + \mathrm{ord}_p((n-1)!) = k + \mathrm{ord}_p((n-1)!)$, we have

$$\begin{aligned}\mathrm{ord}_p(n!) \;&=\; k+\frac{n-1-(k(p-1)+n_k-1+n_{k+1}+\cdots+n_\ell)}{p-1}\\ &=\; \frac{n-1-(n_k-1+n_{k+1}+\cdots+n_\ell)}{p-1} = \frac{n-(n_k+n_{k+1}+\cdots+n_\ell)}{p-1}\\ &=\; \frac{n-\alpha_p(n)}{p-1}.\end{aligned}$$

This completes the proof. □

Two norms are defined to be equivalent if they induce the same topology, that is, they define essentially the same metric. It turns out that all non-trivial metrics on the rational numbers, $\mathbb{Q}$, are equivalent either to the standard metric or to the p-adic metric for some prime p. For the rest of this research project, we shall ignore the existence of the trivial metric on $\mathbb{Q}$.

Usually one completes the rational numbers, $\mathbb{Q}$, to form the real numbers, $\mathbb{R}$, using the standard absolute value by forming the set of all Cauchy sequences with the same limit, with convergence being measured in the sense of the absolute value. Using the p-adic norm the same construction can be carried out. But now instead of ending up with the field of real numbers we end up with the field of p-adic numbers, $\mathbb{Q}_p$.
We can think of a p-adic number as a formal base p expansion which encodes properties modulo higher and higher powers of p. Every nonzero p-adic number, n, can be written in the form

$$n = p^{\alpha}\left(n_0 + n_1 p + n_2 p^2 + \cdots\right) = p^{\alpha}\left(\sum_{i=0}^{\infty} n_i p^i\right), \tag{1.8}$$

where $n_i \in \{0, \ldots, p-1\}$ and $(p, n_0) = 1$, we define $\mathrm{ord}_p = \alpha$ and $|n|_p = p^{-\alpha}$.

Definition 1.1.4. The p-adic integers, $\mathbb{Z}_p$, are defined to be the elements $n \in \mathbb{Q}_p$ with $\alpha = \mathrm{ord}_p n \geq 0$ or equivalently, with $|n|_p = p^{-\alpha} \leq 1$. It is clear that $\mathbb{Z}_p$ contains a copy of $\mathbb{N}$, as elements of $\mathbb{N}$ are just elements of $\mathbb{Z}_p$ for which all but finitely many of the coefficients, n_i, are zero.

It is obvious that one can not hold a p-adic number to infinite precision within a computer's memory, just as one can not hold a real number to infinite precision. It is usual to work to a given accuracy, so we hold a p-adic number as a triple, (α, β, γ), where $\alpha, \beta \in \mathbb{Z}$, $\gamma \in \mathbb{Z} \cup \{\infty\}$ with $\alpha \leq \gamma$ and $0 \leq \beta < p^{\gamma-\alpha}$, where either $\beta = 0$ or $\gcd(p, \beta) = 1$. Such a tipple gives a representation of a p-adic number n up to the p^{γ} digit as follows:

$$n = p^{\alpha}\beta + O(p^{\gamma}). \tag{1.9}$$

1.1.1 Arithmetic in $\mathbb{Q}_p$

We define the basic arithmetic operations on p-adic numbers in a completely elementary and algorithmic way as follows:

For addition and subtraction. Let $z = x \pm y$; then we have

$$\begin{aligned} \gamma_z &= \min\{\gamma_x, \gamma_y\}, \\ \alpha_z &= \mathrm{ord}_p\left(p^{\alpha_x}\beta_x \pm p^{\alpha_y}\beta_y (\mathrm{mod}\ p^{\gamma_z})\right), \\ \beta_z &= p^{-\alpha_x}\left(p^{\alpha_x}\beta_x \pm p^{\alpha_y}\beta_y (\mathrm{mod}\ p^{\gamma_z})\right). \end{aligned} \tag{1.10}$$

For multiplication. Let $z = xy$; then we have

$$\begin{aligned} \alpha_z &= \alpha_x + \alpha_y, \\ \gamma_z &= \min\{\alpha_x + \gamma_y, \alpha_y + \gamma_x\}, \\ \beta_z &= \beta_x \beta_y \left(\mathrm{mod}\ p^{\gamma_z - \alpha_z}\right). \end{aligned} \tag{1.11}$$

For division. Let $z = x/y$. We can only compute this when $\beta_y \neq 0$, otherwise we obtain an undefined object, just like dividing by zero for the real numbers. Note that if $\beta_y = 0$, we may not be dividing by zero but by something we can not recognise as different from zero. So assuming $\beta_y \neq 0$, we have

$$\begin{aligned} \alpha_z &= \alpha_x - \alpha_y, \\ \gamma_z &= \min\{\gamma_x - \alpha_y, \alpha_x + \gamma_y - 2\alpha_y\}, \\ \beta_z &= \beta_x \beta_y^{-1} \left(\mathrm{mod}\ p^{\gamma_z - \alpha_z}\right). \end{aligned} \tag{1.12}$$

Example 1.1.3. [8] For this arithmetic of p-adic numbers, let us consider the 3-adic numbers

$$x = 3 + 2 \cdot 3^2 + 3^4 + O(3^5), \quad y = 2 \cdot 3^{-1} + 3^2 + 2 \cdot 3^3 + O(3^4).$$

Then in this case we have that

$$\begin{aligned} \alpha_x &= 1, \quad \beta_x = 34, \quad \gamma_x = 5, \\ \alpha_y &= -1, \ \beta_y = 191, \quad \gamma_y = 4. \end{aligned}$$

We then, using the rules of arithmetic above, have that

$$\begin{aligned} x + y &= 2 \cdot 3^{-1} + 3 + O(3^4), \\ x \cdot y &= 2 + 3 + 3^2 + O(3^4), \\ x/y &= 2 \cdot 3^2 + 2 \cdot 3^3 + 3^4 + 2 \cdot 3^5 + O(3^6). \end{aligned}$$

Just as one forms algebraic number fields by polynomial extension of rational numbers, so one can form finite extensions of the p-adic numbers by forming polynomial extensions:

$$\mathbb{Q}_p[X] / \left(f\mathbb{Q}_p[X]\right)$$

where $f\mathbb{Q}_p[X]$ is an irreducible polynomial. The problem of computing in such finite extensions is solved in just the same way as when using extensions of the rational

numbers, that is, by using polynomial arithmetic but this time with coefficients in $\mathbb{Q}_p$ represented by the triplets discussed in the p-adic arithmetic above.
The process of taking such finite extensions must eventually terminate in the algebraic closure of $\mathbb{Q}_p$ denoted $\mathbb{Q}_p^a$. However, $\mathbb{Q}_p^a$ is not complete, in the sense that there exist Cauchy sequences in $\mathbb{Q}_p^a$ which do not converge. The completion of $\mathbb{Q}_p^a$, denoted Ω_p, is complete and algebraically closed. The p-adic valuation on $\mathbb{Q}_p$ given by $|x|_p = p^{-\alpha_x}$ extends essentially uniquely to each extension field.
We can arrive at finite extensions of $\mathbb{Q}_p$ in another way which discuss now. Let K be an algebraic number field. Each prime ideal, $\mathfrak{p}$, defines a valuation on K which gives rise to a completion. The valuation is given by looking at the prime ideal factorisation of the principal ideal generated by an element $\phi \in K$. If

$$(\phi) \quad = \quad \mathfrak{p}^{\alpha}\mathfrak{a}/\mathfrak{b},$$

where $\mathfrak{a}$ and $\mathfrak{b}$ are fractional ideals and the ideal $\mathfrak{p}$ is coprime to the ideals $\mathfrak{a}$ and $\mathfrak{b}$. Thus

$$|\phi|_{\mathfrak{p}} \quad = \quad p^{-\alpha f_{\mathfrak{p}}},$$

where $f_{\mathfrak{p}}$ denotes the residue degree of the ideal $\mathfrak{p}$. In what follows below we shall denote $e_{\mathfrak{p}}$ the ramification index of $\mathfrak{p}$ and p is the rational prime lying below $\mathfrak{p}$. Such a completion is a finite extension of $\mathbb{Q}_p$ which contains K, and we shall denote it by $K_{\mathfrak{p}}$. In such a way one obtains an embedding of the number field K into Ω_p. This is similar to the usual $s+t$ embeddings of K into the complex numbers, $\mathbb{C}$, where s denotes the number of real embeddings and t denotes the number of complex conjugate embeddings. Because of the similarity we shall refer to these $s+t$ embeddings into the complex numbers as giving $s+t$ 'infinite' valuations and denote them by $|\cdot|_\infty$. It turns out that all inequivalent valuations on K are in a one to one correspondence with the set of $s+t$ infinite valuations and the valuations arising from each prime ideal.
There are then the following correspondences for a number field $K = \mathbb{Q}(\theta)$ defined by a monic irreducible polynomial $f \in \mathbb{Z}[X]$ such that $f(\theta) = 0$. We let p denote a prime number (which may also include ∞).
Let $f = f_1 \cdots f_r$ denote the factorization of f into irreducible factors in $\mathbb{Q}_p$. Then each non-conjugate embedding into Ω_p (or $\mathbb{C}$) corresponds to one of the $f_i,\ i = 1, \ldots, r$. Two embeddings are said to be conjugate if they map θ into a root of the same f_i. Such conjugate embeddings give the same valuation on K.

If $p \neq \infty$ each factor f_i corresponds to a prime ideal $\mathfrak{p}_i$ lying above p. In such a situation if σ is the corresponding embedding into Ω_p then we have, for any $\phi \in K$, the following identities:

$$|\phi|_{\mathfrak{p}} = |\sigma(\phi)|_p^{e_{\mathfrak{p}} f_{\mathfrak{p}}} = p^{-f_{\mathfrak{p}} e_{\mathfrak{p}} \mathrm{ord}_p(\phi)} = N_{K/\mathbb{Q}}(\mathfrak{p})^{-\mathrm{ord}_{\mathfrak{p}}(\phi)} = |N_{K_{\mathfrak{p}}/\mathbb{Q}_{\mathfrak{p}}}(\phi)|_p. \tag{1.13}$$

The completion of K with respect to a prime ideal, $\mathfrak{p}$, which we shall denote by $K_{\mathfrak{p}}$, is called a 'local field'. The degree of the polynomial f_i, and hence the degree of the extension of the field $\mathbb{Q}_p$, is given by $e_{\mathfrak{p}} f_{\mathfrak{p}}$. The residue field, which we denote by k, is defined to be the quotient field $K_{\mathfrak{p}}/(\mathfrak{p})$. The residue field is a finite field of degree $f_{\mathfrak{p}}$ over $\mathbb{F}_p$.
We now state and prove one of the most important elementary facts in number theory, which is the generalisation of Fermat's little theorem.

Theorem 1.1.1 (Fermat's little theorem [8])**.** *Let $K_{\mathfrak{p}}$ be a local field and let $\alpha \in K_{\mathfrak{p}}$. If $ord_{\mathfrak{p}}(\alpha) = 0$ then*

$$ord_{\mathfrak{p}}\left(\alpha^{N_{K_{\mathfrak{p}}/\mathbb{Q}}(\mathfrak{p})-1} - 1\right) \;>\; 0.$$

In other words,

$$\alpha^{N_{K_{\mathfrak{p}}/\mathbb{Q}}(\mathfrak{p})-1} \;\equiv\; 1 \;(mod\ \mathfrak{p}).$$

Proof. This follows from the fact that the number of elements in the residue field is equal to $N_{K/\mathbb{Q}}(\mathfrak{p}) = p^{f_{\mathfrak{p}}}$. □

An element, $\alpha \in K_{\mathfrak{p}}$, which satisfies $\mathrm{ord}_{\mathfrak{p}}(\alpha) = 0$ is called a unit, because it is an integer of the local field whose multiplicative inverse is also an integer. Therefore, Fermat's little theorem tells us that any unit can be made congruent to 1 (mod $\mathfrak{p}$) by just raising it to some power which is a divisor of $N_{K_{\mathfrak{p}}/\mathbb{Q}}(\mathfrak{p}) - 1$.

Example 1.1.4. Let K be the number field $K = \mathbb{Q}(\theta)$, where $\theta^2 + 1 = 0$. Consider the element $\alpha = 1 + \sqrt{-1}$. As α has norm, $N_{K/\mathbb{Q}}(\alpha) = (1+\sqrt{-1})(1-\sqrt{-1}) = 2$, it is a unit of $K_{\mathfrak{p}}$ for any prime ideal $\mathfrak{p}$ that is not lying above 2. The ideal $\mathfrak{p} = (3)$ is prime in K and thus it has residue degree 2. By Fermat's little theorem if we raise α to a power of some divisor of $3^2 - 1 = 8$, then we obtain an element which is congruent to 1 (mod 3). That is,

$$\begin{aligned}
(1+\sqrt{-1})^2 &= 2\sqrt{-1}, \\
(1+\sqrt{-1})^4 &= (2\sqrt{-1})^2 \;=\; -4 \;=\; 2 + 3 + 2\cdot 3^2 + 2\cdot 3^3 + \cdots, \\
(1+\sqrt{-1})^8 &= (-4)^2 \;=\; 16 \;=\; 1 + 2\cdot 3 + 1\cdot 3^2.
\end{aligned}$$

Hence we need to raise α to the eighth power to achieve the desired result.

1.2 p-adic numerical analysis

In this section we discuss some of the topics which we sometimes meet in a numerical analysis course in the context of real numbers, namely issues of finding roots of polynomials to arbitrary precision using the Newton-Raphson formula, computing solutions to power series equations to arbitrary accuracy and providing algorithms to compute transcendental functions to a given accuracy. The analogues of these problems could all be considered to come from an area of 'p-adic numerical analysis'.

1.2.1 Newton-Raphson method

Suppose we are given a monic polynomial $f(X) \in \mathbb{Z}_p[X]$ and we wish to compute a root to this polynomial in $\mathbb{Z}_p$. One way of doing this would be to mimic the Newton-Raphson method that is used in the real case. This method is so successful and important and it is named after Hensel, who was the first to use it in the p-adic context. Hensel's lemma plays a fundamental role in may algorithms in computer algebra such as polynomial factorisation. Hensel's lemma provides a criterion for when a solution modulo p^n can be made into a solution modulo p^{n+1}, that is, a solution modulo p^n is 'lifted' to a solution modulo p^{n+1}. This process can be repeated to lift the solution modulo p^{n+1} to a solution modulo p^{n+2} and so on.

Remark 1.2.1. In the p-adic integers, $\mathbb{Z}_p$, congruences are approximations. That is, for $a, b \in \mathbb{Z}_p$, $a \equiv b \pmod{p^n}$ is the same as $|a-b|_p \leq p^{-n}$. Thus, turning information modulo one power of p into similar information modulo a higher power of p can be interpreted as improving an approximation.

Theorem 1.2.1 (Hensel's lemma [8])**.** *Let $f(X) \in \mathbb{Z}_p[X]$ be monic and let $a_0 \in \mathbb{Z}_p$ denote an approximation to the value of a root of $f(X)$ such that*

$$|f(a_0)|_p \quad \leq \quad p^{-2\delta-1},$$

where $\delta = ord_p\left(f'(a_0)\right)$ with $f'(a_0) \neq 0$. Then the following sequence tends to a root $a \in \mathbb{Z}_p$:

$$a_{n+1} \quad = \quad a_n - \frac{f(a_n)}{f'(a_n)}.$$

In addition the limit, a, is the unique root of $f(X)$ satisfying

$$|a - a_0|_p < p^{-\delta}.$$

To prove the theorem, we need to first prove the following lemma.

Lemma 1.2.1. *For all $n \in \mathbb{N}$ we have*

$$|f(a_n)|_p \leq p^{-2\delta-n-1}, \quad \textit{and} \quad |a_n - a_{n-1}|_p \leq p^{-\delta-n}.$$

Proof. We prove this lemma by induction, assuming that the result holds for all values less than or equal to N. By the second assumption there is a $b \in \mathbb{Z}_p$ such that

$$a_N = a_{N-1} + p^{\delta+N}b.$$

Then we have, by applying a Taylor series expansion,

$$f(a_N) = f(a_{N-1} + p^{\delta+N}b) = f(a_{N-1}) + O(p^{\delta+N}).$$

So then we have,

$$f'(a_N) = f'(a_{N-1} + p^{\delta+N}b) = f'(a_{N-1}) + O(p^{\delta+N}).$$

But then we have,

$$\mathrm{ord}_p(f'(a_N)) = \mathrm{ord}_p(f'(a_{N-1})) = \cdots = \mathrm{ord}_p(f'(a_0)) = \delta,$$

whence our first assumption implies that

$$\left|\frac{f(a_N)}{f'(a_N)}\right|_p \leq p^{-\delta-N-1}.$$

Therefore,

$$|a_{N+1} - a_N|_p \leq p^{-\delta-(N+1)},$$

which proves the second assertion. To prove the first assertion we need to apply Taylor's theorem,

$$f(a_{N+1}) = f(a_N) - f'\left(\frac{f(a_N)}{f'(a_N)}\right) + c\left(\frac{f(a_N)}{f'(a_N)}\right)^2 = \left(\frac{f(a_N)}{f'(a_N)}\right)^2 c,$$

where $c \in \mathbb{Z}_p$. Hence we find that

$$|f(a_{N+1})|_p \leq p^{-2\delta-2(N+1)} \leq p^{-2\delta-(N+1)-1}.$$

The initial case when $N = 1$ is trivial, so this completes the proof of the lemma. □

Proof. (Hensel's lemma) Using the previous lemma, it is clear that the sequence in Hensel's lemma converges to a zero of the polynomial $f(X)$. Hence we only have to show that this is a unique zero within the required range. Suppose that there is another root α such that

$$|\alpha - a|_p \;\leq\; p^{-\delta-1}.$$

We shall show that $|\alpha - a_N|_p \leq p^{-\delta-N-1}$ implies that $|\alpha - a_{N+1}|_p \leq p^{-\delta-N-2}$, from which the required result will follow. Again using Taylor's theorem we find out that (putting $p^{\delta+N+1}b = \alpha - a_N$ for some $b \in \mathbb{Z}_p$) there is a $c \in \mathbb{Z}_p$ such that

$$f(a_N) + f'(a_N)p^{\delta+N+1}b + p^{2\delta+2N+2}b^2c \;=\; f(\alpha) = 0.$$

Hence we obtain that

$$p^{\delta+N+1}b \;=\; -\frac{f(a_N)}{f'(a_N)} + O(p^{\delta+2N+1}),$$

and thus,

$$\alpha \;=\; a_{N+1} + O(p^{\delta+N+2}).$$

This completes the proof. □

Example 1.2.1. Let p denote an odd prime and consider the polynomial $f(X) = X^2+1$. Clearly a solution of this equation modulo p can be considered as an element, α_0, of $\mathbb{Z}_p$ such that $f'(\alpha_0) \;=\; 2\alpha_0 \;\not\equiv\; 0 \pmod{p}$.
Hence by Hensel's lemma we can 'lift' a solution modulo p to a solution in $\mathbb{Z}_p$. For instance $X^2+1 = 0$ has the following solution in $\mathbb{Z}_5$ using the Hensel lemma algorithm:

$$f(X) = X^2 + 1 \;\Rightarrow f(a_n) = a_n^2 + 1, \quad f'(X) = 2X \;\Rightarrow f'(a_n) = 2a_n.$$

Then, substituting in the formula, we get

$$a_{n+1} \;=\; a_n - \frac{f(a_n)}{f'(a_n)} \;=\; a_n - \frac{a_n^2+1}{2a_n} \;=\; \frac{1}{2}\left(a_n - \frac{1}{a_n}\right).$$

We note that $\alpha_0 = 2$ is a zero of $X^2 + 1 = 0 \pmod 5$, so we take $a_0 = 2 \pmod 5$ as the initial approximation of the root in $\mathbb{Z}_5$. The next solutions are found as follows:

$$\begin{aligned}
a_1 &= \frac{1}{2}\left(a_0 - \frac{1}{a_0}\right) = \frac{3}{4} \equiv 7 = 2 + 1\cdot 5 \pmod{5^2},\\
a_2 &= \frac{1}{2}\left(a_1 - \frac{1}{a_1}\right) = \frac{24}{7} \equiv 57 = 2 + 1\cdot 5 + 2\cdot 5^2 \pmod{5^3},\\
a_3 &= \frac{1}{2}\left(a_2 - \frac{1}{a_2}\right) = \frac{199}{57} \equiv 182 = 2 + 1\cdot 5 + 2\cdot 5^2 + 1\cdot 5^3 \pmod{5^4},
\end{aligned}$$

and so on. Continuing in this fashion we obtain the following solution

$$a = 2 + 1 \cdot 5 + 2 \cdot 5^2 + 1 \cdot 5^3 + 3 \cdot 5^4 + 4 \cdot 5^5 + 2 \cdot 5^6 + 3 \cdot 5^7 + 3 \cdot 5^9 + \cdots .$$

Therefore, Hensel's lemma provides a mechanism to lift an approximate solution modulo an appropriate power of p to a unique solution in $\mathbb{Z}_p$.

1.2.2 Power series in one variable

In this subsection we investigate the properties of power series over the p-adic numbers. We will consider the power series to be polynomials of infinite degree. In practice, when implementing power series in the programs coded in sage, they will be of finite degree, with the remainder specified by $O(x^{n+1})$, depending on the number of variables and the desired precision n.
We begin by investigating convergence properties of sequences and series in $\mathbb{Q}_p$. We recall that since $\mathbb{Q}_p$ is complete, every Cauchy sequence converges. Furthermore, all the axioms of that hold for the usual absolute value in $\mathbb{R}$ also hold in $\mathbb{Q}_p$, so most of the basic theorems still hold in the p-adic sense.

Definition 1.2.1. Let K be a field and let $|\cdot|$ be a non-Archimadean absolute value on K. The subring

$$\mathcal{O}_K = \{x \in K : |x| \leq 1\} \in K \tag{1.14}$$

is called the *valuation ring* of $|\cdot|$. The ideal

$$\mathcal{I}_K = \{x \in K : |x| < 1\} \in \mathcal{O}_K \tag{1.15}$$

is called the *valuation ideal* of $|\cdot|$.

This definition immediately leads to

Definition 1.2.2. The ring of p-adic integers is the valuation ring

$$\mathbb{Z}_p = \{x \in \mathbb{Q}_p : |x|_p \leq 1\}. \tag{1.16}$$

Since the p-adic absolute value is non-Archimedean, the properties of quantities and concepts contingent on the absolute value such as convergence of a Cauchy sequence is likely to differ from that of the real numbers. Indeed, we have

Proposition 1.2.1. A sequence $\{a_n\}$ in $\mathbb{Q}_p$ is a Cauchy sequence, and therefore convergent if and only if it satisfies

$$\lim_{n\to\infty} |a_{n+1} - a_n|_p = 0. \tag{1.17}$$

Proof. If $m = n + r > n$, we get

$$\begin{aligned} |x_m - x_n|_p &= |x_{n+r} - x_{n+r-1} + x_{n+r-1} - x_{n+r-2} + \cdots + x_{n+1} - x_n|_p \\ &\leq \max\left\{|x_{n+r} - x_{n+r-1}|_p, |x_{n+r-1} - x_{n+r-2}|_p, \ldots, |x_{n+1} - x_{n+}|_p\right\}, \end{aligned}$$

since the p-adic absolute value is non-Archimedean. Then the result follows. □

Remark 1.2.2. This result is in clear contrast to analysis in $\mathbb{R}$ where the condition

$$\lim_{n\to\infty} |x_{n+1} - x_n| = 0 \tag{1.18}$$

is not equivalent to the Cauchy condition. For example, consider the harmonic sequence

$$x_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n},$$

for which $|x_{n+1} - x_n| = 1/(n+1)$ which approaches zero as $n \to \infty$. However, it is possible to show that $x_{2^k} \geq (k+2)/2$, hence the sequence is unbounded and does not have a limit.

As a corollary to Proposition 1.2.1, we have

Corollary 1.2.1.1. *An infinite series $\sum a_n$ with $a_n \in \mathbb{Q}_p$ is convergent if and only if $\lim a_n = 0$, in which case we also have*

$$\left|\sum_{n=0}^{\infty} a_n\right|_p \leq \max_n |a_n|_p.$$

Proof. A series converges when the sequence of partial sums converges. We suppose

$$S_N = \sum_{n=0}^{N} a_n.$$

Then $a_n = S_n - S_{n-1}$. If it tends to zero, then it immediately follows from Proposition 1.2.1 that the sequence of partial sums is a Cauchy sequence. The converse direction, assuming the series to be convergent, is trivial. The result about the absolute value of the sum follows from the non-Archimedean property. □

Let p be a prime number. Let a_n denote a sequence of p-adic numbers; then the series $\sum a_i$ converges when $a_n \to 0$, in the p-adic sense. This gives a rather nice convergence criterion for a power series. Let

$$f(X) = \sum_{i\geq 0} a_i X^i = a_0 + a_1 X + a_2 X^2 + \cdots \tag{1.19}$$

denote a power series with p-adic coefficients. Then this series converges at a point x if and only if $a_i x^i \to 0$. Hence it will converge for all values of x if

$$\limsup_{i\to\infty} |a_i|_p^{1/i} \quad = \quad 0, \tag{1.20}$$

that is, the a_i become very highly divisible by p as i increases.
The theorem below is due to Strassman and it is the main result we shall require on power series in one variable. It allows us to bound the number of zeros of such a series of p-adic numbers.

Theorem 1.2.2 (Strassman [8]). *Let a_i be a sequence of p-adic numbers, not all zero, and let*

$$f(X) \quad = \quad \sum_{i\geq 0} a_i X^i \quad = \quad a_0 + a_1 X + a_2 X^2 + \cdots$$

be a power series which converges for all $x \in \mathbb{Z}_p$, that is, $|a_i|_p \to 0$. Define N such that

$$|a_N|_p \quad = \quad \max |a_i|_p, \quad |a_i|_p < |a_N|_p \quad \forall i \geq N.$$

Then there are atmost N elements $\alpha \in \mathbb{Z}_p$ such that $f(\alpha) = 0$.

Proof. We prove the theorem by induction on N. Firstly we prove the initial step and suppose $N = 0$: We know from the condition on N that

$$|a_n|_p \quad > \quad |a_0|_p \quad \text{for all } n > 0.$$

We now assume for the purpose of deriving a contradiction, that there is actually an $\alpha \in \mathbb{Z}_p$ such that $f(\alpha) = 0$. Hence we have

$$\begin{aligned} |a_0|_p \quad &\leq \quad \left| \sum_{i\geq 1} a_i \alpha^i \right|_p, \quad \text{since } \alpha \text{ is a zero,} \\ &\leq \quad \max_{i\geq 1} |a_i|_p, \\ &< \quad |a_0|_p, \quad \text{because } N = 0, \end{aligned}$$

which gives a contradiction.
We now prove the induction step and assume that $N > 0$ and that the theorem is true for $N - 1$. Let α denote a zero of $f(X)$. If there exists is no such α, then we are done. We define a new function $g(X)$ by

$$g(X) \quad = \quad \sum_{i\geq 0} b_i X^i \quad \text{where} \quad b_i \quad = \quad \sum_{j\geq 0} a_{i+1+j} \alpha^j.$$

We then find out that:

1.

$$|b_i|_p \leq \max_{j\geq 0} |a_{i+1+j}|_p \leq |a_N|_p.$$

2.

$$|b_{N-1}|_p \leq \max_{i\geq N} |a_i|_p.$$

Then as $\alpha \in \mathbb{Z}_p^*$ and $N \geq 0$ we find that $|b_{N-1}|_p = |a_N|_p$.

3. If $i \geq N$ we find that

$$|b_i|_p \leq \max_{j\geq N} |a_j|_p \ |a_N|_p.$$

Therefore, we see that the power series $g(X)$ satisfies the conditions of the theorem but for $N-1$. By our inductive hypothesis there are then at most $N-1$ elements $\beta \in \mathbb{Z}_p$ such that $g(\beta) = 0$. We finally have to show that this implies that $f(X) = 0$ has at most N solutions. We already know the existence of one solution, namely, α. But then,

$$f(X) = f(X) - f(\alpha) = \sum_{i\geq 1} a_i \left(X^i - \alpha^i\right) = (X-\alpha)g(X).$$

Whence, any solution of $f(X) = 0$ is either a solution of $g(X) = 0$ or equal to α. So there are at most N solutions to $f(X) = 0$. This completes the proof. □

Example 1.2.2. Consider the (p-adic) power series

$$f(X) = \sum_{n=0}^{\infty} n! X^n.$$

We want to estimate the number of zeros of $f(X)$ by Strassmann's theorem: We set $a_n := n!$, $|a_n|_p = 1$ for all $n \in \{0, \ldots, p-1\}$, $|a_n|_p \leq 1/p$ for all $n \in \{p, \ldots, 2p-1\}$, and so on. So $\limsup |a_n|_p = 1$. Since $|a_n|_p \to 0$, as $n \to \infty$, f converges in $\mathbb{Z}_p$. The number we are looking for is $N = p-1$. We are therefore able to conclude by Strassmann's theorem that f has at most $p-1$ zeros.

Example 1.2.3. In $\mathbb{R}$, Strassmann's theorem is clearly not true. We consider the sine function to see this

$$f(X) = \sin(X) = \sum_{n\geq 0} \frac{(-1)^n}{(2n+1)!} X^{2n+1}.$$

The sequence a_n is given by

$$a_n := \begin{cases} 0, & \text{if } n \text{ is even} \\ \dfrac{(-1)^{\frac{n-1}{2}}}{n!}, & \text{if } n \text{ is odd.} \end{cases}$$

For this, we can see that $N = 1$, from which we can conclude by Strassmann's theorem that $\sin(X) = 0$ has at most one zero. But we know that the sine function has infinitely many zeros and not at most one!.

Definition 1.2.3. Consider the power series $\sum a_n X^n$ where $a_n \in \mathbb{Q}_p$. Then the *radius of convergence* of the series $\sum a_n X^n$ is given by

$$r = \frac{1}{\limsup |a_n|_p^{1/n}}. \tag{1.21}$$

Proposition 1.2.2. The series $\sum a_n X^n$ converges if $|X|_p < r$ and diverges if $|X|_p > r$, where r is the radius of convergence. If for some X_0 with $|X_0|_p = r$ the series $\sum a_n X_0^n$ converges (or diverges), then the series $\sum a_n X_0^n$ converges (or diverges) for all $X \in \mathbb{Q}_p$ with $|X|_p = r$.

Proof. We use our convergence criterion that the series $\sum a_n$ converges if $|a_n|_p \to 0$. Then we first notice that if $|X|_p < r$, then we have

$$|a_n X^n|_p = |a_n|_p |X|_p^n \to 0 \text{ as } n \to \infty.$$

Similarly, if $|X|_p > r$, we have

$$|a_n X^n|_p = |a_n|_p |X|_p^n \nrightarrow 0 \text{ as } n \to \infty.$$

Finally, if there is such an $X_0 \in \mathbb{Q}_p$, then we have

$$|a_n X_0^n|_p = |a_n|_p |X_0|_p^n \to 0 \text{ as } n \to \infty,$$

and thus for every $X \in \mathbb{Q}_p$ with $|X|_p = r$ we have

$$|a_n X^n|_p = |a_n X_0^n|_p = |a_n|_p |X_0|_p^n \to 0 \text{ as } n \to \infty.$$

□

Example 1.2.4. We show that the radii of convergence of the p-adic power series below

$$\exp_p(X) = \sum_{n=0}^{\infty} \frac{X^n}{n!} \quad \text{and} \quad \log_p(X) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1} X^n}{n}$$

are $p^{-1/(p-1)}$ and 1, respectively.
From the previous results, we have

$$|1/n!|_p^{1/n} = p^{(n-\alpha_p(n))/n(p-1)} = p^{(1-\alpha_p(n)/n)/(p-1)},$$

and thus we have

$$\limsup |1/n!|_p^{1/n} = \limsup_{n\to\infty} p^{(1-\alpha_p(n)/n)/(p-1)} = p^{1/(p-1)}.$$

Therefore, the radius of convergence of $\exp_p(X)$ is $p^{-1/(p-1)}$.
Similarly, we have

$$|1/n|_p^{1/n} = p^{\mathrm{ord}_p(n)/n} \Rightarrow \limsup |1/n|_p^{1/n} = 1.$$

Hence, the radius of convergence of $\log_p(X)$ is 1.

1.2.3 Power series in many variables

[8] We shall assume that we are given n power series in n variables with coefficients coming from $\mathbb{Z}_p$. Let $\vec{f}$ denote such a vector of power series. We define the Jacobian matrix of such a system by

$$Jac_{\vec{f}}(\vec{x}) = \left(\frac{\partial f_i}{\partial x_j}\right). \tag{1.22}$$

The determinant of the Jacobian matrix we shall denote by $J_{\vec{f}}(\vec{x})$.
We shall require the following standard result on formal power series

Lemma 1.2.2. *[8] Let $\vec{f}$ denote an n-vector of power series in n variables with no constant term. Suppose $J_{\vec{f}}(\vec{0}) \in \mathbb{Z}_p^*$. Then $\vec{f}$ has an 'inverse' vector of power series with respect to composition of functions.*

Proof. See [5] □

We use this result to prove the following theorem

Theorem 1.2.3 (mult-dimensional Hensel [8])**.** *Let $\vec{f}$ denote an n-vector of power series in n variables. Suppose there is a vector $\vec{a} \in \mathbb{Z}_p^n$ such that*

$$\vec{f}(\vec{a}) \equiv \vec{0}\ (mod\ p^{2\delta+1}), \tag{1.23}$$

where $\delta = ord_p(J_{\vec{f}}(\vec{a})) < \infty$. Then there is a unique zero of the system of power series $\vec{\alpha}$ such that

$$\vec{\alpha} \equiv \vec{a}\ (mod\ p^{\delta+1}). \tag{1.24}$$

This theorem is completely analogous to the standard multi-dimensional version of the Newton-Raphson algorithm in ordinary numerical analysis.

Proof. Just as in the proof of Hensel's lemma we prove this theorem using a Taylor series expansion

$$\vec{f}(\vec{a}+p^{\delta}\vec{X}) \;=\; \vec{f}(\vec{a})+Jac_{\vec{f}}(\vec{a})p^{\delta}\vec{X}+p^{2\delta}\vec{r}(\vec{X}).$$

The remainder power series $\vec{r}(\vec{X})$ will have zero constant and first degree terms. We define the new vector power series

$$\vec{g}(\vec{X}) \;=\; \vec{X}+A\vec{r}(\vec{X}),$$

where A is the unique matrix such that

$$AJac_{\vec{f}}(\vec{a}) \;=\; p^{\delta}I_n.$$

The vector of power series $\vec{g}(\vec{X})$ has an inverse, by Lemma 1.2.2, with respect to composition of functions $\vec{g}^{-1}$; this inverse also has no constant terms. We then find that

$$\begin{aligned}\vec{f}(\vec{a}+p^{\delta}\vec{g}^{-1}(\vec{X})) &= \vec{f}(\vec{a})+Jac_{\vec{f}}(\vec{a})p^{\delta}\vec{g}(\vec{g}^{-1}(\vec{X})),\\ &= \vec{f}(\vec{a})+Jac_{\vec{f}}(\vec{a})p^{\delta}\vec{X}.\end{aligned}$$

We know that $\vec{f}(\vec{a})=p^{2\delta}\vec{b}$, where $\vec{b}$ is a vector congruent to $\vec{0}$ modulo p. We then define

$$\vec{\alpha} \;=\; \vec{a}+p^{\delta}\vec{g}^{-1}(-A\vec{b}).$$

Then we have

$$f(\vec{\alpha}) \;=\; f(\vec{a})-AJac_{\vec{f}}(\vec{a})p^{\delta}\vec{b} \;=\; f(\vec{a})-p^{2\delta}\vec{b} \;=\; 0.$$

That α is the unique such vector follows from the fact that the matrix A has determinant equal to a unit in $\mathbb{Z}_p$. Hence $\vec{x}=-A\vec{b}$ is the unique solution to the equation

$$p^{2\delta}\vec{b}+Jac_{\vec{f}}(\vec{a})p^{\delta}\vec{x} \;=\; 0,$$

and $\vec{x}$ is congruent to $\vec{0}$ modulo p as $\vec{b}$ is. This completes the proof. □

1.2.4 The Iwasawa logarithm

While we are talking about analogues of the results and problems in standard numerical analysis we can also discuss how to compute p-adic logarithms. Firstly we look at the usual Taylor series expansion of the normal real logarithm about the point 1

$$\log(1+x) \quad = \quad \sum_{i\geq 1} \frac{(-1)^{i+1}x^i}{i} \tag{1.25}$$

which satisfies the identity

$$\log((1+x)(1-x)) \quad = \quad \log(1+x)+\log(1-x) \tag{1.26}$$

We could define a p-adic logarithm by taking the above series as a definition. However, we have to worry about convergence problems.
Now if $z \in \Omega_p$ and if $|z-1|_p < 1$, we define the p-adic logarithm by the same series

$$\log_p(z) \quad = \quad -\sum_{i\geq 1} \frac{(1-z)^i}{i}, \tag{1.27}$$

which certainly converges. In such a region of convergence we therefore also have the identity

$$\log_p((1+x)(1-x)) \quad = \quad \log_p(1+x)+\log_p(1-x). \tag{1.28}$$

In the region where $|z|_p < p^{-1/(p-1)}$ we also have that

$$\mathrm{ord}_p\left(\log_p(1+z)\right) \quad = \quad \mathrm{ord}_p z. \tag{1.29}$$

We would like to define the logarithm for the whole of Ω_p. We do this using an idea of Iwasawa with the following rules:

1. For all $x, y \in \Omega_p$ we have $\log_p(xy) = \log_p(x) + \log_p(y)$.
2. If ω is a root of unity in ω_p and $s \in \mathbb{Z}$ then $\log_p(\omega p^s) = 0$.

Using the above definition we can evaluate the p-adic logarithm at any point $\alpha \in \Omega_p$. In our later examples α will be a unit of some $K_{\mathfrak{p}}$ where K is some number field and $\mathfrak{p}$ is a prime ideal. So we shall assume that this case holds for convenience. Note that, since $K_{\mathfrak{p}}$ is complete, then $\alpha \in K_{\mathfrak{p}}$ implies that $\log_p(\alpha) \in K_{\mathfrak{p}}$. We let e denote the ramification index of $\mathfrak{p}$ and f the residue degree. By Fermat's little theorem we know that the order of the image of α in the residue field $\mathbb{F}_{p^f}$ divides $p^f - 1$. We can hence compute the order of the image of α in $\mathbb{F}_{p^f}$, we denote it o.This can be done

by using either the naive method or the Baby-Step-Giant-Step method, see [8]. For elements of large finite fields the determination of o may not be that easy, however in the examples which interest us the finite field will be relatively small.
Now we note that if we choose t such that $p^t > e$, and assume p is odd prime, then

$$(1-\alpha^o)^{p^t} = 1 - p^t\alpha^o + \cdots - \alpha^{op^t} \tag{1.30}$$

and so $\mathrm{ord}_p\left(1-\alpha^{op^t}\right) > \mathrm{ord}_p\left(\mathfrak{p}^{p^t}\right) > 1$. It can be easily verified that the last inequality holds for $p = 2$. Then we have

$$\log_p(\alpha) = \frac{1}{op^t}\log_p\left(\alpha^{op^t}\right) = \frac{-1}{op^t}\sum_{i\geq 1}\frac{\left(1-\alpha^{op^t}\right)^i}{i}. \tag{1.31}$$

We are hence left only with the task of studying how fast such a series converges and developing techniques to speed up the convergence. We shall want to know how many terms to take to obtain a desired level of accuracy, a question which is answered by the following result:

Lemma 1.2.3. *Let $ord_p(1-z) \geq 1$ and let M denote an arbitrary given integer. We let N denote the smallest integer solution of*

$$n \geq \frac{1}{ord_p(1-z)}\left(\frac{\log n}{\log p} + M\right). \tag{1.32}$$

Then we have

$$\log_p(z) = -\sum_{i=1}^{N}\frac{(1-z)^i}{i} + O(p^M). \tag{1.33}$$

Proof. We first note that $\mathrm{ord}_p n \leq \frac{\log n}{\log p}$ for all positive integers n. Now if $n \geq N$, we have

$$\begin{aligned}\mathrm{ord}_p\left(\frac{(1-z)^n}{n}\right) &= n\mathrm{ord}_p(1-z) - \mathrm{ord}_p n \\ &\geq \left(\frac{\log n}{\log p} + M\right) - \leq \frac{\log n}{\log p} \\ &\geq M.\end{aligned}$$

Hence

$$\mathrm{ord}_p\left(-\sum_{i=1}^{N}\frac{(1-z)^i}{i}\right) \geq M.$$

From which the required result follows. □

---

[8] **Algorithm for p-adic logarithms**

`DESCRIPTION:` Finds the p-adic logarithm of the algebraic number $\alpha \in K$ with respect to the embedding of K into Ω_p given by the ideal $\mathfrak{p}$.
α is assumed to be a unit of $K_{\mathfrak{p}}$

`INPUT:` $\alpha \in K$, a prime ideal, $\mathfrak{p}$, of $\mathcal{O}_K$ and a natural number M.

`OUTPUT:` The p-adic logarithm β up to accuracy of p^M.

1. Compute o such that $\mathrm{ord}_{\mathfrak{p}}(\alpha^o - 1) > 0$.
2. Set $\gamma = \alpha^{op^t}$ where t is chosen to be the smallest number such that $m = \mathrm{ord}_p(\gamma - 1) \geq \frac{1}{2}\mathrm{ord}_p(D(\theta)) + 1$.
3. Compute the smallest integer solution, n, to $n \geq \left(\frac{\log n}{\log p} + M\right)/m$.
4. Set $\beta := 0$ and $\delta := 1 - \gamma$.
5. For $i = 1, \ldots, n$ do
6. $\quad \beta := \beta - \delta/i$.
7. $\quad \delta := \delta(1 - \gamma)$.
8. Enddo.
9. $\beta := \beta/(op^t)$.

In such an algorithm we need to take care of any coefficient swell. If $K = \mathbb{Q}(\theta)$ we can write $\gamma - 1$ as a polynomial in θ. We can assume that no coefficient has a denominator divisible by p, hence we can assume that $\gamma - 1 \in \mathbb{Z}_p[\theta]$. By the choice of o and t the polynomials representing β and δ have no coefficients with p-adic value greater than one. For the reason for the choice of t we see the proof of Lemma 1.2.1. Hence we may reduce every coefficient in the logarithm by taking its value modulo

$$p^{M + \frac{\log M}{\log p}}.$$

This allows us to take care of the possible coefficient swell.

Example 1.2.5. Suppose we want to compute the 3-adic logarithm of the rational integer 2. First is we need to compute an exponent o such that $2^o \equiv 1 \pmod 3$.

Clearly we can take $o = 2$, in which case we have

$$\log_3(2) \;=\; \frac{\log_3(4)}{2}.$$

Hence we need to compute $\log_3(4)$, but since $4 \equiv 1 \pmod 3$, this can be done from the series

$$\begin{aligned}\log_3(4) &= -\sum_{i\geq 1}\frac{(1-4)^i}{i}\\ &= -\left(-3+\frac{9}{2}-9+\frac{81}{4}-\frac{243}{5}+\frac{243}{2}+O(3^7)\right)\\ &= 3+2\cdot 3^2+3^3+2\cdot 3^5+2\cdot 3^6+O(3^7).\end{aligned}$$

Therefore, we have

$$\log_3(2) \;=\; 2\cdot 3+2\cdot 3^2+3^5+3^6+O(3^7).$$

Remark 1.2.3. One of the ways to speed up the computation of p-adic logarithms is to use an observation of de Weger [4]. Instead of using the series

$$\log_p z \;=\; -\sum_{i\geq 1}\frac{(1-z)^i}{i}$$

we could use instead the series

$$\log_p\left(\frac{1+z}{1-z}\right) \;=\; 2\left(\sum_{i\geq 0}\frac{z^{2i+1}}{2i+1}\right) \;=\; 2\left(z+\frac{z^3}{3}+\frac{z^5}{5}+\cdots\right).$$

Of course if we make z very close to zero p-adically, then the above series will converge much faster.

Example 1.2.6. As in the Example 1.2.5 above, suppose one wants to compute $\log_3(2)$. Again this is easy once we have computed $\log_3(4)$. We find that

$$\begin{aligned}\log_3(4) &= \log_3\left(\frac{1+\frac{3}{5}}{1-\frac{3}{5}}\right)\\ &= 2\left(\frac{3}{5}+\frac{9}{125}+\frac{243}{15625}+\cdots\right)\\ &= 3+2\cdot 3^2+3^3+2\cdot 3^5+2\cdot 3^6+O(3^7).\end{aligned}$$

Therefore, as before

$$\log_3(2) \;=\; 2\cdot 3+2\cdot 3^2+3^5+3^6+O(3^7).$$

1.2.5 p-adic exponential function

This section would not be complete without a discussion of the p-adic exponential function. This function is defined by

$$\exp_p z = \sum_{i\geq 1} \frac{z^n}{n!}, \tag{1.34}$$

which converges if $\mathrm{ord}_p z > 1/(p-1)$. The function also satisfies the following formulae, in the region in which it is defined:

$$\begin{aligned} (1+z)^a &= \exp_p(a \log_p(1+z)) \\ \exp_p(z_1+z_2) &= \exp_p(z_1)\exp_p(z_2) \\ \mathrm{ord}_p z &= \mathrm{ord}_p(\exp_p(z)-1). \end{aligned}$$

Finally we notice that we have the following result

Lemma 1.2.4. *Let $\alpha \in \Omega_p$ denote a p-adic unit. If*

$$ord_p(\alpha - 1) > \frac{1}{p-1} \tag{1.35}$$

then

$$ord_p(\alpha - 1) = ord_p(\log_p \alpha). \tag{1.36}$$

Proof. The proof of this lemma follows directly from the above formulae satisfied by $\exp_p$ within its region of definition. □

CHAPTER 2

Applications of Local Methods to Diophantine Equations

In this chapter we give some of the local considerations which either allow us to completely solve a diophantine equation, aid us in locating the solutions or give us information about the solutions which can be used in a more advanced method. We show how to apply the p-adic analysis in the previous chapter to find solutions to equations using Skolem's method and then finally we discuss how various pieces of local information can be put together in an algorithmic method using sieving. Sieving is no more than a catch-all phrase for a process meaning applying local considerations one after another to sieve out (or remove) non-solutions [8]. The idea behind sieving is that anything left after we have used a sieve has a good chance of being an actual solution.

2.1 Some useful preliminary results

Lemma 2.1.1. *Let $f \in \mathbb{Q}[X, Y]$ be a homogeneous polynomial in two variables X and Y of degree n such that the degree of $f(X, 1)$ is n as well. Then $f(X, Y)$ is irreducible if and only if $f(X, 1) \in \mathbb{Q}[X]$ is irreducible.*

Proof. Suppose $f(X, Y)$ is irreducible. Then the coefficient for $X^n$is non-zero, so $f(X, 1)$ has degree n. Suppose that $g(X) = f(X, 1)$ were reducible, say $g = h \cdot k$ with $\deg h = m$ and $\deg k = n - m$. Now let h' and k' be the polynomials obtained by adding the power of Y to each monomial such that h' is homogeneous of degree n and its coefficient for $X^j Y^{n-j}$ is the same as the coefficient for X^j in hk. Therefore, we conclude that $h'k' = f(X, Y)$.

Coversely, suppose $f(X,Y)$ were reducible, then a factorization of $f(X,Y)$ includes a factor of X in both factorizations so that $f(X,1)$ would be reducible as well. This completes the proof. □

The following lemma allows us to use the Dirichlet's unit theorem, which is the starting point for Skolem's method

Lemma 2.1.2. *Let $f \in \mathbb{Q}[X,Y]$ be an irreducible homogeneous polynomial in two variables of degree n such that $f(X,Y)$ is monic and of degree n. Then for any $a,b \in \mathbb{Q}$, $f(a,b) = N_{K/\mathbb{Q}}(a - b\theta)$, where θ is a zero of $f(X,1)$ in $\mathbb{C}$ and $K = \mathbb{Q}(\theta)$.*

Proof. Since f is monic, $f(X,1)$ has degree n. Consider

$$f(X,1) \;=\; (X-\alpha_1)(X-\alpha_2)\cdots(X-\alpha_n).$$

Using the same argument as in the proof of the previous lemma, we find that

$$f(X,Y) \;=\; (X-\alpha_1 Y)(X-\alpha_2 Y)\cdots(X-\alpha_n Y).$$

Let $\theta = \alpha_1$ and $K = \mathbb{Q}(\theta)$. Then by Lemma 2.1.1, $[K:\mathbb{Q}] = n$ and we see that the α_i are the Galois conjugates of θ, so that $f(a,b) = N_{K/\mathbb{Q}}(a - b\theta)$ for each $a,b \in \mathbb{Q}$. □

Lemma 2.1.3. *Let K be a number field. An element $a \in \mathcal{O}_K$ is a unit if and only if $N_{K/\mathbb{Q}}(a) = \pm 1$.*

Proof. Suppose $a \in \mathcal{O}_K$ is a unit, then $a^{-1} \in \mathcal{O}_K$ is also a unit, and therefore, since $1 = aa^{-1}$ we have from the properties of the norm that

$$1 \;=\; N_{K/\mathbb{Q}}(a)N_{K/\mathbb{Q}}(a^{-1}).$$

Since both $N_{K/\mathbb{Q}}(a)$ and $N_{K/\mathbb{Q}}(a^{-1})$ are integers, it follows that $N_{K/\mathbb{Q}}(a) = \pm 1$. Conversely, suppose $a\mathcal{O}_K$ and $N_{K/\mathbb{Q}}(a) = \pm 1$, then the equation

$$aa^{-1} \;=\; 1 \;=\; \pm N_{K/\mathbb{Q}}(a),$$

implies that $a^{-1} = N_{K/\mathbb{Q}}(a)/a$. But $N_{K/\mathbb{Q}}(a)$ is the product of the images of a in $\mathbb{C}$ by all embeddings of K into $\mathbb{C}$, therefore, $N_{K/\mathbb{Q}}(a)/a$ is also a product of images of a in $\mathbb{C}$, hence a product of algebraic integers, and thus an algebraic integer. Therefore, $a^{-1} \in \mathcal{O}_K$, which proves that a is a unit. □

Definition 2.1.1. Let K be a number field. The *group of units* U_K associated to a number field K is the group of elements of $\mathcal{O}_K$ that have an inverse in $\mathcal{O}_K$.

Theorem 2.1.1 (Dirichlet, 1846)**.** *Let K be an algebraic number field. The group U_K is the direct product of a finite cyclic group of roots of units with a free abelian group of rank $r+s-1$, where r is the number of real embeddings of K and s is the number of complex conjugate pairs of embeddings of K.*

Proof. See [2] □

2.2 Applications of Strassmann's theorem

In this section we shall now give three examples where we can apply Strassmann's theorem, from the previous chapter, to deduce information about diophantine equations. In all the three cases we derive a p-adic power series and then apply Strassmann's theorem to bound the number of solutions to the diophantine equation. Its range of application is, rather limited.

Example 2.2.1. We show that the Thue equation

$$X^3 + 6Y^3 = \pm 1, \tag{2.1}$$

where we are only interested in integer solutions of the form $(X, Y) \in \mathbb{Z}^2$, has only the trivial solutions $(X, Y) = (\pm 1, 0)$.
Firstly we consider the algebraic number field $K = \mathbb{Q}(\theta)$, where $\theta^3+6 = 0$. The reason why we choose this number field is because it is the one which springs immediately to mind in such a situation as we can write our diophantine equation as

$$N_{K/\mathbb{Q}}(X - \theta Y) = \pm 1. \tag{2.2}$$

The field K is a cubic number field with one real embedding and a single pair of complex conjugate embeddings, it therefore, by Dirichlet's unit theorem, has a single fundamental unit which is given by $1 + 6\theta + 3\theta^2$. Such a fundamental unit can be determined quite easily by either using the modern methods of computing such units or using a computer package to perform the calculation for you. It is clear that the only units of finite order in K are ± 1.
By considering the factorisation our Thue equation

$$(X - \theta Y)(X - \theta\omega Y)(X - \theta\omega^2 Y) = \pm 1, \tag{2.3}$$

where ω is a non-trivial root of unity. We see from the the unique factorization of the ideal $(X - \theta Y)\mathcal{O}_K$ that we must have

$$X - \theta Y = \pm(1 + 6\theta + 3\theta^2)^k. \tag{2.4}$$

We can then formally expand the right hand side of (2.4) as a power series in k using the Taylor series expansion, to obtain

$$\begin{aligned} X - \theta Y &= \pm\left(1 + 3(2\theta+\theta^2)k + \frac{9(2\theta+\theta^2)^2(k^2-k)}{2!} + \frac{27(2\theta+\theta^2)^3(k^3-3k^2+2k)}{3!} + \cdots\right) \\ &= \pm(1 + 3(\theta^2 k + 2\theta k) + 9(2\theta^2(k^2+k)) + 27(\ldots)). \end{aligned}$$

We then notice that X^3+6 is irreducible over $\mathbb{Q}_3$; this is because X^3+6 has a solution mod 3, namely 0, but it can not be extended to a solution mod 9. Thus it does not have a zero in $\mathbb{Q}_3$, and being of degree 3, it is therefore irreducible over $\mathbb{Q}_3$. We can then equate the coefficients of θ in equation (2.5). The coefficients of θ^2 then give us

$$0 = \pm(3k + 9(\ldots)). \tag{2.5}$$

From Strassmann's theorem we then deduce that there is only one 3-adic solution to the above 3-adic power series. But we already know one solution, namely, $k = 0$, which corresponds to our known solutions of the original equation. Hence $(X, Y) = (\pm 1, 0)$ are the only solutions.

Example 2.2.2. We now consider the Thue equation

$$X^3 + 2Y^3 = \pm 1. \tag{2.6}$$

This only has the integral solutions $(X, Y) = \pm(1, 0)$ and $\pm(1, -1)$, which we shall show now.

As in the previous example, we consider the algebraic number field $K = \mathbb{Q}(\theta)$, where $\theta^3 + 2 = 0$. In this field we again have one fundamental unit, namely $-1 - \theta$. By considering the factorization of $X^3 + 2Y^3$, this leads to the equation

$$X - \theta Y = \pm(-1-\theta)^k. \tag{2.7}$$

Unfortunately, applying the method given in the previous example above does not give us any p-adic power series to which we can apply Strassmann's theorem. What worked in the first example was that the fundamental unit was congruent to 1 modulo 3 and hence the power series in k which we obtained converged 3-adically.

By Fermat's little theorem we know that for every algebraic integer, α, of K and every coprime prime ideal, $\mathfrak{p}$, we have $\alpha^o \equiv 1 \pmod{\mathfrak{p}}$, where o divides $p^{f_\mathfrak{p}} - 1$. By raising α^o to the p^t where $p^t > e_\mathfrak{p}$ we obtain, as we previously did in the previous chapter, that

$$\alpha^{op^t} \equiv 1 \pmod{p}. \tag{2.8}$$

In our example if we consider the prime ideal lying above (3), which completely ramifies, using the fact that $\theta^3 = -2$ we see that

$$(-1-\theta)^3 \;=\; -1-3\theta-3\theta^2-\theta^3 \;=\; 1-3\theta(1+\theta). \tag{2.9}$$

Hence we should consider the following three equations

$$X-\theta Y \;=\; \begin{cases} \pm(1-3\theta(1+\theta))^s & \text{if } k=3s, \\ \pm(1+\theta)(1-3\theta(1+\theta))^s & \text{if } k=1+3s, \\ \pm(1+\theta)^2(1-3\theta(1+\theta))^s & \text{if } k=2+3s. \end{cases} \tag{2.10}$$

We then expand the right hand side of these equations in (2.10) as a power series in s and then equate the coefficients of θ^2 as before to obtain three 3-adic power series in s which have to be zero for a solution to our original diophantine equation. The three power series are given by

$$0 \;=\; \begin{cases} 6s+9(\ldots) & \text{if } k=3s, \\ 6s+9(\ldots) & \text{if } k=1+3s, \\ 1+9(\ldots) & \text{if } k=2+3s. \end{cases} \tag{2.11}$$

We therefore deduce that there is at most one solution, s, to the first two 3-adic power series equations and there is no solution to the third equation. By inspection we see that our original diophantine equation has a solution when $k=0$ and $k=1$. Hence these two solutions must be the only solutions. Hence the only solutions are given by $(X,Y) = \pm(1,0)$ and $\pm(1,-1)$.

Example 2.2.3. In this example Strassmann's theorem will also show us where to look for a solution as well. We shall show that the only solutions to the Thue equation

$$X^3+6XY^2-Y^3 \;=\; \pm 1 \tag{2.12}$$

are given by $(X,Y) = \pm(1,0)$, $\pm(0,1)$ and $\pm(1,6)$.
To see this we consider the algebraic number field $K = \mathbb{Q}(\theta)$, where $\theta^3+6\theta-1=0$. In K there is one fundamental unit given by θ. We also notice that

$$\theta^3 \;=\; (1-6\theta) \;\equiv\; 1 \pmod 3$$

and that there is only one ramified prime ideal lying above 3. We look at the three 3-adic power series, given by setting $a = 0$, 1 or 2 in the equation below

$$\begin{aligned} X-\theta Y \;&=\; \theta^k \;=\; \theta^a(1-6\theta)^s, \\ &=\; \theta^a(1-6\theta s+18\theta^2 s^2+27(\ldots)) \end{aligned}$$

from which we deduce that there are at most six solutions; two when $k \equiv 1 (\text{mod } 3)$ and four when $k \equiv 0 (\text{mod } 3)$. We easily find the solutions $k = 0, 1$ which correspond to $(X, Y) = \pm(1, 0), \pm(0, 1)$. The other two solutions must lie in the family $k \equiv 0 (\text{mod } 3)$ which suggests that we look at $k = \pm 3, \pm 6, \ldots$. Fortunately, we find the final two solutions at $k = 3$.

Example 2.2.3 shows how we can use p-adic arguments to locate solutions as well as the bound on the number of actual solution. From these examples it appears that the method works for all examples of cubic Thue equations of negative discriminant. This is however rather optimistic. It also appears from the above examples that we need to use primes for which there is only one prime ideal lying above it. This is not true in general but using such primes makes the presentation neater. For more general primes one needs to decide on which prime ideal to choose and then find a p-adic power series which must be zero for the solution to exist. We cannot just equate coefficients of θ^2 in the general case. We can however find a suitable p-adic power series by, for instance, using Siegel's identity which we discuss in the next section.

2.3 Skolem's method

In the previous section we saw how, if we could produce a p-adic power series in one variable, we could bound the number of solutions to a diophantine equation. However, we could have to be dealing with very small problems for the above method to work all the time. An obvious extension would be to generalize the method to the case when we obtain a power series in many variables. In such a situation we will require many power series as well. The idea behind this solution method, often called Skolem's method, is to generalize Hensel's lemma rather than Strassmann's theorem. Then after a finite amount of 'sieving' we can hopefully locate all the solutions. In any case we will atleast obtain an upper bound on the number of solutions if this method works.

This method dates back to Skolem and his school in the 1930's [8]. Until the 1980's it was the main method used to solve many diophantine equations [8]. However, we shall see later than the modern methods and Skolem's method often share the sieving process in common. The sieving process will turn out to be the major bottleneck. Hence from a computational point of view Skolem's method, when it works, is often no worse than modern methods. We explain this method with the following example.

Example 2.3.1 ([8])**.** We shall now consider that the Thue equation,

$$X^4 - 2Y^4 \;=\; \pm 1, \tag{2.13}$$

has at most 12 integer solutions. To study this equation we first have to consider the quartic number field $K = \mathbb{Q}(\theta)$, where $\theta^4 - 2 = 0$. The unit rank of the ring of integers is two and we can take as a pair of fundamental units the elements

$$\eta_1 \;=\; 1 + \theta^2, \quad \eta_2 \;=\; 1 + \theta. \tag{2.14}$$

We therefore have to determine all possible pairs a_1, a_2 to the equation below

$$X - \theta Y \;=\; \beta \;=\; \pm \eta_1^{a_1} \eta_2^{a_2}. \tag{2.15}$$

The smallest prime number which stays prime in K is 5 and in the residue field the image of η_1 has order 12 and the image of η_2 has order 312, indeed we have

$$\begin{aligned} \eta_1^{12} &= (1+\theta^2)^{12} &&= 1 + 5 \cdot 2\theta^2 + 5^2(\ldots), \\ \eta_2^{312} &= (1+\theta)^{312} &&= 1 + 5(4\theta^2 + 3\theta^3) + 5^2(\ldots). \end{aligned}$$

Therefore, we could equate the coefficients of θ^2 and θ^3 in the identity

$$X - \theta Y \;=\; \beta \;=\; \pm \eta_1^{b_1} \eta_2^{b_2} (1 + (\eta_1^{12} - 1))^{k_1} (1 + (\eta_2^{312} - 1))^{k_2} \tag{2.16}$$

to find two power series in the two variables k_1 and k_2. However, we would have to do this for all possible values of the b_i which range $0 \leq b_1 \leq 11$, $0 \leq b_2 \leq 312$. Hence this looks rather an unpromising situation.
We instead notice that over the algebraic closure of $\mathbb{Q}$ we have four equations of the form

$$X - \theta_i Y \;=\; \beta_i, \quad \text{for } i = 1, 2, 3, 4 \tag{2.17}$$

which correspond to the four roots of our polynomial $X^4 - 2$. Eliminating X and Y from these four equations gives us two equations for the β_i, namely

$$(\theta_i - \theta_2)\beta_1 + (\theta_1 - \theta_i)\beta_2 + (\theta_2 - \theta_1)\beta_i \;=\; 0, \quad \text{for } i = 3, 4. \tag{2.18}$$

This last equation, (2.18), is often referred to as Siegel's identity. Now the prime 7 decomposes in the field K as a product of three prime ideals, one of which has degree 2 and two of degree 1. That is, in modulo 7 the polynomial $x^4 - 2$ factorizes as a product of two linear and one quadratic polynomial:

$$x^4 - 2 \;\equiv\; (x+2)(x+5)(x^2+4) \pmod 7, \tag{2.19}$$

as 7 is not an index divisor. We take θ_1, θ_2 to be the 7-adic roots of $x^4 - 2$ given by

$$\theta_1 \;=\; 2 + 3 \cdot 7 + 7^2(\ldots), \quad \theta_2 \;=\; 5 + 3 \cdot 7 + 7^2(\ldots). \tag{2.20}$$

We then take $\theta_3 = \Omega$ and $\theta_4 = \Omega'$ to be the roots of the 7-adic polynomial

$$g(x) = \frac{(x^4 - 2)}{(x - \theta_1)(x - \theta_2)} = x^2 + 4 + 5 \cdot 7 + 4 \cdot 7^2 + 0 \cdot 7^3 + 5 \cdot 7^4 + O(7^5). \tag{2.21}$$

In the two degree-one 7-adic localizations of K the elements η_i both satisfy:

$$\eta_i^6 \;\equiv\; 1 \pmod 7. \tag{2.22}$$

In the degree-two 7-adic localization of K we find that the η_i satisfy:

$$\eta_1^6 \;\equiv\; 1 \pmod 7, \quad \eta_2^{48} \;\equiv\; 1 \pmod 7. \tag{2.23}$$

We write $a_1 = b_1 + 6k_1$ and $a_2 = b_2 + 48k_2$. We first need to determine which values of $0 \leq b_1 \leq 5$ and $0 \leq b_2 \leq 47$ solve the following congruences, which come from Siegel's identity:

$$(\theta_i - \theta_2)\eta_1^{(1)b_1}\eta_2^{(1)b_2} + (\theta_1 - \theta_i)\eta_1^{(2)b_1}\eta_2^{(2)b_2} + (\theta_2 - \theta_1)\eta_1^{(i)b_1}\eta_2^{(i)b_2} = 0 \pmod 7 \tag{2.24}$$

for $i = 3, 4$. To do this we we need only loop through the 6 possibilities for the b_i and test these in the previous equation. We first find that there are 6 possible pairs (b_1, b_2) given by

$$(b_1, b_2) \;=\; (0,0), (0,1), (2,23), (3,24), (3,25), (5,47).$$

Then given these solutions we need to expand the equations in (2.18) as two 7-adic power series in the variables k_1, k_2. We obtain the following 7-adic power series f_1 and f_2 in each of our six cases:

1. $b_1 = 0 = b_2$.

$$\begin{aligned} f_1 &= 5k_1 + k_2 + 6\Omega k_2 + 7(\ldots), \\ f_2 &= 5k_1 + k_2 + \Omega k_2 + 7(\ldots), \end{aligned}$$

2. $b_1 = 0, b_2 = 1$.

$$\begin{aligned} f_1 &= 5k_1 + 6k_2 + 5\Omega k_2 + 7(\ldots), \\ f_2 &= 5k_1 + 6k_2 + 2\Omega k_2 + 7(\ldots), \end{aligned}$$

3. $b_1 = 2, b_2 = 23$.

$$\begin{aligned} f_1 &= 4 + 5k_1 + 3k_2 + \Omega(2k_1 + 5k_2) + 7(\ldots), \\ f_2 &= 4 + 5k_1 + 3k_2 + \Omega(5k_1 + 2k_2) + 7(\ldots), \end{aligned}$$

4. $b_1 = 3, b_2 = 24$.

$$\begin{aligned} f_1 &= 4 + 2k_1 + 6k_2 + \Omega(4 + k_2) + 7(\ldots), \\ f_2 &= 4 + 2k_1 + 6k_2 + \Omega(3 + 6k_2) + 7(\ldots), \end{aligned}$$

5. $b_1 = 3, b_2 = 25$.

$$\begin{aligned} f_1 &= 5 + 2k_1 + k_2 + \Omega(1 + 2k_1) + 7(\ldots), \\ f_2 &= 5 + 2k_1 + k_2 + \Omega(6 + 5k_1) + 7(\ldots), \end{aligned}$$

6. $b_1 = 5, b_2 = 47$.

$$\begin{aligned} f_1 &= 6 + 2k_1 + 4k_2 + \Omega(5k_1 + 2k_2) + 7(\ldots), \\ f_2 &= 6 + 2k_1 + 4k_2 + \Omega(2k_1 + 5k_2) + 7(\ldots), \end{aligned}$$

In each of the above six cases we apply the multi-dimensional Hensel's theorem, Theorem 1.2.3, to find that in each case there is exactly one possible solution in $\mathbb{Q}_7^2$. As every one of these cases corresponds to two solutions of our Thue equation, 2.13, where we have an upper bound on the number of solutions as 12. A simple search reveals six solutions given in the table below:

b_1	b_2	X	Y
0	0	1	0
0	0	-1	0
0	1	1	-1
0	1	-1	1
5	47	1	1
5	47	-1	-1

Hence there could exist another six possible solutions. To eliminate, or find, these one could either use another prime or apply some of the other modern methods. However, our method has atleast told us that the remaining solution, if they exist at all, lie in the following three families

$$\begin{aligned} &(i) \quad a_1 \equiv 2 \pmod 6, \quad a_2 \equiv 23 \pmod{48} \\ &(ii) \quad a_1 \equiv 3 \pmod 6, \quad a_2 \equiv 24 \pmod{48} \\ &(iii) \quad a_1 \equiv 3 \pmod 6, \quad a_2 \equiv 25 \pmod{48} \end{aligned}$$

We also know that each family can only contain at most one pair of solutions. This idea of finding congruence conditions on the exponents of identities satisfied by the solutions of diophantine equations is discussed more when studying sieving an S-unit equation [8]. We can treat the first part of the above method as 'sieving' out the six possible families (b_1, b_2) from the 6×48 possible families. Hence although Skolem's method has not worked, using the prime 7, it has given us information which we could use in the more advanced modern methods [8].

2.4 Hasse's principle

Sometimes we use local methods to show the non-existence of solutions to certain diophantine equations. Every local field, for instance $\mathbb{R}$, $\mathbb{Q}_p$ or $K_{\mathfrak{p}}$, contains a copy of $\mathbb{Q}$. Therefore, if a solution to a given equation exists then there is a solution in every such local completion. This often gives us a very easy way to check when a diophantine equation is soluble.

Example 2.4.1. As a trivial example, we consider the equation

$$X^2 + Y^2 = -4. \tag{2.25}$$

This equation trivially has no real solutions. It therefore has no rational solutions and hence has no integer solutions.

Example 2.4.2. As another example we consider the equation

$$X^2 - 13Y^2 = 5. \tag{2.26}$$

This equation has no solutions in $\mathbb{Q}_5$ because the congruence $X^2 \equiv 3 \pmod 5$ has no solutions. Hence this equation also has neither rational solutions nor integral solutions.

The previous two examples are special cases of the projective plane curve of degree 2, that is, a ternary quadratic form:

$$aX^2 + bY^2 + cZ^2 = 0. \tag{2.27}$$

Given $a, b, c \in \mathbb{Z}$, we are interested in determining whether such an equation has a solution in relatively prime integers (X, Y, Z) which are not simultaneously equal to zero. Trivially, we need to first check whether it has a solution in $\mathbb{Q}_p$ for every prime number p. A detailed procedure of how to do this is discussed in Chapter IV of [8]. It therefore turns out that this is all we need to do. The following result is due to Hesse.

Theorem 2.4.1 (Hasse [8]). *The equation*

$$aX^2 + bY^2 + cZ^2 = 0. \tag{2.28}$$

has a non-trivial solution in $\mathbb{Z}^3$ if and only if it has a non-trivial solution in $\mathbb{Q}_p^3$ for every prime p (including ∞).

Proof. This is a standard result which can be found in many books which discuss quadratic forms. For the proof, for example, see [3]. □

One should note that this result, Theorem 2.4.1, gives a little more than what is really required. It can be shown that the number of primes (including ∞) for which the equation (2.28) is not locally soluble is always even. Given the above example we have the following definition.

Definition 2.4.1. A diophantine equation is said to satisfy the *Hasse principle* if the existence of rational (global) solutions is guaranteed by the existence of p-adic (local) solutions for every prime (including ∞).

Therefore, the Hasse principle is also often called the *local-global principle*. We see that the equations of the Hasse's theorem satisfy the Hasse principle. However, we are not always lucky. Consider the following example.

Example 2.4.3. The standard example of an equation which does not satisfy the Hasse principle is

$$3X^3 + 4Y^3 + 5Z^3 = 0. \tag{2.29}$$

This equation is due to Selmer [7]. It has no rational solutions but has local solutions in every p-adic field. A detailed study of the failure of the Hasse principle is discussed after one studies elliptic curves which we shall not discuss in this research project.

2.5 Finding small solutions

Sometimes when one solves a diophantine equation one has a bound on the solution space or one is only interested in 'small' solutions. It would be nice if there was a fast method to locate all the solutions up to any given bound. In terms of chapter 2 language we wish to determine all the solutions with bounded logarithmic height. We could just run through all possibilities checking each one in turn. We shall call this the naive method. It is easy to see that this naive method applied to an equation

in two variables would take atleast $O(e^{2B})$ operations, if B were the bound on the logarithmic height. For further details about the naive method, see [8].
We use the information gathered from considering small prime numbers to remove large numbers of non-solutions from consideration. In other words we look at where the solutions could be locally, using modulo p or p-adic arguments. This local information is then put together to deduce information about the location of global solutions. We eliminate as many solutions as possible at the first stage using a single prime. The remaining possible solutions are passed to a second stage where they are checked modulo a different prime q and so on. At each stage one has sieved out a large number of non-solutions.

Example 2.5.1. For instance, suppose we wish to find all rational solutions to the equation

$$Y^2 = aX^4 + bX^3 + cX^2 + dX + e \tag{2.30}$$

with $h(X) \leq B$, where B is some given constant. Firstly we homogenize equation (2.30) by writing X as N/M with $N, M \in \mathbb{Z}$, $(N, M) = 1$ and $M \geq 1$. We then know we must find all solutions to the equation

$$\begin{aligned}(M^2Y)^2 &= F(M,N) \\ &= aN^4 + bMN^3 + cM^2N^2 + dM^3N + eM^4, \quad \max\{|N|, M\} \leq e^B.\end{aligned}$$

Historically we believe that for a given, random prime p the expression $F(M, N)$, for any given M and N, is a square modulo p about half the time. Therefore, looking modulo p, for a single prime number p, will hopefully eliminate half of the solution space.

We define a sieving procedure as follows:

---

[8] **Recussive algorithm for sieving a curve of the form** $Y^2 = F(X)$

`DESCRIPTION:` Sieve (M, N, R):
Find all solutions to the equation $Y^2 = F(X)$
with $h(X) \leq B$.
`INPUT:` $M_0, N_0, R \in \mathbb{Z}$
`OUTPUT:` Solutions to equation (2.30) with $h(X) \leq B$ such that $X = N/M$ with $N - N_0 \equiv M - M_0 \equiv 0 \pmod{R}$.

1. Choose the smallest prime, p, such that $\gcd(p, R) = 1$.
2. For $M_1 = M_0$ to pR step R do
3. For $N_1 = N_0$ to pR step R do
4. If M_1 and N_1 are both divisible by p then
5. If $F(M_1, N_1)$ is a square modulo p then
6. If $pR > 2e^B$ then
7. Check if N_1/M_1 or $(N_1 - pR)/M_1$ really
   is a solution and if so print it.
8. Else
9. Call Sieve (M_1, N_1, pR)
10. Endif
11. Endif
12. Endif
13. Enddo
14. Enddo.

This sieving procedure is called via Sieve$(0,0,1)$. It works in a recursive way by assuming we have a solution (M_0, N_0) modulo R. It then lifts this solution to a new solution modulo pR, where p is a prime such that $(p,R) = 1$. For every solution modulo pR found it calls itself again until the current modulus is greater than $2e^B$. When the current modulus is greater than $2e^B$, the current values are tested to see if they are really global solutions. The method is therefore essentially a *depth first* strategy.

It is probably best for every small primes p to use either prime powers or composite moduli in the loops rather than just the prime p itself. An alternative approach would be to only use primes larger than 5, for instance. Essentially we have combined local information using a Chinese remainder process to obtain information about possible solutions up to the desired bound. However, here, as each prime taken was coprime to the current modulus, the Chinese remaindering needed was trivial.

It remains to discuss how much faster such a sieving technique will be. We first note that

$$\theta(x) \;=\; \sum_{p \le x} \log p \;\approx\; O(x), \tag{2.31}$$

hence the largest prime we need to take is of size $\log e^B = B$, and there are roughly $B/\log B$ primes less than B in size. At each step we eliminate roughly half of the cases modulo p; hence the complexity can be estimated by

$$\begin{aligned}
\text{Time} &\approx p_1^2\left(1+\frac{p_2^2}{2}\left(1+\frac{p_2^3}{2}\left(1+\cdots\right)\right)\right),\\
&\le B^2\left(1+\frac{B^2}{2}\left(1+\frac{B^3}{2}\left(1+\cdots\right)\right)\right),\\
&= \sum_{i=1}^{B/\log B}\frac{B^{2i}}{2^{i-1}},\\
&= 2B^2\frac{\left((B^2/2)^{B/\log B}-1\right)}{B^2-2}, \quad \text{using sum of a G.P,}\\
&= O\left((B^2/2)^{B/\log B}\right).
\end{aligned}$$

Therefore, we see that using sieving gives a slightly better running time than the naive method. Clearly whether sieving is better in practice than the naive method would depend on the implied constants which arise from the implementation. In addition the above analysis of the sieving method has been very pessimistic so as to make the formulae easier to handle [8].

References

[1] A. J. Baker. An introduction to p-adic numbers and p-adic analysis. *University of Glasgow, Internetskript*, 2011.

[2] H. Bass. The Dirichlet unit theorem, induced characters, and Whitehead groups of finite groups. *Topology*, 4(4):391–410, 1966.

[3] J.W.S. Cassels and M.J.T. Guy. On the Hasse principle for cubic surfaces. *Mathematika*, 13(02):111–120, 1966.

[4] B. M. M. de Weger. Algorithms for diophantine equations, CWI-Tract No. 65, Centre for Math. and Comp. *Sci., Amsterdam*, 1989.

[5] M. J. Greenberg. *Lectures on forms in many variables*, volume 31. WA Benjamin New York-Amsterdam, 1969.

[6] B. S. Schmidt. Solutions to systems of multivariate p-adic power series. Master's thesis, Mathematical Institute, University of Oxford, UK, 2015.

[7] E. S. Selmer. The diophantine equation $ax^3 + by^3 + cz^3 = 0$. *Acta Mathematica*, 85(1):203–362, 1951.

[8] N. P. Smart. *The algorithmic resolution of Diophantine equations: a computational cookbook*, volume 41. Cambridge University Press, 1998.