EMS Series of Lectures in Mathematics

Torsten Ekedahl

One Semester of Elliptic Curves

European Mathematical Society

EMS Series of Lectures in Mathematics
Edited by Andrew Ranicki (University of Edinburgh, U.K.)

EMS Series of Lectures in Mathematics is a book series aimed at students, professional mathematicians and scientists. It publishes polished notes arising from seminars or lecture series in all fields of pure and applied mathematics, including the reissue of classic texts of continuing interest. The individual volumes are intended to give a rapid and accessible introduction into their particular subject, guiding the audience to topics of current research and the more advanced and specialized literature.

Previously published in this series:
Katrin Wehrheim, *Uhlenbeck Compactness*

Torsten Ekedahl

One Semester of Elliptic Curves

European Mathematical Society

Author:

Torsten Ekedahl
Department of Mathematics
Stockholm University
SE-106 91 Stockholm
SWEDEN

2000 Mathematics Subject Classification: 14H52; 14K25.

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data are available in the Internet at http://dnb.ddb.de.

ISBN 3-03719-015-9

Contact address:

European Mathematical Society Publishing House
Seminar for Applied Mathematics
ETH-Zentrum FLI C4
CH-8092 Zürich
Switzerland

Phone: +41 (0)1 632 34 36
Email: info@ems-ph.org
Homepage: www.ems-ph.org

Typeset using the author's $\TeX$ files: I. Zimmermann, Freiburg
Printed in Germany

9 8 7 6 5 4 3 2 1

Preface

This book was developed from the notes for a one semester introductory course on elliptic curves. The theory of elliptic curves has a long history (even the name "elliptic curves" was chosen for historical reasons, the actual relation with ellipses is tenuous at best). Today it appears in many different contexts, some of which are far removed from the origins of the subject. Its long history, as well as the current interest in elliptic curves, makes it very difficult to find material to present in an introduction that has not already been covered in other sources. These notes make no pretension of doing that. For a one semester course one must however make a careful and limited choice. The choice made for this course was dictated to a large extent by the heterogeneous nature of the audience. This audience consisted on the one hand of mathematics students, at various stages of their education, but generally with a number theoretic interest, and on the other hand of computer science students interested in the cryptographic aspects of elliptic curves. This led me from the beginning to assume as little background knowledge as possible while at the same time limiting the background material given in the course. For instance, I define the notion of analytic and meromorphic functions and prove the maximal principle, but I only mention without proof some basic results on uniform convergence of holomorphic functions. The basic idea was to give the audience an appreciation of how difficult the applicable results are, while not necessarily proving all of them. Similarly, no knowledge of algebraic geometry was assumed.

No matter the reason for being interested in elliptic curves, knowledge of basic analytic methods is indispensable. Consequently we start by discussing some of the analytic aspects of elliptic curves. We roughly follow the historical development, starting with elliptic integrals and Euler's addition theorem, then turning to elliptic functions where the power of the analytic methods is demonstrated when we prove the fundamental formulas for the Weierstrass $\wp$-function. Preparing for a more algebraic approach we give a short introduction to the projective plane. Then we discuss equivalences between elliptic curves and lattices, leading up to the j-function and its incarnation as both a rational expression in the coefficients of the equation defining the elliptic curve and as an analytic function in the lattice (also defining the elliptic curve).

We then turn to some more specialised subjects reflecting the interests of the audience described above. To prepare for a student presentation of Schoof's algorithm for counting the number of points on an elliptic curve, we went through the basic properties of the division polynomials. We then finished with some additional number theoretical aspects; curves with complex multiplication and the use of modular forms for proving Jacobi's formula for the number of representations of a positive integer as a sum of four squares.

At several points we need to do some reasonably complicated numerical or algebraic calculations. It can be argued that one should avoid such a situation as it obscures what is going on. In most cases these situations are indeed avoidable but at the cost of having to develop more theory. (Note however that in a few cases the calculations seem unavoidable, even using more high powered machinery. The calculations of Subsection 11.1.1 is one such example.) Within the confines of a one semester course this is not however possible. I have chosen the middle road of leaving the heavy calculations to the Mathematica computer algebra system. A Mathematica notebook containing these calculations is available http://www.math.su.se/~teke/undervisning/Elliptisk.nb and the computations can be performed by the reader using Mathematica, but the notebook can also be viewed using the freely available MathReader program (which can be downloaded from http://www.wolfram.com).

Contents

1

Elliptic integrals

As is well known, integrals involving only rational functions in a single variable can always be explicitly integrated: One makes a partial fraction decomposition of the function, all summands of which are directly integrable except those of the form $a(x-r)^{-1}$; they give rise instead to logarithmic terms. However, one quite often wants to integrate more complicated functions. One class of such examples is when the integrand is a rational function in the variable x and in $\sqrt{1-x^2}$. There is a well-known trigonometric substitution which allows for the solution of all such integrals. We shall however indicate an alternative (and one could say more systematic) approach. What we want is first to think of the square root of $1-x^2$ as defining a function $y(x)$ which is characterised (up to a sign) by $y^2 = 1-x^2$. Now we forget that we made y a function of x and instead think of this as a subset of $\mathbb{R}^2$ which more symmetrically is defined by the equation $x^2 + y^2 = 1$, i.e., the circle. Then we find a nice *parametrisation* of the circle in the following way: We pick a fixed point on the circle (we choose to pick $(-1,0)$) and then draw a line with slope k through the fixed point. This line intersects the circle in two points and the fixed point is one of them, so that we get a unique other intersection point. We map the slope k to that point (cf. Fig. 1). Conversely, for a point on the curve we map it to the slope of the line through the fixed point and this point. There are some minor problems with this; on the one hand a line can also have slope equal to ∞, on the other hand the variable point can be equal to the fixed point and then there is not a unique line through the two points. We shall ignore those problems (at least for the moment).

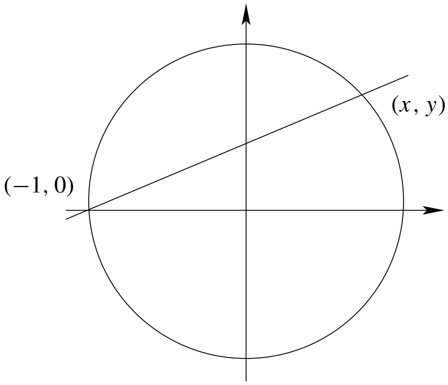

Figure 1. Parametrising the circle.

It is easy to find actual formulas for the map from k to the point. We have the system of equations

$$x^2 + y^2 = 1,$$
$$y = k(x + 1).$$

If we use the second to eliminate y from the first we get

$$(k^2 + 1)x^2 + 2k^2x + k^2 - 1 = 0.$$

This is of degree 2 in x but we know one root already, namely $x = -1$, and if we divide out by it we get a first degree equation which we can solve. There is however a (slightly) faster way. We know that if α and β are the roots of $ax^2 + bx + c$, then $\alpha + \beta = -b/a$ (Exercise). This gives $x - 1 = -2k^2/(k^2 + 1)$, i.e., $x = (1 - k^2)/(1 + k^2)$. Then we substitute this value into the equation for the line and we get $y = 2k/(1 + k^2)$ and hence the map is

$$k \mapsto \left(\frac{1 - k^2}{1 + k^2}, \frac{2k}{1 + k^2}\right).$$

The inverse map is given by the slope of the line which gives

$$(x, y) \mapsto \frac{y}{x + 1}.$$

What happens for infinite slope and the point $(-1, 0)$ is clarified in the following exercise.

Exercise 1. (i) Show that when $k \to \pm\infty$, then (x, y) tends to $(-1, 0)$.
(ii) Show that when $(x, y) \to (-1, 0)$ along the circle, then $k \to \pm\infty$.

This means that if we have an integrand which is a rational function in x and $\sqrt{1 - x^2}$, then we may make the coordinate transformation $x = (1 - k^2)/(1 + k^2)$ and then the integrand is transformed into a rational function in k which we know may be integrated.

Exercise 2. Show that if α and β are the roots of $ax^2 + bx + c$, then $\alpha + \beta = -b/a$.

We may perform this integration not just over a real interval but also over any path in the complex domain. For that we need to recall a few facts about path integration. When integrating over a path (in the plane say), the integrand is an expression of the form $f(x, y)\,dx + g(x, y)\,dy$ and it can be integrated along a parametrised curve γ which by definition is a continuous and piecewise differentiable[1] function $C\colon [a, b] \to \mathbb{R}^2$. The integral of $f(x, y)\,dx + g(x, y)\,dy$ along γ is then defined to be

$$\oint_C f(x, y)\,dx + g(x, y)\,dy := \int_a^b \big(f(x(t), y(t))x'(t) + g(x(t), y(t))y'(t)\big)\,dt,$$

[1] I.e., $[a, b]$ is covered by a finite number of intervals such that C is differentiable on each of them.

where $C(t) = (x(t), y(t))$ when C is differentiable; in general one defines it as the sum over intervals over which it is differentiable. We want to have a formalism that gives various relations between expressions of the form $f(x, y)\,dx + g(x, y)\,dy$. All these relations should have the property that if two expressions are equal, according to the rules we shall introduce, then their integrals along any curve will be equal. To take care of coordinate changes, but also in more general situations, we introduce and interpret expressions of the form $f(x, y)\,dg$, where g is a rational function in x and y, and define

$$\oint_C f(x, y)\,dg := \int_a^b f(x(t), y(t)) \frac{d(g(x(t), y(t)))}{dt}\,dt.$$

It is clear that we may expand $d(g(x(t), y(t)))/dt$ using the chain rule; it is equally clear that the resulting expression can be expressed in terms of partial derivatives of g with respect to x and y and $x'(t)$ and $y'(t)$. The relations for this type of expression that we now introduce model this expansion:

- The expressions are finite sums $\sum_i f_i\,dg_i$ where f_i and g_i are rational functions in x and y. These expressions are called *1-forms in x and y*.
- Multiplication of such expressions by rational functions in x and y fulfil the obvious distributivity and associativity relations.[2]
- $d\lambda = 0$ for $\lambda \in \mathbb{C}$.
- $d(f + g) = df + dg$ for rational functions f and g.
- $d(fg) = f\,dg + g\,df$ for rational functions f and g.

The following set of exercises verifies that these relations are sound as well as complete (in the sense that by using them one can reduce any expression to the form $f\,dx + g\,dy$).

Exercise 3. Show that if two forms can be made equal by using a sequence of these relations, then their integrals along any curve are equal.

Exercise 4. (i) Show that $d(1/f) = -1/f^2\,df$ for any rational function f using only these relations.

(ii) Show that $df(x, y) = f'_x(x, y)\,dx + f'_y(x, y)\,dy$ using only these relations.

(iii) Show that using only these relations, any 1-form can be brought into the form $f(x, y)\,dx + g(x, y)\,dy$.

The following exercise gives an analogue of the chain rule.

Exercise 5. Show (using only these relations) that if $f(x, y)$, $g(x, y)$ and $h(x, y)$ are rational functions, then $dh(f(x, y), g(x, y)) = h'_x(f, g)\,df + h'_y(f, g)\,dg$.

[2] Technically they form a vector space over the field of rational functions.

Having set up our formalism we only need one more observation to be able to continue: There is no reason to assume that C is real valued; we may just as well assume that it is complex valued instead, i.e., a map $C\colon [a,b] \to \mathbb{C}^2$.

If we go back to our original problem we may now try to integrate a rational function in x and $\sqrt{1-x^2}$ along a curve in the complex plane. There is however a problem in that it is less clear what we are to mean by the square root. In the real case there is not much of a problem; only square roots of positive reals are acceptable and then one may choose the positive square root. In the complex case there really is no overall consistent choice. One solution is to make the square root part of the choice. In this way, in order to deal with for instance $\sqrt{1-x^2}$, our path should make a choice not just of a complex number x but also of a square root of $1-x^2$. In this way the path becomes exactly a map $C\colon [a,b] \to \mathbb{C}^2$ such that $x^2(t) + y^2(t) = 1$, where $C(t) = (x(t), y(t))$.

Example 1. Consider the integral $\oint_C dx/\sqrt{1-x^2}$ which we should write as $\oint_C dx/y$. Making the coordinate transformation $(x, y) = ((1-k^2)/(1+k^2), 2k/(1+k^2))$ we get

$$dx = \left(\frac{1-k^2}{1+k^2}\right)' dk = -\frac{4k}{(1+k^2)^2}$$

and

$$\oint_C dx/y = \oint_{C'} \frac{-\frac{4k}{(1+k^2)^2}}{\frac{2k}{1+k^2}}\, dk = \oint_{C'} -\frac{2}{1+k^2}\, dk,$$

where $C'(t) = y(t)/(x(t)+1)$.

Exercise 6. Integrate $\oint_{C'} -\frac{2}{1+k^2}\, dk$ directly and compare the result with the direct integration of $\oint_C dx/\sqrt{1-x^2}$.

This need for choosing a square root is not very serious as the following result attests.

Proposition 1.1. *Let $C\colon [a,b] \to \mathbb{C}^*$ be a continuous map and fix a z_0 such that $z_0^2 = C(a)$. Then there is a unique map $D\colon [a,b] \to \mathbb{C}^*$ such that $D(a) = z_0$ and $D^2(t) = C(t)$ for $t \in [a,b]$.*

Proof. The unicity is made clear by considering two candidates D_1 and D_2 and then putting $E(t) := D_1(t)/D_2(t)$. Then we have that $E^2(t) = D_1^2(t)/D_2^2(t) = C(t)/C(t) = 1$ and thus that $E(t) = \pm 1$. As E also is continuous and is defined on an interval it must be constant, but by assumption $E(a) = 1$. For the existence we let r be the supremum of all $s \in [a,b]$ for which such a D exists. By uniqueness there is such a map on $[a,r]$ and if $r = b$ we are finished. If not we choose an open interval I around r lying in $[a,b]$ for which $|D(t) - D(r)| < |D(r)|$ for all t in I. Now there is a continuous square root $z \mapsto \sqrt{z}$ in the disc $|z-1| < 1$ for which $\sqrt{1} = 1$. It can for instance be defined by the Taylor series for $\sqrt{1+t}$ which converges when

$|t| < 1$. We then define a function D_1 on I by $D_1(t) = D(r)\sqrt{C(t)/C(r)}$. Again by uniqueness it coincides with D on the left of r, and hence we get an extension of D beyond r. □

Note that for this result it is important that the curve does not pass through the origin; if it does then uniqueness is not necessarily true as can be seen, for instance, for $C^2(t) = 1 - t$ on the interval [0, 2] and $C(0) = 1$ (cf. Fig. 2). Note also that we could consider the more general integrand $f(x, y)\,dx + g(x, y)\,dy$. However we are working under the constraint that $x^2 + y^2 = 1$, and hence for any curve $t \mapsto (x(t), y(t))$ we have $x^2(t) + y^2(t) = 1$. If we differentiate we get $2x(t)x'(t) + 2y(t)y'(t) = 0$ which in terms of the differentials means that $2x\,dx + 2y\,dy = 0$. This implies that we can eliminate dy if we want. This computation is part of a more general principle. If we constrain the curve $C\colon [a, b] \to \mathbb{C}^2$ to lie in a subset of $\mathbb{C}^2$, we often get relations between differentials. As we just saw this is true if we constrain the curve to lie in the subset $x^2 + y^2 = 1$; we may then introduce the relation $2x\,dx + 2y\,dy = 0$. The direct recipe for obtaining this relation is that one applies d to the relation: $x^2 + y^2 = 1$ gives $d(x^2 + y^2) = d1$, but $d1 = 0$ and $d(x^2 + y^2) = 2x\,dx + 2y\,dy$. In general one can say that if the constraint is a polynomial relation $p(x, y) = 0$, then one introduces the relation $dp = 0$.

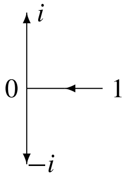

Figure 2. Square root bifurcation.

In particular the same kind of reduction that was used above works for the square root of any quadratic polynomial in x. We shall now continue with what was historically the first example that went beyond square roots of quadratic polynomials, namely square roots of degree 3 and 4 polynomials. In fact the first such integral was the *lemniscate integral*

$$\int_0^r \frac{dt}{\sqrt{1 - t^4}},$$

which appears when one tries to compute the arc length of the *lemniscate*.[3] We shall however consider only degree 3 polynomials (one may actually freely pass between degree 3 and 4 polynomials). We may of course always make coordinate changes in x and thus by "completing the cube" we may assume that the polynomial has the form $x^3 + ax + b$ for some a and b. If this polynomial is divisible by a non-constant square, we may move such a square out of the square root sign. This leaves us with

[3] Those points for which the product of the distances to two fixed points is fixed.

the square root of a linear function which is very simple to handle. We therefore assume that $x^3 + ax + b$ does not have any multiple roots. The problem would now be that of an integration involving rational functions in x and $\sqrt{x^3 + ax + b}$. By the discussion above we should rather think of our polynomial as a rational function in x and y where x and y are subject to the condition $y^2 = x^3 + ax + b$. We shall, for reasons that will not be immediately apparent, restrict ourselves to the integrand dx/y. By way of preliminary motivation for this choice we can point to the analogy with the function $x^2 + y^2 = 1$ where integrating dx/y gives us $\arcsin(x)$. In any case the integral of dx/y along a curve is called an *elliptic integral*.[4] If we try to use the way we handled the case of $x^2 + y^2 = 1$ as a model, we immediately run into the problem that a line will generally intersect the set of solutions to $y^2 = x^3 + ax + b$ in three points. Hence we do not get a parametrisation by looking at all the lines through a fixed point. Using the same idea we do however get something, namely the *addition law* for elliptic integrals, as we shall see now. The idea is to start with two points in $\mathbb{C}^2$ both fulfilling the equation $y^2 = x^3 + ax + b$ and then to construct a third by taking the line through these two points and taking the third intersection point of this line with the set of solutions to $y^2 = x^3 + ax + b$. Let us start by deriving a formula for this third intersection point. Let one point have coordinates (r, s) and the other coordinates (u, v). We seek the third intersection point whose coordinates we call (x, y) (cf. Fig. 3).

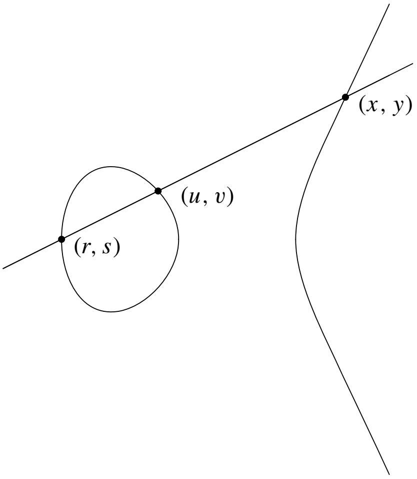

Figure 3. Three points of intersection.

[4] As they appear when one is trying to compute the arc length of an ellipse.

We get the following system of equations for x and y:

$$\begin{aligned} y^2 &= x^3 + ax + b, \\ y - s &= \frac{v-s}{u-r}(x-r). \end{aligned}$$

We use the second equation to eliminate y in the first and get the equation

$$x^3 - \left(\frac{v-s}{u-x}(x-r) + s\right)^2 + \text{lower order terms} = x^3 - \frac{(v-s)^2}{(u-r)^2}x^2 + \text{lower order terms.}$$

Now, if α, β, γ are the roots (counted with multiplicity) of $x^3 - cx^2 + dx - e = 0$, then $\alpha + \beta + \gamma = c$, and thus, as the equation is true for all intersection points, we have

$$u + x + r = \frac{(v-s)^2}{(u-r)^2}.$$

This relation gives rise to a relation among differentials provided we also take into account that we have a relation for the two pairs. That is, our points fulfil the relations

$$\left.\begin{aligned} v^2 &= u^3 + au + b, \\ s^2 &= r^3 + ar + b, \end{aligned}\right\} \tag{1.1}$$

yielding, according to the recipe for constraining relations,

$$\begin{aligned} 2vdv &= (3u^2 + a)du, \\ 2sds &= (3r^2 + a)dr. \end{aligned}$$

If we combine the above relations we obtain

$$\begin{aligned} du + dx + dr &= 2\frac{v-s}{(u-r)^2}dv - 2\frac{(v-s)^2}{(u-r)^3}du + 2\frac{v-s}{(u-r)^2}ds - 2\frac{(v-s)^2}{(u-r)^3}dr \\ &= 2\frac{v-s}{(u-r)^3}\left((u-r)\frac{3u^2+a}{2v} - v + s\right)du \\ &\quad + 2\frac{v-s}{(u-r)^3}\left((u-r)\frac{3r^2+a}{2s} - v + s\right)dr, \end{aligned}$$

which gives

$$\begin{aligned} \frac{dx}{y} &= \frac{1}{y}\left(2\frac{v-s}{(u-r)^3}\left((u-r)\frac{3u^2+a}{2v} - v + s\right) - 1\right)du \\ &\quad + \frac{1}{y}\left(2\frac{v-s}{(u-r)^3}\left((u-r)\frac{3r^2+a}{2s} - v + s\right) - 1\right)dr. \end{aligned}$$

We now want to show that

$$\frac{1}{y}\left(2\frac{v-s}{(u-r)^3}\left((u-r)\frac{3u^2+a}{2v}-v+s\right)-1\right)=-\frac{1}{v}$$

and

$$\frac{1}{y}\left(2\frac{v-s}{(u-r)^3}\left((u-r)\frac{3r^2+a}{2s}-v+s\right)-1\right)=-\frac{1}{s}.$$

The proof of these relations must of course use the relations (1.1) as well as the expansion for y:

$$y=\frac{v-s}{u-r}(x-r)+s=\frac{v-s}{u-r}\left(\left(\frac{v-s}{u-r}\right)^2-u-r-r\right).$$

Verifying the relations above using (1.1) and the expansion for y is now a completely mechanical as well as thoroughly unenlightening process which (in my opinion at least) is best relegated to a computer algebra system.[5] We shall only make the following remark. Using the relations (1.1) we can reduce any rational function in r, s to the form $f(r)+g(r)s$, where f and g are rational functions in r. Similarly, any rational function in r, s, u, v can be reduced to the form $f(r,u)+g(r,u)s+h(r,u)v+k(r,u)sv$. It is this reduction that can be performed mechanically and then all the coefficients f, g, h, and k are identically zero.

In any case admitting these relations gives us the relation

$$\frac{dx}{y}+\frac{du}{v}+\frac{dr}{s}=0, \tag{1.2}$$

which is due to Euler.[6] We shall now use it for the intended purpose, namely to give a relation between integrals. We therefore assume that we have two curves $C\colon [a,b]\to X$ and $D\colon [a,b]\to X$, where $X:=\{(x,y)\in\mathbb{C}^2 \mid y^2=x^3+ax+b\}$. Introducing names for the coordinates of C and D by $C(t)=(r(t),s(t))$ and $D(t)=(u(t),v(t))$, we then define $E\colon [a,b]\to X$ by $E(t)=(x(t),y(t))$, where

$$\begin{aligned} x(t)&=\left(\frac{v(t)-s(t)}{u(t)-r(t)}\right)^2-u(t)-s(t),\\ y(t)&=\frac{v(t)-s(t)}{u(t)-r(t)}(x(t)-r(t))+s(t) \end{aligned}$$

(let us for the moment ignore the possibility that $u(t)$ may be equal to $r(t)$). Then things are arranged exactly so that $\oint_C dx/y=\int_a^b r'(t)/s(t)\,dt$, $\oint_D dx/y=\int_a^b u'(t)/v(t)\,dt$, and $\oint_E dx/y=\int_a^b x'(t)/y(t)\,dt$, where we, somewhat confusingly, have used x and y

[5] A Mathematica notebook that performs these calculations can be obtained via http://www.math.su.se/~teke/undervisning/Elliptisk.nb.

[6] Leonhard Euler, 1707–1783

as the name for the two coordinates in the $\oint$'s, but used different names for them in the expansion of the path integrals as ordinary integrals. It then turns out that Euler's relation gives a relation between these integrals:

Proposition 1.2 ("Addition formula"). *If the paths C, D, and E are defined as above, then we have*

$$\oint_E \frac{dx}{y} + \oint_D \frac{dx}{y} + \oint_C \frac{dx}{y} = 0.$$

Proof. We may assume that C and D are differentiable (and not just piecewise so). Euler's formula (1.2) then has as consequence that

$$\frac{r'(t)}{s(t)} + \frac{u'(t)}{v(t)} + \frac{x'(t)}{y(t)} = 0$$

and integrating this relation immediately gives the desired formula. □

2

Elliptic curves

There are two problems with our proof of the addition formula. The first one is that possibly the denominator of dx/y could become zero, which would seem to make the integral undefined. This can be solved by noticing that we have the relation

$$2y\,dy = (3x^2 + a)\,dx$$

and hence that

$$\frac{dx}{y} = 2\frac{dy}{3x^2 + a},$$

allowing us to replace the integrand dx/y by $2\,dy/(3x^2 + a)$ (possibly only on a part of the interval of integration), which means that the integral is well defined except at points where both $y = 0$ and $3x^2 + a = 0$. However, $y = 0$ implies that $x^3 + ax + b = 0$ and $3x^2 + a$ is the derivative of this polynomial, so that if both quantities were zero, the polynomial would have a multiple root, which has been explicitly excluded.

The other problem seems to be more complicated: We could possibly have that $x(t)$ is undefined because $u(t) = r(t)$. We shall solve this problem by introducing a "point at infinity". It is often useful to introduce such a notion, and for the plane there are in fact several possibilities; one could for instance be interested in keeping track of in which direction one proceeds "to infinity". When we are thinking of the plane as the set of complex numbers, the most common case of interest is however to introduce a single point at infinity. The reason is that the complex number x tends towards infinity exactly when $1/x$ tends towards 0, and there is usually little reason to distinguish between the different directions in which one tends to zero.

We therefore postulate the existence of a point ∞ which is not a complex number and denote by $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ the *extended complex plane*. We then extend the function $x \mapsto 1/x$ to $\overline{\mathbb{C}}$ by defining $1/0 = \infty$ and $1/\infty = 0$. This allows us to transfer all questions involving ∞ to questions involving 0. We say for instance that $x_i \to \infty$ if $1/x_i \to 0$. One way of visualising the extended plane is through *stereographic projection*. Consider the unit sphere (points with distance 1 from the origin) in $\mathbb{R}^3$ and for any point on it different from (0, 0, 1) draw a line through the point and (0, 0, 1) and map the point to the intersection of this line with the xy-plane (cf. Fig. 4). This gives a bijection between the plane and the sphere minus (0, 0, 1), and a point in the plane tends towards ∞ precisely when the corresponding point on the sphere tends towards $(0, 0, 1)$. Hence $\overline{\mathbb{C}}$ can naturally be identified with the 2-dimensional sphere. $\overline{\mathbb{C}}$ is often called the *Riemann sphere*.[1]

[1] Georg Friedrich Bernhard Riemann, 1826–1866

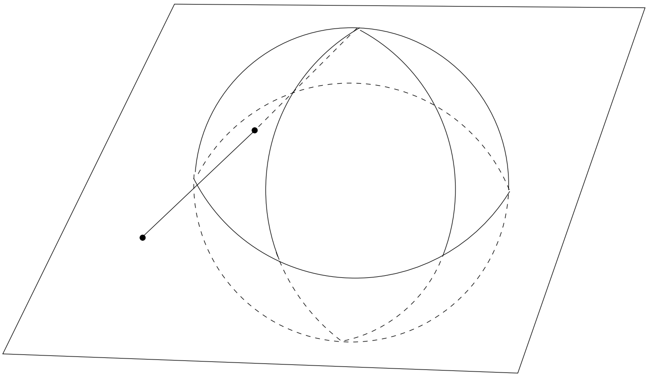

Figure 4. Stereographic projection of the sphere.

Exercise 7. Extend the usual multiplication and addition of $\mathbb{C}$ to *partially defined* operations on $\overline{\mathbb{C}}$ such that if $z \in \mathbb{C}$ and $w \in \mathbb{C} \setminus \{0\}$, then $\infty + z = \infty$, $w\infty = \infty$, $w/\infty = 0$, and $w/0 = \infty$. Let $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ be an invertible 2×2-matrix (with complex entries) and define the *Möbius transformations* by $z \mapsto \frac{az+b}{cz+d}$. Show that with the extended operations the Möbius transformations are well-defined maps from $\overline{\mathbb{C}}$ to itself. Also show that they are invertible and that the inverse and compositions of Möbius transformations are again Möbius transformations.

We shall now apply this philosophy to the equation $y^2 = x^3 + ax + b$ and its set of solutions $X := \{(x, y) \in \mathbb{C}^2 \mid y^2 = x^3 + ax + b\}$. We regard a point (x, y) fulfilling this equation as being mainly determined by x since y takes on at most two values for a given x. We now want to see what happens to the situation when x tends to ∞. According to our recipe we should do this by considering the equation with x replaced by $1/x$ for x close to but not equal to zero. This gives $y^2 = x^{-3} + ax^{-1} + b$ which may be written as $(x^2y)^2 = bx^4 + ax^3 + x$. We may then make the substitution $z = x^2y$ giving the equation $z^2 = bx^4 + ax^3 + x$. This is almost of the same type as the original equation, with the exception that the polynomial in x is (usually) of degree 4. This turns out not to matter very much[2] and more interesting is the fact that the polynomial $bx^4 + ax^3 + x$ has a zero at x so that there is only one possible z for $x = 0$. The situation we have is the following:

We can cover the Riemann sphere by $\mathbb{C} = \overline{\mathbb{C}} \setminus \{\infty\}$ and $\overline{\mathbb{C}} \setminus \{0\}$. The map $x \mapsto 1/x$ maps one to the other and in particular their intersection to itself. This intersection is just $\mathbb{C}^\times := \mathbb{C} \setminus \{0\}$ and the map is of course still $x \mapsto 1/x$. We can describe the situation by saying that we get $\overline{\mathbb{C}}$ by taking two copies of $\mathbb{C}$ and gluing them together along $\mathbb{C}^\times$ in each copy by using the map $x \mapsto 1/x$.

[2] It is, however, not immediately clear what the analogue of Euler's relation should be.

We can do the same for X. We consider on the one hand X and on the other hand $Y := \{(x, z) \in \mathbb{C}^2 \mid z^2 = bx^4 + ax^3 + x\}$. Take $X' \subset X$ defined by the condition that $x \neq 0$ and similarly $Y' \subset Y$, also defined by $x \neq 0$. We then have a bijection $X' \to Y'$ given by $(x, y) \mapsto (1/x, x^2 y)$ and we may glue X and Y together along X' and Y' using this map. Let us denote the result by $\overline{X}$. This is what we shall come to call an *elliptic curve*. Just as $\overline{\mathbb{C}}$ consists of $\mathbb{C}$ and one extra point, the complement of X in $\overline{X}$ consists of a single point. Note that this depends on the fact that we started with a cubic polynomial. Had we started with $y^2 = ax^4 + bx^3 + cx^2 + dx + e$, then the equation on the other part of the curve becomes $z^2 = ex^4 + dx^3 + cx^2 + bx + a$ and if $a \neq 0$ there will be two values of z corresponding to $x = 0$ (on the Y-part).

We may now try to integrate dx/y along a path $C\colon [a, b] \to \overline{X}$ in $\overline{X}$ and not just in X. We first have to handle what it should mean for the path to be continuous and piecewise differentiable. This is however easy. Simply require that $[a, b]$ can be covered with subintervals such that C maps each subinterval into X or Y and the restriction of C to any of them should be continuous and piecewise differentiable as maps into $X, Y \subset \mathbb{C}^2$. What happens then with dx/y in Y? Well, we should express in it $t = 1/x$ and z;

$$\frac{dx}{y} = \frac{d(1/t)}{t^{-2}z} = -\frac{dt}{z}$$

and hence it integrates in the same way (up to the sign) as dx/y does. Hence, there is in fact no problem with one of the paths going through ∞.

2.1 The shape of elliptic curves

When considering a path in $\overline{X}$ it is easiest to think of it as a path C in $\overline{\mathbb{C}}$ together with a continuous choice of square root of $x^3 + ax + b$ at each $x = C(t)$. This description is particularly relevant because of Proposition 1.1 which tells us that as long as we keep away from ∞ and the roots of $x^3 + ax + b$, then the square root along the path is determined by its value at any one point. However it becomes somewhat complicated to keep track of the choice of square root. One may in particular ask what happens for a path that avoids ∞ and the roots of $x^3 + ax + b$ and returns to its starting point. The choice of square root at the end point may be the same or different from that at the starting point. If we could decide which case occurs we can also decide when two paths starting and ending at the same point and starting with the same square root ends with the same square root, since one can simply go forwards by one of them and then return by going backwards with the other. It turns out that quite often the result is largely independent of the curve. To make this precise we start with a definition.

Definition 2.1. Let $U \subseteq \mathbb{C}$ be an open subset. Two paths $C\colon [a, b] \to U$ and $D\colon [a, b] \to U$ with the same end points (i.e., $C(a) = D(a)$ and $C(b) = D(b)$) are *end point homotopic in U* if there is a continuous map $E\colon [a, b] \times [0, 1] \to U$

such that $E(a, s)$ and $E(b, s)$ are both independent of s and $E(t, 0) = C(t)$ and $E(t, 1) = D(t)$ (cf. Fig. 5).

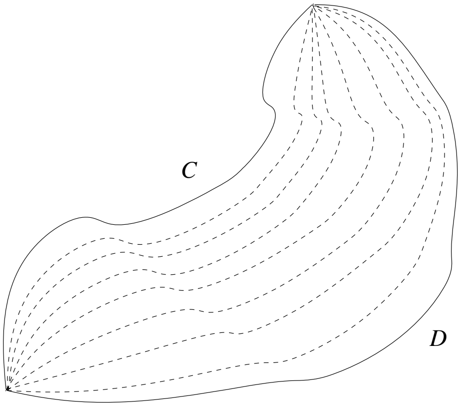

Figure 5. An end point homotopy.

The use of this notion that we shall make at this moment is to show that “square roots depend only on the end point homotopy class of the path”.

Proposition 2.2. *Let $C\colon [a, b] \to \mathbb{C}^\times$ and $D\colon [a, b] \to \mathbb{C}^\times$ be two continuous maps such that $C(a) = D(a)$ and $C^2(t) = D^2(t)$ for all $t \in [a, b]$. If C^2 and D^2 are end point homotopic in $\mathbb{C}^\times$, then $C(b) = D(b)$.*

Proof. We assume that $E'\colon [a, b] \times [0, 1] \to \mathbb{C}^\times$ is an end point homotopy between C^2 and D^2. Suppose we can show that there is a continuous map $E\colon [a, b] \times [0, 1] \to \mathbb{C}^\times$ such that $E^2(s, t) = E'(s, t)$ for all $(s, t) \in [a, b] \times [0, 1]$ and such that $E(a, 0) = C(a)$. Then by the uniqueness of Proposition 1.1 we get first that $E(a, s) = C(a)$, implying that $E(a, 1) = D(a)$ and again by uniqueness $E(t, 0) = C(t)$ and $E(t, 1) = D(t)$. By a final use of uniqueness we have that $E(b, s)$ is constant and hence $C(b) = D(b)$.

The proof that E exists is altogether similar to the existence part of Proposition 1.1. We must however replace taking the supremum by the following argument. A subset T of $[a, b] \times [0, 1]$ will be said to be *ray shaped* if whenever x belongs to T, the closed line segment from $(a, 0)$ to x also lies in T (so that in particular $(a, 0)$ belongs to it). Consider all open ray shaped subsets of $[a, b] \times [0, 1]$ on which an extension exists. The uniqueness of such an extension is proved as in Proposition 1.1, using that the straight line segment from any point in the set to $(a, 0)$ lies, by assumption, in the set, which implies that an extension exists on the union of all such open subsets (which is also open and ray shaped). If this union is not the full rectangle $[a, b] \times [0, 1]$, it has a boundary point and one uses such a boundary point in the same way as the supremum was used in Proposition 1.1. (We can show that one may add a neighbourhood to the line segment from $(a, 0)$ to a boundary point to which the function extends and which is ray shaped.) □

Exercise 8. Fill in the details of this proof.

If we now go back to our current situation, we can look at any path in $\overline{\mathbb{C}}$ which does not pass through the roots of $x^3 + ax + b$ and then we can "transport" the two square roots of $x^3 + ax + b$ along such a path. Notice that this transport gives a bijection between such roots, i.e., if one root y is transported to z, then $-y$ is transported to $-z$. Let us now see what happens when we transport the two roots along a path that moves towards a root z of $x^3 + ax + b$ till it gets to a point z' close to z, then circles (in the positive direction) around z in a circle of small radius r with center at z, and then moves backwards along the same path that we followed to get close to the point (cf. Fig. 6). We now want to show that this closed point will permute the two roots. Transports back and forth to z' are just bijections that are inverses to each other, so what we want to show is that transporting around the circle will permute the roots non-trivially. For this we write $x^3 + ax + b$ as $(x - z)p(x)$, where p is a degree 2 polynomial for which $p(a) \neq 0$, since $x^3 + ax + b$ has no multiple roots. By continuity we can make r so small that $p(x) \neq 0$ when $|x - z| \leq r$. This means that transport of the square root of p along the circle around z of radius r gives the square root we started with, as that circle is end point homotopic to the constant path in the disc around z of radius r and p is $\neq 0$ in that disc. Hence transport of the square root around z of $x^3 + ax + b$ is non-trivial precisely when transport of the square root of $x - z$ is also non-trivial (as the product of a square root of $x - z$ and p is a square root of $x^3 + ax + b$). However, for the square root of $x - z$ we can define an explicit square root; a parametrisation of the circle is given as $t \mapsto re^{it} + z$ for $t \in [0, 2\pi]$ and then $x(t) - z = re^{it}$ and hence a square root is given by $\sqrt{r}e^{it/2}$ and its values at the end points are $\sqrt{r}e^0 = \sqrt{r}$ for $t = 0$ and $\sqrt{r}e^{\pi i} = -\sqrt{r}$ for $t = 2\pi$.

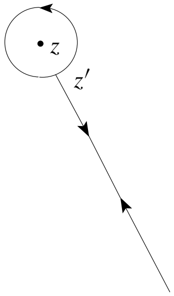

Figure 6. Transporting a square root around a zero.

This calculation can then be used as a basis for further calculation. Consider for example a curve that encircles exactly two zeroes of $x^3 + ax + b$. Such a path is end point homotopic to a path that starts by moving towards one zero, goes around in a circle around it, goes back the same way, and then does the same for the other point (cf. Fig. 7). By Proposition 2.2, transporting the square root around the original path is the same thing as first transporting solely around the first zero and then around the second. Each of these transports changes the square roots, so doing both of them transports a square root back to itself.

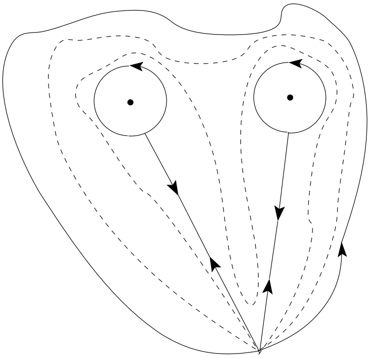

Figure 7. Going around two zeroes.

Let us now encircle the four roots (remember that we also consider ∞ a root) in pairs, so that within each of the two encirclements there are two points, and then remove the encircled areas. The result is the sphere with two "discs" cut out (cf. Fig. 8). Now we can define two square roots of $x^3 + ax + b$ on this cutout sphere S'; pick any point on it as a "base point" and for any other point on it draw a path starting at the base point and ending at that other point. We can now transport any square root along this path and the result is independent of the chosen path. In fact transport along

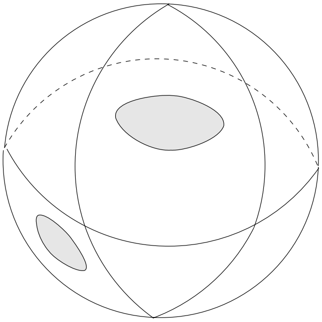

Figure 8. Sphere with two discs cut out of it.

the path which is going from the base point to the chosen point by one path and back by the other transports a square root to itself. As these two globally defined square roots are continuous, this means that the part of X, X' say, which is the inverse image of S' under the projection is the disjoint union of two open[3] subsets both of which map bijectively to S'. Hence X' is the disjoint union of two copies of S'. To recover $\overline{X}$ we must "glue in" the inverse images of the two cut out discs in $\overline{X}$ and hence we must understand these inverse images. If we look at one of these cut out discs we may make a "cut" in it; a path that does not cross itself, starting at one of the points and ending at the other (cf. Fig. 9). If we actually remove the cut we get again two well-defined square roots so that its inverse image is the disjoint union of two copies of it. We must realise however that if we transport along a path that crosses the cut we change the square root and hence pass from one copy (or "sheet" as the classical name for it is) to the other. The actual picture therefore looks something like in Fig. 10. Notice that the two sheets do not actually meet except at the two zeroes and some thought reveals that

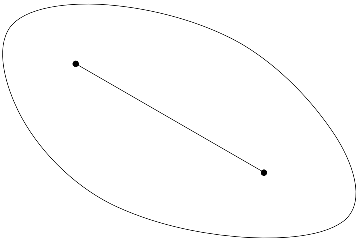

Figure 9. A disc with a cut.

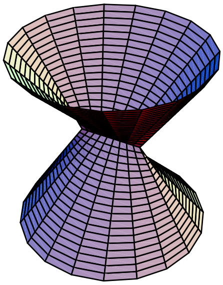

Figure 10. The inverse image of a disc with two zeroes.

[3] Recall that $X \subseteq \mathbb{C}^2 = \mathbb{R}^4$ and hence "open subsets" is a well-defined concept.

what we actually have is a cylinder. The thought process can be aided by considering a cross section of Fig. 10 where the cross section itself should be drawn so to make it clear that there is no intersection at the cut, which means we do not change the shape by switching the upper parts (cf. Fig. 11).

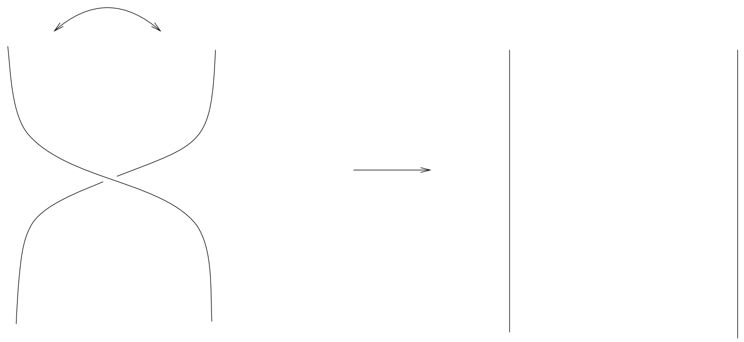

Figure 11. Unfolding to a cylinder, a cross section.

The upshot of it all is that $\overline{X}$ consists of two copies of the sphere with two discs cut out of them and with two cylinders glued in at the boundary circles of the two discs (each cylinder connecting one circle in one copy of the cut out sphere with a circle in the other). By "deflating" the spheres we see that $\overline{X}$ has the shape of a torus (cf. Fig. 12).

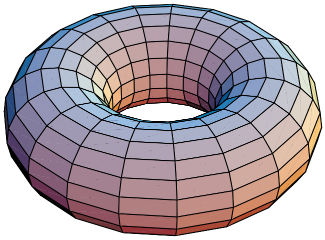

Figure 12. A torus.

Remark 1. Technically $\overline{X}$ is a topological space covered by two closed subsets X and Y of $\mathbb{C}^2$. By having the same shape we mean that they are homeomorphic, i.e., there is a continuous bijection between them with continuous inverse.

Knowing this shape will come in handy when we return to integrals.

2.2 Elliptic integrals revisited

We shall now briefly return to elliptic integrals or more precisely those of the form $\oint_C dx/y$ (which are classically known as "elliptic integrals of the first kind"). Path integrals are most often used to obtain primitive functions of their integrand. That means that we should fix a point and then for a variable point consider a path from the fixed point to the variable one. The result is then supposed to give a primitive function. The only problem is that the integral will in general depend on the path taken to the variable point and not just the point itself. That the result can't vary too much follows from the *Cauchy integral theorem* which says that the integrals of the same integrand (when it is a 1-form in our sense) along two end point homotopic paths are equal.

Remark 2. There are different ways of proving Cauchy's theorem. One way is very close to the proof of Proposition 2.2. In it one solved $y^2 = x$ (or more generally $y^2 = f(x)$) along a path and showed that the result at the end point was the same for end point homotopic paths. Similarly one may solve $y' = f(x)$ along a path and then the integral of $f(x)$ along the path is the difference of the value of y at the two end points. The proof that the value at the end point is the same for end point homotopic paths is then similar to the proof of Proposition 2.2, where the essential point is that a solution y is determined in a neighbourhood of a point by its value at the point.

Still there will be paths that give different values for the integral. It is instructive to first consider a simpler case.

Example 2. Consider the case of dx/y, where $x^2+y^2 = 1$. We know that a primitive function is $\arcsin(x)$ and as this is the inverse of a non-injective function, $\sin(x)$, it is not surprising that something should happen. To understand what happens when we integrate along two different curves, it is enough to understand what happens when we integrate along a *closed curve or path*, i.e., a curve with the same end and starting point. This is because given two curves with the same start and end points, we may integrate along the first forwards and the second backwards, which is integration along a closed curve. This integral is then the difference of the (forwards-moving) integrals along the two curves. Furthermore, the common start and end point for a closed curve is irrelevant: We may choose a fixed curve from one point z_0 to another point z_1. For a closed curve C starting and ending at z_1, we get a curve starting and ending at z_0 by going from z_0 along the fixed curve to z_1, then following C around once and then going back along the fixed curve. As we go back and forth the fixed curve once, the integral along this new curve equals the integral along C.

We now want to add points at ∞ just as we did for $y^2 = x^3 + ax + b$. For that we replace x by $1/x$ obtaining the equation $y^2 + x^{-2} = 1$ and transform it to $z^2 + 1 = x^2$, where $z = xy$. For $x = 0$ there are two choices for z, $\pm i$, and hence, just as in the case of a quartic polynomial, there are two points at ∞ distinguished from each other by the value $\pm i$ of z. We can now continue as for the elliptic case. We cut out a disc containing the two roots of $1 - x^2$ (as we have no "root" at ∞) and then we sew in a cylinder between two copies of the cut out sphere. This gives something looking like a dumbbell but by “smoothing it out” one obtains a sphere.

There is however one more thing one must keep in mind. Contrary to the elliptic case, dx/y behaves badly at infinity; we have $d(1/x)/y = -dx/(xz)$ and this most definitely means that there is a problem at ∞ (i.e., at $x = 0$ for this form). More precisely, near $x = 0$ we have that z is close to $\pm i$ and hence $dx/(xz)$ is very close to $\mp i\, dx/x$ with primitive function $\mp i \log(x)$ which does not have a limit as $x \to 0$. Just as in the elliptic case there is however no problem when $y = 0$; we have $2x\,dx + 2y\,dy = 0$ which gives $dx/y = -dy/x$ and we cannot have both $x = 0$ and $y = 0$.

Let us now consider a path that starts at 0, moves to the left and up, circles -1 and continues below the real axis, encircles 1 and then returns to 0 (cf. Fig. 13). Note

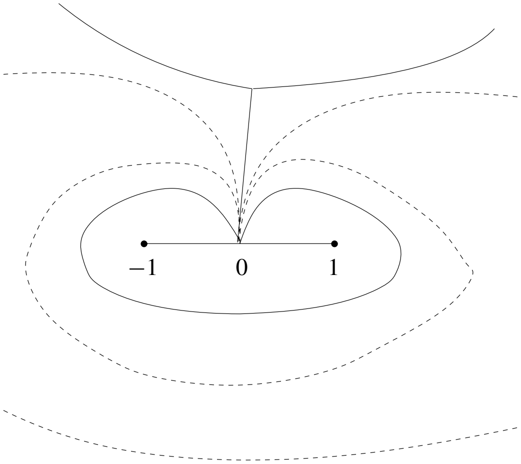

Figure 13. Two paths and homotopies between them.

that on the dumbbell this corresponds to going around once on the cylinder. We may now choose an end point homotopy from this path to a path that goes from 0 along the positive imaginary axis till it gets close to infinity, then circles infinity around a circle in the negative direction and ends by going back to 0 along the positive imaginary axis (cf. Fig. 13).

The part of the path that moves along the imaginary axis is traversed twice in opposite directions and the integrals along these two traversals cancel. Hence the integral along the first path equals the integral around a small circle around infinity. By mapping ∞ to 0 by $x \mapsto 1/x$, one reduces to computing the integral of $-dx/(xz)$ around a small circle traversed in the positive[4] direction. The circle may be made so small that z may be replaced by its value at 0, which is i if we started with $\sqrt{-1 - 0^2} = -i$.

[4] Think about it!

Everything then comes down to computing the integral of dx/x along a small circle which is easily seen to be $2\pi i$. The end result is that the integral along our original path is -2π and in particular is non-zero.

Remark 3. Another way of computing this integral is to let the original path go along the real line, make a small circular turn around -1, continue along the real axis, make a small circular turn around 1 and go back to 0 along the real axis. When the radius of the circles turns to zero, the integral will converge towards

$$\int_0^{-1} \frac{dx}{\sqrt{1-x^2}} + \int_{-1}^{1} \frac{dx}{-\sqrt{1-x^2}} + \int_1^0 \frac{dx}{\sqrt{1-x^2}} = -2\int_{-1}^{1} \frac{dx}{\sqrt{1-x^2}}$$

and is hence equal to it, as all the integrals are the same. (Combining this with the previous computation incidentally gives the well-known fact $\int_{-1}^{1} \frac{dx}{\sqrt{1-x^2}} = \pi$.)

Exercise 9. Two closed paths $C\colon [a,b] \to V$ and $D\colon [a,b] \to V$ are *freely homotopic* if there is a continuous map $E\colon [a,b] \to V$ such that $E(a,s) = E(b,s)$ for all s, $E(t,0)) = C(t)$, and $E(t,1) = D(t)$ for all t. Show that the integrals of a 1-form defined on V along C and D are equal.

If we now return to elliptic integrals it is easy to see that there are examples similar to the one just presented. For instance one may consider the lemniscate integrand $dx/\sqrt{1-x^4}$ and the same path, i.e., going along the real axis and turning around ± 1. However, we want to study the possible values that integrals along closed paths can have. What we want to show is that every closed path is end point homotopic to the composite of a small number of paths, where the *composite* of two closed paths, starting and ending at the same point, is the path that is obtained by first going around the first path and then the second.

Proposition 2.3. *Any closed path on a torus is end point homotopic to a suitable number of composites of the path that goes once around the torus in one direction and the path that goes once around the torus in the orthogonal direction* (*see Fig.* 14).

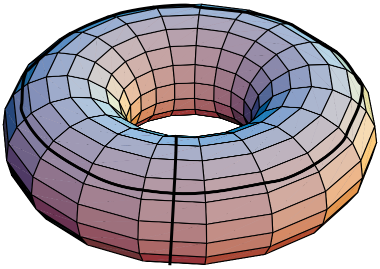

Figure 14. Two paths on the torus.

Sketch of proof. We first note that a torus is in continuous bijection with $S^1 \times S^1$ ($S^1 := \{z \in \mathbb{C} \mid |z| = 1\}$) which can be seen by, for instance, noting that as a torus is obtained by rotating a circle in the xz-plane of radius $1/2$ and centre at $(1, 0, 0)$ around the z-axis. The two coordinates can then be chosen to be the angle θ of the plane containing the point and the z-axis with the xz-plane and the angle ϕ inside of that plane from the centre of the circle with the point and the intersection of the plane and the xy-plane (see Fig. 15). What we then actually prove is that any closed path is end point homotopic with a path that first moves around in the first factor a certain number of times (and is constant in the second) and then does the same in the second factor. For that, one first proves that a closed path $C\colon [a, b] \to S^1$ with, say, $C(a) = C(b) = 1$ lifts to a path $\tilde{C}\colon [a, b] \to \mathbb{R}$, with $\tilde{C}(a) = 0$ and $C(t) = e^{i\tilde{C}(t)}$. This is done by noticing that the map $t \mapsto e^{it}$ is a continuous map which *locally* is a homeomorphism, i.e., note that any point of $\mathbb{R}$ has a neighbourhood on which it is a continuous bijection and then look at the supremum for the values where a lifting is possible, as in the proof of Proposition 1.1. Note that $\tilde{C}(b)$ is not necessarily equal to 0 but must be an integer multiple of 2π. Using this result we get a lifting of a closed path $C\colon [a, b] \to S^1 \times S^1$ to a map $\tilde{C}\colon [a, b] \to \mathbb{R}^2$. Now any two paths $D, E\colon [a, b] \to \mathbb{R}^2$ with the same end points are shown to be end point homotopic by using $F(t, s) = sD(t) + (1 - s)E(t)$. This means that the lifting $\tilde{C}$ is end point homotopic to the path that first moves linearly in the y-direction till it gets to the

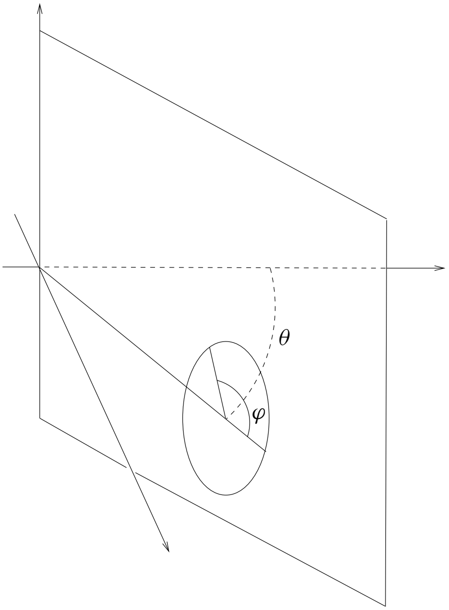

Figure 15. Angular coordinates (θ, φ) on a torus.

same height as $\tilde{C}(b)$ and then moves linearly in the x-direction till it reaches $\tilde{C}(b)$. We then compose this homotopy with the map $(x, y) \mapsto (e^{ix}, e^{iy})$ to get the desired homotopy. □

Exercise 10. Fill in the details of this proof.

Now the integral along a composite of two paths is the sum of the integrals along the two paths. Hence the proposition shows that any integral of the elliptic integrand dx/y along a closed path is an *integral* linear combination of just two such integrals. In general an integral of dx/y along a closed path is called a *period*[5] and having fixed paths as in the proposition so that every closed path is end point homotopic to a composite of them, the integrals along these two paths are called the *fundamental periods*. Hence we may phrase what we have obtained as saying that *a period is an integral linear combination of fundamental periods*. If we want to see explicitly what fundamental periods look like, we can assume that the paths do not pass through any zeroes of $x^3 + ax + b$. Then these paths are completely specified as being paths in the same plane all originating from a fixed point (which we call the *base point*). To get a starting value we have also chosen once and for all a square root of $x^3 + ax + b$ at the base point. If we do that, then such paths can be as in Fig. 16. Note that the paths in the plane intersect at a point different from the base point; as paths in $\overline{X}$ they do not intersect since they represent two different square roots at that point.

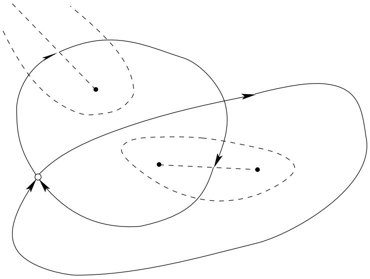

Figure 16. Two closed paths on an elliptic curve.

Just as for the case of $dx/\sqrt{1-x^2}$ we can "compress" the two paths so that they run along the line between two zeroes of $x^3 + ax + b$, turn around one zero in a small circle, move along the line again, circle the other zero and then return to the starting point.[6] Note that we traverse the line between the points twice in different directions but also with different choices of square roots so that the full integral is twice the integral along the line.

[5] We shall soon see why.

[6] We noted before that we may change base point by moving back and forth along a fixed path between the two base points and that the integrals along the fixed path cancel.

This means that we get an even more explicit recipe for the fundamental periods: Order the three zeroes of x^3+ax+b. Then the two fundamental periods are twice the integral of dx/y along the line between the two first and two last zeroes respectively. The integrals along these lines are also important and are called (unsurprisingly) *half-periods* of the integral.

Example 3. Consider the case when $x^3+ax+b = x(x-1)(x-\lambda)$ and $\lambda > 1$. Then the fundamental periods are

$$2\int_0^1 \frac{dx}{\sqrt{x(x-1)(x-\lambda)}}\,dx \quad \text{and} \quad 2\int_1^\lambda \frac{dx}{\sqrt{x(x-1)(x-\lambda)}}\,dx.$$

The fact that the integral gives a many-valued primitive function is not very satisfactory. It was Abel[7] who had the idea of considering the inverse function, i.e., a function f that should fulfil

$$z = \int_{z_0}^{f(z)} \frac{dx}{y},$$

where the integral means the integral along any path from z_0 to z. What Abel showed was that this gives a well-defined function f. Note however that if we can get z as a value on the left-hand side, we can also get $z+\gamma$, where γ is a period, as the value of the integral by changing the path of integration. This forces $f(z)$ to be equal to $f(z+\gamma)$ and hence that f will have γ as a *period.*

Remark 4. This result should be compared with the formula

$$z = \int_0^{\sin(z)} \frac{dx}{\sqrt{1-x^2}}$$

and the fact that 2π is a period of $\sin(z)$.

It turns out to be very natural to use $z_0 = \infty$ as the starting point for our elliptic integrals so that $f(0) = \infty$. In the next section we shall consider the kind of function that one obtains in this way.

[7] Niels Henrik Abel, 1802–1829

3

Elliptic functions

We shall now go the other way, namely we shall start with the properties the putative inverse of an elliptic integral should have and investigate such functions. To begin with the function should be periodic with periods all integral linear combinations of two fundamental periods[1] ω_1 and ω_2. We also must allow ∞ as a value and the function should take ∞ as a value at 0 (and hence at all periods). There is however one further important property of the function which is incorporated in the following definition.

Definition 3.1. Let U be an open subset of $\mathbb{C}$.

i) A function $f\colon U \to \mathbb{C}$ is *holomorphic* if for every $a \in U$ there is a disc $D(a,r) \subseteq U$, $D(z,r) := \{w \in \mathbb{C} \mid |z-w| < r\}$, such that the restriction of f to $D(a,r)$ has the form

$$f(z) = \sum_{n\geq 0} a_n (z-a)^n,$$

where the right-hand side is convergent in $D(a,r)$.

ii) A function $f\colon U \to \overline{\mathbb{C}}$ is *meromorphic* if for every $a \in U$ there is a disc $D(a,r) \subseteq U$, $D(z,r) := \{w \in \mathbb{C} \mid |z-w| < r\}$, such that the restriction of f to $D(a,r)$ has the form

$$f(z) = \sum_{n\geq -N} a_n (z-a)^n,$$

where the right-hand side is convergent in $D(a,r) \setminus \{a\}$ and $f(a) = a_0$ if $a_n = 0$ for $n < 0$ and ∞ otherwise. The points where f takes the value ∞ are called *poles* and the *order* of the pole is minus the least n such that $a_n \neq 0$.

We shall require our functions not only to be periodic with periods ω_1 and ω_2 but also to be meromorphic. Before we go on to doing that we shall deal with an uninteresting case.

Exercise 11. i) Show that if ω_1 and ω_2 are linearly dependent over the rationals, then there is an ω which is an integral linear combination of ω_1 and ω_2 and such that they are integral multiples of ω.

ii) Show that if ω_1 and ω_2 are linearly dependent over the reals but independent over the rationals, then there are non-zero integral linear combinations of them of arbitrarily small absolute value.

iii) Show that a meromorphic function on a connected open subset of $\mathbb{C}$ that is constant in a neighbourhood of a point is constant everywhere.

[1] Classically the periods were of the form $2\omega_1$ and $2\omega_2$ so that ω_1 and ω_2 were the half periods. We have chosen to follow modern conventions.

iv) Show that, given a point and its domain of definitions, a meromorphic function can not have the same value at all points of a set of points that converges to that point.

v) Show that if a meromorphic function $f\colon \mathbb{C} \to \overline{\mathbb{C}}$ has two periods ω_1 and ω_2 that are linearly dependent over the reals but independent over the rationals, then f is constant.

This exercise shows that if the two periods ω_1 and ω_2 lie on the same line (through the origin) we either have a completely trivial case or one which really has only one period. We exclude those cases, so from now on we assume that ω_1 and ω_2 are linearly independent over the reals. We next have to think about the domain of definition of our functions. As they (or at least some of them) are supposed to be inverses of elliptic integrals, their domain of definition will be the set of complex numbers that can be written as an elliptic integral. The following exercise suggests that all complex numbers should be obtainable in this way.

Exercise 12. i) Show that the set of values of a fixed elliptic integral (along all, not just the closed paths) form a subgroup of $\mathbb{C}$.

ii) Show that the set of values of a fixed elliptic integral is an open set.

iii) Show that the set of values of a fixed elliptic integral is all of $\mathbb{C}$.

Taking all these things together we arrive at the following definition: An *elliptic function* with periods ω_1 and ω_2 is a meromorphic function f defined on all $\mathbb{C}$ such that $f(z+\omega_1) = f(z+\omega_2) = f(z)$ for all $z \in \mathbb{C}$.

We also introduce the*fundamental domain* of ω_1 and ω_2,

$$\mathcal{F} = \mathcal{F}_{\omega_1,\omega_2} := \{r\omega_1 + s\omega_2 \mid 0 \leq r, s \leq 1\}$$

(cf. Fig. 17) and the notation $\Gamma = \Gamma_{\omega_1,\omega_2} := \{m\omega_1 + n\omega_2 \mid m, n \in \mathbb{Z}\}$. The significance of these definitions is the following lemma.

Lemma 3.2. *Every $z \in \mathbb{C}$ can be written in the form $z_0 + \gamma$ with $z_0 \in \mathcal{F}_{\omega_1,\omega_2}$ and $\gamma \in \Gamma_{\omega_1,\omega_2}$. This form is unique except for the fact that $r\omega_1 + m\omega_1 + n\omega_2 = r\omega_1 + \omega_2 + m\omega_1 + (n-1)\omega_2$, $s\omega_2 + m\omega_1 + n\omega_2 = \omega_1 + s\omega_2 + (m-1)\omega_1 + n\omega_2$, and $m\omega_1 + n\omega_2 = \omega_1 + \omega_2 + (m-1)\omega_1 + (n-1)\omega_2$ (cf. Fig.* 17).

Proof. This is a statement that only uses the structure of $\mathbb{C}$ as a real vector space and is independent of the choice of basis so that we may choose ω_1 and ω_2 as basis, in which case the statement follows from the corresponding statement for $\mathbb{R}$ and 1 and then it is immediate. □

What this means is that any elliptic function f is completely determined by its values on the fundamental domain, as $f(z) = f(z_0 + \gamma) = f(z_0)$. We shall use that fact to prove our first, seemingly disappointing, result.

Proposition 3.3. *A holomorphic elliptic function is constant.*

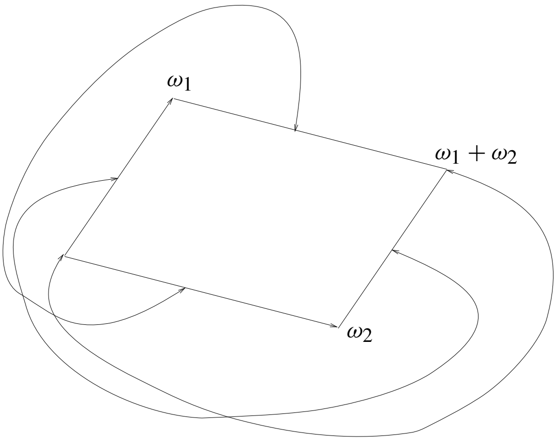

Figure 17. The fundamental domain and its identifications.

Proof. Let f be such a function. A holomorphic function is continuous and hence so is $z \mapsto |f(z)|$. As such it has a maximum on the closed and bounded set $\mathcal{F}$. Let a be a point where the maximum is achieved. As all the values of f are attained already on $\mathcal{F}$, we get that $|f(z)| \leq |f(a)|$ for all z. By definition we may write f as

$$f(z) = \sum_{n=0}^{\infty} a_n (z-a)^n$$

in a disc around a. If all $a_n = 0$ for $n > 0$, then f is constant in the disc and then it is constant everywhere by Exercise 11 iii). Therefore we may assume that some such $a_n \neq 0$ and we may assume that n is the smallest such index. This means that we can write $f(z)$, again in a disc around a, as $a_0 + (z-a)^n h(z)$ with $h(a) \neq 0$. We can not have $a_0 = 0$ because then $|f(z)| \leq |f(a)| = |a_0| = 0$ and f is constant. Hence we can scale f so that $a_0 = 1$. Now, we may find a z such that it lies in the disc and such that $(z-a)^n h(a) = r$ for some $r > 0$. Then, if we put $z_s := s(z-a) + a$ for $0 < s < 1$, we get

$$f(z_s) = 1+(z_s-a)^n h(z_s) = 1+s^n(z-a)^n h(z_s) = 1+s^n r+s^n(z-a)^n(h(z_s)-h(a)).$$

This in turn implies

$$\begin{aligned} 1 = |f(a)| &\geq |f(z_s)| \geq |1+s^n r| - |s^n(z-a)^n(h(z_s)-h(a))| \\ &= 1 + s^n r - |s^n(z-a)^n(h(z_s)-h(a))|, \end{aligned}$$

which gives $s^n|z-a|^n|h(z_s)-h(a)| \geq s^n r$, and dividing by s^n we have $|z-a|^n|h(z_s)-h(a)| \geq r$ but letting $s \to 0$ gives $0 = |a-a|^n \cdot 0 \geq r$ which is a contradiction. □

Remark 5. This would seem to follow more directly from Liouville's[2] theorem that a bounded entire function is constant. However, the result just proved *is* Liouville's theorem, it was Cauchy[3] who generalised it to bounded entire functions. Liouville's proof is different from the one given here and is given in the next exercise.

Exercise 13. Let f be a holomorphic function with periods ω_1 and ω_2 (linearly independent over the reals).

By expanding $g(x, y) = f(x\omega_1 + y\omega_2)$ in a Fourier series and considering the Cauchy–Riemann equations for the resulting series, show that f is constant.

This result is only seemingly disappointing because as we shall see it is a very powerful tool for proving relations between *meromorphic* elliptic functions. In any case it shows that if one wants to construct (non-trivial) elliptic functions they must have poles somewhere. We aim for a pole at the origin and no other than those forced upon us by periodicity. It turns out that this can not be done by having a pole of order 1. The simplest function with a pole at 0 of order greater than 1 is $1/z^2$. It is of course far from being periodic with ω_1 and ω_2 as periods, but one could try to make a periodic function by taking the sum

$$\sum_{\gamma\in\Gamma}\frac{1}{(z+\gamma)^2}.$$

It turns out however that this sum is not convergent, as the following exercise indicates.

Exercise 14. Show that

$$\sum_{(m,n)\in\mathbb{Z}^2\setminus\{(0,0)\}}\frac{1}{m^2+n^2}$$

is divergent and conclude that the suggested sum above is (at least) not absolutely convergent.

There is a trick that can be used to improve convergence, namely to consider instead

$$\wp(z|\omega_1,\omega_2) := \frac{1}{z^2} + \sum_{m,n\in Z}{}' \frac{1}{(z-m\omega_1-n\omega_2)^2} - \frac{1}{(m\omega_1+n\omega_2)^2},$$

where the prime is used to signal the fact that the summand with $m = n = 0$ is left out. This is *Weierstrass' $\wp$-function.*[4] We leave the convergence of this sum including a technical strengthening to the following exercise.

Exercise 15. Show that this sum is uniformly convergent in a neighbourhood of each point where one excludes the polar term when $z = m\omega_1 + n\omega_2$.

[2] Joseph Liouville, 1809–1882

[3] Augustin Louis Cauchy, 1789–1857

[4] Karl Weierstrass, 1815–1897

There are several consequences of the uniform convergence. The first is that this is indeed a meromorphic function (as the uniformly convergent limit of holomorphic functions is holomorphic), a second that one can take the term-wise derivative to obtain the derivative of the function. Note that because of the modification that we were forced to do it is less clear that the obtained function is indeed periodic. This can be shown by manipulating the sum directly, but we shall use another route. We start by using the fact that we may take the term-wise derivative and obtain

$$\wp'(z|\omega_1,\omega_2) = -2\sum_{m,n\in Z}\frac{1}{(z-m\omega_1-n\omega_2)^3}.$$

This sum has the form that we intended for $\wp$ and thus is obviously periodic. Now $\wp(z+\omega_1|\omega_1,\omega_2)$ and $\wp(z|\omega_1,\omega_2)$ are two functions with the same derivative and hence differ by a constant. By putting $z=-\omega_1/2$ we get that this constant equals $\wp(\omega_1/2)-\wp(-\omega_1/2)$. This is zero, however, as $\wp(z)$ is an *even* function, i.e., $\wp(-z)=\wp(z)$ as can be seen from the definition:

$$\begin{aligned}\wp(-z|\omega_1,\omega_2) &= \frac{1}{z^2}+\sum_{m,n\in Z}{}' \frac{1}{(-z-m\omega_1-n\omega_2)^2}-\frac{1}{(m\omega_1+n\omega_2)^2}\\ &= \frac{1}{z^2}+\sum_{m,n\in Z}{}' \frac{1}{(z-(-m)\omega_1-(-n)\omega_2)^2}-\frac{1}{((-m)\omega_1+(-n)\omega_2)^2}\\ &= \wp(z|\omega_1,\omega_2).\end{aligned}$$

We thus see that $\wp(z)$ is an even elliptic function and $\wp'$, being the derivative of an even function, is odd. Both of them have poles only at the points of Γ. We shall now see that we can use Liouville's theorem to find a relation between them. Consider on the one hand $\wp'^2(z)$ and on the other hand $4\wp^3(z)$. Both of them are even elliptic functions with a pole of order 6 at 0. The $1/z^3$-term for the power series expansion of $\wp'^2(z)$ at 0 is $4/z^6$ and this is true also for $4\wp^3(z)$. That means that $\wp'^2(z)-4\wp^3(z)$ is an even elliptic function with a pole of order at most 4 at the origin and no poles outside of Γ. Let a be the coefficient of $1/z^4$ in the power series expansion of this difference around 0. Then $\wp'^2(z)-4\wp^3(z)-a\wp^2(z)$ is an even elliptic function with a pole of order at most 2 at the origin and no poles outside of Γ. Let b be the coefficient of $1/z^2$ of this function. Then $\wp'^2(z)-4\wp^3(z)-a\wp^2(z)-b\wp(z)$ is an even elliptic function with no poles on or off Γ and is hence a constant c. This means that we have the relation

$$\wp'^2(z)=4\wp^3(z)+a\wp^2(z)+b\wp(z)+c.$$

We shall now analyse this a little bit more carefully to obtain explicit forms for the constants a, b, and c. We start by expanding $\wp(z)$ in a Taylor series around the origin.

Proposition 3.4. *We have an expansion, for $|z|<\min_{\gamma\in\Gamma\setminus\{0\}}|\gamma|$,*

$$\wp(z|\omega_1,\omega_2)=\frac{1}{z^2}+\sum_{n=2}^{\infty}(2n-1)G_n(\omega_1,\omega_2)z^{2n-2},$$

where

$$G_k(\omega_1, \omega_2) := \sum_{m,n\in\mathbb{Z}}{}' \frac{1}{(m\omega_1 + n\omega_2)^{2k}}$$

is the Eisenstein series.[5]

Proof. We write the summands of the non-polar part as

$$\begin{aligned}\frac{1}{(z-\gamma)^2} - \frac{1}{\gamma^2} &= \frac{1}{\gamma^2}\left(\frac{1}{(1-z/\gamma)^2} - 1\right) = \frac{1}{\gamma^2}\left(\left(1 + \frac{z}{\gamma} + \frac{z^2}{\gamma^2} + \cdots\right)^2 - 1\right)\\ &= \sum_{n=1}^{\infty} \frac{(n+1)z^n}{\gamma^{n+2}}\end{aligned}$$

for every $0 \neq \gamma \in \Gamma$. Summing over all such γ and exchanging summation signs gives

$$\wp(z|\omega_1, \omega_2) = \frac{1}{z^2} + \sum_{n=2}^{\infty}(2n-1)G_n(\omega_1, \omega_2)z^{2n-2},$$

as the odd degree parts sum up to zero because $\gamma \mapsto -\gamma$ leaves the sum invariant. That exchanging the summation signs is allowable is easily checked.

An alternative proof is obtained by considering $\wp(z|\omega_1, \omega_1) - 1/z^2$ and determining its Taylor expansion by taking derivatives and substituting $z = 0$. □

Exercise 16. Justify the formal manipulations in the proof.

Exercise 17. Show that

$$\lim_{\Im(z)\to\infty} G_k(z, 1) = 2\zeta(2k),$$

where $\zeta(s) := \sum_{n=1}^{\infty} 1/n^s$.

We now, for reasons that will become clear momentarily, put $g_2 := 60G_2$ and $g_3 := 140G_2$. We then have that

$$\wp(z) = \frac{1}{z^2} + \frac{g_2}{20}z^2 + \frac{g_3}{28}z^4 + \cdots .$$

This gives[6]

$$\begin{aligned}\wp(z) &= \frac{1}{z^2} + \frac{g_2}{20}z^2 + \frac{g_3}{28}z^4 + \cdots ,\\ \wp'(z) &= -\frac{2}{z^3} + \frac{g_2}{10}z + \frac{g_3}{7}z^3 + \cdots ,\\ \wp'^2(z) &= \frac{4}{z^6} - \frac{2g_2}{5}\frac{1}{z^2} + \frac{4g_3}{7} + \cdots .\end{aligned}$$

[5] Ferdinand Gotthold Max Eisenstein, 1823–1852

[6] See http://www.math.su.se/~teke/undervisning/Elliptisk.nb for verification.

From this we see that the Taylor expansion of $\wp'^2(z) - (4\wp^3(z) - g_2\wp(z) - g_3)$ has a z^2-term as first possible non-zero term and then by Liouville's theorem it is actually equal to zero. This gives a differential equation for $\wp$:

$$\wp'^2(z) = 4\wp^3(z) - g_2\wp(z) - g_3. \tag{3.1}$$

Exercise 18. Show, by substituting the Taylor expansion of $\wp(z)$ directly into this differential equation, that the coefficients of that Taylor expansion are polynomials in g_2 and g_3. Conclude that $G_n(\omega_1, \omega_2)$ is a polynomial[7] (with rational coefficients) in $G_2(\omega_1, \omega_2)$ and $G_3(\omega_1, \omega_2)$.

i) Show that $\zeta(2k)$, $1 < k \in \mathbb{Z}$, is a polynomial in the $\zeta(4)$ and $\zeta(6)$.

Exercise 19. Show that $\wp^{(3)}(z) = 12\wp(z)\wp'(z)$ for all z.

If we now consider the elliptic integral $\int dx/\sqrt{4x^3 - g_2x - g_3}$, then we may make the variable substitution $x = \wp(z)$ and get

$$\int \frac{dx}{\sqrt{4x^3 - g_2x - g_3}} = \int \frac{\wp'(z)}{\sqrt{\wp'^2 2(z)}}\, dz = \int dz.$$

If we do this more carefully we arrive at the following result.

Proposition 3.5. *Let $\overline{X}$ be obtained from $\{(x, y) \in \mathbb{C}^2 \mid y^2 = 4x^3 - g_2x - g_3\}$ by adding a point over ∞. For every $z \in \mathbb{C}$ there is a path C in $\overline{X}$ starting at ∞ and ending at $(\wp(z), \wp'(z))$ such that*

$$\oint_C \frac{dx}{y} = z.$$

Proof. We choose, for instance, the curve $C(t) = (\wp(tz), \wp'(tz))$ for $t \in [0, 1]$. Then

$$\oint_C \frac{dx}{y} = \int_0^1 \frac{\wp'(tz)z}{\wp'(tz)}\, dt = z. \qquad \square$$

In somewhat more naive terms this can be formulated as

$$z = \int_\infty^{\wp(z)} \frac{dx}{\sqrt{4x^3 - g_2x - g_3}},$$

which means that we have inverted an elliptic integral in Abel's sense.

We shall now see that the addition formula can be derived in a similar fashion.

[7] See http://www.math.su.se/~teke/undervisning/Elliptisk.nb for formulas.

Proposition 3.6. *If $u, v, w \in \mathbb{C} \setminus \Gamma$ with $u + v + w = 0$, then we have the following relation:*

$$\begin{vmatrix} \wp(u) & \wp'(u) & 1 \\ \wp(v) & \wp'(v) & 1 \\ \wp(w) & \wp'(w) & 1 \end{vmatrix} = 0.$$

Proof. We solve directly for w; $w = -u - v$ and consider the expression as a meromorphic function in u. We start by expanding $\wp(u)$ and $\wp(-u - v)$ as Taylor series:

$$\wp(u) = \frac{1}{u^2} + O(u^2),$$
$$\wp(-u - v) = \wp(v) + \wp'(v)u + \wp''(v)\frac{u}{2} + O(u^3).$$

Expanding[8] the determinant gives

$$\begin{vmatrix} \wp(u) & \wp'(u) & 1 \\ \wp(v) & \wp'(v) & 1 \\ \wp(w) & \wp'(w) & 1 \end{vmatrix} = O(1)$$

and hence the determinant is holomorphic at $u = 0$. At $u = -v$ we may introduce $u' = -u - v$ and then $u = -u' - v$ and the determinant becomes the same, but with top and bottom row interchanged (which of course only changes the sign). The same argument then gives that the determinant is holomorphic at $u = -v$. By Liouville's theorem this means that the determinant is independent of u. However, exchanging the rôles of u and v shows that it is also independent of v and hence is a constant. However, putting $u = v$ gives two equal rows and thus the constant value is zero. □

Remark 6. We shall see a simpler proof of this fact later.

What does this relation mean? Well, we have the following linear algebra exercise.

Exercise 20. Show that the three pairs (r, s), (u, v), and (x, y) of $\mathbb{C}^2$ (or $\mathbb{R}^2$) lie on a line precisely when

$$\begin{vmatrix} r & s & 1 \\ u & v & 1 \\ x & y & 1 \end{vmatrix} = 0.$$

Hence what the relation says is that if $u + v + w = 0$, then $(\wp(u), \wp'(u))$, $(\wp(v), \wp'(v))$, and $(\wp(w), \wp'(w))$ lie on a line. Now, u, v, w is the value of the elliptic integral and we know by the addition formula that $u + v + w = 0$ if $(\wp(u), \wp'(u))$, $(\wp(v), \wp'(v))$, and $(\wp(w), \wp'(w))$ lie on a line. Our result thus provides a converse of that. Trying to understand the relations between these two results makes it natural to ask when a point on the elliptic curve $y^2 = 4x^3 - g_2x - g_3$ is of the form $(\wp(z), \wp'(z))$ for some z. We shall deal with this question and others in the next subsection.

[8] See http://www.math.su.se/~teke/undervisning/Elliptisk.nb for this expansion.

3.1 Poles and zeroes

We have seen that each pole of a meromorphic function is *isolated*, i.e., if z is a pole, then there is a disc $D(z, r)$ around z such that z is the only pole in it. The same argument shows that also the set of zeroes of a meromorphic function which is not identically 0 is isolated (and in fact the set consisting of the zeroes and the poles is isolated). For an elliptic function this implies that they are finite in number:

Exercise 21. Show that a non-zero elliptic function has only a finite number of zeroes and poles in a fundamental domain.

We now want to count the number of zeroes and poles. Just as for polynomials and rational functions it will be necessary to count them with multiplicity. Hence if f is a non-zero meromorphic function, we may write it close to a point a as

$$f(z) = \sum_{n \geq N} a_n z^n$$

with $a_N \neq 0$. If $N < 0$ we have a pole and we say that it has *multiplicity* $-N$, similarly if $N > 0$ we have a zero of *multiplicity* N. We should also take care not to over-count the number of zeroes. Thus the phrase *the number of non-congruent zeroes counted with multiplicity* means that we take the sum of the multiplicities over a subset of the zeroes in the fundamental domain where we take all the zeroes and throw out all but one of a set of zeroes that differ by a period. Similarly for poles. We now have the following result.

Proposition 3.7. *The number of non-congruent zeroes counted with multiplicity of a non-zero elliptic function f equals the number of non-congruent poles counted with multiplicity.*

Proof. We consider the 1-form df/f and integrate along the boundary of the fundamental domain in counter-clockwise fashion (cf. Fig. 18). On the one hand this integral is zero, as f (and its derivative) has the same values on opposite sides of the fundamental domain and the path goes in opposite directions on opposing sides. On the other hand we may deform this path as shown in Fig. 18 to a path that circles around the poles and zeroes in small circles and a number of paths that are traversed twice in opposite directions; by Cauchy's integral theorem these integrals are equal. For the second path, paths going in opposite directions cancel. What remains of the second integral is the sum of the integrals along the small circle, but if

$$\frac{f'(z)}{f(z)} = \sum_{n \geq -N} a_n (z-a)^n,$$

a a zero or pole, then this integral is easily seen to be $2\pi i a_{-1}$. However, if $f(z) = (z-a)^N h(z)$ with $h(a) \neq 0$, then $f'(z)/f(z) = N/(z-a) + h'(z)/h(z)$ which

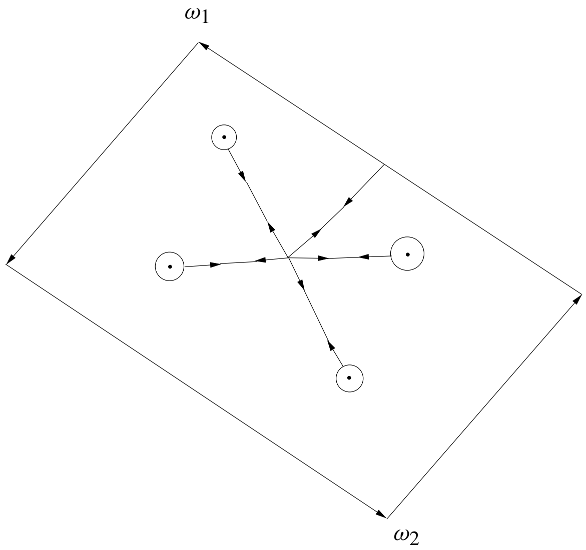

Figure 18. Integrating around the fundamental domain.

shows that a_{-1} equals the multiplicity of the zero or minus the multiplicity of the pole. As these values sum up to the integral which has been shown to be zero, we get the result. □

Exercise 22. The proof works as it stands only if f has no zeroes or poles on the boundary of the fundamental domain. Work out the case when this is no longer true.

Exercise 23. Show that a rational function has the same number of poles and zeroes counted with multiplicity on the Riemann sphere.

Example 4. Consider $\wp'(z)$. It has a triple pole at each point of Γ and no other, so the number of non-congruent poles counted with multiplicity is 3 and hence there should be three zeroes. Now, $\wp'(\omega_1/2) = \wp'(\omega_1/2 - \omega_1) = \wp'(-\omega_1/2) = -\wp'(\omega_1/2)$ which means that $\wp'(\omega_1/2) = 0$ and the same is true for $\wp'(\omega_2/2)$ and $\wp'((\omega_1 + \omega_2)/2)$. These three zeroes do not differ from any of the others by an element of Γ and hence they give three different zeros. Thus they are all the zeroes and they are zeroes of multiplicity 1.

It is somewhat awkward to use Proposition 3.7 as one has to be sure to count only zeroes and poles in the fundamental domain and then make sure not to count them twice if they lie on its boundary. It turns out to be much more convenient to be more abstract and consider instead an elliptic function with periods ω_1 and ω_2 as a function on the *quotient group*[9] $E_{\omega_1,\omega_2} := \mathbb{C}/\Gamma_{\omega_1,\omega_2}$. Here we are in much better shape as each zero and pole occurs exactly once on E_{ω_1,ω_2}. There is also no problem in defining the multiplicity, simply take any representative of the equivalence class and look at the multiplicity at that point. Note that the fundamental domain maps surjectively onto $E = E_{\omega_1,\omega_2}$ and two points of the fundamental domain map to the same point precisely when they lie on opposite sides of it and differ by a translation by a fundamental period. A moment's thought shows that, gluing like this, the opposite side of a fundamental parallelogram gives rise to a torus (cf. Fig. 19). This is not surprising as an elliptic curve also has the shape of a torus. Proposition 3.7 can then be formulated as saying that the sums of the zeroes, resp. poles, of f as a function on E counted with multiplicities are the same. Another advantage of E is that the next result has its most natural formulation using it.

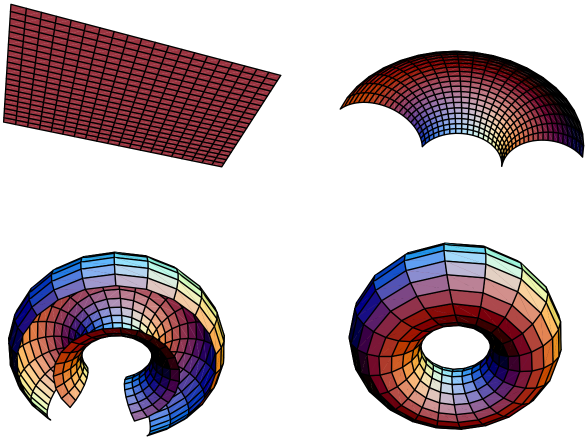

Figure 19. Gluing a torus together.

Theorem 3.8. *Let ω_1 and ω_2 be complex numbers linearly independent over the reals and let $g_2 = g_2(\omega_1,\omega_2)$ and $g_3 = g_3(\omega_1,\omega_2)$.*

i) *The polynomial $p(x) := 4x^3 - g_2x - g_3$ has the roots $\wp(\omega_1/2)$, $\wp(\omega_2/2)$, and $\wp((\omega_1+\omega_2)/2)$ and they are distinct. In particular $p(x)$ does not have multiple roots.*

ii) *Let $\overline{X}$ be the set $X := \{(x,y) \in \mathbb{C}^2 \mid y^2 = 4x^3 - g_2x - g_3\}$, with the point at infinity adjoined. Then the map $P\colon E_{\omega_1,\omega_2} \to \overline{X}$ which takes $[z]$ to $(\wp(z), \wp'(z))$ if $[z] \neq [0]$ and to the point at infinity if $[z] = [0]$ is a bijection.*

Proof. We have seen in the example above (Example 4) that $\wp'(\omega_1/2) = \wp'(\omega_2/2) = \wp'((\omega_1+\omega_2)/2) = 0$ and by (3.1) $\wp(\omega_1/2)$, $\wp(\omega_2/2)$, and $\wp((\omega_1+\omega_2)/2)$ are roots

[9] I.e., the set of equivalence classes under the equivalence relation $z \sim w \iff z - w \in \Gamma$. We define $[z] \pm [w] := [z \pm w]$.

of $p(x)$ and to prove the first part it remains to show that they are distinct. We do this by proving a more general fact (that we shall also use in the second part):

Assume that $\wp([z_0]) = \wp([z_1])$. If the common value is ∞, then $[z_0] = [0] = [z_1]$ and if not $f(z) := \wp(z) - \wp(z_0)$ has a double pole on E and is zero for z_0 and $-z_0$. If $[z_0] \neq -[z_0]$, then $[z_0]$ and $-[z_0]$ are two distinct zeroes of f and by Proposition 3.7 there are no more, so that $[z_1] = \pm[z_0]$. If $[z_0] = -[z_0]$, then $g(z) := f(z + z_0)$ is an even function as $g(-z) = f(-z + z_0)$ which is equal to $f(-z - z_0)$ as $[z_0] = -[z_0]$ which in turn equals $f(z + z_0)$ as f is even. That means that for the Taylor expansion of f around z_0, only even terms appear and therefore z_0 is a double root of f and again there are no more roots.

This can be rephrased as saying that $\wp([z]) = \wp([w])$ implies that $[w] = \pm[z]$. Applied to $z = \omega_1/2, \omega_2/2, (\omega_1 + \omega_2)/2$ this finishes the proof of the first part.

As for the second part we start by noticing that the point at infinity lies in the image. Assume now that $(x, y) \in X$. As $\wp(z) - x$ has a double pole, it has a zero by Proposition 3.7. By (3.1) we then have that $\wp'^2(z) = y^2$ and hence $y = \pm\wp'(z)$. But $\wp(-z) = \wp(z)$ and $\wp'^2(-z) = -\wp'(z)$, so after possibly replacing z by $-z$ we get $(x, y) = (\wp(z), \wp'(z))$. This proves surjectivity. Assume now that $(\wp([z]), \wp'([z])) = (\wp([w]), \wp'([w]))$. By what was just proved from the equality of first factors we get $[w] = \pm[z]$ and if $[w] \neq [z]$ we get $\wp'([z]) = \wp'([w]) = \wp'(-[z]) = -\wp'([z])$, i.e., $\wp'([z]) = 0$. By the first part this implies that $[z] = -[z]$ and hence $[w] = [z]$. □

The theorem gives an almost complete correspondence between the analytic and algebraic sides:

- We have a bijection between the set of solutions of $y^2 = 4x^3 - g_2x - g_3$ (with a point at infinity added) and $\mathbb{C}/\Gamma$.
- The functions $\wp([z])$ and $\wp'([z])$ correspond to the functions x and y, i.e., the projections on the two factors.
- The 1-form dx/y corresponds to the 1-form dz.

This, however, does not answer all questions. Three that might come to mind are:

- What is the precise relation between the condition $[u] + [v] + [w] = [0]$ and $P(u)$, $P(v)$, and $P(w)$ lying on a line?
- Which are the elliptic functions (for given periods)?
- Do we get all elliptic curves (with complex coefficients) in this way?

For the first question we have proven that $[u] + [v] + [w] = [0]$ implies that $P(u)$, $P(v)$, and $P(w)$ lies on a line provided that $[u]$, $[v]$, and $[w]$ are all different from $[0]$, but we do not know what to say if one of them equals $[0]$. On the other hand Euler's addition formula would seem to give a converse. Let us first deal with the case that one of them, $[u]$ say, is 0. This means that $[w] = -[w]$ which implies that $\wp([z]) = \wp([w])$ and $\wp'([z]) = -\wp'([w])$. From the algebraic side this means that the x-coordinates are equal and the y-coordinates are then the two square roots of $4x^3 - g_2x - g_3$. From the point of view of E_{ω_1,ω_2} we thus see that lines of the form $x = c$ that intersect the elliptic curve should be thought of as lines that also pass through the point at infinity. We shall postulate this for now but later we shall also see that from the algebraic point of view this is not an arbitrary but in fact very natural choice. We have to make a decision when three points, some of which may be equal, lie on a line. When we are dealing with a line that is not parallel to the y-axis, there is no problem. As we saw we get a cubic equation in x when we intersect the elliptic curve with a line and this equation may have multiple roots. If we count these roots with multiplicity, then we always have three points on the intersection between such a line. For lines parallel to the y-axis we count, as defined previously, ∞ as one of the intersection points but we also count any point with $y = 0$ twice on the line parallel to the y-axis on which it lies. Finally, which seems to be the most ad hoc choice of them all but will be vindicated later, we say that ∞ lies on a line that meets the curve three times at ∞ and at no other points on the curve.

With these definitions Proposition 3.6 remains true even when $[u]$, $[v]$, or $[w]$ are zero, at least in the sense that the images under P (which is the map with $\wp(z)$ and $\wp'(z)$ as its two components) lie on a line. We shall not prove this now as it will become much clearer after some more foot work. We can however at this point give a positive answer to the first question.

Proposition 3.9. *The images of three points* $[u], [v], [w] \in E_{\omega_1,\omega_2}$ *under the map* $[z] \mapsto (\wp(z), \wp'(z))$ *lie on a line precisely when* $[u] + [v] + [w] = 0$.

Proof. One direction follows from Proposition 3.6 together with the claim that we can extend its validity when one of the three elements is $[0]$ and some arguments needed to prove that, if two of the points $[u]$, $[v]$, or $[w]$ are equal, then their images under P are the three points on the line in the sense where multiplicity is taken into account. Assume conversely that $P([u])$, $P([v])$, and $P([w])$ lie on a line. If the line is the line that only the point at infinity lies on, then we have $[0] + [0] + [0] = 0$. If the line is parallel to the y-axis, then the two points on it that are distinct from ∞ are of the form $P([z])$ and $P([-z])$ and we have $[0] + [z] + [-z] = 0$. Finally, assume that the line is one that does not contain the point at ∞, then two of the points are of the form $P([u])$ and $P([v])$. Then the third point on the line through $P([u])$ and $P([v])$ is, by Proposition 3.6, $P([-u - v])$, but that accounts for all three points on the only line that passes through all three of them (again there are some special cases to consider when there are coincidences between the points). □

Exercise 24. Work through all the special cases of the proof.

When it comes to describing all elliptic functions we shall now see that they are all rational functions in $\wp(z)$ and $\wp'(z)$.

Proposition 3.10. *All elliptic functions of periods ω_1 and ω_2 are of the form $f(\wp(z))+g(\wp(z))\wp'(z)$, where $f(x)$ and $g(x)$ are rational functions.*

Proof. Let h be such an elliptic function. We start by writing it as a sum of an even and an odd function:

$$h(z) = \frac{h(z)+h(-z)}{2} + \frac{h(z)-h(-z)}{2}.$$

Both parts are still elliptic with the same periods so we are reduced to the case when h is odd or even. Assume that we have proved that if h is even, then it is a rational function in $\wp(z)$. If h is odd, then we divide h by $\wp'(z)$. The quotient is still meromorphic and clearly periodic with periods ω_1 and ω_2 and is also even and hence by assumption a rational function in $\wp(z)$ which proves the result.

We are thus left with the case when h is even. Consider an $\alpha \in E_{\omega_1,\omega_2}$. If $\alpha \neq -\alpha$, then because h is even, the multiplicity of α as a zero or pole of h is the same as that of $-\alpha$ as is seen by transforming a Taylor expansion of $h(z)$ at α to one for $h(-z)$ at $-\alpha$. On the other hand, if $\alpha = -\alpha$, then the same argument gives that the multiplicity at α is even. In particular the multiplicity of 0 as a pole or zero is even. Consider now $\wp^n(z)h(z)$, where n is half the multiplicity of 0 as a zero, if it is a zero, and $-n$ is half the multiplicity of 0 as a pole, if it is a pole, and is 0 otherwise. Then $\wp^n(z)h(z)$ has neither a zero nor a pole at 0 and we are thus reduced to h having that property. Now, the fact that any α with $\alpha = -\alpha$ is a pole or zero with even multiplicity means that we may enumerate some of the zeroes of h: $[a_1], [a_2], \dots, [a_k]$ (possibly with multiplicity) such that $[a_1], [a_2], \dots, [a_k], -[a_1], -[a_2], \dots, -[a_k]$ is the set of zeroes of h counted with multiplicity. We may also choose a sequence $[b_1], [b_2], \dots, [b_k]$ which does the same thing for the poles (and these sequences have the same length as the number of poles and zeroes are the same). Now put

$$g(z) := \frac{\prod_i \wp(z) - \wp(a_i)}{\prod_i \wp(z) - \wp(b_i)}.$$

This elliptic function has the same pole and zero sets (counted, of course, with multiplicities) as h, as the poles at $[0]$ from the different factors cancel, which means that $h(z)/g(z)$ has neither zeroes nor poles and is hence a constant which finishes the proof. □

Exercise 25. Give an alternative proof of the second part by producing a rational function in $\wp(z)$ that at all poles has the same "polar part" as the given even function.

The proof of Proposition 3.7 can be slightly modified to give another useful result which in particular further constrains the possible zero and pole sets of an elliptic function.

Proposition 3.11. *Let h be a non-zero elliptic function with periods ω_1 and ω_2 and let $\alpha_1, \alpha_2, \dots, \alpha_n \in E_{\omega_1,\omega_2}$ be the set of zeroes of h counted with multiplicity and*

$\beta_1, \beta_2, \dots, \beta_n \in E_{\omega_1,\omega_2}$ the set of poles of h counted with multiplicity. Then

$$\sum_i \alpha_i = \sum_i \beta_i.$$

Proof. This time we integrate zdh/h around the fundamental domain. This time the integrals along the opposite sides (and in opposite directions) do not cancel because of the z-factor. However, if we look at the opposite sides in the ω_1-direction (cf. Fig. 18) we get the contribution

$$\int_{\omega_2}^{\omega_1+\omega_2} z\frac{dh}{h} - \int_0^{\omega_1} z\frac{dh}{h} = \omega_2 \int_0^{\omega_1} \frac{dh}{h} + \int_0^{\omega_1} z\frac{dh}{h} - \int_0^{\omega_1} z\frac{dh}{h}.$$

Now, $\int_0^{\omega_1} \frac{dh}{h}$ is always $2\pi i$ times an integer (cf. Exercise 26) so that the contribution divided by $2\pi i$ is a period and similarly for the contribution in the ω_2-direction. As an element in E, the integral divided by $2\pi i$ then gives 0. On the other hand, if we deform the path of integration as in Proposition 3.7 we are left with a contribution from each zero or pole coming from the integration of zdh/h around a small circle with center at the zero or pole a. As in Proposition 3.7 we write h as $h(z) = (z-a)^n g(z)$ where g is non-zero at a and then $zdh/h = zn/(z-a) + z\,dg/g$ and the integral becomes $2\pi ina$. Dividing by $2\pi i$ and summing up in E using that the sum is zero gives the result. □

Exercise 26. Show that if C is a path and f is holomorphic in a neighbourhood of C with the same value at the two endpoints, then $\oint_C df/f$ is $2\pi i$ times an integer.

The importance of this result will be seen later but for the moment we shall give only one example of its use.

Example 5. Consider again the addition formula for $\wp(z)$. Start with two numbers $u, v \in \mathbb{C}$ which are distinct modulo Γ and for which neither u, v, nor $u+v$ is contained in Γ. We can find two complex numbers A and B such that

$$\wp'(u) + A\wp(u) + B = 0,$$
$$\wp'(v) + A\wp(v) + B = 0,$$

because if $\wp(u) = \wp(v)$, then $u - v$ or $u + v$ is in Γ by Theorem 3.8 which is not the case by assumption. This means that $f(z) := \wp'(z) + A\wp(z) + B$ is an elliptic function with a triple pole at [0] and no other poles. By construction it has zeroes at $[u]$ and $[v]$ which are distinct points. By Proposition 3.11, if w is the third zero, then $[u] + [v] + [w] = 0$ which means that $[w] = [-u-v]$ so that we also have

$$\wp'(-v-u) + A\wp(-u-v) + B = 0.$$

This however means that the matrix

$$\begin{pmatrix} \wp'(u) & \wp(u) & 1 \\ \wp'(v) & \wp(v) & 1 \\ \wp'(-u-v) & \wp(-u-v) & 1 \end{pmatrix}$$

has a non-zero null space and hence has determinant 0. What should we do if $[u] = [v]$? Well, a natural thing to do seems to insist that $\wp'(z) + A\wp(z) + B$ have a double root at u. This on the one hand means that the derivative should also be zero at u, i.e., $\wp''(u) + A\wp'(u) + B \cdot 0 = 0$. Then we get that the third root should fulfil $[2u + v] = 0$ so that $\wp'(-2u) + A\wp(-2u) + B = 0$ (assuming $2u \notin \Gamma$) and hence

$$\begin{vmatrix} \wp'(u) & \wp(u) & 1 \\ \wp''(u) & \wp'(u) & 0 \\ \wp'(-2u) & \wp(-2u) & 1 \end{vmatrix} = 0.$$

We can use the differential equation to further simplify this, as the following exercise shows, which also makes the previous calculation of the third intersection point with the line clearer.

Exercise 27. i) Use the differential equation to show that in the first case the three roots $\wp(u)$, $\wp(v)$, and $\wp(w)$ are roots of the polynomial $4x^3 - g_2x - g_3 - (Ax + B)^2 = 0$. Show that on the other hand $A(\wp(u) - \wp(v)) = \wp'(u) - \wp'(v)$ and conclude a formula for $\wp(u) + \wp(v) + \wp(w)$.

ii) Imitate the proof of the first part or pass to the limit to show that

$$\wp(2u) = -2\wp(u) + \frac{1}{4}\left(\frac{\wp''(u)}{\wp'(u)}\right)^2.$$

We shall finish this section by making a short discussion of how one, using more techniques than we are willing to admit in these notes, can answer the last question. Recall that the problem with the integral of dx/y was that, given a fixed starting point, its value did not depend only on the end point but also on the path taken to the end point. We can solve this problem by letting the integral take values not in $\mathbb{C}$ but in $\mathbb{C}$ divided by the subgroup of values of integrals along closed paths. This subgroup Γ is, as we have seen, generated by the two fundamental periods, integrals along two specific closed curves on the elliptic curve, $\overline{X}$ say. If we knew that these two periods were linearly independent over the reals, Γ would be a lattice in our sense and we would get a map $\overline{X} \to \mathbb{C}/\Gamma$. This is true and the map is a bijection. In the following exercises we sketch a proof of this fact.

Exercise 28. A real-valued function f on an elliptic curve $\overline{X}$ (over the complex numbers) is *harmonic* if for every point $(x, y) \in X$ there is a neighbourhood of it such that, if $y \neq 0$, f is the composite of a harmonic function defined in some open subset of $\mathbb{C}$ and x; if instead $y = 0$, f is the composite of a harmonic function defined in some open subset of $\mathbb{C}$ and y. Impose also similar conditions at ∞.

i) Show that a harmonic function on $\overline{X}$ is constant.

ii) Show that if the two fundamental periods of dx/y are linearly dependent over the reals, then there is a $0 \neq \lambda \in \mathbb{C}$ such that $\Re(\lambda \int dx/y)$ gives a well-defined function on $\overline{X}$. Show that it is harmonic and conclude that it is constant and get a contradiction from this.

Exercise 29. Show that the map $\overline{X} \to \mathbb{C}/\,\Gamma$ is a covering map and then conclude that it is an isomorphism.

4

A projective interlude

Our aim now is to see how much of what we have proven with analytic methods makes algebraic sense (and can be proven algebraically). Before we get seriously into that, let us first solve one problem. We saw that it was natural to add a point at infinity to the set of solutions to $y^2 = x^3 + ax + b$. When we did that we saw that the shape of the result was a torus which gives one justification for adding the point at infinity, as otherwise we would have gotten the torus minus one point. This choice was further vindicated when we saw that $\mathbb{C}/\Gamma_{\omega_1,\omega_2}$ was also a torus. One problem remains however. We have seen that the addition theorem is closely related to intersecting the elliptic curve with lines. This view-point can not naturally be fit into this framework. We shall now see that we may modify our constructions so that it does fit. In order to start being more algebraic let us consider an arbitrary field K.

The *projective plane*, $\mathbb{P}^2(K)$, over K is the set of 1-dimensional subspaces of K^3. Any such subspace has a basis element $(x, y, z) \neq 0$ and two such vectors (x, y, z) and (x', y', z') give rise to the same subspace exactly when there is a non-zero scalar λ such that $(x', y', z') = (\lambda x, \lambda y, \lambda z)$. This defines an equivalence relation on the set of non-zero vectors in K^3 and the projective plane is identified with the set of equivalence classes under this equivalence relation. The equivalence class containing (x, y, z) is denoted $(x : y : z)$ and is called a *homogeneous* coordinate of the element (or *point* as we shall also call the elements of the projective plane).

We may embed the ordinary, or *affine*, plane, K^2, in the projective plane by $(x, y) \mapsto (x : y : 1)$. The image consists exactly of the homogeneous coordinates $(x : y : z)$ for which $z \neq 0$ and then $(x : y : z) = (x/z : y/z : 1)$ so that $(x : y : z) \mapsto (x/z, y/z)$ is the inverse map. Over the reals this embedding can be easily (and nicely) envisaged. One considers the plane $z = 1$ which is parallel to the xy-plane. Then every line through the origin that does not lie in the xy-plane meets this plane at a unique point (cf. Fig. 20). This intersection point is of course exactly the representative of the form $(x : y : 1)$ and the lines in the xy-plane have homogeneous coordinates $(x : y : 0)$. Now, if ℓ is a line in the plane $z = 1$, then the union of all the 1-dimensional spaces intersecting the plane in that line is the intersection of the plane with a 2-dimensional subspace (cf. Fig. 20). Hence it is natural to say that a *line* in the projective plane is the set of 1-dimensional subspaces that lie in a given 2-dimensional subspace. Note that such a linear subspace, being of codimension 1, is of the form $\{(x : y : z) \mid ax + by + cz = 0\}$ for some $(a, b, c) \neq 0$.

Example 6. i) When do three points, $(x_i : y_i : z_i)$, $i = 1, 2, 3$ in the projective plane lie on a line? Well, that happens precisely when they span a subspace of dimension ≤ 2, i.e., when they are linearly dependent. That in turn is equivalent with the vanishing of

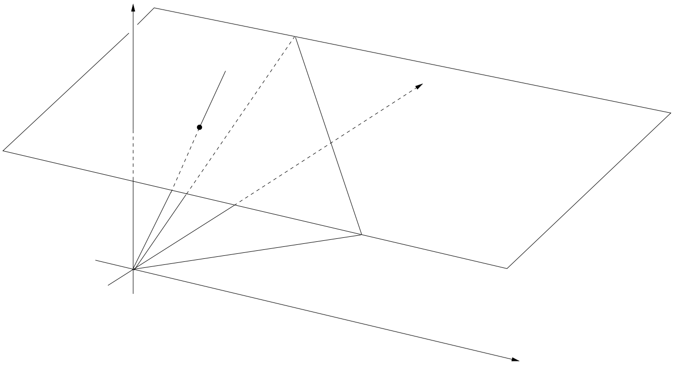

Figure 20. The real projective plane.

the determinant

$$\begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{vmatrix}.$$

When $z_i = 1$ for all i, this is recognisable as the condition used in the second proof of the addition formula.

ii) Consider the projective plane over the field $\mathbb{Z}/2$ of two elements. There is a unique non-zero vector in each 1-dimensional subspace so that there are in all $2^3 - 1 = 7$ points in this projective plane which is also called the *Fano plane*.[1] Similarly there are 7 lines with $2^2 - 1 = 3$ points on each of them. We can draw the following picture of it (see Fig. 21) where there are seven points and the lines are the points that lie on a line in the picture or on the circle.[2]

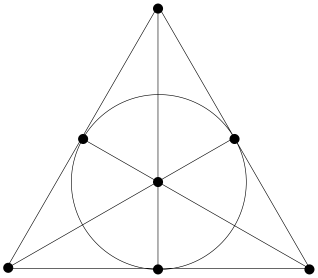

Figure 21. The Fano plane, $\mathbb{P}^2(\mathbb{Z}/2)$.

[1] Gino Fano, 1871–1952

[2] It can be shown that the Fano plane can not be realised in the real plane by points and lines alone.

iii) We may similarly introduce the *projective line* whose points are the 1-dimensional subspaces of K^2. Here one embeds the usual line by $x \mapsto (x:1)$ and the image consists of those $(x:z)$ with $z \neq 0$ and the inverse map is given by $(x:z) \mapsto x/z$. There is only one point outside it, $\infty := (1:0)$. The Riemann sphere then can be identified with $\mathbb{P}^1(\mathbb{C})$.

iv) It is clear that one cannot in general reduce a rational number modulo a prime p as the denominator may be divisible by p. However, a point in $\mathbb{P}^1(\mathbb{Q})$ can be reduced; a non-zero pair (x, y) of rational numbers can be written as $r(m, n)$ where r is a non-zero rational number and m and n are relatively prime integers. This decomposition is unique except that $r(m, n) = -r(-m, -n)$. We then have that $(x:y) = (m:n)$ and the reduction modulo p, $(\bar{m}:\bar{n})$ is well defined (i.e., $(\bar{m}, \bar{n})$ is not the null vector and $(\bar{m}:\bar{n})$ depends only on $(x:y)$). Note that if $r \in \mathbb{Q}$ we may consider $(r:1) \in \mathbb{P}^1(\mathbb{Q})$ which then may be reduced modulo p. It is easy to see that if $r = m/n$ with m and n relatively prime integers, then this reduction is $(\bar{m}/\bar{n}:1)$ if $\bar{n} \neq 0$ and ∞ if $\bar{n} = 0$. Similarly, one may reduce a point in $\mathbb{P}^2(\mathbb{Q})$ to a point in $\mathbb{P}^2(\mathbb{Z}/p)$.

Exercise 30. i) Introduce projective n-space for an arbitrary $n \geq 1$.

ii) Show that an invertible linear transformation $g \in \mathrm{GL}_{n+1}(K)$ of K^{n+1} acts on projective n-space (giving rise to a *projective transformation*).

iii) Show that the elements of $\mathrm{GL}_{n+1}(K)$ acting trivially on $\mathbb{P}^n(K)$ are the scalar matrices.

iv) Show that if we write $\mathbb{P}^1(K)$ as $K \cup \{\infty\}$, then $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ acts by the Möbius transformation $x \mapsto (ax + b)/(cx + d)$.

There is of course nothing special about the last coordinate. One may instead look at the plane $x = 1$ or $y = 1$ which embed as the set of homogeneous coordinates with $x \neq 0$ or $y \neq 0$. They are called the *coordinate patches* and the mappings to the affine plane are given by $(x:y:z) \mapsto (y/x, z/x)$ resp. $(x:y:z) \mapsto (x/y, z/y)$. Note that a point may lie in several patches and then it corresponds to a different point in the affine plane under the different coordinate patches (cf. Fig. 22). For instance going from the $x \neq 0$ to the $z \neq 0$ patch is given by $(x, y) \mapsto (1:x:y) \mapsto (1/y, x/y)$.

Example 7. If we consider instead the projective line, we have two patches given by $x \neq 0$ and $y \neq 0$. The points on the affine line which also lie in the other patch are the non-zero numbers and the transformation from one patch to the other is given by $x \mapsto (x:1) \mapsto 1/x$. This is exactly the transformation we considered previously when we were considering the relations between the two parts of $\overline{\mathbb{C}}$, which thus deserves to also be denoted $\mathbb{P}^1(\mathbb{C})$.

We now consider a point $(x:y:z)$ in the affine patch $z \neq 0$ which as such lies in $X := \{(x, y) \mid y^2 = x^3 + ax + b\}$. The affine point is $(x/z, y/z)$, so that means that $(y/z)^2 = (x/z)^3 + ax/z + b$. We can clear denominators in a minimal fashion to get the equivalent equation $y^2z = x^3 + axz^2 + bz^3$. This equation then actually makes sense on all of the projective plane, as all the monomials appearing in it have the same degree, so whether or not (x, y, z) fulfills it is equivalent with

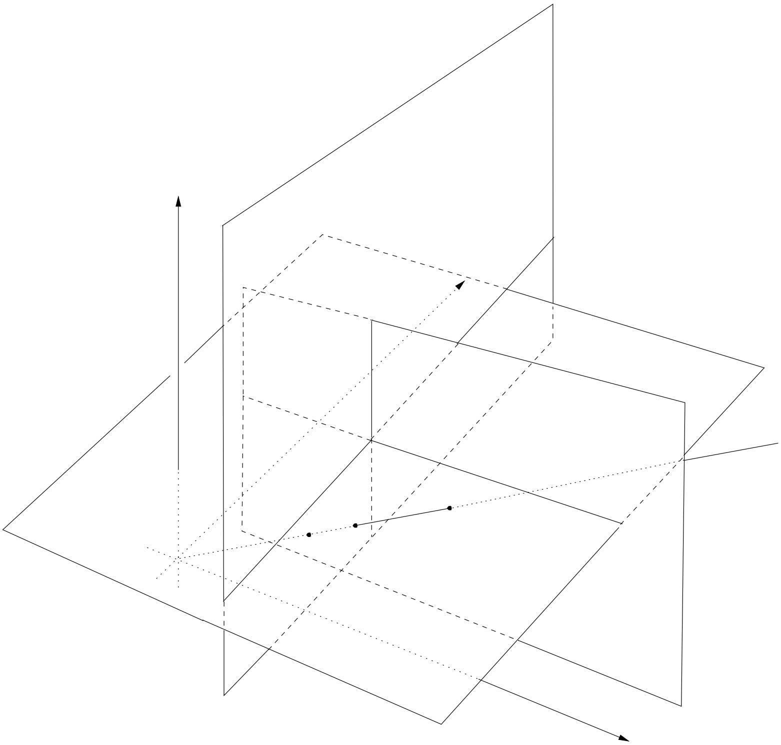

Figure 22. The three coordinate patches of the projective plane.

whether or not $(\lambda x, \lambda y, \lambda z)$ does, where $0 \neq \lambda \in K$. This means that we can define $\overline{X} := \{(x:y:z) \in \mathbb{P}^2(K) \mid y^2z = x^3 + axz^2 + bz^3\}$ which is called the *projective completion* of X. Then the intersection of $\overline{X}$ with the affine patch $z \neq 0$ will be X. To investigate the complement of X in $\overline{X}$ we simply put $z = 0$ giving the equation $x^3 = 0$, i.e., $x = 0$ and hence the complement consists only of the point $(0:1:0)$. This shows that the $\overline{X}$ adds one point to X just as the $\overline{X}$ we have considered previously. That we get another way to add a point at infinity is maybe not so impressive but things become more interesting when we look at the intersection of $\overline{X}$ with a line. We have seen that each line in the affine patch $z \neq 0$ is the intersection of that patch with a projective line given as $\{(x:y:z) \mid ax + by + cz = 0\}$ for some $(a, b, c) \neq 0$. We get an extra (i.e., outside of the affine patch) intersection point precisely when the only extra point of $\overline{X}$ lies on the line, i.e., when $a \cdot 0 + b \cdot 1 + c \cdot 0 = 0$, i.e., when $b = 0$. This gives two possibilities. One is that $a = 0$ and then $c \neq 0$ so the equation becomes $z = 0$. In that case we have seen that $(0:1:0)$ is the only point on the intersection. The other is that $a \neq 0$ and then the equation has the form $x = -c/az$ and to get the intersection points with $\overline{X}$ we only have to look at the affine patch (because we already know that $(0:1:0)$ lies on it) and a point $(x:y:1)$ lies on the line precisely when $x=-a/b$, i.e., the lines in question are those for which x is constant. This fits very well with what we were doing when we were discussing the addition theorem where we postulated that lines parallel with the y-axis intersected the curve in the point at infinity (apart from the intersection with X). Furthermore, we also introduced a fictitious line that only intersected $\overline{X}$ in the point at infinity. From the point of view of the projective plane these postulations become facts. Furthermore, we can make more aspects of that discussion precise. Consider for this a line ℓ. If we choose two points $(x_1:y_1:z_1)$ and $(x_2:y_2:z_2)$ on it we can parametrise the line $\mathbb{P}^1(K) \xrightarrow{\sim} \ell$ by $(s:t) \mapsto (sx_1+tx_2:sy_1+ty_2:sz_1+tz_2)$.

Exercise 31. i) Show that given an ordered set of three points on a line $\ell \subseteq \mathbb{P}^2(K)$ there is a unique parametrisation that takes $0=(0:1)$, $1=(1:1)$, and $\infty=(1:0)$ to the points in the given order.

ii) Show that for an ordered set of three distinct points in $\mathbb{P}^1(K)$ there is a unique Möbius transformation (cf. Exercise 30) taking them in the given order to 0, 1, and ∞.

If we want to determine which parameter values give values in $\overline{X}$, then we should substitute $x=sx_1+tx_2$, $y=sy_1+ty_2$, and $z=sz_1+tz_2$ in the equation $y^2z=x^3+axz^2+bz^3$ or better yet in the polynomial $f(x,y,z)=y^2z-(x^3+axz^2+bz^3)$. This gives a polynomial $F(s,t)=f(sx_1+tx_2,sy_1+ty_2,sz_1+tz_2)$. It is clearly *homogeneous* of degree 3 (i.e., all its monomials are of degree 3). The homogeneity implies that the condition $F(s,t)=0$ depends only on the homogeneous coordinate $(s:t)$ (as it should). By construction the zeroes of F, seen as points on $\mathbb{P}^1(k)$ and then as points on ℓ, are the points of intersection between ℓ and $\overline{X}$ but we may also attach a multiplicity to each such point.

Exercise 32. i) Show that every homogeneous polynomial $F(s,t)\neq 0$ over an algebraically closed field factors as $\prod_i(b_is-a_it)^{n_i}$, where the points $(a_i:b_i)$ are the distinct zeroes of F and the *multiplicities* n_i are uniquely determined by the zero $(a_i:b_i)$. Show in particular that if we choose another parametrisation, then the multiplicity of a zero is the same.

ii) Show that if $f(x,y,z)$ is a non-zero homogeneous polynomial, then if $f(sx_1+tx_2,sy_1+ty_2,sz_1+tz_2)$ is the zero-polynomial for two distinct points $(x_1:y_1:z_1)$ and $(x_2:y_2:z_2)$ on a line ℓ, then f is divisible by an equation $ax+by+cz$ for ℓ.

It is clear from this exercise that we can attach a multiplicity to each intersection point between $\overline{X}$ and a line (once we have shown no linear polynomial divides $y^2z-(x^3+axz^2+bz^3)$). It is also clear that the number of intersection points counted with multiplicity equals 3. This does in fact take care of the special cases we encountered in the proof of the complete addition formula (some of which were left to Exercise 24). However, in order to see what these special cases mean we shall need not just to accept the possibility of having multiple intersection points, we shall also need to enforce it.

Example 8. Let $p = (x : y : z) \in \overline{X}$ and assume that it is an intersection point of $\overline{X}$ and the line ℓ. In order to investigate what this means let us parametrise ℓ by choosing $(x : y : z)$ and another point $(x' : y' : z')$ on ℓ. Hence the polynomial to consider is $f(sx + tx', sy + ty', sz + tz')$, where $f(x, y, z) = y^2z - (x^3 + axz^2 + bz^3)$. Now, the multiplicity p in the intersection is equal to the power by which t divides this polynomial. We may expand this using Taylor's formula.

Exercise 33. Prove that with the formal definition of partial derivatives of a polynomial, Taylor's formula is valid for polynomials over any field.

We get

$$\begin{aligned} &f(sx + tx', sy + ty', sz + tz') \\ &\qquad = s^3 f(x, y, z) + s^2 t(f'_x(x, y, z)x' + f'_y(x, y, z)y' + f'_z(x, y, z)z') + O(t^2), \end{aligned}$$

and as $f(x, y, z) = 0$ we get that p appears with multiplicity ≥ 2 precisely when $f'_x(x, y, z)x' + f'_y(x, y, z)y' + f'_z(x, y, z)z' = 0$. We have

$$\begin{aligned} f'_x(x, y, z) &= -3x^2 - az^2, \\ f'_y(x, y, z) &= 2yz, \\ f'_z(x, y, z) &= y^2 - 2axz - 3z^2. \end{aligned}$$

Assume now that the characteristic of the field K is different from 2. Then these partial derivatives are not all equal to zero because if they are we get that either $z = 0$ or $y = 0$ from $f'_y = 0$. If $z = 0$ we have seen that $p = (0 : 1 : 0)$, as it is the only such point on $\overline{X}$, but $f'_z(0, 1, 0) = 1$. If $y = 0$ and $z \neq 0$ we may scale the homogeneous coordinate for p so that $z = 1$, and then we get $3x^2 + a$ from the x-derivative and $x^3 + ax + b$ from the fact that we have a point in X. Together these would force $x^3 + ax + b$ to have a multiple root, which is excluded by assumption. Hence we get only a single line for which the multiplicity is ≥ 2, namely the one given by the equation $f'_x(x, y, z)x' + f'_y(x, y, z)y' + f'_z(x, y, z)z' = 0$. This (for rather obvious reasons) will be called the *tangent* to $\overline{X}$ at p.

Exercise 34. i) Show that if $f(x, y, z)$ is homogeneous of degree n, then $xf'_x(x, y, z) + yf'_y(x, y, z) + f'_z(x, y, z) = nf(x, y, z)$ ("Euler's formula").

ii) Show that if $f(x, y, 1) = 0$ and $f'_x(x, y, 1) = f'_y(x, y, 1) = 0$, then also $f'_z(x, y, 1) = 0$.

iii) Show that if $f(x, y, 1) = 0$ and not all partial derivatives of f in $(x, y, 1)$ are zero, then $t \mapsto (x + tf'_y(x, y, 1) : y - tf'_x(x, y, 1) : 1)$ is a parametrisation of the tangent of f at $(x : y : 1)$.

We are now prepared to investigate the more precise significance of three points on X lying on a line.

5

The group structure on an elliptic curve

Let us fix two fundamental periods $\omega_1, \omega_2 \in \mathbb{C}$ (which thus are assumed to be linearly independent over the reals) and let

$$\overline{X} := \{(x:y:z) \in \mathbb{P}^2(\mathbb{C}) \mid y^2z = 4x^3 - g_2(\omega_1,\omega_2)xz^2 - g_3(\omega_1,\omega_2)z^3\}.$$

We have defined a map $\mathbb{C} \to \overline{X}$ given for $z \notin \Gamma_{\omega_1,\omega_2}$ by the formula $z \mapsto (\wp(z) : \wp'(z) : 1)$ and mapping Γ to the point at infinity. The latter definition seemed rather ad hoc but we shall now see that this extension to the point at infinity is obtained by including "continuity" in the generic definition. Thus we pick $\gamma \in \Gamma$ and note that when $z \notin \Gamma$ we have

$$(\wp(z):\wp'(z):1) = ((z-\gamma)^3\wp(z):(z-\gamma)^3\wp'(z):(z-\gamma)^3).$$

Now, both $(z-\gamma)^3\wp(z)$ and $(z-\gamma)^3\wp'(z)$ have holomorphic extensions to $z=\gamma$, the first one with a zero at γ and the second one with value -2 at γ. Hence it seems very natural to extend the map to γ by putting it equal to $(0:-2:0) = (0:1:0)$ which is exactly the point at infinity of $\overline{X}$ (which of course was what we did in our first definition). The following exercise shows that this extension really is an extension by continuity.

Exercise 35. Say that a subset of $\mathbb{P}^2(\mathbb{C})$ is *open* precisely when its intersection with the affine coordinate patches are open.

i) Show that a subset of $\mathbb{P}^2(\mathbb{C})$ is open precisely when its inverse image under the map $\mathbb{C}^3 \setminus \{0\} \to \mathbb{P}^2(\mathbb{C})$ given by $(x,y,z) \mapsto (x:y:z)$ is open.

ii) Show that the map

$$\mathbb{C}\setminus\Gamma \to \mathbb{P}^2(\mathbb{C}), \quad z \mapsto (\wp(z):\wp'(z):1),$$

has a unique continuous extension to $\mathbb{C}$ and that it takes Γ to $(0:1:0)$.

The formulation of the addition theorem that takes care of all the special cases is then: *The images of $\alpha, \beta, \gamma \in \mathbb{C}/\Gamma$ under the map to $\overline{X}$ are the three intersection points (counted with multiplicity) of $\overline{X}$ with a line precisely when $\alpha+\beta+\gamma=0$.* The map $\mathbb{C}/\Gamma \to \overline{X}$ is a bijection and what we have found is that the group structure on $\mathbb{C}/\Gamma$ seems intimately connected with the geometry of $\overline{X}$. In fact it seems that we should be able to define the group structure induced on $\overline{X}$ by that on $\mathbb{C}/\Gamma$. This turns out to be true and follows from just two conditions:

- $\infty = (0:1:0)$ is the identity element.
- If $p_1, p_2, p_3 \in \overline{X}$ are the three intersection points (counted with multiplicity) of a line ℓ and $\overline{X}$, then $p_1+p_2+p_3=0$.

Let us show how this determines the group structure. If p_1 and p_2 are two points on $\overline{X}$, then we draw a line through them (or the tangent to $\overline{X}$ at them if they are equal) and let p_3 be the third intersection point. Then we have $p_1 + p_2 + p_3 = 0$, i.e., $p_3 = -(p_1 + p_2)$ so that in order to get the sum of $p_1 + p_2$ we need to find the inverse of p_3. For this, if $p_3 = \infty$ we note that it is its own inverse, being the identity element, otherwise we draw the line through p_3 and ∞ (i.e., the line given by its x-coordinate being equal to that of p_3) and let p_4 be the third point of intersection. Then we have $p_3 + p_4 = \infty + p_3 + p_4 = 0$ so that $p_4 = -p_3 = p_1 + p_2$ (cf. Fig. 23). Note that this prescription makes sense over any field (not of characteristic 2 as we saw that tangents might not work then). It is natural here to accept any cubic polynomial $p(x) = ax^3 + bx^2 + cx + d$ without double roots so that for instance $4x^3 - g_2x - g_3$ is accepted along with $x^3 + ax + b$. However, we may make a (linear) coordinate change $x \mapsto \lambda x$ and $y \mapsto \mu y$, $\lambda, \mu \neq 0$. This changes the equation $y^2 = ax^3 + bx^2 + cx + d$ to $y^2 = \mu^{-2}\lambda^3 ax^3 + \mu^{-2}\lambda^2 bx^2 + \mu^{-2}\lambda cx + \mu^{-2}d$. (We have also made the simplification to deal only with the non-homogeneous equation, where $z = 1$, which makes it trivial to pass back and forth between the homogeneous and non-homogeneous versions.) In particular we may put $\mu = \lambda = a^{-1}$ which reduces to $a = 1$. This also means that there would be no loss of generality if we assumed $a = 1$ (though it would somewhat hide the relations with $\wp$). If the characteristic is different from 3, we can also make a coordinate change $x \mapsto x - b/3$ which gets rid of the quadratic term, hence somewhat justifying that we have only used the form $x^3 + ax + b$ for our polynomial (we shall return to what happens in the case of characteristic 3 – as well as 2).

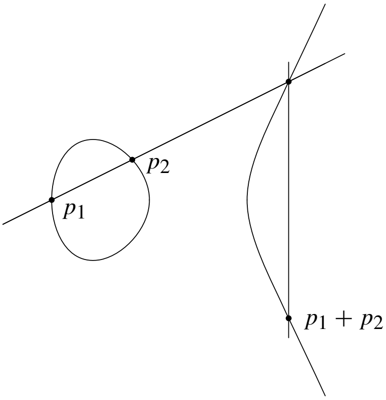

Figure 23. Addition on an elliptic curve.

Let us now look at formulas. If one of the points is ∞, the sum is just the other point. Hence we may consider (r, s) and (u, v) fulfilling $s^2 = r^3 + ar + b$ and $v^2 = u^3 + au + b$. Assume first that they are distinct points. If $u = r$, then $v = -s$

and their sum is ∞ so assume that $u \neq r$. In that case we have already performed the calculation and the sum is (x, y) defined by

$$\begin{aligned} x &= \frac{(v-s)^2}{(u-r)^2} - u - r, \\ y &= \frac{v-s}{u-r}(x-r) + s. \end{aligned}$$

If instead the points are the same we should use the tangent as the line and it has the parametric form $(x, y) = (r + t(2s), s - t(3r^2 + b))$. Substituting gives

$$(s - t(3r^2 + b))^2 = (r + 2st)^3 + a(r + 2st) + b.$$

Here $t = 0$ is a double root and the relation between the coefficients and the roots can be used here giving the third root

$$t = \frac{b^2 + 6br^2 + 9r^4 - 12rs^2}{8s^3}.$$

This gives

$$\begin{aligned} x &= r + \frac{b^2 + 6br^2 + 9r^4 - 12rs^2}{4s^2}, \\ y6 &= s - \frac{(3r^2 + b)(b^2 + 6br^2 + 9r^4 - 12rs^2)}{8s^3}. \end{aligned}$$

This could be further reduced by using $s^2 = r^3 + ar + b$. Figure 24 gives an example where one starts with one point and then takes multiples of it. Note that the inverse of a point $(x, y) \in X$ is given by $(x, -y)$.

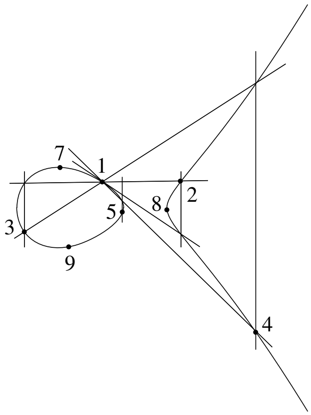

Figure 24. A point and some of its multiples.

We have not actually proved that the binary operation thus defined gives us a commutative group. For those elliptic curves over the complex numbers that actually come from doubly periodic functions, it is clear as we have seen that the binary operation corresponds to addition of complex numbers modulo a lattice. We shall now sketch a proof of the fact that it is true in general.

Proposition 5.1. *Let $y^2 = x^3 + ax + b$ define an elliptic curve over a field K (of characteristic different from 2). Then the binary operation defined above gives the solutions of $y^2 = x^3 + ax + b$ plus the point at infinity the structure of a commutative abelian group.*

Proof. The only non-trivial part is the associativity and we shall sketch one of several possible proofs of it. If $K = \mathbb{C}$ and there are fundamental periods ω_1 and ω_2 such that $g_2(\omega_1, \omega_2) = -4a$ and $g_4(\omega_1, \omega_2) = -4b$, then associativity is clear since the operation corresponds to addition in $\mathbb{C}/\Gamma_{\omega_1,\omega_2}$. We shall show later that we can choose ω_1 and ω_2 such that $-g_2(\omega_1, \omega_2)/4$ and $-g_4(\omega_1, \omega_2)/4$ are algebraically independent (i.e., fulfills no non-trivial polynomial in two variables with rational coefficients). Assume now that none of the three points u, v, and w involved in proving associativity equal 0, nor do any of $u + v$, $v + w$, $u + (v + w)$ or $(u + v) + w$ equal zero, nor do $u = v$, $v = w$, $u = v + w$, or $u + v = w$ (these special cases can be handled either directly or in a similar fashion). The associativity then amounts to equality between two expressions which are rational functions in the coordinates of u, v, and w together with a and b and where the equality is modulo the relations giving that u, v, and w are points on the curve. We may clear out the denominators so that we get an equality of polynomials or that certain polynomials should vanish. These polynomials have integer coefficients and all are still modulo relations of the type $y^2 = x^3 + ax + b$, where x and y are the coordinates of one of u, v, and w. We may use these relations to get rid of all the higher powers of the second coordinates of all the points, so that we get an expression q which is a polynomial in these second coordinates of degree ≤ 1 in each of them and whose coefficients are integer polynomials in the first coordinates and a and b. We shall now show that these coefficient polynomials are each equal to zero, which means that the whole expression is zero when evaluated at any element over any field.

To prove that they are zero as polynomials we first choose ω_1 and ω_2 so that the corresponding a and b are algebraically independent. Then we choose u, v, and w such that their first coordinates together with a and b are algebraically independent (this is always possible as $\mathbb{C}$ is algebraically closed and has infinity transcendence degree over the rational numbers). Then as the result is true for these particular values the polynomials vanish for them, but they are algebraically independent (there is a small subtlety in that we only get that q becomes zero and need to conclude that also its coefficients are zero, but one can use algebraic independence to show that). □

Exercise 36. Consider the elliptic curve defined by $y^2z = x^3 + xz^2 + 2z^3$ over the field $\mathbb{Z}/5$. Write down the points (over $\mathbb{Z}/5$) on it and determine the group structure.

Knowing that the *group of complex points* for a complex elliptic curve associated to the fundamental periods ω_1 and ω_2 is isomorphic to $\mathbb{C}/\,\Gamma_{\omega_1,\omega_2}$ doesn't in general tell you very much. One piece of information that can be extracted from this description is however very interesting. One can namely use it to determine the structure of the kernel of multiplication by a positive integer n: We have that $n[z] = 0$ precisely when $nz \in \Gamma$, i.e., $z \in 1/n\Gamma$ so that the kernel by multiplication by n is given by $1/n\Gamma/\,\Gamma$. Now, it is clear that the map $(\mathbb{Z}/\mathbb{Z}n)^2 \to 1/n\Gamma/\,\Gamma$ which takes (r,s) to $r\omega_1/n + s\omega_2/n$ mod Γ is an isomorphism. This is interesting because multiplication by n is also described by polynomial functions in our algebraic model.

Example 9. i) $2p = 0$ means that $p = -p$ and if $p \neq \infty$ then that means that the y-coordinate of p is 0, and then the x-coordinate is a root of $x^3 + ax + b$. By assumption there are three such roots and hence in all four solutions of the equation $2p = 0$. The structure of the group is then necessarily $(\mathbb{Z}/2)^2$.

ii) $3p = 0$ means that $2p = -p$. If we assume that $p \neq \infty$ and write $p = (r,s)$ as above, this means

$$\begin{cases} r = r + \frac{b^2+6br^2+9r^4-12rs^2}{4s^2}, \\ s = -s + \frac{(3r^2+b)(b^2+6br^2+9r^4-12rs^2)}{8s^3}, \end{cases}$$

the first of which gives $0 = b^2+6br^2+9r^4-12rs^2 = b^2+6br^2+9r^4-12r(r^3+ar+b)$ which is equivalent to $3r^4 + (6b-12a)r^2 - 12br + b^2 = 0$. In characteristic different from 3 this gives four different values for r (and one can show that there are no multiple roots) and then $s^2 = r^3 + ar + b$ gives two values for s for each value of r (as r being a root of $r^3 + ar + b$ leads to a point of order 2). This seems to give in all $2 \cdot 4 + 1 = 9$ solutions to $3p = 0$ (what could go wrong are multiple roots for r). In characteristic 3 things go wrong and there are only three or one solutions (depending on a and b).

6

Equivalence

We shall now see that given an elliptic curve one can find other elliptic curves which are essentially the same, i.e., we shall arrive at a notion of equivalent elliptic curves. We shall also see a similar thing for fundamental periods where equivalent fundamental periods will give rise to essentially the same theory of elliptic functions. We shall then compare these two notions.

6.1 Equivalences of elliptic curves

If we have an elliptic curve with affine equation $y^2 = f(x)$ where this time f is a polynomial of degree 3 or 4 without multiple roots, then we may introduce a new variable x' related to x by a Möbius transformation $x = (ax' + b)/(cx' + d)$, where $ad \neq bc$. This gives the equation $y^2 = f((ax'+b)/(cx'+d))$ which we may rewrite as $y'^2 = g(x')$ where $g(x') = (cx'+d)^4 f((ax'+b)/(cx'+d))$ and $y' = (cx'+d)^2 y$. Here, of course, g is a new polynomial of degree 3 or 4. This gives a bijection between the solutions to the two equations except possibly for exceptions such as $cx' + d = 0$ and at points at infinity. To understand what is happening at those points it is convenient to homogenise our situation another way. We put $f(x, z) := z^4 f(x/z)$ which is a homogeneous polynomial of degree 4 and then consider the equation $y^2 = f(x, z)$. Now, f is not a well-defined function on the projective line so this equation must be interpreted properly: We define the *weighted projective space* (of weights $(1, 2, 1)$) as the set of equivalence classes of elements of $K^3 \setminus \{(0, 0, 0)\}$ under the equivalence relation $(x, y, z) \sim (\lambda x, \lambda^2 y, \lambda z)$ for all $0 \neq \lambda \in K$. Then $y^2 = f(x, z)$ defines a subset $\overline{X}$ of the weighted projective space. Projection onto the x- and z-coordinates gives a well-defined mapping $X \to \mathbb{P}^1$ since if $x = z = 0$, then also $y = 0$ from the equation. Over a point $(x : 1) \in \mathbb{P}^1$ we get the equation $y^2 = f(x, 1) = f(x)$ hence the usual finite part of the elliptic curve (when the degree of f is 3). Over $(1 : x)$ we get the equation $y^2 = f(1, x) = x^4 f(1/x)$ which when $x \neq 0$ gives $(x^{-2}y)^2 = f(1/x)$, but $(1, y, x) \sim (1/x, x^{-2}y, 1)$ which is the coordinate change we already have considered. Finally for $(1:0)$ we get $y^2 = 0$ if the degree of f is 3 and an equation $y^2 = a$, where a is the highest coefficient of f, if not. When f is of degree 3 we recover our earlier picture with one point at infinity added to the affine picture, and in the case of degree 4 we instead get two points at infinity.

It is now much easier to perform the Möbius transformation, in fact we simply apply the corresponding (cf. Exercise 30) linear transformation giving $(x, z) = (ax' + bz', cx' + dz')$ to f and get $y^2 = f(ax' + bz', cx' + dz') = g(x', z')$ where as before

$g(x') = (cx' + d)^4 f((ax' + b)/(cx' + d))$. Note that we get back our original form if we start with a solution of the form $(x, y, 1)$, transform it as above to $(ax' + b, y, cx' + d)$ and then get it back to the same form (provided $cx' + d \neq 0$) which gives $((ax'+b)/(cx'+d), (cx'+d)^{-2}y, 1)$. In this way we get a bijection between the points (including the points at infinity) of the elliptic curve given by $y^2 = f(x)$ and the curve given by $y'^2 = g(x')$. (Here we use the name "elliptic curve" even when the degree of the polynomial is 4. We shall see that this mostly makes sense.) It is not difficult to show (though not completely obvious as we work with a homogenisation that is not suited to describing lines in the projective plane) that this bijection respects the group structure as well. Hence it is reasonable to say that the curves given by $y^2 = f(x, z)$ and by $y^2 = g(x, z)$ are equivalent[1] if there are $a, b, c, d \in K$ with $ad \neq bc$ and $g(x, z) = f(ax + bz, cx + dz)$. How does one then classify elliptic curves up to equivalence? Well, first one can look at the zeroes of f in $\mathbb{P}^1(K)$ (we assume now that K is algebraically closed so that f has four zeroes in $\mathbb{P}^1(K)$). These zeroes determine the equivalence class of f since any other polynomial with the same zeroes will differ under multiplication by a non-zero constant and $f(\mu x, \mu z) = \mu^4 f(x, z)$ (and K is algebraically closed so that all its elements are fourth powers). If we transform f by $(x, z) = (ax'+bz', cx'+dz')$, then the zeroes transform by the Möbius transformation $x' = (ax + b)/(cx + d)$ and hence the equivalence classes of elliptic curves are in bijection with the equivalence classes (i.e., orbits) of four unordered points of $\mathbb{P}^1(k)$ under the action of the group of Möbius transformations. If we first look at four *ordered* points we have the following result.

Exercise 37. Show that given three distinct points $a, b, c \in \mathbb{P}^1(K)$, there is a unique Möbius transformation g such that $g(a) = \infty$, $g(b) = 0$, and $g(c) = 1$.

From this exercise it follows that for distinct points $a, b, c, d \in \mathbb{P}^1(K)$ there is a unique Möbius transformation taking the three first to $\infty, 0, 1$ (in that order) and then by definition the fourth is taken to the *cross ratio*,[2] $\{a, b; c, d\}$.

It follows from this exercise that we may begin by assuming that one of the roots of $f(x, z)$ is at ∞. This is equivalent to $f(x) = f(x, 1)$ being of degree 3 and justifies our (implicit) assertion that we may reduce to the case that the degree of f is 3. If we continue we may put two other roots at 0 and 1 and then the polynomial f becomes (once we have scaled it so that it becomes monic) $f(x) = x(x - 1)(x - \lambda)$. We have thus shown that every elliptic curve is equivalent to one of the form $E_\lambda\colon y^2 = x(x - 1)(x - \lambda)$ for some $\lambda \neq 0, 1$. Now this is not a complete classification, as we have chosen a specific order in which to take the roots. If we choose another order we get a different λ.

Exercise 38. Show that if one changes the order in which the distinct points $a, b, c, d \in \mathbb{P}^1(K)$ are taken, then $\lambda := \{a, b; c, d\}$ is changed into one of λ, $1-\lambda$, $1/\lambda$, $1/(1-\lambda)$, $\lambda/(\lambda - 1)$, or $(\lambda - 1)/\lambda$.

[1] Actually isomorphic once one has introduced enough notions to make sense of that.

[2] There are some differing conventions about the order in which one should take $\infty, 1, 0$.

Using this exercise one obtains a condition for when two curves E_λ and $E_{\lambda'}$ are equivalent. In a subsequent subsection we shall find an explicit way of deciding when this happens. First we shall however consider the question of equivalence of fundamental periods.

Exercise 39. i) Show that the Möbius transformations that preserve ∞ are the *affine transformations*, $x \mapsto ax + b$, $a \neq 0$.

ii) Show that an affine transformation of one elliptic curve associated to a degree 3 polynomial into another is induced by a projective transformation of the projective plane.

6.2 Equivalence of periods

Given a pair of fundamental periods ω_1 and ω_2, whether or not they are periods of a meromorphic function depends only on the subgroup $\Gamma_{\omega_1,\omega_2}$. Hence, if ω_1' and ω_2' is another pair of generators of $\Gamma_{\omega_1,\omega_2}$, then the elliptic functions with respect to ω_1' and ω_2' are the same as those for ω_1 and ω_2. That ω_1' and ω_2' belong to $\Gamma_{\omega_1,\omega_2}$ means that there are integers a, b, c, d such that $\omega_1' = a\omega_1 + b\omega_2$ and $\omega_2' = c\omega_1 + d\omega_2$ and the fact that they are generators means that ω_1 and ω_2 may be similarly expressed in terms of ω_1' and ω_2'. This in turn is easily seen to be equivalent to $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ being an invertible integer matrix, i.e., belonging to $\mathrm{GL}_2(\mathbb{Z})$. On the other hand, there is a very simple relation between elliptic functions for ω_1 and ω_2 and elliptic functions for $\lambda\omega_1$ and $\lambda\omega_2$, where $\lambda \in \mathbb{C}^\times$: If f is elliptic for $\lambda\omega_1$ and $\lambda\omega_2$, then $g(z) := f(\lambda z)$ is elliptic for ω_1 and ω_2 and if f is elliptic for ω_1 and ω_2, then $g(z) := f(\lambda^{-1}z)$ is elliptic for $\lambda\omega_1$ and $\lambda\omega_2$. Note that applying this to the Weierstrass $\wp$-function and its derivative almost gives back functions of the same type: $\wp(\lambda z|\lambda\omega_1, \lambda\omega_2) = \lambda^{-2}\wp(z|\omega_1, \omega_2)$ and $\wp'(\lambda z|\lambda\omega_1, \lambda\omega_2) = \lambda^{-3}\wp'(z|\omega_1, \omega_2)$. Furthermore, we have that $g_2(\lambda\omega_1, \lambda\omega_2) = \lambda^{-4}g_2(\omega_1, \omega_2)$ and $g_3(\lambda\omega_1, \lambda\omega_2) = \lambda^{-6}g_3(\omega_1, \omega_2)$. This means that when one multiplies the periods by λ, then the equation $y^2 = 4x^3 - g_2x - g_3$ changes to $y^2 = 4x^3 - \lambda^{-2}g_2x - \lambda^{-3}g_3$ which is equivalent to the original one under the coordinate change $(x, y) \mapsto (\lambda^2x, \lambda^3y)$. This of course induces a bijection on the set of points of the elliptic curves, and if we transfer it back to $\mathbb{C}/\Gamma_{\omega_1,\omega_2}$ and $\mathbb{C}/\Gamma_{\lambda\omega_1,\lambda\omega_2}$ it becomes just multiplication by λ on $\mathbb{C}$ (note that $\Gamma_{\lambda\omega_1,\lambda\omega_2} = \lambda\Gamma_{\omega_1,\omega_2}$). This leads to the following classification problem: A *lattice* in $\mathbb{C}$ is a subgroup generated by a real basis for $\mathbb{C}$. Two lattices Γ and Λ are *similar* if there is $\lambda \neq 0 \in \mathbb{C}$ such that $\Lambda = \lambda\Gamma$. The problem is then to classify lattices up to similarity. Note that any lattice is of the form $\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2 = \Gamma_{\omega_1,\omega_2}$ so that our discussion above shows that classifying lattices up to similarity is the same thing as classifying pairs (ω_1, ω_2) of non-zero complex numbers such that $\omega_2/\omega_1 \notin \mathbb{R}$ up to the action of the group $\mathrm{GL}_2(\mathbb{Z}) \times \mathbb{C}^\times$ acting as

$$\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \lambda\right)(\omega_1, \omega_2) = (\lambda(a\omega_1 + b\omega_2), \lambda(c\omega_1 + d\omega_2)).$$

Now we can start by multiplying such a pair by $1/\omega_2$ taking it into a pair of the form $(\tau, 1)$, where $\tau \notin \mathbb{R}$. If $\Im(\tau) < 0$ we can replace τ by $-\tau$ so we may assume that $\Im(\tau) > 0$. Now, if $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{GL}_2(\mathbb{Z})$ it takes $(\tau, 1)$ to $(a\tau + b, c\tau + d)$ and then scaling gives $(\tau', 1)$, where $\tau' = (a\tau + b)/(c\tau + d)$. We then use the following exercise:

Exercise 40. Show that if $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{GL}_2(\mathbb{R})$ and $\tau \in \mathbb{H}$, where $\mathbb{H} := \{z \in \mathbb{C} \mid \Im(z) > 0\}$, the *upper half plane*, then $(a\tau + b)/(c\tau + d) \in \mathbb{H}$ precisely when $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{GL}_2^+(\mathbb{R})$, the group of matrices with (strictly) positive determinant.

From this it follows that the set of equivalence classes of lattices up to similarity is in natural bijection with the set of equivalence classes of $\mathbb{H}$ under the action of $\mathrm{SL}_2(\mathbb{Z})$ acting on $\mathbb{H}$ by Möbius transformations. We shall now determine a *fundamental domain* for this action of $\mathrm{SL}_2(\mathbb{Z})$, i.e., a closed subset of $\mathbb{H}$ such that each point of $z \in \mathbb{H}$ is equivalent with a point of it and only points on the boundary are equivalent with some other point of the domain.

Proposition 6.1. *Let D be the subset $\{z \in \mathbb{H} \mid |z| \geq 1,\ |\Im(z)| \leq 1/2\}$ of $\mathbb{H}$ (see Fig. 25).*

i) *For every $z \in \mathbb{H}$ there is a $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ such that $\gamma(z) \in D$.*

ii) *If $z \neq w \in \mathbb{H}$ and if $w = \gamma(z)$ for some $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ different from $\pm$ the identity matrix, then $|\Im(z)| = 1/2$ and $w = z \pm 1$, or $|z| = 1$ and $w = -1/z$, or $z = e^{2\pi i/3}, e^{\pi i/3}$ and $\gamma = \pm\left(\begin{smallmatrix} 0 & 1 \\ -1 & \pm 1 \end{smallmatrix}\right)$ or $\pm$ the square of this matrix.*

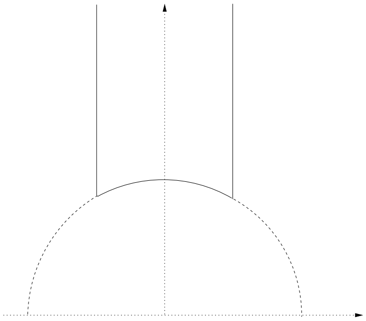

Figure 25. A fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$.

Proof. We start by noticing that if $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, then it is easily checked that

$$\Im(\gamma z) = \frac{\Im(z)}{|cz + d|^2}. \tag{6.1}$$

Now, as c and d are integers, for a fixed z there are only a finite number of possibilities for $|cz + d|$ below a given number. Hence given a $z \in \mathbb{H}$ there is an element of

$\{\gamma z \mid \gamma \in \mathrm{SL}_2(\mathbb{Z})\}$ of largest imaginary part and by replacing z by it, we may assume that z has the largest imaginary part among the elements of the set. We may then further replace z by some $z+m$, $m \in \mathbb{Z}$, and may assume that $|\Im(z)| \le 1/2$. In that case we must have $|z| \ge 1$ as otherwise $-1/z$ has a larger imaginary part.

Assume now that $z \in D$ and that $\gamma z \in D$ for some $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$. After possibly replacing (z, γ) by $(\gamma z, \gamma^{-1})$ we may assume that $\Im(\gamma z) \ge \Im(z)$. From (6.1) it then follows that $|cz+d| \le 1$. Writing $z = x+iy$ this implies $\{cx\}^2 + (cy)^2 \le 1$, where $\{r\}$ is the distance of r to the closest integer. On the other hand, $z \in D$ means $x^2+y^2 \ge 1$ and $|x| \le 1/2$ which gives $y^2 \ge 3/4$. As we on the other hand have $c^2y^2 \le 1$ we get $c = 0, \pm 1$. If $c = 0$ we get $\gamma z = z+b$ and as $\gamma z \in D$ we get $|x| = 1/2$ and $b = \pm 1$. If on the other hand we have $c = \pm 1$ we get $\{cx\} = |x|$ and hence $x^2+y^2 = \{cx\}^2+(cy)^2 \le 1$ which together with $x^2+y^2 \ge 1$ gives $x^2+y^2 = 1$. If $d = 0$ we directly get $\gamma z = -1/z$. If not, the fact that $1 = x^2+y^2 = |cz+d|^2 \le 1$ gives $|cx+d| = |x|$ which implies $x = \pm 1/2$ which together with $x^2+y^2 = 1$ and $y > 0$ gives $z = e^{2\pi i/3}$ or $z = e^{\pi i/3}$. A simple computation leaves us with only the two stated possibilities for γ. □

This gives us a very clear picture of the equivalence classes of lattices; they correspond to points in the fundamental domain with only a few points on the boundary being identified. We have seen that if we start with two similar lattices, then the corresponding elliptic curves are equivalent. Two questions arise: Do all equivalence classes of curves arise in this way, and if the curves are equivalent are the lattices also equivalent? These questions will be the topics of the next subsection.

As an aside, this determination of the fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$ also gives a way of determining the class group for imaginary quadratic fields.

Exercise 41. Let $K = \mathbb{Q}(-D)$, D a square-free positive integer, be a quadratic imaginary number field and $R = \mathbb{Z}[\alpha]$ the ring of algebraic integers, where $\alpha = \sqrt{-D}$ if $D \equiv 2, 1 \bmod 4$ and $\alpha = (1+\sqrt{-D})/2$ if not. A *fractional ideal* of R is an R-submodule of K with a finite number of generators. Two fractional ideals are equivalent if they differ under multiplication by a non-zero element of K.

i) Show that a fractional ideal is a lattice in $\mathbb{C}$ if we think of K as a subset of $\mathbb{C}$.

ii) Show that a lattice in $\mathbb{C}$ is equal to a fractionary ideal times a non-zero complex number precisely when it is stable under multiplication by α.

iii) Show that every fractional ideal is equivalent to a unique fractional ideal of the form $\mathbb{Z}+\mathbb{Z}\tau$, where $\tau \in K$, $-1/2 < \Re\tau \le 1/2$ and $|\tau| \ge 1$ and $|\tau| = 1$ implies that $\Re\tau \ge 0$.

iv) Show that there are only a finite number of equivalence classes of fractional ideals.

v) Determine representatives for ideal classes for $D = 23$.

The elliptic curves and in particular their j-invariants whose lattices are fractional ideals as in the previous exercise have some remarkable properties, some of which are indicated in the following exercise (we shall return to that subject later, the exercise may be too difficult before we have done that).

Exercise 42. i) Use the q-expansion of j to compute $j(\tau)$ numerically (for definition of which see next section) for τ giving representatives for the class group for $\mathbb{Q}(\sqrt{-23})$ as in the previous exercise. Show that the coefficients of the monic polynomial $p(t)$ that has these values as roots are, up to high precision, integers and determine these integers.

ii) Show that the splitting field of $p(t)$ contains $\mathbb{Q}(\sqrt{-23})$.

iii) Show that the splitting field of $p(t)$ is unramified over $\mathbb{Q}(\sqrt{-23})$.

6.3 The j-function

In this section we assume that the characteristic of the base field K is different from 2 *and* 3. *We also assume that K is algebraically closed.*

Returning to the classification of elliptic curves, we have seen that every elliptic curve is equivalent to one of the form $y^2 = x(x-1)(x-\lambda)$ and we also showed (see Exercise 38) that if it is equivalent with $y^2 = x(x-1)(x-\lambda')$, then λ' is one of λ, $1-\lambda$, $1/\lambda$, $1/(1-\lambda)$, $\lambda/(\lambda-1)$, or $(\lambda-1)/\lambda$. We would like to have an expression in λ that is a *complete invariant*, i.e., it should take the same value on all equivalent λ's (which says that it is an invariant) and if it takes the same value on two λ's these two values should give equivalent curves. Let us first look for invariants and let us aim for the simplest possible ones, namely a rational function $f(\lambda) \in K(\lambda)$. What it then means for it to be an invariant is that, as rational functions, $f(\lambda') = f(\lambda)$, where λ' is one of λ, $1-\lambda$, $1/\lambda$, $1/(1-\lambda)$, or $\lambda/(\lambda-1)$. Now, these six Möbius transformations actually form a subgroup G of the group of Möbius transformations. This can be seen by direct verification but the real reason for this fact is contained in the following exercise.

Exercise 43. Let Σ_3 be the group of permutations on three letters. Consider the map that associates to any $\sigma \in \Sigma_3$ the Möbius transformation T_σ which takes $(\infty, 0, 1)$ to $\sigma(0, 1, \infty)$ (where σ permutes the coordinates). Show that this map is an injective group homomorphism from Σ_3 to the group of Möbius transformations. Show further that the image consists exactly of the Möbius transformations of the form λ, $1-\lambda$, $1/\lambda$, $1/(1-\lambda)$, and $\lambda/(\lambda-1)$.

What we are asking for is thus an f which is invariant under the group G. It is clear that the subset consisting of all invariants under G form a subfield (containing K of course). Even though we shall not use it, it follows from Galois theory that $K(\lambda)$ is of degree $|G| = 6$ over this field of invariants. It also follows from a theorem of Lüroth[3] that there is a single invariant such that any other invariant is a rational function in this invariant. This suggests that there could indeed be a single complete invariant (though it does not quite prove it).

[3] Jacob Lüroth, 1844–1910

In any case in such a situation there is a systematic way of constructing invariants. We start with some element in $K(\lambda)$, e.g. λ, and then consider the polynomial that has as roots the transformations of the element; $\prod_{g\in G}(t-g\lambda)$. As applying a $g\in G$ to this polynomial just permutes its roots, its coefficients will be invariants. Concretely we have[4]

$$\begin{aligned}p(x,\lambda) &:= (x-\lambda)(x-(1-\lambda))\left(x-\frac{1}{\lambda}\right)\left(x-\frac{1}{1-\lambda}\right)\left(x-\frac{\lambda}{\lambda-1}\right)\left(x-\frac{\lambda-1}{\lambda}\right)\\ &= x^6-3x^5-\frac{\lambda^6-3\lambda^5+5\lambda^3-3\lambda+1}{(\lambda-1)^2\lambda^2}x^4\\ &\quad+\frac{2\lambda^6-6\lambda^5+5\lambda^4+5\lambda^2-6\lambda+2}{(\lambda-1)^2\lambda^2}x^3\\ &\quad-\frac{\lambda^6-3\lambda^5+5\lambda^3-3\lambda+1}{(\lambda-1)^2\lambda^2}x^2-3x+1.\end{aligned}$$

We only get the following two distinct non-constant coefficients $\frac{\lambda^6-3\lambda^5+5\lambda^3-3\lambda+1}{(\lambda-1)^2\lambda^2}$ and $\frac{2\lambda^6-6\lambda^5+5\lambda^4+5\lambda^2-6\lambda+2}{(\lambda-1)^2\lambda^2}$. However, it is easy to see that

$$2\frac{\lambda^6-3\lambda^5+5\lambda^3-3\lambda+1}{(\lambda-1)^2\lambda^2}-\frac{2\lambda^6-6\lambda^5+5\lambda^4+5\lambda^2-6\lambda+2}{(\lambda-1)^2\lambda^2}=-5$$

so that they essentially give the same invariants and we need only consider one of them. It turns out that any of them actually has the property that any other invariant is a rational function in it.

Exercise 44. Let $L\subseteq K(\lambda)$ be the subfield generated by K and the coefficients of the polynomial $\prod_{g\in G}(t-g\lambda)$. Show that $K(\lambda)$ has degree ≤ 6 over L. Show that $K(\lambda)$ has degree 6 over the field of invariants of G. Conclude that L equals the field of invariants.

However we shall not use one of the coefficients but rather use

$$6+\frac{\lambda^6-3\lambda^5+5\lambda^3-3\lambda+1}{(\lambda-1)^2\lambda^2}=\frac{(\lambda^2-\lambda+1)^3}{\lambda^2(\lambda-1)^2}$$

which has a nicer numerator. For various reasons we also want to consider a rather peculiar multiple of it and put

$$j(\lambda):=2^8\frac{(\lambda^2-\lambda+1)^3}{\lambda^2(\lambda-1)^2}$$

which (for obvious reasons) is called the *j-invariant* of the elliptic curve. Note that in terms of the j-function we have

$$p(x,\lambda)=x^6-3x^5-(2^{-7}j(\lambda)-6)x^4+(2^{-7}j(\lambda)-7)-(2^{-7}j(\lambda)-6)-3x+1. \quad (6.2)$$

[4] See http://www.math.su.se/~teke/undervisning/Elliptisk.nb for this calculation.

By construction j is now an invariant but the next question is if it is a complete invariant. This would mean that if $\lambda, \lambda' \in K \setminus \{0, 1\}$ and if $j(\lambda) = j(\lambda')$, then there is a $g \in G$ such that $\lambda' = g\lambda$. This is what we now want to prove. We get from (6.2) that $p(x, \lambda) = p(x, \lambda')$ and as $0 = p(\lambda, \lambda)$ we get that $p(\lambda, \lambda') = 0$ and by the definition of p this means that $\lambda' = g\lambda$ for some g.

The definition of j is fine but not easy to apply to a general equation $y^2 = ax^3 + bx^2 + cx + d$. If we should happen to start with such an equation we can always after scaling y assume that $a = 1$. After that the recipe is to consider the cross ratio $\lambda = \{\infty, \alpha, \beta, \gamma\}$, where α, β, and γ are the three roots of $f(x) := x^3 + bx^2 + cx + d$ and then consider $j(\lambda)$. By construction $j(\lambda)$ is independent of the order in which we enumerated the roots. Now the *fundamental theorem of symmetric functions* says that any function in the roots of a polynomial that is symmetric in them (i.e., is invariant when the arguments of the function is permuted) is a function in the coefficients of the polynomial. Usually "function" is taken to mean polynomial but the case of rational functions follows directly from this. Hence we should be able to express j as a rational function of b, c and d and this is what we shall do. We know however that if we replace f by $f(rx+s)$ for $0 \neq r, s \in K$, then j will be unchanged. Thus we should be looking for rational functions in b, c and d which are unchanged under such transformations. If we start by looking at invariance under $x \mapsto x + s$, then it is easy to see[5] that

$$\begin{aligned} g_2 &:= \frac{4b^2}{3} - 4c, \\ g_3 &:= -\frac{8b^3}{27} + \frac{4bc}{3} - 4d \end{aligned}$$

are invariant under the transformations $x \mapsto x+s$. It is more easily verified that for the polynomial $x^3 - a/4x - b/4$ we have $g_2 = a$ and $g_3 = b$ and hence for the polynomial constructed out of the fundamental periods ω_1 and ω_2 we have $g_2 = g_2(\omega_1, \omega_2)$ and $g_3 = g_3(\omega_1, \omega_2)$. The fact that g_2 and g_3 are invariant means that any polynomial and any rational function in them are invariant. In order for such a rational function to be invariant under all $x \mapsto rx + s$ it is then enough that it is invariant also under $x \mapsto rx$, since if it is invariant under two transformations it is invariant under their composite, and $x \mapsto rx + s$ is the composite of $x \mapsto rx$ and $x \mapsto x + s$. Now $x \mapsto rx$ takes g_2 to $r^{-2}g_2$ and g_3 to $r^{-3}g_3$. Hence if we say that a monomial $g_2^i g_3^j$ has *degree* $2i + 3j$ and that a polynomial of g_2 and g_3 is *homogeneous* of degree n if all the monomials appear in it have degree n, then a rational function in g_2 and g_3 is invariant under $x \mapsto rx$ for all r precisely when it is the quotient of homogeneous polynomials of the same degree. Such a rational function is a rational function of the roots which is invariant under reordering of the roots and (simultaneous) affine transformations of the roots. This means that it is determined by its restriction to polynomials of the form $x(x-1)(x-\lambda)$. Hence to check that such a rational function is equal to j it is enough

[5] See http://www.math.su.se/~teke/undervisning/Elliptisk.nb.

to check it on such polynomials. For them we have[6]

$$\begin{aligned} g_2 &:= \frac{4}{3}(\lambda^2 - \lambda + 1),\\ g_3 &:= \frac{4}{27}(\lambda - 2)(\lambda + 1)(2\lambda - 1). \end{aligned}$$

One now checks that

$$g_2^3 - 27g_3^2 = 16\lambda^2(\lambda - 1)^2$$

and combining everything we get

$$j(g_2, g_3) = 1728\frac{g_2^3}{g_2^3 - 27g_3^2}.$$

Notice that even though $g_2^3 - 27g_3^2$ is not invariant under affine transformation, it is multiplied by a non-zero element under such a transformation. As function of λ it is zero precisely when λ is equal to one of the other roots. From this it is easy to conclude that it vanishes for an arbitrary polynomial precisely when that polynomial has a multiple zero. It is called the *discriminant* of the polynomial.

Exercise 45. Show that a rational function in g_2 and g_3 invariant under $x \mapsto rx$ is a rational function in j.

On the purely algebraic side we have a very satisfactory picture; two elliptic curves are equivalent precisely when their j-invariants are equal. The picture can then be completed by noticing that over an algebraically closed field, $j(\lambda) = j$ always has a solution λ. However, we can do even a little bit better.

Exercise 46. Show that if $j \neq 0$, 1728, then the equation

$$y^2 = x^3 - 1/2x^2 - 36/(j - 1728)x - 1/(j - 1728)$$

defines an elliptic curve with j-invariant j. Show that $y^2 = x^3 - 1$ has j-invariant 0 and $y^2 = x^3 - x$ has j-invariant 1728.

On the analytic side things are less satisfactory. For a lattice Γ we get a curve $y^2 = 4x^3 - g_2(\Gamma)x - g_3(\Gamma)$ and equivalent lattices are mapped to equivalent curves. We do not know however if this is a bijection. By composing with j, these questions are equivalent to the following: If

$$j(\omega_1, \omega_2) = 1728\frac{g_2(\omega_1, \omega_2)^3}{g_2(\omega_1, \omega_2)^3 - 27g_3(\omega_1, \omega_2)^2}$$

is equal to $j(\omega_1', \omega_2')$, are then the two lattices spanned by them similar? Is any complex number of the form $j(\omega_1, \omega_2)$? We shall now study these questions. Note

[6] As usual see http://www.math.su.se/~teke/undervisning/Elliptisk.nb.

that by construction $j(\omega_1, \omega_2)$ depends only on the similarity classes of the lattice generated by ω_1 and ω_2 and by the analysis performed previously j is determined by the function $j(\tau) := j(\tau, 1)$. That function, as a function $j\colon \mathbb{H} \to \mathbb{C}$ fulfils

$$j\left(\frac{a\tau + b}{c\tau + d}\right) = j(\tau) \quad \text{for } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}),$$

and because of this it is determined by its values on the fundamental domain of the action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathbb{H}$.

We shall now show that j is indeed a bijection. The situation is very similar to the one for $z \mapsto (\wp(z), \wp'(z))$ and our proof will also be analogous. Hence we could try to integrate $j'(\tau)/(j(\tau) + c)$ around the boundary of the fundamental domain. There are several problems with this approach but the most important is that the fundamental domain is no longer compact, i.e., closed and bounded. It certainly is closed but not bounded, more precisely $\Im(\tau)$ is not bounded on it. We deal with this as follows. We start by a general meromorphic function $f\colon \mathbb{H} \to \overline{\mathbb{C}}$ fulfilling the condition $f(\gamma\tau) = f(\tau)$ for all $\gamma \in \mathrm{SL}_2(\mathbb{Z})$. In particular we have that $f(\tau + 1) = f(\tau)$, i.e., f is periodic of period 1. One holomorphic function of period 1 is the map $q\colon \mathbb{H} \to \mathbb{D}$, where $\mathbb{D} := \{z \in \mathbb{C} \mid |z| < 1\}$, given by $q(\tau) = e^{2\pi i\tau}$. We shall now sketch a proof of the fact that for every holomorphic $h\colon \mathbb{H}_h \to \mathbb{C}$, $\mathbb{H}_h := \{z \in \mathbb{C} \mid \Im(z) > h\}$, which is periodic of period 1, there is a holomorphic function $H\colon \mathbb{D}_t^* \to \mathbb{C}$, where $t = e^{-2\pi h}$ and $\mathbb{D}_t^* := \{0 \neq z \in \mathbb{C} \mid |z| < t\}$, such that $h(\tau) = H(q)$, where we write $q = e^{2\pi i\tau}$. One way of showing this is to consider $H(q) := h(\log(q)/2\pi i)$. The logarithm is well defined (and holomorphic) up to integer multiples of $2\pi i$. Thus $\log(q)/2\pi i$ is well defined up to an integer but as h is periodic of period 1, $H(q)$ is actually well defined and holomorphic, being locally the composite of two holomorphic functions.

Exercise 47. i) Fill in the details of this proof.

ii) Give an alternative proof by first expanding $x \mapsto f(x + yi)$ in a Fourier series whose coefficients are functions in y and then use Cauchy–Riemann's equations to show that these coefficients as functions of y has the right form.

Now, we say that f is *meromorphic* (*holomorphic*) at infinity if there is an $h > 0$ such that f is without poles on $\mathbb{H}_h$ and that the F given by $F(q) = f(\tau)$ for $q \in \mathbb{D}_{e^{-2\pi h}}$ has a convergent expansion

$$F(q) = \sum_{n \geq N} a_n q^n$$

for some $N \in \mathbb{Z}$ (resp. $N \geq 0$). (We shall in the future not introduce a new symbol for the function in q given by a periodic function in τ but rather write $f(q)$.) A function f that is

- meromorphic on $\mathbb{H}$,
- invariant under $\mathrm{SL}_2(\mathbb{Z})$, i.e., $f(\gamma\tau) = f(\tau)$ for $\gamma \in \mathrm{SL}_2(\mathbb{Z})$, and
- meromorphic at infinity

will be called a *modular function*.

Exercise 48. Show that a modular function without poles (including at infinity) is constant.

We now want to prove that, just as for elliptic functions, a modular function has the same number of poles as zeroes. However, in the elliptic case the proof of this property was done by showing that the sum of the multiplicity of the zeroes and the poles, where the poles were counted with negative multiplicity, was equal to zero. We shall formalise this description (a formalisation which actually is also convenient for elliptic functions). Assume that f is not identically 0. Then we almost define $z \in \mathbb{H}$ $v_z(f)$ to be equal to the multiplicity of z as a zero if it is holomorphic at z and minus the multiplicity of z as a pole if not. By "almost" we mean that we must modify it at some points (a modification which is reasonable *only* if f is modular). More precisely, if z is congruent under $\mathrm{SL}_2(\mathbb{Z})$ to i, we divide the multiplicity of the zero (or minus the multiplicity of the pole) by 2 and if z is congruent to $e^{\pi i/3}$ we divide by 3. We also define $v_\infty(f)$ to be the multiplicity of the zero or minus the multiplicity of the pole of $f(q)$ at $q = 0$.

These modifications may at first sight seem quite ad hoc but we shall see that they fit very well with the proof of the next result. The following exercise gives a further motivation for the modifications.

Exercise 49. Show that for a non-zero modular function f, $v_z(f) \in \mathbb{Z}$ for $z \in \mathbb{H}$.

We are now almost ready to formulate our result on the zeroes and poles of a modular function. What is needed is first to note that a non-zero modular function has only a finite number of zeroes or poles in the fundamental domain. This is not quite as easy as the elliptic case, as the fundamental domain is not compact. However, from the fact that $f(q)$ is meromorphic in the whole unit disc we get that there is some concentric disc of radius smaller than 1 that only contains a finite number of zeroes and poles and then as $D \cap \{z \in \mathbb{H} \mid \Im(z) \leq R\}$ is compact we get in all just a finite number. Secondly we must take care to count each zero or pole once. Hence by a *complete set of non-congruent zeroes and poles* of f we mean the choice of a set of poles and zeroes of f (including possibly ∞) such that each zero or pole is congruent modulo $\mathrm{SL}_2(\mathbb{Z})$ with exactly one of them (where ∞ is only congruent with itself). One can make a specific choice of representatives of the congruence classes by insisting that they should lie in the fundamental domain minus the line $\Im(z) = -1/2$ and the arc $|z| = 1,\ \pi/2 < \arg(z) \leq 2\pi/3$ but we shall not insist on doing so.

Proposition 6.2. *If f is a non-zero modular function we have that*

$$\sum_z v_z(f) = 0$$

where the sum runs over a complete set of non-congruent zeroes and poles of f.

Proof. We assume that, except possibly for z equal to i, $e^{\pi i/3}$, or , $e^{2\pi i/3}$, there are no zeroes or poles on the boundary of the fundamental domain, leaving the general case to an exercise. Now pick an R such that $D \cap \{z \in \mathbb{C} \mid \Im(z) \geq R\}$ contains no zeroes or poles and consider the path of Fig. 26 where the horisontal part is at $\Im(z) = R$ and the vertical at $\Re(z) = \pm 1/2$. As usual we can find a free homotopy of this path to one

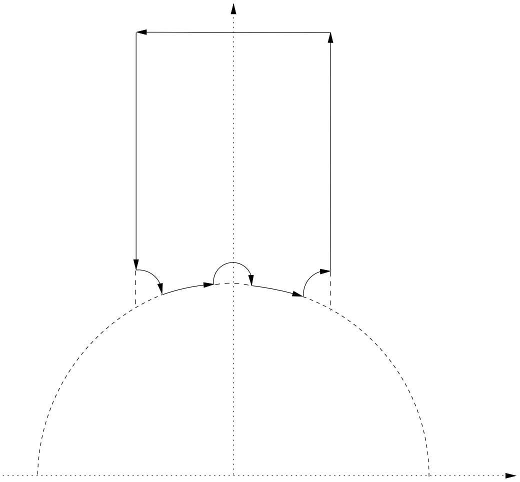

Figure 26. A path encircling zeroes and poles.

encircling the poles and zeroes of f in the interior of D plus some connecting paths traversed in both directions. Hence if we integrate df/f over this path we get $2\pi i$ times the sum of the $v_z(f)$ for z in the interior of D. Hence what needs to be proven is that this integral equals $-2\pi i(v_\infty(f) + v_i(f) + v_\zeta(f))$, where $\zeta = e^{\pi i/3}$. We now look at the contribution of different parts of the path to the integral:

- The integral along the left-hand vertical part cancels the integral along the right-hand vertical part as $f(\tau + 1) = f(\tau)$ and as the left-hand maps to the opposite of the right-hand under $\tau \mapsto \tau + 1$.

- The integral along the left-hand arc on the unit circle cancels the integral along the right-hand arc on the unit circle as $f(-1/\tau) = f(\tau)$ and as these arcs map to the opposite of each other under $\tau \mapsto -1/\tau$ (provided the arcs have been chosen with the end points mapping to each other under the map $\tau \mapsto -1/\tau$).

- The integral along the horisontal part corresponds to going once around the origin on the circle of radius $e^{-2\pi R}$ in *clockwise* direction in the q coordinate. As $q = 0$ is the only possible pole or zero, we get that the integral along it is $-2\pi i v_\infty(f)$.

- The integral along the right-hand arc going from the unit circle to the line $\Im(z) = 1/2$ is an arc with center at ζ. The integral would seem to depend on the radius (though it can be seen that it doesn't) but in any case, when the radius tends to 0 it tends to $-6\pi i v_\zeta(f)$ times the limit proportion of the length of the arc to the length of the full circle. That proportion is the angle between the unit circle and the line $\Im(z) = 1/2$ divided by 2π which is $1/6$ with total contribution $-\pi i v_\zeta(f)$. (We are integrating a function with a simple pole with $(z-\zeta)^{-1}$-coefficient equal to $3v_\zeta(f)$ along such an arc and the contribution from all the other terms of Taylor series tends to zero.)

- The integral along the left-hand arc going from the line $\Im(z) = -1/2$ to the unit circle tends, when the radius tends to zero, by the same argument to $-\pi i v_{-\zeta^2}(f)$. However, $f(-1/\tau) = f(\tau)$ and $\tau \mapsto -1/\tau$ takes $-\zeta^2$ to ζ so that $v_{-\zeta^2}(f) = v_\zeta(f)$ and the total contribution of both arcs is $-2\pi i v_\zeta(f)$.
- The integral along the arc with center at i tends, with the same argument as for ζ, to $-2\pi i v_i(f)$ when the radius tends to zero.

The different contributions thus add up to exactly what we want. □

Exercise 50. Prove the case when there could be zeroes and poles anywhere on the boundary fundamental domain.

We would like to be able to use this result in a fashion very similar to the way it was used for $\wp$ to show that j does indeed take similarity classes of lattices bijectively to complex numbers. That would follow if we knew that j had a simple pole at infinity. We shall see that this is indeed the case but it requires computing first the "q-expansion" of the G_k. This makes sense because even though G_k is not a modular function it is periodic of period 1 and any such holomorphic function can be so expanded. A closer look at the definition of G_k reveals that it can be written as an infinite sum of functions that are periodic of period 1 (where we as for j write $G_k(\tau) := G_k(\tau, 1)$):

$$G_k(\tau) = \sum_{m\in\mathbb{Z}}{}' \frac{1}{m^{2k}} + 2\sum_{m\geq 1}\sum_{n\in\mathbb{Z}} \frac{1}{(m\tau+n)^{2k}}. \tag{6.3}$$

We see that not only is the inner term periodic; it is in fact the evaluation at $m\tau$ of the function

$$f_{2k}(z) := \sum_{n\in\mathbb{Z}} \frac{1}{(z+n)^{2k}}.$$

This expression makes sense also for odd exponents and even for exponent 1 if one sums symmetrically around the origin. In the following two exercises we shall now show that

$$\pi\cot(\pi z) = \frac{1}{z} + \sum_{n\geq 1} \frac{1}{z+n} + \frac{1}{z-n}. \tag{6.4}$$

Exercise 51. Show that the difference of the right- and left-hand side of (6.4) is a holomorphic function on $\mathbb{C}$ periodic of period 1 and bounded when $\Im(z) \to \infty$. Conclude that the difference is a constant and compute that constant by putting $z = 0$.

Exercise 52. Put $z = x + iy$ and expand the right-hand series of (6.4) as a Fourier series in x with fixed $y > 0$ by judiciously exchanging sum and integral, and then compute the resulting integral using residue calculus. Compare the result with the q-expansion of $\pi \cot(\pi z)$ obtained below.

We now use Euler's formula to write

$$\pi \cot(\pi\tau) = \pi i \frac{e^{\pi i\tau} + e^{-\pi i\tau}}{e^{\pi i\tau} - e^{-\pi i\tau}} = \pi i \frac{q+1}{q-1} = \pi i - 2\pi i \sum_{n=0}^{\infty} q^n,$$

and then taking derivatives repeatedly we get for $k \geq 2$,

$$\sum_{n\in\mathbb{Z}} \frac{1}{(\tau+n)^k} = \frac{(-2\pi i)^k}{(k-1)!} \sum_{m=1}^{\infty} m^{k-1} q^m.$$

Applying that to (6.3), treating the $n = 0$-term specially, and using that $e^{2\pi i m\tau} = q^m$, we get

$$\begin{aligned} G_k(\tau) &= 2\zeta(2k) + 2\sum_{m\geq 1} \frac{(-2\pi i)^{2k}}{(2k-1)!} \sum_{r=1}^{\infty} r^{2k-1} (q^m)^r \\ &= 2\zeta(2k) + \frac{(-2\pi i)^{2k}}{(2k-1)!} 2 \sum_{n=1}^{\infty} \sum_{r|n} r^{2k-1} q^n, \end{aligned}$$

and if we introduce the (standard) notation $\sigma_k(n) := \sum_{d|n} d^k$ we get

$$G_k(\tau) = 2\zeta(2k) + \frac{(-1)^k 2(2\pi)^{2k}}{(2k-1)!} \sum_{n=1}^{\infty} \sigma_{2k-1}(n) q^n. \tag{6.5}$$

Now, in order to get the q-expansion for j we need to compute the constant term in this expansion (or for that application really only for $k = 2, 3$). We can do this by using (6.4) once more. We start by defining[7] the *Bernoulli numbers*[8] by the equality

$$\frac{x}{e^x - 1} = 1 - \frac{x}{2} + \sum_{k=1}^{\infty} (-1)^{k+1} B_k \frac{x^{2k}}{(2k)!}$$

(for this to make sense one needs to show that $x/(e^x - 1) + x/2$ is an even function which follows from the transformation below). It is clear that the B_k are rational

[7] There are some differing conventions concerning Bernoulli numbers, in particular it is in some contexts convenient not to build into the definition that most of the odd terms vanish in the expansion below. If that is done the Bernoulli numbers here would appear with even exponents.

[8] Jacob Bernoulli, 1654–1705

numbers and one computes $B_1 = 1/6$, $B_2 = 1/30$, and $B_3 = 1/42$. Putting $x = 2\pi iz$ gives

$$\pi z \cot(\pi z) = 1 - \sum_{k=1}^{\infty} B_k \frac{(2\pi)^{2k} z^{2k}}{(2k)!}$$

(and now it is clear that the left-hand side is an even function). On the other hand from (6.4) we get (where we temporarily let $\infty \cdot 0 = 0$)

$$\begin{aligned}\pi z \cot(\pi z) &= 1 + \sum_{n\geq 1} \frac{z/n}{1+z/n} - \frac{z/n}{1-z/n}\\ &= 1 + \sum_{k\geq 1}\sum_{n=1}^{\infty} \frac{1}{n^k}((-1)^{k-1} - 1)z^k = 1 - 2\sum_{k\geq 1} \zeta(2k) z^{2k}\end{aligned}$$

and comparing coefficients we get

$$2\zeta(2k) = B_k \frac{(2\pi)^{2k}}{(2k)!}.$$

Now using this formula we may factor out the $2\zeta(2k)$-factor in the q-expansion to obtain

$$E_k := \frac{G_k}{2\zeta(2k)} = 1 + (-1)^k \frac{4k}{B_k} \sum_{n\geq 1} \sigma_{2k-1}(n) q^n.$$

Exercise 53. i) Use Exercise 18 to show that $E_2^2 = E_4$ and that $E_2 E_3 = E_5$.

ii) Show that

$$\sigma_7(n) = \sigma_3(n) + 120 \sum_{m=1}^{n-1} \sigma_3(m)\sigma_3(n-m)$$

and

$$11\sigma_9(n) = 21\sigma_5(n) - 10\sigma_3(n) + 5040 \sum_{m=1}^{n-1} \sigma_3(m)\sigma_5(n-m)$$

for all n.

We can now use this for $k = 2, 3$ to get[9] the q-expansion for j:

$$j(q) = \frac{1}{q} + 744 + 196884q + 21493760q^2 + 864299970q^3 + \cdots.$$

In particular it shows that j has a simple pole at infinity (and with q^{-1}-coefficient equal to 1 which is one of the reasons for the strange coefficient in the definition of $j(\lambda)$). We can now prove the promised theorem on the bijectivity of j.

[9] See http://www.math.su.se/~teke/undervisning/Elliptisk.nb for the computation.

Proposition 6.3. *The j-function maps the upper half plane $\mathbb{H}$ surjectively onto the complex plane and $j(\tau) = j(\tau')$ precisely when there is a $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ such that $\tau' = \gamma(\tau)$.*

Proof. Given $c \in \mathbb{C}$ we can consider $j - c$ which is a modular function with a simple pole at infinity and which is holomorphic on $\mathbb{H}$. By Proposition 6.2 it must have at least one zero and hence there is a $\tau \in \mathbb{H}$ such that $j(\tau) = c$. Assume now that $c := j(\tau) = j(\tau')$ with τ and τ' non-congruent modulo $\mathrm{SL}_2(\mathbb{Z})$. As $j - c$ has a simple pole at infinity and no other poles, hence by Proposition 6.2 we have that $v_\tau(j - c) + v_{\tau'}(j - c) \leq 1$. As both of them are positive and integers unless they are congruent to i or ζ we get a contradiction, unless one of them is congruent to i and the other to ζ and there are no other zeroes. However, we can not write 1 as a positive integer linear combination of $1/2$ and $1/3$. (We can also refer to Exercise 49, which shows that $v_z(j - c)$ is always an integer, or Exercise 54.) □

Exercise 54. i) Show that $g_2(\zeta, 1) = 0$ and conclude that $j(\zeta) = 0$.
ii) Show that $g_3(i, 1) = 0$ and conclude that $j(i) = 1728$.

Exercise 55. Using Proposition 6.2 and Exercise 49 show that each modular function is a rational function in j.

7

Formulaire

We shall now consider the general form of an elliptic curve. We have already seen that there are problems with the form that we have been using in characteristic 2 and that there are some problems also in characteristic 3. However, for some problems we need to be even more general, e.g., arithmetic problems. In the arithmetic case we would like to allow at least the coefficients of the equation to be integers. Hence we assume for the moment that we allow coefficients from an arbitrary commutative ring R. The equation defining an elliptic curve then has the general form, the so-called *Weierstrass form*,

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

where thus $a_i \in R$. We shall allow some coordinate transformations which are postulated to take a curve into an equivalent curve:

$$\left.\begin{aligned} x &= u^2x' + r, \\ y &= u^3y' + su^2x' + t, \end{aligned}\right\} \tag{7.1}$$

where $r, s, t \in R$ and $u \in R^\times$ (the invertible elements). It is clear that under composition these transformations form a group and the equivalence classes of curves are just the orbits under that group. If 2 is invertible in R, then we can use the transformation $x = x'$ and $y = y' - a_1/2x' - a_3$ to reduce to the case when $a_1 = a_3 = 0$, and when 3 is invertible we can use $x = x' - a_2/3$ and $y = y'$ to reduce to the case when $a_2 = 0$. If both 2 and 3 are invertible, then we do the first and then the second transformation to reduce to the case when $a_1 = a_3 = a_2 = 0$, which is the case we have been studying previously. We can then consider the discriminant and the j-invariant which will become a polynomial, resp. a rational function, in the original variables. We claim[1] that they are given by the following formulas, where the last definition of course assumes that Δ is an invertible element of R. (This is not necessarily assumed but it is reasonable to assume that it is a non-zero divisor which is equivalent to the condition that R can be embedded in a ring where it is invertible).

$$\begin{aligned} b_2 &= a_1^2 + 4a_2, \\ b_4 &= a_1a_3 + 2a_4, \\ b_6 &= a_3^2 + 4a_6, \\ b_8 &= a_1^2a_6 - a_1a_3a_4 + 4a_2a_6 + a_2a_3^2 - a_4^2, \end{aligned}$$

[1] See http://www.math.su.se/~teke/undervisning/Elliptisk.nb for the calculation.

$$\begin{aligned}
c_4 &= b_2^2 - 24b_4,\\
c_6 &= -b_2^3 + 36b_2b_4 - 216b_6,\\
\Delta &= -b_2^2b_8 - 8b_4^3 - 27b_6^2 + 9b_2b_4b_6,\\
j &= c_4^3/\Delta.
\end{aligned}$$

Exercise 56. Show that an element t of a commutative ring R is a non-zero divisor precisely when R can be embedded in a commutative ring where t is invertible.

Note that some of the definitions can be simplified under further conditions. We have for instance that

$$4b_8 = b_2b_6 - b_4^2, \quad 1728\Delta = c_4^3 - c_6^2,$$

so that if 2 and 3 are invertible we can dispense with b_8 and get a simpler definition for Δ.

From now on we shall use the above definitions for Δ and j. It is clearly interesting to see what happens when we apply the coordinate transformation of (7.1) to the defined quantities. If we first look at the a's we get[2]

$$\begin{aligned}
ua_1' &= a_1 + 2s,\\
u^2a_2' &= a_2 - sa_1 - s^2 + 3r,\\
u^3a_3' &= a_3 + ra_1 + 2t,\\
u^4a_4' &= a_4 - sa_3 + 2ra_2 - (t + rs)a_1 + 3r^2 - 2st,\\
u^6a_6' &= a_6 + ra_4 - ta_3 + r^2a_2 - rta_1 + r^3 - t^2,
\end{aligned}$$

and for the b's we get

$$\begin{aligned}
u^2b_2' &= b_2 + 12r,\\
u^4b_4' &= b_4 + rb_2 + 6r^2,\\
u^6b_6' &= b_6 + 2rb_4 + r^2b_2 + 4r^3,\\
u^8b_8' &= b_8 + 3rb_6 + 3r^2b_4 + r^3b_2 + 3r^4,
\end{aligned}$$

and finally for the c's and Δ,

$$u^4c_4' = c_4, \quad u^6c_6' = c_6, \quad u^{12}\Delta' = \Delta.$$

From these last transformations we get that $j' = j$. For an algebraically closed field of characteristic different from 2 we showed that if $j' = j$ for two curves, then they are equivalent if $a_1 = a_1' = a_3 = a_3' = 0$.

Exercise 57. i) Show that if the characteristic of the field K is different from 2, then any elliptic curve over K is equivalent to one with $a_1 = a_3 = 0$.

ii) Show that if the characteristic of the algebraically closed field K is equal to 2, then if $j = j'$ for two elliptic curves, they are equivalent.

[2] For these and the subsequent ones see http://www.math.su.se/~teke/undervisning/Elliptisk.nb.

8

Finite fields

We shall now discuss the particular properties of elliptic curves over finite fields. Of course the most particular property of such curves is that the number of points of the curve over the field is always finite. As such it has a number of invariants whose evaluations should be of interest. The first such invariant is the order and we start by discussing it. Let us for simplicity consider the case of odd characteristic and consider the curve E given by the equation $y^2 = x^3 + ax + b$ with $a, b \in \mathbb{F}_q$. We start by considering the practical problem of counting the number of points $E(\mathbb{F}_q)$ on the curve over $\mathbb{F}_q$. We always have the point at infinity and apart from it we have all pairs $(s, t) \in \mathbb{F}_q$ for which $t^2 = s^3 + as + b$. That however would mean going through q^2 such pairs which rather quickly becomes unwieldy. In fact, there is no reason to go through all pairs if we just want to count the number of solutions. We are instead interested in knowing for a fixed x how many square roots in $\mathbb{F}_q$ the expression $x^3 + ax + b$ has. This number is 1 if it is zero, 2 if it is a non-zero square, and 0 if it is not. Hence if we define $\chi : \mathbb{F}_q \to \mathbb{Z}$ by

$$\chi(z) := \begin{cases} 0 & \text{if } z = 0, \\ 1 & \text{if } z \neq 0 \text{ and } z \text{ is a square,} \\ -1 & \text{if } z \neq 0 \text{ and } z \text{ is not a square,} \end{cases}$$

we have that the number of points in $E(\mathbb{F}_q)$ equals

$$1 + \sum_{x \in \mathbb{F}_q} \big(\chi(x^3 + ax + b) + 1\big) = 1 + q + \sum_{x \in \mathbb{F}_q} \chi(x^3 + ax + b).$$

This can be used to improve the speed of calculation but it also has theoretical implications. One heuristic implication is that we get a feeling for the size of the number of points of $E(\mathbb{F}_q)$. For this we note that there are at most three choices of x which are zeroes of $x^3 + ax + b$ and except for those values $\chi(x^3 + ax + b)$ takes on the values 1 and -1. Further, there are as many non-zero squares as there are non-squares in $\mathbb{F}_q$ and there seems to be no reason why $x^3 + ax + b$ should “prefer” to be square or non-square. If so the sum in the last term should be a sum of (approximately) q randomly (with even distribution) chosen ± 1’s. From the central limit theorem we get (in particular) that the expected size of the sum should be on the order of $\sqrt{q}$. We can check this for a curve with integer coefficients (in this case $y^2 = x^3 + x + 3$) and reduce it modulo different primes. In order to see the size of the sum we let $c_p := (p+1-|E(\mathbb{F}_p)|)/(2\sqrt{p})$:

p	c_p	p	c_p	p	c_p	p	c_p
5	0.447	71	0.474	149	−0.983	229	0.528
7	0.377	73	−0.351	151	−0.610	233	−0.622
11	−0.904	79	−0.225	157	0.119	239	−0.776
17	0.121	83	0.329	163	−0.939	241	0.354
23	−0.312	89	0.529	167	−0.502	251	0.378
29	−0.557	97	0.050	173	0.228	257	−0.187
31	−0.808	101	0.746	179	−0.074	263	0.369
37	−0.082	103	−0.788	181	−0.445	269	−0.487
41	0.234	107	−0.870	191	0.542	271	0.121
43	−0.228	109	−0.047	193	0.071	277	0.570
47	−0.437	113	0.564	197	0.641	281	−0.447
53	0.412	127	−0.709	199	0.106	283	0.118
59	0.325	131	0.305	211	−0.550	293	0.438
61	0.448	137	−0.598	223	0.669	307	0.313
67	0.183	139	0.381	227	0.796	311	0.453

We see that the sum does indeed seem to be of the order of $\sqrt{p}$. However, if the sum was truly a sum of independent evenly distributed variables, in about 5% of the cases c_p should have a value outside of $[-1, 1]$ (again by the central limit theorem). A look at the table shows however that we get outside of $[-1, 1]$ in none of the 60 cases. This is of course not conclusive evidence (even from a heuristic point of view) as we would only expect three exceptions, but the frequency plot in Figure 27 plots about 20, 000 primes and we see that there is still no value outside of $[-1, 1]$. This certainly suggests that we never get outside of $[-1, 1]$ and a result of Hasse gives in fact that we never do. Furthermore, Hasse's result is more precise in that it gives a relation between the number of points over a field $\mathbb{F}_q$ and the number of points over an extension field.

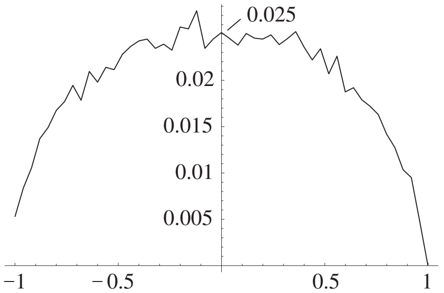

Figure 27. Distribution of points on elliptic curves.

Proposition 8.1. *Let $E\colon y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6$ be an elliptic curve over the finite field $\mathbb{F}_q$. Then there are complex numbers α and β with $\alpha\beta = q$ and $|\alpha| = |\beta| = \sqrt{q}$ and such that for any $n \geq 1$ we have*

$$|E(\mathbb{F}_{q^n})| = q^n + 1 - (\alpha^n + \beta^n).$$

In particular we have that $\big|q + 1 - |E(\mathbb{F}_{q^n})|\big| \leq 2q^{n/2}$.

Proof. That the inequality follows from the formulas is clear as $|\alpha^n + \beta^n| \leq |\alpha|^n + |\beta|^n = 2q^{n/2}$. The proof of this formula will not be taken up in these notes. □

Example 10. i) Consider the equation $y^2 + xy = x^3 + 1$ over $\mathbb{F}_2$. It has the solutions $(0, 1)$, $(1, 0)$, and $(1, 1)$ (plus of course the point at ∞) and hence $|E(\mathbb{F}_2)| = 4$ and we get $4 = 2 + 1 - (\alpha + \beta)$ which gives $\alpha + \beta = -1$ and as $\alpha\beta = 2$ we get that α and β are solutions to the equation $x^2 + x + 2 = 0$ which has the solutions $\alpha, \beta = \frac{-1\pm\sqrt{7}i}{2}$ and hence

$$|E(\mathbb{F}_{2^n})| = 2^n + 1 - \frac{(-1 + \sqrt{7}i)^n + (-1 - \sqrt{7}i)^n}{2^n}.$$

ii) Consider the equation $y^2 + xy = x^3 + x^2 + 1$ over $\mathbb{F}_2$. It has the solution $(0, 1)$ and hence $|E(\mathbb{F}_2)| = 2$ giving that α and β are solutions to $x^2 - x + 2 = 0$ giving $\alpha, \beta = \frac{1\pm\sqrt{7}i}{2}$.

iii) Consider the equation $y^2 + y = x^3$ over $\mathbb{F}_2$. It has solutions $(0, 0)$ and $(0, 1)$. Hence $|E(\mathbb{F}_2)| = 3$ and we get $\alpha + \beta = 0$ which gives $\alpha, \beta = \pm\sqrt{2}i$.

iv) Consider the equation $y^2 = x^3 + 1$ over a finite field $\mathbb{F}_q$ of characteristic different from 2 and 3 and for which $3 \not| q - 1$. In that case $x \mapsto x^3$ is a bijection from $\mathbb{F}_q$ to itself and hence for each y there is a unique x for which $y^2 = x^3 + 1$. That means that $|E(\mathbb{F}_q)| = q + 1$ and $\alpha + \beta = 0$ so that $\alpha, \beta = \pm\sqrt{q}i$.

There is a way to get the number of points modulo p that we shall now describe, when the characteristic is $\neq 2, 3$. It uses two facts:

- For $z \in \mathbb{F}_q$ we have $\chi(z) = z^{(q-1)/2}$ in $\mathbb{F}_q$ (Exercise).
- If $q - 1 \not| n$, then $\sum_{z\in\mathbb{F}_q} z^n = 0$ and if it does, then $\sum_{z\in\mathbb{F}_q} z^n = -1$ (Exercise).

Exercise 58. i) Show that if $z \in \mathbb{F}_q$, then $\chi(z) = z^{(q-1)/2}$ in $\mathbb{F}_q$.

ii) Show that if $q - 1 \not| n$, then $\sum_{z\in\mathbb{F}_q} z^n = 0$ and if it does, then $\sum_{z\in\mathbb{F}_q} z^n = -1$.

We now compute in $\mathbb{F}_q$ and get

$$\sum_{x\in\mathbb{F}_q} \chi(x^3 + ax + b) = \sum_{x\in\mathbb{F}_q} (x^3 + ax + b)^{(q-1)/2}.$$

However we may expand

$$(x^3 + ax + b)^{(q-1)/2} = \sum_{i+j+k=(q-1)/2} \binom{(q-1)/2}{i, j, k} a^j b^k x^{3i+j}$$

and exchanging the order of summation we get

$$\sum_{x\in\mathbb{F}_q}(x^3+ax+b)^{(q-1)/2} = \sum_{i+j+k=(q-1)/2}\binom{(q-1)/2}{i,j,k}a^jb^k\sum_{x\in\mathbb{F}_q}x^{3i+j}.$$

However, the inner sum vanishes unless $q-1|3i+j$ and not both i and j are zero, but we always have $3i+j = 2i+(i+j) < 4((q-1)/2)$ which gives $3i+j = q-1$ so that the sum becomes

$$-\sum_{(q-1)/4\le i\le(q-1)/3}\binom{(q-1)/2}{i,q-1-3i,2i-(q-1)/2}a^{q-1-3i}b^{2i-\frac{q-1}{2}}.$$

This thus computes $|E(\mathbb{F}_q)|-(q+1)$ modulo the prime p for which $q=p^m$. Note that if $m=1$, then $|E(\mathbb{F}_q)|-(q+1)$ is an integer of size $\le 2\sqrt{p}$ so that if $p > 4\sqrt{p}$, the residue modulo p determines that integer.

8.1 The curve $y^2 = x^3 + 1$

Consider the curve $E\colon y^2 = x^3+1$, i.e., $a=0$ and $b=1$ in the form used above and assume that $q=p^1$. (We have already treated the case when $q\equiv 3 \bmod 4$.) A summand of the mod p formula then gives a zero contribution unless $3i = q-1$ and that term equals

$$\binom{(p-1)/2}{(p-1)/3,0,(p-1)/6} = \binom{(p-1)/2}{(p-1)/3} = \frac{p+2}{3}\frac{p+5}{3}\cdots\frac{p-1}{2}.$$

As we have seen, this formula determines the number of points when $p\ge 4\sqrt{p}$, i.e., $p\ge 16$. We can do better however. If $y^2 = x^3+1$ is a solution with $x\ne 0$, then there are exactly two other choices for x that also gives a solution with the same y as $3|p-1$ and there are therefore three cube roots of unity in $\mathbb{F}_p$. There are exactly two solutions with $x=0$ and thus $|E(\mathbb{F}_p)|\equiv 0 \bmod 3$. Thus we actually know $|E(\mathbb{F}_p)|$ modulo $3p$ and hence it is determined once $3p\ge 4\sqrt{p}$, i.e., $p\ge 16/9$.

Example 11. i) Consider $p=7$. We have $\binom{3}{2}=3$ and hence $|E(\mathbb{F}_p)|-8\equiv -3 \bmod 7$. On the other hand $|E(\mathbb{F}_p)|-8\equiv 0-8\equiv 1 \bmod 3$. This implies that $|E(\mathbb{F}_p)|-8\equiv 4 \bmod 21$ and hence by Hasse's theorem that $|E(\mathbb{F}_p)|-8=4$ and hence $|E(\mathbb{F}_p)|=12$. Furthermore, we get $\alpha+\beta=-4$ and hence α and β are roots of the polynomial x^2+4x+7 giving $\alpha,\beta = -2\pm\sqrt{3}i$.

ii) Consider $p=13$. We have $\binom{6}{4}=15$ and thus $|E(\mathbb{F}_p)|-14\equiv -15 \bmod 13$. Together with $|E(\mathbb{F}_p)|-14\equiv 1 \bmod 3$ this gives $|E(\mathbb{F}_p)|-14\equiv -2 \bmod 39$ and hence $|E(\mathbb{F}_p)|-14=-2$. This gives that α and β are roots of $x^2-2x+13$ and hence $\alpha,\beta = 1\pm 2\sqrt{3}i$.

It is seen in these two examples that $\alpha, \beta \in \mathbb{Z}[\zeta]$, where $\zeta = \frac{-1+\sqrt{3}i}{2}$ and $\mathbb{Z}[\zeta]$ denotes the ring of integral linear combinations $a + b\zeta$, $a, b \in \mathbb{Z}$.[1] This is no coincidence and is true for all $p \equiv 1 \bmod 3$ but is on the other hand very special for the equation $y^2 = x^3 + 1$. We shall not prove it here (but see exercise 60) but show that assuming it one can find a quicker way to determine $|E(\mathbb{F}_p)|$. We start by noticing that as $|\alpha| = \sqrt{p}$ we have $\alpha\overline{\alpha} = p$ and as $\alpha\beta = p$ this gives $\beta = \overline{\alpha}$. Hence, α and β are complex conjugates of each other. Furthermore, we have $\alpha\overline{\alpha} = p$ and from the theory of arithmetic in the ring $\mathbb{Z}[\zeta]$ this implies that α is determined up to complex conjugation (which only exchanges α and β) and multiplication by $\pm\zeta^i$, $i = 0, 1, 2$.

Exercise 59. Prove that if $\alpha, \beta \in \mathbb{Z}[\zeta]$ fulfil $\alpha\overline{\alpha} = p = \beta\overline{\beta}$ for a prime p, then $\beta = \pm\zeta^i\alpha$ or $\beta = \pm\zeta^i\overline{\alpha}$.

Exercise 60. Let p be a prime with $3|p-1$. Let $\chi_0\colon \mathbb{Z}/\mathbb{Z}p \to \mathbb{C}$ be the function with constant value 1 and $\chi_1\colon \mathbb{Z}/\mathbb{Z}p \to \mathbb{C}$ be the function taking only values 0, 1, or -1 and for which $\chi_1(x) \equiv x^{(p-1)/2} \bmod p$. Finally, let $\chi_2\colon \mathbb{Z}/\mathbb{Z}p \to \mathbb{C}$ be a function which only takes values 0, 1, ζ, and ζ^2, which is 0 on 0 and whose restriction to $(\mathbb{Z}/\mathbb{Z}p)^\times$ is a non-trivial group homomorphism into $\{1, \zeta, \zeta^2\}$. Let χ_3 be the complex conjugate of χ_2.

i) Show that χ_2 and χ_3 are the only functions fulfilling the condition that was imposed on χ_2.

ii) Show that the number of solutions to $x^2 = c$ equals $\chi_0(c) + \chi_1(c)$ and that the number of solutions to $x^3 = d$ equals $\chi_0(d) + \chi_2(d) + \chi_3(d)$.

iii) For $b \in \mathbb{Z}/\mathbb{Z}p$ show that the number of solutions to $y^2 = x^3 + b$ in $\mathbb{Z}/\mathbb{Z}p$ equals

$$p + \sum_{d \in \mathbb{Z}/\mathbb{Z}p} \chi_1(d+b)\chi_2(d) + \sum_{d \in \mathbb{Z}/\mathbb{Z}p} \chi_1(d+b)\chi_3(d)$$

and show that the two sums are complex conjugates of each other.

iii) (Difficult) Prove that

$$\sum_{d \in \mathbb{Z}/\mathbb{Z}p} \chi_1(d+b)\chi_2(d) \sum_{d \in \mathbb{Z}/\mathbb{Z}p} \chi_1(d+b)\chi_3(d) = p$$

if $b \neq 0$.

We shall now see that we may inject further information that determines α up to complex conjugation. To simplify the arguments we put $\mathrm{Tr}(\alpha) := \alpha + \overline{\alpha}$ for $\alpha \in \mathbb{Z}[\zeta]$. Note that Tr is additive with $\mathrm{Tr}(1) = 2$ and $\mathrm{Tr}(\zeta) = \mathrm{Tr}(\zeta^2) = -1$. If $\alpha = a + b\zeta$, $a, b \in \mathbb{Z}$, then $|E(\mathbb{F}_p)| = p + 1 - \mathrm{Tr}(\alpha) = p + 1 - (2a - b)$. The elements $(x, 0)$, where x is a root of $x^3 + 1$ gives the elements of order 2 of E. One such solution is $x = -1$ and there are two others as there are three cube roots of unity in $\mathbb{F}_p$. Hence $E(\mathbb{F}_p)$ contains a subgroup of order 4 and therefore 4 divides $|E(\mathbb{F}_p)|$. That means that $|E(\mathbb{F}_p)| = p + 1 - (2a - b) \equiv 0 \bmod 4$. In particular b is even. As

[1] They actually lie in $\mathbb{Z}[\sqrt{3}i]$ but that would not be true for instance for $y^2 = x^3 + 2$.

$2\zeta = -1+\sqrt{3}i$ this gives that $\alpha = c + d\sqrt{3}i$ for $c = a - b/2$ and $d = b/2$. On the other hand we know $0 \equiv |E(\mathbb{F}_p)| = p + 1 - 2c \bmod 3$ giving $2c \equiv 2 \bmod 3$, i.e., $c \equiv 1 \bmod 3$. It now turns out that these conditions cannot simultaneously be true for α and some $\pm\zeta^i\alpha$ distinct from α. Indeed, $\zeta(a + b\zeta) = -b + (a - b)\zeta$ which shows that $\alpha + \overline{\alpha}$ and $\zeta\alpha + \overline{\zeta\alpha}$ are integers that are congruent modulo 3, hence $\pm\zeta^i\alpha$ fulfils the mod 3 condition only if the plus sign is chosen. Similarly we have that $\mathrm{Tr}(\zeta(a+b\zeta)) = -a-b$ and if $2a-b \equiv -a-b \bmod 4$ we have $3a \equiv 0 \bmod 4$, i.e., $4|a$ but as $2|b$ we get $\alpha/2 \in \mathbb{Z}[\zeta]$ which would give $p = \alpha\overline{\alpha} = 4\mathbb{Z}$ which is not possible. In the same manner $\mathrm{Tr}(\zeta^2(a + b\zeta)) = -a + 2b$ giving that $2a - b \equiv -a + 2b \bmod 4$ implies $a \equiv b \bmod 4$ which again gives $2|a, b$.

We have thus seen that there is a unique (unordered) pair $\alpha, \overline{\alpha} \in \mathbb{Z}[\zeta]$ such that $\alpha\overline{\alpha} = p$, $\mathrm{Tr}(\alpha) \equiv p + 1 \bmod 12$, and that under these conditions $\alpha \in \mathbb{Z}[\sqrt{3}i]$. This means that $\alpha = a + b\sqrt{3}i$ and then $a^2 + 3b^2 = p$ and $a \equiv (p + 1)/2 \bmod 6$, in fact the implied condition $a \equiv 1 \bmod 3$ is enough to determine a.

Example 12. i) Assume again $p = 7$. We are then looking for a solution to $a^2+3b^2 = 7$ which can be reformulated as saying that $7 - 3b^2$ is a square. Checking for $b = 1$ gives the solution $(-2)^2 + 3 \cdot 1^2 = 13$ and $\alpha = -2 + \sqrt{3}i$.

ii) Assume $p = 13$. This time we want $13 - 3b^2$ to be a square. Checking for $b = 1$ and 2 gives the solution $1^2 + 3 \cdot 2^2 = 13$ and $\alpha = 1 + 2\sqrt{3}i$.

iii) Assume $p = 31$. Checking for $31 - 3b^2$ to be a square gives the solution $(-2)^2 + 3 \cdot 3^2$ and $\alpha = -2 + 3\sqrt{3}i$.

This method clearly cuts down on the time needed to compute the number points but when p gets large it is still quite slow (more precisely it requires $O(\sqrt{p})$ arithmetic operations on integers of the size $O(\sqrt{p})$). There is an algorithm of Cornacchia that is much faster (requiring $O(\log^n p)$ operations for a small n).

8.2 Normal forms and twists

We do not need to consider the general Weierstrass form for curves over a finite field. Consider first the case of a curve over a field of characteristic different from 2 and 3. (Note that we do not assume that the field is algebraically closed.) We can start by "completing the square", mapping $y \mapsto y - a_1/2x - a_3/2$, to find an equivalent curve for which the equation has y^2 as left-hand side, i.e., $a_1 = a_3 = 0$. We can then complete the cube, mapping $x \mapsto x - a_2/3$, to obtain an equation with $a_2 = 0$ leaving us with an equation of the form $y^2 = x^3 + ax + b$. We can then use the j-function to determine when two such curves are equivalent over the *algebraic closure* of the base field. However, that does not mean that two curves with the same j-invariants are equivalent over the base field $\mathbb{F}_q$. In fact, let $u \in \mathbb{F}_q$ be a non-square and consider the equation $uy^2 = x^3 + ax + b$. It is of course not in Weierstrass form but from this form it is clear that over a larger field where $u = v^2$ we can make the coordinate change

$y \mapsto v^{-1}y$ to transform this to the original form. On the other hand, the transformation $(x, y) \mapsto (ux, uy)$ transforms the equation $uy^2 = x^3 + ax + b$ to the Weierstrass form $y^2 = x^3 + u^{-2}x + u^{-3}b$. This is called the *quadratic twist* of the original curve. It can then be shown that, with two exceptions, this new curve is not equivalent to the original one over $\mathbb{F}_q$ and that these two curves represent the equivalence classes over $\mathbb{F}_q$ of curves that become equivalent over the algebraic closure of $\mathbb{F}_q$. The exceptions are the curves of j-invariant 0 and 1728. A curve with j-invariant 0 is one given by the equation $y^2 = x^3 + 1$. Apart from the quadratic twist one may also perform a cubic twist by picking a non-cube $u \in \mathbb{F}_q$ and considering $y^2 = ux^3 + 1$ which is transformed to $y^2 = x^3 + u^2$. More generally one may consider $y^2 = x^3 + b$, $b \neq 0$, and then all such curves are equivalent over the algebraic closure of $\mathbb{F}_q$ and those associated to a and a' are equivalent over $\mathbb{F}_q$ precisely when a and a' differ by multiplication by a sixth power.

Similarly, a curve with j-invariant 1728 is $y^2 = x^3 + x$ and twists, this time called bi-quadratic, are obtained by again considering $y^2 = x^3 + ux$ and two such twists are equivalent if they differ by a fourth power.

Exercise 61. Show that if E' is the quadratic twist of E, then $|E(\mathbb{F}_q)| - (q + 1) = -(|E'(\mathbb{F}_q)| - (q + 1))$.

The case of characteristic 3 is quite similar. If the j-invariant is different from $0 = 1728$ we can transform the curve into the form $y^2 = x^3 + ax^2 + b$ and we may perform a quadratic twist in the usual way. On the other hand, $y^2 = x^3 - x$ has j-invariant 0 and we may perform a bi-quadratic twist to obtain $y^2 = x^3 - ax$. There is however a further possible twist. We can consider $y^2 = x^3 - x - a$. If we pick a b, possibly in some extension of $\mathbb{F}_q$, such that $b^3 - b = a$, then a coordinate change $x \mapsto x - b$ transforms the equation into $y^2 = x^3 - x$. The most general form that still gives j-invariant 0 is $y^2 = x^3 - bx - a$. To describe the equivalent curves over $\mathbb{F}_q$ is a little bit involved.

The case of characteristic 2 is quite different. If $a_1 \neq 0$, then we may get rid of a_3 by translating in x and then scale y so that $a_1 = 1$ leading to $y^2 + xy = x^3 + a_2x^2 + a_4x + a_6$. By translating $y \mapsto y + b$, where $b \in \mathbb{F}_q$ fulfils $b^2 = a_6$ we can get rid of a_6, and by translating $y \mapsto y + tx$ we can replace a_2 by $a_2 + t^2 + t + a_2$. Over an extension field we can always choose t so that this makes $a_2 = 0$ and over $\mathbb{F}_q$ $z \mapsto z^2 + z$ is an additive map with kernel consisting of 0 and 1, so that its image has index 2 and there is an element $m \in \mathbb{F}_q$ not of the form $t^2 + t$ so that we may transform a_2 into 0 or m. Hence we have the basic form $y^2 + xy = x^3 + ax$ and its quadratic twist $y^2 + xy = x^3 + mx^2 + ax$ (quadratic this time as we can get rid of m after a quadratic extension).

Exercise 62. Show that if $m \in \mathbb{F}_q$ is not of the form $n^2 + n$ for $n \in \mathbb{F}_q$, then the curves $E\colon y^2 + xy = x^3 + ax$ and $E'\colon y^2 + xy = x^3 + mx^2 + ax$ fulfil $|E(\mathbb{F}_q)| - (q + 1) = -(|E'(\mathbb{F}_q)| - (q + 1))$.

If instead $a_1 = 0$, then one can scale x and y so that $a_3 = 1$ (if $a_3 = 0$ the discriminant is zero and we do not have an elliptic curve). We can complete the cube in x and hence get to the form $y^2 + y = x^3 + a_4x + a_6$. We can make the coordinate change $(x, y) \mapsto (x+b^2, y+bx)$ which changes the equation to $y^2+y = x^3+(b^4+b+a_4)x+b^6+a_4bb^2+a_6$, and over an extension field we can thus make $a_4 = 0$. We can then after that make also a_6 disappear by a coordinate change of the type $y \mapsto y+r$. Thus all the curves $y^2 + y = x^3 + a_4x + a_6$ are twists of the single one $y^2 + y = x^3$. To determine when two such twists are equivalent over the base field is reasonably involved.

9

Division polynomials

We shall now investigate the map from an elliptic curve to itself given by multiplication by a positive integer n. We start by investigating the condition that a point is mapped to 0, the identity element. First we consider the case when the base field is the field of complex numbers. For an abelian group A and a positive integer n we let ${}_nA$ denote $\{a \in A \mid na = 0\}$ and consider

$$n^2 \prod_{0 \neq a \in {}_n(\mathbb{C}/\Gamma)} \wp(z) - \wp(a),$$

where as usual $\Gamma = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$. This is clearly a meromorphic function with periods the elements of Γ. We shall now see that it is in fact the square of such a function. Indeed, if $a \in {}_n(\mathbb{C}/\Gamma)$ and $a \neq -a$, then $(\wp(z) - \wp(a))(\wp(z) - \wp(-a)) = (\wp(z) - \wp(a))^2$ and thus the pair $\{a, -a\}$ contributes a square. If n is odd that proves the claim. If n is even we are left with a being equal to one of the residues of $\omega_1/2$, $\omega_2/2$, or $(\omega_1 + \omega_2)/2$. However, it follows from Theorem 3.8 i) that $4\wp^3(z) - g_2\wp - g_3 = 4(\wp(z) - \wp(\omega_1/2))(\wp(z) - \wp(\omega_2/2))(\wp(z) - \wp((\omega_1 + \omega_2)/2))$ and hence $4(\wp(z) - \wp(\omega_1/2))(\wp(z) - \wp(\omega_2/2))(\wp(z) - \wp((\omega_1 + \omega_2)/2)) = \wp'^2(z)$. We thus can write

$$f_n^2(z) := n^2 \prod_{0 \neq a \in {}_n(\mathbb{C}/\Gamma)} \wp(z) - \wp(a),$$

where f_n is meromorphic with Γ as periods. To be more precise we have seen that $f_n(z)$ can be chosen to have the form $P_n(\wp(z))$ if n is odd and $1/2\wp'(z)P_n(\wp(z))$ if n is even, where the P_n are polynomials with leading coefficient n (and of degree $(n^2 - 1)/2$, resp. $(n^2 - 4)/2$). Our first result is a description of $\wp(nz)$ in terms of these functions.

Proposition 9.1. *The following relation holds for $m > n$:*

$$\wp(nz) - \wp(mz) = \frac{f_{m+n}(z) f_{m-n}(z)}{f_m^2(z) f_n^2(z)}$$

and in particular for $n > 1$

$$\wp(z) - \wp(nz) = \frac{f_{n+1}(z) f_{n-1}(z)}{f_n^2(z)}.$$

Proof. We have that the poles of $\wp(nz)$ are exactly the points of ${}_n(\mathbb{C}/\Gamma)$. It is clear that if f is a meromorphic function on $\mathbb{C}$, then the multiplicity of a zero or pole of $f(nz)$

at a is the same as that of $f(z)$ at na. Hence we get that the multiplicity of the poles of $\wp(nz)-\wp(mz)$ is 2 at each point of ${}_m(\mathbb{C}/\Gamma)\bigcup{}_n(\mathbb{C}/\Gamma)$ as the coefficient of $\wp(nz)$ in front of lowest order term is $1/n^2$ so that the $1/z^2$-terms for $\wp(nz)$ and $\wp(mz)$ never cancel. Hence there are in all $2(m^2+n^2-d^2)$ poles counted with multiplicity, where $d := (m,n)$. On the other hand, from the argument above it follows that $f_n(z)$ has a simple zero at each $0 \neq a \in {}_n(\mathbb{C}/\Gamma)$ and no other zeroes, so that $f_n^2(z)$ has a double zero at each of those elements and no other zeroes. Further, $\wp(nz)-\wp(mz)$ has a zero exactly at all $z \in \mathbb{C}/\Gamma$ for which $mz \neq 0$, $nz \neq 0$, and $nz = \pm mz$ as $\wp$ takes the same values only at z and $-z$ by the proof of Theorem 3.8. That means that there are zeroes at the $z \in \mathbb{C}/\Gamma$ for which $(m \pm n)z = 0$, but $mz, nz \neq 0$. If $m-n$ is odd, then $(m+n, m-n) = d$ and $(m+n)z = (m-n)z = 0$ is equivalent to $mz = nz = 0$. So the number of zeroes is $(m+n)^2 + (m-n)^2 - 2d^2 = 2(m^2+n^2-d^2)$ and the number of zeroes equals the number of poles counted with multiplicity and thus each zero has multiplicity 1.

If instead $m-n$ is even we have that $(m-n, m+n) = 2d$ and hence there are points for which $(m+n)z = (m-n)z = 0$, yet $mz, nz \neq 0$, i.e., those for which $2dz = 0$ but $dz \neq 0$. If a is such a point, then $\wp(n(z+a))-\wp(m(z+a))$ is an even function of z and thus $\wp(nz)-\wp(mz)$ has a zero with even multiplicity at a. If all the other zeroes are counted with multiplicity 1, we get in all $(m+n)^2+(m-n)^2-2(2d)^2+2((2d)^2-d^2) = m^2+n^2-d^2$ so again we have found the right multiplicities.

We have thus found the zeroes and the poles of $\wp(z)-\wp(nz)$ and we have also seen that $f_n(z)$ has a simple zero at all $0 \neq z \in \mathbb{C}/\Gamma$ with $nz = 0$ and it then follows that

$$\frac{f_{m+n}(z)f_{m-n}(z)}{f_m^2(z)f_n^2(z)}$$

has the same poles and zeroes (counted with multiplicity) as $\wp(nz)-\wp(mz)$. In fact the only possibly problematic case[1] is $z = 0$ but it is easy to see that the expansion of f_n at $z = 0$ starts off as

$$f_n(z) = (-1)^{n+1}\frac{n}{z^{n^2-1}} + \dots, \tag{9.1}$$

where the sign comes from the choice of sign for f_n in terms of the expansions as $P_n(\wp(z))$ resp. $\wp'(z)P_n(\wp(z))$. This means that the quotient between the right- and left-hand side of the formula to be proven has no poles or zeroes and hence by Liouville's theorem is a non-zero constant. That constant is determined by looking at the lowest term of the expansion around zero, using that $\wp(nz)$ has lowest term $1/n^2 1/z^2$ and the expansion just used for f_n. □

Exercise 63. Show directly that $\wp(nz)-\wp(mz)$ has a simple zero at the non-zero points where $(n \pm 1)z = 0$ and $2z \neq 0$ and a double zero where $(n \pm 1)z = 2z = 0$.

[1] This case does not have to be treated as if an elliptic function has just a single possible pole or zero, then it does not have any poles or zeroes as the sum of the multiplicities is zero. On the other hand we need to expand at 0 anyway to determine the constant.

Example 13. i) We have that $f_1 = 1$ (and $P_n = 1$).

ii) We have $f_2(z) = \wp'(z)$ and hence $P_2(x) = 2$.

iii) We have from the proposition that

$$\wp(z) - \wp(2z) = \frac{f_3(z)f_1(z)}{f_2^2(z)} = \frac{f_3(z)}{\wp'^2(z)}. \tag{9.2}$$

Now, from Exercise 27 ii) we get that

$$\wp(2z) = -2\wp(z) + \frac{1}{4}\left(\frac{\wp''(z)}{\wp'(z)}\right)^2$$

and by differentiating the differential equation (3.1) we get

$$2\wp''(z)\wp'(z) = (12\wp^2(z) - g_2)\wp'(z),$$

which gives

$$\wp''(z) = 6\wp^2(z) - \frac{1}{2}g_2.$$

Hence using once more (3.1),

$$\begin{aligned}\wp(z) - \wp(2z) &= 3\wp(z) - \frac{1}{4}\left(\frac{6\wp^2(z) - 1/2g_2}{\wp'(z)}\right)^2\\ &= \frac{3(4\wp^4(z) - g_2\wp^2(z) - g_3\wp(z)) - 1/4(36\wp^4(z) - 6g_2\wp^2(z) + 1/4g_2^2)}{\wp'^2(z)}\\ &= \frac{3\wp^4(z) - 3/2g_2\wp^2(z) - 3g_3\wp(z) - 1/16g_2^2}{\wp'^2(z)},\end{aligned}$$

and this gives

$$f_3(z) = 3\wp^4(z) - \frac{3g_2}{2}\wp^2(z) - 3g_3\wp(z) - \frac{g_2^2}{16}.$$

Thus we get $P_3(x) = 3x^4 - 3/2g_2x^2 - 3g_3x - 1/16g_2^2$.

iv) We have the formula

$$\wp(u+v) = -\wp(u) - \wp(v) + \frac{1}{4}\left(\frac{\wp'(u) - \wp'(v)}{\wp(u) - \wp(v)}\right)^2$$

by Exercise 27 i) and by changing the sign of v we get

$$\wp(u-v) = -\wp(u) - \wp(v) + \frac{1}{4}\left(\frac{\wp'(u) + \wp'(v)}{\wp(u) - \wp(v)}\right)^2.$$

Subtracting these two expressions we get

$$\wp(u+v) - \wp(u-v) = -\frac{\wp'(u)\wp'(v)}{(\wp(u) - \wp(v))^2}$$

and putting $u = 2z$ and $v = z$ this gives, using the determination of f_n, $n = 1, 2$,

$$\frac{\wp'(z)f_4(z)}{f_3^2(z)} = \frac{f_2(z)f_4(z)}{f_3^2(z)} = \wp(z) - \wp(3z) = \frac{\wp'(2z)\wp'(z)}{(\wp(2z)-\wp(z))^2} = \frac{\wp'(2z)\wp'^5(z)}{f_3^2(z)}$$

which gives $\wp'(2z)\wp'^4(z) = f_4(z)$. Differentiating (see Exercise 27 ii)) and using our formulas for $\wp''(z)$ and $\wp^{(3)}(z)$ we get

$$\begin{aligned}\wp'(2z)\wp'^4(z) =&\wp'(z)\Bigg(-\wp'^4(z) + 18\wp^2(z)\wp'^2(z) - \frac{3g_2}{2}\wp(z)\wp'^2(z)\\ &-54\wp^3(z) + \frac{27g_2}{2}\wp^2(z) - \frac{9g_2^2}{8}\wp(z) + \frac{g_2^3}{32}\Bigg),\end{aligned}$$

and using the differential equation for $\wp(z)$ we get

$$\begin{aligned}f_4(z) = \frac{1}{2}\wp'(z)\Bigg(&4\wp^6(z) - 5g_2\wp^4(z) - 20g_3\wp^3(z)\\ &-\frac{5g_2^2}{4}\wp^2(z) - g_2g_3\wp(z) + \frac{g_2^3}{16} - 2g_3^2\Bigg).\end{aligned}$$

This gives

$$P_4(x) = 4x^6 - 5g_2x^4 - 20g_3x^3 - \frac{5g_2^2}{4}x^2 - g_2g_3x + \frac{g_2^3}{16} - 2g_3^2.$$

We have now computed enough steps to compute all the f_n using a recursion formula that we are now about to prove.

Proposition 9.2. *We have for $m > n$ that*

$$f_{m+n}f_{m-n} = f_{m+1}f_{m-1}f_n^2 - f_{n+1}f_{n-1}f_m^2.$$

Proof. We have by Proposition 9.1 that

$$\wp(z) - \wp(nz) = \frac{f_{n+1}(z)f_{n-1}(z)}{f_n^2(z)} \quad \text{and} \quad \wp(z) - \wp(mz) = \frac{f_{m+1}(z)f_{m-1}(z)}{f_m^2(z)}$$

which gives

$$\wp(mz) - \wp(nz) = \frac{f_{m+1}(z)f_{m-1}(z)}{f_m^2(z)} - \frac{f_{n+1}(z)f_{n-1}(z)}{f_n^2(z)}.$$

However, the formula

$$\wp(mz) - \wp(nz) = \frac{f_{m+n}(z)f_{m-n}(z)}{f_m^2(z)f_n^2(z)}$$

is also part of Proposition 9.1. □

Applying this formula for $(m, n) = (n + 1, n - 1)$ and $(m, n) = (n + 1, n)$ gives

$$\wp'(z) f_{2n}(z) = f_n(z)\big(f_{n+2}(z) f_{n-1}^2(z) - f_{n-2}(z) f_{n+1}^2(z)\big), \tag{9.3}$$
$$f_{2n+1}(z) = f_{n+2}(z) f_n^3(z) - f_{n+1}^3(z) f_{n-1}(z). \tag{9.4}$$

Exercise 64. Show that the number of polynomial operations needed for computing P_n using these formulas is $O(\log^2 n)$.

9.1 Algebraic version

The formulas we obtained above would seem to have something to do only with the values of Weierstrass' $\wp$-function and its derivative. However, under the map $z \mapsto (\wp(z), \wp'(z))$ we have seen that $z \mapsto nz$ is taken to multiplication by n, in the sense of iterated addition, on the curve $E\colon y^2 = 4x^3 - g_2x - g_3$. Hence $(\wp(z), \wp'(z)) \mapsto \wp(nz)$ corresponds to the function $x \circ n$, the composition of multiplication by n, $n\colon E(\mathbb{C}) \to E(\mathbb{C})$, and $x\colon E(\mathbb{C}) \to \mathbb{P}^1(\mathbb{C})$. Similarly, $(\wp(z), \wp'(z)) \mapsto \wp'(nz)$ corresponds to $y \circ n$. Hence, all the functions involved make perfect algebraic sense and the question is if the formulas are true. The existence of f_n uses only the algebraic properties of the functions involved so that existence is assured. Before we continue let us set up the situation we shall consider. We stick for simplicity to a field of characteristic different from 2 and 3 and the elliptic curve $y^2 = x^3 + ax + b$. In the complex case we have the relations $x = \wp(z)$, $y = 1/2\wp'(z)$, $g_2 = -4a$, and $g_3 = -4b$. Sticking first to the case when the base field is the complex numbers, and using that every complex a and b comes in this way from a lattice $\Gamma_{\omega_1,\omega_2}$, we have that the formulas proved for $\wp$ immediately translate into identities, where we have put $x_n := x \circ n$:

$$\begin{aligned}
\psi_1 &= 1,\\
\psi_2 &= 2y,\\
\psi_3 &= 3x^4 + 6ax^2 + 12bx - a^2,\\
\psi_4 &= 4y(x^6 + 5ax^4 + 20bx^3 - 5a^2x^2 - 4abx - a^3 - 8b^2),\\
\psi_n^2 &= n^2 \prod_{u \in_n E(\mathbb{C})} (x - x(u)),\\
x - x_n &= \frac{\psi_{n-1}\psi_{n+1}}{\psi_n^2},\\
\psi_{m+n}\psi_{m-n} &= \psi_{m+1}\psi_{m-1}\psi_n^2 - \psi_{n+1}\psi_{n-1}\psi_m^2,\\
\psi_{2n+1} &= \psi_{n+2}\psi_n^3 - \psi_{n+1}^3\psi_{n-1},\\
2y\psi_{2n} &= \psi_n\big(\psi_{n+2}\psi_{n-1}^2 - \psi_{n-2}\psi_{n+1}^2\big).
\end{aligned} \tag{9.5}$$

Here we have put $\psi_{2n+1} := P_{2n+1}(x)$ and $\psi_{2n} := yP_{2n}(x)$. All these formulas make sense for any field K (where the defining formula for ψ_n needs to take the product over ${}_nE(\overline{K})$ where $\overline{K}$ is an algebraic closure of K) and we shall now see that they are, with some modifications in characteristic p, in fact true in that context. For the moment however we stick to the complex numbers as base field. We start by getting a formula for multiplication by n in E. For this we first need the following formula:

$$\wp'(nz) = \frac{f_{2n}(z)}{f_n^4(z)}. \tag{9.6}$$

This formula is proved as in Proposition 9.1; $\wp'(nz)$ is zero, with multiplicity 1 at the $z \in \mathbb{C}/\Gamma$ with $nz \in {}_2(\mathbb{C}/\Gamma) \setminus \{0\}$, i.e., $2nz = 0$ and $nz \neq 0$, and has triple poles at all the $z \in \mathbb{C}/\Gamma$ with $nz = 0$. It is easily verified that so does the right-hand side of the formula and one then uses (9.1) to get the value of the quotient.

We are now going to introduce some polynomials in x and y other than the ψ's.

$$\begin{aligned} \varphi_n &:= x\psi_n^2 - \psi_{n-1}\psi_{n+1}, \\ 4y\omega_n &:= \psi_{n+2}\psi_{n-1}^2 - \psi_{n-2}\psi_{n+1}^2. \end{aligned}$$

Note for these polynomials as well as for ψ_n we shall always reduce them modulo the relation $y^2 = x^3 + ax + b$ so that y appears at most to the first power.[2] We also want to be explicit about the dependency on a and b and note that ψ_n, φ_n, and ω_n are functions not only of x and y but also of a and b. We start by recording the fact that as functions of x, y, a and b they are of a simple form.

Proposition 9.3. *For all n, $\varphi_n \in \mathbb{Z}[a, b, x]$, $\psi_n \in \mathbb{Z}[a, b, x]$ if n is odd and $\psi_n \in y\mathbb{Z}[a, b, x]$ if n is even, $\omega_n \in y\mathbb{Z}[a, b, x]$ if n is even, and $\omega_n \in \mathbb{Z}[a, b, x]$ if n is odd. Furthermore, φ_n is monic of degree n^2 in x and ψ_n^2 is of degree $n^2 - 1$ with top coefficient n^2.*

Proof. For ψ_n this follows from the recursive formulas for ψ_{2n} and ψ_{2n+1} (where the fact that it is either a polynomial in x or y times such a polynomial is known a priori but also follows from the recursion). Using this, the first result for φ_n and ω_n follows from the definitions. The degree and top coefficient for ψ_n^2 follows from the definition of ψ_n whereas the statement for φ_n follows from this and the definition of φ_n in terms of the ψ. □

Keeping the assumption that we work over the complex numbers, we now show that if $z = (x, y)$ does not belong to ${}_nE(\mathbb{C})$, then we have that

$$n(x, y) = \left(\frac{\varphi_n}{\psi_n^2}, \frac{\omega_n}{\psi_n^3}\right). \tag{9.7}$$

[2] For computational purposes this may not always be the best thing to do.

Indeed, this follows from the formula

$$x_n = x - \frac{\psi_{n-1}\psi_{n+1}}{\psi_n^2}$$

for the x-component and from (9.6) and

$$y\psi_{2n} = \psi_n\bigl(\psi_{n+2}\psi_{n-1}^2 - \psi_{n+1}^2\psi_{n+2}\bigr)$$

for the y-component. Our aim is now to extend this formula both to any field (of characteristic different from 2 and 3) and to $z \in {}_nE(\mathbb{C})$. If we begin by considering the last problem we should remember that (x, y) is just a short form for the homogeneous coordinate $(x:y:1)$ and we could try to clear the denominators in our formula. That is certainly possible but at the price of getting ψ_n^3 in the last component and $\psi_n\varphi_n$ in the first which is somewhat awkward. We can get a nicer expression by considering yet another weighted projective space, this time with weights $(2, 3, 1)$, i.e., the quotient of $K^3 \setminus \{0\}$ under the equivalence relation $(x, y, z) \sim (\lambda^2 x, \lambda^3 y, \lambda z)$ for $\lambda \in K^\times$. This means that any class with $z \neq 0$ is equivalent to a unique tuple $(x, y, 1)$. Furthermore, the homogenisation of the equation $y^2 = x^3 + ax + b$ defining the elliptic curve is $y^2 = x^3+axz^4+bz^6$ which thus gives a well-defined subset of the weighted projective space. For $z \neq 0$ we get as we have seen the "finite points" of the elliptic curve. The "points at infinity" in this version then corresponds to $z = 0$ and the equation $y^2 = x^3$ and putting $\lambda = yx^{-2}$ gives $(x, y, 0) \sim (y^2x^{-3}, y^3x^{-6}y, 0) = (1, 1, 0)$ and hence there is exactly one point at infinity.

Remark 7. There are very good reasons to claim that this is indeed the most natural weighting of the variables defining an elliptic curve.

We now want to clear out denominators (using of course the equivalence relation for homogeneous coordinates) in (9.7) to get that

$$n(x:y:1) = (\varphi_n(x, y):\omega_n(x, y):\psi_n(x, y)).$$

Over the complex numbers and when $\psi_n(x, y) \neq 0$ this follows from (9.7) and the equivalence relation on homogeneous coordinates of weights $(2, 3, 1)$. Let us continue to work over the complex numbers but assume that $\psi_n(x, y) = 0$. In that case the right-hand side of our formula will be the point at infinity (as it should) if it is in fact defined. Clearly the problem is that possibly $\varphi_n(x, y)$, $\omega_n(x, y)$, and $\psi_n(x, y)$ could simultaneously be zero. We shall prove that this is not so and also show that the formula makes sense and is true over any field (of characteristic different from 2 and 3). One problem that has to be overcome is that the addition is defined by cases. The major such division is whether or not the points to be added are equal and it can not resolved. However, some smaller divisions can be overcome using $(2, 3, 1)$-homogeneous coordinates as we shall do now. The easy case is inversion. We have

$$-(x:y:z) = (x:-y:z) = (x:y:-z)$$

and this formula holds true for all points. Continuing to addition we have the non-homogeneous formula established previously

$$(x_1\!:\!y_1\!:\!1) + (x_2\!:\!y_2\!:\!1) = \left(\left(\frac{y_1 - y_2}{x_1 - x_2}\right)^2 - x_1 - x_2 \!:\! \frac{y_1 - y_2}{x_1 - x_2}(x_3 - x_1) + y_1 \!:\! -1\right),$$

provided that $x_1 \neq x_2$, where x_3 is the first coordinate of the result. We may first scale this by the scale factor $x_1 - x_2$ and then start with the points $(x_1 : y_1 : z_1)$ and $(x_2 : y_2 : z_2)$ where $z_1, z_2 \neq 0$, scale to get the last coordinates equal to 1, use the formula and then scale by the smallest possible power of $z_1 z_2$ to get just polynomials in the variables. This gives the following (unappetising) formula

$$\begin{aligned}
&(x_1\!:\!y_1\!:\!z_1) + (x_2\!:\!y_2\!:\!z_2) = \\
&\quad \big(x_1^2 x_2 z_2^2 + x_1 x_2^2 z_1^2 - 2y_1 y_2 z_1 z_2 + a x_1 z_1^2 z_2^4 + a x_2 z_1^4 z_2^2 + 2b z_1^4 z_2^4 : \\
&\qquad 3x_1 x_2^2 y_1 z_1 - 3x_1^2 x_2 y_2 z_1^2 + x_2^3 y_1 z_1^3 - x_1^3 y_2 z_2^3 + a x_1 y_1 z_1 z_2^6 - a x_2 y_2 z_1^6 z_2 \\
&\qquad + 3a x_2 y_1 z_1^3 z_2^4 - 3a x_1 y_2 z_1^4 z_2^3 + 4b y_1 z_1^3 z_2^6 - 4b y_2 z_1^6 z_2^3 : x_2 z_1^2 - x_1 z_2^2\big)
\end{aligned} \tag{9.8}$$

It is easily verified that this formula works for all points except in the case when $(x_1\!:\!y_1\!:\!z_1) = (x_2\!:\!y_2\!:\!z_2)$. If we do the same thing for the doubling formula we get

$$\begin{aligned}
2(x\!:\!y\!:\!z) = \big(&x^4 - 2ax^2z^4 - 8bx + a^2z^8 : x^6 + 5ax^4z^4 \\
&+ 20bx^3z^6 - 5a^2x^2z^8 - 4abxz^{10} - (8b^2 + a^3)z^{12} : 2yz^4\big)
\end{aligned} \tag{9.9}$$

and this formula works for all points on a curve.

We are now ready to formulate our general result.

Proposition 9.4. *Let E be the elliptic curve defined by $y^2 = x^3 + ax + b$ over some field K of characteristic different from 2 or 3, let $(r : s : 1) \in E(K)$, and let n be an integer ≥ 1. Then not all of $\varphi_n(r, s)$, $\omega_n(r, s)$, and $\psi_n(r, s)$ are zero and we have*

$$n(r\!:\!s\!:\!1) = (\varphi_n(r, s)\!:\!\omega_n(r, s)\!:\!\psi_n(r, s)).$$

Proof. We prove this by induction on n, $n = 1$ being easily verified. Suppose first that n is even, $n = 2m$. By induction what we need to prove is that

$$2(\varphi_m(r, s)\!:\!\omega_m(r, s)\!:\!\psi_m(r, s)) = (\varphi_n(r, s)\!:\!\omega_n(r, s)\!:\!\psi_n(r, s))$$

and we may use formula (9.9) to expand the left-hand side. That means that the left-hand side will have the form

$$(\Phi_n(r, s)\!:\!\Omega_n(r, s)\!:\!\Psi_n(r, s)),$$

where the Φ_n, Ω_n, and Ψ_n are polynomials in a, b, x, y obtained by substituting the polynomials φ_n, ω_n, and ψ_n in formula (9.9). We shall prove the more precise relationship $\Phi_n(r,s) = \varphi_n(r,s)$, $\Omega_n(r,s) = \omega_n(r,s)$, and $\Psi_n(r,s) = \psi_n(r,s)$ which then also will prove that not all of $\varphi_n(r,s)$, $\omega_n(r,s)$, and $\psi_n(r,s)$ are zero.

As we have the relation $y^2 = x^3 + ax + b$, Φ_n, Ω_n, and Ψ_n can be viewed as elements in $\mathbb{Z}[a,b,x,y]/(y^2 - (x^3 + ax + b))$ and so can φ_n, ω_n, and ψ_n. If we can show that $\Phi_n = \varphi_n$, $\Omega_n = \omega_n$, and $\Psi_n = \psi_n$ in this ring, then we can evaluate them at (r,s) and get what we want. We start by showing that $\Psi_n = \psi_n$. Indeed, it follows from (9.9) that $\Psi_n = 2\omega_m\psi_m = 1/y(\psi_{m+2}\psi_{m-1}^2 - \psi_{m+1}^2\psi_{m-2})\psi_m = y/y\psi_n$ (it's OK to divide by y as it is easily seen that $\mathbb{Z}[a,b,x,y]/(y^2 - (x^3 + ax + b))$ is a domain).

Choose now complex numbers A, B, and X such that they are algebraically independent, i.e., do not fulfil any non-trivial polynomial equation with rational coefficients. As $\mathbb{C}$ is algebraically closed we can find a $Y \in \mathbb{C}$ such that $Y^2 = X^3 + AX + B$. This gives us a ring homomorphism $\mathbb{Z}[a,b,x,y]/(y^2 - (x^3 + ax + b)) \to \mathbb{C}$ given by the condition that it take $(a,b,x,y) \mapsto (A,B,X,Y)$. I claim that it is injective. To begin with it is injective on $\mathbb{Z}[a,b,x]$ by assumption and from that it is easy to see that it is non-injective on $\mathbb{Z}[a,b,x,y]/(y^2 - (x^3 + ax + b))$ only if $x^3 + ax + b$ is a square in $\mathbb{Z}[a,b,x]$. Now $(X:Y:1) \in E'(\mathbb{C})$, where E' is the complex elliptic curve given by $y^2 = x^3 + Ax + B$. We do not have that $n(X:Y:1) = (1:1:0)$ because if so we would have $P_n(X) = 0$, but the coefficients of $P_n(X)$ are polynomials in A and B and A, B, and X are algebraically independent. Hence by (9.7) we get that

$$n(X:Y:1) = \left(\frac{\varphi_n(X,Y)}{\psi_n^2(X,Y)} : \frac{\omega_n(X,Y)}{\psi_n^3(X,Y)} : 1\right) = (\varphi_n(X,Y) : \omega_n(X,Y) : \psi_n(X,Y)),$$

and this together with the equality $\Psi_n(X,Y) = \psi_n(X,Y)$ gives $\Phi_n(X,Y) = \varphi_n(X,Y)$ and $\Omega_n(X,Y) = \omega_n(X,Y)$.

Exercise 65. Complete the argument that $\mathbb{Z}[a,b,x,y]/(y^2 - (x^3 + ax + b)) \to \mathbb{C}$ is injective as $x^3 + ax + b$ is not a square.

The case when $n = 2m + 1$ is almost identical. We write np, where $p := (r:s:1)$, as $mp + (m+1)p$ and note that $mp = (m+1)p$ gives $p = 0$ which is not true. Hence we can use (9.8) to expand $\Psi_n(r,s)$ as $\omega_m(r,s)\psi_{m+1}^2(r,s) - \omega_{m+1}(r,s)\psi_m^2(r,s)$ which expands to $\psi_m^3(r,s)\psi_{m+2}(r,s) - \psi_{n-1}(r,s)\psi_{n+1}^3(r,s)$ by the definition of ω and this then equals $\psi_n(r,s)$ by (9.5). The rest of the proof is then identical. □

Remark 8. One can turn this proof around and use the constructions as an inductive definition of the ψ_n and the φ_n and ω_n. With some work one can then use these formulas to express φ_n and ω_n in terms of ψ_n. This bypasses our use of "transcendental methods",[3] i.e., methods that use complex numbers and analysis. On the other hand one doesn't really need transcendental methods for the arguments we gave that seemingly used them. The major property that we used was that an elliptic function without zeroes and poles is a non-zero constant. Elliptic functions with given periods

[3] As opposed to algebraic methods.

are, as we have seen, exactly the rational functions in $\wp$ and $\wp'$ which in algebraic language are x and y. One may define algebraically the notion of pole and zero and then show that only constants have no zeroes and poles. One may in fact also show that the number of poles counted with multiplicity equals the number of zeroes. On the third hand there are a number of formulas that are most easily obtained by using transcendental methods.

10

Torsion points

We shall now use the previous section to study the group of torsion points (i.e., the points for which $np = 0$ for some $n \geq 1$, these form a subgroup) in $E(K)$ for an elliptic curve over an algebraically closed field. If we look at the case of $K = \mathbb{C}$ and pick a lattice Γ such that E is equivalent to $\mathbb{C}/\Gamma$, then we see that the torsion group is $\mathbb{Q}\Gamma/\Gamma$, which as Γ is a free abelian group of rank 2 is isomorphic to $(\mathbb{Q}/\mathbb{Z})^2$. That, of course, is equivalent to ${}_nE(\mathbb{C})$ being isomorphic to $(\mathbb{Z}/\mathbb{Z}n)^2$ for all non-zero integers n. We shall now show that almost the same is true over any algebraically closed field, the only difference is with the p-torsion for an algebraically closed field. We start by the following result.

Proposition 10.1. *Let E be an elliptic curve over an algebraically closed field K. Then, for every integer $n \geq 1$, $n\colon E(K) \to E(K)$ is surjective, i.e., $E(K)$ is a* divisible *abelian group. Furthermore, the kernel has at most n^2 elements.*

Proof. That the point at infinity is in the image is clear as $n\infty = \infty$, so it is enough to show that we can solve $nq = p$ for $p = (r : s : 1) \in E(K)$. It is in fact enough to find q so that $r = a/c^2$, where $nq = (a : b : c)$, since if $s \neq a/c^3$ then $s = -a/c^3$, as both points are on the curve, and then $n(-q) = -nq = p$. Now setting $q = (x : y : 1)$ we get $nq = (\varphi_n(x, y) : \omega_n(x, y) : \psi_n(x, y))$ and this leads us to the equation $\varphi_n(x, y) - r\psi_n^2(x, y) = 0$ and this is, by Proposition 9.3, a polynomial of degree n^2 in x and hence always has a solution as K is algebraically closed. We can then find a y by solving a quadratic equation so that $(x : y : 1) \in E(K)$. Such a solution can not have $\psi_n(x, y) = 0$ because if so then, as $\varphi_n(x, y) - r\psi_n^2(x, y) = 0$, we would also get $\varphi_n(x, y) = 0$, but $(\varphi_n(x, y) : \omega_n(x, y) : \psi_n(x, y))$ is a well-defined point on E and $(0 : 1 : 0)$ is not. Finally, as multiplication by n is a group homomorphism, the cardinality of the kernel is equal to the cardinality of the inverse image of any point (now that we know that it is surjective). We see that we get at most n^2 possible values for the x-coordinate, as $\varphi_n(x, y) - r\psi_n^2(x, y)$ is a polynomial of degree n^2 in x. However, if we choose a point whose y-coordinate is different from zero, then there is only one possible y for given x in the inverse image as $n(-p) = -np$. □

The fact that a group is divisible gives strong restrictions on the kernel under multiplication by an integer n. We summarise these restrictions in the following exercise.

Exercise 66. An abelian group is a *torsion group* if every element in it has finite order. It is a *p-torsion group*, p a prime, if every element in it has order a power of p. For an abelian group A its *torsion subgroup* is the set of elements of finite order. For an abelian group A its *p-torsion subgroup* is the set of elements of order a power of p. If $\{A_\alpha\}$ is a (possibly infinite) collection of subgroups of an abelian group A, then A is the *direct sum* of the A_α if for each $a \in A$ there is a unique tuple (a_α) with $a_\alpha \in A_\alpha$ and all but a finite number of the a_α equal to the zero element such that $a = \sum_\alpha a_\alpha$.

i) Show that the (p-)torsion subgroup of an abelian group is a subgroup.

ii) Show that a torsion group is a direct sum of its p-torsion subgroups for all primes p.

iii) Show that the (p-)torsion subgroup of a divisible group is divisible.

iv) Show that a torsion group is divisible precisely when its p-torsion subgroups are.

v) Show that a p-torsion group is divisible precisely when multiplication by p is surjective.

vi) Show that a p-torsion group for which the kernel of multiplication by p is finite is divisible precisely when there is a k such that the kernel of multiplication by p^m is isomorphic to $(\mathbb{Z}/\mathbb{Z}p^m)^k$ for all k.

It then follows from the theorem and the exercise that if E is an elliptic curve over an algebraically closed field K, then there is for each prime ℓ a natural number k_ℓ such that ${}_{\ell^m}E(K) \cong (\mathbb{Z}/\mathbb{Z}\ell^m)^{k_\ell}$ for all m and furthermore that $0 \leq k_\ell \leq 2$ for all ℓ. We shall show that $k_\ell = 2$ if ℓ is distinct from the characteristic of K and $k_\ell = 0, 1$ if it isn't.

To prepare for the proof let us first say a few words about the formulas that have been given for the group addition. We have one formula, (9.8), when the two arguments are distinct and a different one, (9.9), when they are equal. The problem is that even though one of them works for any pair of arguments, there is no guarantee that either of them will work simultaneously for, e.g., two pairs of arguments. The reason why we would prefer for something like that to be true is that we would like to work with algebraic formulas. There is a way of solving this problem however. We can do this by writing $a + b = (a + (b + x)) - x$. For the right-hand side we can use the addition, as opposed to the doubling, formula as soon as $a \neq b + x$, $a \neq b + x$, and $a + b + x \neq -x$. This then gives an algebraic formula (which of course will depend on x). To be able to use the same algebraic formula for any finite number of pairs we may use the following result.

Exercise 67. Show that for a finite number (a_i, b_i) of pairs of points on an elliptic curve E over an algebraically closed field K there is an $x \in E(K)$ such that $a_i \neq b_i + x$, $a_i \neq b_i + x$, and $a_i + b_i + x \neq -x$ for all i.

This means that we may at will assume that we have an algebraic formula that works for any finite number of pairs. Note incidentally, that if we make two choices of x's, then the formulas are also related algebraically on their common domains of definition.

The next step is to extend our algebraic formula for nP to the point at infinity. It is easily checked that the components of

$$(z^{2n^2}\varphi_n(x/z^2, y/z^3) : z^{3n^2}(x/z^2, y/z^3) : z^{n^2}\psi_n(x/z^2, y/z^3))$$

are polynomials in x, y, and z and that the first one is not divisible by z whereas the last one is. This means that for $(1:1:0)$ we do indeed get $(1:1:0)$ so that using this as our formula we have found an extension.

Our first goal is to show that ψ_n has no multiple roots if n is not divisible by the characteristic. When n is odd, ψ_n is a polynomial in x so it is at least clear what that means. When n is even we have that $\psi_n(x, y) = yP_n(x)$ and what our statement is supposed to mean is that $P_n(x)$ has no multiple roots and no roots in common with $x^3 + ax + b$. If this is true, then we get that over an algebraically closed field K there are n^2 points $P \in E(K)$ for which $nP = 0$. If we for the moment consider the case when n is odd, then what we have to prove is that $\psi_n(x)$ and $\psi_n'(x)$ have no root in common. Now, one way of computing the derivative[1] of a polynomial f is to consider $f(x+t)$ and start expanding it as a polynomial in t; $f(x+t) = f(x) + tf'(x) + \cdots$. As all powers of t strictly greater than 1 are to be ignored we may go one algebraic step further and consider the residue class δ of t in $K[t]/(t^2)$. One usually denotes this quotient ring $K[\delta]$ and calls it the *ring of dual numbers* over K. We then have $f(x+\delta) = f(x) + \delta f'(x)$. Note that we have a ring homomorphism $K[\delta] \to K$ taking $a + b\delta$ to a and we shall denote by $(r)_0$ the image of an element under it. We can use the dual numbers to deal efficiently with tangent spaces. Hence a *tangent vector* in the projective plane is an equivalence class of tuples $(x, y, z) \in K[\delta]^3$ such that $((x)_0, (y)_0, (z)_0) \neq (0, 0, 0)$ under the equivalence relation $(x, y, z) \sim (\lambda x, \lambda y, \lambda z)$ for $\lambda \in K[\delta]$ with $(\lambda)_0 \neq 0$ (or equivalently, λ is invertible). (We shall use the notation $(x:y:z)$ for its equivalence class and let $(x:y:z)_0 := ((x)_0 : (y)_0 : (z)_0)$ and call this point of $\mathbb{P}^2(K)$ the *starting point* of the tangent vector.) The set of tangent vectors with starting point $p \in \mathbb{P}^2(K)$ will be called the *tangent space* at p. The tangent vector that is equivalent to one where $x, y, z \in K$ will be called the *zero tangent vector*. In general if $x = x_0 + \delta x_1$, $y = y_0 + \delta y_1$, and $z = z_0 + \delta z_1$, then the vector space spanned by (x_0, y_0, z_0) and (x_1, y_1, z_1) depends only on the class of (x, y, z). It is 1-dimensional precisely when the tangent vector is the zero vector. When the tangent vector is non-zero we hence get a line in the projective plane which we shall call the *line spanned by the tangent vector*. Furthermore, if $f(x, y, z)$ is a homogeneous polynomial, then we say that a tangent vector $(x : y : z)$ is *tangent to* the curve defined by f if $f(x, y, z) = 0$ (in $K[\delta]$). The following exercise treats the most pertinent properties of tangent vectors.

Exercise 68. i) Show that the condition for being a tangent vector is independent of the choice of homogeneous representative.

ii) Assume that $p = (x : y : z)$ is a tangent vector for which the z-coordinate of p_0 is non-zero. Show that p may be written as $(x' : y' : 1)$ with uniquely determined

[1] In fact this is the best way to define it!

$x', y' \in K[\delta]$. Show that in this fashion tangent vectors at p_0 can be identified with K^2 and that if also the x- or y-coordinate of p_0 is non-zero, then the corresponding identifications with K^2 differ from the one using the z-coordinate by an invertible linear transformation.

iii) Show that for tangent vectors at $(x:y:z) \in \mathbb{P}^2(k)$ the vector space structure on the tangent space provided by the previous part is given by $\lambda(x+u\delta : y+v\delta : z+w\delta) = (x+\lambda u\delta : y+\lambda v\delta : z+\lambda w\delta)$ and $(x+u\delta : y+v\delta : z+w\delta)+(x+u'\delta : y+v'\delta : z+w'\delta) = (x+(u+u')\delta : y+(v+v')\delta : z+(w+w')\delta)$.

iv) Show that the set of tangent vectors starting at a fixed point on the curve given by a homogeneous polynomial is a sub-vector space of the space of all tangent vectors at the point.

v) Let $f(x,y,z)$ be a homogeneous polynomial and let $h_1(x,y,z)$, $h_2(x,y,z)$, and $h_3(x,y,z)$ be homogeneous polynomials all of the same degree and such that for no points on the curve $\{f(x,y,z)=0\}$ all the h_i are zero. Show that the formula $(x:y:z) \mapsto (h_1(x,y,z) : h_2(x,y,z) : h_3(x,y,z))$ gives a map both on points on the curve $\{f(x,y,z)=0\}$ for which the h_i are not all zero and on tangent vectors starting at such points. Show that this map, restricted to tangent vectors starting at a fixed point, is linear.

vi) Show that if p is a point of the curve $\{f(x,y,z)=0\}$, then a tangent vector q with $q_0 = p$ is tangent to the curve precisely when it is either the zero vector or the line spanned by it is a tangent line in the previously given sense.

Note now that if we have two tangent vectors p and q to an elliptic curve E, then we may define $p+q$ as follows: If $p_0 \neq q_0$, then the formula for addition of distinct points makes sense also for p and q. If not we choose $x \in E(K)$ such that the steps in $(p_0+(q_0+x))-x$ are defined. If we let X be the zero tangent vector starting at x, then we may define $p+q$ by $(p+(q+X))-X$. All the properties of this sum, such as independence of the choice of x and associativity, comes down to polynomial identities which must be true as they are true when evaluated for points over the algebraically closed field K. Hence we get the structure of an abelian group on the set $E(K[\delta])$ of tangent vectors to E. Furthermore, the map $q \mapsto q_0$ is a group homomorphism and in particular the set T_0E of tangent vectors starting at the identity element of E is a subgroup.

Exercise 69. i) Show that if $T_0E \times T_0E \to T_0E$ is the addition for this group structure, then it is a linear map.

ii) Show that if V is a vector space and $(V, *)$ is a group structure on V such that $*$ is a linear map, then $*$ is equal to the addition of the vector space structure of V.

It follows from this exercise that the map $T_0E \to T_0E$ induced by multiplication by n on E is multiplication by n in the vector space structure. In particular, if n is not divisible by the characteristic of the ground field K, then the map is an isomorphism. Assume now that $0 \neq \alpha \in E(K)$ with $n\alpha = 0$. Then $x \mapsto x+\alpha$ and $x \mapsto x-\alpha$ are inverses to each other and in particular they induce isomorphisms between T_0E and $T_\alpha E$, the space of tangent vectors starting at α. As we have that $n(x+\alpha) = nx$ we get that the map $T_\alpha E \to T_0 E$ induced by multiplication by n is also a bijection (provided that n is not divisible by the characteristic). We now have that $n(x:y:1) = (\varphi_n(x,y):\omega_n(x,y):\psi_n(x,y))$. Assume that there is a tangent vector $p = (a:b:c)$ starting at α and for which $\psi_n(a,b) = 0$, Thus $(\varphi_n(a,b):\omega_n(a,b):\psi_n(a,b)) = (c:d:0)$. By scaling we may assume that $(\varphi_n(a,b))_0 = (\omega_n(a,b))_0 = 1$. If $c = 1 + \delta c'$, then by scaling by $\lambda = 1 - \delta c'/3$ we may assume that $c'' = 0$. As $(c:d:0)$ is a $K[\delta]$-point of E we have that $d^2 = c^3 = 1$ and if $d = 1 + \delta d'$ this gives $2d' = 0$ and hence $d' = 0$. Thus we get that $n(a:b:1)$ is the zero tangent vector and by the bijectivity proved, p is the zero tangent vector.

Assume now that $\alpha = (\varepsilon : \phi : 1)$ with $\phi \neq 0$. Then for any $\tau \in K$ there is a (unique) $\sigma \in K$ such that $(\varepsilon + \delta\tau : \phi + \delta\sigma : 1)$ is a tangent vector to E. Indeed, if the equation for E is $y^2 = f(x)$, the condition that it be a tangent vector is that $(\phi + \delta\sigma)^2 = f(\varepsilon + \delta\tau)$; i.e., $\phi^2 + 2\delta\phi\sigma = f(\varepsilon) + \delta\tau f'(\varepsilon)$ and as α is a point on E this is equivalent to $2\delta\phi\sigma = \delta\tau f'(\varepsilon)$ which means $2\phi\sigma = \tau f'(\varepsilon)$. As $\phi \neq 0$ this determines σ uniquely. Choose now such a tangent vector with $\tau \neq 1$. If we evaluate it at ψ_n, we get $0 \neq (\phi + \delta\sigma)^g P_n(\varepsilon + \delta)$, with $g = 0, 1$, as n is even resp. odd. As $\phi \neq 0$ this gives $0 \neq P_n(\phi + \delta\sigma) = \delta P_n'(\varepsilon)$ and hence $P_n'(\varepsilon) \neq 0$ and P_n does not have multiple roots at ε.

If n is odd, then we do indeed never have $\phi = 0$ as the latter condition gives points of order 2. Hence $P_n(x) = \psi_n(x,y)$ is without multiple roots and as it has degree $(n^2-1)/2$ (as the top coefficient is $n \neq 0$) and as each root of $P_n(x)$ gives two α (by the fact that $\phi \neq 0$) we get that there are in all $2(n^2-1)/2 + 1 = n^2$ points with $np = 0$. If instead n is even we have, if still $\phi \neq 0$, again that $P_n'(\varepsilon) \neq 0$ and this gives in all $2(n^2-4)/2 = n^2 - 4$ points. Adding the three points with $\phi = 0$ and the point at ∞ gives a total of n^2 points. This almost proves the following.

Proposition 10.2. *Let E be an elliptic curve over an algebraically closed field K.*

i) *If n is not divisible by the characteristic of K, then ${}_nE(K)$ is isomorphic to $(\mathbb{Z}/\mathbb{Z}n)^2$.*

ii) *If the characteristic of K is a prime p, then there is an $r = 0, 1$ such that ${}_{p^k}E(K)$ is isomorphic to $(\mathbb{Z}/\mathbb{Z}p^k)^r$ for all k.*

Proof. The first part follows from the divisibility of $E(K)$ together with the fact that ${}_nE(K)$ has n^2 points. For the second part we need to prove that ${}_{p^k}E(K)$ has less than p^{2k} points. This however follows from the fact that the $(n^2-1)/2$ th, resp. $(n^2-4)/2$ th coefficient of $P_n(x)$ is n which is zero if p divides n. □

Remark 9. The statement is true also in characteristic 2 and 3.

Exercise 70. We used implicitly results on tangent vectors for weighted homogeneous coordinates that we have proven (mostly in exercises) only for ordinary homogeneous coordinates. Prove these results for weighted coordinates (with weights (2, 3, 1)) for characteristic different from 2 and 3.

One might reasonably wonder how to decide if r is equal to 0 or 1 in the case of positive characteristic p. We have $r = 0$ exactly when there are no points of order p in $E(K)$. For a curve over a finite field this is equivalent to p never dividing the order of $E(F)$ for any finite extension field F. We know how to compute the number modulo p of points; it is equal to

$$1 - \sum_{(q-1)/4 \le i \le (q-1)/3} \binom{(q-1)/2}{i, q-1-3i, 2i-(q-1)/2} a^{q-1-3i} b^{2i-\frac{q-1}{2}}.$$

If this is never zero not only for the field itself but also for all finite extension fields, then we conclude that $r = 0$. This may look like an infinite amount of work but the following exercise shows that it is not.

Exercise 71. i) Show that if

$$\alpha_n := \sum_{(q^n-1)/4 \le i \le (q^n-1)/3} \binom{(q^n-1)/2}{i, q^n-1-3i, 2i-(q^n-1)/2} a^{q^n-1-3i} b^{2i-\frac{q^n-1}{2}}$$

for $a, b \in \mathbb{F}_q$, then $\alpha_n = \alpha_1^n$.

ii) Show that $r = 0$ precisely when $\alpha_1 = 0$.

11

Lattice inclusions

Let us return to lattices in $\mathbb{C}$. We shall be interested in when one lattice is included in another. So let us assume that we have two lattices $\Gamma' \subseteq \Gamma \subset \mathbb{C}$. We shall also assume that for no $d > 1$ do we have $\Gamma' \subseteq d\Gamma$. This condition means (by for instance the structure theorem for finite abelian groups) that the quotient group Γ/Γ' is cyclic. For simplicity we shall assume that the order of this group is a prime p. Now, it is clear that $p\Gamma \subset \Gamma' \subset \Gamma$ so that Γ' corresponds bijectively to a subgroup of order p of $\Gamma/p\Gamma$. If we chose a basis for Γ we may identify the set of such subgroups with $\mathbb{P}^1(\mathbb{Z}/p)$ and hence they are described by homogeneous coordinates. If we specifically let $\Gamma = \mathbb{Z}+\mathbb{Z}\tau$, then we see that Γ' is of the form $\mathbb{Z}+\mathbb{Z}p\tau$ corresponding to $(1:0)$ or $\mathbb{Z}p+\mathbb{Z}(\tau+a)$ corresponding to $(a:1)$, for $0 \le a < p$. This gives in all $p+1$ lattices Γ_α for $\alpha \in \mathbb{P}^1(\mathbb{Z}/p)$. The question that we pose ourselves is what we can say about the $j_\alpha(\tau) := j(\Gamma_\alpha)$. They are clearly analytic functions of τ. Considering the expansion at ∞ we have for $\alpha = (a:1)$ that $j_\alpha(\tau) = j(\tau p + a/p)$ and for $\alpha = (1:0)$ that $j_\alpha(\tau) = j(p\tau)$. For the latter we thus have that as a function of q, $j_{(1:0)}(q) = j(q^p)$ and for the former that $j_{(a:1)}(q) = j(\zeta^a q^{1/p})$, where $\zeta = e^{2\pi i/p}$. (The last formula may seem somewhat dubious as the pth root of the complex number q is not well defined. However it should be interpreted in the sense that $q^{1/p} := e^{2\pi i\tau/p}$. Hence what the formula says is that $j_{(a:1)}(\tau)$ is obtained by substituting $\zeta^a q^{1/p}$ in the q-expansion for j.) Hence the q-expansion of j_α is in powers of q^p when $\alpha = (1:0)$ and in powers of $q^{1/p}$ when it is not. In the latter case this means that j_α most definitely is not a modular function. We shall now see that all is not lost. We start by noticing that for every $\alpha \in \mathbb{P}^1(\mathbb{Z}/p)$ there is an integer matrix M_α of determinant p such that $j_\alpha(\tau) = j(M_\alpha(\tau))$, where matrices act on $\mathbb{H}$ by the corresponding Möbius transformation. Indeed, we may put

$$M_\alpha := \begin{cases} \begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix} & \text{if } \alpha = (1:0), \\ \begin{pmatrix} 1 & a \\ 0 & p \end{pmatrix} & \text{if } \alpha = (a:1). \end{cases}$$

If now $M \in \mathrm{SL}_2(\mathbb{Z})$, then $j_\alpha(M(\tau)) = j(M_\alpha M(\tau))$. Note that for $N \in \mathrm{SL}_2(\mathbb{Z})$ we have $j(\tau) = j(N(\tau))$ and in particular $j(M_\alpha M(\tau)) = j(NM_\alpha M(\tau))$. We may then apply the following exercise.

Exercise 72. An *elementary integer* 2×2*-matrix* is a matrix of the form $\left(\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix}\right)$ or $\left(\begin{smallmatrix} 1 & 0 \\ n & 1 \end{smallmatrix}\right)$ for an integer n.

i) Show that every integer 2×2-matrix of non-zero determinant is of the form AM, where A is a product of elementary matrices and M is of the form $\left(\begin{smallmatrix} m & d \\ 0 & n \end{smallmatrix}\right)$, where m is a non-zero integer, n a positive integer and $0 \le d < 0$.

ii) Show that every integer 2×2-matrix of determinant p is of the form NM_α for unique $N \in \mathrm{SL}_2(\mathbb{Z})$ and $\alpha \in \mathbb{P}^1(\mathbb{Z})$.

iii) Show that for every $M \in \mathrm{SL}_2(\mathbb{Z})$ and every $\alpha \in \mathbb{P}^1(\mathbb{Z})$ there is an $N \in \mathrm{SL}_2(\mathbb{Z})$ such that $M_\alpha M = NM_\beta$, where $\beta = M^{-1}\alpha$ and where $\mathrm{SL}_2(\mathbb{Z})$ acts on $\mathbb{P}^1(\mathbb{Z}/p)$ by reduction modulo p and the natural action of $\mathrm{SL}_2(\mathbb{Z}/p)$ on $\mathbb{P}^1(\mathbb{Z}/p)$.

It follows from this exercise that for every $M \in \mathrm{SL}_2(\mathbb{Z})$ we have $j_\alpha(M(\tau)) = j_\beta(\tau)$, where $\beta = M^{-1}\alpha$. In particular the j_α are permuted under the action of $\mathrm{SL}_2(\mathbb{Z})$. Consider therefore the polynomial

$$\prod_{\alpha\in\mathbb{P}^1(\mathbb{Z}/p)} (t - j_\alpha).$$

Its coefficients are invariant under the action of $\mathrm{SL}_2(\mathbb{Z})$ which makes the following proposition highly plausible.

Proposition 11.1. i) *There is a* (*unique*) *polynomial* $\Phi_p(s,t)$ *in* t *and* j *with integer coefficients such that*

$$\Phi_p(t,j) = \prod_{\alpha\in\mathbb{P}^1(\mathbb{Z}/p)} (t - j_\alpha).$$

ii) *Seen as a polynomial in* s*, the* s^{p+1}*-coefficient of* Φ_p *is equal to* 1*, the constant coefficient is a monic polynomial in* t *of degree* $p+1$*, the* s^p*-coefficient is a monic polynomial in* t *of degree* p*, and the* s^i*-coefficients* $0 < i < p$ *are polynomials of degree* $\leq p$ *in* t*.*

iii) Φ_p *is an irreducible polynomial* (*as a polynomial in* s).

Proof. We start by some general preliminaries. We let L be the field (as it were) of meromorphic functions on $\mathbb{H}$. We have an action of $\mathrm{SL}_2(\mathbb{Z})$ on L by field automorphisms given by $Gf(\tau) := f(G\tau)$. We have a particular element j of L giving a ring homomorphism $\mathbb{C}[s] \to L$ taking $p(s)$ to $p(j)$. This map is injective, which is seen for instance by considering lowest terms of the q-expansion, and we denote the image $\mathbb{C}[j]$. By the injectivity we also get an extension to an embedding of the fraction field $\mathbb{C}(s)$ in L whose image is denoted $\mathbb{C}(j)$. Every element in $\mathbb{C}(j)$ is fixed under the action of $\mathrm{SL}_2(\mathbb{Z})$ and by Exercise 55 the elements of $\mathbb{C}(j)$ are exactly the elements of L fixed under $\mathrm{SL}_2(\mathbb{Z})$. We also have $j_\alpha \in L$ and as we shall see they are algebraic over $\mathbb{C}(j)$.

We have already noticed that the coefficients of $H(s) := \prod_{\alpha\in\mathbb{P}^1(\mathbb{Z}/p)}(s - j_\alpha)$ as polynomials in t are invariant under $\mathrm{SL}_2(\mathbb{Z})$. They are holomorphic functions on $\mathbb{H}$, being sums of products of such functions. If we consider the q-expansion, then each j_α has a q-expansion in $q^{1/p}$ with only a finite number of negative powers (in fact at most one). The same is then true for finite products and sums of such functions and hence for the coefficients of H. However, the fact that these coefficients are invariant under $\tau \mapsto \tau + 1$ implies that these expansions are in fact expansions in integral powers of q and hence are the q-expansions of these functions. As only a finite number of negative powers are present, we have proven that the coefficients are modular functions and hence rational functions in j by Exercise 55. As these coefficients are also holomorphic on $\mathbb{H}$ it is easy to see that the rational function is actually a polynomial. However, it is easier to do this directly. Thus, let f be a modular function which is holomorphic on $\mathbb{H}$. We show that it is a polynomial in j by induction on the order of the pole of f at ∞. If the order is zero, then f is constant and we are finished. If the order is $n > 0$ and the q-expansion of f starts as aq^{-n} we consider $f - aj^n$ which has lower order. This gives the existence of the polynomial $\Phi_p(s, t)$ (and the unicity follows from the fact that j is transcendental over $\mathbb{C}$).

We can use this argument to get the more precise result that Φ_p has integer coefficients. For this we first have to prove that the q-expansion of j has integer coefficients. Now we have

$$E_2(q) = 1 + 240 \sum_{k\geq 1} \sigma_3(k)q^k,$$
$$E_3(q) = 1 + 504 \sum_{k\geq 1} \sigma_5(k)q^k,$$

so that their q-expansions have integer coefficients. However, we also have

$$j(q) = 1728 \frac{E_2^3(q)}{E_2^3(q) - E_3^2(q)},$$

and as $E_2^3(q) - E_3^2(q) = 1728q + \cdots$ it will be enough to show that the coefficients of $E_2^3(q) - E_3^2(q)$ are divisible by 1728. The quickest way (for us at least) to do this is to quote the well-known result that

$$E_2^3(q) - E_3^2(q) = 1728q \prod_{k\geq 1}(1 - q^k)^{24}.$$

(See Exercise 76 for a more elementary proof.) Using this and the induction above it follows that the coefficients used when expressing f above as a polynomial in j will lie in the group generated by the coefficients of the q-expansion of f. Applied to $\Phi_p(s, t)$ this gives that the coefficients of Φ_p will lie in $\mathbb{Z}[\zeta]$, the ring generated by ζ. We shall now conclude by proving two things: The coefficients of Φ_p are rational numbers and $\mathbb{Z}[\zeta] \cap \mathbb{Q} = \mathbb{Z}$. For the first part we argue as follows. We start by noticing that as this just involves the coefficients of the q-expansion we may consider the q-expansion of j as a formal power series. Thus

$$\prod_{0\leq a<p} (s - j(\zeta^a q^{1/p}))$$

may be considered as a formal power series in $q^{1/p}$ whose coefficients lie in $\mathbb{Z}[\zeta][s]$. Now rewrite $s - j(q)$ as $q^{-1}(sq - qj(q))$ giving

$$\prod_{0\leq a<p} (s - j(\zeta^a q^{1/p})) = \prod_{0\leq a<p} \zeta^a q^{1/p} \prod_{0\leq a<p} (s\zeta q^{1/p} - q^{1/p} j(q^{1/p}))$$

and as $\prod_{0\le a<p}\zeta^a = (-1)^{p+1}$ we have that $\prod_{0\le a<p}\zeta^a q^{1/p} = (-1)^{p+1}q$ and hence it is enough to show that

$$\prod_{0\le a<p}(s\zeta^a q^{1/p} - q^{1/p}j(q^{1/p}))$$

as a power series in $q^{1/p}$ has coefficients in $\mathbb{Q}[s]$. For this we notice that $sq - qj(q) = 1 + O(q)$ and hence the logarithm of $s\zeta q^{1/p} - q^{1/p}j(q^{1/p})$ converges as a formal power series giving

$$\log\Big(\prod_{0\le a<p}(s\zeta^a q^{1/p} - q^{1/p}j(q^{1/p}))\Big) = \sum_{0\le a<p}\log\Big(s\zeta^a q^{1/p} - q^{1/p}j(q^{1/p})\Big).$$

Putting

$$\log(sq - qj(q)) = \sum_{k\ge 1} c_k q^k$$

this sum equals

$$\sum_{0\le a<p}\sum_{k\ge 1} c_k\zeta^{ak}q^{k/p} = \sum_{k\ge 1}\Big(\sum_{0\le a<p}\zeta^{ak}\Big)c_k q^{k/p}.$$

The sum $\sum_{0\le a<p}\zeta^{ak}$ being a geometric series equals $(\zeta^{kp}-1)/(\zeta-1) = 0$ if $p\nmid k$ and p if $p|k$. This results in the series

$$\sum_{k\ge 1} c_{pk}q^k,$$

and as its coefficients lie in $\mathbb{Q}[s]$ so does its exponential which implies that

$$\prod_{0\le a<p}(s\zeta q^{1/p} - q^{1/p}j(q^{1/p}))$$

also has its coefficients in $\mathbb{Q}[s]$.

As for the fact that $\mathbb{Z}[\zeta]\cap\mathbb{Q} = \mathbb{Z}$, this is a standard fact of algebraic number theory since $\mathbb{Z}[\zeta]$ consists of algebraic numbers. An ad hoc proof is given in Exercise 75. By multiplying $\prod_{0\le a<p}(s - j(\zeta^a q^{1/p}))$ by $t - j(q^p)$ we then get that the coefficients of Φ_p are integers.

Next we show the statements on the degrees and top coefficients of Φ_p as polynomial in s. That the s^{p+1}-coefficient is 1 is clear. For the constant term its q-expansion equals

$$(-1)^{p+1}j(q^p)\prod_{0\le a<p}j(\zeta^a q^{1/p}).$$

The lowest power of q appearing in this product is $(-1)^{p+1}\prod_{0\le a<p}\zeta^a q^{-p-1} = q{-}p - 1$ which means that the highest power of j that appears when writing it as a polynomial in j is $p+1$ and that the coefficient in front of that power is 1.

Consider similarly the s^p-coefficient in its q-expansion:

$$\sum_{0\leq a<p} j(\zeta^a q^{1/p}) + j(q^p).$$

The term with the lowest power of q is q^{-p} and hence the highest power of j is p with coefficient 1. Finally, considering the t^i-coefficient for $0 < i$ we see that the lowest possible q-exponent appearing is $-p - (p-i)/p$ and as we know that only integer exponents appear, the lowest possible exponent is in fact $-p$ giving p as the highest exponent of j.

What remains to show is the irreducibility. Assume that $\Phi_p(s,t)$ has a non-trivial factorisation $G(s,t)H(s,t)$. Since Φ_p is monic as a polynomial in s both factors have positive degree in s. We now note that all the j_α are distinct as they have distinct q-expansions. This means that the factorisation

$$G(s,j)H(s,j) = \prod_{\alpha\in\mathbb{P}^1(\mathbb{Z}/p)} (s - j_\alpha)$$

divides $\mathbb{P}^1(\mathbb{Z}/p)$ into two disjoint non-empty subsets S and T according to as j_α is a root of G or H. However, $\mathrm{SL}_2(\mathbb{Z})$ preserves G and H as they are polynomials in j and hence $\mathrm{SL}_2(\mathbb{Z})$ preserves S and T under its action by $\alpha \mapsto G^{-1}\alpha$. Now, the reduction modulo p $\mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/p)$ is surjective (Exercise 73) and $\mathrm{SL}_2(\mathbb{Z}/p)$ acts transitively[1] on $\mathbb{P}^1(\mathbb{Z}/p)$ and hence can not preserve non-trivial subsets. □

Exercise 73. Show that $\mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/p)$ is surjective.

Exercise 74. Show that there are polynomials $f_i \in \mathbb{Z}[a_1,\dots,a_i]$ for each i such that if $f(q) = 1 + \sum_{i\geq1} a_i q^i$ is a formal power series, then $\prod_{0\leq a<p} f(\zeta^a q^{1/p}) = 1 + \sum_{i\geq1} f_i(a_1,\dots,a_i)q^i$.

Exercise 75. i) Show that $\mathbb{Z}[\zeta]$ has 1, ζ, …, ζ^{p-2} as a $\mathbb{Z}$-basis.

ii) For $\alpha \in \mathbb{Z}[\zeta]$ define $\mathrm{Tr}(\alpha)$ by taking the trace of the matrix describing multiplication by α on $\mathbb{Z}[\zeta]$ in the basis 1, ζ, …, ζ^{p-2}. Show that $\mathrm{Tr}(\zeta^i) = 0$ for $0 < i < p$ and $\mathrm{Tr}(1) = p - 1$

iii) Conclude that $\mathbb{Z}[\zeta] \cap \mathbb{Q} = \mathbb{Z}$.

Let us consider $p = 2, 3$ as examples.

Example 14. For $p = 2$ we have[2] the following expansion:

$$\begin{aligned}
&(t - j(\sqrt{q}))(t - j(-\sqrt{q}))(t - j(q^2)) \\
&\quad = q^{-3} - (t^2 - 1488t + 159768)q^{-2} + (42987519t + 8509195260)q^{-1} \\
&\qquad + t^3 - 2232t^2 + 40492979352t - 151107596045760 \\
&\qquad - (42987520t^2 - 8527520120556t - 1668458962374390)q + \mathrm{O}(q^2).
\end{aligned}$$

[1] I.e., every element may be taken to any other element.

[2] For the calculations see http://www.math.su.se/~teke/undervisning/Elliptisk.nb for $p = 2$ and http://www.math.su.se/~teke/undervisning/j.mg for $p = 3$.

This can then be used to express this polynomial as a polynomial in t and j,

$$\begin{aligned}
&(t-j(\sqrt{q}))(t-j(-\sqrt{q}))(t-j(q^2))\\
&\quad = j^3(q)-(t^2-1488t+162000)j^2(q)\\
&\qquad + (1488t^2+40773375t+8748000000)j(q)\\
&\qquad + t^3-162000t^2+8748000000t-157464000000000
\end{aligned}$$

thus giving

$$\begin{aligned}
\Phi_2(s,t) = {} & s^3+t^3-s^2t^2+1488(s^2t+st^2)-162000(s^2+t^2)\\
& + 40773375st+8748000000(s+t)-157464000000000.
\end{aligned}$$

Similarly one has

$$\begin{aligned}
\Phi_3(s,t) = {} & s^4+t^4-s^3t^3+2232(s^3t^2+s^2t^3)-1069956(s^3t+st^3)\\
& + 36864000(s^3+t^3)+2587918086s^2t^2+8900222976000(s^2t+st^2)\\
& + 452984832000000(s^2+t^2)-770845966336000000st\\
& + 1855425871872000000000(s+t).
\end{aligned}$$

We shall now use the results obtained so far to obtain a symmetry result that can be suspected from the examples above.

Proposition 11.2. *We have that* $\Phi_p(s,t)=\Phi_p(t,s)$.

Proof. We have by construction that $\Phi_p(j(p\tau'),j(\tau'))=0$ for all $\tau'\in\mathbb{H}$ and by putting $\tau'=\tau/p$ we get that $\Phi_p(j(\tau),j(\tau/p))=0$, i.e., that $\Phi_p(j,j_{(0:1)})=0$. That means that as polynomials in L, the field of meromorphic functions on $\mathbb{H}$, $\Phi_p(j,t)$ is divisible by $t-j_{(0:1)}$ but as it is a polynomial with coefficients in $\mathbb{C}(j)$ it must be divisible by the minimal polynomial of $j_{(0:1)}$ which by Proposition 11.1 means that it is divisible by $\Phi_p(t,j)$. However, again by Proposition 11.1, $\Phi_p(j,t)$ is monic of degree $p+1$ just as $\Phi_p(t,j)$ and hence they must be equal. □

Exercise 76. i) Show that $\sigma_3(n)\equiv\sigma_5(n)$ mod 12 for all n.

ii) Show directly that $E_2^3\equiv E_3^2$ mod 1728.

Exercise 77. i) Show that any integer 2×2-matrix with non-zero determinant is of the form GAG', where G and G' are products of elementary matrices and A has the form $\left(\begin{smallmatrix} m & 0\\ 0 & n\end{smallmatrix}\right)$ with $m>0$ and $n|m$.

ii) Give another proof of the fact that $\mathrm{SL}_2(\mathbb{Z})$ permutes the j_α transitively.

Exercise 78. Extend the definition and the properties of Φ_p to Φ_n for any $n\geq 2$.

11.1 Complex multiplication

In this section we pose ourselves the following question having fixed $n > 1$. For which lattices $\Gamma \subset \mathbb{C}$ does there exist a sublattice Γ' equivalent to Γ such that Γ/Γ' is cyclic of order n? We start with the more direct answer to this question. That Γ' is equivalent to Γ means that there is a $\lambda \in \mathbb{C}$ such that $\Gamma' = \lambda\Gamma$ so that $\lambda\Gamma \subseteq \Gamma$. We may assume that $\Gamma = \mathbb{Z} + \mathbb{Z}\tau$ for $\tau \in \mathbb{H}$ and then the fact that $\lambda = \lambda \cdot 1 \in \Gamma$ implies that $\lambda = a + b\tau$ for $a, b \in \mathbb{Z}$. We must have $b \neq 0$, as if not $\Gamma/\lambda\Gamma$ equals $\Gamma/a\Gamma$ which is not cyclic (unless $a = \pm 1$ in which case its order is not > 1). We also have $\lambda\tau \in \Gamma$ which means that there are integers $c, d \in \mathbb{Z}$ such that $(a + b\tau)\tau = c + d\tau$ which gives that $b\tau^2 + (a - d)\tau - c = 0$ and, as $b \neq 0$ and as τ is not real, that $\tau \in \mathbb{Q}(\sqrt{-D})$ for some square-free integer $D > 0$. The fact that $\lambda = a + b\tau$ implies that also $\lambda \in \mathbb{Q}(\sqrt{-D})$. However, λ fulfils stronger conditions. The following exercise gives the necessary preliminaries for establishing these conditions.

Exercise 79. The *characteristic polynomial* of $\lambda \in \mathbb{Q}(\sqrt{-D})$ is the characteristic polynomial of the matrix in some basis of the linear transformation given by multiplication by λ on $\mathbb{Q}(\sqrt{-D})$ considered as a (2-dimensional) $\mathbb{Q}$-vector space.

i) Show that the characteristic polynomial is independent of the choice of basis.

ii) Show that the trace of multiplication by λ equals $\lambda + \bar{\lambda}$ and the determinant equals $\lambda\bar{\lambda}$ and that the characteristic polynomial equals $t^2 - (\lambda + \bar{\lambda})t + \lambda\bar{\lambda}$

iii) Show that if $t^2 - at + b$ is the characteristic polynomial of λ, then $\lambda^2 - a\lambda + b = 0$ in $\mathbb{Q}(\sqrt{-D})$.

iv) Show that if there is a finitely generated subgroup $\Gamma \subset \mathbb{Q}(\sqrt{-D})$ such that $\lambda\Gamma \subseteq \Gamma$, then the characteristic polynomial of λ has integer coefficients.

Exercise 80. Let $M \in M_2(\mathbb{Z})$ be a matrix of non-zero determinant. Show that the quotient $\mathbb{Z}^2/M(\mathbb{Z}^2)$ is a finite group of order $\det M$.

It follows from this exercise that $\lambda + \bar{\lambda}, \lambda\bar{\lambda} \in \mathbb{Z}$. This is true not just for λ but also for any complex number which multiplies Γ into itself. The set of such numbers form a subring of $\mathbb{Q}(\sqrt{-D})$ which is finitely generated as an abelian group because it is also a subring of the ring of endomorphisms of Γ which in turn is isomorphic to $M_2(\mathbb{Z})$. This motivates the following definition.

Definition 11.3. An *imaginary quadratic order* is a subring of $\mathbb{Q}(\sqrt{-D})$ for some positive square-free integer D which is finitely generated as an abelian group.

Thus what we have proved is that for a given lattice $\Gamma \subset \mathbb{C}$, the set of $\lambda \in \mathbb{C}$ such that $\lambda\Gamma \subseteq \Gamma$ either consists just of the integers or is an imaginary quadratic order. It is therefore clear that the following exercise should be relevant.

Exercise 81. If D is a positive integer $\equiv -1, 0$, let R_D be $\mathbb{Z} + \mathbb{Z}\alpha$, where $\alpha = \sqrt{-D/4}$ if $D \equiv 0 \bmod 4$ and $\alpha = \frac{1+\sqrt{-D}}{2}$ if $D \equiv -1 \bmod 4$.

Show that R_D is an imaginary quadratic order and that all such orders are obtained in this way.

For a lattice $\Gamma \subset \mathbb{C}$ we define the *set of complex multiplications* of Γ, $\mathrm{End}_{\mathbb{C}}(\Gamma)$, as $\{\lambda \in \mathbb{C} \mid \lambda\Gamma \subseteq \Gamma\}$. We shall say that Γ *has complex multiplication* if $\mathrm{End}_{\mathbb{C}}(\Gamma) \neq \mathbb{Z}$ and then, by Exercise 81, $\mathrm{End}_{\mathbb{C}}(\Gamma) = R_D$ for some $D > 0$. We also let $\mathcal{C}_D$ denote the set of equivalence classes of lattices with $\mathrm{End}_{\mathbb{C}}(\Gamma) = R_D$. Our first goal is to show that the polynomial $\prod_{\Gamma \in \mathcal{C}_D}(t - j(\Gamma))$ is a polynomial with integer coefficients. The following lemma gathers some of the results needed for this. First let us introduce some definitions. We say that $\lambda \in R_D$ is *primitive* if it is not divisible by an integer > 1 as an element of R_D. A *primitive norm* of R_D is an integer of the form $N(\lambda) := \lambda\bar{\lambda}$ for a primitive element of R_D. The *primitive norm spectrum* of R_D is the set of primitive norms.

Lemma 11.4. i) *If* $0 \neq \lambda \in \mathrm{End}_{\mathbb{C}}(\Gamma)$ *for a lattice* $\Gamma \subset \mathbb{C}$*, then the order of* $\Gamma/\lambda\Gamma$ *is* $N(\lambda)$.

ii) *If* Γ *has complex multiplication, then* $\lambda \in \mathrm{End}_{\mathbb{C}}(\Gamma)$ *is primitive precisely when for no* $d > 1$ *do we have* $\lambda\Gamma \subseteq d\Gamma$.

iii) *If* $\lambda \in \mathrm{End}_{\mathbb{C}}(\Gamma)$ *is primitive with* $n := N(\lambda) > 1$*, then* $\Phi_n(j(\Gamma), j(\Gamma)) = 0$. *Conversely, if* $\Phi_n(j(\Gamma), j(\Gamma)) = 0$*, then there is a primitive* $\lambda \in \mathrm{End}_{\mathbb{C}}(\Gamma)$ *with* $n = N(\lambda)$.

iv) $-\Phi_n(t,t)$ *is a monic integer polynomial.*

v) *If* R_D *and* $R_{D'}$ *have the same primitive norm spectrum, then* $D = D'$.

Proof. Starting with i) it follows directly from Exercise 79, which gives that the determinant of multiplication $\lambda\colon \Gamma \to \Gamma$ is $N(\lambda)$, and Exercise 80.

As for ii) it is clear that if λ is not primitive, then $\lambda\Gamma \subseteq d\Gamma$ for some integer $d > 1$. Conversely, if $\lambda\Gamma \subseteq d\Gamma$, then $\lambda/d\Gamma \subseteq \Gamma$ and hence $\lambda/d \in \mathrm{End}_{\mathbb{C}}(\Gamma)$ and $\lambda = d \cdot \lambda/d$.

Continuing with iii) it follows from i) and ii) that if λ is primitive, then $\Gamma/\lambda\Gamma$ is cyclic of order n. That means that $j(\lambda\Gamma) = j(\Gamma)$ is one of the roots of $\Phi_n(t, j(\Gamma))$, i.e., $\Phi_n(j(\Gamma), j(\Gamma)) = 0$. Conversely, if $\Phi_n(j(\Gamma), j(\Gamma)) = 0$, then by construction there is a sublattice $\Gamma' \subseteq \Gamma$ with Γ/Γ' cyclic of order n such that $j(\Gamma') = j(\Gamma)$. This implies by Proposition 6.3 that $\Gamma' = \lambda\Gamma$ for some $\lambda \in \mathbb{C}$. This implies that $\lambda\Gamma \subseteq \Gamma$ and hence that $\lambda \in \mathrm{End}_{\mathbb{C}}(\Gamma)$ and by i) and ii) this implies that λ is primitive and $N(\lambda) = n$.

Assume finally that the primitive root spectra of R_D and $R_{D'}$ are the same. If $4|D$ the primitive norm spectrum of R_D consists of all integers > 1 of the form $a^2 + b^2d$, where $d = D/4$ and $(a, b) = 1$. Its least element is clearly d or 2 if $d = 1$. Similarly, when $4 \not| D$ the primitive norm spectrum are the integers > 1 of the form $(a^2 + b^2D)/4$ with $a \equiv b \bmod 2$ and $(a, b) = 1, 2$. This time the least element is $(D + 1)/4$ when $D > 3$ and 3 if $D = 3$. Hence, by looking at the least element we see that if both D and D' are divisible by 4 or none of them is, then the spectra differ except when $D = 3$ and $D = 11$ or $D = 4$ and $d' = 8$ (or, of course vice versa). In the first exceptional case we see that $4 = (2^2 + 2^2 3)/4$ is in the spectrum of R_3 but not of R_{11}. In the second we have that $5 = 1^2 + 2^2$ is in the spectrum of R_4 but not in that of R_8.

We are hence left with the case when, say, $D = 4d$ and $D' = 4d' - 1$ and to begin with the minimal elements of the spectra have to be equal. Excluding for the moment the cases when either $d = 1$ or $d' = 1$ this means that $d = d'$. However, in that case $(D' + 9)/4 = d' + 2$ is in the spectrum of $R_{D'}$ but not in that of R_D. When $d = 1$ we get instead that $d' = 2$, but in that case $5 = 1^2 + 2^2$ belongs to the spectrum of R_4 but not to that of R_7. When $d' = 1$ we get $d = 3$ and then $4 = 1^2 + 1^2 3$ is in the spectrum of R_{12} but is not in the *primitive* spectrum of R_3, as the only solutions to $a^2 + 3b^2 = 16$ are $a, b = \pm 2$. □

For each positive integer D we now define a monic integer polynomial $H_D(t)$ as follows:

We start by constructing the greatest common divisor $H'_D(t,t)$ of the $-\Phi_n(t,t)$, where n runs over the primitive norm spectrum of R_D. (Note that as the $-\Phi_n(t,t)$ are monic integer polynomials this common divisor is a monic integer polynomial by Gauss' lemma.) Then for each n that is *not* in the primitive norm spectrum of R_D we remove from $H'_D(t)$ all the factors it has in common with $-\Phi_n(t,t)$. The resulting polynomial (which again by Gauss' lemma is monic integral) is $H_D(t)$.

Theorem 11.5. *The roots of $H_D(t)$ are exactly the j-invariants of the elements of $\mathcal{C}_D$.*

Proof. By construction $j(\Gamma)$ is a root precisely when $\mathrm{End}_{\mathbb{C}}(\Gamma)$ has the same primitive norm spectrum as R_D. However, by Lemma 11.4 v) this means that $\mathrm{End}_{\mathbb{C}}(\Gamma) = R_D$. □

11.1.1 Complex multiplication lattices. We shall now consider the determination of elements of $\mathcal{C}_D$. We choose a lattice $\Gamma = \mathbb{Z} + \mathbb{Z}\tau$ with $\mathrm{End}_{\mathbb{C}}(\Gamma) = R_D$. We furthermore assume that τ lies in the fundamental domain with the added conditions that $-1/2 < \Re(\tau) \leq 1/2$ and $\Re(\tau) \geq 0$ if $|\tau| = 1$ (which makes the lattice be unique in its equivalence class). We have seen that τ is the root of a quadratic polynomial with rational coefficients. We can specify a unique equation by demanding that $a\tau^2 + b\tau + c =$ for integers a, b, c with $a > 0$ and $(a, b, c) = 1$. This implies that

$$\tau = \frac{b + \sqrt{-D}}{2a},$$

where $D = 4ac - b^2$ (where $D > 0$ as τ is not real). The condition that τ belong to the fundamental domain means

$$-a < b \leq a,$$
$$b^2 + D \geq 4a^2$$

(with the added conditions that $b \geq 0$ if $b^2 + D = 4a^2$). This implies that $D \geq 3a^2$ so that for fixed D there are only a finite number of τ's and furthermore that it is very simple to find all the possible τ's.

Exercise 82. Show that if τ is of this type, then $\mathrm{End}_{\mathbb{C}}(\Gamma) = R_D$ so that the class of Γ is in $\mathcal{C}_D$.

Example 15. Consider the case $D = 23$. Thus $23 \geq 3a^2$ which gives $a = 1, 2$ and $a = 1$ gives $b = 0, 1$ but $b = 0$ that contradicts $D = 4ac - b^2$ and $a = 2$ gives $b = 0, \pm 1$ but again $b = 0$ that contradicts $D = 4ac - b^2$. This gives the solutions

$$\tau = \frac{1+\sqrt{-23}}{2}, \frac{1+\sqrt{-23}}{4}, \frac{-1+\sqrt{-23}}{4}.$$

We now go on to discuss how to compute the j-invariants. We know that $\prod_{\tau \in \mathcal{C}_D}(t - j(\tau))$ is a monic polynomial with integer coefficients. We have seen (at least when D is not too big) that we may enumerate the τ's in question. We may then compute $j(\tau)$ numerically using for instance[3] the q-expansion with enough precision so that the error in computing the coefficients of $\prod_{\tau \in \mathcal{C}_D}(t - j(\tau))$ is less than 0.5 which then identifies them as integers. (Note that as $\Im(\tau) \geq \sqrt{3}/2$ we will have $|q| \leq e^{-\pi\sqrt{3}} \approx 0.0043$ which makes convergence very fast.)

Example 16. i) We have that[4]

$$-\Phi_2(t,t) = t^4 - 2978t^3 - 40449375t^2 - 17496000000t + 157464000000000$$

and if we factor it we get

$$-\Phi_2(t,t) = (t - 8000)(t - 1728)(t + 3375)^2.$$

We recognise 1728 as $j(i)$ and indeed $N(1+i) = (1+i)(1-i) = 2$ so that $(1+i)\Gamma$ is of index 2 in $\Gamma = \mathbb{Z} + \mathbb{Z}i$. Furthermore, we have that $N(\sqrt{-2}) = 2$ and $N((1+\sqrt{-7})/2) = 2$. That means that $j(\sqrt{-2})$ and $j((1+\sqrt{-7})/2)$ are roots of $-\Phi_2(t,t)$ as the lattices $\mathbb{Z} + \mathbb{Z}\sqrt{-2}$ and $\mathbb{Z} + \mathbb{Z}(1+\sqrt{-7})/2$ are stable under multiplication with $\sqrt{-2}$ resp. $(1+\sqrt{-7})/2$. This accounts for all the three roots of $-\Phi_2(t,t)$. On the other hand making approximate calculations we have

$$\begin{aligned} j(i) &= 1728.000\ldots, \\ j(\sqrt{-2}) &= 8000.000\ldots, \\ j((1+\sqrt{-7})/2) &= -3375.000\ldots, \end{aligned}$$

and we have thus proved that $j(\sqrt{-2}) = 8000$ and $j((1+\sqrt{-7})/2) = -3375$.

ii) Consider again the case $D = 23$. This time we have

$$\begin{aligned} j\left(\frac{1+\sqrt{-23}}{2}\right) &\approx 3493225.6999699, \\ j\left(\frac{-1+\sqrt{-23}}{4}\right) &\approx 737.849984 + 1764.018938i, \\ j\left(\frac{1+\sqrt{-23}}{4}\right) &\approx 737.849984 - 1764.018938i. \end{aligned}$$

[3] There are more efficient methods.

[4] For this and subsequent calculations see as usual http://www.math.su.se/~teke/undervisning/Elliptisk.nb.

Note that it is clear a priori that the first value is real as $e^{2\pi i(1/2+\sqrt{-23}/2)} = -e^{-\pi\sqrt{23}}$ and the coefficients of the q-expansion are integers. Similarly, the two last values are complex conjugates of each other as $e^{2\pi i(\pm 1/4+\sqrt{-23}/4)} = \pm i e^{-\pi\sqrt{23}/2}$. In any case we have

$$\left(t - j\left(\frac{1+\sqrt{-23}}{2}\right)\right)\left(t - j\left(\frac{-1+\sqrt{-23}}{4}\right)\right)\left(t - j\left(\frac{1+\sqrt{-23}}{4}\right)\right)$$
$$\approx t^3 + 3491750.00000t^2 - 5151296875.00000t + 12771880859375.00000$$

which gives us the exact polynomial

$$t^3 + 3491750t^2 - 5151296875t + 12771880859375.$$

iii) One shows without difficulty that there is only one element of $\mathcal{C}_{163}$ which means that $j((1+\sqrt{-163})/2)$ is an integer. On the other hand,

$$j(q) = \frac{1}{q} + 744 + O(q)$$

and $e^{-\pi\sqrt{163}} \approx 3.81 \times 10^{-18}$ is very small which means that $e^{\pi\sqrt{163}}$ is close to an integer and indeed

$$e^{\pi\sqrt{163}} \approx 262537412640768743.99999999999250.$$

11.2 Elliptic curve interpretation

In this section we have obtained a number of interesting results on lattices and their j-invariants without involving the associated elliptic curves. We shall now see that the curves do indeed fit very naturally into our results.

Let us start with two lattices $\Gamma' \subseteq \Gamma$. If f is an elliptic function with the elements of Γ' as periods, then for every $\gamma \in \Gamma$, the function $f_\gamma(z) := f(z+\gamma)$ is also elliptic with the elements of Γ' and furthermore it only depends on the residue $\bar{\gamma}$ of γ in Γ/Γ'. We are therefore going to use also this residue as index. It is clear that $f_\alpha(z+\beta) = f_{\alpha+\beta}(z)$ and hence

$$T_{\Gamma/\Gamma'} f(z) := \sum_{\alpha\in\Gamma/\Gamma'} f_\alpha$$

has all the elements of Γ as periods. If we apply this first to $\wp'(z|\Gamma')$ we get

$$T_{\Gamma/\Gamma'}(\wp'(-|\Gamma'))(z) = \sum_{\alpha\in\Gamma/\Gamma'} -2\sum_{\gamma\in\Gamma'} \frac{1}{(z+\alpha-\gamma)^3} = -2\sum_{\gamma\in\Gamma}\frac{1}{(z-\gamma)^3} = \wp'(z|\Gamma).$$

(This can also be seen by just noting that $T_{\Gamma/\Gamma'}(\wp'(-|\Gamma'))$ has poles at the right places and with the right polar parts.) As clearly $T_{\Gamma/\Gamma'}$ commutes with d/dz we get as a consequence that $T_{\Gamma/\Gamma'}(\wp(-|\Gamma'))(z) = \wp(z|\Gamma) + \beta$, where β is a constant. This constant can be determined by looking at the constant terms for the Taylor series at $z = 0$ which gives $\beta = \sum'_{\alpha\in\Gamma/\Gamma'}\wp(\alpha|\Gamma')$. Note that we have $n\Gamma \subseteq \Gamma'$ for some positive integer n and then for every $0 \neq \alpha \in \Gamma/\Gamma'$ we have that $\psi_n(\wp(\alpha|\Gamma'), \wp'(\alpha|\Gamma')) = 0$ so that all the values in the sum $\sum'_{\alpha\in\Gamma/\Gamma'}\wp(\alpha|\Gamma')$ are algebraic over $\mathbb{Q}(g_2(\Gamma'), g_3(\Gamma'))$. We also have that $\wp(z+\alpha)$ and $\wp'(z+\alpha)$ may be expanded using the addition formulas. Together with the formula just established this gives formulas for $\wp(z|\Gamma)$ and $\wp'(z|\Gamma)$ as rational functions $P_{\Gamma/\Gamma'}$, resp. $Q_{\Gamma/\Gamma'}$, in $\wp(z|\Gamma')$ and $\wp'(z|\Gamma')$ whose coefficients are rational functions in $\wp(\alpha|\Gamma')$ and $\wp'(\alpha|\Gamma')$ for $0 \neq \alpha \in \Gamma/\Gamma'$. That means that $(x : y : 1) \mapsto (P(x, y) : Q(x, y) : 1)$ gives a map from $E'(\mathbb{C}) \to E(\mathbb{C})$, where E', resp. E, are the elliptic curves corresponding to Γ', resp. Γ, such that it corresponds to $\mathbb{C}/\Gamma' \to \mathbb{C}/\Gamma$ induced by the identity map on $\mathbb{C}$. The map $(x : y : 1) \mapsto (P(x, y) : Q(x, y) : 1)$ is a priori not defined at the points of Γ/Γ' but it can be extended using the methods we have already employed. There is a converse of this result. For its formulation we need to introduce some definitions. If $x \in \mathbb{C}/\Gamma$, a *coordinate neighbourhood* of it consists of the choice of a point $y \in \mathbb{C}$ which maps to x under the quotient map $\mathbb{C} \to \mathbb{C}/\Gamma$ and an open connected neighbourhood U of y which maps injectively to $\mathbb{C}/\Gamma$. It is easy to see that U is determined by y and its image in $\mathbb{C}/\Gamma$ and as the choice of y is largely irrelevant we shall identify a coordinate neighbourhood with this image. If Γ and Γ' are two lattices, then a map $f: \mathbb{C}/\Gamma' \to \mathbb{C}/\Gamma$ will be said to be *holomorphic* if for every $x \in \mathbb{C}/\Gamma'$ there are coordinate neighbourhoods U and V of x resp $f(x)$ such that $f(U) \subseteq V$ and f considered as a map $U \to V$ and U and V considered as open subsets of $\mathbb{C}$ are holomorphic.

Exercise 83. Let $f: \mathbb{C}/\Gamma' \to \mathbb{C}/\Gamma$ be a holomorphic map.

i) Show that there is a holomorphic map $\tilde{f}: \mathbb{C} \to \mathbb{C}$ such that for each $\gamma \in \Gamma'$ there is an $\ell_\gamma \in \Gamma$ such that $\tilde{f}(z+\gamma) = \tilde{f}(z) + \ell_\gamma$.

ii) Show that $\tilde{f}$ is of the form $\tilde{f}(z) = \lambda z + \alpha$.

iii) Show that if $f(0) = 0$, then we may choose $\tilde{f}$ such that $\tilde{f}(z) = \lambda z$ and we have $\lambda\Gamma' \subseteq \Gamma$.

Exercise 84. i) Let $f: \mathbb{C}/\Gamma' \to \mathbb{C}/\Gamma$ be a holomorphic map and let $(x_0 : y_0 : 1) \in E' := \mathbb{C}/\Gamma'$ with $y_0 \neq 0$. Then the map $x \mapsto f((x : \sqrt{4x^3 - g_2'x - g_3'} : 1))$, where the square root has been chosen so that it is continuous and equal to y_0 for $x = x_0$, for x close to x_0 can be written in the form $x \mapsto (f(x) : g(x) : h(x))$ where f, g, and h are holomorphic functions (defined close to x_0) not all vanishing for an x close to x_0.

ii) Show that if $(x_0 : 0 : 1) \in E'$, then there is a (unique) holomorphic function t defined near 0 and with $t(0) = x_0$ and such that $y^2 = t^3(y) - g_2't(y) - g_3'$. Show that the composite $x \mapsto f((t(y) : y : 1))$ is of the form $y \mapsto (f(y) : g(y) : h(y))$ where f, g, and h are holomorphic functions (defined close to 0) not all vanishing for any y close to 0.

iii) Formulate a result similar to the two previous ones for $(0:1:0) \in E'$.

iv) Show that a function $f\colon E' \to E$ that fulfils the conclusions of i)-iii) is holomorphic.

It seems reasonable that an algebraically defined map $E' \to E$ (a notion for which we have not given a precise definition) should be a holomorphic map $\mathbb{C}/\Gamma' \to \mathbb{C}/\Gamma$. Conversely it seems equally reasonable that the map above given by P and Q should be an algebraically defined map.

A problem with our candidate algebraically defined maps is that they are not defined at some (finite in number) points. Once the notions have been set up properly it turns out there is a unique extension to an everywhere defined algebraic map. We shall solve this problem by simply ignoring it and hence allow the map to be undefined at a finite number of points. As an algebraically defined map in the case of a curve over the complex numbers should give rise to a holomorphic map with at most poles at the missing points, the following exercises give some substance to our extension claim.

Exercise 85. Let $U \subseteq \mathbb{C}$ be an open subset. A map $F\colon U \to \mathbb{P}^2(\mathbb{C})$ is *holomorphic* if for every point $x_0 \in U$ there is an open subset $x_0 \in V \subseteq U$ and holomorphic functions $f, g, h\colon V \to \mathbb{C}$ such that for each $x \in V$ one of $f(x)$, $g(x)$, and $h(x)$ is non-zero and $F(x) = (f(x) : g(x) : h(x))$. A *meromorphic map* from U to $\mathbb{P}^2$ is a holomorphic function $F\colon U' \to \mathbb{P}^2$ such that

- $U' = U \setminus S$ with $S \subset U$ *discrete*, i.e., for each $x \in S$ there is an open $x \in V \subseteq U$ with $S \cap V = \{x\}$,
- for each $x \in S$ there is an open $x \in V \subseteq U$ and holomorphic functions $f, g, h\colon V \setminus S \to \mathbb{C}$,
- for each $x \in V$ one of $f(x)$, $g(x)$, and $h(x)$ is non-zero and $F(x) = (f(x) : g(x) : h(x))$, and
- f, g, and h are meromorphic on V.

Show that for every meromorphic map $F\colon U \to \mathbb{P}^2(\mathbb{C})$ there is a unique holomorphic map $G\colon U \to \mathbb{P}^2(\mathbb{C})$ which coincides with F whenever the latter is defined.

Exercise 86. A *meromorphic map* from $E = \mathbb{C}/\Gamma$, an elliptic curve over $\mathbb{C}$, to $\mathbb{P}^2(\mathbb{C})$ is a holomorphic map from $E \setminus S$ to $\mathbb{P}^2(\mathbb{C})$, where S is a finite set, such that for every $x \in S$ there is a coordinate neighbourhood U of x such that the induced map $U \setminus \{x\} \to \mathbb{P}^2(\mathbb{C})$ is meromorphic.

A meromorphic map between two elliptic curves E and E' is a function $E \setminus S \to E'$ such that the composite $E \setminus S \to E' \hookrightarrow \mathbb{P}^2(\mathbb{C})$ is meromorphic. Show that every such meromorphic map has an extension to a holomorphic map $E \to E'$.

Exercise 87. An *algebraic map* $\mathbb{P}^1 \to \mathbb{P}^2$ is given by a collection $f(s,t)$, $g(s,t)$, and $h(s,t)$ of homogeneous polynomials of the same degree without a common non-trivial zero. The map then is $(s:t) \mapsto (f(s,t):g(s,t):h(s,t))$. It gives rise to a point $(f(x,1):g(x,1):h(x,1)) \in \mathbb{P}^2(K(x))$, $K(x)$ being the field of rational functions, the field of fractions of $K[x]$. Show that every such point comes from a (unique) algebraic map $\mathbb{P}^1 \to \mathbb{P}^2$.

Remark 10. The fact, indicated by these exercises, that algebraic (or analytic) maps extend over poles is very special for maps from *curves*.

We use the indications above, as well as examples obtained previously, to *define* an algebraic map from an elliptic curve E to another elliptic curve E' to be a map $(x:y:1) \mapsto (f(x,y):g(x,y):1)$ given by *rational* functions f and g. Hence we should have $g^2 = f^3 + a'f + b'$ but this should not be an equality of rational functions but an equality modulo the relation $y^2 = x^3 + ax + b$. This means that we should consider f and g as elements of the fraction field, $K(E)$, of $K[x,y]/(y^2 - (x^3 + ax + b))$.

Example 17. Show that if x^3+ax+b is without multiple roots, then $y^2-(x^3+ax+b)$ is irreducible and hence $K[x,y]/(y^2 - (x^3 + ax + b))$ is a domain.

Hence we get a ring homomorphism $K[x',y']/(y'^2 - (x'^3 + ax' + b)) \to K(E)$.

Exercise 88. Show that if f and g are not constant, then this map is injective.

By the exercise we get a (necessarily injective) mapping of K-field extensions $K(E') \to K(E)$ and conversely such a map gives a map $E \to E'$. Hence, the non-constant algebraic maps $E \to E'$ correspond exactly to K-field extensions $K(E') \to K(E)$. An *algebraic homomorphism* is then such an algebraic map which is also a group homomorphism on points (over an algebraically closed over-field). (From the argument of Exercise 83 it follows that when $K = \mathbb{C}$, then every map that takes ∞ to ∞ is a homomorphism, this is actually true in general.) In the complex case the algebraic homomorphisms plus the map with constant value ∞ forms a ring which is equal to $\mathrm{End}_{\mathbb{C}}(\Gamma)$ when $E = \mathbb{C}/\Gamma$ (and for that reason we shall allow the map with constant value ∞ also as an algebraic homomorphism).

If we apply this to when we have an inclusion $\Gamma' \subseteq \Gamma$, then the formulas for $\wp(z|\Gamma)$ and $\wp'(z|\Gamma)$ in terms of $\wp(z|\Gamma')$ and $\wp'(z|\Gamma')$ give us an algebraically defined map $E \to E'$ where the elliptic curves $\mathbb{C}/\Gamma'$ and $\mathbb{C}/\Gamma$ are equivalent to E', resp. E. Note furthermore that Γ/Γ' is equal to the kernel of the group homomorphism $E' \to E$. Combining what we know so far we see that there is an algebraic group homomorphism $E' \to E$ whose kernel is cyclic of order n precisely when $\Phi_n(j(E'), j(E)) = 0$. As the kernel of $E' \to E$ is of exponent n, the map factors through $E \to E' \to E$ with the composite being multiplication by n (this corresponds to the inclusion $n\Gamma \subseteq \Gamma' \subseteq \Gamma$ of lattices). It's is easy to see that the kernel of $E \to E'$ is again cyclic of order n. It seems reasonable (and is true) that the map $E \to E'$ is again algebraic. If so we should also have $\Phi_n(j(E), j(E')) = 0$ which gives a perhaps clearer explanation for the first step in the proof of Proposition 11.2.

More interesting however is that these ideas can be used to give a more conceptual proof of the fact that the monic polynomial whose roots are the j-invariants of the elements of $\mathcal{C}_D$ has integer coefficients. Indeed, as those j-invariants are roots of some appropriate $-\Phi_n(t,t)$, they are algebraic integers and hence, by some elementary algebraic number theory, it is enough to show that the coefficients are rational numbers. Now, the following are basic facts of Galois theory (and not too difficult to establish): An *algebraic conjugate* of a complex number α is the image of α under an automorphism of $\mathbb{C}$. A monic polynomial without multiple roots has rational coefficients precisely when any algebraic conjugate of a root of it is again a root.

Hence to prove our statement it is enough to prove that if $\Gamma \in \mathcal{C}_D$ and if $\sigma\colon \mathbb{C} \to \mathbb{C}$ is an automorphism, then $\sigma(j(\Gamma))$ is the j-invariant of some $\Gamma' \in \mathcal{C}_D$. If we put $E = \mathbb{C}/\Gamma$ we have seen above that the ring (under point wise addition and composition) $\operatorname{End}(E)$ of algebraic endomorphisms of E is equal to $\operatorname{End}_{\mathbb{C}}(\Gamma) = R_D$. Now, E is defined by the equation $y^2 = 4x^3 - g_2x - g_3$ and we let E^σ be the curve defined by $y^2 = 4x^3 - \sigma(g_2)x - \sigma(g_3)$. We have a ring automorphism $\mathbb{C}[x,y] \to \mathbb{C}[x,y]$ taking $\sum_{ij} a_{ij}x^iy^j$ to $\sum_{ij}\sigma(a_{ij})x^iy^j$. It induces a ring isomorphism

$$\mathbb{C}[x,y]/(y^2 - (4x^3 - g_2x - g_3)) \to \mathbb{C}[x,y]/(y^2 - (4x^3 - \sigma(g_2)x - \sigma(g_3)))$$

which in turn induces an isomorphism between fraction fields $\mathbb{C}(E) \xrightarrow{\sim} \mathbb{C}(E^\sigma)$. (Note that this map is *not* $\mathbb{C}$-linear but is rather σ on $\mathbb{C}$.) An algebraic endomorphism $\lambda\colon E \to E$ different from 0 corresponds to a $\mathbb{C}$-linear ring homomorphism $\lambda\colon \mathbb{C}(E) \to \mathbb{C}(E)$. We now let $\lambda^\sigma := \sigma^{-1} \circ \lambda \circ \sigma\colon \mathbb{C}(E') \to \mathbb{C}(E')$. A moment's thought reveals that it is the identity on $\mathbb{C}$ and hence corresponds to an algebraic map $\lambda^\sigma\colon E' \to E'$. It is now not difficult to show that it is an algebraic homomorphism.

Exercise 89. i) Show that λ^σ is an algebraic homomorphism.

ii) Show that $\lambda \mapsto \lambda^\sigma$ is a ring homomorphism $\operatorname{End}(E) \to \operatorname{End}(E')$ and that it is also a bijection.

iii) Show that if R_D and $R_{D'}$ are isomorphic as rings, then $D = D'$.

It follows from this exercise that writing E' as $\mathbb{C}/\Gamma'$ we have that $\Gamma' \in \mathcal{C}_D$. Now, $j(E)$ and $j(E')$ are the same rational function in g_2, g_3 resp. $\sigma(g_2), \sigma(g_3)$ and hence $j(\Gamma') = j(E') = \sigma(j(E))$ and we have just seen that $\Gamma' \in \mathcal{C}_D$ which proves what we want.

12

Modular forms

The invariance property of a modular function f says that $f(M\tau) = f(\tau)$ for every $M \in \mathrm{SL}_2(\mathbb{Z})$. This is the formulation which is the easiest for calculation but there is a more conceptual formulation. We have seen that for any lattice $\Gamma \subset \mathbb{C}$ we can define $f(\Gamma)$ by choosing a τ such that Γ is equivalent to $\mathbb{Z} + \mathbb{Z}\tau$ and then put $f(\Gamma) := f(\tau)$. This gives a bijective correspondence between modular functions and functions f on lattices such that $f(\Gamma) = f(\lambda\Gamma)$ for $\lambda \in \mathbb{C}^\times$ and such that $\tau \mapsto f(\mathbb{Z} + \mathbb{Z}\tau)$ is a holomorphic function on the upper half plane and for which the expansion in $q := e^{2\pi i\tau}$ only has a finite number of terms of negative exponent. Clearly, the most important modular function is the j-function. As we have seen it is the quotient of two other functions E_2^3 and $(E_2^3 - E_3^2)/1728$. These in turn are *not* modular functions. They, or rather the functions G_2 and G_3, are by definition functions on lattices. However, they do not take the same value on Γ and $\lambda\Gamma$ but instead $G_k(\lambda\Gamma) = \lambda^{-2k}G_k(\Gamma)$. On the other hand their behaviour at infinity is nicer than that of j in that their q-expansions are holomorphic. This leads to the following definition.

Definition 12.1. A *modular form of weight k*, k an integer, is a function f on the set of lattices $\Gamma \subset \mathbb{C}$ such that

- $f(\lambda\Gamma) = \lambda^{-2k}f(\Gamma)$ for all Γ and $\lambda \in \mathbb{C}^\times$,
- the function $\tau \mapsto f(\mathbb{Z} + \mathbb{Z}\tau)$ is holomorphic in the upper half plane, and
- the function in q given by $q = e^{2\pi i\tau} \mapsto f(\mathbb{Z}+\mathbb{Z}\tau)$ has a holomorphic extension to $q = 0$.

If also $f(\mathbb{Z} + \mathbb{Z}\tau) \to 0$ when $\Im\tau \to \infty$, we say that f is a *cusp form*.

Note that the holomorphicity in q may be replaced by seemingly weaker conditions such as being bounded when $\Im\tau \to \infty$ (by the Riemann removability theorem).

Exercise 90. i) Show that if one replaces the first condition with the condition that $f(\lambda\Gamma) = \lambda^k f(\Gamma)$ for an odd k, then such a function is identically zero.

ii) Show that the function E_k is a modular form of weight k.

iii) Show that modular forms of a fixed weight form a $\mathbb{C}$-vector space. Show that every modular form of weight k is the unique sum of a multiple of E_k and a cusp form.

iv) Show that the product of a modular form of weight k and one of weight ℓ is a modular form of weight $k+\ell$ and that the set of finite sums of modular forms (possibly of different weights) forms a ring. Show that such finite sums can be written uniquely as a sum of forms of distinct weights.

It is clear that a modular form is determined by its restriction to the lattices $\mathbb{Z}+\mathbb{Z}\tau$ and the following exercise gives a characterisation of modular forms in terms of the function $\tau \mapsto f(\mathbb{Z}+\mathbb{Z}\tau)$.

Exercise 91. i) If F is a modular function of weight k, show that the function $f(\tau) := F(\mathbb{Z}+\mathbb{Z}\tau)$ is holomorphic on $\mathbb{H}$ fulfilling the condition

$$f\left(\frac{a\tau+b}{c\tau+d}\right) = (c\tau+d)^{2k} f(\tau), \quad \text{for all } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}),$$

which tends to a finite limit when $\Im\tau \to \infty$. Show conversely that a function fulfilling these properties comes from a unique modular form of weight k.

ii) Show that a function $f\colon \mathbb{H} \to \mathbb{C}$ which tends to a finite limit when $\Im\tau \to \infty$ and for which $f(\tau+1) = f(\tau)$ and $f(-1/\tau) = \tau^{2k} f(\tau)$ for all τ is a modular form of weight k.

There are many reasons why modular forms are important. One is that the quotient of a modular form and a non-zero modular form of the same weight is a modular function (as we have seen in the case of the j-function). There are also intrinsic reasons of which we shall see examples. Some further reasons are indicated in the following exercises.

Exercise 92. Let $f_0, f_2, \dots, f_n$ be modular forms of the same weight not all of which are identically zero. Let S be the set of points $\tau \in \mathbb{H}$ for which all the $f_i(\tau) := f_i(\mathbb{Z}+\mathbb{Z}\tau)$ are zero.

i) Show that the map $\mathbb{H} \setminus S \to \mathbb{P}^n(\mathbb{C})$ given by $\tau \mapsto \mathbb{P}^n(\mathbb{C})$ extends to a holomorphic map $f\colon \mathbb{H} \to \mathbb{P}^n(\mathbb{C})$, where "holomorphic" is defined analogously to Exercise 85.

ii) Show that f is invariant under $\mathrm{SL}_2(\mathbb{Z})$, i.e., $f(M\tau) = f(\tau)$ for all $M \in \mathrm{SL}_2(\mathbb{Z})$.

iii) (Difficult) From the previous part we get a map $g\colon \mathbb{C} \to \mathbb{P}^n(\mathbb{C})$ such that $g(j(\tau)) = f(\tau)$. Show that g is holomorphic.

We have seen that E_k is a modular form of weight k. By Exercise 90 their products are also modular forms and so are linear combinations of such products of the same weight. We have already seen, by using the differential equation of $\wp(z)$, that the E_k are polynomials in E_2 and E_3 so that we need only use these two forms. Our intent is to show that any modular form is a polynomial in E_2 and E_3. (This will actually give a new proof of the fact that the E_k are polynomials in E_2 and E_3.) We begin with the following exercise which generalises Proposition 6.2. Note first that, if f is a non-zero modular form and if $z \in \mathbb{H} \cup \{\infty\}$, we can define $v_z(f)$ in the same way as for modular functions.

Exercise 93. Show that if f is a non-zero modular form, then

$$\sum_z v_z(f) = \frac{k}{6}.$$

Example 18. Notice that for a non-zero modular form f we have that $v_z(f) \geq 0$ for all z as f does not have any poles.

i) If the weight k of f is (strictly) negative we have $0 \leq \sum_z v_z(f) = k/6 < 0$ so that any modular form of negative weight is zero.

ii) If the weight is zero we are dealing with a modular function without poles. By subtracting the constant value at some point we can assume that it has a zero. Thus $0 < \sum_z v_z(f) = 0/6 = 0$ so that every modular form of weight zero is constant.

iii) If the weight is 1 we have that $v_z(f) \geq 1/3$ and $\sum_z v_z(f) = 1/6$. This is not possible so that the form is zero.

iv) If the weight is 2, we may by subtracting of a multiple of E_2 assume that it is a cusp form, i.e., that $v_\infty(f) \geq 1$. As $\sum_z v_z(f) = 2/6$ this implies that every form of weight 2 is a multiple of E_2.

v) The same argument for weight 3, 4, and 5 shows that the form is a multiple of E_3, E_2^2, and E_2E_3 respectively.

vi) If the weight is 6, $\Delta := (E_2^3 - E_3^2)/1728$ is a cusp form. For any non-zero cusp form we have that $v_\infty(f) + \sum_{z \neq \infty} v_z(f) = 6/6 = 1$ and as $v_\infty(f) \geq 1$ we get that $v_\infty(f) = 1$ and that the form has no zeroes in $\mathbb{H}$. This means that the space of cusp forms is 1-dimensional, for if not we could for instance get a non-zero form with a double zero at ∞. In particular Δ has a simple zero at ∞ and no other zeroes.

We are now ready to show our promised result.

Proposition 12.2. *Every modular form of weight k is a homogeneous polynomial of weight k in E_2 and E_3 (where E_2, resp. E_3, are given weights* 2*, resp.* 3*).*

Proof. In the previous example we showed this for $k < 6$ (actually also for $k = 6$) and we continue by induction on k. It is easily seen that for every $k \geq 6$ there is a monomial m in E_2 and E_3 of weight k. It has value 1 at ∞. Now, given a modular form f of weight $k \geq 6$ we may choose c such that $f - cm$ is a cusp form. That means that $(f - cm)/\Delta$ is holomorphic at ∞ and as Δ has no zeroes in $\mathbb{H}$, $(f - cm)/\Delta$ is holomorphic everywhere. It is also easily seen that it fulfils the transformation properties for a modular form of weight $k - 6$. By induction $(f - cm)/\Delta$ is a polynomial in E_2 and E_3 and hence so is $f = ((f - cm)/\Delta)\Delta + cm$ as $\Delta = (E_2^3 - E_3^2)/1728$. □

Exercise 94. i) Show that the dimension of the space of modular forms of weight $k \geq 0$ is $[k/6] + 1$ if $k \not\equiv 1 \bmod 6$ and $[k/6]$ if $k \equiv 1 \bmod 6$.

ii) Show that E_2 and E_3 are algebraically independent over $\mathbb{C}$, i.e., they fulfil no non-trivial polynomial relation with complex coefficients.

12.1 Elliptic curve interpretation

One would (for many different and good reasons) like to make algebraic sense of modular forms. For that we need to say a little bit more about 1-forms. To begin with we have only defined 1-forms for rational functions. However a look at the definition (given on p. 3) reveals that one only needs a field containing the complex numbers and one may indeed replace the complex numbers by any other base field so that we are dealing with an arbitrary field extension $K \to L$. Hence we get the vector L-space $\Omega^1_{L/K}$ of 1-forms for L/K spanned by elements df, $f \in L$ and fulfilling the conditions from page 3 with the field of rational functions replaced by L and the field of complex numbers replaced by K.

Example 19. Consider an elliptic curve $E\colon y^2 = x^3 + ax + b$ and the corresponding field $K(E)$ of rational functions, i.e., the fraction field of $K[x, y]/(y^2-(x^3+ax+b))$. Every element of $K(E)$ can be (uniquely) written in the form $f(x) + yg(x)$, with f and g rational functions in x. We then have that $d(f(x) + yg(x)) = f'(x)\,dx + yg'(x)\,dx + g(x)\,dy$ which shows that $\Omega^1_{K(E)/K}$ is spanned by dx and dy. We also have $2y\,dy = (3x^2 + a)\,dx$ (characteristic different from 2 and 3) which means that dx and dy are $K(E)$-multiples of each other so that $\Omega^1_{K(E)/K} = K(E)dx$. To show that $dx \neq 0$ we define a $K(E)$-linear map $\Omega^1_{K(E)/K} \to K(E)$ taking dx to 1. We define it on dh, $h = f(x) + yg(x)$, as $f'(x) + yg'(x) + (3x^2 + a)/(2y)g(x)$. It then only remains to show that this map preserves the relations in $\Omega^1_{K(E)/K}$ which is easily done.

Exercise 95. Show that L-linear maps $T\colon \Omega^1_{L/K} \to V$, V an L-vector space, may be identified with K-linear maps $t\colon L \to V$ such that $t(fg) = ft(g) + gt(f)$ for all $f, g \in L$. The identification is supposed to fulfil $T(df) = t(f)$.

We have already encountered the 1-form dx/y. What distinguishes this from other forms is that it has no poles. This is elucidated in the following definition.

Definition 12.3. Let E be an elliptic curve.

i) An element $f \in K(E)$ is *regular* at a point $p \in E(K)$ if it satisfies the following:

- when $p = (x_0 : y_0 : 1)$, f has the form g/h, where $g, h \in K[x, y]/(p(x, y))$, p is the equation of E, and $h(x_0, y_0) \neq 0$,
- when $p = (0 : 1 : 0)$, f has the form g/h where g and h are polynomials $g(y/x^2, 1/x)$ and $f(y/x^2, 1/x)$ in y/x^2 and $1/x$ (of $K(E)$) and $h(0, 0) \neq 0$.

ii) A 1-form $\omega \in \Omega^1_{K(E)/L}$ is *regular* at a point $p \in E(K)$ if it is a sum of 1-forms of the type $f\,dg$ where f and g are regular at p.

iii) A 1-form is *regular* (when K is algebraically closed) if it is regular at all points.

Remark 11. The definition of regularity at the point at infinity is motivated by the fact that in homogeneous coordinates of weights $(1, 2, 1)$ a point "close" to it has the form $(1 : r : s)$ and r and s should thus be considered regular at ∞ but $(1 : y/x^2 : 1/x) = (x : y : 1)$ so that $r = y/x^2$ and $s = 1/x$.

Example 20. i) When the equation of E is $y^2 = x^3 + ax + b$, dx/y is regular. This is immediate at all points except those for which $y = 0$ and the point at infinity. We have however that $dx/y = 2\,dy/(3x^2 + a)$ so that at finite points the only points of contention are those for which $y = 0 = 3x^2 + a$, but there are no such points as $x^3 + ax + b$ has no multiple roots. At the point at infinity we have, with $r = y/x^2$ and $s = 1/x$, that $x = 1/s$ and $y = r/s^2$ so that

$$\frac{dx}{y} = \frac{d(1/s)}{r/s^2} = -\frac{ds}{r} = \frac{2dr}{1 + as^3 + s^4},$$

as we have $r^2 = 1 + as^3 + s^4$. As $s(\infty) = r(\infty) = 0$ we have presented dx/y in the form required for regularity.

ii) Let E be the complex curve of equation $y^2 = 4x^3 - g_2x - g_3$ and let Γ be a lattice with $g_2 = g_2(\Gamma)$ and $g_3 = g_3(\Gamma)$. If $\omega \in \Omega^1_{\mathbb{C}(E)/\mathbb{C}}$ we define an elliptic function ω/dz with Γ as periods as follows. For $\omega = dg$ we write g as a rational function $g(x, y)$ and put

$$\frac{\omega}{dz} := \frac{dg(\wp(z), \wp'(z))}{dz}. \tag{12.1}$$

By Exercise 96 this gives a $\mathbb{C}(E)$-linear map from $\Omega^1_{\mathbb{C}(E)/\mathbb{C}}$ to elliptic functions with Γ as periods (a field that is isomorphic to $\mathbb{C}(E)$). We shall now prove first that if $f \in \mathbb{C}(E)$ is regular, then $f(\wp(z), \wp'(z))$ is holomorphic and hence constant. To show this, consider a point $p = (\wp(z_0) : \wp'(z_0) : 1)$ (or $p = \infty$ and $z_0 = 0$). It is clear that $\wp(z)$ and $\wp'(z)$ (resp. $\wp'(z)/\wp^2(z)$ and $1/\wp(z)$) are holomorphic at $z = z_0$ and so then is any polynomial in them. Furthermore, the value of such a polynomial at z_0 is its value at p (resp. its constant term). From this it follows that if $f \in \mathbb{C}(E)$ is regular at p, then $f(\wp(z), \wp'(z))$ is holomorphic at z_0.

From this it immediately follows that if a form $\omega \in \Omega^1_{\mathbb{C}(E)/\mathbb{C}}$ is regular at p, then ω/dz is holomorphic at z_0. In particular, if ω is regular, then ω/dz is a holomorphic elliptic function and hence a constant. As we have seen, any ω in $\Omega^1_{\mathbb{C}(E)/E}$ has the form $f(x, y)\,dx$ and then $\omega/dz = f(\wp(z), \wp'(z))\wp'(z) \in \mathbb{C}$ and as $f \mapsto f(\wp(z), \wp'(z))$ gives an isomorphism of fields from $\mathbb{C}(E)$ to the field of elliptic functions (with Γ as periods) we get that λ/y and $\omega = \lambda dx/y$. Hence the regular forms are exactly the multiples of dx/y. This can be shown to be true for any field.

Exercise 96. Show that the formula (12.1) fulfils the conditions of Exercise 95.

We have seen (at least in the complex case) that any regular 1-form is a multiple of dx/y. Note that the exact 1-form dx/y depends on the precise equation and not just on the equivalence class of the curve. Indeed, if E is given by $y^2 = x^3 + ax + b$, then it is also given by $y'^2 = x'^3 + a\lambda^4x' + b\lambda^6$, where $x' = \lambda^2x$ and $y' = \lambda^3y$ and we have

$$\frac{dx}{y} = \lambda\frac{dx'}{y'}.$$

On the other hand, going back to the complex case, choosing a regular non-zero 1-form ω specifies a lattice Γ. One simply gets Γ as the integrals along closed curves (a.k.a. "periods") of the form ω. Replacing ω by a non-zero multiple $\lambda\omega$ then scales the lattice of periods by λ. We can thus identify modular forms of weight k with functions on (equivalence classes of) complex elliptic curves with a chosen non-zero regular 1-form such that $f((E, \lambda\omega)) = \lambda^{-2k} f((E, \omega))$ and such that, restricted to $(\mathbb{C}/\mathbb{Z}+\mathbb{Z}\tau, dx/y)$, it has the required analytic properties. It is not difficult to imagine (and it is in any case true) that these analytic properties can be given a purely algebraic characterisation.

Example 21. If we stay away from characteristic 2 and 3, any elliptic curve is equivalent to one with equation $y^2 = x^3 + ax + b$ and the only way to get an equivalent curve is to make the coordinate change $x' = \lambda^2 x$ and $y' = \lambda^3 y$ giving the equation $y'^2 = x'^3 + a\lambda^4 x' + b\lambda^6$. As we have seen if we do that we have the relation $dx/y = \lambda dx'/y'$. This means that for any elliptic curve E with non-zero regular 1-form ω there is a *unique* equation of the form $y^2 = x^3 + ax + b$ giving a curve equivalent to E and for which $\omega = dx/y$. Hence, a and b are well-defined functions on (E, ω) and when we change (E, ω) to $(E, \lambda\omega)$, we change a to $\lambda^{-4}a$ and b to $\lambda^{-6}b$. Hence a and b should be (and are) modular forms of weight 2 and 3 respectively.

12.2 Modular forms of higher level

We have seen examples of functions which almost but not quite are modular functions, specifically the j_α, $\alpha \in \mathbb{P}^1(\mathbb{Z}/p)$. There are similar examples for modular forms. Consider for instance the roots of $\wp'(z|\Gamma)$. We have seen that these are just the elements $1/2\Gamma/\Gamma \setminus \{0\}$. If we then consider their values for $\wp(z|\Gamma)$ we get $e_1(\omega_1, \omega_2) := \wp(\omega_1/2|\Gamma)$, $e_2(\omega_1, \omega_2) := \wp(\omega_2/2|\Gamma)$, and $e_3(\omega_1, \omega_2) := \wp((\omega_1+\omega_2)/2|\Gamma)$, where $\Gamma = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$. We have that $e_i(\lambda\omega_1, \lambda\omega_2) = \lambda^{-2} e_i(\omega_1, \omega_2)$ which makes them look like modular forms of weight 1. However, these are not modular forms[1] as they really depend on the choice of generators ω_1 and ω_2 and not just on the lattice Γ. More, precisely if we use $M \in \mathrm{SL}_2(\mathbb{Z})$ to change the basis ω_1 and ω_2, the functions e_i will be permuted among themselves (more precisely they are permuted in the same way that the elements of $\mathbb{P}^1(\mathbb{Z}/2)$ are). That means that by considering the coefficients of the polynomial $(t - e_1)(t - e_2)(t - e_3)$ we get modular forms of weight 1, 2, and 3 respectively.

Exercise 97. Show that $e_1 + e_2 + e_3 = 0$, $e_1e_2 + e_1e_3 + e_2e_3 = -g_2/4$, and $e_1e_2e_3 = g_3/4$.

There is however an alternative (and ultimately better) way of looking at these functions: We lower our ambitions and make the group smaller. Hence, for a case

[1] There are as we have seen no modular forms of weight 1.

such as e_1, we do not allow all base changes of $\mathbb{Z}\omega_1+\mathbb{Z}\omega_2$ but only those in the subgroup $\Gamma_0(2) := \left\{ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z}) \mid c \equiv 0 \bmod 2 \right\}$. If we on the other hand want all of the e_i to be modular we should choose the group $\Gamma(2) := \ker(\mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/2))$. Note that this group has index 6 in $\Gamma(2)$.

Exercise 98. Show that $\mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/2)$ is surjective.

There are some problems about what is meant by the holomorphicity conditions. To understand them we must first realise that if we are dealing with a modular form for $\Gamma(2)$, then the fundamental domain $\mathbb{D} := \{z \in \mathbb{H} \mid |z| \geq 1, -1/2 \leq \Re(z) \leq 1/2\}$ is not really relevant. The reason that it was relevant for $\mathrm{SL}_2(\mathbb{Z})$ was that we had

$$\mathbb{H} = \cup_{g \in \mathrm{SL}_2(\mathbb{Z})} g\mathbb{D}$$

with the union essentially disjoint,[2] i.e., with intersection only at the boundaries. Hence a modular form would be determined by its values on $\mathbb{D}$. These results are no longer true if we let g run over $\Gamma(2)$, only “a sixth” of $\mathbb{H}$ will be covered in this way. More precisely if we let $g_1, g_2, \dots, g_6$ be coset representatives for $\Gamma(2)$ in $\mathrm{SL}_2(\mathbb{Z})$, then

$$\mathbb{H} = \cup_{g \in \mathrm{SL}_2(\mathbb{Z})} g\mathbb{D} = \cup_{g \in \Gamma(2)} \cup_{1 \leq i \leq 6}\, g g_i \mathbb{D} = \cup_{g \in \Gamma(2)} g \left(\cup_{1 \leq i \leq 6} g_i \mathbb{D}\right) = \cup_{g \in \Gamma(2)} g\mathbb{D}',$$

where we have put $\mathbb{D}' := \cup_{1 \leq i \leq 6} g_i \mathbb{D}$. This means that $\mathbb{D}'$ should work as a fundamental domain for $\Gamma(2)$. If we are careful about choosing the g_i we can make $\mathbb{D}'$ as nice as $\mathbb{D}$: Define the following Möbius transformations induced from elements of $\mathrm{SL}_2(\mathbb{Z})$ by $g(\tau) := \tau + 1$ and $h(\tau) := -1/\tau$ and then put $g_1 = \mathrm{id}$, $g_2 = g$, $g_3 = h$, $g_4 = gh$, $g_5 = hg$, and $g_6 = ghg$. With that choice we get the picture of Figure 28. Three

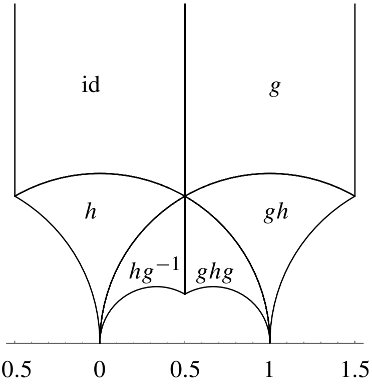

Figure 28. The fundamental domain for $\Gamma(2)$.

[2] For this we should really let g run over $\mathrm{SL}_2(\mathbb{Z})/\{\pm 1\}$.

things are clear from this picture. The first is that we do indeed get a nice fundamental domain for $\Gamma(2)$. The second is that when we try to investigate the holomorphicity of a form at ∞ by considering the q-expansion, we are faced with the problem that $\tau \mapsto \tau+1$ does not come from an element of $\Gamma(2)$. On the other hand $\tau \mapsto \tau+2$ *does* come from such an element and hence the form can be expanded in powers of $e^{\pi i\tau}$, a function of a type that we have already encountered when dealing with the j_α and then called $q^{1/2}$. Once this has been noted there are however no problems; holomorphicity at ∞ means simply that there are no negative powers in the q-expansion (now in half-integer powers of q).

The third thing to note also has to do with the behaviour at infinity. The reason that we have insisted on holomorphicity (for forms) and meromorphicity (for modular functions) is that we need it to have control over these functions. Without them we would not have that modular functions are rational functions in j or that modular forms are polynomials in E_2 and E_3. We see from Figure 28 that we should expect the need for control not just at infinity but also at 0 and 1, both of which are points at the boundary of the domain of definition of the form and also at the boundary of the fundamental domain. We shall deal with this after we have given a more formal definition of modular forms for groups such as $\Gamma(2)$.

We thus let $G \subseteq \mathrm{SL}_2(\mathbb{Z})$ be a subgroup of finite index. For technical reasons we need a subgroup $N \subseteq G$ which is normal and of finite index in $\mathrm{SL}_2(\mathbb{Z})$. Such a subgroup always exists; one may for instance let N be the kernel of the group homomorphism from G to the group of permutations on the right cosets G/H of G with respect to H. A *modular form for G of weight k* is a function f on pairs (ω_1, ω_2) of complex numbers linearly independent over the reals such that

- $f(M(\omega_1, \omega_2)) = f(\omega_1, \omega_2)$ for all $M \in G$,
- $f(\lambda\omega_1, \lambda\omega_2) = \lambda^{-2k} f(\omega_1, \omega_2)$,
- the function $\tau \mapsto f((1, \tau))$ from the $\mathbb{H}$ to $\mathbb{C}$ is holomorphic, and
- let $0 \neq n \in \mathbb{Z}$ such that $\left(\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right)^n = \left(\begin{smallmatrix}1&0\\n&1\end{smallmatrix}\right)$ belongs to the group N (which exists as N has finite index), then for every $H \in \mathrm{SL}_2(\mathbb{Z})$, the function $f_H(\tau) := f(H(1, \tau))$ is periodic of period n as
$$\begin{aligned} f_H(\tau+n) &= f\left(H\begin{pmatrix}1&0\\n&1\end{pmatrix}(1,\tau)\right) = f\left(H\begin{pmatrix}1&0\\n&1\end{pmatrix}H^{-1}H(1,\tau)\right) \\ &= f(H(1,\tau)) = f_H(\tau), \end{aligned}$$
where we use that N is normal in G, and we require that the expansion in powers of $q^{1/n}$ only contains positive powers.

If the q-expansion of each f_H is without constant term for all $H \in \mathrm{SL}_2(\mathbb{Z})$ we say that the form is a *cusp form*.

Remark 12. Note that we allow k to be "half integer" weights, i.e., $2k \in \mathbb{Z}$ and $k \notin \mathbb{Z}$ and though we have seen that all such forms for $\mathrm{SL}_2(\mathbb{Z})$ are zero this is not so for a general group.

Exercise 99. Show that the conditions are independent of the choice of N.

Exercise 100. Formulate the conditions on a modular form for G in terms of $F(\tau) := f(1, \tau)$ similar to the one for forms for $\mathrm{SL}_2(\mathbb{Z})$.

Exercise 101. i) Show that in order to check holomorphicity at the cusps it is enough to check $f_H(\tau)$ for a set of H's with the following property: For every $M \in \mathrm{SL}_2(\mathbb{Z})$ there is an H in the set and a $P \in G$ such that $M(\infty) = PH(\infty)$.

ii) Show that $\mathrm{SL}_2(\mathbb{Z})$ acts transitively on $\mathbb{P}^1(\mathbb{Q})$.

iii) Let S be the set of orbits of G on $\mathbb{P}^1(\mathbb{Q})$. Show that a set of H's fulfils the property of the first part precisely when for every $s \in S$ there is an H such that $H(\infty) \in s$.

Example 22. i) For $\Gamma_0(2)$ there are exactly 2 orbits on $\mathbb{P}^1(\mathbb{Q})$, represented by ∞ and 0. Hence we may let H be the identity matrix or a matrix giving the Möbius transformation $\tau \mapsto -1/\tau$.

ii) For $\Gamma(2)$ there are exactly three orbits represented by 0, 1, and ∞. As indicated by this example it is in general possible to choose a fundamental domain which meets the boundary of $\mathbb{H}$ (including the point at ∞) in precisely the different orbits.

We shall now consider the ring of modular forms for $\Gamma(2)$. As we have seen there are three cusps that may be chosen to be 0, 1, and ∞. These are identified with all the points of $\mathbb{P}^1(\mathbb{Z}/2)$ by reduction modulo 2 and the action of $\mathrm{SL}_2(\mathbb{Z})$ on the orbits of $\Gamma(2)$ on $\mathbb{P}^1(\mathbb{Q})$ are given exactly by the reduction $\mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/2)$ and the action of $\mathrm{SL}_2(\mathbb{Z}/2)$ on $\mathbb{P}^1(\mathbb{Z}/2)$. We have seen that e_i, $i = 1, 2, 3$, are candidates for modular forms for $\Gamma(2)$ (in fact individually they are candidates for $\Gamma_0(2)$ and its conjugates). We have seen that they fulfil the invariance properties and it is clear that they are holomorphic in $\mathbb{H}$ and it thus remains to check behaviour at cusps. Now, the e_i are permuted under the action of $\mathrm{SL}_2(\mathbb{Z})$ and hence it is enough to investigate all of them at a single cusp which we choose to be ∞. To show that they are holomorphic at ∞ it is, as we have noted,[3] enough to show that it has a limit when $\Im(\tau) \to \infty$,

[3] Technically, a holomorphic function in a punctured neighbourhood of a point can be extended to a holomorphic function at the point if it is bounded in that neighbourhood; this is the Riemann removability theorem.

that limit is then equal to the value at ∞. Now, we have the following formulas:

$$e_1(\tau) = \frac{1}{(1/2)^2} + {\sum_{m,n}}' \frac{1}{(1/2+m\tau+n)^2} - \frac{1}{(m\tau+n)^2},$$
$$e_2(\tau) = \frac{1}{(\tau/2)^2} + {\sum_{m,n}}' \frac{1}{(\tau/2+m\tau+n)^2} - \frac{1}{(m\tau+n)^2},$$
$$e_3(\tau) = \frac{1}{((\tau+1)/2)^2} + {\sum_{m,n}}' \frac{1}{((\tau+1)/2+m\tau+n)^2} - \frac{1}{(m\tau+n)^2}.$$

All the series converge uniformly so that we may take the limit term-wise which gives

$$\begin{aligned}e_1(\infty) &= \frac{1}{(1/2)^2} + {\sum_{n\in\mathbb{Z}}}' \frac{1}{(1/2+n)^2} - \frac{1}{n^2}\\ &= -\frac{\pi^2}{3} + 8\sum_{n\geq 0}\frac{1}{(2n+1)^2} = -\frac{\pi^2}{3} + \pi^2 = \frac{2\pi^2}{3},\\ e_2(\infty) &= -{\sum_n}' \frac{1}{n^2} = -\frac{\pi^2}{3},\\ e_3(\infty) &= -{\sum_n}' \frac{1}{n^2} = -\frac{\pi^2}{3}.\end{aligned}$$

(Note that these values are consistent with the fact that $e_1 + e_2 + e_3 = 0$.) We shall however need the q-expansions of these functions. To formulate the result in a convenient fashion we introduce some variants of the σ_k-functions:

$$\sigma_k^{ec}(n) := \sum_{\substack{d|n\\ d \text{ even}}} d^k, \qquad \sigma_k^{oc}(n) := \sum_{\substack{d|n\\ d \text{ odd}}} d^k,$$
$$\sigma_k^{e}(n) := \sum_{\substack{d|n\\ d \text{ even}}} \left(\frac{n}{d}\right)^k, \qquad \sigma_k^{o}(n) := \sum_{\substack{d|n\\ d \text{ odd}}} \left(\frac{n}{d}\right)^k.$$

We clearly have that

$$\sigma_k(n) = \sigma_k^{ec}(n) + \sigma_k^{oc}(n),$$
$$\sigma_k(n) = \sigma_k^{e}(n) + \sigma_k^{o}(n),$$

and if n is odd, $\sigma_k^e(n) = \sigma_k^{ec}(n) = 0$, while if n is even $\sigma_k^e(n) = \sigma_k(n/2)$ and $\sigma_k^{ec}(n) = 2^k\sigma_k(n/2)$. Hence in all cases $\sigma_k^{ec}(n) = 2^k\sigma_k^e(n)$ and that gives $\sigma_k^{oc}(n) = \sigma_k(n) - 2^k\sigma_k^o(n)$. Now to get the q-expansion we shall, just as for the G_k, use (30) (in our case only for $k = 2$):

$$\sum_{n\in\mathbb{Z}} \frac{1}{(\tau+n)^2} = -4\pi^2\sum_{k=1}^{\infty} kq^k.$$

This gives, using that $e^{2\pi i(1/2+m\tau)} = -q^m$,

$$\begin{aligned}
e_1(\tau) &= \frac{2\pi^2}{3} + 2\sum_{m\geq 1}\sum_{n\in\mathbb{Z}} \frac{1}{(1/2+m\tau+n)^2} - \frac{1}{(m\tau+n)^2}\\
&= \frac{2\pi^2}{3} - 8\pi^2 \sum_{m\geq 1}\sum_{k=1}^{\infty} k((-q^m)^k - (q^m)^k)\\
&= \frac{2\pi^2}{3} - 8\pi^2 \sum_{m\geq 1}\sum_{k=1}^{\infty} k((-1)^k - 1)q^{mk}\\
&= \frac{\pi^2}{3}\Big(2 + 48\sum_{n=1}^{\infty} \sigma_1^{oc}(n)q^n\Big).
\end{aligned}$$

In the same manner we have, using that $e^{2\pi i(\tau/2+m\tau)} = q^{m+1/2}$,

$$\begin{aligned}
e_2(\tau) &= -\frac{\pi^2}{3} + \sum_{n\in\mathbb{Z}} \frac{1}{(\tau/2+n)^2} + \sum_{m\neq 0}\sum_{n\in\mathbb{Z}} \frac{1}{(\tau/2+m\tau+n)^2} - \frac{1}{(m\tau+n)^2}\\
&= -\frac{\pi^2}{3} - 4\pi^2\Big(2\sum_{m\geq 0}\sum_{k=1}^{\infty} k(q^{1/2})^{(2m+1)k} - \sum_{m\geq 1}\sum_{k=1}^{\infty} k(q^{1/2})^{2mk}\Big)\\
&= -\frac{\pi^2}{3} - 4\pi^2\sum_{n\geq 1}\Big(2\sigma_1^o(n) - \sigma_1^e(n)\Big)q^{n/2}\\
&= \frac{\pi^2}{3}\Big(-1 - 12\sum_{n\geq 1}\big(3\sigma_1^o(n) - \sigma_1(n)\big)q^{n/2}\Big).
\end{aligned}$$

Similarly for e_3 (where $q^{1/2}$ must be replaced by $-q^{1/2}$)

$$\begin{aligned}
e_2(\tau) &= -\frac{\pi^2}{3} + \sum_{n\in\mathbb{Z}} \frac{1}{((\tau+1)/2+n)^2}\\
&\quad + \sum_{m\neq 0}\sum_{n\in\mathbb{Z}} \frac{1}{((\tau+1)/2+m\tau+n)^2} - \frac{1}{(m\tau+n)^2}\\
&= -\frac{\pi^2}{3} - 4\pi^2\Big(2\sum_{m\geq 0}\sum_{k=1}^{\infty} k(-q^{1/2})^{(2m+1)k} - \sum_{m\geq 1}\sum_{k=1}^{\infty} k(q^{1/2})^{2mk}\Big)\\
&= -\frac{\pi^2}{3} - 4\pi^2\sum_{n\geq 1}\Big(2(-1)^n\sigma_1^o(n) - \sigma_1^e(n)\Big)q^{n/2}\\
&= \frac{\pi^2}{3}\Big(-1 - 12\sum_{n\geq 1}\big((1+2(-1)^n)\sigma_1^o(n) - \sigma_1(n)\big)q^{n/2}\Big).
\end{aligned}$$

This means that each e_i is non-zero at all cusps with the same value at two of them and another at the third (which is ∞ for $i = 1$, 0 for $i = 2$, and 1 for $i = 3$). Hence if we put $f_k := e_i - e_j$, where $\{1, 2, 3\} = \{i, j, k\}$ and $i < j$, then they are modular forms that are zero at one of the cusps and non-zero at the two others and the cusp where it is zero is different for different k. Furthermore, at the cusp at which one of them is zero it has a simple zero (where simple means that its q-expansion starts with $q^{1/2}$) as can be seen by the formulas above as the two non-zero $q^{1/2}$-terms have opposite signs. This is what is needed for the following result.

Proposition 12.4. *Any modular form for* $\Gamma(2)$ *is a polynomial in the* e_i (*and as* $e_1 + e_2 + e_3 = 0$ *thus in any two of them*).

Proof. A modular form of weight 0 is constant just as for the $\mathrm{SL}_2(\mathbb{Z})$-case. Forms of half-integer weights $k/2$ are zero, as one would have $f(\tau) = f\big(\big(\begin{smallmatrix}-1 & 0\\ 0 & -1\end{smallmatrix}\big)\tau\big) = (-1)^k f(\tau) = -f(\tau)$ just as in the $\mathrm{SL}_2(\mathbb{Z})$-case. Let us for the moment leave aside the case of weight < 0. We know (cf. Theorem 3.8 i)) that $e_i(\tau) \neq e_j(\tau)$ for any τ and $i \neq j$. Hence f_1 is non-zero on all of $\mathbb{H}$, at the cusps 0 and 1 and has a simple zero at ∞. We now prove our result by induction on the weight k. The case $k = 0$ has just been proved. Consider now a form f of (integer) weight $k > 0$. By subtracting of a multiple of e_1^k we may assume that it is zero at ∞ and then f/f_1 is holomorphic everywhere and thus a modular form of weight $k - 1$. By induction it is a polynomial in the e_i's. If f has negative weight $-k$, then ff_1^k has weight 0 and is hence a constant. As f_1 is zero at ∞ and f finite there, we get that the constant is 0 which implies that $f = 0$. □

Exercise 102. Show that if f is a modular form for $\Gamma(2)$ of weight k, then $\sum_z v_z(f) = k$ where z runs over the fundamental domain (without duplications on the boundary) and the three cusps, and $v_z(f)$ is the multiplicity of the zero of f at z (without weights this time for points in the fundamental domain and with the multiplicity of the zero at the cusps being the power of the first power of $q^{1/2}$ that occurs in the q-expansion of the cusp).

Exercise 103. Show that the modular forms for $\Gamma_0(2)$ are polynomials in e_1 and e_2e_3.

12.2.1 Sums of four squares. We shall now use the results obtained to give a formula for the number of ways to write a positive integer as a sum of four squares. We shall do this by showing that the "generating series"

$$\sum_{x,y,z,w\in\mathbb{Z}} q^{(x^2+y^2+z^2+w^2)/2} = \sum_{n\geq 0} a_n q^{n/2},$$

where a_n is the number of ways of writing n as a sum of squares, is the q-expansion at ∞ of a modular form of weight 1 for $\Gamma(2)$. If we know this, then it is a linear combination of the e_1 and e_2 and, looking at a few coefficients in the q-expansion, the coefficients of such a linear combination may be determined. However, we shall also see that it has a further transformation property which will determine it up to a multiple, that multiple can then be determined by looking at just the constant term in the q-expansion.

To show this we start by studying the *θ-function*

$$\theta(\tau) := \sum_{n\in\mathbb{Z}} q^{n^2/2} = 1 + 2\sum_{k\geq 1} q^{k/2},$$

the point being that

$$\theta^4(\tau) = \sum_{x,y,z,w\in\mathbb{Z}} q^{(x^2+y^2+z^2+w^2)/2}$$

and it turns out to be easier to study θ. What we specifically are going to show is that $\theta(-1/\tau) = (-i\tau)^{1/2}\theta(\tau)$, where the square root has been chosen such that it is continuous in $\mathbb{H}$ and takes the value 1 when $\tau = i$. There are several ways of proving this formula but we shall choose one which is related to elliptic functions.

For that we (re)pose the seemingly unrelated question of producing doubly periodic functions of periods 1 and $\tau \in \mathbb{H}$. To make it periodic of period 1 is easy, we simply make it a function of $u := e^{2\pi i z}$. We then assume that the function may be expanded as a convergent series $f(z) = \sum_{n\in\mathbb{Z}} a_n u^n$. Of course we know that it is not possible to get a non-constant example of a doubly periodic function in this way as it would be holomorphic for all z. We shall settle for a little less and consider

$$\theta(z,\tau) := \sum_{n\in\mathbb{Z}} u^n q^{n^2/2}$$

so that $\theta(\tau) = \theta(0,\tau)$. This function has the property that

$$\begin{aligned}
&\theta(z+1,\tau) = \theta(z,\tau),\\
&\theta(z+\tau,\tau) = \sum_{n\in\mathbb{Z}}(uq)^n q^{n^2/2} = (uq^{1/2})^{-1}\sum_{n\in\mathbb{Z}}(u)^{n+1}q^{(n+1)^2/2} = (uq^{1/2})^{-1}\theta(z,\tau).
\end{aligned}$$

The key to our proof of the functional equation for $\theta(\tau)$ is the following result.

Lemma 12.5. *$\theta(z,\tau)$ is the unique function which is holomorphic on all of $\mathbb{C}$ and fulfils the relations*

$$\begin{aligned}
\theta(z+1,\tau) &= \theta(z,\tau),\\
\theta(z+\tau,\tau) &= (uq^{1/2})^{-1}\theta(z,\tau),\\
\int_0^1 \theta(t,\tau)\,dt &= 1.
\end{aligned}$$

Proof. We have checked the first two conditions for θ. As for the third we have

$$\int_0^1 u^k(t)\,dt = \int_0^1 e^{2\pi i k t}\,dt = \begin{cases} 1 & \text{if } k = 0,\\ 0 & \text{if } k \neq 0,\end{cases}$$

and the sum defining θ converges uniformly so that one may exchange sum and integral.

Now let $f(z)$ fulfil the conditions. As $f(z+1) = f(z)$ we may expand f as a Laurent series in $u = e^{2\pi iz}$:

$$f(z) = \sum_{n\in\mathbb{Z}} a_n u^n$$

and the condition $f(z+\tau) = (uq^{1/2})^{-1} f(z)$ then translates into

$$\sum_{n\in\mathbb{Z}} a_n u^n = ((uq^{1/2})^{-1} \sum_{n\in\mathbb{Z}} a_n q^n u^n$$

or $a_n = q^{n+1/2} a_{n+1}$, which shows that the function f is determined by the value of a_0. The last condition then fixes that value. □

We now try to use $\theta(z,\tau)$ to construct a function that behaves like $\theta(z, -1/\tau)$. Taking the lead from what happens for doubly periodic functions we consider $f(z) := \theta(\tau z, \tau)$. It fulfils $f(z - 1/\tau) = \theta(\tau z - 1, \tau) = f(z)$ which is not what we want. We therefore make a slight modification and put instead $f(z) := e^{\pi i\tau z^2}\theta(\tau z, \tau)$. We then have

$$\begin{aligned} f(z+1) &= e^{\pi i\tau(z+1)^2}\theta(\tau z+\tau,\tau) = e^{\pi i\tau(z+1)^2}e^{-2\pi i\tau z-\pi i\tau}\theta(\tau z,\tau) = f(z),\\ f(z-1/\tau) &= e^{\pi i\tau(z-1/\tau)^2}\theta(\tau z-1,\tau) = e^{-2\pi iz+\pi i/\tau}e^{\pi i\tau z^2}\theta(\tau z,\tau)\\ &= e^{-2\pi iz+\pi i/\tau}f(z), \end{aligned}$$

and this means that f fulfils the conditions of Lemma 12.5 for $\theta(z, -1/\tau)$ except for the last normalisation condition. Hence we get that

$$e^{\pi i\tau z^2}\theta(\tau z,\tau) = \varphi(\tau)\theta(z,\tau)$$

for some function $\varphi(\tau)$. More precisely we have (where the interchange between integral and sum is trivially justified)

$$\begin{aligned} \varphi(\tau) &= \int_0^1 e^{\pi i\tau s^2}\theta(\tau s,\tau)\,ds = \sum_{n\in\mathbb{Z}}\int_0^1 e^{\pi i\tau s^2+2\pi ins\tau+\pi in^2\tau}\,ds\\ &= \sum_{n\in\mathbb{Z}}\int_0^1 e^{\pi i\tau(s^2+2ns+n^2)}\,ds = \sum_{n\in\mathbb{Z}}\int_0^1 e^{\pi i\tau(s+n)^2}\,ds\\ &= \sum_{n\in\mathbb{Z}}\int_n^{n+1} e^{\pi i\tau s^2}\,ds = \int_{-\infty}^{\infty} e^{\pi i\tau s^2}\,ds. \end{aligned}$$

What remains to be shown is that $\int_{-\infty}^{\infty} e^{\pi i\tau s^2}\,ds = (-i\tau)^{1/2}$. The proof of this is in two steps. First we note that both sides are analytic functions in the upper half plane and hence it is, by analytic continuation, enough to prove the equality for $\tau = it$,

$t > 0$. By definition the right-hand side is equal to $t^{1/2}$ (i.e., the positive square root). For the left-hand side we have with $r = t^{1/2}s$,

$$\int_{-\infty}^{\infty} e^{-\pi t s^2}\, ds = t^{1/2}\int_{-\infty}^{\infty} e^{-\pi r^2}\, dr = Ct^{1/2},$$

where $C := \int_{-\infty}^{\infty} e^{-\pi r^2}\, dr$ (which is well known to be 1 but we shall give a new proof of it). Now putting $z = 0$ in our equality we get $\theta(\tau) = C(-i\tau)^{1/2}\theta(-1/\tau)$ and putting $\tau = i$ we get $\theta(i) = C\theta(i)$, but $\theta(i) \neq 0$ as it is a sum of positive terms so we get $C = 1$.

Remark 13. It could be seen to be somewhat disappointing that $\theta(z, \tau)$ is unique up to a constant factor, since the quotient of two functions fulfilling the first two properties of Lemma 12.5 would be a doubly periodic function. One may however obtain non-trivial examples by replacing the second condition by $f(z + \tau) = (u^n q^{1/2})^{-1} f(z)$.

Raising the functional equation to the fourth power we get

$$\begin{aligned}\theta^4(\tau + 2) &= \theta^4(\tau),\\ \theta^4(-1/\tau) &= -\tau^2\theta^4(\tau).\end{aligned}$$

Note that the second functional equation is not of the type required for a modular form of weight 1 as that would be $f(-1/\tau) = \tau^2 f(\tau)$. We shall ignore this point for the moment and note that in any case the functional equations for the elements $\left(\begin{smallmatrix}1 & 2\\ 0 & 1\end{smallmatrix}\right)$ and $\left(\begin{smallmatrix}0 & -1\\ 1 & 0\end{smallmatrix}\right)$ of $\mathrm{SL}_2(\mathbb{Z})$ gives by iteration some functional equation for all elements of the subgroup generated by them. That subgroup is easily determined.

Exercise 104. Show that the subgroup of $\mathrm{SL}_2(\mathbb{Z})$ generated by $\left(\begin{smallmatrix}1 & 2\\ 0 & 1\end{smallmatrix}\right)$ and $\left(\begin{smallmatrix}0 & -1\\ 1 & 0\end{smallmatrix}\right)$ is equal to the subgroup $\Gamma_{1,2}$ of elements that are congruent modulo 2 to $\left(\begin{smallmatrix}0 & -1\\ 1 & 0\end{smallmatrix}\right)$ or the identity matrix.

Hence we know for which elements we get a functional equation and we shall now determine what these functional equations are. For that it is more convenient to deal with the homogeneous version, that is we consider $\Theta(\tau) := \theta^4(\tau)$ instead as a function of the two generators ω_1 and ω_2 of the lattice. In that formulation the functional equations proven say that $(\omega_1, \omega_2) \mapsto (\omega_1, \omega_1 + 2\omega_2)$ fixes Θ whereas $(\omega_1, \omega_2) \mapsto (-\omega_2, \omega_1)$ takes Θ to $-\Theta$. Hence an arbitrary element of $\Gamma_{1,2}$ takes Θ to $\pm\Theta$ and the mapping $M \mapsto M(\Theta)/\Theta \in \{\pm 1\}$ is a group homomorphism. This group homomorphism is characterised by the fact that it takes $(\omega_1, \omega_2) \mapsto (\omega_1, \omega_1 + 2\omega_2)$ and $(\omega_1, \omega_2) \mapsto (-\omega_2, \omega_1)$ to -1. Now, it is easily shown that $\left(\begin{smallmatrix}a & b\\ c & d\end{smallmatrix}\right) \mapsto (-1)^c$ is a group homomorphism with the same property and hence we have shown that

$$\Theta\left(\frac{a\tau + b}{c\tau + d}\right) = (-1)^c (c\tau + d)^2\Theta(\tau), \quad \text{for all } \begin{pmatrix}a & b\\ c & d\end{pmatrix} \in \Gamma_{1,2}.$$

In particular Θ is a modular form of weight 1 for $\Gamma(2)$ and is hence a linear combination of e_1 and e_2. However, we have that $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ permutes e_1 and e_2 and as this matrix takes Θ to $-\Theta$ we have that Θ is a multiple of $e_1 - e_2$. Now, for the q-expansion at infinity we have

$$e_1 - e_2 = \frac{\pi^2}{3}\big(3 + O\big(q^{1/2}\big)\big)$$

and as $\Theta(\tau) = 1 + O(q^{1/2})$ we get $\Theta = 1/\pi^2(e_1 - e_2)$. Now substituting the q-expansion of e_1 and e_2 we get

$$\frac{1}{\pi^2}(e_1 - e_2) = 1 + \sum_{\substack{n\geq 1 \\ n \text{ odd}}} 8\sigma_1(n)q^{n/2} + \sum_{\substack{n\geq 1 \\ n \text{ even}}} \big(16\sigma_1^{oc}(n/2) + 4(3\sigma_1^{o}(n) - \sigma_1(n))\big)q^{n/2}.$$

If we generally write $\theta^k(\tau) = \sum_{n=0}^{\infty} r_k(n)q^{n/2}$, then it is clear that $r_k(n)$ is equal to the cardinality of the set $\{(n_i) \in \mathbb{Z}^k \mid \sum_i n_i^2 = n\}$. In particular we have that $r_k(0) = 1$ and for the case $k = 4$ we conclude that

$$\theta^4 = 1 + \sum_{\substack{n\geq 1 \\ n \text{ odd}}} 8\sigma_1(n)q^{n/2} + \sum_{\substack{n\geq 1 \\ n \text{ even}}} \big(16\sigma_1^{oc}(n/2) + 4(3\sigma_1^{o}(n) - \sigma_1(n))\big)q^{n/2}.$$

We also get that

$$r_4(n) = \begin{cases} 8\sigma_1(n), & \text{if } n \text{ is even,} \\ 16\sigma_1^{oc}(n/2) + 4(3\sigma_1^{o}(n) - \sigma_1(n)), & \text{if } n \text{ is odd.} \end{cases}$$

Hints to exercises

Hint 9. Add the path $E(a,s)$ to the beginning of D and $E(a,1-s)$ to the end to get end point homotopic paths.

Hint 11. ii): Use the pigeon hole principle.

iii): Show that the set of points for which the function is constant in a neighbourhood of it is open and closed.

iv): Write a meromorphic function as $(z-a)^n h(z)$ in a disc around a where h is holomorphic with $h(a) \neq 0$.

Hint 12. i): Use the addition theorem.

Hint 14. Compare with the integral

$$\int_1^\infty \int_1^\infty \frac{dxdy}{x^2+y^2}.$$

Hint 19. Either show that the difference is holomorphic with constant term 0 at the origin or use (3.1).[4]

Hint 21. Use that the fundamental domain is closed and bounded.

Hint 22. Translate slightly the fundamental domain.

Hint 24. One way is to formulate the conclusion so that it is "closed under limits" and then write all the special cases as limits of non-special cases.

Hint 26. Consider $s \mapsto f(C_s)/\exp(\oint_{C_s} df/f)$, where $C_s = C_{|[a,s]}$ and take its derivative.

Hint 28. i): There is a local maximum and then f is constant by the maximum principle for harmonic functions.

ii): Use that a holomorphic function that only takes on imaginary values in the neighbourhood of a point is constant.

Hint 29. The derivative of the map is the 1-form dx/y which one shows is everywhere non-zero. Because of compactness the map is a covering map. Use covering space theory to show that the map then corresponds to a subgroup of Γ and by unfolding definitions that all periods land in this subgroup.

Hint 40. One of several ways is to use that $\mathrm{GL}_2(\mathbb{R})$ maps the real axis plus ∞ onto itself and thus either takes the upper half plane into itself or into the lower half plane. Then use the fact that $\mathrm{GL}_2^+(\mathbb{R})$, the matrices of positive determinant, is connected.

[4] See http://www.math.su.se/~teke/undervisning/Elliptisk.nb for the latter calculation.

Hint 41. iii,iv): Write out the condition that the lattice is stable under multiplication with α.

Hint 42. ii): Compute the discriminant of p.

Hint 44. The polynomial is of degree 6 and has λ as a root.

Hint 45. Show that a rational function in g_2 and g_3 invariant under $x \mapsto rx$ is a rational function in g_2^3/g_2^3.

Hint 48. Use the theorem that is called "Liouville's theorem".

Hint 49. For $z = i$ use that $f(-1/z) = f(z)$ and consider the Taylor expansion of this at $z = i$ together with the fact that the value of the derivative of $z \mapsto -1/z$ at i is -1. For $z = e^{\pi i/3}$ use instead $z \mapsto (z-1)/z$.

Hint 50. Make a small indentation of the path at a zero or pole making sure that they "match up" at congruent points.

Hint 54. i): Use that multiplication by ζ permutes the non-zero elements of $\Gamma_{\zeta,1}$.
ii): Use that multiplication by i permutes the non-zero elements of $\Gamma_{i,1}$.

Hint 55. Imitate the proof of Proposition 3.10.

Hint 56. Show that the natural ring homomorphism $R \to R[x]/(tx-1)$ is injective.

Hint 58. i: Show that the group homomorphism $x \mapsto x^{(q-1)/2}$ from $\mathbb{F}_q^\times$ to $\{\pm 1\}$ is surjective.
ii: When $q - 1 \not| n$ multiply the sum with some $x^{(q-1)/2} \neq 0, 1$.

Hint 59. Use that $\mathbb{Z}[\zeta]$ is a unique factorisation domain.

Hint 60. iv: Multiply out the product and make a change of summation variables.

Hint 62. First transform the equations, for $x \neq 0$, to $Y^2 + Y = x + ax^{-1}$ resp. $Y^2 + Y = x + m + ax^{-1}$. Then show that $y^2 + y = t$ has a solution in $\mathbb{F}_q$ precisely when $t + t^2 + \cdots + t^{2^{n-1}} = 0$, where $q = 2^n$.

Hint 63. Look at the value at z of the first (and second derivatives).

Hint 65. Write an element in the kernel as the residue of a polynomial $a(x) + b(x)y$.

Hint 66. ii: Use the Chinese remainder theorem.

Hint 73. Show that $\mathrm{SL}_2(\mathbb{Z}/p)$ is generated by elementary matrices.

Hint 80. Use Exercise 77.

Hint 93. Analyse the proof of Proposition 6.2; it has to be modified at exactly one point and that modification gives rise to the factor $k/6$.

Hint 103. Identify modular forms for $\Gamma_0(2)$ with modular forms for $\Gamma(2)$ that are invariant under the action of $\Gamma_0(2)/\Gamma(2)$.

Hint 104. Use the two elements to make row operations on a given matrix of $\Gamma_{1,2}$.

Solutions to exercises

Solution 27. i):

$$\wp(u)+\wp(v)+\wp(w)=\frac{1}{4}\left(\frac{\wp'(u)-\wp'(v)}{\wp(u)-\wp(v)}\right)^2.$$

Solution 36. The points are $\{(0:1:0), (1:2:1), (4:0:1), (1:3:1)\}$ and the group is cyclic of order 4.

Solution 41. v):

$$\tau=\frac{1+\sqrt{-23}}{2}, \frac{1+\sqrt{-23}}{4}, \frac{-1+\sqrt{-23}}{4}.$$

Solution 42. i): $p(t)=t^3+3491750t^2-5151296875t+12771880859375$.

Solution 103. Seen as functions on pairs (ω_1,ω_2), modular forms for $\Gamma_0(2)$ are modular forms for $\Gamma(2)$, i.e., functions invariant under $f \mapsto f \circ H$, $H \in \Gamma(2)$, that are also invariant under $\Gamma_0(2)$. As a modular form for $\Gamma(2)$ is already invariant under $\Gamma(2)$, the group $\Gamma_0(2)$ acts on them through the quotient $\Gamma_0(2)/\Gamma(2) \to \mathrm{SL}_2(\mathbb{Z}/2)$ whose only non-trivial element is $\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$. It fixes e_1 and permutes e_2 and e_3. As we know that the modular forms are polynomials in e_2 and e_2 and as it is easily seen that e_2 and e_3 fulfil no non-trivial polynomial identities, what this amounts to is that the modular forms for $\Gamma_0(2)$ are the elements of $\mathbb{C}[e_2,e_3]$ that are invariant under $e_2 \leftrightarrow e_3$. It is clear (and a special case of the fundamental theorem of symmetric functions) that those are polynomials in e_2+e_3 and e_2e_3.

Some further reading

The literature on elliptic curves is vast and includes by now a fair number of introductory books. We shall restrict ourselves to indicate some further directions of study.

Today arithmetic aspects of the theory of elliptic curves is a large and active area of research. A nice introduction to it is provided by [4]. A somewhat surprising development, with relations to the arithmetic theory, in recent years is the emergence of elliptic curves over finite fields as candidates for use in public key cryptography. [1] contains a very application oriented introduction to the use of elliptic curves in cryptography. Modular forms play an important rôle in the further development of elliptic curves as well as in generalisations. [2] and [3] contain very pleasant introductions to some aspects of these developments.

[1] I. F. Blake, G. Seroussi, and N. P. Smart, *Elliptic curves in cryptography*. London Mathematical Society Lecture Note Series 265, Cambridge University Press, Cambridge, 2000.

[2] David Mumford, *Tata lectures on theta.* I. With the assistance of C. Musili, M. Nori, E. Previato and M. Stillman, Progress in Mathematics 28, Birkhäuser, Boston, MA, 1983.

[3] David Mumford, *Tata lectures on theta.* II. *Jacobian theta functions and differential equations*. With the collaboration of C. Musili, M. Nori, E. Previato, M. Stillman and H. Umemura, Progress in Mathematics 43, Birkhäuser, Boston, MA, 1984.

[4] Joseph H. Silverman, *The arithmetic of elliptic curves*. Corrected reprint of the 1986 original, Graduate Texts in Mathematics 106, Springer-Verlag, New York, 1992.

Index