Utrecht University

Double Bachelor's Thesis
Mathematics and Physics

Elliptic Curves and the Yang-Baxter Equation

Author:
Corijn Rudrum

Supervisors:
dr. Lennart Meier
dr. Dirk Schuricht

June 11, 2020

Abstract

The two main topics of this thesis are elliptic curves and the Yang-Baxter equation, whose common theme is their connection to elliptic functions. Elliptic curves, on the one hand, are algebraic objects with applications in for example number theory and cryptography, and the subject of much current mathematical research. We prove a number theoretical result due to Gauss about a specific elliptic curve, and we show that elliptic curves can be parameterised using elliptic functions. The Yang-Baxter equation, on the other hand, comes up when studying the scattering of identical particles in one dimensional integrable systems. We derive a solution for this equation using the Jacobi elliptic functions, and we discuss how this solution can be used to show that the one-dimensional XYZ Heisenberg model has infinitely many conservation laws.

Acknowledgements

First and foremost, I would like to thank my supervisors dr. Lennart Meier and dr. Dirk Schuricht, who were always available for questions and took the time to explain things I did not understand. Secondly, I would like to thank all the friends who have helped me with all my other questions, specially Joppe Stokvis, Dennis Hilhorst and Artim Bassant. Lastly, I would like to thank Casper van Hal, for being genuinely interested and wanting to read my thesis before it was even finished. I could not think of a better motivation for writing this thesis.

Introduction

The study of Diophantine equations, that is finding integer or rational solutions of polynomial equations, has interested mathematicians for over two thousand years, and is still an active area of research. One of the appealing things about Diophantine equations, is that one can ask questions that are easy to understand, but very hard to answer. For example, consider the statement that the equation

$$X^n + Y^n = Z^n \tag{0.0.1}$$

has no integer solutions where X, Y and Z are all non-zero for $n \geq 3$. This was conjectured by Fermat in the seventeenth century, but even though the statement can be understood by a high school student, it took more than three hundred years before the British mathematician Andrew Wiles managed to prove that it was in fact true. The statement is known as Fermat's Last Theorem. One of the two main topics of this thesis are elliptic curves, which, as we will see, are closely related to cubic Diophantine equations in two variables. Much current research is done on elliptic curves, and they have applications in fields like number theory and cryptography. In fact, in his proof of Fermat's Last Theorem, Andrew Wiles made ingenious use of elliptic curves.

The other main topic of this thesis is the Yang-Baxter equation. This relation comes up when studying the scattering of particles in one dimension, and ensures that we can decompose the scattering of multiple particles into a series of two-particle scatterings. A very nice application of the Yang-Baxter equation is in the quantum inverse-scattering method. Most of the time in physics, we define a system and then try to derive properties of this system. For example when we study a pendulum, we first write down the Hamiltonian or Lagrangian for the system, and from there we derive the equations of motion for the pendulum. However, in the quantum inverse-scattering method we turn this process around. It turns out that we can show that if we find a solution to the Yang-Baxter equation, then we can use this solution to find some Hamiltonian of a system with infinitely many conservation laws. So then we know that we are going to find a system with infinitely many conservation laws, even before we know what system it will be.

The aim of this thesis is to give an introduction to both elliptic curves and the Yang-Baxter equation, but also to show that these topics are not completely unrelated. It turns out that both elliptic curves and the Yang-Baxter equation are related to so-called elliptic functions. We will connect elliptic curves and the Yang-Baxter equation by exploring their links to elliptic functions.

Outline of the Thesis

In the first chapter, we will give an introduction to the projective plane and projective curves. This is all set-up for the second chapter, where we introduce elliptic curves. After defining what elliptic curves are, we will show how the points on an elliptic curve in so-called Weierstrass normal form together form a group. In Chapter 3 we will prove a number theoretical result originally proven by Gauss. Subsequently, we will use this result to determine the group structure of a specific elliptic curve. Chapter 4 forms the connection between the first three chapters and the last two. In this chapter we will introduce elliptic functions and show how they can be used to parameterise elliptic curves in Weierstrass normal form. Further, we will introduce the Jacobi elliptic functions, which we will use to solve the Yang-Baxter equation in the subsequent chapters. In this chapter we also explain where the name ‘elliptic’ comes from. In Chapter 5 we derive the Yang-Baxter equation by looking at the scattering of identical particles in a one-dimensional system. Lastly, in Chapter 6 we will give a solution of the Yang-Baxter equation using the Jacobi elliptic functions from Chapter 4. Further, we will use this solution to construct the Hamiltonian of a system with infinitely many conservation laws via the quantum inverse-scattering method. The Hamiltonian we construct is that of the one-dimensional XYZ Heisenberg model.

Contents

Chapter 1

Projective Geometry

In this chapter we will introduce the projective plane, following Appendix A from Silverman and Tate [1]. We only consider the properties that we need in the rest of this thesis.

1.1 A Motivating Example

We start with a famous problem from number theory. We want to find the solutions (x, y) of the equation

$$x^n + y^n = 1, \tag{1.1.1}$$

with $x, y \in \mathbb{Q}$ and $n \geq 3$. We call such a solution non-trivial if both x and y are non-zero. Suppose that we have a solution $x = a/c$ and $y = b/d$, with $a, b \in \mathbb{Z}$ and $c, d \in \mathbb{Z} \setminus \{0\}$. Then we can assume the fractions to be simplified such that c and d are both positive, and $\gcd(a, c) = \gcd(b, d) = 1$. Here we write gcd() for the greatest common divisor of two integers. Filling in this solution, we can rewrite Equation (1.1.1) to

$$a^n d^n + b^n c^n = c^n d^n. \tag{1.1.2}$$

This implies that $c^n \mid a^n d^n$. By assumption we have $\gcd(a, c) = 1$, so we get $c^n \mid d^n$, and therefore $c \mid d$. Similarly, we have $d^n \mid b^n c^n$ and $\gcd(b, d) = 1$, so $d \mid c$. Hence, we find $c = \pm d$. We assumed c and d to be both positive, so then $c = d$. Therefore every solution of Equation (1.1.1) is of the form $(a/c, b/c)$. Replacing d by c, we can divide Equation (1.1.2) by c^n to get

$$a^n + b^n = c^n. \tag{1.1.3}$$

So we find that every non-trivial rational solution $(a/c, b/c)$ of Equation (1.1.1) gives us a non-trivial integer solution (a, b, c) of

$$X^n + Y^n = Z^n. \tag{1.1.4}$$

Here non-trivial means that X, Y and Z are all non-zero. Conversely, if we have a non-trivial integer solution (a, b, c) for Equation (1.1.4) with $c \neq 0$, then $(a/c, b/c)$ is a non-trivial solution of Equation (1.1.1). Now, one might think that we have a one-to-one correspondence between the non-trivial solutions of Equation (1.1.1) and the non-trivial solutions of Equation (1.1.4). This is in fact not true, because different integer solutions (a, b, c) may lead to the same rational solution $(a/c, b/c)$. Indeed, suppose that we have an integer solution (a, b, c) with $c \neq 0$ of Equation (1.1.4). Then for every non-zero integer t, the triple (ta, tb, tc) is also a non-trivial integer solution. But we have $(ta/tc, tb/tc) = (a/c, b/c)$, so their corresponding solutions of Equation

(1.1.1) are the same. We can solve this problem by looking at equivalence classes of solutions of Equation (1.1.4), where we identify solutions (a, b, c) and (a', b', c') if there exists a non-zero integer t such that $(a', b', c') = (ta, tb, tc)$. In that case we do have a one-to-one correspondence between non-trivial rational solutions of Equation (1.1.1) and equivalence classes of non-trivial integer solutions of Equation (1.1.4) with $Z \neq 0$.

Our original question was to find non-trivial rational solutions of Equation (1.1.1) for $n \geq 3$, but now we know that this is equivalent to finding (equivalence classes of) non-trivial integer solutions of Equation (1.1.4). We call Equation (1.1.4) the *homogenization* of Equation (1.1.1), which we will explain further in Section 1.3. You may recognize Equation (1.1.4) from Fermat's Last Theorem, FLT. This theorem states that for $n \geq 3$, there exist no non-trivial solutions to Equation (1.1.4)[1]. Therefore we see that for $n \geq 3$, there exist no non-trivial rational solutions of Equation (1.1.1).

Now one could ask: What about the solutions of Equation (1.1.4) where $Z = 0$? Indeed, for odd n we also have the integer solutions $(t, -t, 0)$ for $t \in \mathbb{Z} \setminus \{0\}$, but these do not correspond to any solutions of Equation (1.1.1). To understand where they come from, consider an infinite sequence of solutions

$$(a_0, b_0, c_0),\ \ (a_1, b_1, c_1),\ \ (a_2, b_2, c_2),\ \ \ldots \tag{1.1.5}$$

such that $c_i \neq 0$ for all $i \in \mathbb{N}$, and

$$\lim_{i \to \infty} (a_i, b_i, c_i) = (t, -t, 0) \tag{1.1.6}$$

for some $t \in \mathbb{Z} \setminus \{0\}$. Here the triples (a_i, b_i, c_i) consist of real numbers that form a solution of Equation (1.1.4). As $c_i \neq 0$, we have for each triple a corresponding real solution $(a_i/c_i, b_i/c_i)$ of Equation (1.1.1). From Equation (1.1.6) we see that $(a_i/c_i, b_i/c_i)$ approaches $(\infty, -\infty)$ as $i \to \infty$. So somehow the solutions $(t, -t, 0)$ of Equation (1.1.4), i.e. the equivalence class of $(1, -1, 0)$, correspond to a solution of Equation (1.1.1) that lies "at infinity". In the rest of this chapter we will see that the theory of solutions of polynomial equations becomes more elegant if we include these solutions "at infinity", which is exactly what we do in projective geometry.

1.2 The Projective Plane $\mathbb{P}^2$

There are multiple ways in which you can interpret the projective plane. Here, we will give an algebraic definition of a projective space. After this definition, we will present a more geometrical way to look at the two-dimensional projective space, also called the projective plane.

Let K be a field. If you are not familiar with fields as algebraic objects, you should think of K as for example the real numbers $\mathbb{R}$, complex numbers $\mathbb{C}$, or rational numbers $\mathbb{Q}$. We define the set

$$S_n = \{(a_0, \ldots, a_n) \mid a_0, \ldots, a_n \in K, (a_0, \ldots, a_n) \neq (0, \ldots, 0)\} \tag{1.2.1}$$

as the set of all $n + 1$-tuples of elements in K except for $(0, \ldots, 0)$. We define the equivalence relation $\sim_n$ on S_n as follows: $(a_0, \ldots, a_n) \sim_n (a'_0, \ldots, a'_n)$ if and only if there exists a non-zero $t \in K$ such that $a'_0 = ta_0,\ \ldots,\ a'_n = ta_n$. We denote the equivalence class of a tuple $(a_0, \ldots, a_n)$ by $[a_0, \ldots, a_n]$. Now, we define the n-dimensional projective space as follows.

[1] The first person to state this theorem was Fermat, writing it down in the margin of a book. He also added that he had found a truly marvelous proof for this, but sadly it did not fit in the margin of the page. For more than three hundred years the theorem remained unproven, until in 1995 a correct proof was given by Andrew Wiles. In fact, elliptic curves played a major role in his proof.

Definition 1.2.1. *Let K be a field, and let the set S_n and relation $\sim_n$ be as described above for some integer $n \geq 1$. Then we define **the n-dimensional projective space** $\mathbb{P}^n(K)$ as the set of equivalence classes $[a_0, \ldots, a_n]$ of tuples $(a_0, \ldots, a_n)$ in S_n. In formula:*

$$\mathbb{P}^n(K) = S_n / \sim_n . \tag{1.2.2}$$

*We call the $a_0, \ldots, a_n$ **homogeneous coordinates** for the point $[a_0, \ldots, a_n]$.*

In this thesis we will mostly be interested in the two-dimensional projective space $\mathbb{P}^2(K)$, called the *projective plane*. There each point $[a, b, c] \in \mathbb{P}^2(K)$ has three homogeneous coordinates $a, b, c \in K$.

Now, we want to be able to do geometry in the projective plane. In the next section we will define curves in $\mathbb{P}^2(K)$ in general, but here we already give the definition of a *line*.

Definition 1.2.2. *Let $L \subset \mathbb{P}^2(K)$ be a subset of the projective plane. We say that L is a **line** in $\mathbb{P}^2(K)$ if there exists an equation of the form*

$$\alpha X + \beta Y + \gamma Z = 0 \tag{1.2.3}$$

for some constants $\alpha, \beta, \gamma \in K$ not all zero, such that for every point $[a, b, c] \in L$ we have that $(X, Y, Z) = (a, b, c)$ is a solution of this equation, and every solution of this equation corresponds to a point in L.

Note that if a triple (a, b, c) is a solution of Equation (1.2.3), then the triple (ta, tb, tc) for $t \neq 0$ is also a solution. So in our definition of a line, it doesn't matter which representant of $[a, b, c] \in L$ we choose. Therefore everything is indeed well-defined.

From the above definitions, it is not immediately clear what the projective plane and its lines actually look like. It turns out that there is a one-to-one correspondence between the projective plane $\mathbb{P}^2(K)$ and the set $K^2 \cup \mathbb{P}^1(K)$. We will call K^2 the *affine part* of the projective plane, and the set $\mathbb{P}^1(K)$ we call the set of *points at infinity*.

In K^2, we know that lines are given by equations of the form

$$\alpha y = \beta x + \gamma. \tag{1.2.4}$$

It would be nice if the lines in $\mathbb{P}^2(K)$ would correspond to lines in K^2, with perhaps some extra points in $\mathbb{P}^1(K)$, i.e. points at infinity. To see what happens, we first make the correspondence between $\mathbb{P}^2(K)$ and $K^2 \cup \mathbb{P}^1(K)$ more precise. We define the bijective map $f : \mathbb{P}^2(K) \to K^2 \cup \mathbb{P}^1$ by

$$f([a, b, c]) = \begin{cases} \left(\frac{a}{c}, \frac{b}{c}\right) \in K^2 & \text{if } c \neq 0 \\ [a, b] \in \mathbb{P}^1(K) & \text{if } c = 0 \end{cases}. \tag{1.2.5}$$

To see that f is a bijection, we construct another map $g : K^2 \cup \mathbb{P}^1(K) \to \mathbb{P}^2(K)$ as

$$g(p) = \begin{cases} g_1(p) & \text{if } p \in K^2 \\ g_2(p) & \text{if } p \in \mathbb{P}^1(K) \end{cases} \tag{1.2.6}$$

Here the maps $g_1 : K^2 \to \mathbb{P}^2(K)$ and $g_2 : \mathbb{P}^1(K) \to \mathbb{P}^2(K)$ are given by

$$\begin{aligned} g_1((x, y)) &= [x, y, 1], \\ g_2([a, b]) &= [a, b, 0]. \end{aligned} \tag{1.2.7}$$

Then the maps f and g are inverses of each other. Indeed, if we let $a, b, c \in K$ with $c \neq 0$, then

$$g \circ f([a,b,c]) = g\left(\left(\frac{a}{c}, \frac{b}{c}\right)\right) = \left[\frac{a}{c}, \frac{b}{c}, 1\right] = [a,b,c]. \tag{1.2.8}$$

In the case $c = 0$ we get

$$g \circ f([a,b,0]) = g([a,b]) = [a,b,0]. \tag{1.2.9}$$

Conversely, if we let $(x,y) \in K^2$ then

$$f \circ g((x,y)) = f([x,y,1]) = (x,y), \tag{1.2.10}$$

and for $[a,b] \in \mathbb{P}^1(K)$ we find

$$f \circ g([a,b]) = f([a,b,0]) = [a,b]. \tag{1.2.11}$$

This proves that f and g are inverses of each other, and therefore f is indeed a bijection.

This way, we have found a one-to-one correspondence between the points in the projective plane $\mathbb{P}^2(K)$ and the points in $K^2 \cup \mathbb{P}^1(K)$. Now suppose that we have a line L in $\mathbb{P}^2(K)$. Then, according to Definition 1.2.2, there exist constants $\alpha, \beta, \gamma \in K$ not all zero such that the homogeneous coordinates of points in L satisfy the equation

$$\alpha X + \beta Y + \gamma Z = 0. \tag{1.2.12}$$

Suppose first that α and β are not both zero. Then if we have a point $[a,b,c] \in L$ with $c \neq 0$, we see that

$$f([a,b,c]) = \left(\frac{a}{c}, \frac{b}{c}\right). \tag{1.2.13}$$

This is a point on the line $\alpha x + \beta y + \gamma = 0$ in K^2. We call this line L'. Note that for every point in $(x,y) \in L'$ we have a corresponding point $[x,y,1] \in L$. There is exactly one point $[a,b,c] \in L$ with $c = 0$, namely the point $[-\beta, \alpha, 0]$. We find

$$f([-\beta, \alpha, 0]) = [-\beta, \alpha] \in \mathbb{P}^1(K). \tag{1.2.14}$$

So we see that L corresponds to a line L' in K^2 together with one extra point in $\mathbb{P}^1(K)$. We call L' the *affine part* of L. So, if we view L as a line in $K^2 \cup \mathbb{P}^1$, we see that it consists of an affine part L' and one point at infinity. By changing α, β and γ, while making sure that α and β are not both zero, we see that for every line L' in K^2 there exists a unique projective line $L \subset \mathbb{P}^2(K)$ such that L' is the affine part of L. So the point at infinity on a projective line is determined by the affine part of the line.

Now suppose that α and β are both 0. Then the coordinates of points in L satisfy the equation

$$Z = 0. \tag{1.2.15}$$

Therefore the points on L are precisely the points of the form $[a,b,0]$. From Equation (1.2.5) we see that these points correspond precisely to the set of points at infinity. In $K^2 \cup \mathbb{P}^1(K)$ we define the *line at infinity*, denoted by L_∞, as the set of all points at infinity. Then we see that every line in $\mathbb{P}^2(K)$ corresponds to either the line at infinity, or a unique line in K^2 together with one point at infinity.

The following proposition gives an interesting property of lines in the projective plane.

Proposition 1.2.3. *Let K be a field. Then any two distinct lines in the projective plane $\mathbb{P}^2(K)$ intersect in exactly one point, and for any two distinct points in $\mathbb{P}^2(K)$ there is a unique line through both points.*

Proof. For the proof of this proposition, see the first section of Appendix A of Silverman and Tate [1]. □

1.3 Curves in the Projective Plane

In the previous section we already gave the definition of a line in the projective plane. In this section we will generalize this to the definition of a curve.

Before looking at the projective plane, we first look at curves in K^2 for some field K.

Definition 1.3.1. *Let K be a field. We define an* ***algebraic curve*** *in K^2 as the set of solutions to a polynomial equation in two variables,*

$$f(x,y) = 0. \tag{1.3.1}$$

For example, the polynomial equation $x^2 + y^2 - 1 = 0$ defines a circle in K^2, and $2x - 3y^2 + 1 = 0$ gives a parabola. Circles and parabolas are therefore examples of algebraic curves.

Now, we want to have a similar definition for curves in the projective plane. We will first use our definition of $\mathbb{P}^2(K)$ with homogeneous coordinates. Then we will look what happens if we view the projective plane as $K^2 \cup \mathbb{P}^1(K)$. Using homogeneous coordinates, a point in the projective plane has three coordinates instead of two. Therefore we will have to use polynomials in three variables instead of two. However, one point in $\mathbb{P}^2(K)$ can be represented by different choices for the homogeneous coordinates. Therefore we only want to look at polynomials $F(X,Y,Z)$ with the property that $F(a,b,c) = 0$ implies that $F(ta,tb,tc) = 0$ for every non-zero $t \in K$. These polynomials turn out to be the *homogeneous polynomials*, given by the following definition.

Definition 1.3.2. *Let K be a field. A polynomial $F(X,Y,Z)$ in three variables is called a* ***homogeneous polynomial of degree*** *d, if and only if it is a linear combination of mononomials $X^iY^jZ^k$ with $i + j + k = d$.*

From this definition it follows that for a homogeneous polynomial $F(X,Y,Z)$ of degree d we have

$$F(tX,tY,tZ) = t^dF(X,Y,Z) \tag{1.3.2}$$

for any $t \in K$. Therefore if F is homogeneous we find that indeed $F(a,b,c) = 0$ implies $F(ta,tb,tc) = 0$ for every non-zero $t \in K$. Note that in Definition 1.3.2 the natural numbers i, j and k are allowed to be 0. For example, we see that the polynomial $X^2 + YZ + 5Z^2$ is homogeneous of degree 2, but the polynomial $2XZ^2 + 3Y^2$ is not homogeneous.

Now we can give the definition of a curve in the projective plane, which we will call a *projective curve.*

Definition 1.3.3. *Let K be a field. We define a* ***projective curve*** *C in the projective plane $\mathbb{P}^2(K)$ as the set of points whose homogeneous coordinates form a solution to a polynomial equation*

$$C : F(X,Y,Z) = 0, \tag{1.3.3}$$

where F is a non-constant homogeneous polynomial. If it is clear from the context that we are working with projective curves, we may also call C an algebraic curve or just curve.

By the *degree* of a projective curve we mean the degree of its corresponding homogeneous polynomial. For instance, the projective curve

$$C : X^2 + YZ + 5Z^2 = 0 \tag{1.3.4}$$

has degree 2.

We will now show that it is indeed well-defined whether a point in the projective plane lies on a given curve or not. Let $p \in \mathbb{P}^2(K)$ be a point in the projective plane. Suppose we have two triples of homogeneous coordinates representing the point p:

$$p = [a, b, c] \quad \text{and} \quad p = [a', b', c']. \tag{1.3.5}$$

Then there is a non-zero $t \in K$ such that $(a', b', c') = (ta, tb, tc)$. Let $C : F(X, Y, Z) = 0$ be a projective curve of degree d. Because F is a homogeneous polynomial, we find

$$F(a', b', c') = t^d F(a, b, c). \tag{1.3.6}$$

Therefore we see that $F(a, b, c) = 0$ if and only if $F(a', b', c') = 0$, which is exactly what we wanted to show.

To see what a projective curve looks like if we view the projective plane as $K^2 \cup \mathbb{P}^1(K)$, we use the identification of our two versions of $\mathbb{P}^2(K)$ given in the previous section. Let $C : F(X, Y, Z) = 0$ be a projective curve of degree d. Suppose we have a point $p = [a, b, c] \in C$ with $c \neq 0$. Then the point p corresponds to the point

$$\left(\frac{a}{c}, \frac{b}{c}\right) \in K^2 \subset K^2 \cup \mathbb{P}^1(K). \tag{1.3.7}$$

We know that F is homogeneous and $F(a, b, c) = 0$. Therefore we have

$$0 = \frac{1}{c^d} F(a, b, c) = F\left(\frac{a}{c}, \frac{b}{c}, 1\right). \tag{1.3.8}$$

We define the polynomial $f(x, y)$ by

$$f(x, y) = F(x, y, 1). \tag{1.3.9}$$

Then the subset of points (x, y) in K^2 for which $f(x, y) = 0$ is an algebraic curve in K^2. We call this curve C'. Using Equations (1.3.7) and (1.3.8) we see that for every point $p = [a, b, c] \in C$ with $c \neq 0$, its corresponding point in $K^2 \cup \mathbb{P}^1(K)$ lies on the curve $C' \subset K^2$. Conversely, if we have a point (x, y) on C', then $[x, y, 1]$ lies on the curve C. Therefore the points $[a, b, c]$ on C with $c \neq 0$ correspond to a curve C' in K^2. We call C' the *affine part* of C.

The points $[a, b, c]$ on C for which $c = 0$ correspond to points in $\mathbb{P}^1(K) \subset K^2 \cup \mathbb{P}^1(K)$, which we called points at infinity. We conclude that if we have a projective curve $C : F(X, Y, Z) = 0$, we can write it as the union of its affine part C' and its points at infinity. Here the affine curve C' is given by

$$C' : f(x, y) = F(x, y, 1) = 0. \tag{1.3.10}$$

The points at infinity on C are the points of the form $[a, b, 0] \in C$.

The process of turning a homogeneous polynomial $F(X, Y, Z)$ into the polynomial $f(x, y) = F(x, y, 1)$ is called *dehomogenization (with respect to the variable Z)*. This way we can find the affine part of a projective curve. Now, we would like to reverse this process. Given an algebraic curve C' in K^2, we want to find a projective curve that has the curve C' as its affine part.

Suppose that we have an algebraic curve C' in K^2 given by

$$C' : f(x,y) = 0. \tag{1.3.11}$$

To find the projective curve that has C' as its affine part, we must construct a homogeneous polynomial $F(X,Y,Z)$ such that $f(x,y) = F(x,y,1)$. We can write the polynomial $f(x,y)$ as

$$f(x,y) = \sum_{i,j} a_{ij} x^i y^j. \tag{1.3.12}$$

We define the *degree* of f as the largest value of $i+j$ for which a_{ij} is not zero, denoted by $\deg(f)$. For example

$$\deg(x^3 - 6xy + y + x) = 3, \quad \text{and} \quad \deg(y^3 - y^2x^2 + x) = 4. \tag{1.3.13}$$

Definition 1.3.4. *For a polynomial in two variables $f(x,y) = \sum_{i,j} a_{ij}x^iy^j$ with degree d, we define its* ***homogenization*** *$F(X,Y,Z)$ as*

$$F(X,Y,Z) = \sum_{i,j} a_{ij} X^i Y^j Z^{d-i-j}. \tag{1.3.14}$$

From this definition we see immediately that the homogenization $F(X,Y,Z)$ of a polynomial $f(x,y)$ of degree d is homogeneous of degree d, and $F(x,y,1) = f(x,y)$. Note that $F(x,y,1) = f(x,y)$ would also be the true if we defined $F(X,Y,Z)$ as

$$F(X,Y,Z) = \sum_{i,j} a_{ij} X^i Y^j Z^{k-i-j}, \qquad \text{for any } k > d. \tag{1.3.15}$$

This way $F(X,Y,Z)$ is also homogeneous, but of degree $k > d$. The problem with this definition is that then we would have $F(X,Y,0) = 0$, so the curve $C : F(X,Y,Z) = 0$ would contain all the points at infinity. To avoid this we defined the homogenization of a polynomial to be homogeneous of the same degree as the polynomial itself. Then the homogenization always contains a mononomial without the variable Z, and therefore $F(X,Y,0) \neq 0$. This way, for a given curve $C' : f(x,y) = 0$ in K^2, we find the unique curve $C : F(X,Y,Z) = 0$ that has affine part C' and does not contain all points at infinity. Note that by using homogenization and dehomogenization, we get a one-to-one correspondence between the algebraic curves in K^2 and the projective curves in $\mathbb{P}^2(K)$ that do not contain the whole line at infinity.

It is important to note that in principle there is nothing special about the variable Z. Given a homogeneous polynomial $F(X,Y,Z)$ we can dehomogenize it with respect to any of the variables X, Y or Z. For example dehomogenizing with respect to Y would give the polynomial $f(x,z) = F(x,1,z)$. If we have a projective curve $C : F(X,Y,Z) = 0$, then the set of points on C that lie at infinity depends on how we dehomogenize $F(X,Y,Z)$. For example, consider the projective curve

$$C : F(X,Y,Z) = Y^2Z - X^3 - Z^3 = 0 \tag{1.3.16}$$

and the point $p = [0,1,0] \in C$. If we dehomogenize $F(X,Y,Z)$ with respect to Z, then the point p corresponds to the point $[0,1] \in \mathbb{P}^1(K)$ at infinity. In this case the affine part of C is given by the curve

$$C'_Z : y^2 - x^3 - 1 = 0. \tag{1.3.17}$$

If we instead homogenize $F(X,Y,Z)$ with respect to Y, then the point p corresponds to the point $(x,z) = (0,0) \in K^2$ in the affine part of C. The whole affine part of C is in this case the curve

$$C'_Y : z - x^3 - z^3 = 0. \tag{1.3.18}$$

When studying points on a projective curve C, it is often easier if the point lies in the affine part of C. So then it is sometimes better to dehomogenize to another variable than Z, or even use different dehomogenizations for different points.

Now let us go back to our example in Section 1.1. There we wanted to find the points $(x, y) \in \mathbb{Q}^2$ that satisfied the equation

$$x^n + y^n = 1. \tag{1.3.19}$$

We found that these solutions were all of the form $(a/c, b/c)$, for integers a, b, c. Further, if we had such a solution, it corresponded to an equivalence class of solutions of the equation

$$X^n + Y^n = Z^n \tag{1.3.20}$$

with $Z \neq 0$. Now, we see that we were actually studying the affine part

$$C' : f(x, y) = x^n + y^n - 1 = 0 \tag{1.3.21}$$

of the projective curve

$$C : F(X, Y, Z) = X^n + Y^n - Z^n = 0. \tag{1.3.22}$$

The equivalence class of solutions corresponding to a solution $(a/c, b/c) \in C'$ is just the point $[a, b, c] \in C \subset \mathbb{P}^2(\mathbb{Q})$. The extra solutions of Equation (1.3.20) we got for $Z = 0$, are in fact the points at infinity on the curve C. Remember that we were only interested in integer solutions of Equation (1.3.20). So in principle we want the points C to have integer coordinates, but $C \subset \mathbb{P}^2(\mathbb{Q})$. However, because points in the projective plane are given by homogeneous coordinates, we can for every point $p \in \mathbb{P}^2(\mathbb{Q})$ clear the denominators of its coordinates to get integer coordinates for p. So we find that we can indeed represent every point in $\mathbb{P}^2(\mathbb{Q})$ using integer homogeneous coordinates.

1.4 Tangent Lines

Given a point p on a curve C, algebraic or projective, an interesting question would be whether the curve has a tangent line at p and how we can construct it. We will answer this question by showing how to find the tangent line, and then we immediately see when this does not work.

We start by looking at algebraic curves in $\mathbb{R}^2$. We can then of course view these curves as the affine part of a projective curve. Suppose we have the curve

$$C' : f(x, y) = 0, \tag{1.4.1}$$

and a point $p = (p_x, p_y) \in C'$. If we want to construct the tangent line at p, this comes down to finding the slope of the curve at the point p. For this we need the following result from the *implicit function theorem.*

Theorem 1.4.1. *Suppose that we have the polynomial equation*

$$g(x, y) = 0, \tag{1.4.2}$$

and a point $q = (a, b)$ that satisfies the equation. Further, suppose that

$$\left.\frac{\partial g}{\partial y}\right|_q \neq 0. \tag{1.4.3}$$

Then there exists a neighborhood U of q such that in U we can write the variable y as a function of x.

Proof. This theorem is a special case of the implicit function theorem, which can be found on for example page 733 of Adams and Essex [2]. □

Now, we can find the slope of C' using a technique called *implicit differentiation*. We take the polynomial equation $f(x, y) = 0$, and take its derivative to the variable x. However, and this is the special thing about implicit differentiation, we do this wile viewing the variable y as a function of x. Then the derivative $\frac{dy}{dx} = y'$ gives the slope of the curve. We know from the above theorem that this is locally possible around any point $p \in C'$, as long as

$$\left.\frac{\partial f}{\partial y}\right|_p \neq 0. \tag{1.4.4}$$

After taking the implicit derivative with respect to x of the equation $f(x, y) = 0$, we end up with a polynomial equation in the variables x, y and y'. According to the chain rule for functions of two variables we get

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} y' = 0, \tag{1.4.5}$$

see page 708 of Adams and Essex [2]. Suppose that we have solved this equation for y'. Then filling in the coordinates of our point $p = (p_x, p_y)$, we obtain the slope $y'(p_x, p_y)$ of C' at the point p. Now the tangent line to C' at p is the unique line through p with slope $y'(p_x, p_y)$. This line is given by

$$y = y'(p_x, p_y)(x - p_x) + p_y. \tag{1.4.6}$$

We can rewrite this to

$$-y'(p_x, p_y)(x - p_x) + y - p_y = 0. \tag{1.4.7}$$

Then, if we multiply this equation with $\frac{\partial f}{\partial y}(p_x, p_y)$ and use Equation (1.4.5) evaluated in the point p, we obtain

$$\frac{\partial f}{\partial x}(p_x, p_y)(x - p_x) + \frac{\partial f}{\partial y}(p_x, p_y)(y - p_y) = 0. \tag{1.4.8}$$

This equation gives the tangent line to C' at the point p. Remember that we could not find the slope y' using implicit differentiation in points p where

$$\left.\frac{\partial f}{\partial y}\right|_p \neq 0. \tag{1.4.9}$$

However, we see that Equation (1.4.8) also works for these points, as long as not also

$$\left.\frac{\partial f}{\partial x}\right|_p \neq 0. \tag{1.4.10}$$

In that case the tangent line at p will be a vertical line, so therefore its slope is indeed not defined. Because the partial derivative of a polynomial is defined for polynomials over any field, we can use Equation (1.4.8) to give a definition of the tangent line to a curve in the plane K^2 for any field K.

Definition 1.4.2. *Let K be a field. Suppose that we have the affine curve $C' \subset K^2$ given by*

$$C' : f(x, y) = 0, \tag{1.4.11}$$

together with a point $p = (p_x, p_y) \in C'$. Then the ***tangent line*** *to C' at the point p is given by Equation* (1.4.8).

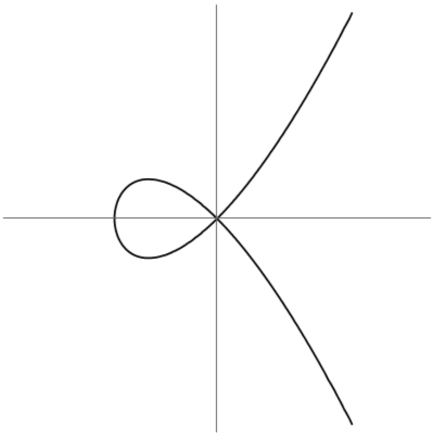

Figure 1.1: The curve $C_1 : y^2 = x^3 + x^2$ in $\mathbb{R}^2$

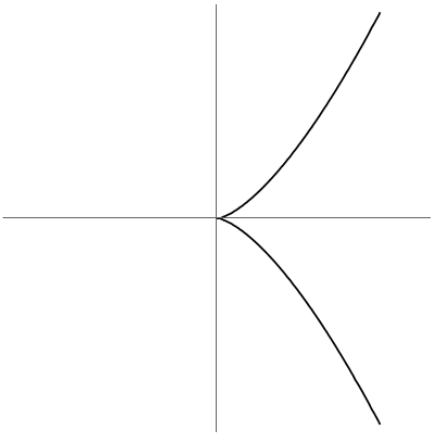

Figure 1.2: The curve $C_2 : y^2 = x^3$ in $\mathbb{R}^2$

We see that the above definition gives the tangent line in almost every point of a curve

$$C' : f(x, y) = 0 \tag{1.4.12}$$

in the plane K^2 for some field K. The only problematic points are when both partial derivatives of f are 0. We call these points *singular points*. Other points are called *non-singular points*. For example, the curves

$$C_1 : y^2 = x^3 + x^2 \qquad \text{and} \qquad C_2 : y^2 = x^3 \tag{1.4.13}$$

both have the point $p = (0, 0)$ as a singular point. If we sketch these curves in $\mathbb{R}^2$, we see that C_1 makes a loop and crosses itself at p, see Figure 1.1. Therefore the tangent line is not well-defined because there are two distinct tangent directions. The curve C_2, on the other hand, has a so-called cusp at p. The curve forms a sharp point with the point p at the tip, see Figure 1.2.

Definition 1.4.3. *Let C be an algebraic curve. We say that C is a* ***non-singular curve*** *if every point of C is non-singular. A non-singular curve is also called a* ***smooth curve****.*

From this definition, we immediately see that a non-singular curve has a well-defined tangent line in all of its points.

Now what about projective curves? Suppose we have a projective curve $C : F(X, Y, Z) = 0$ and a point $p = [a, b, c] \in C$. If $c \neq 0$, then p lies on the affine part of C if we dehomogenize F with respect to Z. Then the point p corresponds to the point

$$p' = \left(\frac{a}{c}, \frac{b}{c}\right) \quad \text{on the affine curve} \quad C' : F(x, y, 1) = 0. \tag{1.4.14}$$

We say that p is a singular point of C if and only if p' is a singular point of C'. In the case that $c = 0$, we find that p corresponds to a point at infinity if we dehomogenize F with respect to Z. We do not really know how to check whether a point at infinity is a singular point, but we can avoid this problem by just dehomogenizing F with respect to another variable. For example if $a \neq 0$, we can dehomogenize F with respect to X. Then p corresponds to the point

$$p'' = \left(\frac{b}{a}, \frac{c}{a}\right) \quad \text{on the affine curve} \quad C'' : F(1, y, z) = 0, \tag{1.4.15}$$

and we call p singular if and only if p'' is singular. For this to be well defined, we have to show that it is independent of the homogenization we choose whether a point is singular or not. We will now prove that this is the case. Indeed, consider the curve

$$C : F(X, Y, Z) = 0. \tag{1.4.16}$$

Recall from Equation (1.3.2) that for any $t \in K$ we have

$$F(tX, tY, tZ) = t^d F(X, Y, Z), \tag{1.4.17}$$

where d is the degree of F. We can differentiate the above expression with respect to t. Then, again using the multivariable chain rule [2], we get

$$X\frac{\partial F}{\partial X}(tX, tY, tZ)+Y\frac{\partial F}{\partial Y}(tX, tY, tZ)+Z\frac{\partial F}{\partial Z}(tX, tY, tZ) = dt^{d-1}F(X, Y, Z). \tag{1.4.18}$$

If we evaluate this expression in $t = 1$ we obtain

$$X\frac{\partial F}{\partial X}(X, Y, Z) + Y\frac{\partial F}{\partial Y}(X, Y, Z) + Z\frac{\partial F}{\partial Z}(X, Y, Z) = dF(X, Y, Z). \tag{1.4.19}$$

Note that the partial derivatives of a homogeneous polynomial of degree d are themselves homogeneous polynomials of degree $d - 1$. Now, let $p = [a, b, c] \in C$ be a point on the curve. Then we have $F(a, b, c) = 0$. Filling this point in into the above equation yields

$$a\frac{\partial F}{\partial X}(a, b, c) + b\frac{\partial F}{\partial Y}(a, b, c) + c\frac{\partial F}{\partial Z}(a, b, c) = 0. \tag{1.4.20}$$

Because $[a, b, c]$ is a point in the projective plane we know that at least one of its homogeneous coordinates must be non-zero. We assume $c \neq 0$. Then we can dehomogenize the curve with respect to Z such that p corresponds to the point $(a/c, b/c)$ on the affine part of the curve. The affine part of the curve is then given by

$$C' : F(x, y, 1) = 0. \tag{1.4.21}$$

Then, by definition, the point p is singular if and only if

$$\frac{\partial F}{\partial x}\left(\frac{a}{c}, \frac{b}{c}, 1\right) = \frac{\partial F}{\partial y}\left(\frac{a}{c}, \frac{b}{c}, 1\right) = 0. \tag{1.4.22}$$

Now, note that

$$\frac{\partial F}{\partial x} = \frac{\partial F}{\partial X} \quad \text{and} \quad \frac{\partial F}{\partial y} = \frac{\partial F}{\partial Y}. \tag{1.4.23}$$

Then, using the fact that the partial derivatives of F are homogeneous polynomials, we see that Equation (1.4.22) implies

$$\frac{\partial F}{\partial X}(a, b, c) = \frac{\partial F}{\partial Y}(a, b, c) = 0. \tag{1.4.24}$$

Remember that we assumed $c \neq 0$, so then it follows from Equation (1.4.20) that also

$$\frac{\partial F}{\partial Z}(a, b, c) = 0. \tag{1.4.25}$$

Therefore, we see that all three partial derivatives of F vanish at $p = [a, b, c]$. This implies that if we would have dehomogenized the curve with respect to the variables X or Y, we would still find that $[a, b, c]$ is a singular point on C. In this derivation we assumed $c \neq 0$, but the cases $a \neq 0$ and $b \neq 0$ are completely analogous. So we conclude that when we check whether a point on a projective curve is singular, the result is indeed independent from the dehomogenization we work with.

Similar to our definition for algebraic curves, we call a projective curve C *non-singular* or *smooth* if all of its points, including those at infinity, are non-singular. In particular this implies that the affine part of a smooth projective curve is smooth.

Chapter 2

Elliptic Curves

In this chapter we will give the definition of an elliptic curve and derive some properties for elliptic curves. In particular we show how the points on an elliptic curve in Weierstrass normal form together form an abelian group, and we derive explicit formulas for the addition of two points. In this chapter we follow Chapter I of Silverman and Tate [1].

2.1 Elliptic Curves and Weierstrass Normal Form

Now we are ready to start looking at elliptic curves, which are just special examples of projective curves.

Definition 2.1.1. *Let K be a field. Let*

$$C : F(X, Y, Z) = 0 \tag{2.1.1}$$

be a smooth projective curve in $\mathbb{P}^2(K)$ with at least one point $p \in C$. If the polynomial $F(X, Y, Z)$ is homogeneous of degree 3, then we call C an ***elliptic curve****.*

A curve of degree 3 is also called a *cubic*. So elliptic curves are special examples of cubics. Note that in the above definition we require an elliptic curve to contain at least one point. This is needed because there exist in fact cubic curves that do not contain any points. This happens for instance with the curve

$$3x^3 + 4Y^3 + 5Z^3 = 0 \tag{2.1.2}$$

in $\mathbb{P}^2(\mathbb{Q})$, as shown by Selmer [1]. Remember that for homogeneous polynomials the projective integer and rational solutions coincide. Selmer showed that the only integer solution to the above equation is $(X, Y, Z) = (0, 0, 0)$, but that is not a point in $\mathbb{P}^2(\mathbb{Q})$. Therefore the curve has no points in $\mathbb{P}^2(\mathbb{Q})$.

An important result for elliptic curves, is that any elliptic curve can be transformed into an elliptic curve in so-called *Weierstrass normal form* [1]. This transformation is such that there is a one-to-one correspondence between the points on the original curve and the transformed curve. If we are working in $\mathbb{P}^2(K)$ where the characteristic of the field K is unequal to 2, then an elliptic curve in Weierstrass normal form is of the form

$$Y^2Z = X^3 + aX^2Z + bXZ^2 + cZ^3. \tag{2.1.3}$$

If we dehomogenize with respect to Z, we see that the affine part of a curve in normal form is given by an equation of the form

$$y^2 = f(x) = x^3 + ax^2 + bx + c. \tag{2.1.4}$$

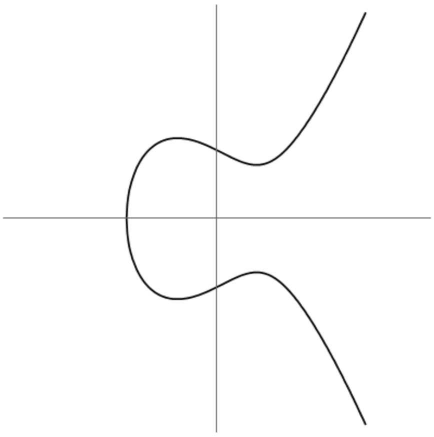

Figure 2.1: The smooth curve C_1 : $y^2 = x^3 - x + 1$ in $\mathbb{R}^2$.

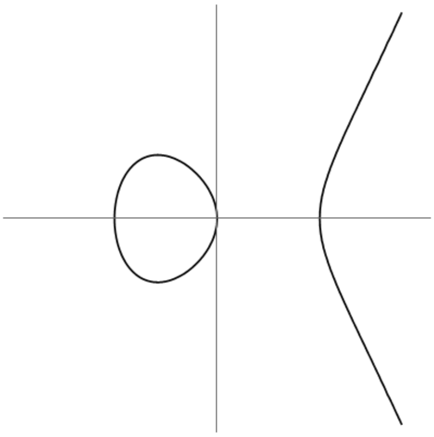

Figure 2.2: The smooth curve C_2 : $y^2 = x^3 - x$ in $\mathbb{R}^2$.

The affine part of an elliptic curve in Weierstrass normal form typically looks like one of the curves in Figures 2.1 and 2.2, depending on whether $f(x)$ has one or three roots in K. The curves in Figures 1.1 and 1.2 from the previous chapter are not elliptic curves, as they are not smooth.

Proposition 2.1.2. *A curve in K^2 given by an equation of the form*

$$y^2 = f(x) = x^3 + ax^2 + bx + c \tag{2.1.5}$$

is smooth if and only if the discriminant of f is non-zero. In other words, it is smooth if and only if f has three distinct roots in the algebraic closure of K.

Proof. For the proof of this proposition, see Proposition III.1.4 in Silverman [3]. □

See Appendix A if you are not familiar with the discriminant of a polynomial. To see which points at infinity lie on a curve in normal form, we fill in $Z = 0$ in Equation (2.1.3). Then we get

$$0 = X^3, \tag{2.1.6}$$

which has only one solution $X = 0$. Therefore an elliptic curve in normal form has exactly one point at infinity, namely $[0, 1, 0]$. This is the point where vertical lines in K^2 meet. Note that because we found one point already, every smooth projective curve given by an equation of the form in Equation (2.1.3) is an elliptic curve. So by Proposition 2.1.2 we see that every affine curve given by an equation of the form in Equation (2.1.4) defines the affine part of an elliptic curve if and only if the discriminant of f is non-zero.

2.2 Intersections of a Line and a Curve

In the next section we will show that there is a group structure on the points of an elliptic curve. But before we do this, we have to introduce the notion of *intersection multiplicity*.

We start by looking at the unit circle $\mathbb{S}^1 \subset \mathbb{R}^2$. Note that this is not an elliptic curve. Suppose that we take one point p in the unit circle and draw a line through p. Then most of the time this line will intersect the circle in exactly one other point. However, there is one exception, namely when we take the tangent line to $\mathbb{S}^1$ at p. This tangent line has only one intersection with $\mathbb{S}^1$. Another way to think of this is

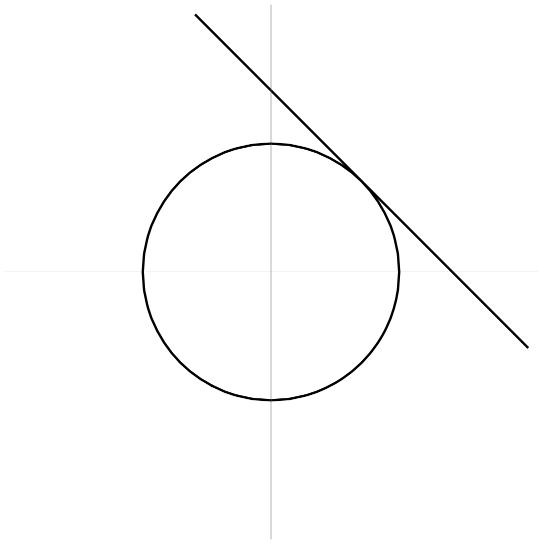

Figure 2.3: Unit circle with tangent line at $p = (\sqrt{2}/2, \sqrt{2}/2)$.

that the tangent line in p does intersect $\mathbb{S}^1$ twice, but both times in the same point p. Intuitively this makes sense because suppose that we have a sequence of points $\{q_i\}_{i \in \mathbb{N}}$ on $\mathbb{S}^1$ that converges to p. Then if we let i go to infinity, the line through p and q_i will approach the tangent line to $\mathbb{S}^1$ at p. However, we can also motivate this from an algebraic point of view. The unit circle $\mathbb{S}^1$ is the curve given by the equation

$$x^2 + y^2 = 1. \tag{2.2.1}$$

Now suppose that p is the point $(\sqrt{2}/2, \sqrt{2}/2)$. Then the tangent line to $\mathbb{S}^1$ through p is given by the equation

$$y = -x + \sqrt{2}. \tag{2.2.2}$$

Now we can determine the intersections of the line and $\mathbb{S}^1$ by substituting $y = -x + \sqrt{2}$ into the equation for $\mathbb{S}^1$. Then we find

$$2x^2 - 2\sqrt{2}x + 1 = 0, \tag{2.2.3}$$

which we can also write as

$$\left(\sqrt{2}x - 1\right)^2 = 0. \tag{2.2.4}$$

Hence, we see that $x = \sqrt{2}/2$ is a double solution of this equation. In other words, it is a root with multiplicity two. Of course, this solution corresponds to the point p. Therefore we say that the tangent line at p intersects $\mathbb{S}^1$ at p with multiplicity 2.

Now, suppose we have a cubic curve $C : F(X, Y, Z) = 0$, together with a point $p \in C$. Then if we draw a line through p, this line will in general intersect the curve C in two more points. Just as for the circle, we can find these points algebraically. Then we obtain a cubic equation for the coordinates of the intersection points. Note that for a projective curve one of these points could lie at infinity. If we find a double root of this equation, it means that the line is tangent to the curve in one point, and has one other normal intersection point. However, this time we could also have a triple root. This happens if we take the tangent line in an inflection point (Dutch: buigpunt) of the curve. Therefore we say that the tangent line at an inflection point of C intersects the curve with multiplicity 3. Note that if we find two roots of the equation for the coordinates of intersection points (counting multiplicities), then we automatically also get a third root. This is because we can factor out the the two known roots to obtain a linear equation for the third root. However, if we find only one root, then it could be that there is only one root. For example the cubic

$$(x+1)(x^2+1) = 0 \tag{2.2.5}$$

has only one root in $\mathbb{R}$, namely $x = -1$. The other two roots are complex. Therefore we find that for any point p on a cubic curve, a line through p has either one or three intersections with the curve, counting multiplicities. At least of these intersections is of course at p itself.

2.3 The Group of Points on an Elliptic Curve

In this section we will define a group structure on the set of points on elliptic curves. Remember that we can transform every elliptic curve into an elliptic curve in Weierstrass normal form, in such a way that there is a one-to-one correspondence between the points on the curves. Therefore we will only consider elliptic curves in Weierstrass normal form.

Suppose that we have an elliptic curve in normal form

$$C : Y^2Z = X^3 + aX^2Z + bXZ^2 + cZ^3. \tag{2.3.1}$$

At the end of Section 2.1 we showed that if we dehomogenize with respect to Z, then this curve has exactly one point at infinity, namely $[0, 1, 0]$. We did this by looking at the intersections of the line at infinity with the curve. Now using what we saw in the previous section, we find that the line at infinity intersects the curve in $[0, 1, 0]$ with multiplicity 3. So the point $[0, 1, 0]$ is actually an inflection point of the curve. We call this point $\mathcal{O}$, and it will be the identity element of the group of points on the curve.

Now, the group structure on the set of points on an elliptic curve comes from the following idea: Suppose that we have an elliptic curve C together with two points $p, q \in C$. Then we can use these two points to construct a third point on C. Namely, if we draw the line through p and q, then we saw in the previous section that this line will intersect the curve in a third point. We denote this third point by $p * q$. If we have $p = q$, then we say that the line through p and itself is the tangent line to the curve at p. This is motivated by the fact that this tangent line intersects the curve with multiplicity at least 2, as we saw in the previous section.

So this gives us an operation $*$ that takes two points on an elliptic curve and spits out a third one. Then one might think that the set of points on the curve together with this operation forms a group. But it is relatively simple to see that this is not the case, because there can not be a point that acts as the identity element for the operation $*$. However, it turns out that if we modify the operation a bit, then we can turn the set of points on an elliptic curve into an abelian group. Because the group will be abelian, we denote the group operation by $+$. We define the operation as follows:

Definition 2.3.1 (The addition of points on an elliptic curve)**.** *Given two points p, q on an elliptic curve C, we can construct the point $p * q$ as described above. Then we define the **sum** $p + q$ as the third intersection point with C of the line through $\mathcal{O}$ and $p * q$. So we have $p + q = \mathcal{O} * (p * q)$.*

In Figures 2.4 and 2.5 it is illustrated how the addition of points on an elliptic curve works. Remember that $\mathcal{O}$ is the point $[0, 1, 0]$ on the elliptic curve. We dehomogenized the curve with respect to Z, so $\mathcal{O}$ is a point at infinity. Then how do we draw a line

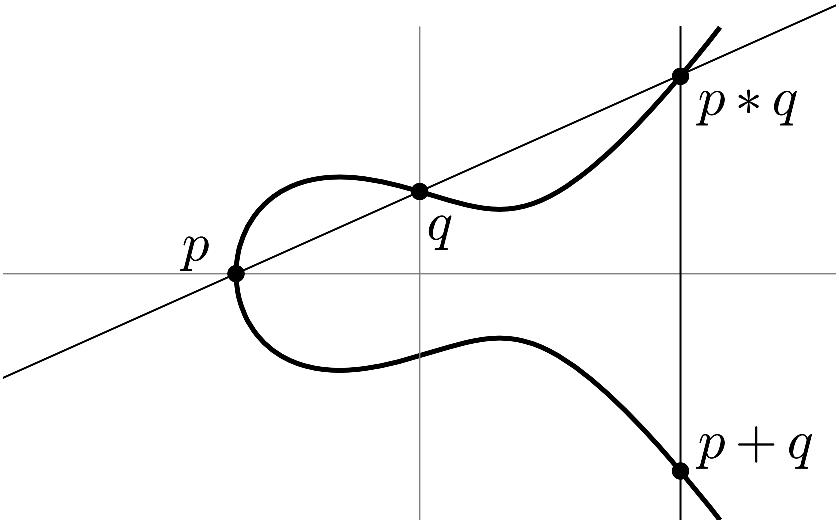

Figure 2.4: The addition of two points on an elliptic curve in Weierstrass normal form.

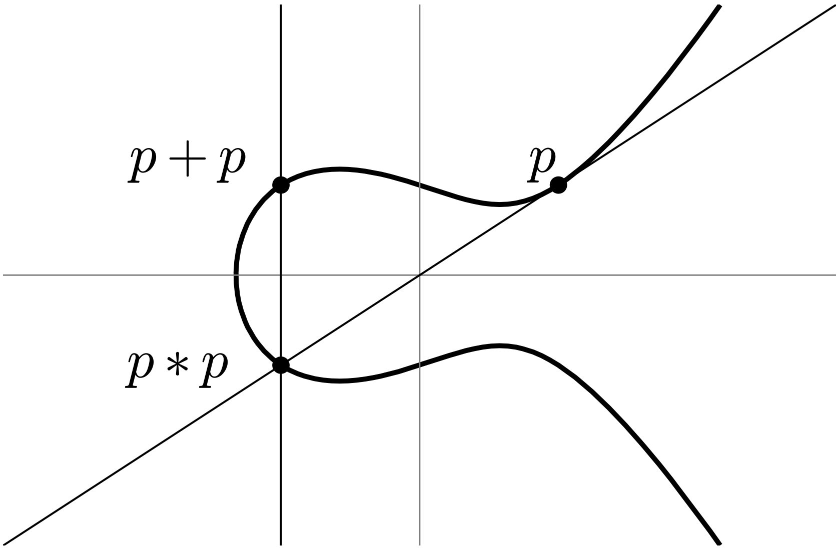

Figure 2.5: Adding a point on an elliptic curve in Weierstrass normal form to itself.

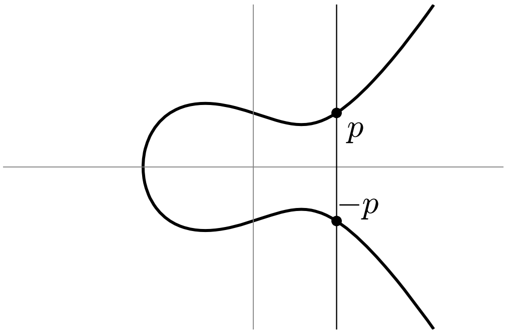

Figure 2.6: Finding the inverse of a point on an elliptic curve in Weierstrass normal form.

through $\mathcal{O}$ and another point on the curve? It turns out that the projective lines containing $\mathcal{O}$ are precisely the lines that have a vertical line as affine part. So if we have a point p on the affine part of an elliptic curve and we want to draw the line through p and $\mathcal{O}$, then the affine part of this line will be the vertical line through p.

Theorem 2.3.2. *Together with the operation + defined in Definition 2.3.1, the points on an elliptic curve in Weierstrass normal form form an abelian group.*

Proof. To show that together with the operation + the set of points of an elliptic curve indeed forms an abelian group, we have to prove four things. Namely, the operation + should be commutative, there must be an identity element, every point must have an inverse, and the operation + should be associative. If we have two points p, q on an elliptic curve, then the line through p and q is the same as the line through q and p. Therefore we find $p * q = q * p$, which implies $p + q = q + p$. So the operation + is indeed commutative. As we claimed before, the identity element of the group is $\mathcal{O}$. To see this, suppose that we add any point p to $\mathcal{O}$. Then we get

$$p + \mathcal{O} = \mathcal{O} * (p * \mathcal{O}). \tag{2.3.2}$$

The point $p * \mathcal{O}$ is the third intersection point of the line through p and $\mathcal{O}$ with the curve. But then the line through $\mathcal{O}$ and $p * \mathcal{O}$ will be the same line. So the point $\mathcal{O} * (p * \mathcal{O})$ is again the point p. Hence $p + \mathcal{O} = p$ for any point p on the curve, so $\mathcal{O}$ is indeed the identity element of the group. The inverse $-p$ of a point p is given by $-p = p * (\mathcal{O} * \mathcal{O})$. Indeed, we get $p * (-p) = (\mathcal{O} * \mathcal{O})$, so $p + (-p) = \mathcal{O} * (\mathcal{O} * \mathcal{O}) = \mathcal{O}$. Now the only thing left to prove in order to show that the points form a group is the associativity of the group operation. To prove that the operation + is associative is very cumbersome and not particularly interesting. In the next section we will give explicit formulas for the operation +, then in principle one could check associativity just by writing everything out. We will not work out the calculation in this thesis. □

It turns out that finding the inverse of a point on a curve in Weierstrass normal form is pretty easy, as shown in the following proposition. See also Figure 2.6.

Proposition 2.3.3. *Let $p = (x, y) \in K^2$ be a point on the affine part of an elliptic curve. Then we have*

$$-p = p * \mathcal{O} = (x, -y). \tag{2.3.3}$$

Proof. In the previous proof we saw that

$$-p = p * (\mathcal{O} * \mathcal{O}). \tag{2.3.4}$$

However, we can actually simplify this expression using the fact that we chose $\mathcal{O}$ such that it is an inflection point of the curve. Namely, this implies that $\mathcal{O} * \mathcal{O} = \mathcal{O}$. Therefore we see that the inverse of a point p is given by $p * \mathcal{O}$. By assumption the point p lies in the affine part of the curve. Remember that in that case the line through p and $\mathcal{O}$ is the vertical line through p. From Equation (2.1.4) we see that the affine part of elliptic curve in Weierstrass normal form is symmetric around the x-axis. Therefore the inverse of a point p is just the point we get when we reflect p across the x-axis. Hence, we find that indeed

$$-p = p * \mathcal{O} = (x, -y). \tag{2.3.5}$$

□

Corollary 2.3.4. *From the above Proposition it follows that for two points p, q on an elliptic curve we have*

$$p + q = -(p * q). \tag{2.3.6}$$

2.4 Explicit Formulas for the Group Operation

In this section we will derive explicit formulas for the addition of two points on an elliptic curve in Weierstrass normal form. We do this by writing the points in coordinates in the plane K^2, such that we can explicitly calculate the line through two points. However, there is one point on an elliptic curve in normal form that does not lie in K^2, namely the point $\mathcal{O}$. But we know that $\mathcal{O}$ is the identity element of the group. So if we add $\mathcal{O}$ to a any point p on the curve, then we just obtain p again.

Now, suppose that we have two points p_1 and p_2 both unequal to $\mathcal{O}$ that we want to add to each other. The first case we consider is when $p_2 = -p_1$. Then we get $p_1 + p_2 = \mathcal{O}$. Now, consider the case when $-p_1 \neq p_2$. Then both $p_1 * p_2$ and $p_1 + p_2$ are unequal to $\mathcal{O}$. Therefore we can write these points in coordinates as

$$p_1 = (x_1, y_1), \quad p_2 = (x_2, y_2), \quad p_1 * p_2 = (x_3, -y_3). \tag{2.4.1}$$

We chose the y-coordinate of $p_1 * p_2$ as $-y_3$, because then $p_1 + p_2$ will be equal to (x_3, y_3). Indeed, this follows directly from Corollary 2.3.4 and Proposition 2.3.3.

Given the coordinates (x_1, y_1) and (x_2, y_2), we want to compute (x_3, y_3). Suppose first that $p_1 \neq p_2$. Then we can construct the line through p_1 and p_2. We write the equation for this line as

$$y = \lambda x + \nu. \tag{2.4.2}$$

Then we find

$$\lambda = \frac{y_2 - y_1}{x_2 - x_1} \quad \text{and} \quad \nu = y_1 - \lambda x_1 = y_2 - \lambda x_2. \tag{2.4.3}$$

Now, the point $p_1 * p_2 = (x_3, -y_3)$ is the third intersection point of this line with the curve. The affine part of the elliptic curve is given by the equation

$$y^2 = x^3 + ax^2 + bx + c, \tag{2.4.4}$$

for some constants $a, b, c \in K$. To find the third intersection point we substitute

$$y^2 = (\lambda x + \nu)^2 = x^3 + ax^2 + bx + c. \tag{2.4.5}$$

We can rewrite this to

$$x^3 + (a - \lambda^2)x^2 + (b - 2\lambda\nu)x + (c - \nu^2) = 0. \tag{2.4.6}$$

Then the roots of this equation will be x_1, x_2 and x_3, the x-coordinates of the intersection points. Therefore we get

$$x^3 + (a - \lambda^2)x^2 + (b - 2\lambda\nu)x + (c - \nu^2) = (x - x_1)(x - x_2)(x - x_3). \tag{2.4.7}$$

This polynomial equation has to hold for any value of x. Therefore the coefficients in front of each power of x must be the same. Equating the coefficients of the x^2 term of both sides yields

$$\lambda^2 - a = x_1 + x_2 + x_3. \tag{2.4.8}$$

Hence, we find that the coordinates of the point $p_1 + p_2 = (x_3, y_3)$ are given by

$$x_3 = \lambda^2 - a - x_1 - x_2, \qquad y_3 = -(\lambda x_3 + \nu). \tag{2.4.9}$$

Now, the only case left to consider is when $p_2 = p_1$. In that case we want to construct the tangent line to the curve at p_1, and calculate its third intersection point with the curve. If we let

$$f(x, y) = y^2 - x^3 - ax^2 - bx - c, \tag{2.4.10}$$

then the affine part of the curve is given by $f(x, y) = 0$. Then, according to Definition 1.4.2 the tangent line to the curve at the point $p_1 = (x_1, y_1)$ is given by

$$\frac{\partial f}{\partial x}(x_1, y_1)(x - x_1) + \frac{\partial f}{\partial y}(x_1, y_1)(y - y_1) = 0. \tag{2.4.11}$$

The partial derivatives of $f(x, y)$ at the point p_1 are

$$\frac{\partial f}{\partial x}(x_1, y_1) = -3x_1^2 - 2ax_1 - b \quad \text{and} \quad \frac{\partial f}{\partial y}(x_1, y_1) = 2y_1. \tag{2.4.12}$$

Therefore the equation for the tangent line at p becomes

$$(-3x_1^2 - 2ax_1 - b)x + (3x_1^3 + 2ax_1^2 + bx_1) + 2y_1y - 2y_1^2 = 0. \tag{2.4.13}$$

Remember that we assumed that $p_1 \neq -p_2$. So in this case, $p_1 \neq -p_1$. The elliptic curve is smooth and symmetric about the x-axis. Therefore the tangent line at a point on the curve with y-coordinate equal to 0 must be a vertical line. Then the other intersection point of this line with the curve is $\mathcal{O}$. So every intersection point of the curve with the x-axis is its own inverse. Therefore, our assumption that $p_1 \neq -p_1$ implies that $y_1 \neq 0$. Hence, we can divide the equation for the tangent line by $2y_1$. Then we can rewrite it to

$$y = \frac{3x_1^2 + 2ax_1 + b}{2y_1}x - \frac{3x_1^3 + 2ax_1^2 + bx_1}{2y_1} + y_1. \tag{2.4.14}$$

So, if we write this line as $y = \lambda x + \nu$, we get

$$\lambda = \frac{3x_1^2 + 2ax_1 + b}{2y_1}, \qquad \nu = -\frac{3x_1^3 + 2ax_1^2 + bx_1}{2y_1} + y_1. \tag{2.4.15}$$

Then, using these values for λ and ν, we can use Equation (2.4.9) to find the coordinates of the point $p_1 + p_2 = (x_3, y_3)$ in the case that $p_1 = p_2$.

2.5 Elliptic Curves over Finite Fields

For a prime number p, we denote by $\mathbb{F}_p$ the finite field of integers modulo p. In this section we will take a closer look at elliptic curves in the projective plane $\mathbb{P}^2(\mathbb{F}_p)$, following Section IV.1 of Silverman and Tate [1].

The field $\mathbb{F}_p$ consists of p elements, $[0], \ldots, [p-1]$. These elements are equivalence classes of integers, and two integers $m, n \in \mathbb{Z}$ belong to the same equivalence class if and only if $m \equiv n \pmod{p}$. Because $\mathbb{F}_p$ is a field, we can obtain its multiplicative group $\mathbb{F}_p^*$ by leaving out 0 and forgetting about addition. For this group we have the following theorem:

Theorem 2.5.1. *Let p be a prime number. Then the multiplicative group $\mathbb{F}_p^*$ of the finite field $\mathbb{F}_p$ is cyclic, i.e. $\mathbb{F}_p^* \cong \mathbb{Z}/(p-1)\mathbb{Z}$.*

Proof. The proof of this theorem can be found in for example the proof of Theorem IV.1.9 of Lang [4], or in Section 7.4 of Beukers [5] (Dutch). The statement of the theorem is equivalent to saying that for every prime p there exists a primitive root modulo p. □

Suppose that we have an elliptic curve in $\mathbb{P}^2(\mathbb{F}_p)$

$$C : F(X, Y, Z) = 0. \tag{2.5.1}$$

We can dehomogenize this curve with respect to for instance Z, such that C becomes an affine curve in $\mathbb{F}_p^2$ together with one point at infinity (we assume the curve to be in normal form). Everything we derived in the previous section for the group of points on an elliptic curve also holds for curves in $\mathbb{P}^2(\mathbb{F}_p)$. That is, as long as $p \neq 2$, because we need the characteristic of the field to be unequal to 2 to transform it into Weierstrass normal form. So the set of points on C forms an abelian group. In the case of a finite field, it is hard to visualize the group operation with lines and intersections. However, we do not really need to visualize anything. We can just stick to the formulas and use the explicit formulas for the group operation given in the previous section.

As an example, we consider the curve

$$C : Y^2Z = X^3 + XZ^2 + Z^3 \tag{2.5.2}$$

in the projective space $\mathbb{P}^2(\mathbb{F}_5)$. Note that it is already in Weierstrass normal form. If we dehomogenize with respect to Z, then the affine part of this curve is given by

$$y^2 = x^3 + x + 1, \tag{2.5.3}$$

and the point $\mathcal{O}$ at infinity is $[0, 1, 0]$ as always. Now, the field $\mathbb{F}_5$ only has 5 elements, so to find the points on the affine part C we can just try every pair of points in $\mathbb{F}_5$ in the equation. Then we find

$$C = \{\mathcal{O}, (0, \pm1), (2, \pm1), (3, \pm1), (4, \pm2)\}. \tag{2.5.4}$$

We find that there are nine point on C, so these points form an abelian group of order nine. Then this group is either isomorphic to $\mathbb{Z}/9\mathbb{Z}$ or $\mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$. To see which group it is, we look at multiples of the point $q = (0, 1) \in C$. We write $2q = q + q$, $3q = q + q + q$ and so on. We want to calculate these multiples using the formulas we gave in Equation (2.4.9). We can write the tangent line to C at the point q as

$$y = \lambda x + \nu. \tag{2.5.5}$$

From Equation (2.4.15) we find that

$$\lambda = -\frac{1}{2}, \qquad \nu = 1. \tag{2.5.6}$$

We are working in $\mathbb{F}_5$, so $1/2$ means 1 multiplied with the multiplicative inverse of 2. We have $2 \cdot 3 = 6 \equiv 1 \pmod 5$, so $1/2 = 3$. Then using the explicit formulas, while remembering that we are working modulo 5, we get

$$2q = (4,2), \quad 3q = (2,1), \quad 4q = (3,-1). \tag{2.5.7}$$

We see that the order of q is greater than 3. Hence the group of points on C cannot be isomorphic to $\mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$, so it is isomorphic to $\mathbb{Z}/9\mathbb{Z}$.

When the prime nuber p becomes larger, trying for every point in $\mathbb{P}^2(\mathbb{F}_p)$ separately whether it lies on a given curve quickly becomes a lot of work. One thing that we would like to know without trying every point, is how many points there are on a given curve. In the next chapter we will look at the number of points on a specific elliptic curve over a finite field. Here we will look at the number of points on a general line in $\mathbb{P}^2(\mathbb{F}_p)$.

Suppose that we have a line $L \subset \mathbb{P}^2(\mathbb{F}_p)$. After dehomogenizing, we can write this line as the union of one point at infinity and an affine part given by the equation

$$y = ax + b \tag{2.5.8}$$

for some constants $a, b \in \mathbb{F}_p$. Now, for every value of x that we fill into this equation, we get exactly one solution for y. Therefore the number of points on the affine part of L is just the size of $\mathbb{F}_p$, which is p. Then the total number of points on L is $p+1$, because we have one extra point at infinity. So every line in $\mathbb{P}^2(\mathbb{F}_p)$ has $p+1$ points.

Chapter 3

A Theorem of Gauss

The main topic of this chapter will be a number theoretic result due to Gauss. In the first section we will state and prove this result, and in the second section we will use it to determine the group structure of the points on a specific elliptic curve in $\mathbb{P}^2(\mathbb{F}_{19})$. The results we derive in this chapter will not be used in the rest of this thesis.

3.1 A Theorem of Gauss

In this section we will state and prove a theorem of Gauss following Section IV.2 of Silverman and Tate [1].

We are interested in the number of solutions to the Fermat equation

$$x^3 + y^3 = 1, \tag{3.1.1}$$

for $x, y \in \mathbb{F}_p$. In homogeneous form this equation reads

$$X^3 + Y^3 = Z^3. \tag{3.1.2}$$

We will consider the projective solutions $[X, Y, Z] \in \mathbb{P}^2(\mathbb{F}_p)$. Note that the solutions to this equation are in one-to-one correspondence with solutions to the equation

$$X^3 + Y^3 + Z^3 = 0, \tag{3.1.3}$$

by taking $[X, Y, Z] \to [X, Y, -Z]$. Now the theorem of Gauss can be stated as follows:

Theorem 3.1.1 (Gauss). *Let M_p be the number of projective solutions $[X, Y, Z] \in \mathbb{P}^2(\mathbb{F}_p)$ to* (3.1.3).
(a) *If $p \not\equiv 1 \pmod 3$, then $M_p = p + 1$.*
(b) *If $p \equiv 1 \pmod 3$, then there exist integers A, B such that*

$$4p = A^2 + 27B^2. \tag{3.1.4}$$

Here A and B are unique up to changing their signs, and if we fix the sign of A such that $A \equiv 1 \pmod 3$, then

$$M_p = p + 1 + A \tag{3.1.5}$$

Note that in the case that $p \equiv 1 \pmod 3$ and $4p = A^2 + 27B^2$, it follows that $A^2 \equiv 1 \pmod 3$. Therefore $A \equiv \pm 1 \pmod 3$, so we can indeed always fix the sign of A such that $A \equiv 1 \pmod 3$.

Theorem 3.1.1 consists of two parts, (a) and (b). We start our proof of the theorem by proving part (a), which is the easier one.

Proof of Theorem 3.1.1(a). So we consider the case $p \not\equiv 1 \pmod 3$. Remember from Theorem 2.5.1 that $\mathbb{F}_p^*$ is a cyclic group. Therefore, there exists an element $g \in \mathbb{F}_p^*$ that generates the whole group. For $p \not\equiv 1 \pmod 3$ we have $3 \not| \, p-1$. We will show that this implies that the map

$$f : \mathbb{F}_p^* \to \mathbb{F}_p^*, \quad f(x) = x^3 \tag{3.1.6}$$

is an isomorphism.

Suppose that we have $x, y \in \mathbb{F}_p^*$ such that $x^3 = y^3$. Then $x = g^m$ and $y = g^n$ for some $n, m < p-1$. From $x^3 = y^3$ we find $g^{3m} = g^{3n}$, which implies $3m \equiv 3n \pmod{p-1}$. Now, because $3 \not| \, p-1$ and 3 is prime, this is equivalent to $m \equiv n \pmod{p-1}$. Therefore we find $m = n$, so $x = y$. Hence f is an injection from a finite set to itself and therefore a bijection. It is also a homomorphism because for any $u, v \in \mathbb{F}_p^*$ we have $f(uv) = (uv)^3 = u^3v^3 = f(u)f(v)$, where we use that $\mathbb{F}_p^*$ is commutative. We conclude that f is indeed an isomorphism.

Note that $0^3 = 0$, so every element of $\mathbb{F}_p$ has a unique cubic root. Therefore the number of solutions of $x^3 + y^3 + z^3 = 0$ in $\mathbb{F}_p$ is equal to the number of solutions of $x + y + z = 0$. For both equations we have that for any solution (x, y, z), the triple (ax, ay, az) is also a solution for every $a \in \mathbb{F}_p$. Therefore we obtain their number of projective solutions by leaving out the trivial solution $(0, 0, 0)$ and dividing by $|\mathbb{F}_p^*|$. This implies that also the number of projective solutions must be the same for both equations. The equation $x + y + z = 0$ defines a line in $\mathbb{P}^2(\mathbb{F}_p)$. Using our previous result that lines contain exactly $p + 1$ projective points, we conclude that indeed $M_p = p + 1$. □

The proof of part (b) takes more work, so we will first prove four lemmas. In the rest of this section we let p be a prime number such that $p \equiv 1 \pmod 3$. We write $p = 3m + 1$. We will still denote by M_p the number of projective solutions to $x^3 + y^3 + z^3 = 0$. We also introduce a symbol. Let X, Y, Z be subsets of $\mathbb{F}_p$. We denote by $[X, Y, Z]$ or $[XYZ]$ the number of triples (x, y, z) such that

$$x \in X, \quad y \in Y, \quad z \in Z, \quad \text{and} \quad x + y + z = 0. \tag{3.1.7}$$

This symbol has the following properties, which follow immediately from its definition:

$$\begin{aligned} [XY(Z \cup W)] &= [XYZ] + [XYW] \quad \text{if } Z \cap W = \emptyset, \\ [XYZ] &= [aX, aY, aZ] \quad \text{for any } a \neq 0, \text{ where } aX = \{ax \mid x \in X\}, \\ [XYZ] &= [XZY] = [YXZ] = [YZX] = [ZXY] = [ZYX], \\ [XY\mathbb{F}_p] &= |X||Y|. \end{aligned} \tag{3.1.8}$$

Note that for the second property we use that for $x, y \in \mathbb{F}_p$ and $a \in \mathbb{F}_p$, $a \neq 0$, we have $ax = ay$ if and only if $x = y$. This follows from the fact that $a \in \mathbb{F}_p^*$ and that $\mathbb{F}_p^*$ is a group, because that implies that a has a multiplicative inverse a^{-1}. Then we can just multiply $ax = ay$ with a^{-1} to obtain $x = y$.

Lemma 3.1.2. *Let*

$$R = \{x^3 \mid x \in \mathbb{F}_p^*\} \tag{3.1.9}$$

be the set of cubic residues in $\mathbb{F}_p^*$. Then R is a subgroup of $\mathbb{F}_p^*$ with index 3, and if we denote its other cosets by S and T we have

$$M_p = 9\frac{[RTS]}{m}. \tag{3.1.10}$$

Proof. We start by proving that R is a subgroup of $\mathbb{F}_p^*$ with index 3. As we have seen before, the map

$$f : \mathbb{F}_p^* \to \mathbb{F}_p^*, \quad f(x) = x^3 \tag{3.1.11}$$

is a group homomorphism. We saw that for $p \not\equiv 1 \pmod 3$ it was even an isomorphism. Now we have $p = 3m + 1$. Then by theorem 2.5.1 we have $\mathbb{F}_p^* \cong \mathbb{Z}/3m\mathbb{Z}$. This implies that there are exactly two elements $u, v \in \mathbb{F}_p^*$ different from 1 with order 3, and they are related via $v = u^2$. Therefore in this case the homomorphism f is not injective, and its kernel is given by $K = \{1, u, u^2\}$. We have that $R = f(\mathbb{F}_p^*)$. So R is the image of a homomorphism, and therefore a subgroup of $\mathbb{F}_p^*$. Now, according to the first isomorphism theorem [6], we have

$$R \cong \mathbb{F}_p^*/K. \tag{3.1.12}$$

Therefore $|R| = m$, so R has indeed index 3 in $\mathbb{F}_p^*$.

We write S and T for the other two cosets of R in $\mathbb{F}_p^*$. Because $\mathbb{F}_p^*$ is abelian we have that R is a normal subgroup of $\mathbb{F}_p^*$. Therefore the cosets R, S, T form a group, called the quotient group $\mathbb{F}_p^*/R$. In $\mathbb{F}_p^*/R$ multiplication works as follows: $(gR)(hR) = (gh)R$ for $g, h \in \mathbb{F}_p^*$. This group has order 3, so we must have $\mathbb{F}_p^*/R \cong \mathbb{Z}/3\mathbb{Z}$. Also $|R| = |S| = |T| = m$, and

$$\mathbb{F}_p = \{0\} \sqcup R \sqcup S \sqcup T. \tag{3.1.13}$$

Note that $(-1)^3 = -1$, so $-1 \in R$. That implies $R = -R$, and therefore we also have $S = -S$ and $T = -T$.

We want to express M_p in terms of R, S and T. For this, we use the symbol $[XYZ]$ we introduced. In terms of this symbol we find that the number of ways to write 0 as the sum of three non-zero cubes is $[RRR]$. But for every non-zero cube $x^3 \in R$, we have $x^3 = (ux)^3 = (u^2x)^3$. The polynomial $y^3 = x^3$ in the variable y has degree three. Therefore it cannot have more than three roots, see Corollary 3.1.3 in [7]. Hence, we see that there cannot be another element $y \in \mathbb{F}_p$ such that $y^3 = x^3$. So every non-zero cube has exactly 3 different roots in $\mathbb{F}_p$. Hence, we find that the number of solutions (x, y, z) to $x^3 + y^3 + z^3 = 0$ with x, y, z all non-zero is $27[RRR]$. Here we count (x, y, z) and (ax, ay, az) for some $a \in \mathbb{F}_p^*$ as different solutions, but we are interested in the number of projective solutions. If we have a solution (x, y, z) then the triple (ax, ay, az) is a solution for every $a \in \mathbb{F}_p^*$. Therefore the number of projective solutions of $x^3 + y^3 + z^3 = 0$ where x, y, z are all non-zero is equal to

$$\frac{27[RRR]}{3m} = \frac{9[RRR]}{m}. \tag{3.1.14}$$

Now, we want to calculate the number of projective solutions to $x^3 + y^3 + z^3 = 0$ where at least one of x, y, z is zero. Note that in that case we have that exactly one of x, y, z is zero. Because if two of them are zero it follows that the third must also be zero, and the triple $(0, 0, 0)$ is not a projective solution. Suppose $z = 0$. Then the equation becomes $y^3 = -x^3$. As we have seen before, for every $x \in \mathbb{F}_p^*$ this equation has exactly 3 solutions for y. Therefore the total number of solutions $(x, y, 0)$ is $3(p - 1)$. Again, we are only interested in projective solutions, and $(ax, ay, 0)$ is a solution for every solution $(x, y, 0)$ and $a \in \mathbb{F}_p^*$. So dividing by $p - 1$ we find that the number of projective solutions of $x^3 + y^3 = 0$ is 3. We can do exactly the same for the cases $y = 0$ and $x = 0$, so in total there are 9 projective solutions of $x^3 + y^3 + z^3 = 0$ where one of x, y, z is 0.

Taking this all together we find that

$$M_p = \frac{9[RRR]}{m} + 9 = 9\left(\frac{[RRR]}{m} + 1\right). \tag{3.1.15}$$

Now using the first and fourth properties in (3.1.8) and the fact that $\mathbb{F}_p = \{0\} \sqcup R \sqcup S \sqcup T$, we find

$$[RR\{0\}] + [RRR] + [RRS] + [RRT] = [RR\mathbb{F}_p] = m^2. \tag{3.1.16}$$

Let $s \in S$ and $t \in T$. Remember that S and T were the other two cosets of R in $\mathbb{F}_p^*$, and $\mathbb{F}_p^*/R \cong \mathbb{Z}/3\mathbb{Z}$. We have $R = 1R = 1^2R = R^2$, so R is the identity element of $\mathbb{F}_p^*/R$. From $\mathbb{F}_p^*/R \cong \mathbb{Z}/3\mathbb{Z}$ it follows that S and T have order 3, so $S^2 = T$ and $T^2 = S$. Note that $1 \in R$, so $sR = S$ and $s^2R = (sR)(sR) = S^2 = T$. Similarly we find $tR = T$ and $t^2R = S$. Now, using the second property in (3.1.8), we get $[RRS] = [sR, sR, sS] = [SST]$ and $[RRT] = [tR, tR, tT] = [TTS]$. Filling this in into Equation (3.1.16) we obtain

$$[RR\{0\}] + [RRR] + [SST] + [TTS] = m^2. \tag{3.1.17}$$

Similar to (3.1.16) we find

$$[\{0\}TS] + [RTS] + [STS] + [TTS] = [\mathbb{F}_pTS] = m^2. \tag{3.1.18}$$

The term $[\{0\}TS]$ is the number of solutions to $t^3 + s^3 = 0$ for $t \in T$ and $s \in S$. But this is 0, because $S = -S$ and $S \cap T = \emptyset$. Also $[RR\{0\}] = m$ because $R = -R$. Subtracting Equation (3.1.18) from (3.1.17) we obtain

$$m + [RRR] = [RTS]. \tag{3.1.19}$$

Now filling this in in Equation (3.1.15) we get indeed

$$M_p = 9\frac{[RTS]}{m}. \tag{3.1.20}$$

□

In the rest of this section we will stick to the notation R, T and S for the set of cubic residues in $\mathbb{F}_p^*$ and its cosets respectively. Now Lemma 3.1.2 tells us that in order to find M_p, we have to compute $[RTS]$. To do this we will introduce some complex numbers called cubic Gauss sums. But before we do that, we first have know what the p^{th} *roots of unity* are.

We define the complex number $\zeta = e^{2\pi i/p}$. Then the p^{th} roots of unity are $\zeta^0, \zeta^1, \ldots, \zeta^{p-1}$. Note that if $a, b \in \mathbb{Z}$, then $\zeta^{a+b} = \zeta^a\zeta^b$. Further, we have that $\zeta^a = \zeta^b$ if and only if $a \equiv b \pmod{p}$). So we see that the map from $\mathbb{F}_p$ to $\{\zeta^0, \ldots, \zeta^{p-1}\}$ given by

$$[a] \mapsto \zeta^a \tag{3.1.21}$$

is a group isomorphism, sending addition in $\mathbb{F}_p$ to multiplication in $\{\zeta^0, \ldots, \zeta^{p-1}\}$. Note that $(\zeta^a)^p = 1$, hence the name p^{th} roots of unity.

Now we define three complex numbers $\alpha_1, \alpha_2, \alpha_3$ as follows:

$$\alpha_1 = \sum_{r \in R} \zeta^r, \qquad \alpha_2 = \sum_{s \in S} \zeta^s, \qquad \alpha_3 = \sum_{t \in T} \zeta^t. \tag{3.1.22}$$

These $\alpha_1, \alpha_2, \alpha_3$ are called *cubic Gauss sums*. For $x \in \mathbb{F}_p$ we denote the number of pairs (s, t) with $s \in S$ and $t \in T$ such that $s + t = x$ by N_x. So we have $N_x = [ST\{-x\}]$. Note that for $r \in R$ we have $rR = R$ and therefore $rS = S$ and $rT = T$. Using this we find

$$N_x = [ST\{-x\}] = [rS, rT, \{-rx\}] = [ST\{-rx\}] = N_{rx}. \tag{3.1.23}$$

So $N_x = N_y$ for all $y \in xR$. Thus N_x only depends on the coset of R in which x is contained, R, S or T. Note that we used here that the right and left cosets xR and Rx are the same for all $x \in \mathbb{F}_p$ because $\mathbb{F}_p$ is abelian. We obtain

$$mN_x = [S, T, xR] = \begin{cases} [STR] & \text{if } x \in R, \\ [STS] & \text{if } x \in S, \\ [STT] & \text{if } x \in T. \end{cases} \tag{3.1.24}$$

Because $N_x \in \mathbb{N}$ we can define $a, b, c \in \mathbb{N}$ by

$$[STR] = ma, \qquad [STS] = mb, \qquad [STT] = mc. \tag{3.1.25}$$

Or equivalently

$$N_x = \begin{cases} a & \text{if } x \in R, \\ b & \text{if } x \in S, \\ c & \text{if } x \in T. \end{cases} \tag{3.1.26}$$

Lastly we also define the integer

$$k = 3a - m. \tag{3.1.27}$$

The following lemma gives some useful relations between the complex numbers α_1, α_2 and α_3 and integers a, b, c and k.

Lemma 3.1.3. *The following relations hold:*

$$\begin{aligned} \alpha_2\alpha_3 &= a\alpha_1 + b\alpha_2 + c\alpha_3, \\ \alpha_1\alpha_3 &= a\alpha_2 + b\alpha_3 + c\alpha_1, \\ \alpha_1\alpha_2 &= a\alpha_3 + b\alpha_1 + c\alpha_2, \end{aligned} \tag{3.1.28}$$

$$\alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_3 = -m, \tag{3.1.29}$$

$$\alpha_1^2 + \alpha_2^2 + \alpha_3^2 = 1 + 2m, \tag{3.1.30}$$

$$\alpha_1\alpha_2\alpha_3 = \frac{a + km}{3}. \tag{3.1.31}$$

Proof. We will prove the relations in the order they are given.

Using the above definition of N_x, we find

$$\alpha_2\alpha_3 = \sum_{s \in S} \sum_{t \in T} \zeta^s \zeta^t = \sum_{x \in \mathbb{F}_p} N_x \zeta^x. \tag{3.1.32}$$

Now using our definitions of a, b and c we find indeed the first relation:

$$\begin{aligned} \alpha_2\alpha_3 &= a \sum_{x \in R} \zeta^x + b \sum_{x \in S} \zeta^x + c \sum_{x \in T} \zeta^x \\ &= a\alpha_1 + b\alpha_2 + c\alpha_3. \end{aligned} \tag{3.1.33}$$

Similarly, for α_1 and α_3 we find

$$\alpha_1\alpha_3 = \sum_{x \in \mathbb{F}_p} N'_x \zeta^x, \tag{3.1.34}$$

where N'_x is the number of pairs (r,t) with $r \in R$ and $t \in T$, such that $r+t=x$. Again, N'_x only depends on the coset R, S or T in which x lies, so

$$mN'_x = [R,T,xR] = \begin{cases} [RTR] & \text{if } x \in R, \\ [RTS] & \text{if } x \in S, \\ [RTT] & \text{if } x \in T. \end{cases} \tag{3.1.35}$$

Now if we let $t \in T$ we have $T = tR$ and $S = t^2R$, as we have seen before. Then we find

$$\begin{aligned} &[RTR] = [TRR] = [tT,tR,tR] = [STT] = mc, \\ &[RTS] = [STR] = ma, \\ &[RTT] = [TRT] = [tT,tR,tT] = [STS] = mb. \end{aligned} \tag{3.1.36}$$

So similarly to equation (3.1.33) we obtain

$$\alpha_1\alpha_3 = a\alpha_2 + b\alpha_3 + c\alpha_1. \tag{3.1.37}$$

For α_1 and α_2, a similar calculation gives

$$\alpha_1\alpha_2 = a\alpha_3 + b\alpha_1 + c\alpha_2. \tag{3.1.38}$$

Adding up equations (3.1.33), (3.1.37) and (3.1.38) we get

$$\alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_3 = (a+b+c)(\alpha_1+\alpha_2+\alpha_3). \tag{3.1.39}$$

Now using $\mathbb{F}_p = \{0\} \sqcup R \sqcup S \sqcup T$ and $[ST\{0\}] = 0$, we have

$$m(a+b+c) = [STR] + [STS] + [STT] = [ST\mathbb{F}_p] - [ST\{0\}] = m^2. \tag{3.1.40}$$

Therefore $(a+b+c) = m$. Note that

$$0 = \zeta^p - 1 = (\zeta - 1)(\zeta^{p-1} + \zeta^{p-2} + \ldots + \zeta + 1). \tag{3.1.41}$$

Since $\zeta \neq 1$, we find that $\zeta^{p-1} + \zeta^{p-2} \ldots + \zeta + 1 = 0$. Hence

$$\alpha_1 + \alpha_2 + \alpha_3 = \sum_{x \in \mathbb{F}_p^*} \zeta^x = -1. \tag{3.1.42}$$

Therefore we find indeed

$$\alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_3 = -m. \tag{3.1.43}$$

Now, because we know the sum of the alpha's and the sum of their pairwise products, computing the sum of their squares is relatively easy:

$$\alpha_1^2 + \alpha_2^2 + \alpha_3^2 = (\alpha_1+\alpha_2+\alpha_3)^2 - 2(\alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_3) = 1 + 2m. \tag{3.1.44}$$

Lastly, we want to know $\alpha_1\alpha_2\alpha_3$. we have

$$\alpha_1\alpha_2\alpha_3 = \begin{cases} \alpha_1(\alpha_2\alpha_3) = \alpha_1(a\alpha_1 + b\alpha_2 + c\alpha_3), \\ \alpha_2(\alpha_1\alpha_3) = \alpha_2(a\alpha_2 + b\alpha_3 + c\alpha_1), \\ \alpha_3(\alpha_1\alpha_2) = \alpha_3(a\alpha_3 + b\alpha_1 + c\alpha_2). \end{cases} \tag{3.1.45}$$

Adding the three equations on the right and using equations (3.1.43) and (3.1.44), we see that

$$\begin{aligned}3\alpha_1\alpha_2\alpha_3 &= a(\alpha_1^2+\alpha_2^2+\alpha_3^2)+(b+c)(\alpha_1\alpha_2+\alpha_1\alpha_3+\alpha_2\alpha_3),\\ &= a(1+2m)-m(b+c),\\ &= a+(2a-b-c)m,\\ &= a+(3a-m)m.\end{aligned}\tag{3.1.46}$$

Using our definition $k = 3a - m$ we find that indeed

$$\alpha_1\alpha_2\alpha_3 = \frac{a+km}{3}.\tag{3.1.47}$$

□

Using the relations from Lemma 3.1.3, we prove the following lemma:

Lemma 3.1.4. *Let $A = 3k-2$ and $B = b-c$. Then A and B are integers, and we have*

$$M_p = p+1+A,\tag{3.1.48}$$

and

$$4p = A^2+27B^2.\tag{3.1.49}$$

Proof. We let $A = 3k-2$ and $B = b-c$. Now, because k, b and c are integers, it follows that A and B are both integers. From Lemma 3.1.2 we know

$$M_p = 9\frac{[RTS]}{m}.\tag{3.1.50}$$

We also defined the integer a by $am = [STR] = [RTS]$. Therefore we find $M_p = 9a$. We can write a in terms of k as $3a = k+m$. Hence $M_p = 9a = 3k+3m$. Now, if we let $A = 3k-2$, we obtain indeed

$$M_p = A+2+3m = p+1+A.\tag{3.1.51}$$

We define the polynomial $F(t)$ with roots $\alpha_1, \alpha_2, \alpha_3$ as

$$\begin{aligned}F(t) &:= (t-\alpha_1)(t-\alpha_2)(t-\alpha_3)\\ &= t^3+t^2(-\alpha_1-\alpha_2-\alpha_3)+t(\alpha_1\alpha_2+\alpha_1\alpha_3+\alpha_2\alpha_3)-\alpha_1\alpha_2\alpha_3\\ &= t^3+t^2-mt-\frac{a+km}{3},\end{aligned}\tag{3.1.52}$$

where we used the relations from Lemma 3.1.3. We write D_F for the discriminant of F. See Appendix A if you are not familiar with discriminants. Using the definition of the discriminant given in Equation (A.2.3) and Lemma 3.1.3, we find that the square root of D_F is equal to

$$\begin{aligned}\pm\sqrt{D_F} &= (\alpha_1-\alpha_2)(\alpha_1-\alpha_3)(\alpha_2-\alpha_3)\\ &= \alpha_1^2(\alpha_2-\alpha_3)-\alpha_2^2(\alpha_1-\alpha_3)+\alpha_3^2(\alpha_1-\alpha_2)\\ &= \alpha_2\alpha_3(\alpha_2-\alpha_3)+\alpha_1\alpha_3(\alpha_3-\alpha_1)+\alpha_1\alpha_2(\alpha_1-\alpha_2)\\ &= (a\alpha_1+b\alpha_2+c\alpha_3)(\alpha_2-\alpha_3)+(a\alpha_2+b\alpha_3+c\alpha_1)(\alpha_3-\alpha_1)\\ &\quad+(a\alpha_3+b\alpha_1+c\alpha_2)(\alpha_1-\alpha_2)\\ &= a(\alpha_1(\alpha_2-\alpha_3)+\alpha_2(\alpha_3-\alpha_1)+\alpha_3(\alpha_1-\alpha_2))\\ &\quad+b(\alpha_2(\alpha_2-\alpha_3)+\alpha_3(\alpha_3-\alpha_1)+\alpha_1(\alpha_1-\alpha_2))\\ &\quad+c(\alpha_3(\alpha_2-\alpha_3)+\alpha_1(\alpha_3-\alpha_1)+\alpha_2(\alpha_1-\alpha_2))\\ &= a(0)+(b-c)(\alpha_1^2+\alpha_2^2+\alpha_3^2-\alpha_1\alpha_2-\alpha_1\alpha_3-\alpha_2\alpha_3)\\ &= (b-c)(1+3m)\\ &= Bp.\end{aligned}\tag{3.1.53}$$

So we have $D_F = B^2p^2$.

Now we define

$$\beta_i := 1 + 3\alpha_i, \tag{3.1.54}$$

where $i = 1, 2, 3$. Then

$$\beta_1 + \beta_2 + \beta_3 = 3 + 3(-1) = 0, \tag{3.1.55}$$

and

$$\begin{aligned}\beta_1\beta_2 + \beta_1\beta_3 + \beta_2\beta_3 &= 3 + 6(\alpha_1 + \alpha_2 + \alpha_3) + 9(\alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_3)\\ &= -3 - 9m\\ &= 3p,\end{aligned} \tag{3.1.56}$$

and

$$\begin{aligned}\beta_1\beta_2\beta_3 &= 1 + 3(\alpha_1 + \alpha_2 + \alpha_3) + 9(\alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_3) + 27\alpha_1\alpha_2\alpha_3\\ &= -2 - 9m + 9(a + km)\\ &= -2 - 6m + 3k + 9km\\ &= (3k - 2)(3m + 1)\\ &= Ap.\end{aligned} \tag{3.1.57}$$

We define the polynomial $G(t)$ with roots $\beta_1, \beta_2, \beta_3$ as

$$\begin{aligned}G(t) &:= (t - \beta_1)(t - \beta_2)(t - \beta_3)\\ &= t^3 + t^2(-\beta_1 - \beta_2 - \beta_3) + t(\beta_1\beta_2 + \beta_1\beta_3 + \beta_2\beta_3) - \beta_1\beta_2\beta_3\\ &= t^3 - 3pt - Ap.\end{aligned} \tag{3.1.58}$$

Let D_G be the discriminant of G. From Equation (A.2.8) it follows that

$$D_G = -4(-3p)^3 - 27(Ap)^2 = 4 \cdot 27p^3 - 27A^2p^2. \tag{3.1.59}$$

But, using the definition of the discriminant A.2.3 and the fact that $\beta_i - \beta_j = 3(\alpha_i - \alpha_j)$, we obtain $D_G = 27^2 D_F$. Hence

$$D_G = 27^2B^2p^2. \tag{3.1.60}$$

So combining (3.1.59) and (3.1.60) we get

$$4 \cdot 27p^3 - 27A^2p^2 = 27^2B^2p^2. \tag{3.1.61}$$

Now, dividing by $27p^2$, we find indeed

$$4p = A^2 + 27B^2. \tag{3.1.62}$$

□

With Lemma 3.1.4 we have found integers A and B satisfying almost all the conditions from Theorem 3.1.1(b). The only thing left to show is that these A and B we found are in fact unique as a solution for Equation (3.1.4) up to changing their signs. This uniqueness is ensured by the following lemma.

Lemma 3.1.5. *Suppose we have two integers A, B such that*

$$4p = A^2 + 27B^2. \tag{3.1.63}$$

Then the pair (A, B) is uniquely determined up to sign changes of A and B.

Proof. Let A, B be two integers such that

$$4p = A^2 + 27B^2. \tag{3.1.64}$$

Suppose we have another pair (A_1, B_1) such that

$$4p = A_1^2 + 27B_1^2. \tag{3.1.65}$$

Then we have

$$\begin{aligned} 4p(B_1^2 - B^2) &= (A^2 + 27B^2)B_1^2 - (A_1^2 + 27B_1^2)B^2 \\ &= (AB_1 + A_1B)(AB_1 - A_1B). \end{aligned} \tag{3.1.66}$$

Since p divides the left-hand side, at least one of the factors on the right-hand side is divisible by p.

First suppose $p \mid (AB_1 - A_1B)$. Multiplying Equations (3.1.64) and (3.1.65) we get

$$16p^2 = A^2A_1^2 + 27B^2A_1^2 + 27B_1^2A^2 + 27^2B^2B_1^2. \tag{3.1.67}$$

We can rewrite this to

$$16p^2 - (AA_1 + 27BB_1)^2 = 27(AB_1 - A_1B)^2. \tag{3.1.68}$$

Using that $p \mid (AB_1 - A_1B)$ we find that $p \mid (AA_1 + 27BB_1)$. Dividing by p^2 we obtain

$$16 - \left(\frac{AA_1 + 27BB_1}{p}\right)^2 = 27\left(\frac{AB_1 - A_1B}{p}\right)^2. \tag{3.1.69}$$

Notice that the left hand side is not greater than 16, but the right hand side is 27 times the square of an integer. This implies that both sides must be equal to 0. Therefore, we have $AB_1 = A_1B$. Let

$$\lambda = \frac{A_1}{A} = \frac{B_1}{B}. \tag{3.1.70}$$

Then $A_1 = \lambda A$ and $B_1 = \lambda B$, so we get

$$4p = A_1^2 + 27B_1^2 = \lambda^2(A^2 + 27B^2) = \lambda^2 4p, \tag{3.1.71}$$

which implies $\lambda = \pm 1$. So the pair A_1, B_1 is equal to A, B, up to sign changes.

Now suppose $p \mid (AB_1 + A_1B)$. This time, we rewrite Equation (3.1.67) to

$$16p^2 - (AA_1 - 27BB_1)^2 = 27(AB_1 + A_1B)^2. \tag{3.1.72}$$

Using that $p \mid (AB_1 + A_1B)$, we find $p \mid (AA_1 - 27BB_1)$. Now, with the same reasoning as before, we get $AB_1 = -A_1B$. If we let

$$\lambda' = \frac{A_1}{A} = -\frac{B_1}{B}, \tag{3.1.73}$$

then $A_1 = \lambda' A$ and $B_1 = -\lambda' B$. However, doing the same as in (3.1.71) we find $4p = (\lambda')^2 4p$. Therefore we have $\lambda' = \pm 1$. Hence the pair (A_1, B_1) is equal to (A, B) up to sign changes.

We conclude that (A, B) is indeed unique as as a solution pair of Equation (3.1.64) up to changing the signs of A and B. □

Now the proof of part (b) of Theorem 3.1.1 only consists of invoking the last two lemmas.

Proof of Theorem 3.1.1(b). From Lemma 3.1.4 we know that if we let $A = 3k - 2$ and $B = b - c$, then they satisfy

$$4p = A^2 + 27B^2. \tag{3.1.74}$$

From Lemma 3.1.5 it follows that the pair (A, B) is the unique solution to (3.1.74) up to changing the signs of A and B. By letting $A = 3k - 2$ we have fixed the sign of A, and we fixed it such that $A \equiv 1 \pmod 3$. From Lemma 3.1.4 we know that in this case

$$M_p = p + 1 + A, \tag{3.1.75}$$

which completes the proof of the theorem. □

3.2 The Group of Points on a Curve in $\mathbb{P}^2(\mathbb{F}_{19})$

In this section we will consider the elliptic curve

$$C(\mathbb{F}_{19}) : X^3 + Y^3 + Z^3 = 0 \tag{3.2.1}$$

in the projective plane $\mathbb{P}^2(\mathbb{F}_{19})$. It can be shown that $C(\mathbb{F}_{19})$ is smooth, and we have $[1, -1, 0] \in C(\mathbb{F}_{19})$. Hence, the curve $C(\mathbb{F}_{19})$ is indeed an elliptic curve. This curve is not in Weierstrass normal form, but from the previous chapter we know that it is possible to transform $C(\mathbb{F}_{19})$ to a curve in Weierstrass normal form, which we will denote by $C_1(\mathbb{F}_{19})$. Recall that this transformation is a one-to-one correspondence between the points on $C(\mathbb{F}_{19})$ and $C_1(\mathbb{F}_{19})$. First we will determine how many points there are on $C(\mathbb{F}_{19})$, and then we will use this to find the group structure of the points on $C_1(\mathbb{F}_{19})$.

Just like before, we denote the number of points on $C(\mathbb{F}_{19})$ by M_{19}. We know that $19 \equiv 1 \pmod 3$. Therefore, using the theorem we proved in the previous section, we know that there exist integers A, B such that

$$4 \cdot 19 = A^2 + 27B^2, \tag{3.2.2}$$

and A and B are unique up to changing their signs. Further, we can fix the sign of A such that $A \equiv 1 \pmod 3$, and then we have

$$M_{19} = 20 + A. \tag{3.2.3}$$

For large primes p, it can be quite hard to actually find the integers A and B such that

$$4p = A^2 + 27B^2. \tag{3.2.4}$$

However, in our case that $p = 19$, one quickly finds the solution

$$A = 7, \qquad B = 1. \tag{3.2.5}$$

Then we find $M_{19} = 27$. So the curve $C(\mathbb{F}_{19})$ consists of 27 points, and therefore also $|C_1(\mathbb{F}_{19})| = 27$. Now, we want to determine the group structure of $C_1(\mathbb{F}_{19})$. We know that the group will be abelian, so it must be isomorphic to a product of cyclic groups [6]. For a group with 27 elements, we get three options

$$\mathbb{Z}/27\mathbb{Z}, \qquad \mathbb{Z}/9\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}, \qquad \mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}. \tag{3.2.6}$$

We write $\overline{\mathbb{F}}_{19}$ for the algebraic closure of $\mathbb{F}_{19}$, and we denote by $C_1(\mathbb{F}_{19})[n]$ for some $n \in \mathbb{N}$ the subset of $C_1(\mathbb{F}_{19})$ consisting of the points x such that $nx = x + \cdots + x = \mathcal{O}$. To see which one of the groups above is isomorphic to $C_1(\mathbb{F}_{19})$, we will use the following theorem for elliptic curves.

Theorem 3.2.1. *Let $E(\mathbb{F}_p)$ be an elliptic curve in $\mathbb{P}^2(\mathbb{F}_p)$ for some prime p, and let $n \in \mathbb{N}$ be an integer such that $\gcd(n,p) = 1$. Then we have*

$$E(\overline{\mathbb{F}}_p)[n] \cong \mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}, \tag{3.2.7}$$

where $E(\overline{\mathbb{F}}_p)$ is the elliptic curve in $\mathbb{P}^2(\mathbb{F}_p)$ given by the same equation as $E(\mathbb{F}_p)$.

Proof. For the proof of this theorem, see Corollary III.6.4 in Silverman [3]. □

Now, because $|C_1(\mathbb{F}_{19})| = 27$, we know that every for every point $x \in C_1(\mathbb{F}_{19})$ we have $27x = \mathcal{O}$. In other words, we have $C_1(\mathbb{F}_{19})[27] = C_1(\mathbb{F}_{19})$. On the other hand, we know that $C_1(\mathbb{F}_{19})[27] \leq C_1(\overline{\mathbb{F}}_{19})[27]$, because $\mathbb{F}_{19} \subset \overline{\mathbb{F}}_{19}$. Hence, using the above theorem, we find

$$C_1(\mathbb{F}_{19}) \leq \mathbb{Z}/27\mathbb{Z} \times \mathbb{Z}/27\mathbb{Z}. \tag{3.2.8}$$

Now, note that the group on the right-hand side of the above equation has exactly nine elements of order three (also including $(0,0)$). So therefore $C_1(\mathbb{F}_{19})$ can have at most 9 points of order three. However, the group $\mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/3Z$ has 27 points of order three, if we also count $(0,0,0)$. Hence, we see that $C_1(\mathbb{F}_{19})$ cannot be isomorphic to $\mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/3Z$.

Now there are two options left for the group structure of $C_1(\mathbb{F}_{19})$, namely $\mathbb{Z}/27\mathbb{Z}$ and $\mathbb{Z}/9\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$. Note that

$$|(\mathbb{Z}/27\mathbb{Z})[3]| = 3, \quad \text{whereas} \quad |(\mathbb{Z}/9\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z})[3]| = 9. \tag{3.2.9}$$

We will show that $C_1(\mathbb{F}_{19})$ has a subgroup isomorphic to $\mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$, which implies

$$|C_1(\mathbb{F}_{19})[3]| \geq 9. \tag{3.2.10}$$

So then it follows that $C_1(\mathbb{F}_{19}) \cong \mathbb{Z}/9\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$.

In the previous section we saw that if $p \equiv 1 \pmod 3$, then every non-zero cube has exactly three different roots in $\mathbb{F}_p$. We have $1^3 = 1$ and $(-1)^3 = (-1)$, so both 1 and -1 must have two more cubic roots in $\mathbb{F}_p$. In $\mathbb{F}_{19}$, we have

$$1^3 = 7^3 = 11^3 = 1, \quad \text{and} \quad (-1)^3 = (-7)^3 = (-11)^3 = -1. \tag{3.2.11}$$

From Equation (3.2.1), it follows that if we take one of the variables X, Y and Z to be zero and the other two to be cubic roots of 1 and -1, then $[X,Y,Z] \in C(\mathbb{F}_{19})$. Now, for such a point $[X,Y,Z]$, there are three options where you can place 0. After choosing which variable is 0, there are two places left. In one of them we must place a cubic root of 1, and in the other a cubic root of -1. Because we are working with homogeneous coordinates, we can always multiply the coordinates with a constant such that the first non-zero coordinate will be a 1. Then the other coordinate must be a cubic root of -1, for which there are three options. So in total we find 9 different points on the curve of this form, given by

$$\begin{array}{rrr} [1,-1,0] & [1,0,-1] & [0,1,-1] \\ [1,8,0] & [1,0,8] & [0,1,8] \\ [1,12,0] & [1,0,12] & [0,1,12]. \end{array} \tag{3.2.12}$$

Here we used that $-7 = 12$ and $-11 = 8$ in $\mathbb{F}_{19}$.

Proposition 3.2.2. *The nine points in $C_1(\mathbb{F}_{19})$ corresponding to the nine points on $C(\mathbb{F}_{19})$ given in Equation (3.2.12) form a subgroup of $C_1(\mathbb{F}_{19})$ isomorphic to $\mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$.*

Proof. We will prove this proposition using the explicit formulas for the group law we gave in the previous chapter. To do this, we first have to transform $C(\mathbb{F}_{19})$ into Weierstrass normal form, so that we have an explicit formula for $C_1(\mathbb{F}_{19})$. We can do this by taking

$$X_1 = 7Z, \qquad Y_1 = -2(X-Y), \qquad Z_1 = X+Y. \tag{3.2.13}$$

We can rewrite the equation $X^3+Y^3+Z^3=0$ as

$$4(X^3+Y^3-XY^2-YX^2) = Z^3+5(X^3+Y^3+3XY^2+3YX^2). \tag{3.2.14}$$

Remember that we are working modulo 19, so $15 \equiv -4 \pmod{19}$. Now, using Equation (3.2.13), we see that the above equation is equivalent to

$$Y_1^2 Z_1 = X_1^3 + 5Z_1^3. \tag{3.2.15}$$

Therefore, using the coordinate transformation in Equation (3.2.13), we have transformed the curve $C(\mathbb{F}_{19})$ into the curve

$$C_1(\mathbb{F}_{19}) : Y_1^2 Z_1 = X_1^3 + 5Z_1^3 \tag{3.2.16}$$

which is in Weierstrass normal form. One can check that the transformations in Equation (3.2.13) indeed give a one-to-one correspondence between the points on $C(\mathbb{F}_{19})$ and $C_1(\mathbb{F}_{19})$. Using Equation (3.2.13), we can determine the points on $C_1(\mathbb{F}_{19})$ that correspond to the points in Equation (3.2.12). Keeping the same order as in Equation (3.2.12), we get the following points in $C_1(\mathbb{F}_{19})$:

$$\begin{array}{rrr} \mathcal{O} & [-7,-2,1] & [-7,2,1] \\ [0,-5,9] & [-1,-2,1] & [-1,2,1] \\ [0,3,-6] & [8,-2,1] & [8,2,1]. \end{array} \tag{3.2.17}$$

Here we write $\mathcal{O} = [0,1,0]$ just like in the previous chapter. As we saw before, the point $\mathcal{O}$ will be the identity element of $C_1(\mathbb{F}_{19})$. Now, in order to use our explicit formulas for the addition of two points, we have to dehomogenize the curve $C_1(\mathbb{F}_{19})$ with respect to Z_1. Then we get that the affine part of $C_1(\mathbb{F}_{19})$ is given by

$$C_1'(\mathbb{F}_{19}) : y^2 = x^3 + 5, \tag{3.2.18}$$

and the nine points above become

$$\begin{array}{rrr} \mathcal{O} & (-7,-2) & (-7,2) \\ (0,10) & (-1,-2) & (-1,2) \\ (0,9) & (8,-2) & (8,2). \end{array} \tag{3.2.19}$$

Now, remember that if we have two points on the affine part of a curve in Weierstrass normal form with the same x-coordinate but different y-coordinates, then these points are inverses of each other. So we see that apart from $\mathcal{O}$, we have four pairs of points that are each others inverse. Using Equations (2.4.9) and (2.4.15), it is easily verified that if we take one of these points p, then $2p = -p$. So all of these points have order three. We conclude that these points form indeed a subgroup of $C_1(\mathbb{F}_{19})$ isomorphic to $\mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$. □

Corollary 3.2.3. *We have*

$$C_1(\mathbb{F}_{19}) \cong \mathbb{Z}/9\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}. \tag{3.2.20}$$

In this section we determined the group structure only for the curve $C_1(\mathbb{F}_{19})$ in Weierstrass normal form. It turns out that in fact it is possible to also define a group structure on the curve $C(\mathbb{F}_{19})$ itself, similar to how we defined it for curves in Weierstrass normal form. Then we have to make a choice for the identity element of this group. Suppose we fix the point $[1, -1, 0] \in C(\mathbb{F}_{19})$ as the identity element. Then it turns out that there exists a group isomorphism between $C(\mathbb{F}_{19})$ and $C_1(\mathbb{F}_{19})$. This isomorphism then has to send $[1, -1, 0]$ to $\mathcal{O}$. To explain this in detail would lead too far, but it is important to realize that for every elliptic curve we can define a group structure on its set of points, not just for curves in Weierstrass normal form [3].

Chapter 4

Elliptic Functions

In this chapter we will consider elliptic functions. First we will give some background in complex analysis, and then we will give the definition of an elliptic function and consider the Weierstrass $\wp$-function as an example. In Section 4.4 we will show the connection between elliptic curves and elliptic functions. Lastly, in Section 4.5 we introduce the Jacobi elliptic functions which we will use in Chapter 6, and we explain where the name 'elliptic' comes from.

4.1 Basic Results From Complex Analysis

In this part of the thesis we will look at complex-valued functions. We do not want to dive to deep into complex analysis, but we will give some (simplified) basic definitions and results. For a more detailed introduction to complex analysis one could for example look at the first chapters of Lang [8].

Definition 4.1.1. *A function $f : \mathbb{C} \to \mathbb{C}$ is called **holomorphic** on an open subset $U \subset \mathbb{C}$ if it has a complex derivative at every point $z_0 \in U$.*

Note that is not immediately clear what it means for a function to have a complex derivative at a point, but here we will skip over that. If a function $f(z) : \mathbb{C} \to \mathbb{C}$ is holomorphic, then we can calculate $f'(z)$ completely similar to how we calculate the derivative of a real-valued function. For instance the sum rule, product rule and chain rule all still hold.

Definition 4.1.2. *Let $f : \mathbb{C} \to \mathbb{C}$ be a complex-valued function, and let $U \subset \mathbb{C}$ be an open subset of $\mathbb{C}$. Suppose that for every point $z_0 \in U$ there exists a power series*

$$\sum_{n=0}^{\infty} a_n(z - z_0)^n \tag{4.1.1}$$

*and some $r > 0$ such that the series converges absolutely to $f(z)$ for $|z - z_0| < r$. Then the function f is called **analytic** on U*

Now a very important result in complex analysis is that it actually means the same for a function to be holomorphic or analytic.

Theorem 4.1.3. *Let $f(z) : \mathbb{C} \to \mathbb{C}$ be a function. Then f is analytic on an open subset $U \subset \mathbb{C}$ if and only if it is holomorphic on U.*

Proof. We will not give the entire proof of this theorem, but we will show that every analytic function is holomorphic. The converse is shown in Theorem III.7.2 of Lang [8].

Suppose f is analytic in a point z_0 in U. Then in a neighborhood of z_0, we can write f as a power series. A power series is differentiable, because we can just differentiate term by term. Therefore f is also holomorphic at z_0. So if f is analytic on U, then f is holomorphic on U. □

A function that is holomorphic on all of $\mathbb{C}$ is called *entire*. We have the following theorem for bounded entire functions.

Theorem 4.1.4 (Liouville's Theorem)**.** *A bounded entire function is constant.*

Proof. For the proof of this theorem, see Theorem III.7.5 in Lang [8]. □

Now, suppose that we have a function that is holomorphic on an open subset $U \subset \mathbb{C}$, except for one point $z_0 \in U$ where f is not defined. We call such a point z_0 a *singular point*. For example this happens with the function $1/z$ if U is a disk centered at 0 and $z_0 = 0$.

Definition 4.1.5. *Suppose that we have a function f that has a singular point z_0, such that in a neighborhood of z_0 we can write*

$$f(z) = \frac{a_{-m}}{(z-z_0)^m} + \cdots + a_0 + a_1(z-z_0) + \cdots . \tag{4.1.2}$$

*So we can write f as a power series with a finite amount of terms with a negative exponent, in this case m. Then we say that f has a **pole of order** m at the point z_0. The coefficient a_{-1} is called the **residue** of the pole. A pole of order 1 is also called a **simple pole**.*

We denote the order and residue of a pole of a function f at the point z_0 by $\mathrm{Ord}_{z_0} f$ and $\mathrm{Res}_{z_0} f$, respectively. Note that from the above definition it immediately follows that a simple pole has non-zero residue.

Next, we define what is means for a function to be *meromorphic*.

Definition 4.1.6. *Let f be a function defined on an open subset $U \subset \mathbb{C}$, except for a discrete set of points $S \subset U$ that are poles of f. Suppose that f is holomorphic on $U \setminus S$. Then the function f is called **meromorphic** on U.*

We finish this section with a theorem that ensures the uniqueness of an analytic continuation.

Theorem 4.1.7. *Let $D \subset \mathbb{C}$ be a subset of the complex numbers, and let $M \subset D$ be a subset of D that is not discrete. Let $f : M \to \mathbb{C}$ be a function. Then if there exists an analytic function $\widetilde{f} : D \to \mathbb{C}$ such that $\widetilde{f}(z) = f(z)$ for $z \in M$, then $\widetilde{f}$ is unique with this property. We call $\widetilde{f}$ the analytic continuation of f.*

Proof. For the proof of this theorem, see the beginning of Section III.2 in Freitag and Busam [9]. □

4.2 Doubly Periodic Functions

Now we will start looking at elliptic functions. This part of the chapter, up to and including Section 4.4, is mainly based on Chapter VI of Silverman [3], Chapter V up to Section V.4 of Freitag and Busam [9], Sections I.3 to I.6 of Koblitz [10] and Sections 9.1 and 9.2 of Washington [11]. The material covered in these sources is very similar, but different enough to complement each other.

Let ω_1 and ω_2 be two non-zero complex numbers. We say that ω_1 and ω_2 are *linearly independent over* $\mathbb{R}$ if there exists no real number $r \in \mathbb{R}$ such that $\omega_1 = r\omega_2$. Note that this is equivalent to saying that ω_1 and ω_2 are both non-zero and their quotient ω_1/ω_2 is not a real number.

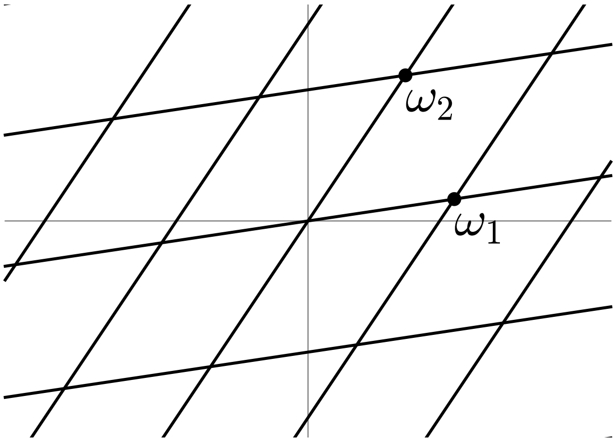

Figure 4.1: The lattice L generated by ω_1 and ω_2 in the complex plane.

Definition 4.2.1. *Let $\omega_1, \omega_2 \in \mathbb{C}$ be two complex numbers that are linearly independent over $\mathbb{R}$. Then the set*

$$L = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2 = \{n_1\omega_1 + n_2\omega_2 \mid n_1, n_2 \in \mathbb{Z}\} \tag{4.2.1}$$

is called a ***lattice****. We say that ω_1 and ω_2 generate L. See Figure 4.1.*

Note that a lattice only consists of the intersection points of the lines in Figure 4.1. We see that a lattice divides the complex plane into parallelograms. The parallelogram that has both ω_1 and ω_2 in its boundary is called the *fundamental parallelogram* for the lattice L, denoted by $\mathcal{F}$. So we have

$$\mathcal{F} = \{a_1\omega_1 + a_2\omega_2 \mid a_1, a_2 \in \mathbb{R},\ \ 0 \leq a_1, a_2 < 1\}\,. \tag{4.2.2}$$

Note that we defined $\mathcal{F}$ with two closed edges and two open edges. We denote the closure of $\mathcal{F}$ in $\mathbb{C}$ by $\overline{\mathcal{F}}$.

We can take the quotient of the complex plane with a lattice L, denoted by $\mathbb{C}/L$. This space is obtained by quotienting out the following equivalence relation on $\mathbb{C}$:

$$z_1 \sim z_2 \quad \text{if and only if} \quad z_1 - z_2 \in L. \tag{4.2.3}$$

This essentially means that we identify all the parallelograms with each other in their original orientation. Therefore the space $\mathbb{C}/L$ can be obtained from the (closure of the) fundamental parallelogram $\mathcal{F}$ for the lattice L by identifying (or "gluing") the opposite sides of $\mathcal{F}$. Hence we see that the space $\mathbb{C}/L$ is topologically equivalent to a torus. Also note that for every element $[z] \in \mathbb{C}/L$ there is a unique point $z_0 \in \mathcal{F}$ such that $[z_0] = [z]$. Here we denote by $[z]$ the equivalence class in $\mathbb{C}/L$ of a point $z \in \mathbb{C}$. Conversely, for every point $z_0 \in \mathcal{F}$ there exists a unique equivalence class $[z] \in \mathbb{C}/L$ such that $[z_0] = [z]$. Therefore we see that $\mathcal{F}$ is isomorphic to $\mathbb{C}/L$.

Definition 4.2.2. *Let $\omega_1, \omega_2 \in \mathbb{C}$ be two linearly independent complex numbers, and let $L \subset \mathbb{C}$ be their corresponding lattice. An* ***elliptic function*** *for the lattice L is a meromorphic function $f : \mathbb{C} \to \mathbb{C}$ with the property*

$$f(z + \omega) = f(z) \quad \textit{for} \quad \omega \in L \textit{ and } z \in \mathbb{C}. \tag{4.2.4}$$

Note that in order to have $f(z+\omega) = f(z)$ for all $\omega \in L$, it is enough to check that

$$f(z+\omega_1) = f(z+\omega_2) = f(z). \tag{4.2.5}$$

So an elliptic function f has two linearly independent periods ω_1 and ω_2. Therefore elliptic functions are also called *doubly periodic*. This also implies that in each parallelogram in Figure 4.1 the elliptic function f with periods ω_1 and ω_2 attains exactly the same values. So when studying an elliptic fuction, we only have to look at how it behaves on its corresponding fundamental parallelogram. Using this, we can prove the following theorem for elliptic functions.

Theorem 4.2.3. *Any elliptic function without poles is constant. Similarly, any elliptic function without zeros is constant.*

Proof. Suppose that we have an elliptic function with periods $\omega_1, \omega_2 \in \mathbb{C}$ that has no poles, i.e. f is an entire function. We write L for the lattice generated by ω_1 and ω_2, and $\mathcal{F}$ for its fundamental parallelogram. As we noted before, for every value $a \in \mathbb{C}$ the function f can take there is a point $z_0 \in \mathcal{F}$ such that $f(z_0) = a$. The closure of the fundamental parallelogram is a closed and bounded subset of $\mathbb{C}$, and therefore compact. Hence, the function f attains a maximum on $\overline{\mathcal{F}}$, so f is bounded on $\overline{\mathcal{F}}$. But we also have $\mathcal{F} \subset \overline{\mathcal{F}}$. Therefore we see that f is a bounded entire function. Then it follows from Theorem 4.1.4 that f is constant.

The proof for the case that f is an elliptic function without zeros goes exactly the same, by noting that in that case $1/f$ is an entire elliptic function. □

Theorem 4.2.4. *An elliptic function has only finitely many poles in its corresponding fundamental parallelogram, and the sum of the residues of these poles vanishes:*

$$\sum_{z_i} \operatorname{Res}_{z_i} f = 0. \tag{4.2.6}$$

Here we write $\{z_i\}$ for the finite set of poles of f in its fundamental parallelogram.

Proof. We will only prove the first part of this theorem. The rest of the proof uses more complex analysis and can be found in for example the proof of Theorem V.1.4 in Freitag and Busam [9] or Theorem VI.2.2 in Silverman [3].

Per definition, the set of poles S of an elliptic function is discrete. The closure of the fundamental parallelogram $\overline{\mathcal{F}}$ is compact, so therefore the intersection $\overline{\mathcal{F}} \cap S$ contains only finitely many elements. Hence $\mathcal{F}$ itself also contains only finitely many poles. □

This last theorem has an important consequence, but for that we first need the following definition.

Definition 4.2.5. *We define the* ***order of an elliptic function*** *f, denoted by* $\operatorname{Ord} f$*, as the sum of the orders of its poles in its fundamental parallelogram. In formula:*

$$\operatorname{Ord} f = \sum_{z_i} \operatorname{Ord}_{zi} f. \tag{4.2.7}$$

Here we write again $\{z_i\}$ for the finite set of poles of f in its fundamental parallelogram. If f has no poles, then we say $\operatorname{Ord} f = 0$.

Then we can restate Theorem 4.2.3 as

$$\operatorname{Ord} f = 0 \quad \text{if and only if} \quad f \text{ is constant.} \tag{4.2.8}$$

Further, Theorem 4.2.4 implies that the order of an elliptic function can never be 1, because a simple pole has non-zero residue. This is an important result, so we state:

Proposition 4.2.6. *There exist no elliptic functions of order* 1.

This implies in particular that a non-constant elliptic function must have order at least 2. In the next section we will look for the simplest example of such a function.

4.3 The Weierstrass ℘-Function

In this section we want to find the simplest example of an elliptic curve with given periods $\omega_1, \omega_2 \in \mathbb{C}$. In the previous section we showed that the order of a non-constant elliptic function must be at least 2. For a function of order 2, we have two options: a function with two simple poles in its fundamental parallelogram $\mathcal{F}$, or a function with one pole of order 2 with residue zero in $\mathcal{F}$. We choose to look for the latter. Further, we want our elliptic function to have its pole at the point 0. Then because of the periodicity of an elliptic function, it will have poles precisely at the points in the lattice generated by ω_1 and ω_2. It turns out that the so-called *Weierstrass ℘-function* satisfies these requirements.

Definition 4.3.1. *Let $\omega_1, \omega_2 \in \mathbb{C}$ be two linearly independent complex numbers, and let L be the lattice they generate. Then we define the* ***Weierstrass ℘-function*** *$\wp : \mathbb{C} \setminus L \to \mathbb{C}$ as*

$$\wp(z) = \wp(z; L) = \frac{1}{z^2} + \sum_{\substack{\omega \in L \\ \omega \neq 0}} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right). \tag{4.3.1}$$

For this definition to be of any value, we need the following theorem.

Theorem 4.3.2. *Let L be the lattice generated by $\omega_1, \omega_2 \in \mathbb{C}$. The series*

$$\sum_{\substack{\omega \in L \\ \omega \neq 0}} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right) \tag{4.3.2}$$

converges absolutely and uniformly on compact subsets of $\mathbb{C} \setminus L$, and there it defines a holomorphic function.

Proof. The proof of this theorem can be found in for example the proof of Lemma V.2.3 in Freitag and Busam [9] or Theorem VI.3.1(b) in Silverman [3]. Note that after having proved the absolute uniform convergence of the series, it is clear that it defines a holomorphic function on $\mathbb{C} \setminus L$, because every term is holomorphic on $\mathbb{C} \setminus L$. □

From its definition, we can immediately see that the function $\wp$ has poles of order 2 with residue zero at every point in L. Therefore, using the above theorem, we find that $\wp$ is a meromorphic function on $\mathbb{C}$ with a pole of order 2 with residue zero at each point of L. The following theorem then ensures that $\wp$ is indeed the elliptic function we were looking for.

Theorem 4.3.3. *The Weierstrass ℘-function is an even elliptic function.*

Proof. To see that $\wp$ is an even function, we look at $\wp(-z)$. Note that for every $\omega \in L$ we have that $-\omega \in L$, so we can replace ω by $-\omega$ in Equation (4.3.1). Then, filling in $-z$ and $-\omega$ we see that $\wp(-z) = \wp(z)$, so $\wp$ is indeed even.

The only thing left to show to prove that $\wp$ is an elliptic function, is that it is doubly periodic. For this, we look at its derivative $\wp'(z)$. Because the series in Equation (4.3.1) converges uniformly, we can differentiate $\wp$ term by term. We find

$$\wp'(z) = -2\sum_{\omega\in L} \frac{1}{(z-\omega)^3}. \tag{4.3.3}$$

We see that

$$\wp'(z+\omega) = \wp'(z) \quad \text{for all } \omega \in L. \tag{4.3.4}$$

We can integrate this equality to find

$$\wp(z+\omega) = \wp(z) + c(\omega), \tag{4.3.5}$$

where $c(\omega)$ does not depend on z. We will show that $c(\omega_i) = 0$, for $i \in \{1,2\}$. Then because ω_1 and ω_2 generate L we get that $c(\omega) = 0$ for all $\omega \in L$. Note that $-\omega_i/2 \notin L$, so we can fill in $z = -\omega_i/2$ in the above equation. Then we get

$$\wp\left(-\frac{\omega_i}{2} + \omega_i\right) = \wp\left(-\frac{\omega_i}{2}\right) + c(\omega_i). \tag{4.3.6}$$

This implies that

$$\wp\left(\frac{\omega_i}{2}\right) - \wp\left(-\frac{\omega_i}{2}\right) = c(\omega_i), \tag{4.3.7}$$

and then from the fact that $\wp$ is even it follows that $c(\omega_i) = 0$. Hence, we see that $\wp$ is periodic with ω_1 and ω_2 as periods. We already knew that $\wp$ is a meromorphic function on $\mathbb{C}$ with a pole of order 2 with residue zero at each point in L, so we conclude that $\wp$ is indeed an elliptic function of order 2. □

In the proof of the above theorem, we used the derivative of $\wp$ to say something about $\wp$. However, it turns out the the derivative $\wp'(z)$ is very interesting on its own.

Theorem 4.3.4. *The derivative of the Weierstrass $\wp$-function $\wp'(z)$ is an odd elliptic function of order 3.*

Proof. Recall from Equation (4.3.4) that

$$\wp'(z+\omega) = \wp'(z) \quad \text{for all } \omega \in L. \tag{4.3.8}$$

It is also clear from Equation (4.3.3) that $\wp'(z)$ is meromorphic on $\mathbb{C}$, because it is holomorphic on $\mathbb{C} \setminus L$ and it has a pole of order three in each latticepoint. Hence, $\wp'(z)$ is indeed an elliptic function of order 3. □

We will denote by $\mathbb{C}(L)$ the set of all elliptic functions with period lattice L. For every elliptic function $f \in \mathbb{C}(L)$, we can write f as the sum of an even and an odd elliptic function as follows:

$$f(z) = \frac{f(z)+f(-z)}{2} + \frac{f(z)-f(-z)}{2}. \tag{4.3.9}$$

It turns out that we can write every even elliptic function in terms of the Weierstrass $\wp$-function.

Theorem 4.3.5. *Let $f \in \mathbb{C}(L)$ be an even elliptic function, and let $\wp$ be the Weierstrass $\wp$-function for the lattice L. Then we can write f as*

$$f(z) = \frac{a_n\wp(z)^n + \cdots + a_1\wp(z) + a_0}{b_m\wp(z)^m + \cdots + b_1\wp(z) + b_0} \qquad a_i, b_j \in \mathbb{C}. \tag{4.3.10}$$

In other words, every even elliptic function can be written as a rational function of the Weierstrass $\wp$-function.

Proof. For the proof of this theorem see Propositions V.3.1 and V.3.2 in Freitag and Busam [9], or Theorem 3.2 in Silverman [3]. □

Now suppose that we have an odd elliptic function $f \in \mathbb{C}(L)$. We already know one odd elliptic function, namely $\wp'(z) \in \mathbb{C}(L)$. The quotient of two elliptic functions, where the denominator is not constant 0, is again an elliptic function [9]. Therefore, we get that $g(z) = f(z)/\wp'(z)$ is an elliptic function. Because $g(z)$ is the quotient of two odd functions, it follows that $g(z)$ is even. Therefore, by the theorem above, we can write $g(z)$ as a rational function of $\wp(z)$. Then we find $f(z) = \wp'(z)g(z)$. This leads to the following theorem.

Theorem 4.3.6. *Let $f \in \mathbb{C}(L)$ be an elliptic function. Then there exist rational functions R and S with complex coefficients, such that*

$$f = R(\wp) + \wp' S(\wp). \tag{4.3.11}$$

Proof. Let $f \in \mathbb{C}(L)$ be an elliptic function. Then, using Equation (4.3.9), we can write f as the sum of an even function f_1 and an odd function f_2

$$f = f_1 + f_2. \tag{4.3.12}$$

Now, as we saw before, we can write any odd function as the product of $\wp'$ and an even function g. So we find

$$f_2 = \wp' g. \tag{4.3.13}$$

Then we use Theorem 4.3.5 to write the even functions f_1 and g as rational functions of $\wp$. We write $f_1 = R(\wp)$ and $g = S(\wp)$. Filling this in into Equation (4.3.12) finishes the proof of the theorem. □

4.4 From Elliptic Functions to Elliptic Curves

In this section we will explore the link between elliptic functions and elliptic curves. We start by showing that the Weierstrass $\wp$-function satisfies a certain differential equation.

Definition 4.4.1. *Let L be a lattice. We define the **Eisenstein series** of weight $2k$ for the lattice L as*

$$G_{2k}(L) = \sum_{\substack{\omega \in L \\ \omega \neq 0}} \omega^{-2k}. \tag{4.4.1}$$

It turns out that for every lattice L, the Eisenstein series $G_{2k}(L)$ is absolutely convergent for every $k > 1$. For a proof of this statement, see Theorem VI.3.1(a) in Silverman [3].

Theorem 4.4.2. *In a neighborhood of the point $z = 0$, we can write the Weierstrass $\wp$-function as*

$$\wp(z) = \frac{1}{z^2} + \sum_{k=1}^{\infty} (2k+1) G_{2k+2} z^{2k}. \tag{4.4.2}$$

This expression is the so-called Laurent expansion of $\wp(z)$ at the point $z = 0$.

Proof. For the proof of this theorem, see Theorem VI.3.5(a) in Silverman [3]. □

Following standard notation, we define for a given lattice L the constants $g_2(L)$ and $g_3(L)$ as

$$g_2 = g_2(L) = 60G_4(L) \qquad \text{and} \qquad g_3 = g_3(L) = 140G_6(L). \tag{4.4.3}$$

Now, the following theorem is an important result for the Weierstrass $\wp$-function, and is the key to understanding the link between elliptic functions and elliptic curves.

Theorem 4.4.3. *The Weierstrass $\wp$-function satisfies the differential equation*

$$\wp'(z)^2 = 4\wp(z)^3 - g_2\wp(z) - g_3. \tag{4.4.4}$$

Proof. From Theorem 4.4.2 we know that in a neighborhood of the point $z = 0$ we can write

$$\wp(z) = \frac{1}{z^2} + \sum_{k=1}^{\infty}(2k+1)G_{2k+2}z^{2k}. \tag{4.4.5}$$

We have that the function $\wp$ is analytic on $\mathbb{C} \setminus L$, and so is the right-hand side of the above equation. Also, a neighborhood of a point is not discrete. Therefore, it follows from Theorem 4.1.7 that $\wp(z)$ is equal to its Laurent expansion at the point 0 for every point $z \in \mathbb{C} \setminus L$. So for every point $z \in C \setminus L$ we have

$$\wp(z) = z^{-2} + 3G_4z^2 + 5G_6z^4 + \cdots \quad . \tag{4.4.6}$$

Here, and in the following equations, the dots at the end stand for higher order terms. We can differentiate the above expression to find

$$\wp'(z) = -2z^{-3} + 6G_4z + 20G_6z^3 + \cdots \quad . \tag{4.4.7}$$

Then taking the square gives

$$\wp'(z)^2 = 4z^{-6} - 24G_4z^{-2} - 80G_6 + \cdots \quad . \tag{4.4.8}$$

Similarly, we compute

$$\wp(z)^2 = z^{-4} + 6G_4 + 10G_6z^2 + \cdots \quad , \tag{4.4.9}$$

$$\wp(z)^3 = z^{-6} + 9G_4z^{-2} + 15G_6 + \cdots \quad . \tag{4.4.10}$$

Then we see that

$$\wp'(z)^2 - 4\wp(z)^3 = -60G_4z^{-2} - 140G_6, \tag{4.4.11}$$

and therefore

$$\wp'(z)^2 - 4\wp(z)^3 + 60G_4\wp(z) = -140G_6 + \cdots \quad . \tag{4.4.12}$$

Now on the left-hand side we have an elliptic function, but on the right-hand side we have a power series in z without any terms with negative power. So this elliptic function has no poles. Then by Theorem 4.2.3 the function must be constant. Therefore the higher order terms in the dots on the right-hand side are all zero. Then, using our definitions of g_2 and g_3, we obtain indeed

$$\wp'(z)^2 = 4\wp(z)^3 - g_2\wp(z) - g_3. \tag{4.4.13}$$

□

Now, what does this theorem have to do with elliptic curves? Remember that if we have an elliptic curve in Weierstrass normal form and we dehomogenized with respect to Z, then the affine part of the curve is given by an equation of the form

$$y^2 = f(x) = x^3 + ax^2 + bx + c, \tag{4.4.14}$$

where the discriminant of $f(x)$ is non-zero, i.e. f is smooth. Further, for elliptic curves in $\mathbb{C}$ we can even transform every elliptic curve into an elliptic curve of the form

$$y^2 = f(x) = 4x^3 - bx - c, \tag{4.4.15}$$

where $f(x)$ is again smooth. This form is called Weierstrass normal form as well. If we compare this to Equation (4.4.4), it is clear where this name comes from. On the right-hand side of Equation (4.4.4) we have a polynomial in $\wp(z)$ of degree three. However, we do not know yet whether this polynomial is smooth. It turns out that this is the case.

Theorem 4.4.4. *The discriminant of the polynomial*

$$f(x) = 4x^3 - g_2x - g_3 \tag{4.4.16}$$

is non-zero, i.e. $f(x)$ *is smooth.*

Proof. For the proof, see proposition VI.3.6(a) of Silverman [3]. □

Corollary 4.4.5. *The curve*

$$C' : y^2 = 4x^3 - g_2x - g_3 \tag{4.4.17}$$

is the affine part of an elliptic curve $E \subset \mathbb{P}^2(\mathbb{C})$ *in Weierstrass normal form when dehomogenized with respect to* Z*. Its point at infinity is then* $[0, 1, 0]$*.*

In the above corollary we call the elliptic curve E instead of C, because we will look at this specific elliptic curve in the rest of this section.

Now, from Theorem 4.4.3 it follows that for every $z \in \mathbb{C} \setminus L$ we have that $[\wp(z), \wp'(z), 1] \in E$. Then there is one point on E that we certainly do not reach, namely the point $\mathcal{O} = [0, 1, 0]$. As we saw before, an elliptic function takes all its values in its fundamental parallelogram. The Weierstras $\wp$-function has one point in its fundamental parallelogram $\mathcal{F}$ where it is not defined, namely the point $z = 0$. We define the function $\phi : \mathcal{F} \to E$ as

$$\phi(z) = \begin{cases} [\wp(z), \wp'(z), 1] & \text{if } z \neq 0 \\ [0, 1, 0] & \text{if } z = 0 \end{cases}. \tag{4.4.18}$$

As we noted before, the fundamental parallelogram $\mathcal{F}$ is isomorphic to the space $\mathbb{C}/L$ (which was a torus). Therefore, we can also view ϕ as a function from $\mathbb{C}/L$ to E. The complex numbers $\mathbb{C}$ form a group under addition, and taking the quotient with L is a group homomorphism with respect to this addition. We can see this by viewing $\mathbb{C}/L$ as $\mathcal{F}$. Then if we take two numbers $z_1, z_2 \in \mathcal{F}$ we can just add them in $\mathbb{C}$, and then we say $[z_1] + [z_2] = [z_1 + z_2]$. So $\mathbb{C}/L$ has a group structure. In Chapter 2 we defined a group structure on the set of points of an elliptic curve in Weierstrass normal form. So the function ϕ is a map between groups.

Theorem 4.4.6. *The map* $\phi : \mathcal{F} \to E$ *defined in Equation* (4.4.18) *is a group isomorphism.*

Proof. For the proof of this theorem, see Theorem VI.3.6(b) of Silverman [3]. □

The above theorem implies that the affine part of the elliptic curve E can be parameterised by taking $x = \wp(z)$ and $y = \wp'(z)$. Up to this moment, we have only proven that this works for the specific elliptic curve E from Corollary 4.4.5. This curve depends on the choice of the lattice L, but it is not immediately clear that by varying L we can parameterise every elliptic curve in Weierstrass normal form using the Weierstrass elliptic function. It turns out that in fact we can, which is also the reason that elliptic curves are called elliptic curves.

Theorem 4.4.7. *Let $C \subset \mathbb{P}^2(\mathbb{C})$ be an elliptic curve in Weierstrass normal form. Then there exists a lattice $L \subset \mathbb{C}$ such that the function $\phi : \mathcal{F} \to C$ defined in Equation* (4.4.18) *is a group isomorphism.*

For the proof of this theorem, remember that for elliptic curves in $\mathbb{P}^2(\mathbb{C})$ we have the more strict definition of Weierstrass normal form (2.1.3) with $a = 0$. Let $C \subset \mathbb{P}^2(\mathbb{C})$ be an elliptic curve in Weierstrass normal form. Then, if we dehomogenize with respect to Z, the affine part C' of C is of the form

$$C' : y^2 = f(x) = x^3 + bx + c, \tag{4.4.19}$$

where the polynomial $f(x)$ is smooth. Then we can parameterise C using the map ϕ if we can find a lattice $L \subset \mathbb{C}$ such that $b = -g_2(L)$ and $c = -g_3(L)$. Therefore, in order to prove Theorem 4.4.7, is enough to prove the following proposition.

Proposition 4.4.8. *Let $b, c \in \mathbb{C}$ be two complex numbers such that the polynomial*

$$f(x) = 4x^3 - bx - c \tag{4.4.20}$$

is smooth. Then there exists a unique lattice $L \subset \mathbb{C}$ such that

$$g_2(L) = b \quad \text{and} \quad g_3(L) = c. \tag{4.4.21}$$

Proof. This is Theorem VI.5.1 in Silverman [3]. For the proof, he refers to for example VII Proposition 5 in Serre [12] □

4.5 Elliptic Integrals and Jacobi Elliptic Functions

In this section we will look at elliptic integrals, which come up when calculating the arc length of an ellipse. These elliptic integrals are closely related to elliptic functions, which is also the reason that elliptic functions are called elliptic. Further, using elliptic integrals we will introduce some special examples of elliptic functions, namely the Jacobi elliptic functions. In the last chapter we will use these functions to write down a solution of the Yang-Baxter equation. Lastly, I will give a short summary as to why elliptic curves are called elliptic. The Subsection 4.5.1 is based on Exercise 1.16 in Silverman and Tate [1] and the section "From ellipses to elliptic integrals" in Rice and Brown [13]. In Subsection 4.5.2 we follow Section 2.8 of Prasolov and Solovyev [14].

4.5.1 The Arc Length of an Ellipse

Suppose that we have an ellipse centered at the origin. Then this ellipse is a curve C given by

$$C : \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \tag{4.5.1}$$

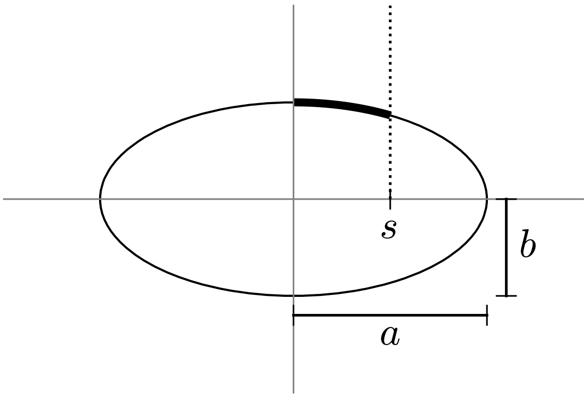

Figure 4.2: A picture of an ellipse centered at the origin. The arc whose length we compute in Equation (4.5.5) is indicated with a thicker line.

for some constants $0 < a, b$. Then the width of the ellipse is $2a$ and its height is $2b$. We assume $b \leq a$. See Figure 4.2 for a picture of the ellipse. In the positive quadrant of the plane, we have that the curve C is given by

$$y = \frac{b}{a}\sqrt{a^2 - x^2}. \tag{4.5.2}$$

Then y is a function of x, so we can calculate its derivative

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{b}{a}\frac{x}{\sqrt{a^2 - x^2}}. \tag{4.5.3}$$

Recall that we can calculate the arc length of a function f between x_1 and x_2 with the integral

$$\int_{x_1}^{x_2} \sqrt{1 + \left(\frac{\mathrm{d}f}{\mathrm{d}x}\right)^2}\,\mathrm{d}x. \tag{4.5.4}$$

Therefore, we can calculate the arc length of the ellipse in the positive quadrant between $x = 0$ and $x = s < a$ as

$$\int_0^s \sqrt{1 + \left(-\frac{b}{a}\frac{x}{\sqrt{a^2 - x^2}}\right)^2}\,\mathrm{d}x. \tag{4.5.5}$$

See also Figure 4.2. We can simplify this integral a bit by substituting $x = a\sin\theta$. Then we find $\mathrm{d}\theta = 1/(a\cos\theta)\ \mathrm{d}x$, and $a^2 - x^2 = a^2\cos^2\theta$. Using this we get that the above integral becomes

$$\int_0^{\arcsin s/a} a\sqrt{1 - \left(1 - \frac{b^2}{a^2}\right)\sin^2\theta}\ \mathrm{d}\theta. \tag{4.5.6}$$

Here the function $\arcsin : [-1, 1] \to [-\pi/2, \pi/2]$ is the inverse function of the restriction of the sine function $\sin : [-\pi/2, \pi/2] \to [-1, 1]$. Note that $s < a$, so we can take the inverse sine of s/a. For s ranging between 0 and a, the value of $\arcsin s/a$ will range from 0 to $\pi/2$. Now, we assumed $0 < b \leq a$, so $1 - b^2/a^2 \geq 0$. Therefore we can define k as $k = \sqrt{1 - a^2/b^2}$. We also write $t = \arcsin s/a$. Then we get the integral

$$a\int_0^t \sqrt{1 - k^2\sin^2\theta}\ \mathrm{d}\theta, \tag{4.5.7}$$

which is a function of t. The arc length L of the total ellipse is four times its arc length in the positive quadrant, so we find

$$L = 4a \int_0^{\pi/2} \sqrt{1 - k^2 \sin^2\theta}\ \mathrm{d}\theta. \tag{4.5.8}$$

The integral in Equation (4.5.7) is an example of a so-called *elliptic integral*, named for the fact that it appears when calculating the arc length of an ellipse. Even though this integral looks fairly simple, it turns out that it cannot be expressed in elementary functions. Here with elementary functions we mean rational functions, trigonometric functions, exponentials and logarithms. Therefore we have to see the elliptic integrals as "new" functions.

4.5.2 The Jacobi Elliptic Functions

Another example of an elliptic integral would be

$$F(s) = \int_0^s \frac{1}{\sqrt{(1-x^2)(1-kx^2)}}\ \mathrm{d}x = \int_0^t \frac{1}{\sqrt{1-k^2\sin^2\theta}}\ \mathrm{d}\theta, \tag{4.5.9}$$

where $x = \sin\theta$, and $t = \arcsin s$. If we set $k = 0$, which in the previous subsection would correspond to the case where the ellipse is a circle, we obtain the integral

$$F(s) = \int_0^s \frac{1}{\sqrt{1-x^2}}\ \mathrm{d}x = \int_0^t \mathrm{d}\theta = t. \tag{4.5.10}$$

So in that case we find $F(s) = \arcsin s$. It is often easier to work with sin than with arcsin. This motivates us to consider the inverse function of $F(s)$ for arbitrary k. However, instead of $F(s)$, we will from now on take the integral on the right hand side of Equation (4.5.9) as the definition of $F(t)$, where $t = \arcsin s$, and look at the inverse of $F(t)$. Note that we expect this inverse function to exist, because for $0 \leq k < 1$ the integral on the right-hand side of Equation (4.5.9) is a strictly increasing continuous function of t defined on the interval $[0, \pi/2]$, and therefore a bijection onto its image. We will write $u(t) = F(t)$, because that is the notation Jacobi used, and it is still used in most of the literature [14]. Then we call the inverse function $t(u)$ of $u(t)$ the *amplitude* of u, denoted by $t(u) = \operatorname{am} u$. Note that if $k = 0$, then $u(t) = t$ so $\operatorname{am} u = u$.

Definition 4.5.1. *For $u = F$ as given in Equation* (4.5.9)*, we define the **Jacobi elliptic functions** $\operatorname{sn} u$, $\operatorname{cn} u$ and $\operatorname{dn} u$ as*

$$\operatorname{sn} u = \sin(\operatorname{am} u), \qquad \operatorname{cn} u = \cos(\operatorname{am} u), \qquad \operatorname{dn} u = \sqrt{1 - k^2 \operatorname{sn}^2 u}. \tag{4.5.11}$$

These functions depend on the number k in Equation (4.5.9)*, which is called the **modulus** [15]. The functions* sn *and* cn *are also called the **Jacobi sine function** and **Jacobi cosine function**, respectively.*

Note that because in Equation (4.5.9) we have $t = \arcsin s$, it follows from the above definition that $\operatorname{sn} u = s$. So the Jacobi sine function is the inverse function of the elliptic integral on the left-hand side of Equation (4.5.9).

Theorem 4.5.2. *The function* sn *is an even elliptic function, and the functions* cn *and* dn *are odd elliptic functions.*

Proof. For the proof, see Section 2.8 of Prasolov and Solovyev [14]. □

Rember that if the modulus k is zero, then we have $\operatorname{am} u = u$. Therefore, for $k = 0$, we find

$$\operatorname{sn} u = \sin u, \qquad \operatorname{cn} u = \cos u, \qquad \operatorname{dn} u = 0. \tag{4.5.12}$$

So the Jacobi elliptic functions are in fact doubly periodic generalizations of the usual sine and cosine functions. Their similarity is also obvious from the following results we have for the Jacobi elliptic functions.

Theorem 4.5.3. *The following identities hold for any value of the modulus* $0 \leq k < 1$*:*

$$\operatorname{sn}^2 + \operatorname{cn}^2 = 1, \qquad \operatorname{dn}^2 + k^2 \operatorname{sn}^2 = 1. \tag{4.5.13}$$

Proof. These identities are a direct result of Definition 4.5.1 and the fact that $\sin^2 + \cos^2 = 1$. □

Also in the point $u = 0$ the Jacobi elliptic functions behave as you would expect.

Theorem 4.5.4. *The following identities hold*

$$\operatorname{sn} 0 = 0, \qquad \operatorname{cn} 0 = \operatorname{dn} 0 = 1. \tag{4.5.14}$$

Proof. From the way we defined the amplitude $\operatorname{am} u$, it follows that $\operatorname{am} 0 = 0$. Hence, from Definition 4.5.1 we see that indeed $\operatorname{sn} 0 = 0$ and $\operatorname{cn} 0 = 1$, which also implies $\operatorname{dn} 0 = 1$. □

Then there are two more results that we need in the last chapter

Theorem 4.5.5. *We have the following results for the derivatives of the Jacobi elliptic functions*

$$\begin{aligned} \frac{d}{du} \operatorname{sn} u &= \operatorname{cn} u \operatorname{dn} u, \\ \frac{d}{du} \operatorname{cn} u &= -\operatorname{sn} u \operatorname{dn} u, \\ \frac{d}{du} \operatorname{dn} u &= -k^2 \operatorname{sn} u \operatorname{cn} u. \end{aligned} \tag{4.5.15}$$

Proof. For the proof, see page 493 of Whittaker and Watson [15]. □

Theorem 4.5.6 (Addition formulas)**.** *We have the following addition formulas for the Jacobi elliptic functions*

$$\begin{aligned} \operatorname{sn}(u+v) &= \frac{\operatorname{sn} u \operatorname{cn} v \operatorname{dn} v + \operatorname{sn} v \operatorname{cn} u \operatorname{dn} u}{1 - k^2 \operatorname{sn}^2 u \operatorname{sn}^2 v}, \\ \operatorname{cn}(u+v) &= \frac{\operatorname{cn} u \operatorname{cn} v - \operatorname{sn} u \operatorname{sn} v \operatorname{dn} u \operatorname{dn} v}{1 - k^2 \operatorname{sn}^2 u \operatorname{sn}^2 v}, \\ \operatorname{dn}(u+v) &= \frac{\operatorname{dn} u \operatorname{dn} v - k^2 \operatorname{sn} u \operatorname{sn} v \operatorname{cn} u \operatorname{cn} v}{1 - k^2 \operatorname{sn}^2 u \operatorname{sn}^2 v}. \end{aligned} \tag{4.5.16}$$

Proof. For the proof of this theorem, see Section 2.8 of Prasolov and Solovyev [14]. □

4.5.3 Why Elliptic Curves Are Called Elliptic

In this subsection I give a short explanation why elliptic curves are called elliptic, following Rice and Brown [13]. See their article for more information on the history of elliptic curves.

Up to this moment, we have seen three objects that are called ‘elliptic’: elliptic curves, elliptic functions and elliptic integrals. We introduced them in this order, but this order might be confusing when trying to understand where their names come from. From a historical perspective, the order should be: elliptic integrals, elliptic functions, elliptic curves [13].

Elliptic integrals are called elliptic because they are encountered when trying to calculate the arc length of an ellipse, as we saw in Subsection 4.5.1. However, it turned out that it was more convenient to study the inverse functions of elliptic integrals [13]. Because of their relation with elliptic integrals, these inverse functions were called elliptic functions. Lastly, it turned out that these elliptic functions could be used to parameterise certain cubic curves, and therefore these curves are named elliptic curves.

Chapter 5

The Yang-Baxter Equation

In this chapter and the next, we will look at an application of elliptic functions in physics. In principle, the reader should be able to understand these chapters without having read the rest of this thesis, except for Subsection 4.5.2 on the Jacobi elliptic functions. In this chapter we will introduce the Yang-Baxter equation, and in the next chapter we will derive a solution for this equation using the Jacobi elliptic functions we saw in the previous chapter. The Yang-Baxter equation comes up when studying the scattering of particles in a one dimensional integrable system, so that is where we will start.

5.1 General Formalism

In this section we will set up a scattering problem from which we will derive the Yang-Baxter equation. We follow Sections 1.1 and 5.1 of Šamaj and Bajnok [16] and Sections 5.1.2 and 5.1.3 from Giamarchi [17].

Suppose that we have a one-dimensional system of N identical particles with mass m and no internal degrees of freedom. Suppose that the particles interact pairwisely by a symmetric potential $v(x_i, x_j)$ that only depends on the distance between the particles and approaches zero for large distances. We set $\hbar = 2m = 1$. Then we get the following Hamiltonian:

$$H = -\sum_{j=1}^{N} \frac{\partial^2}{\partial x_j^2} + \sum_{j<k=1}^{N} v(x_j, x_k). \tag{5.1.1}$$

The time-independent Schrödinger equation then reads

$$H\psi(x_1, \ldots, x_N) = E\psi(x_1, \ldots, x_N). \tag{5.1.2}$$

In the case $N = 1$ we get the plane wave solution

$$\psi(x) = A \exp{(ikx)}, \tag{5.1.3}$$

where k is the wavenumber and A a normalization constant. If we calculate the expectation value of the momentum for this wave function, we find that the momentum is equal to k. Therefore we also call the wavenumber k the momentum of a particle. We set $\hbar = 2m = 1$, so the corresponding energy is $E = k^2$.

Now if $N > 1$ we have particles that can interact with each other. The potential v is such that the interaction only takes place at short distances. We call this interaction of particles *scattering*. Also, if we have multiple particles in one dimension then they must have a certain order. We will make this explicit by introducing ordering sectors.

5.1.1 Ordering Sectors

Say we have N particles. Our system is only one-dimensional, so there is a well-defined ordering on our particles. A possible way to order the particles we call an *ordering sector*. Let S_N be the symmetric group of degree N which consists of permutations of $(1, 2, \ldots, N)$. Then there is a one-to-one relation between the ordering sectors and elements of S_N. For $q \in S_N$ we get the ordering sector $Q = (Q1, \ldots, QN) = q(1, \ldots, N)$. Then in the sector Q we have $x_{Q1} < \ldots < x_{QN}$. The *fundamental ordering sector* I is the one identified with the identity element of S_N, and therefore gives us the ordering $x_1 < \ldots < x_N$. For a wavefunction in ordering sector Q we write $\psi_Q(x_1, \ldots, x_N)$. The particles in our system are identical so the wavefunction should be either symmetric or antisymmetric for bosons or fermions respectively. Therefore, for bosons we get

$$\psi_Q(x_1, \ldots, x_N) = \psi_I(x_{Q1}, \ldots, x_{QN}) \tag{5.1.4}$$

and for fermions we find

$$\psi_Q(x_1, \ldots, x_N) = \text{sign}(Q)\psi_I(x_{Q1}, \ldots, x_{QN}), \tag{5.1.5}$$

where $\text{sign}(Q)$ equals 1 if Q consists of an even number of transpositions, and -1 if Q consists of an odd number of transpositions.

5.1.2 Scattering and the Bethe-Ansatz

We want to find the wavefunction for a system with multiple particles. Then we have to understand what actually happens if particles scatter off of each other. Basically, the scattering of two particles can be described as follows:

We have two particles moving towards each other. At first, the distance between the particles is large and therefore they do not interact. At close distances however, the particles interact (for example via a delta potential). Then afterwards they move away from each other. During the interaction it is possible for the particles to change momenta and other internal degrees of freedom.

So Let us make this more precise in our case. We have two particles with wavenumbers k_1 and k_2. At $t = t_0$ the particles are far away from each other but move towards each other. When they get close enough they interact in some way and afterwards they move away from each other. At time $t = t_1$ the particles are far away from each other and move in opposite directions with wavenumbers k_1' and k_2'.

We assume the scattering to be elastic, so between the times t_0 and t_1 the total momentum and total energy should be conserved. Therefore we have $k_1 + k_2 = k_1' + k_2'$ and $k_1^2 + k_2^2 = k_1'^2 + k_2'^2$. Together this leaves only two solutions: $(k_1, k_2) = (k_1', k_2')$ or $(k_1, k_2) = (k_2', k_1')$. Therefore, for large distances between the particles, the wavefunction of the particles (in the fundamental ordering sector) has to be of the form [17]

$$\psi_I(x_1, x_2) = Ae^{i(k_1x_1+k_2x_2)} + Be^{i(k_1x_2+k_2x_1)} \tag{5.1.6}$$

The *Bethe ansatz* for two particles is the assumption that this wavefunction holds for any distance between the particles.

Now, for an arbitrary amount of particles N we can do something similar. We have N initial momenta $k_1, \ldots, k_N$ and N post-interaction momenta $k_1', \ldots, k_N'$. Again the total momentum and energy should be conserved. However, in the case of three or more particles the conservation of momentum and energy is in general not enough to state that $(k_1, \ldots, k_N) = q((k_1', \ldots, k_N'))$ for some $q \in S_N$. We could have a so-called *diffractive scattering*, where this is not the case. Therefore we look at *integrable* *models*. In integrable models we have the following conservation laws [16]:

$$\sum_{j=1}^{N} (k_j)^m = \sum_{j=1}^{N} \left(k_j'\right)^m, \quad m \in \mathbb{N}. \tag{5.1.7}$$

The cases $m = 1$ and $m = 2$ correspond to momentum and energy conservation respectively. Note that integrable models have infinitely many conservation laws. Together, these constraints imply that the scattering of multiple particles is never diffractive in integrable models. So we get $(k_1, \ldots, k_N) = q((k_1', \ldots, k_N'))$ for some $q \in S_N$. Then, just as for $N = 2$, we get that the wavefunction for large distances between the particles must be of the form [17]

$$\psi_I(x_1, \ldots, x_N) = \sum_{P \in S_N} A(k_{P_1}, \ldots, k_{P_N}) \exp\left(i \sum_{j=1}^{N} k_{P_j} x_j\right). \tag{5.1.8}$$

Now the Bethe-ansatz is again the assumption that the above wavefunction holds for all positions of the particles. We will use it.

Of course it could happen that particles do have internal degrees of freedom $\sigma = 1, \ldots, l$, which we call *colors*. For an electron we have two colors ($l = 2$): spin up and spin down. In the case of colored particles each particle is described by its position and color. We can implement this in our wavefunction to find the general wavefunction

$$\psi_I(\sigma_1, x_1, \ldots, \sigma_N, x_N) = \sum_{P \in S_N} A_{\sigma_1, \ldots, \sigma_N}(k_{P_1}, \ldots, k_{P_N}) \exp\left(i \sum_{j=1}^{N} k_{P_j} x_j\right). \tag{5.1.9}$$

If we want the wavefunction in another ordering sector Q we can just use equations (4) and (5). Then for bosons we find

$$\psi_Q(\sigma_1, x_1, \ldots, \sigma_N, x_N) = \sum_{P \in S_N} A_{\sigma_{Q_1}, \ldots, \sigma_{Q_N}}(k_{P_1}, \ldots, k_{P_N}) \exp\left(i \sum_{j=1}^{N} k_{P_j} x_{Q_j}\right), \tag{5.1.10}$$

and for fermions we get

$$\psi_Q(\sigma_1, x_1, \ldots, \sigma_N, x_N) = \sum_{P \in S_N} \operatorname{sign}(Q) A_{\sigma_{Q_1}, \ldots, \sigma_{Q_N}}(k_{P_1}, \ldots, k_{P_N}) \exp\left(i \sum_{j=1}^{N} k_{P_j} x_{Q_j}\right). \tag{5.1.11}$$

5.2 The Scattering Matrix

In this section we will introduce the scattering matrix that describes the scattering of two particles. However, before we do that, in the next subsection we will introduce a new notation for the elements of a matrix in a product space. There we will also introduce the Einstein summation convention, which we will use in the rest of this thesis.

5.2.1 Matrix Notation and the Einstein Summation Convention

Suppose that we have a square matrix A of dimension n (almost every matrix in this thesis will be square). Then, as you have probably seen before, we can denote its set of elements by $\{a_{ij}\}_{1\leq i,j\leq n}$. Here the i index says in which row the element is, and the j index represents the column. For both i and j we have n options, so there are n^2 elements in total. Now, in this thesis, we will use a slightly different notation. Instead of a_{ij}, we write a^i_j. So the row is represented by the upper index, and the column by the lower index. Let B be another square matrix of dimension n with elements b^i_j, and $AB = C$ with elements c^i_j. Then we can relate the elements of C to those of A and B by

$$c^i_j = \sum_{k=1}^{n} a^i_k b^k_j. \tag{5.2.1}$$

Now this is not yet very different from the notation we had. The advantage of this new notation becomes clear if we consider operators that act on the product space of multiple vector spaces.

Suppose we have two vector spaces, V_a of dimension n_a and V_b of dimension n_b, and a linear operator T acting on $W = V_a \otimes V_b$. If we choose bases $\{e_{a_i}\}_{1\leq i\leq n_a}$ and $\{e_{b_j}\}_{1\leq j\leq n_b}$ for V_a and V_b respectively, we get the basis $\{e_{a_i} \otimes e_{b_j}\}$ where $1 \leq i \leq n_a$ and $1 \leq j \leq n_b$ for W. The product space W has finite dimension $m = n_a n_b$, so we can represent T by an m-dimensional square matrix. In principle we could denote the set of elements of T by $\{T^i_j\}_{1\leq i,j\leq m}$, where we label each row or column by a single index between 1 and m. However, each row of T corresponds to a basisvector of W, and the same for each column. If we look at our basis for W, we see that we have precisely one basisvector for every ordered pair (i, j) with $1 \leq i \leq n_a$ and $1 \leq j \leq n_b$. Therefore it makes more sense two use two indices to indicate the row of T, and also two two to indicate the column. So we denote the elements of T by $T^{ij}_{i'j'}$, where $1 \leq i, i' \leq n_a$ and $1 \leq j, j' \leq n_b$.

It should be clear how we can generalize this notation to operators acting on the product space of more than two vector spaces. For example if the linear operator T acts on the space $V_1 \otimes \cdots \otimes V_k$, then we denote the elements of its matrix by

$$T^{i_1\ldots i_k}_{i'_1\ldots i'_k}. \tag{5.2.2}$$

This looks much more compact than if we would indicate both the row and column by lower indices. Further, we can immediately see which row and column indices correspond to the same vector space, as they are positioned directly above each other.

In the rest of this thesis, we will encounter many equations where we take the product of operators. As we saw in Equation (5.2.1), we can write down matrix multiplication element-wise with a summation. However, if we multiply multiple matrices, the notation can become quite tedious. Therefore we will use the *Einstein summation convention*, or just summation convention. Essentially, this means that we will not write down the summation signs anymore. The idea is that it should be clear from the indices whether we take the sum over them. The rule of the summation convention is: if we multiply two terms with the same index, then we sum over this index.

For example, suppose that we take the matrix T from Equation (5.2.2) and multiply it with itself. Then with summation signs we can write this product element-wise as

$$[TT]^{i_1\ldots i_k}_{i'_1\ldots i'_k} = \sum_{i''_1\ldots i''_k} T^{i_1\ldots i_k}_{i''_1\ldots i''_k} T^{i''_1\ldots i''_k}_{i'_1\ldots i'_k}. \tag{5.2.3}$$

We see that on the right-hand side inside the sum we have the product of two terms that share the indices $i''_1, \ldots, i''_k$. These are exactly the indices that we sum over. Therefore we see that if we use the summation convention, the above equation is

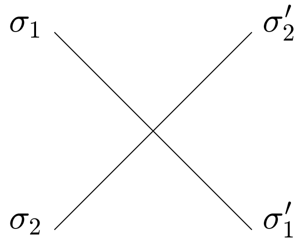

Figure 5.1: The scattering of two particles.

equivalent to the one we get if we just remove the summation sign:

$$[TT]^{i_1\ldots i_k}_{i'_1\ldots i'_k} = T^{i_1\ldots i_k}_{i''_1\ldots i''_k} T^{i''_1\ldots i''_k}_{i'_1\ldots i'_k}. \tag{5.2.4}$$

In the next sections, there will be many equations where we sum over indices. It can be hard to get used to the summation convention, so I will sometimes (but not always!) mention where we use it.

5.2.2 The Scattering Matrix

Now we are ready to introduce the scattering matrix. We follow the second half of Section 5.1 of Šamaj and Bajnok [16].

In both the boson and fermion cases (5.1.10) and (5.1.11), we can describe the relation between the A-amplitudes for different permutations (P, Q) via the so-called *scattering matrix*. For $N = 2$ particles, we define the scattering matrix S element-wise by

$$A_{\sigma_j\sigma_i}(k_v, k_u) = S^{\sigma_i\sigma_j}_{\sigma'_i\sigma'_j}(k_u, k_v) A_{\sigma'_i\sigma'_j}(k_u, k_v), \tag{5.2.5}$$

where $(i, j), (u, v) \in \{(12), (21)\}$. Note that we use the summation convention, so we sum over the indices σ'_i and σ'_j. We assume the scattering of two particles to be elastic, so there is no energy loss in the scattering process. As we have seen before this implies that the individual momenta of the particles are conserved. However, the particles can change their color σ_i in the scattering process. We can think of our one dimensional system with N particles as a chain of N sites. Then every site is occupied by exactly one particle, and the sites correspond to positions in the ordering sector. This way we discretise our system. Then the scattering of two particles should be understood as two adjacent particles that switch sites, possibly changing their colors in the process. From now on we will refer to our system as the chain of N sites. See Figure 5.1 for a visualization of the scattering of two particles, which is described by the scattering matrix.

The scattering matrix S given in Equation (5.2.5) acts on the product space of two vector spaces $V_i \otimes V_j$. Here V_i and V_j are l-dimensional vectorspaces, each corresponding to one of the two particles involved in the scattering described by S. Each dimension in the vectorspace V_i corresponds to a color its corresponding particle may have, i.e. a value of σ_i. Similarly, each dimension in V_j corresponds to a value of σ_j. Therefore we use the two indices σ_i and σ_j to indicate the row of S, and two other indices σ'_i and σ'_j to indicate the column. Note that the scattering matrix S for two particles has dimensions l^2, because both σ_1 and σ_2 can take l values.

We define the *permutation operator* $\mathscr{P}$ of dimension l^2 as

$$\mathscr{P}^{\sigma_1\sigma_2}_{\sigma'_1\sigma'_2} = \delta^{\sigma_1}_{\sigma'_2}\delta^{\sigma_2}_{\sigma'_1}. \tag{5.2.6}$$

Here we write δ^i_j for the Kronecker delta. So we have

$$\delta^i_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} . \tag{5.2.7}$$

This notation can be motivated by the fact that we can view the Kronecker delta as elements of the identity matrix. We have $\delta^i_j = I^i_j$, where I is in this case the identity matrix of dimension l. This is not to be confused with the fundamental ordering sector I. To see why $\mathscr{P}$ is called the permutation operator, we look at what happens if we multiply it with another matrix, say the scattering matrix S. We get

$$[\mathscr{P}S]^{\sigma_1\sigma_2}_{\sigma_1'\sigma_2'} = \delta^{\sigma_1}_{\sigma_2''}\delta^{\sigma_2}_{\sigma_1''}S^{\sigma_1''\sigma_2''}_{\sigma_1'\sigma_2'} = S^{\sigma_2\sigma_1}_{\sigma_1'\sigma_2'}. \tag{5.2.8}$$

Note that here we use the summation convention. Similarly we find $[S\mathscr{P}]^{\sigma_1\sigma_2}_{\sigma_1'\sigma_2} = S^{\sigma_1\sigma_2}_{\sigma_2'\sigma_1'}$. So the permutation operator permutes the indices of S, by switching two of them. Clearly, we have $\mathscr{P}^2 = I$.

Looking at Equation (5.2.5), we can derive the following properties of the S matrix:

- Normalization: If we let $k_u = k_v = k$, then we have

$$S(k, k) = \mathscr{P}. \tag{5.2.9}$$

- Unitarity: we have

$$S(k_u, k_v)S(k_v, k_u) = I. \tag{5.2.10}$$

  This can be seen by applying Equation (5.2.5) to itself. Then we get

$$\begin{aligned} A_{\sigma_j\sigma_i}(k_v, k_u) &= S^{\sigma_i\sigma_j}_{\sigma_i'\sigma_j'}(k_u, k_v)S^{\sigma_i'\sigma_j'}_{\sigma_i''\sigma_j''}(k_v, k_u)A_{\sigma_j''\sigma_i''}(k_v, k_u), \\ &= [S(k_u, k_v)S(k_v, k_u)]^{\sigma_i\sigma_j}_{\sigma_i''\sigma_j''}A_{\sigma_j''\sigma_i''}(k_v, k_u). \end{aligned} \tag{5.2.11}$$

  This implies that

$$[S(k_u, k_v)S(k_v, k_u)]^{\sigma_i\sigma_j}_{\sigma_i''\sigma_j''} = \delta^{\sigma_i}_{\sigma_i''}\delta^{\sigma_j}_{\sigma_j''} = \delta^{\sigma_i\sigma_j}_{\sigma_i''\sigma_j''}. \tag{5.2.12}$$

  Therefore we see that indeed $S(k_u, k_v)S(k_v, k_u) = I$.

In addition to these properties, we also make the following assumption:

- Symmetries: We assume

$$S^{\sigma_i\sigma_j}_{\sigma_i'\sigma_j'}(k_u, k_v) = S^{\sigma_i'\sigma_j'}_{\sigma_i\sigma_j}(k_u, k_v) = S^{\sigma_j\sigma_i}_{\sigma_j'\sigma_i'}(k_u, k_v). \tag{5.2.13}$$

Now, we claimed before that we can describe the relation between the A-amplitudes for different permutations (P, Q) via the scattering matrix. For the general case of N scattering particles, this comes from the fact that in integral models we can decompose the scattering into two-particle scatterings. This is the fundamental property of integrable systems [16]. Indeed, suppose we have two permutation sectors (P, Q) and $(\widetilde{P}, \widetilde{Q})$, that differ from each other only by a transposition of two nearest neighbours. So for some index $1 \leq j < N$ we have $Q_i = \widetilde{Q}_i$ for $i \neq j, j+1$, and $Q_j = \widetilde{Q}_{j+1}$, $Q_{j+1} = \widetilde{Q}_j$. Similarly, for some index $1 \leq u < N$ we have the same for P and $\widetilde{P}$. Then the scattering process that corresponds to going from permutation sector (P, Q) to $(\widetilde{P}, \widetilde{Q})$ is just a two-particle scattering process. Therefore we can use the two-particle scattering matrix to relate the corresponding A-amplitudes. Writing $j+1=k$ and $u+1=v$, we get:

$$A_{\dots\sigma_{Q_k}\sigma_{Q_j}\dots}(\dots,k_{P_v},k_{P_u},\dots) = S^{\sigma_{Q_j}\sigma_{Q_k}}_{\sigma'_{Q_j}\sigma'_{Q_k}}(k_{P_u},k_{P_v})A_{\dots\sigma'_{Q_j}\sigma'_{Q_k}\dots}(\dots,k_{P_u},k_{P_v},\dots). \tag{5.2.14}$$

Now, note that every permutation $Q \in S_N$ can be written as a sequence of next-neighbour transpositions, see Theorem 6.2(b) in Armstrong [6]. Therefore, if we have any amplitude $A_{\sigma_{Q_1}\dots\sigma_{Q_N}}(k_{P_1}\dots k_{P_N})$ in ordering sector (P,Q), we can convert it to an amplitude in the sector (P',I) by applying (5.2.14) successively. Note that P' is in general not equal to either P or I.

5.3 The Yang-Baxter Equation

In this section we will derive the Yang-Baxter equation, following section 5.2 of Šamaj and Bajnok [16].

The fact that in integrable systems the scattering of N particles can be decomposed into two-particle scatterings, gives a restriction on the possible forms of the scattering matrix. This can already be seen from the case $N=3$.

Suppose that we have three particles with positions x_1,x_2,x_3 and momenta k_1,k_2,k_3. We start in the ordering sector $Q=(3,2,1)$, so we have $x_3 \leq x_2 \leq x_1$. Then they undergo a scattering after which we end up in the ordering sector $Q'=I=(1,2,3)$. If we decompose this scattering into two-particle scatterings, we see that there are two possible decompositions:

(a) $(3,2,1) \to (3,1,2) \to (1,3,2) \to (1,2,3)$,

(b) $(3,2,1) \to (2,3,1) \to (2,1,3) \to (1,2,3)$.

Using Equation (5.2.14), decomposition (a) gives us

$$\begin{aligned} A_{\sigma_3\sigma_2\sigma_1}(k_3,k_2,k_1) &= S^{\sigma_1\sigma_2}_{\sigma'_1\sigma'_2}(k_1,k_2)A_{\sigma_3\sigma'_1\sigma'_2}(k_3,k_1,k_2), \\ &= S^{\sigma_1\sigma_2}_{\sigma'_1\sigma'_2}(k_1,k_2)S^{\sigma'_1\sigma_3}_{\sigma''_1\sigma'_3}(k_1,k_3)A_{\sigma''_1\sigma'_3\sigma'_2}(k_1,k_3,k_2), \\ &= S^{\sigma_1\sigma_2}_{\sigma'_1\sigma'_2}(k_1,k_2)S^{\sigma'_1\sigma_3}_{\sigma''_1\sigma'_3}(k_1,k_3)S^{\sigma'_2\sigma'_3}_{\sigma''_2\sigma''_3}(k_2,k_3)A_{\sigma''_1\sigma''_2\sigma''_3}(k_1,k_2,k_3). \end{aligned} \tag{5.3.1}$$

Remember that we use the summation convention, so in the last expression on the right-hand side we sum over six indices, namely $\sigma'_1,\sigma'_2,\sigma'_3,\sigma''_1,\sigma''_2$ and σ''_3.

Doing the same thing for decomposition (b) we get

$$\begin{aligned} A_{\sigma_3\sigma_2\sigma_1}(k_3,k_2,k_1) &= S^{\sigma_2\sigma_3}_{\sigma'_2\sigma'_3}(k_2,k_3)A_{\sigma'_2\sigma'_3\sigma_1}(k_2,k_3,k_1), \\ &= S^{\sigma_2\sigma_3}_{\sigma'_2\sigma'_3}(k_2,k_3)S^{\sigma_1\sigma'_3}_{\sigma'_1\sigma''_3}(k_1,k_3)A_{\sigma'_2\sigma'_1\sigma''_3}(k_2,k_1,k_3), \\ &= S^{\sigma_2\sigma_3}_{\sigma'_2\sigma'_3}(k_2,k_3)S^{\sigma_1\sigma'_3}_{\sigma'_1\sigma''_3}(k_1,k_3)S^{\sigma'_1\sigma'_2}_{\sigma''_1\sigma''_2}(k_1,k_2)A_{\sigma''_1\sigma''_2\sigma''_3}(k_1,k_2,k_3). \end{aligned} \tag{5.3.2}$$

Of course, if we claim that the scattering of three particles can be decomposed into two-particle scatterings, we need the result to be independent of the decomposition we choose. Therefore in Equations (5.3.1) and (5.3.2) we must get the same amplitude. This implies that

$$\begin{aligned} &S^{\sigma_1\sigma_2}_{\sigma'_1\sigma'_2}(k_1,k_2)S^{\sigma'_1\sigma_3}_{\sigma''_1\sigma'_3}(k_1,k_3)S^{\sigma'_2\sigma'_3}_{\sigma''_2\sigma''_3}(k_2,k_3)A_{\sigma''_1\sigma''_2\sigma''_3}(k_1,k_2,k_3) \\ &\quad = S^{\sigma_2\sigma_3}_{\sigma'_2\sigma'_3}(k_2,k_3)S^{\sigma_1\sigma'_3}_{\sigma'_1\sigma''_3}(k_1,k_3)S^{\sigma'_1\sigma'_2}_{\sigma''_1\sigma''_2}(k_1,k_2)A_{\sigma''_1\sigma''_2\sigma''_3}(k_1,k_2,k_3). \end{aligned} \tag{5.3.3}$$

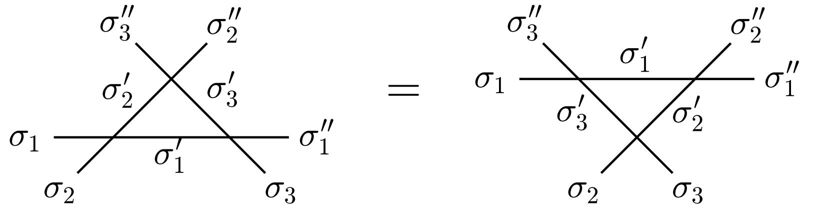

Figure 5.2: A visualisation of the Yang-Baxter equation.

Now, remember that the A-amplitudes were just some constants in our wavefunction, see Equations (5.1.10) and (5.1.11). In principle, for different permutations $(ijk) \in S_3$, we can choose the amplitudes $A_{\sigma_i''\sigma_j''\sigma_k''}$ in any way we want, and Equation (5.3.3) has to hold for any choice of these amplitudes. This means in particular that we can choose all the amplitudes to be 0 except $A_{\sigma_i''\sigma_j''\sigma_k''}$ for one specific permutation $(ijk) \in S_3$. But we can do this for every permutation (ijk), and therefore equation (5.3.3) implies the following relation:

$$S^{\sigma_1\sigma_2}_{\sigma_1'\sigma_2'}(k_1,k_2)S^{\sigma_1'\sigma_3}_{\sigma_1''\sigma_3'}(k_1,k_3)S^{\sigma_2'\sigma_3'}_{\sigma_2''\sigma_3''}(k_2,k_3) = S^{\sigma_2\sigma_3}_{\sigma_2'\sigma_3'}(k_2,k_3)S^{\sigma_1\sigma_3'}_{\sigma_1'\sigma_3''}(k_1,k_3)S^{\sigma_1'\sigma_2'}_{\sigma_1''\sigma_2''}(k_1,k_2), \tag{5.3.4}$$

for every choice of $\sigma_1, \sigma_2, \sigma_3, \sigma_1'', \sigma_2''$ and σ_3'' (we sum over $\sigma_1', \sigma_2', \sigma_3'$). This set of l^6 equations is known as the *Yang-Baxter equation (YBE)*. Conversely, if the YBE is satisfied, then Equation (5.3.3) holds. Therefore the condition that three-particle scatterings can be decomposed into two-particle scatterings independently of the order of the two-particle scatterings, is equivalent to saying that the elements of the scattering matrix have to satisfy the YBE. In fact, if for any N we have two different decompositions of an N-particle scattering into two-particle scatterings, then they are equivalent if and only if the elements of the scattering matrix satisfy the YBE. This is because we can turn the different decompisitions into each other by succesively applying the YBE for three particles. Therefore the YBE is also called the *integrability condition.* Similar to Figure 5.1, the YBE can be visualised as shown in Figure 5.2

The scattering matrix has dimension l^2, and therefore l^4 elements. The YBE gives l^6 equations for these elements. Therefore it would in fact be quite surprising if we can find a solution to the YBE, because we have l^6 equations for l^4 variables. However, in the next chapter we will show that in fact there exists a solution using the Jacobi elliptic functions.

Up to this point, we always wrote the scattering matrix $S(k_u, k_v)$ as a function of the wavenumbers of the two particles corresponding to that scattering matrix. However, it turns out that we can always paramaterize the wavenumbers k_u, k_v in terms of so-called spectral parameters λ_u, λ_v, $k_u = k_u(\lambda_u)$ and $k_v = k_v(\lambda_v)$, such that the scattering matrix only depends on the difference $\lambda_u - \lambda_v$. Then we have $S(k_u, k_v) = S(\lambda_u - \lambda_v)$. In terms of these spectral parameters, we can rewrite the normalization (5.2.9) and unitarity (5.2.10) properties of the scattering matrix as

$$S(0) = \mathscr{P} \qquad \text{and} \qquad S(\lambda)S(-\lambda) = I. \tag{5.3.5}$$

We want to rewrite the YBE in Equation (5.3.4) in terms of spectral parameters. Therefore we parameterize the wavenumbers as $k_1 = k_1(\lambda_1)$, $k_2 = k_2(\lambda_2)$ and $k_3 = k_3(\lambda_3)$. Then, if we define $\lambda = \lambda_1 - \lambda_3$ and $\mu = \lambda_2 - \lambda_3$, we can write the YBE as

$$S^{\sigma_1\sigma_2}_{\sigma_1'\sigma_2'}(\lambda-\mu)S^{\sigma_1'\sigma_3}_{\sigma_1''\sigma_3'}(\lambda)S^{\sigma_2'\sigma_3'}_{\sigma_2''\sigma_3''}(\mu) = S^{\sigma_2\sigma_3}_{\sigma_2'\sigma_3'}(\mu)S^{\sigma_1\sigma_3'}_{\sigma_1'\sigma_3''}(\lambda)S^{\sigma_1'\sigma_2'}_{\sigma_1''\sigma_2''}(\lambda-\mu). \tag{5.3.6}$$

Note that the only difference between two scattering matrices $S(\nu_1)$ and $S(\nu_2)$, is the difference between ν_1 and ν_2.

It would also be nice if we could write the YBE in Equation (5.3.6) as a matrix equation, instead of having to write all these indices all the time. As we said before, the scattering matrix is the matrix representation of an operator acting on the tensor product of two l-dimensional vector spaces, V_1, V_2, one for each particle. For the vector spaces V_1 and V_2 we have the bases $\{e_{\sigma_1} \mid \sigma_1 = 1, \ldots, l\}$ and $\{e_{\sigma_2} \mid \sigma_2 = 1, \ldots, l\}$, respectively. Then we obtain the basis

$$\{(e_{\sigma_1} \otimes e_{\sigma_2}) \mid 1 \leq \sigma_1, \sigma_2 \leq l\} \tag{5.3.7}$$

of $V_1 \otimes V_2$. From usual matrix multiplication we find

$$S(\nu)\left(e_{\sigma_1} \otimes e_{\sigma_2}\right) = S^{\sigma_1\sigma_2}_{\sigma_1'\sigma_2'}(\nu)\left(e_{\sigma_1'} \otimes e_{\sigma_2'}\right). \tag{5.3.8}$$

Remember that we use the summation convention. So in the equation above we sum over σ_1' and σ_2'.

Now, suppose that we have three particles labeled $1, 2$ and 3, with corresponding vector spaces V_1, V_2 and V_3. The scattering of these three particles can be decomposed into three two-particle scatterings, and we get corresponding scattering matrices. These scattering matrices are essentially all the same two-particle scattering matrix S, but acting on different vector spaces. For instance, the scattering matrix corresponding to the scattering of particles 1 and 2 acts on $V_1 \otimes V_2$, but the matrix corresponding to the scattering of particles 1 and 3 acts on $V_1 \otimes V_3$. To be able to write a product of the scattering matrices, we want to view the scattering matrices as operators on $V_1 \otimes V_2 \otimes V_3$. For the scattering matrix $S(\nu)$ for particles i and j, we define the operator S_{ij} acting on $V_1 \otimes V_2 \otimes V_3$ as the operator that acts on V_i and V_j as the scattering matrix S, and on the remaining vector space V_k it acts as the identity. So we have

$$S_{12}(\nu_1) = S(\nu_1) \otimes I \qquad \text{and} \qquad S_{23}(\nu_2) = I \otimes S(\nu_2). \tag{5.3.9}$$

However, we see that for S_{13} we have a problem, because we can not write it as the tensor product of a scattering matrix and an identity operator. We will fix this problem later by introducing the R-matrix. For the moment, it turns out that it does not really matter if we just want to rewrite the YBE (5.3.6) in terms of S_{12}, S_{13} and S_{23}. Similar to Equation (5.3.8), we have

$$\begin{aligned}
S_{12}(\nu_1)(e_{\sigma_1} \otimes e_{\sigma_2} \otimes e_{\sigma_3}) &= S^{\sigma_1\sigma_2}_{\sigma_1'\sigma_2'}(\nu_1)(e_{\sigma_1'} \otimes e_{\sigma_2'} \otimes e_{\sigma_3}),\\
S_{13}(\nu_2)(e_{\sigma_1} \otimes e_{\sigma_2} \otimes e_{\sigma_3}) &= S^{\sigma_1\sigma_3}_{\sigma_1'\sigma_3'}(\nu_2)(e_{\sigma_1'} \otimes e_{\sigma_2} \otimes e_{\sigma_3'}),\\
S_{23}(\nu_3)(e_{\sigma_1} \otimes e_{\sigma_2} \otimes e_{\sigma_3}) &= S^{\sigma_2\sigma_3}_{\sigma_2'\sigma_3'}(\nu_3)(e_{\sigma_1} \otimes e_{\sigma_2'} \otimes e_{\sigma_3'}).
\end{aligned} \tag{5.3.10}$$

Now, using this, we will show that the YBE (5.3.6) is equivalent to

$$S_{12}(\lambda - \mu)S_{13}(\lambda)S_{23}(\mu) = S_{23}(\mu)S_{13}(\lambda)S_{12}(\lambda - \mu). \tag{5.3.11}$$

To see this, we first take the left-hand side of Equation (5.3.11), and we let it act on the basisvector $e_{\sigma_1} \otimes e_{\sigma_2} \otimes e_{\sigma_3}$. Then we get

$$\begin{aligned} S_{12}(\lambda-\mu)S_{13}(\lambda)S_{23}(\mu)\left(e_{\sigma_1} \otimes e_{\sigma_2} \otimes e_{\sigma_3}\right) \\ &= S_{12}(\lambda-\mu)S_{13}(\lambda)\left[S^{\sigma_2\sigma_3}_{\sigma'_2\sigma'_3}(\mu)(e_{\sigma_1} \otimes e_{\sigma'_2} \otimes e_{\sigma'_3})\right] \\ &= S_{12}(\lambda-\mu)\left[S^{\sigma_2\sigma_3}_{\sigma'_2\sigma'_3}(\mu)S_{13}(\lambda)(e_{\sigma_1} \otimes e_{\sigma'_2} \otimes e_{\sigma'_3})\right] \\ &= S_{12}(\lambda-\mu)\left[S^{\sigma_2\sigma_3}_{\sigma'_2\sigma'_3}(\mu)S^{\sigma_1\sigma'_3}_{\sigma'_1\sigma''_3}(\lambda)(e_{\sigma'_1} \otimes e_{\sigma'_2} \otimes e_{\sigma''_3})\right] \\ &= S^{\sigma_2\sigma_3}_{\sigma'_2\sigma'_3}(\mu)S^{\sigma_1\sigma'_3}_{\sigma'_1\sigma''_3}(\lambda)S_{12}(\lambda-\mu)(e_{\sigma'_1} \otimes e_{\sigma'_2} \otimes e_{\sigma''_3}) \\ &= S^{\sigma_2\sigma_3}_{\sigma'_2\sigma'_3}(\mu)S^{\sigma_1\sigma'_3}_{\sigma'_1\sigma''_3}(\lambda)S^{\sigma'_1\sigma'_2}_{\sigma''_1\sigma''_2}(\lambda-\mu)(e_{\sigma''_1} \otimes e_{\sigma''_2} \otimes e_{\sigma''_3}). \end{aligned} \tag{5.3.12}$$

Doing the same thing for the right-hand side of Equation (5.3.11), we find

$$\begin{aligned} S_{23}(\mu)S_{13}(\lambda)S_{12}(\lambda-\mu)\left(e_{\sigma_1} \otimes e_{\sigma_2} \otimes e_{\sigma_3}\right) \\ &= S^{\sigma_1\sigma_2}_{\sigma'_1\sigma'_2}(\lambda-\mu)S^{\sigma'_1\sigma_3}_{\sigma''_1\sigma'_3}(\lambda)S^{\sigma'_2\sigma'_3}_{\sigma''_2\sigma''_3}(\mu)(e_{\sigma''_1} \otimes e_{\sigma''_2} \otimes e_{\sigma''_3}). \end{aligned} \tag{5.3.13}$$

Therefore Equation (5.3.11) implies that for any $\sigma_1, \sigma_2, \sigma_3$ we have

$$\begin{aligned} S^{\sigma_2\sigma_3}_{\sigma'_2\sigma'_3}(\mu)S^{\sigma_1\sigma'_3}_{\sigma'_1\sigma''_3}(\lambda)S^{\sigma'_1\sigma'_2}_{\sigma''_1\sigma''_2}(\lambda-\mu)(e_{\sigma''_1} \otimes e_{\sigma''_2} \otimes e_{\sigma''_3}) \\ &= S^{\sigma_1\sigma_2}_{\sigma'_1\sigma'_2}(\lambda-\mu)S^{\sigma'_1\sigma_3}_{\sigma''_1\sigma'_3}(\lambda)S^{\sigma'_2\sigma'_3}_{\sigma''_2\sigma''_3}(\mu)(e_{\sigma''_1} \otimes e_{\sigma''_2} \otimes e_{\sigma''_3}) \end{aligned} \tag{5.3.14}$$

Now, note that on both sides of this equation we sum over σ''_1, σ''_2 and σ''_3, but the summand consists of a number times the basisvector $e_{\sigma''_1} \otimes e_{\sigma''_2} \otimes e_{\sigma''_3}$. These basisvectors are linearly independent, so therefore we see that for every choice of σ''_1, σ''_2 and σ''_3 we must have

$$S^{\sigma_2\sigma_3}_{\sigma'_2\sigma'_3}(\mu)S^{\sigma_1\sigma'_3}_{\sigma'_1\sigma''_3}(\lambda)S^{\sigma'_1\sigma'_2}_{\sigma''_1\sigma''_2}(\lambda-\mu) = S^{\sigma_1\sigma_2}_{\sigma'_1\sigma'_2}(\lambda-\mu)S^{\sigma'_1\sigma_3}_{\sigma''_1\sigma'_3}(\lambda)S^{\sigma'_2\sigma'_3}_{\sigma''_2\sigma''_3}(\mu). \tag{5.3.15}$$

This is precisely the YBE as in Equation (5.3.6). This way Equation (5.3.11) implies the YBE. To see that they are in fact equivalent, we note that by reversing the steps we just did, it follows from Equation (5.3.15) that

$$S_{12}(\lambda-\mu)S_{13}(\lambda)S_{23}(\mu)\left(e_{\sigma_1} \otimes e_{\sigma_2} \otimes e_{\sigma_3}\right) = S_{23}(\mu)S_{13}(\lambda)S_{12}(\lambda-\mu)\left(e_{\sigma_1} \otimes e_{\sigma_2} \otimes e_{\sigma_3}\right), \tag{5.3.16}$$

for any choice of σ_1, σ_2 and σ_3. Because an operator acting on a vector space is completely determined by how it acts on the basisvectors of that space, we find indeed that the above equation implies Equation (5.3.11). Therefore we can view (5.3.11) as the YBE written in matrix form.

From our normalization condition $S(0) = \mathscr{P}$ (also called *initial condition*), it follows that if we fill in $\lambda = \mu = 0$ in the YBE (5.3.11), we get

$$\mathscr{P}_{12}\mathscr{P}_{13}\mathscr{P}_{23} = \mathscr{P}_{23}\mathscr{P}_{13}\mathscr{P}_{12}. \tag{5.3.17}$$

Here $\mathscr{P}_{ij}$ is an operator on $V_1 \otimes V_2 \otimes V_3$ defined in the same way as S_{ij}. The above identity is always true, as both sides are equal to $\mathscr{P}_{13}$. Therefore, for $\lambda = \mu = 0$, the YBE is always satisfied.

We finish this section with introducing the R-matrix. Remember that we had a "problem" for the operator S_{13}, because we could not write it as the tensor product of a scattering matrix and an identity operator. Writing the operators in terms of the R-matrix will remove this problem. We define the R-matrix as

$$R(\nu) = \mathscr{P}S(\nu). \tag{5.3.18}$$

Then the initial condition $S(0) = \mathscr{P}$ becomes

$$R(0) = I. \tag{5.3.19}$$

Similar to Equation (5.3.8), we find for the permutation matrix that

$$\begin{aligned}\mathscr{P}(e_{\sigma_1} \otimes e_{\sigma_2}) &= \sum_{\sigma_1'\sigma_2'} \delta^{\sigma_1}_{\sigma_2'} \delta^{\sigma_2}_{\sigma_1'} (e_{\sigma_1'} \otimes e_{\sigma_2'}) \\ &= (e_{\sigma_2} \otimes e_{\sigma_1})\end{aligned} \tag{5.3.20}$$

Now we can do the same thing for the R-matrix to find

$$\begin{aligned}R(\nu)\,(e_{\sigma_1} \otimes e_{\sigma_2}) &= \mathscr{P}\,[S(\nu)\,(e_{\sigma_1} \otimes e_{\sigma_2})] \\ &= \mathscr{P}\left[S^{\sigma_1\sigma_2}_{\sigma_1'\sigma_2'}(\nu)\left(e_{\sigma_1'} \otimes e_{\sigma_2'}\right)\right] \\ &= S^{\sigma_1\sigma_2}_{\sigma_1'\sigma_2'}(\nu)\mathscr{P}\left(e_{\sigma_1'} \otimes e_{\sigma_2'}\right) \\ &= S^{\sigma_1\sigma_2}_{\sigma_1'\sigma_2'}(\nu)\left(e_{\sigma_2'} \otimes e_{\sigma_1'}\right)\end{aligned} \tag{5.3.21}$$

If we compare this to Equation (5.3.8), we see that $R^{\sigma_1\sigma_2}_{\sigma_1'\sigma_2'} = S^{\sigma_1\sigma_2}_{\sigma_2'\sigma_1'}$. However, in Equation (5.2.8) we found that

$$R^{\sigma_1\sigma_2}_{\sigma_1'\sigma_2'} = S^{\sigma_2\sigma_1}_{\sigma_1'\sigma_2'}. \tag{5.3.22}$$

These statements are indeed equivalent, because we assumed the symmetry $S^{\sigma_1\sigma_2}_{\sigma_1'\sigma_2'} = S^{\sigma_2\sigma_1}_{\sigma_2'\sigma_1'}$ for S.

Now, we can write the YBE (5.3.11) in terms of the R-matrix as

$$[I \otimes R(\lambda-\mu)]\,[R(\lambda) \otimes I]\,[I \otimes R(\mu)] = [R(\mu) \otimes I]\,[I \otimes R(\lambda)]\,[R(\lambda-\mu) \otimes I]\,. \tag{5.3.23}$$

To see that this is true, we compare the left-hand side of the above Equation to that of Equation (5.3.11). We can split the R-matrix in a permutation operator and a scattering matrix. Therefore we can write the left-hand side of Equation (5.3.23) as

$$[I \otimes \mathscr{P}]\,[I \otimes S(\lambda-\mu)]\,[\mathscr{P} \otimes I]\,[S(\lambda) \otimes I]\,[I \otimes \mathscr{P}]\,[I \otimes S(\mu)] \tag{5.3.24}$$

This expression is an operator on the space $V_1 \otimes V_2 \otimes V_3$. We see that we first act with the operator $[I \otimes S(\mu)]$, which is exactly $S_{23}(\mu)$ as in the left-hand side of Equation (5.3.11). But then we act with $[I \otimes \mathscr{P}]$. From Equation (5.3.20) it follows that the operator $[I \otimes \mathscr{P}]$ "switches" the spaces V_2 and V_3. So then the next operator, in this case $[S(\lambda) \otimes I]$, acts on the space $V_1 \otimes V_3 \otimes V_2$. Therefore we see that it acts as $S_{13}(\lambda)$ as in Equation (5.3.11)! After that, the operator $[\mathscr{P} \otimes I]$ switches V_1 and V_3 such that $[I \otimes S(\lambda-\mu)]$ acts on $V_3 \otimes V_1 \otimes V_2$, exactly like $S_{12}(\lambda-\mu)$ in Equation (5.3.11). The last term $[I \otimes \mathscr{P}]$ then switches V_1 and V_2 to get the space $V_3 \otimes V_2 \otimes V_1$. So the left-hand sides of Equations (5.3.11) and (5.3.23) are operators that act in exactly the same way, except that in Equation (5.3.23) we change the space $V_1 \otimes V_2 \otimes V_3$ into $V_3 \otimes V_2 \otimes V_1$.

Now for the right-hand sides we can do the same thing. Then we find that in Equations (5.3.11) and (5.3.23) we have essentially the same operator, but that in Equation (5.3.23) changes the space $V_1 \otimes V_2 \otimes V_3$ into $V_3 \otimes V_2 \otimes V_1$. So that is just like the left-hand side. Therefore we find that Equations (5.3.11) and (5.3.23) are indeed equivalent in a sort of trivial way, because they say exactly the same thing. So now we have written the YBE as a matrix equation in terms of matrices we understand (unlike S_{13}).

5.4 Lax Operators and Transfer Matrices

In this section we will use the scattering matrix to define larger operators that act on the entire chain of N sites at once. Then we will derive relations for these operators from the Yang-Baxter equation. We follow Section 5.3 of Šamaj and Bajnok [16].

Remember that the system we are studying is a chain of N sites. In each site we have exactly one particle. We say that the state of a site is the color of the particle on that site, for which there are l possibilities. So the n'th site has state index $\sigma_n = 1, \dots, l$, where σ_n represents the color of the n'th particle, just as before. Now, we add two *auxiliary sites* ξ and η to the chain, with state indices γ_ξ and γ_η that also have have the possible values $1, \dots, l$. These auxiliary sites have no physical meaning, but we will treat them just like the other sites on the chain. The point of these extra sites is that they make the notation more convenient, nothing more.

We define the *Lax operators* $L_{\xi n}$ for $n = 1, 2, \dots, N$ as follows:

$$L_{\xi n}(\lambda)^{\gamma_\xi \sigma_1 \dots \sigma_N}_{\gamma'_\xi \sigma'_1 \dots \sigma'_N} = S_{\xi n}(\lambda)^{\gamma_\xi \sigma_n}_{\gamma'_\xi \sigma'_n} \delta^{\sigma_1}_{\sigma'_1} \cdots \delta^{\sigma_{n-1}}_{\sigma'_{n-1}} \delta^{\sigma_{n+1}}_{\sigma'_{n+1}} \cdots \delta^{\sigma_N}_{\sigma'_N}. \tag{5.4.1}$$

Each Lax operator acts on the entire chain of N sites and one auxiliary site ξ, so it has dimension l^{N+1}. We see that $L_{\xi n}$ acts trivially on the chain of N sites, except for the n'th site. Similarly we define $L_{\eta n}$. To lighten the notation we write

$$\begin{aligned} \delta^{\sigma_1}_{\sigma'_1} \cdots \delta^{\sigma_{n-1}}_{\sigma'_{n-1}} \delta^{\sigma_{n+1}}_{\sigma'_{n+1}} \cdots \delta^{\sigma_N}_{\sigma'_N} &= \Delta_{\prime}, \\ \delta^{\sigma'_1}_{\sigma''_1} \cdots \delta^{\sigma'_{n-1}}_{\sigma''_{n-1}} \delta^{\sigma'_{n+1}}_{\sigma''_{n+1}} \cdots \delta^{\sigma'_N}_{\sigma''_N} &= \Delta^{\prime}_{\prime\prime}. \end{aligned} \tag{5.4.2}$$

Note that $\Delta_{\prime}$ and $\Delta^{\prime}_{\prime\prime}$ depend on n. Also $\Delta_{\prime}\Delta^{\prime}_{\prime\prime} = \Delta_{\prime\prime}$ (we use the summation convention), where $\Delta_{\prime\prime}$ is defined similarly to (5.4.2).

We can write the YBE (5.3.6) in terms of Lax operators as

$$S_{\xi\eta}(\lambda - \mu) L_{\xi n}(\lambda) L_{\eta n}(\mu) = L_{\eta n}(\mu) L_{\xi n}(\lambda) S_{\xi\eta}. \tag{5.4.3}$$

This can be seen by writing it out in coördinates. The following derivation consists of five equations, but they are too wide to be displayed on one line. Each equation follows from the one before it.

$$\begin{aligned} &[S_{\xi\eta}(\lambda - \mu) L_{\xi n}(\lambda) L_{\eta n}(\mu)]^{\gamma_\xi \gamma_\eta \sigma_1 \dots \sigma_N}_{\gamma''_\xi \gamma''_\eta \sigma''_1 \dots \sigma''_N} \\ &\qquad\qquad = [L_{\eta n}(\mu)\, L_{\xi n}(\lambda) S_{\xi\eta}]^{\gamma_\xi \gamma_\eta \sigma_1 \dots \sigma_N}_{\gamma''_\xi \gamma''_\eta \sigma''_1 \dots \sigma''_N} \\ &S_{\xi\eta}(\lambda - \mu)^{\gamma_\xi \gamma_\eta}_{\gamma'_\xi \gamma'_\eta} L_{\xi n}(\lambda)^{\gamma'_\xi \sigma_1 \dots \sigma_N}_{\gamma''_\xi \sigma'_1 \dots \sigma'_N} L_{\eta n}(\mu)^{\gamma'_\eta \sigma'_1 \dots \sigma'_N}_{\gamma''_\eta \sigma''_1 \dots \sigma''_N} \\ &\qquad\qquad = L_{\eta n}(\mu)^{\gamma_\eta \sigma_1 \dots \sigma_N}_{\gamma'_\eta \sigma'_1 \dots \sigma'_N} L_{\xi n}(\lambda)^{\gamma_\xi \sigma'_1 \dots \sigma'_N}_{\gamma'_\xi \sigma''_1 \dots \sigma''_N} S_{\xi\eta}(\lambda - \mu)^{\gamma'_\xi \gamma'_\eta}_{\gamma''_\xi \gamma''_\eta} \\ &S_{\xi\eta}(\lambda - \mu)^{\gamma_\xi \gamma_\eta}_{\gamma'_\xi \gamma'_\eta} S_{\xi n}(\lambda)^{\gamma'_\xi \sigma_n}_{\gamma''_\xi \sigma'_n} \Delta_{\prime} S_{\eta n}(\mu)^{\gamma'_\eta \sigma'_n}_{\gamma''_\eta \sigma''_n} \Delta^{\prime}_{\prime\prime} \\ &\qquad\qquad = S_{\eta n}(\mu)^{\gamma_\eta \sigma_n}_{\gamma'_\eta \sigma'_n} \Delta_{\prime} S_{\xi n}(\lambda)^{\gamma_\xi \sigma'_n}_{\gamma'_\xi \sigma''_n} \Delta^{\prime}_{\prime\prime} S_{\xi\eta}(\lambda - \mu)^{\gamma'_\xi \gamma'_\eta}_{\gamma''_\xi \gamma''_\eta} \\ &S_{\xi\eta}(\lambda - \mu)^{\gamma_\xi \gamma_\eta}_{\gamma'_\xi \gamma'_\eta} S_{\xi n}(\lambda)^{\gamma'_\xi \sigma_n}_{\gamma''_\xi \sigma'_n} S_{\eta n}(\mu)^{\gamma'_\eta \sigma'_n}_{\gamma''_\eta \sigma''_n} \Delta_{\prime\prime} \\ &\qquad\qquad = S_{\eta n}(\mu)^{\gamma_\eta \sigma_n}_{\gamma'_\eta \sigma'_n} S_{\xi n}(\lambda)^{\gamma_\xi \sigma'_n}_{\gamma'_\xi \sigma''_n} S_{\xi\eta}(\lambda - \mu)^{\gamma'_\xi \gamma'_\eta}_{\gamma''_\xi \gamma''_\eta} \Delta_{\prime\prime} \\ &S_{\xi\eta}(\lambda - \mu)^{\gamma_\xi \gamma_\eta}_{\gamma'_\xi \gamma'_\eta} S_{\xi n}(\lambda)^{\gamma'_\xi \sigma_n}_{\gamma''_\xi \sigma'_n} S_{\eta n}(\mu)^{\gamma'_\eta \sigma'_n}_{\gamma''_\eta \sigma''_n} \\ &\qquad\qquad = S_{\eta n}(\mu)^{\gamma_\eta \sigma_n}_{\gamma'_\eta \sigma'_n} S_{\xi n}(\lambda)^{\gamma_\xi \sigma'_n}_{\gamma'_\xi \sigma''_n} S_{\xi\eta}(\lambda - \mu)^{\gamma'_\xi \gamma'_\eta}_{\gamma''_\xi \gamma''_\eta}. \end{aligned} \tag{5.4.4}$$

Here we used that in this case the Δ's commute with the scattering matrices, as they act on different sites. In the second to last equation we have a $\Delta_{''}$ on both sides. This $\Delta_{''}$ acts on all sites except for the n'th one and the auxiliary sites. The scattering matrices, however, act only on these three sites. So in the last step we just choose coordinates $\sigma_i = \sigma'_i$ for $i \neq n$, such that $\Delta_{''} = 1$. Then we obtain indeed the YBE as in Equation (5.3.6). To show that (5.4.3) follows from (5.3.6) we just reverse the order in (5.4.4).

To make sense of equation (5.4.3), we want to view the Lax operators as operators on their corresponding auxiliary space. So $L_{\xi n}$ acts on ξ-space. Then their dimension will be just l, but their elements will be operators of dimension l^N acting on the chain of N sites:

$$\left[L_{\xi n}(\lambda)^{\gamma_\xi}_{\gamma'_\xi}\right]^{\sigma_1\ldots\sigma_N}_{\sigma'_1\ldots\sigma'_N} = L_{\xi n}(\lambda)^{\gamma_\xi\sigma_1\ldots\sigma_N}_{\gamma'_\xi\sigma'_1\ldots\sigma'_N} = S_{\xi n}(\lambda)^{\gamma_\xi\sigma_n}_{\gamma'_\xi\sigma'_n}\Delta_{'}. \tag{5.4.5}$$

Writing out equation (5.4.3) in coordinates in (ξ,η)-space, we obtain:

$$S_{\xi\eta}(\lambda-\mu)^{\gamma_\xi\gamma_\eta}_{\gamma'_\xi\gamma'_\eta} L_{\xi n}(\lambda)^{\gamma'_\xi}_{\gamma''_\xi} L_{\eta n}(\mu)^{\gamma'_\eta}_{\gamma''_\eta} = L_{\eta n}(\mu)^{\gamma_\eta}_{\gamma'_\eta} L_{\xi n}(\lambda)^{\gamma_\xi}_{\gamma'_\xi} S_{\xi\eta}(\lambda-\mu)^{\gamma'_\xi\gamma'_\eta}_{\gamma''_\xi\gamma''_\eta}. \tag{5.4.6}$$

To get the YBE in terms of the R matrix, we act with $\mathscr{P}_{\xi\eta}$ from the left to find:

$$\begin{aligned}
\mathscr{P}^{\gamma_\xi\gamma_\eta}_{\gamma'''_\xi\gamma'''_\eta} S_{\xi\eta}(\lambda-\mu)^{\gamma'''_\xi\gamma'''_\eta}_{\gamma'_\xi\gamma'_\eta} L_{\xi n}(\lambda)^{\gamma'_\xi}_{\gamma''_\xi} L_{\eta n}(\mu)^{\gamma'_\eta}_{\gamma''_\eta} &= \mathscr{P}^{\gamma_\xi\gamma_\eta}_{\gamma'''_\xi\gamma'''_\eta} L_{\eta n}(\mu)^{\gamma'''_\eta}_{\gamma'_\eta} L_{\xi n}(\lambda)^{\gamma'''_\xi}_{\gamma'_\xi} S_{\xi\eta}(\lambda-\mu)^{\gamma'_\xi\gamma'_\eta}_{\gamma''_\xi\gamma''_\eta} \\
R_{\xi\eta}(\lambda-\mu)^{\gamma_\xi\gamma_\eta}_{\gamma'_\xi\gamma'_\eta} L_{\xi n}(\lambda)^{\gamma'_\xi}_{\gamma''_\xi} L_{\eta n}(\mu)^{\gamma'_\eta}_{\gamma''_\eta} &= \delta^{\gamma_\xi}_{\gamma'''_\eta}\delta^{\gamma_\eta}_{\gamma'''_\xi} L_{\eta n}(\mu)^{\gamma'''_\eta}_{\gamma'_\eta} L_{\xi n}(\lambda)^{\gamma'''_\xi}_{\gamma'_\xi} R_{\xi\eta}(\lambda-\mu)^{\gamma'_\eta\gamma'_\xi}_{\gamma''_\xi\gamma''_\eta} \\
R_{\xi\eta}(\lambda-\mu)^{\gamma_\xi\gamma_\eta}_{\gamma'_\xi\gamma'_\eta} L_{\xi n}(\lambda)^{\gamma'_\xi}_{\gamma''_\xi} L_{\eta n}(\mu)^{\gamma'_\eta}_{\gamma''_\eta} &= L_{\eta n}(\mu)^{\gamma_\xi}_{\gamma'_\eta} L_{\xi n}(\lambda)^{\gamma_\eta}_{\gamma'_\xi} R_{\xi\eta}(\lambda-\mu)^{\gamma'_\eta\gamma'_\xi}_{\gamma''_\xi\gamma''_\eta}.
\end{aligned} \tag{5.4.7}$$

Here we used that $S_{\xi\eta}(\lambda-\mu)^{\gamma'_\xi\gamma'_\eta}_{\gamma''_\xi\gamma''_\eta} = S_{\xi\eta}(\lambda-\mu)^{\gamma'_\eta\gamma'_\xi}_{\gamma''_\eta\gamma''_\xi} = R_{\xi\eta}(\lambda-\mu)^{\gamma'_\eta\gamma'_\xi}_{\gamma''_\xi\gamma''_\eta}$. Note that γ_η and γ_ξ are both just indices with possible values $1,\ldots,l$. Therefore $L_{\eta n}(\mu)^{\gamma_\xi}_{\gamma'_\eta}$ does make sense.

We define the *monodromy matrix* $\mathscr{T}_\xi$ as follows:

$$\mathscr{T}_\xi(\lambda)^{\gamma_\xi\sigma_1\ldots\sigma_N}_{\gamma'_\xi\sigma'_1\ldots\sigma'_N} = S^{\sigma_1\gamma_\xi}_{\sigma'_1\gamma_2}(\lambda) S^{\sigma_2\gamma_2}_{\sigma'_2\gamma_3}(\lambda)\cdots S^{\sigma_N\gamma_N}_{\sigma'_N\gamma'_\xi}. \tag{5.4.8}$$

We can write the monodromy matrix as a product of Lax operators in the following way:

$$\mathscr{T}_\xi(\lambda) = L_{\xi 1}(\lambda)L_{\xi 2}(\lambda)\cdots L_{\xi N}(\lambda). \tag{5.4.9}$$

We can see this by writing it out in coördinates:

$$\begin{aligned}
[L_{\xi 1}(\lambda)L_{\xi 2}(\lambda)\cdots L_{\xi N}(\lambda)]^{\gamma_\xi\sigma_1\ldots\sigma_N}_{\gamma'_\xi\sigma'_1\ldots\sigma'_N} &= \Big[\Big(S^{\gamma_\xi\sigma_1}_{\gamma_2(\sigma_1)_2}(\lambda)\delta^{\sigma_2}_{(\sigma_2)_2}\cdots\delta^{\sigma_N}_{(\sigma_N)_2}\Big) \\
&\qquad \Big(S^{\gamma_2(\sigma_2)_2}_{\gamma_3(\sigma_2)_3}(\lambda)\delta^{(\sigma_1)_2}_{(\sigma_1)_3}\delta^{(\sigma_3)_2}_{(\sigma_3)_3}\cdots\delta^{(\sigma_N)_2}_{(\sigma_N)_3}\Big) \\
&\qquad \cdots\Big(S^{\gamma_N(\sigma_N)_N}_{\gamma'_\xi\sigma'_N}(\lambda)\delta^{(\sigma_1)_N}_{\sigma'_1}\cdots\delta^{(\sigma_{N-1})_N}_{\sigma'_{N-1}}\Big)\Big] \\
&= S^{\gamma_\xi\sigma_1}_{\gamma_2\sigma'_1}(\lambda)S^{\gamma_2\sigma_2}_{\gamma_3\sigma'_2}(\lambda)\cdots S^{\gamma_N\sigma_N}_{\gamma'_\xi\sigma'_N} \\
&= \mathscr{T}_\xi(\lambda)^{\gamma_\xi\sigma_1\ldots\sigma_N}_{\gamma'_\xi\sigma'_1\ldots\sigma'_N}.
\end{aligned} \tag{5.4.10}$$

In the last step we used that $S^{ab}_{a'b'} = S^{ba}_{b'a'}$. Now, analogous to (5.4.3) for the Lax operators, we have the following relation for the monodromy matrix:

$$S_{\xi\eta}(\lambda-\mu)\mathscr{T}_\xi(\lambda)\mathscr{T}_\eta(\mu) = \mathscr{T}_\eta(\mu)\mathscr{T}_\xi(\lambda)S_{\xi\eta}(\lambda-\mu). \tag{5.4.11}$$

To show this, we note that $L_{\xi n}$ and $L_{\eta m}$ commute for $n \neq m$, because they act non-trivially only on different sites. Using this fact and equation (5.4.9) we find

$$\mathscr{T}_\xi(\lambda)\mathscr{T}_\eta(\mu) = L_{\xi 1}(\lambda)L_{\eta 1}(\mu)\cdots L_{\xi N}(\lambda)L_{\eta N}(\mu). \tag{5.4.12}$$

Now, multiplying from the left with $S_{\xi\eta}(\lambda-\mu)$ and using equation (5.4.3) N times, we see that indeed

$$\begin{aligned} S_{\xi\eta}(\lambda-\mu)\mathscr{T}_\xi(\lambda)\mathscr{T}_\eta(\mu) &= S_{\xi\eta}(\lambda-\mu)L_{\xi 1}(\lambda)L_{\eta 1}(\mu)\cdots L_{\xi N}(\lambda)L_{\eta N}(\mu) \\ &= L_{\eta 1}(\mu)L_{\xi 1}(\lambda)\cdots L_{\eta N}(\mu)L_{\xi N}(\lambda)S_{\xi\eta}(\lambda-\mu) \\ &= \mathscr{T}_\eta(\mu)\mathscr{T}_\xi(\lambda)S_{\xi\eta}(\lambda-\mu). \end{aligned} \tag{5.4.13}$$

As for equation (5.4.3), we want to look at equation (5.4.11) coördinate-wise in just (ξ,η)-space. So we see $\mathscr{T}_\xi$ and $\mathscr{T}_\eta$ as an l-dimensional operators in ξ or η space respectively. Their elements will be l^N-dimensional operators acting on the chain of N sites:

$$\left[\mathscr{T}_\xi(\lambda)^{\gamma_\xi}_{\gamma'_\xi}\right]^{\sigma_1\ldots\sigma_N}_{\sigma'_1\ldots\sigma'_N} = \left[L_{\xi 1}(\lambda)^{\gamma_\xi}_{\gamma_2}L_{\xi 2}(\lambda)^{\gamma_2}_{\gamma_3}\cdots L_{\xi N}(\lambda)^{\gamma_N}_{\gamma'_\xi}\right]^{\sigma_1\ldots\sigma_N}_{\sigma'_1\ldots\sigma'_N}. \tag{5.4.14}$$

Then equation (5.4.11) becomes

$$S_{\xi\eta}(\lambda-\mu)^{\gamma_\xi\gamma_\eta}_{\gamma'_\xi\gamma'_\eta}\mathscr{T}_\xi(\lambda)^{\gamma'_\xi}_{\gamma''_\xi}\mathscr{T}_\eta(\mu)^{\gamma'_\eta}_{\gamma''_\eta} = \mathscr{T}_\eta(\mu)^{\gamma_\eta}_{\gamma'_\eta}\mathscr{T}_\xi(\lambda)^{\gamma_\xi}_{\gamma'_\xi}S_{\xi\eta}(\lambda-\mu)^{\gamma'_\xi\gamma'_\eta}_{\gamma''_\xi\gamma''_\eta}. \tag{5.4.15}$$

Completely analogous to (5.4.7), we can act with the permutation matrix from the left to get the relation in terms of the R matrix:

$$R_{\xi\eta}(\lambda-\mu)^{\gamma_\xi\gamma_\eta}_{\gamma'_\xi\gamma'_\eta}\mathscr{T}_\xi(\lambda)^{\gamma'_\xi}_{\gamma''_\xi}\mathscr{T}_\eta(\mu)^{\gamma'_\eta}_{\gamma''_\eta} = \mathscr{T}_\eta(\mu)^{\gamma_\xi}_{\gamma'_\eta}\mathscr{T}_\xi(\lambda)^{\gamma_\eta}_{\gamma'_\xi}R_{\xi\eta}(\lambda-\mu)^{\gamma'_\eta\gamma'_\xi}_{\gamma''_\xi\gamma''_\eta}. \tag{5.4.16}$$

We define the *transfer matrix* T of dimension l^N as

$$T(\lambda)^{\sigma_1\ldots\sigma_N}_{\sigma'_1\ldots\sigma'_N} = S^{\sigma_1\gamma_1}_{\sigma'_1\gamma_2}(\lambda)S^{\sigma_2\gamma_2}_{\sigma'_2\gamma_3}(\lambda)\cdots S^{\sigma_N\gamma_N}_{\sigma'_N\gamma_1}(\lambda) \tag{5.4.17}$$

If we let $\mathrm{Tr}_\xi\cdots = \sum_{\gamma_\xi,\gamma'_\xi}\delta^{\gamma_\xi}_{\gamma'_\xi}\cdots$ the trace in ξ-space, then we have

$$T(\lambda) = \mathrm{Tr}_\xi\ \mathscr{T}_\xi(\lambda). \tag{5.4.18}$$

The transfer matrix is interesting mainly because of the following result:

$$[T(\lambda),T(\mu)] = 0 \qquad \text{for any } \lambda,\mu\in\mathbb{C}. \tag{5.4.19}$$

Here $[A,B] = AB - BA$ is the commutator of A and B. So the statement above is equivalent to saying that transfer matrices depending on arbitrary spectral parameters commute with each other. The important consequence of this, is that those transfer matrices are therefore *simultaneously diagonalizable*, see Exercise 8.21 in Roman [18]. This means that there exists a single invertible matrix P, such that $P^{-1}T(\lambda)P$ is a diagonal matrix for any λ. In the next chapter we will construct a Hamiltonian from the Transfer matrix, and then because of the fact that the transfer matrices are simultaneously diagonalizable we can find the whole energy spectrum of this Hamiltonian. For now, we will just show that (5.4.19) is true. In order to see this we consider Equation (5.4.16). Note that the matrix $R_{\xi\eta}(\lambda-\mu)$ is invertible, and its inverse is given by

$$R^{-1}_{\xi\eta}(\lambda-\mu) = R_{\xi\eta}(\mu-\lambda). \tag{5.4.20}$$

This can be seen from the fact that $R = \mathscr{P}S = S\mathscr{P}$. Then we find

$$\begin{aligned} R_{\xi\eta}(\lambda-\mu)R_{\xi\eta}(\mu-\lambda) &= S_{\xi\eta}(\lambda-\mu)\mathscr{P}\mathscr{P}S_{\xi\eta}(\mu-\lambda) \\ &= S_{\xi\eta}(\lambda-\mu)S_{\xi\eta}(\mu-\lambda) \\ &= I. \end{aligned} \tag{5.4.21}$$

For elements of the inverse of the R-matrix we have then

$$R_{\xi\eta}(\lambda-\mu)^{\gamma_\xi\gamma_\eta}_{\gamma'_\xi\gamma'_\eta} R^{-1}_{\xi\eta}(\lambda-\mu)^{\gamma'_\xi\gamma'_\eta}_{\gamma''_\xi\gamma''_\eta} = \delta^{\gamma_\xi}_{\gamma''_\xi}\delta^{\gamma_\eta}_{\gamma''_\eta}. \tag{5.4.22}$$

Now, we act on Equation (5.4.16) from the right with the inverse of the R-matrix. Then we get

$$\begin{aligned} &R_{\xi\eta}(\lambda-\mu)^{\gamma_\xi\gamma_\eta}_{\gamma'_\xi\gamma'_\eta}\mathscr{T}_\xi(\lambda)^{\gamma'_\xi}_{\gamma''_\xi}\mathscr{T}_\eta(\mu)^{\gamma'_\eta}_{\gamma''_\eta}R^{-1}_{\xi\eta}(\lambda-\mu)^{\gamma''_\xi\gamma''_\eta}_{\gamma'''_\xi\gamma'''_\eta} \\ &\qquad = \mathscr{T}_\eta(\mu)^{\gamma_\xi}_{\gamma'_\eta}\mathscr{T}_\xi(\lambda)^{\gamma_\eta}_{\gamma'_\xi}R_{\xi\eta}(\lambda-\mu)^{\gamma'_\eta\gamma'_\xi}_{\gamma''_\xi\gamma''_\eta}R^{-1}_{\xi\eta}(\lambda-\mu)^{\gamma''_\xi\gamma''_\eta}_{\gamma'''_\xi\gamma'''_\eta}. \end{aligned} \tag{5.4.23}$$

Remember that we use the summation convention. We can now take the traces $\mathrm{Tr}_\xi \cdots = \sum_{\gamma_\xi\gamma'''_\xi}\delta^{\gamma_\xi}_{\gamma'''_\xi}\cdots$ and $\mathrm{Tr}_\eta \cdots = \sum_{\gamma_\eta\gamma'''_\eta}\delta^{\gamma_\eta}_{\gamma'''_\eta}\cdots$ on both sides of the equation to get

$$\begin{aligned} &R_{\xi\eta}(\lambda-\mu)^{\gamma_\xi\gamma_\eta}_{\gamma'_\xi\gamma'_\eta}\mathscr{T}_\xi(\lambda)^{\gamma'_\xi}_{\gamma''_\xi}\mathscr{T}_\eta(\mu)^{\gamma'_\eta}_{\gamma''_\eta}R^{-1}_{\xi\eta}(\lambda-\mu)^{\gamma''_\xi\gamma''_\eta}_{\gamma_\xi\gamma_\eta} \\ &\qquad = \mathscr{T}_\eta(\mu)^{\gamma_\xi}_{\gamma'_\eta}\mathscr{T}_\xi(\lambda)^{\gamma_\eta}_{\gamma'_\xi}R_{\xi\eta}(\lambda-\mu)^{\gamma'_\eta\gamma'_\xi}_{\gamma''_\xi\gamma''_\eta}R^{-1}_{\xi\eta}(\lambda-\mu)^{\gamma''_\xi\gamma''_\eta}_{\gamma_\xi\gamma_\eta} \end{aligned} \tag{5.4.24}$$

Now, in these equations, the elements of the R-matrices are just numbers. So they commute with everything. Therefore we can rewrite the left-hand side of the above equation to

$$\begin{aligned} R^{-1}_{\xi\eta}(\lambda-\mu)^{\gamma''_\xi\gamma''_\eta}_{\gamma_\xi\gamma_\eta}R_{\xi\eta}(\lambda-\mu)^{\gamma_\xi\gamma_\eta}_{\gamma'_\xi\gamma'_\eta}\mathscr{T}_\xi(\lambda)^{\gamma'_\xi}_{\gamma''_\xi}\mathscr{T}_\eta(\mu)^{\gamma'_\eta}_{\gamma''_\eta} &= \delta^{\gamma''_\xi}_{\gamma'_\xi}\delta^{\gamma''_\eta}_{\gamma'_\eta}\mathscr{T}_\xi(\lambda)^{\gamma'_\xi}_{\gamma''_\xi}\mathscr{T}_\eta(\mu)^{\gamma'_\eta}_{\gamma''_\eta} \\ &= \mathscr{T}_\xi(\lambda)^{\gamma''_\xi}_{\gamma''_\xi}\mathscr{T}_\eta(\mu)^{\gamma''_\eta}_{\gamma''_\eta} \\ &= T(\lambda)T(\mu). \end{aligned} \tag{5.4.25}$$

For the right-hand side of Equation (5.4.24), we find that it is equal to

$$\begin{aligned} \mathscr{T}_\eta(\mu)^{\gamma_\xi}_{\gamma'_\eta}\mathscr{T}_\xi(\lambda)^{\gamma_\eta}_{\gamma'_\xi}\delta^{\gamma'_\eta}_{\gamma_\xi}\delta^{\gamma'_\xi}_{\gamma_\eta} &= \mathscr{T}_\eta(\mu)^{\gamma_\xi}_{\gamma_\xi}\mathscr{T}_\xi(\lambda)^{\gamma_\eta}_{\gamma_\eta} \\ &= T(\mu)T(\lambda). \end{aligned} \tag{5.4.26}$$

Therefore we find that $T(\lambda)T(\mu) = T(\mu)(T\lambda)$ for all $\lambda, \mu \in \mathbb{C}$, which proves Equation (5.4.19).

Chapter 6

The Quantum Inverse-Scattering Method

In this chapter we will derive a solution for the Yang-Baxter Equation. Then, using this solution, we will construct a Hamiltonian from the transfer matrix via the so-called quantum inverse-scattering method. From the commutation property of the transfer matrices it then follows that we obtain infinitely many operators that commute with this Hamiltonian. The physical interpretation of this is that we find a system with infinitely many quantities that are conserved in time. The Hamiltonian we find will be that of the one-dimensional XYZ Heisenberg model.

6.1 A Solution to the Yang-Baxter Equation

In this section we will derive a solution for the Yang-Baxter equation in the simplest non-trivial case, which is when particles can take 2 colors. So we have $l = 2$. Remember that we assumed that in every site on our chain we have exactly one particle, so then every site has two possible states. This is for example the case when the particles on our chain are electrons. Electrons have the internal variable spin, which can be either up or down. We begin the derivation following Section 5.4 of Šamaj and Bajnok [16]. However, from Equation (6.1.10) until the end of the section we follow Section 10.4 of Baxter [19].

The solution of the Yang-Baxter equation will be a scattering matrix. We will look for a solution of the form

$$S(\lambda) = \sum_{j=0}^{3} w_j(\lambda) \ \sigma^j \otimes \sigma^j. \tag{6.1.1}$$

Here $\sigma^0 = I$, and the matrices $\sigma^1 = \sigma^x$, $\sigma^2 = \sigma^y$ and $\sigma^3 = \sigma^z$ are the Pauli matrices. Our use of the Pauli matrices comes from the fact that together with $\sigma^0 = I$ the Pauli matrices form a basis for 2×2 hermitian matrices. The coefficients $w_j(\lambda)$ are for now unknown functions. By calculating the tensor products of the Pauli matrices with themselves, we can write Equation (6.1.1) as

$$S(\lambda) = \begin{pmatrix} a(\lambda) & 0 & 0 & d(\lambda) \\ 0 & b(\lambda) & c(\lambda) & 0 \\ 0 & c(\lambda) & b(\lambda) & 0 \\ d(\lambda) & 0 & 0 & a(\lambda) \end{pmatrix}, \tag{6.1.2}$$

where

$$\begin{aligned} a(\lambda) &= w_0(\lambda) + w_3(\lambda) \\ b(\lambda) &= w_0(\lambda) - w_3(\lambda) \\ c(\lambda) &= w_1(\lambda) + w_2(\lambda) \\ d(\lambda) &= w_1(\lambda) - w_2(\lambda). \end{aligned} \tag{6.1.3}$$

The 4 × 4 permutation operator $\mathscr{P}$ is given by

$$\mathscr{P} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{6.1.4}$$

Therefore we find that the R-matrix $R(\lambda) = \mathscr{P}S(\lambda)$ is given by

$$R(\lambda) = \begin{pmatrix} a(\lambda) & 0 & 0 & d(\lambda) \\ 0 & c(\lambda) & b(\lambda) & 0 \\ 0 & b(\lambda) & c(\lambda) & 0 \\ d(\lambda) & 0 & 0 & a(\lambda) \end{pmatrix} \tag{6.1.5}$$

From Equation (6.1.1) and the definition of the Lax operator, we find that the Lax operator $L_{\xi n}$ is given by

$$L_{\xi n} = \sum_{j=0}^{3} w_j(\lambda)\ \sigma^j \otimes \boldsymbol{\sigma}_n^j. \tag{6.1.6}$$

Here we write $\boldsymbol{\sigma}_n^j$ for the operator acting on the chain of N particles, that acts on the n'th site as σ^j and as the identity operator on the rest of the chain. So in the above equation, the σ^j on the left-hand side of the ⊗ symbol acts on the auxiliary space ξ. We can write $L_{\xi n}$ as a matrix in ξ-space by writing out this σ^j as a matrix. We get

$$L_{\xi n} = \begin{pmatrix} w_0(\lambda)\boldsymbol{\sigma}_n^0 + w_3(\lambda)\boldsymbol{\sigma}_n^z & w_1(\lambda)\boldsymbol{\sigma}_n^x - iw_2(\lambda)\boldsymbol{\sigma}_n^y \\ w_1(\lambda)\boldsymbol{\sigma}_n^x + iw_2(\lambda)\boldsymbol{\sigma}_n^y & w_0(\lambda)\boldsymbol{\sigma}_n^0 - w_3(\lambda)\boldsymbol{\sigma}_n^z \end{pmatrix}. \tag{6.1.7}$$

For the Lax operator $L_{\eta n}$ we can of course do exactly the same thing. Now, with these explicit expressions for the R-matrix and the Lax operators, we can fill them in into the YBE (5.4.7) and start looking for a solution. In the end we want to find expressions for the functions $a(\lambda), b(\lambda), c(\lambda)$ and $d(\lambda)$, such that the scattering matrix in Equation (6.1.2) is a solution to the YBE.

When we fill in the above expressions for R, $L_{\xi n}$ and $L_{\eta n}$ into the YBE (5.4.7), then it can be shown that the YBE is satisfied if the equation

$$w_m w'_l w''_j - w_l w'_m w''_k + w_k w'_j w''_l - w_j w'_k w''_m = 0 \tag{6.1.8}$$

holds for any permutation (j, k, l, m) of $(0, 1, 2, 3)$ [16]. Here we write

$$w_j = w_j(\lambda), \quad w'_j = w_j(\mu), \quad w''_j = w_j(\lambda - \mu). \tag{6.1.9}$$

There are 24 possible permutations of $(0, 1, 2, 3)$, so Equation (6.1.8) consists of 24 equations. However, suppose that we have one of those equations for a certain permutation (j, k, l, m) of $(0, 1, 2, 3)$. Then it is not hard to see that for the permutations (k, j, m, l), (l, m, j, k) and (m, l, k, j) we get the exact same equation. Therefore we see that Equation (6.1.8) consists of at most 6 independent equations. Now, similar to Equation (6.1.9) we write

$$a = a(\lambda), \quad a' = a(\mu), \quad a'' = a(\lambda - \mu) \tag{6.1.10}$$

and similarly for b, c and d. Then it turns out that the 6 independent equations from (6.1.8) are equivalent to [19]

$$\begin{aligned} ac'a'' + da'd'' - bc'b'' - ca'c'' &= 0 \\ ab'c'' + dd'b'' - ba'c'' - cc'b'' &= 0 \\ cb'a'' + bd'd'' - ca'b'' - bc'c'' &= 0 \\ ad'b'' + db'c'' - bd'a'' - cb'd'' &= 0 \\ aa'd'' + dc'a'' - bb'd'' - cd'a'' &= 0 \\ da'a'' + ac'd'' - db'b'' - ad'c'' &= 0. \end{aligned} \tag{6.1.11}$$

These equations are linear homogeneous equations in a'', b'', c'' and d''. Therefore, if we take four of these equations, we can write them as one matrix equation $Av = 0$, where $v = (a'', b'', c'', d'')^T$. Then in order to get a non-trivial solution ($v \neq 0$), we must have $\det A = 0$. If we take the first, third, fourth and sixth equations from (6.1.11), then the determinant of the corresponding matrix is given by [19]

$$(cda'b' - abc'd')\left((a^2 - b^2)(c'^2 - d'^2) + (c^2 - d^2)(a'^2 - b'^2)\right). \tag{6.1.12}$$

This determinant is the product of two factors, so it is zero if and only if at least one of the two factors is zero. Our aim is to find one solution of the YBE, and not necessarily all the solutions. Therefore we can just require that the first factor in the above determinant must be 0. Then we get

$$\frac{cd}{ab} = \frac{c'd'}{a'b'}. \tag{6.1.13}$$

Using this relation, we can solve the system of the first, third, fourth and sixth equation for a'', b'', c'' and d''. The solutions we obtain are [19]

$$\begin{aligned} a'' &= a(cc' - dd')(b^2c'^2 - c^2a'^2)/c \\ b'' &= b(dc' - cd')(a^2c'^2 - d^2a'^2)/d \\ c'' &= c(bb' - aa')(a^2c'^2 - d^2a'^2)/a \\ d'' &= d(ab' - ba')(b^2c'^2 - c^2a'^2)/b. \end{aligned} \tag{6.1.14}$$

Now, we obtained these solutions using the information from the first, third, fourth and sixth equations from (6.1.11). If we fill them in into the second or fifth equation, one can derive the relation [19]

$$\frac{a^2 + b^2 - c^2 - d^2}{ab} = \frac{a'^2 + b'^2 - c'^2 - d'^2}{a'b'}. \tag{6.1.15}$$

Now, we define

$$\begin{aligned} \Delta &= \frac{a^2 + b^2 - c^2 - d^2}{2(ab + cd)} \\ \Gamma &= \frac{ab - cd}{ab + cd}, \end{aligned} \tag{6.1.16}$$

and similarly we define Δ' and Γ' by replacing a, b, c, d by a', b', c', d'. Then the relations (6.1.13) and (6.1.15) are equivalent to

$$\Gamma = \Gamma', \qquad \Delta = \Delta'. \tag{6.1.17}$$

Indeed, if we let

$$\gamma = \frac{1 - \Gamma}{1 + \Gamma} = \frac{cd}{ab}, \tag{6.1.18}$$

then we see that $\Gamma = \Gamma'$ implies (6.1.13). Conversely, we see that

$$\gamma + 1 = \frac{2}{1+\Gamma}, \tag{6.1.19}$$

and therefore

$$\Gamma = \frac{2}{\gamma+1} - 1 = \frac{2}{\frac{ac}{bc}+1} - 1. \tag{6.1.20}$$

Hence (6.1.13) implies $\Gamma = \Gamma'$, so these statements are equivalent. For Δ we find

$$\begin{aligned} 1+\Delta &= \frac{a^2+b^2-c^2-d^2+2ab+2cd}{2(ab+cd)} \\ &= \frac{\frac{a^2+b^2-c^2-d^2}{ab}+2+2\frac{cd}{ab}}{2+\frac{cd}{ab}} \end{aligned} \tag{6.1.21}$$

So the relations (6.1.13) and (6.1.15) together imply that indeed $\Delta = \Delta'$. If we assume $\Gamma = \Gamma'$, then we know that (6.1.13) holds. Thus, in that case it follows from Equation (6.1.21) that also (6.1.15) must hold. Therefore we see that indeed the relations (6.1.13) and (6.1.15) together are equivalent to Equation (6.1.17).

So what we have found up till now is that if we can find a, b, c, d and a', b', c', d' such that Equation (6.1.17) holds, then we have found a solution to the YBE. Remember from Equation (6.1.10) that a, b, c, d are all functions of a spectral parameter, and the difference between a, a' and a'' is just the value of their spectral parameter. So if we can find four functions $a(\lambda), b(\lambda), c(\lambda), d(\lambda)$ such that the corresponding Δ and Γ are independent of λ, then the scattering matrix $S(\lambda)$ in Equation (6.1.2) is a solution to the YBE. We will do this by parameterising a, b, c and d using the Jacobi elliptic functions.

We can combine the two equations in (6.1.16) to get an equation without the variable d as follows

$$\begin{aligned} 2\Delta(1+\gamma)ab &= 2\Delta\left(\frac{ab}{ab}+\frac{cd}{ab}\right)ab \\ &= 2\Delta(ab+cd) \\ &= a^2+b^2-c^2-d^2 \\ &= a^2+b^2-c^2-a^2b^2\gamma^2c^{-2}. \end{aligned} \tag{6.1.22}$$

If we divide this equation by c^2, we obtain

$$2\Delta(1+\gamma)\frac{b}{c}\frac{a}{c} = \frac{a^2}{c^2}+\frac{b^2}{c^2}-1-\gamma^2\frac{b^2}{c^2}\frac{a^2}{c^2} \tag{6.1.23}$$

This equation is a quadratic equation in a/c and in b/c. Suppose that b/c is given, then we can write it as a quadratic equation in a/c in the standard form

$$\left(1-\gamma^2\frac{b^2}{c^2}\right)\frac{a^2}{c^2} - 2\Delta(1+\gamma)\frac{b}{c}\frac{a}{c} + \frac{b^2}{c^2} - 1 = 0. \tag{6.1.24}$$

In order to find a solution for a/c, we calculate the discriminant of the above equation. This discriminant is given by

$$4\Delta^2(1+\gamma)^2\frac{b^2}{c^2} - 4\left(1-\gamma^2\frac{b^2}{c^2}\right)\left(\frac{b^2}{c^2}-1\right). \tag{6.1.25}$$

The above expression can be rewritten as [19]

$$4\left(1-y^2\frac{b^2}{c^2}\right)\left(1-k^2y^2\frac{b^2}{c^2}\right), \tag{6.1.26}$$

where we introduced the variables k, y that are uniquely defined by [19]

$$\begin{aligned} k^2y^4 &= \gamma^2 \\ (1+k^2)y^2 &= 1+\gamma^2-\Delta^2(1+\gamma)^2. \end{aligned} \tag{6.1.27}$$

Now, we can parameterise b/c as a function of a new variable u, so that the discriminant in Equation (6.1.26) takes a simpler form. We say

$$\frac{b}{c} = y^{-1}\operatorname{sn} u, \tag{6.1.28}$$

where we take sn to be the Jacobi elliptic sine function with modulus k. Remember from Theorem 4.5.3 that

$$\operatorname{sn}^2 u + \operatorname{cn}^2 u = 1, \qquad k^2\operatorname{sn}^2 u + \operatorname{dn}^2 u = 1. \tag{6.1.29}$$

Using this, we find that the discriminant in Equation (6.1.26) becomes

$$4\operatorname{cn}^2 u\operatorname{dn}^2 u \tag{6.1.30}$$

Now, we can use this discriminant and Equation (6.1.28) to write down a solution of Equation (6.1.24) in terms of elliptic functions. We get

$$\begin{aligned} \frac{a}{c} &= \frac{2\Delta(1+\gamma)y^{-1}\operatorname{sn} u + 2\operatorname{cn} u\operatorname{dn} u}{2\left(1-\gamma^2y^{-2}\operatorname{sn}^2 u\right)} \\ &= \frac{\Delta(1+\gamma)y\operatorname{sn} u + y^2\operatorname{cn} u\operatorname{dn} u}{y^2-\gamma^2\operatorname{sn}^2 u} \end{aligned} \tag{6.1.31}$$

In order to simplify this expression we define η by

$$k\operatorname{sn}\eta = -\frac{\gamma}{y}. \tag{6.1.32}$$

Note that η only depends on k, γ and y, which in turn only depend on Δ and Γ. Now, if you work out the algebra using equations (6.1.18) and (6.1.27), one can show that [19]

$$y = \operatorname{sn}\eta, \qquad \gamma = -k\operatorname{sn}^2\eta \tag{6.1.33}$$

and [19]

$$\begin{aligned} \Gamma &= \frac{1+k\operatorname{sn}^2\eta}{1-k\operatorname{sn}^2\eta} \\ \Delta &= -\frac{\operatorname{cn}\eta\operatorname{dn}\eta}{1-k\operatorname{sn}^2\eta}. \end{aligned} \tag{6.1.34}$$

So we see that indeed Δ and Γ are independent of the variable u. Therefore, if we can write a, b, c, d as functions of u, we are done.

Using the addition formula for sn we gave in Theorem 4.5.6, we can rewrite Equation (6.1.31) as

$$\frac{a}{c} = \frac{\operatorname{sn}(\eta-u)}{\operatorname{sn}\eta}. \tag{6.1.35}$$

Then using equations (6.1.28), (6.1.33) and (6.1.35), we can rewrite Equation (6.1.18) as

$$\begin{aligned} -k\operatorname{sn}^2\eta &= \frac{cd}{ab} = \frac{c}{b}\frac{c}{a}\frac{d}{c} \\ &= \frac{\operatorname{sn}\eta}{\operatorname{sn} u}\frac{\operatorname{sn}\eta}{\operatorname{sn}(\eta-u)}\frac{d}{c}. \end{aligned} \tag{6.1.36}$$

Hence, we find

$$\frac{d}{c} = -k \operatorname{sn} u \operatorname{sn}(\eta - u). \tag{6.1.37}$$

Then finally, from equations (6.1.28), (6.1.33), (6.1.35) and (6.1.37) we can derive a parameterisation for a, b, c and d in terms of u, up to a common factor. Because we parameterise them up to a common factor, we can fix $c(u) = \operatorname{sn} \eta$ (so c is independent of u). Then multiplying those equations by c gives us the elliptic parameterisation

$$\begin{aligned} a(u) &= \operatorname{sn}(\eta - u) \\ b(u) &= \operatorname{sn} u \\ c(u) &= \operatorname{sn} \eta \\ d(u) &= -k \operatorname{sn} \eta \operatorname{sn} u \operatorname{sn}(\eta - u), \end{aligned} \tag{6.1.38}$$

up to a common factor. Normalizing the scattering matrix will determine this common factor, but in many problems the normalization is not really important so we can just work with the parameterisations above. So now we have found a solution to the Yang-Baxter equation.

It is important to note that the elliptic solution in Equation (6.1.38) is not the simplest solution of the Yang-Baxter equation. If we let the modulus k for the elliptic functions in (6.1.38) tend to 0, then remember from Equation (4.5.12) that we get

$$\begin{aligned} &\operatorname{sn} \to \sin, \\ &\operatorname{cn} \to \cos, \qquad \text{for } k \to 0. \\ &\operatorname{dn} \to 1, \end{aligned} \tag{6.1.39}$$

That way, if we let $k \to 0$ we obtain from Equation (6.1.38) the trigonometric parameterisation

$$a(u) = \sin(\eta - u), \qquad b(u) = \sin u, \qquad c(u) = \sin \eta, \qquad d(u) = 0. \tag{6.1.40}$$

We see that this solution to the YBE is a special case of the elliptic solution. We can even go further by rescaling the spectral parameter u as $u \to \eta u$. Then, if we let η become very small, the above trigonometric parameterisation becomes the rational parameterisation

$$a(u) = \eta - \eta u, \qquad b(u) = \eta u, \qquad c(u) = \eta, \qquad d(u) = 0. \tag{6.1.41}$$

Here we used that $\sin x \to x$ if $x \to 0$. All these parameterisations are up to a common factor, so we can simplify Equation (6.1.41) a bit more by dividing through η. We obtain

$$a(u) = 1 - u, \qquad b(u) = u, \qquad c(u) = 1, \qquad d(u) = 0. \tag{6.1.42}$$

6.2 The Quantum Inverse-Scattering Method

In this section we will use the elliptic solution to the YBE we found in the previous section to construct the Hamiltonian of a system with infinitely many conserved quantities. This way of constructing a system with infinitely many conservation laws from a solution of the Yang-Baxter equation is called the *quantum inverse-scattering method*. The Hamiltonian we find will be that of the one-dimensional XYZ Heisenberg model. In Subsection 6.2.1 we follow Section 6.2 of Šamaj and Bajnok [16], with the help of handwritten notes on the quantum inverse-scattering method from Dirk Schuricht [20].

6.2.1 Constructing a Hamiltonian from the Transfer Matrix

The elliptic solution for the Yang-Baxter equation we found in the previous section was given by

$$a(u) = \operatorname{sn}(\eta - u), \quad b(u) = \operatorname{sn} u, \quad c(u) = \operatorname{sn}\eta, \quad d(u) = -k \operatorname{sn}\eta \operatorname{sn} u \operatorname{sn}(\eta - u). \tag{6.2.1}$$

Here sn is the Jacobi sine function with modulus k. We leave out the common normalization constant because it turns out to be not important [16]. We can find the scattering matrix $S(u)$ corresponding to these elements using Equation (6.1.2). We have $\operatorname{sn} 0 = 0$, because sn is an even function. Therefore, we see that $S(u = 0)$ is given by

$$S(0) = \operatorname{sn}\eta \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \operatorname{sn}\eta \mathscr{P}. \tag{6.2.2}$$

So in coordinates we have

$$S^{\sigma_1\sigma_2}_{\sigma'_1\sigma'_2}(0) = \operatorname{sn}\eta \; \delta^{\sigma_1}_{\sigma'_2}\delta^{\sigma_2}_{\sigma'_1}. \tag{6.2.3}$$

Now, we consider the transfer matrix. In the previous chapter we defined the transfer matrix $T(u)$ as

$$T(u)^{\sigma_1\ldots\sigma_N}_{\sigma'_1\ldots\sigma'_N} = S^{\sigma_1\gamma_1}_{\sigma'_1\gamma_2}(u) S^{\sigma_2\gamma_2}_{\sigma'_2\gamma_3}(u) \cdots S^{\sigma_N\gamma_N}_{\sigma'_N\gamma_1}(u). \tag{6.2.4}$$

Remember that we use the summation convention, so here we sum over all the γ_i for $1 \leq i \leq N$. Using Equation (6.2.3), we see that

$$\begin{aligned} T(0)^{\sigma_1\ldots\sigma_N}_{\sigma'_1\ldots\sigma_N} &= (\operatorname{sn}\eta)^N \delta^{\sigma_1}_{\gamma_2}\delta^{\gamma_1}_{\sigma'_1}\delta^{\sigma_2}_{\gamma_3}\delta^{\gamma_2}_{\sigma'_2}\cdots\delta^{\sigma_N}_{\gamma_1}\delta^{\gamma_N}_{\sigma'_N} \\ &= (\operatorname{sn}\eta)^N \delta^{\sigma_1}_{\sigma'_2}\delta^{\sigma_2}_{\sigma'_3}\cdots\delta^{\sigma_N}_{\sigma'_1}. \end{aligned} \tag{6.2.5}$$

If we ignore the constant $(\operatorname{sn}\eta)^N$, then we see that $T(0)$ acts in some sense like a cyclic permutation. If we let $T(0)$ act on a vector with N coordinates that correspond to sites on the chain, then what $T(0)$ does is shift all the sites one place in the positive direction (and scale with the constant $(\operatorname{sn}\eta)^N$). The N'th site is send to the first, which corresponds to the term $\delta^{\sigma_N}_{\sigma'_1}$, so we have periodic boundary conditions. Therefore, as long as $\operatorname{sn}\eta \neq 0$, the operator $T(0)$ clearly has an inverse operator, which just shifts the sites back. This inverse is then given by

$$T^{-1}(0)^{\sigma_1\ldots\sigma_N}_{\sigma'_1\ldots\sigma'_N} = (\operatorname{sn}\eta)^{-N}\delta^{\sigma_1}_{\sigma'_N}\delta^{\sigma_2}_{\sigma'_1}\cdots\delta^{\sigma_N}_{\sigma'_{N-1}}. \tag{6.2.6}$$

We can calculate the derivative of the transfer matrix evaluated at $u = 0$ as follows

$$\begin{aligned} \left[\frac{\mathrm{d}}{\mathrm{d}u}T(u)\right]^{\sigma_1\ldots\sigma_N}_{\sigma'_1\ldots\sigma'_N}\Bigg|_{u=0} &= \sum_{n=1}^{N} S^{\sigma_1\gamma_1}_{\sigma'_1\gamma_2}(0)\cdots S^{\sigma_{n-1}\gamma_{n-1}}_{\sigma'_{n-1}\gamma_n}(0)\;\frac{\mathrm{d}}{\mathrm{d}u}S^{\sigma_n\gamma_n}_{\sigma'_n\gamma_{n+1}}(u)\Big|_{u=0} \\ &\quad S^{\sigma_{n+1}\gamma_{n+1}}_{\sigma'_{n+1}\gamma_{n+2}}(0)\cdots S^{\sigma_N\gamma_N}_{\sigma'_N\gamma_1} \\ &= (\operatorname{sn}\eta)^{N-1}\sum_{n=1}^{N}\delta^{\sigma_1}_{\gamma_2}\delta^{\gamma_1}_{\sigma'_1}\cdots\delta^{\sigma_{n-1}}_{\gamma_n}\delta^{\gamma_{n-1}}_{\sigma'_{n-1}}\;\frac{\mathrm{d}}{\mathrm{d}u}S^{\sigma_n\gamma_n}_{\sigma'_n\gamma_{n+1}}(u)\Big|_{u=0} \\ &\quad \delta^{\sigma_{n+1}}_{\gamma_{n+2}}\delta^{\gamma_{n+1}}_{\sigma'_{n+1}}\cdots\delta^{\sigma_N}_{\gamma_1}\delta^{\gamma_N}_{\sigma'_N} \\ &= (\operatorname{sn}\eta)^{N-1}\sum_{n=1}^{N}\delta^{\sigma_1}_{\sigma'_2}\cdots\delta^{\sigma_{n-2}}_{\sigma'_{n-1}}\;\frac{\mathrm{d}}{\mathrm{d}u}S^{\sigma_n\sigma_{n-1}}_{\sigma'_n\sigma'_{n+1}}(u)\Big|_{u=0}\delta^{\sigma^{n+1}}_{\sigma'_{n+2}}\cdots\delta^{\sigma_N}_{\sigma'_1}. \end{aligned} \tag{6.2.7}$$

Here we use the periodic boundary condition $N+1=1$. Note that we have to write the summation sign explicitly in the above equations, because otherwise from the rules of the summation convention it would not be clear that we sum over n. However, in the upper two equations on the right-hand side we also sum over the γ_i of course, using the summation convention.

Now, consider the product of $T(0)$ and $\frac{\mathrm{d}}{\mathrm{d}u}T(u)\big|_{u=0}$. Remember that $T(0)$ acts like a cyclic permutation on the chain of N sites, while also multiplying with a constant. The direction of this cyclic permutation depends on whether you act with $T(0)$ from the left or from the right. Now, because the chain is completely translationally invariant (using periodic boundary conditions), it follows that it does not matter whether you act with $T(0)$ on $\frac{\mathrm{d}}{\mathrm{d}u}T(u)\big|_{u=0}$ from the left or from the right. Hence, these operators commute.

Now suppose that we have a function f of the form

$$f(x)=\sum_{n=0}^{\infty}a_n x^n, \tag{6.2.8}$$

and a matrix $A(u)$ such that $A(0)$ and $\frac{\mathrm{d}}{\mathrm{d}u}A(u)\big|_{u=0}$ commute with each other. We have

$$f'(x)=\sum_{n=1}^{\infty}n a_n x^{n-1}, \tag{6.2.9}$$

so using that $A(0)$ and $\frac{\mathrm{d}}{\mathrm{d}u}A(u)\big|_{u=0}$ commute we find

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}u}f(A(u))\bigg|_{u=0} &= \left(\sum_{n=1}^{\infty}n a_n A^{n-1}(0)\right)\frac{\mathrm{d}}{\mathrm{d}u}A(u)\bigg|_{u=0}\\ &= f'(A(0))\,\frac{\mathrm{d}}{\mathrm{d}u}A(u)\bigg|_{u=0}\end{aligned} \tag{6.2.10}$$

Here in the first step we were able to take the term $\frac{\mathrm{d}}{\mathrm{d}u}A(u)\big|_{u=0}$ out of the sum because this term commutes with $A(0)$. Otherwise this would not be possible. Note that we could just as well have taken the term out of the sum on the left-hand side.

We know that the natural logarithm can be written as

$$\ln x=\sum_{n=1}^{\infty}\frac{(-1)^{n-1}(x-1)^n}{n}, \tag{6.2.11}$$

and $\frac{\mathrm{d}}{\mathrm{d}x}\ln x = x^{-1}$. So, because $T(0)$ and $\frac{\mathrm{d}}{\mathrm{d}u}T(u)\big|_{u=0}$ commute, we find

$$\frac{\mathrm{d}}{\mathrm{d}u}\ln T(u)\bigg|_{u=0}=T^{-1}(0)\,\frac{\mathrm{d}}{\mathrm{d}u}T(u)\bigg|_{u=0}. \tag{6.2.12}$$

Then, using equations (6.2.6) and (6.2.7) we find

$$\begin{aligned}\left[\frac{\mathrm{d}}{\mathrm{d}u}\ln T(u)\right]^{\sigma_1\ldots\sigma_N}_{\sigma'_1\ldots\sigma'_N}\Bigg|_{u=0} &= \frac{1}{\operatorname{sn}\eta}\delta^{\sigma_1}_{\sigma''_N}\delta^{\sigma_2}_{\sigma''_1}\cdots\delta^{\sigma_N}_{\sigma''_{N-1}}\sum_{n=1}^{N}\delta^{\sigma''_1}_{\sigma'_2}\cdots\delta^{\sigma''_{n-2}}_{\sigma'_{n-1}}\,\frac{\mathrm{d}}{\mathrm{d}u}S^{\sigma''_n\sigma''_{n-1}}_{\sigma'_n\sigma'_{n+1}}(u)\bigg|_{u=0}\\ &\quad \delta^{\sigma''_{n+1}}_{\sigma'_{n+2}}\cdots\delta^{\sigma''_N}_{\sigma'_1}\\ &= \frac{1}{\operatorname{sn}\eta}\sum_{n=1}^{N}\delta^{\sigma_1}_{\sigma'_1}\cdots\delta^{\sigma_{n-1}}_{\sigma'_{n-1}}\,\frac{\mathrm{d}}{\mathrm{d}u}S^{\sigma_{n+1}\sigma_n}_{\sigma'_n\sigma'_{n+1}}(u)\bigg|_{u=0}\delta^{\sigma_{n+2}}_{\sigma'_{n+2}}\cdots\delta^{\sigma_N}_{\sigma'_N}.\end{aligned} \tag{6.2.13}$$

Now, the ansatz form of the scattering matrix we assumed in Equation (6.1.1) is equivalent to

$$S^{\sigma_1\sigma_2}_{\sigma_1'\sigma_2'}(u) = \sum_{j=0}^{3} p_j(u)(\sigma^j)^{\sigma_1}_{\sigma_2'}(\sigma^j)^{\sigma_2}_{\sigma_1'}, \tag{6.2.14}$$

where

$$\begin{aligned} p_0 &= \frac{1}{2}(w_0 + w_1 + w_2 + w_3) = \frac{1}{2}(a + c) \\ p_1 &= \frac{1}{2}(w_0 + w_1 - w_2 - w_3) = \frac{1}{2}(b + d) \\ p_2 &= \frac{1}{2}(w_0 - w_1 + w_2 - w_3) = \frac{1}{2}(b - d) \\ p_3 &= \frac{1}{2}(w_0 - w_1 - w_2 + w_3) = \frac{1}{2}(a - c). \end{aligned} \tag{6.2.15}$$

This can be seen by just writing it out. For example, using the above equations we find

$$S^{11}_{11}(u) = p_0(u) + p_3(u) = a(u), \tag{6.2.16}$$

which is exactly what we wanted according to Equation (6.1.2). Using Equation (6.2.14) we can rewrite (6.2.13) to

$$\left.\left[\frac{\mathrm{d}}{\mathrm{d}u}\ln T(u)\right]^{\sigma_1\ldots\sigma_N}_{\sigma_1'\ldots\sigma_N'}\right|_{u=0} = \frac{1}{\operatorname{sn}\eta}\sum_{n=1}^{N}\delta^{\sigma_1}_{\sigma_1'}\cdots\delta^{\sigma_{n-1}}_{\sigma_{n-1}'}\left(\sum_{j=0}^{3}\left.\frac{\partial p_j}{\partial u}\right|_{u=0}(\sigma^j)^{\sigma_n}_{\sigma_n'}(\sigma^j)^{\sigma_{n+1}}_{\sigma_{n+1}'}\right) \\ \delta^{\sigma_{n+2}}_{\sigma_{n+2}'}\cdots\delta^{\sigma_N}_{\sigma_N'}. \tag{6.2.17}$$

Now, remember that in Equation (6.1.6) we wrote $\boldsymbol{\sigma}^j_n$ for the operator on the chain of N particles that acts on the n'th site as σ^j, and trivially on the rest of the chain. We can write down $\boldsymbol{\sigma}^j_n$ explicitly as

$$\left(\boldsymbol{\sigma}^j_n\right)^{\sigma_1\ldots\sigma_n\ldots\sigma_N}_{\sigma_1'\ldots\sigma_n'\ldots\sigma_N'} = \delta^{\sigma_1}_{\sigma_1'}\cdots(\sigma^j)^{\sigma_n}_{\sigma_n'}\cdots\delta^{\sigma_N}_{\sigma_N'}. \tag{6.2.18}$$

Therefore, recalling that $\sigma^0 = I$, $\sigma^1 = \sigma^x$, $\sigma^2 = \sigma^y$ and $\sigma^3 = \sigma^z$, we see that Equation (6.2.17) can be rewritten as

$$\operatorname{sn}\eta\left.\left[\frac{\mathrm{d}}{\mathrm{d}u}\ln T(u)\right]^{\sigma_1\ldots\sigma_N}_{\sigma_1'\ldots\sigma_N'}\right|_{u=0} = \frac{1}{2}\sum_{n=1}^{N}\left(J_x\boldsymbol{\sigma}^x_n\boldsymbol{\sigma}^x_{n+1} + J_y\boldsymbol{\sigma}^y_n\boldsymbol{\sigma}^y_{n+1} + J_z\boldsymbol{\sigma}^z_n\boldsymbol{\sigma}^z_{n+1}\right) + \frac{N}{2}J_0\boldsymbol{I}, \tag{6.2.19}$$

where

$$J_x = \left.\frac{\partial p_1}{\partial u}\right|_{u=0}, \qquad J_y = \left.\frac{\partial p_2}{\partial u}\right|_{u=0}, \qquad J_z = \left.\frac{\partial p_3}{\partial u}\right|_{u=0}, \qquad J_0 = \left.\frac{\partial p_0}{\partial u}\right|_{u=0}. \tag{6.2.20}$$

We also distinguish between the identity operators I and $\boldsymbol{I}$, where the first acts on one site of the chain and the second on the whole chain. Using the parameterisation in Equation (6.2.1) and Theorems 4.5.4 and 4.5.5 for the Jacobi elliptic functions, we find

$$\begin{aligned} J_x &= \left.\frac{\partial}{\partial u}(\operatorname{sn}u - k\operatorname{sn}\eta\operatorname{sn}u\operatorname{sn}(\eta - u))\right|_{u=0} \\ &= \left.\operatorname{cn}u\operatorname{dn}u(1 - k\operatorname{sn}^2\eta)\right|_{u=0} \\ &= 1 - k\operatorname{sn}^2\eta. \end{aligned} \tag{6.2.21}$$

Similarly, we obtain

$$J_y = 1 + k\,\mathrm{sn}^2\,\eta \quad \text{and} \quad J_z = J_0 = -\,\mathrm{cn}\,\eta\,\mathrm{dn}\,\eta. \tag{6.2.22}$$

The right-hand side of Equation (6.2.19) is the Hamiltonian of the so-called *one-dimensional XYZ Heisenberg model.* The Pauli matrices must be thought of as the components of the spin operator. Therefore we call this Hamiltonian a spin Hamiltonian. So Equation (6.2.19) tells us how we can construct this Hamiltonian from the transfer matrix corresponding to the elliptic solution of the Yang-Baxter equation.

6.2.2 The One-Dimensional XYZ Heisenberg Model

The one-dimensional XYZ Heisenberg model describes a system of N identical particles with spin on a one-dimensional lattice. It is called the XYZ model because the spin of each particle has three components x, y and z. The particles have a nearest-neighbor interaction with the periodic boundary condition $N+1 \equiv 1$. The XYZ Heisenberg model is a relatively simple Hamiltonian that can be used to describe the magnetism of solids, because magnetism partially originates from the relative spin alignment of the particles in a solid [6].

To get a feeling for this model, we consider the case that the particles on the chain are spin-$\frac{1}{2}$ particles. Remember from quantum mechanics that the Hilbert space of a single spin-$\frac{1}{2}$ particle is $\mathbb{C}^2$. Therefore the Hilbert space of the chain is the tensor product of N copies of $\mathbb{C}^2$:

$$\underbrace{\mathbb{C}^2 \otimes \cdots \otimes \mathbb{C}^2}_{N \text{ times}}. \tag{6.2.23}$$

The spin operator $\boldsymbol{\sigma}_n^j$ acts on this Hilbert space as follows: it acts as σ^j on the n'th copy of $\mathbb{C}^2$, corresponding to the n'th particle in the chain, and it acts trivially on the rest of the chain.

From Equations (6.2.19) and (6.2.22) we know that the Hamiltonian of the one-dimensional XYZ Heisenberg model is given by

$$H = \frac{1}{2}\sum_{n=1}^{N} \left(J_x \boldsymbol{\sigma}_n^x \boldsymbol{\sigma}_{n+1}^x + J_y \boldsymbol{\sigma}_n^y \boldsymbol{\sigma}_{n+1}^y + J_z \boldsymbol{\sigma}_n^z \boldsymbol{\sigma}_{n+1}^z\right) + \frac{N}{2} J_z \boldsymbol{I}. \tag{6.2.24}$$

The constants J_x, J_y and J_z are called the *coupling constants*. Suppose that J_y and J_z are both zero. Note that with our elliptic parameterisations in Equations (6.2.21) and (6.2.22) this is actually not possible, because $0 \leq k < 1$ and therefore $J_y \geq 1$. However, considering impossible cases like this will help us to understand how the model works. Then the Hamiltonian takes the form

$$H_x = \frac{1}{2}\sum_{n=1}^{N} J_x \boldsymbol{\sigma}_n^x \boldsymbol{\sigma}_{n+1}^x. \tag{6.2.25}$$

The Pauli matrix σ^x has two eigenvalues, $+1$ and -1. The corresponding eigenvectors are

$$\psi_+ = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}, \qquad \psi_- = \frac{1}{\sqrt{2}}\begin{pmatrix}-1\\1\end{pmatrix} \tag{6.2.26}$$

The states ψ_+ and ψ_- are also called spin up and spin down in the x-direction, respectively. We see that if the particles on the sites n and $n+1$ are both in the same eigenstate of σ^x, then the term $\boldsymbol{\sigma}_n^x \boldsymbol{\sigma}_{n+1}^x$ gives a 1 when we act with H_1 on this state of the chain. Conversely, if the particles are both in an eigenstate of σ^x but their

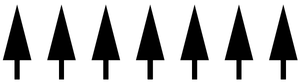

Figure 6.1: Ferromagnetic alignment of spins.

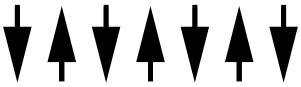

Figure 6.2: Antiferromagnetic alignment of spins.

states are not the same, the the term $\boldsymbol{\sigma}_n^x\boldsymbol{\sigma}_{n+1}^x$ gives -1. Therefore, if J_x is negative, there are two ground states of H_x with energy $NJ_x/2$, given by

$$\underbrace{\psi_+ \otimes \cdots \otimes \psi_+}_{N \text{ times}} \quad \text{and} \quad \underbrace{\psi_- \otimes \cdots \otimes \psi_-}_{N \text{ times}}. \tag{6.2.27}$$

In these states, the spins of the particles are all aligned in the same direction. This is called parallel alignment or *ferromagnetic ordering* of the spins, see Figure 6.1. On the other hand, if J_x is positive, the two ground states of H_x are

$$\underbrace{\psi_+ \otimes \psi_- \otimes \psi_+ \otimes \cdots \otimes \psi_\pm}_{N \text{ times}} \quad \text{and} \quad \underbrace{\psi_- \otimes \psi_+ \otimes \psi_- \otimes \cdots \otimes \psi_\mp}_{N \text{ times}}, \tag{6.2.28}$$

with energy $-NJ_x/2$. In these states, the spins of two neighboring particles point in the opposite direction. This is called antiparallel alignment or *antiferromagnetic ordering* of the spins, see also Figure 6.2.

Now, suppose that instead of $J_y = J_z = 0$ we have $J_x = J_y = 0$. Then we get the Hamiltonian

$$H_z = \frac{1}{2}\sum_{n=1}^{N} J_z \boldsymbol{\sigma}_n^z \boldsymbol{\sigma}_{n+1}^z. \tag{6.2.29}$$

The Pauli matrix σ^z also has the two eigenvalues $+1$ and -1, but the eigenvectors are different from those of σ^x. They are given by

$$\phi_+ = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \phi_- = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{6.2.30}$$

Then, completely analogous to Equations (6.2.27) and (6.2.28), we find that if $J_z < 0$ then the ground states of H_z are given by

$$\underbrace{\phi_+ \otimes \cdots \otimes \phi_+}_{N \text{ times}} \quad \text{and} \quad \underbrace{\phi_- \otimes \cdots \otimes \phi_-}_{N \text{ times}}, \tag{6.2.31}$$

and if $J > 0$ they are given by

$$\underbrace{\phi_+ \otimes \phi_- \otimes \phi_+ \otimes \cdots \otimes \phi_\pm}_{N \text{ times}} \quad \text{and} \quad \underbrace{\phi_- \otimes \phi_+ \otimes \phi_- \otimes \cdots \otimes \phi_\mp}_{N \text{ times}}. \tag{6.2.32}$$

In both cases, the energy of the ground states is $-N|J_z|/2$.

So in both of the cases we considered above where two of the coupling constants are 0, it is easy to determine the ground state energy of the system. However, if we let two or even three of the coupling constants to be non-zero, then this becomes much harder. For example, suppose that only $J_y = 0$. Then we get the Hamiltonian

$$H_{xz} = \frac{1}{2}\sum_{n=1}^{N} J_x \boldsymbol{\sigma}_n^x \boldsymbol{\sigma}_{n+1}^x + J_z \boldsymbol{\sigma}_n^z \boldsymbol{\sigma}_{n+1}^z. \tag{6.2.33}$$

Determining the ground state energy for this Hamiltonian is much more complicated than in the previous cases, because the J_x and the J_z term cannot be minimized independently. This is because the x- and z-components of the spin cannot be changed independently from each other. It turns out that the ground state of H_{xz} is not one of the states we found where either the J_x term or the J_z term is minimized, but something in between. We will show this by approximating the ground state energy up to first order using degenerate perturbation theory. We follow section 7.2.1 of Griffiths and Schroeter [21].

We write

$$H_{xz} = H^0 + H', \tag{6.2.34}$$

where $H^0 = H_x$ and $H' = H_z$. We denote the ground state energy of H_{xz} by E_0. The first order approximation of E_0 is given by

$$E_0^0 + E_0^1, \tag{6.2.35}$$

where E_0^0 is the ground state energy of H^0 and E_0^1 is the first order correction to the energy. We have $H^0 = H_x$, so $E_0^0 = NJ_x/2$. We write χ_+ and χ_- for the ground states of H^0 given in Equation (6.2.27). Then the first order correction to the ground state energy of H_{xz} is given by [21]

$$E_0^1 = \frac{1}{2}\left(W_{++} + W_{--} - \sqrt{(W_{++} - W_{--})^2 + 4|W_{+-}|^2}\right), \tag{6.2.36}$$

where

$$W_{ij} = \langle \chi_i \mid H' \mid \chi_j \rangle, \quad \text{for } i, j \in \{+, -\}. \tag{6.2.37}$$

Now, using the fact that

$$\langle \psi_+ \mid \sigma^z \mid \psi_+ \rangle = \langle \psi_- \mid \sigma^z \mid \psi_- \rangle = 0, \tag{6.2.38}$$

we see that

$$W_{++} = W_{--} = 0. \tag{6.2.39}$$

Then Equation (6.2.36) simplifies to

$$E_0^1 = -|W_{+-}|. \tag{6.2.40}$$

A short calculation gives

$$\langle \psi_+ \mid \sigma^z \mid \psi_- \rangle = -1. \tag{6.2.41}$$

Therefore, we get

$$W_{+-} = \frac{NJ_z}{2}. \tag{6.2.42}$$

Hence, using the fact that $J_z < 0$, we find that the ground state energy of H_{xz} up to first order is given by

$$E_0^0 + E_0^1 = \frac{NJ_x}{2} + \frac{NJ_z}{2} = \frac{N}{2}(J_x + J_z). \tag{6.2.43}$$

So we see that the ground state energy of H_{xz} is not equal to $NJ_x/2$ or $NJ_z/2$. Therefore, we find indeed that the ground state of H_{xz} is not equal to one of the ground states of H_x or H_z.

The main message from the previous calculations is that deriving properties of the one-dimensional XYZ Heisenberg model is a complicated task. However, we can derive a lot of information about the XYZ model from the fact that we obtained it via the Quantum Inverse Scattering Method. From Equation (6.2.19) we see that the Hamiltonian is equal to the derivative of the logarithm of the transfer matrix $T(u)$, evaluated at $u = 0$. Now, in the previous chapter we showed that transfer matrices for different values of the spectral parameter all commute with each other. Using that fact, it can be shown that we obtain also infinitely operators that commute with $\left.\left[\frac{\mathrm{d}}{\mathrm{d}u} \ln T(u)\right]^{\sigma_1 \ldots \sigma_N}_{\sigma'_1 \ldots \sigma'_N}\right|_{u=0}$ [16]. So from Equation (6.2.19) we see that we actually have infinitely many operators that commute with the Hamiltonian of the one-dimensional XYZ Heisenberg model. These linear operators correspond to measurable quantities of the system, and the fact that they commute with the Hamiltonian implies that they are conserved in time. Hence, we obtain infinitely many conservation laws for the one-dimensional XYZ Heisenberg model.

Further, we also mentioned that this commuting set of transfer matrices can be simultaneously diagonalized. Using this, we can find the eigenvectors and -values of the transfer matrix, from which we then can derive the eigenvalues of the XYZ Hamiltonian [16]. Therefore, we can calculate the entire energy spectrum of the one-dimensional XYZ Heisenberg model. If we know all the energy levels of a system, we can calculate its partition function. As you probably remember from statistical physics, once you know the partition function of a system, you basically know everything. For example we can use it to calculate the free energy of the system.

Conclusion

Finally, let us recapitulate what we have done in this thesis. After the introduction to projective geometry in the first chapter, we defined elliptic curves as a special kind of projective curves. Subsequently, we defined a group structure on the set of points of an elliptic curve in Weierstrass normal form. In the third chapter we proved a number theoretical result due to Gauss, and we saw that we can use this result to determine the group structure of a specific elliptic curve in $\mathbb{P}^2(\mathbb{F}_{19})$. Then, in Chapter 4 we looked at elliptic functions. Specifically, we introduced the Weierstrass $\wp$-function, and we saw how we can use this function to parameterise elliptic curves in Weierstrass normal form. Furthermore, we introduced elliptic integrals and the Jacobi elliptic functions, and we discussed the relation between elliptic curves, elliptic functions, elliptic integrals and ellipses. In Chapter 5 we derived the Yang-Baxter equation as the integrability condition. Then we constructed the transfer matrix from the scattering matrix, and we showed that the Yang-Baxter equation implies that transfer matrices for different values of the spectral parameter commute with each other. This way, we obtained an infinite set of commuting transfer matrices. In the last chapter we derived a solution for the Yang-Baxter equation using the Jacobi elliptic functions. Lastly, by means of the quantum inverse-scattering method we used this solution to construct the Hamiltonian of the one-dimensional XYZ Heisenberg model from the transfer matrix. From the commutation property of the transfer matrices it then followed that this model has infinitely conservation laws.

Outlook

The theory of elliptic functions is vast, and there are many topics one could dive into after having read this thesis. For example, one could look at Lenstra's elliptic curve factorisation algorithm. This algorithm makes use of elliptic curves to determine the prime factorization of integers. See for example Section IV.4 in Silverman and Tate [1]. Another interesting topic is the group structure of an elliptic curve. In this thesis we showed that the points on an elliptic curve form a group, but we only know what this group looks like for the specific example we gave in Section 3.2. One could for instance look at Mordell's Theorem, which states that the group of points on an elliptic curve is in general finitely generated. This implies that we only need a finite amount of points on the curve, and then we can construct all the other points by adding points to each other.

Also for the cases of the Yang-Baxter equation and the quantum inverse-scattering method, we by no means covered everything there is to be said about them in this thesis. In Section 6.1 we derived a solution for the Yang-Baxter equation, but this was only for the very specific case that the particles can have two possible colors. As a follow-up, one could for example consider the case $l = 3$. Then the solution to the Yang-Baxter equation would be a 9×9 scattering matrix instead of 4×4. Also, in the last chapter we mentioned that the fact that the transfer matrices are simultaneously diagonalizable can be used to determine the complete energy spectrum of the one-dimensional XYZ Heisenberg model, but we did not see how this works exactly. Readers interested in the details could for example look at Sections 6.3 and 16.1 in Šamaj and Bajnok [16].

Appendix A

The Discriminant

In this appendix we will define the determinant of a polynomial, following Section IV.6 of Lang [4]. We start by looking at symmetric polynomials.

A.1 Symmetric Polynomials

Let $P(x_1, \ldots, x_n)$ be a polynomial in n variables with coefficients in some commutative ring R, i.e. $P(x_1, \ldots, x_n) \in R[x_1, \ldots, x_2]$. We say that P is *symmetric* if for every permutation $g \in S_n$ we have

$$P(x_1, \ldots, x_n) = P(x_{g(1)}, \ldots, x_{g(n)}). \tag{A.1.1}$$

So for example the polynomial $ax_1 + bx_2$ is symmetric if and only if $a = b$, because then switching x_1 and x_2 does not change the polynomial.

Now, consider the polynomial $P(y, x_1, \ldots, x_n) \in R[y, x_1, \ldots, x_n]$ given by

$$P(y, x_1, \ldots, x_n) = (y - x_1) \cdots (y - x_n). \tag{A.1.2}$$

Then we can write

$$P(y, x_1, \ldots, x_n) = y^n - s_1 y^{n-1} + \ldots + (-1)^n s_n, \tag{A.1.3}$$

where each s_i is an element of $R[x_1, \ldots, x_n]$, so they are polynomials in the variables $x_1, \ldots, x_n$. For example we have that $s_1 = x_1 + \ldots + x_n$ and $s_n = x_1 \cdots x_n$. The polynomials s_i we call the *elementary symmetric polynomials* of $x_1, \ldots, x_n$. By looking at Equation (A.1.2), we see that P stays the same if we permute the $x_1, \ldots, x_n$. Hence the polynomials s_i are indeed symmetric, as their name suggests. The following theorem is an important result for symmetric polynomials, and the reason why we call the s_i the *elementary* symmetric polynomials.

Theorem A.1.1. *Let $P(x_1, \ldots, x_n) \in R[x_1, \ldots, x_n]$ be a symmetric polynomial. Then there exists a unique polynomial $Q(t_1, \ldots, t_n) \in R[t_1, \ldots, t_n]$ such that*

$$P(x_1, \ldots, x_n) = Q(s_1, \ldots, s_n), \tag{A.1.4}$$

where the s_i are the elementary symmetric polynomials of $x_1, \ldots, x_n$.

Proof. For a proof of this theorem, see for example the proof of Theorem 6.1 in Chapter IV of Lang [4]. Note that the uniqueness of Q is not proven there, but this follows immediately from the fact that the elementary symmetric polynomials are algebraically independent, which is shown directly after Theorem 6.1. □

As an example look at $P(x_1, x_2) = x_1^2 + x_2^2 \in \mathbb{C}[x_1, x_2]$. If we take Equation (A.1.2) for $n = 2$ and rewrite it as Equation (A.1.3), we see that there are two elementary symmetric polynomials of x_1, x_2, given by $s_1 = x_1 + x_2$ and $s_2 = x_1 x_2$. Now, we see that we can write the polynomial $P(x_1, x_2)$ as

$$P(x_1, x_2) = x_1^2 + x_2^2 = (x_1 + x_2)^2 - 2x_1x_2 = s_1(x_1, x_2)^2 - 2s_2(x_1, x_2). \tag{A.1.5}$$

So if we let $Q(t_1, t_2) = t_1^2 - 2t_2 \in R[t_1, t_2]$, then we find indeed that

$$P(x_1, x_2) = Q(s_1(x_1, x_2), s_2(x_1, x_2)). \tag{A.1.6}$$

A.2 The Discriminant

Now, let us consider polynomials $P(x) \in R[x]$ in one variable of the form

$$P(x) = (x - r_1) \cdots (x - r_n), \tag{A.2.1}$$

where $r_1, \ldots, r_n \in R$. We define $\delta \in R$ as

$$\delta = \prod_{i<j} (r_i - r_j). \tag{A.2.2}$$

Then δ is just an element of R, but we can view δ as a polynomial with variables $r_1, \ldots, r_n$. From the above definition it is clear that permuting $r_1, \ldots, r_n$ will only change the sign of δ. Therefore δ^2 is a symmetric polynomial. We call δ^2 the *discriminant* of P, denoted by D_P. Because $D_P(r_1, \ldots, r_n)$ is symmetric, we know from Theorem (A.1.1) that there exists a unique polynomial $D(t_1, \ldots, t_n) \in R[t_1, \ldots, t_n]$ such that $D_P(r_1, \ldots, r_n) = D(s_1, \ldots, s_n)$. Here the s_i are the elementary polynomials of $r_1, \ldots, r_n$. So, for the discriminant D_P of a polynomial P given by Equation (A.2.1) and letting $s_1, \ldots, s_n$ be the elementary symmetric polynomials of $r_1, \ldots, r_n$, we have

$$D_P = D(s_1, \ldots, s_n) = \prod_{i<j} (r_i - r_j)^2. \tag{A.2.3}$$

Note that the discriminant of a polynomial P is non-zero if and only if the roots of P are all distinct.

Suppose that we have a polynomial $P(x)$ of degree d in one variable with complex coëfficients, so $P(x) \in \mathbb{C}[x]$. We write $P(x) = a_d x^d + \ldots + a_0$. We assume $P(x)$ to be monic, so $a_d = 1$. Then by the Fundamental Theorem of Algebra we know that there exist complex numbers $r_1, \ldots, r_d$ such that

$$P(x) = (x - r_1) \ldots (x - r_d), \tag{A.2.4}$$

see also Corollary 3.1.5 in [7]. Then the discriminant of $P(x)$ is equal to

$$D_P = D(s_1, \ldots, s_d) = \prod_{i<j} (r_i - r_j)^2, \tag{A.2.5}$$

where the s_i are the elementary symmetric polynomials of $r_1, \ldots, r_d$.

For example in the case $d = 2$, we have

$$P(x) = x^2 + bx + c = (x - r_1)(x - r_2). \tag{A.2.6}$$

Then we find $b = -(r_1 + r_2) = -s_1(r_1, r_2)$ and $c = r_1 r_2 = s_2(r_1, r_2)$. The discriminant of P is then given by

$$D_P = (r_1 - r_2)^2 = (r_1 + r_2)^2 - 4r_1r_2 = b^2 - 4c, \tag{A.2.7}$$

which is the expression we would expect from what we learned in high school.

In the case $d = 3$, we write $P(x) = x^3 + ax^2 + bx + c$. Then it can be checked that [1]

$$D_P = -4a^3c + a^2b^2 + 18abc - 4b^3 - 27c^2. \tag{A.2.8}$$

Bibliography

[1] Joseph H. Silverman and John Tate. *Rational Points on Elliptic Curves.* Springer-Verlag, 1992.

[2] Robert A. Adams and Christopher Essex. *Calculus.* Pearson, eighth edition, 2014.

[3] Joseph H. Silverman. *The Arithmetic of Elliptic Curves.* Springer-Verlag, second edition, 2009.

[4] Serge Lang. *Algebra.* Springer-Verlag, revised third edition, 2002.

[5] Frits Beukers. *Getaltheorie; Een inleiding. (Dutch)* [*Number Theory; An Introduction*]. Epsilon Uitgaven, 2008.

[6] Mark A. Armstrong. *Groups and Symmetry.* Springer-Verlag, 1988.

[7] Frits Beukers. *Rings and Galois Theory.* Department of Mathematics, Utrecht University, 2018. Lecture notes for the course *Rings and Galois Theory.*

[8] Serge Lang. *Complex Analysis.* Springer-Verlag, fourth edition, 1999.

[9] Eberhard Freitag and Rolf Busam. *Complex Analysis.* Springer-Verlag, 2005.

[10] Neil Koblitz. *Introduction to Elliptic Curves and Modular Forms.* Springer-Verlag, second edition, 1993.

[11] Lawrence C. Washington. *Elliptic Curves; Number Theory and Cryptography.* Taylor & Francis Group, second edition, 2008.

[12] Jean-Pierre Serre. *A course in arithmetic.* Springer-Verlag, 1973.

[13] Adrian Rice and Ezra Brown. Why ellipses are not elliptic curves. *Mathematics Magazine*, 85(3):163–176, 2012. `https://personal.math.vt.edu//brown/doc/ellipses_not_ecs.pdf`.

[14] Viktor Prasolov and Yuri Solovyev. *Elliptic Functions and Elliptic Integrals.* American Mathematical Society, 1997.

[15] Edmund T. Whittaker and George N. Watson. *A Course of Modern Analysis.* Cambridge University Press, fourth edition, 1927.

[16] Ladislav Šamaj and Zoltán Bajnok. *Introduction to the Statistical Physics of Integrable Many-body Systems.* Cambridge University Press, 2013.

[17] Thierry Giamarchi. *Quantum Physics in One Dimension.* Oxford University Press, 2003.

[18] Steven Roman. *Advanced Linear Algebra.* Springer-Verlag, third edition, 2008.

[19] Rodney J. Baxter. *Exactly Solved Models in Statistical Mechanics.* Dover Publications, 2007.

[20] Dirk Schuricht. Notes on QISM. Private communication.

[21] David J. Griffiths and Darrel F. Schroeter. *Introduction to Quantum Mechanics.* Cambridge University Press, third edition, 2018.

Index