ELLIPTIC CURVES AND MODULAR FORMS

EHUD DE SHALIT

1. Complex Elliptic functions

See scanned notes on my home page.

2. The Abel-Jacobi theorem

See scanned notes on my home page.

3. Elliptic Theta functions and periods

3.1. The Weierstrass σ function.

The function

$$\zeta(z)=\frac{1}{z}+\sum_{\omega\in\Lambda}{}^{'}\left(\frac{1}{z-\omega}+\frac{1}{\omega}+\frac{z}{\omega^2}\right) \tag{3.1}$$

(check that the series converges) is the unique *odd* primitive of $-\wp(z)$ (the minus sign is there only for historical reasons; every other primitive will differ by a constant and will not be odd). It is not Λ-periodic anymore, but for every $\omega\in\Lambda$

$$-\int_{z_0}^{z_0+\omega}\wp(z)dz=\zeta(z_0+\omega)-\zeta(z_0)=\eta(\omega) \tag{3.2}$$

is independent of z_0, because of the periodicity of $\wp$.

The function $\zeta(z)$ has simple poles at Λ with residue 1. It follows that for a fixed z_0, the integral $\int_{z_0}^{z}\zeta(t)dt$ (along any path avoiding Λ) is well-defined modulo $2\pi i$, so

$$\sigma(z)=C\exp\left(\int_{z_0}^{z}\zeta(t)dt\right) \tag{3.3}$$

is well-defined. A change in the choice of z_0 or C results in rescaling σ by a multiplicative constant. Observe that

$$\zeta=\frac{\sigma'}{\sigma}. \tag{3.4}$$

In other words, ζ is the logarithmic derivative of σ, so the divisor of σ can be read from the residues of ζ at its poles: σ is everywhere analytic, and has simple zeros at the points of the lattice Λ. We normalize C so that $\sigma'(0)=1$.

Exercise 3.1. *Check that*

$$\sigma(z)=z\prod_{\omega\in\Lambda}{}^{'}\left(1-\frac{z}{\omega}\right)\exp\left(\frac{z}{\omega}+\frac{z^2}{2\omega^2}\right). \tag{3.5}$$

and that $\sigma(z)$ is odd.

By the definition of σ,

$$\begin{aligned}\frac{\sigma(z+\omega)}{\sigma(z)} &= \exp\left(\int_z^{z+\omega}\zeta(t)dt\right)\\ &= \exp\left(\eta(\omega)z+\xi(\omega)\right) \qquad (3.6)\end{aligned}$$

for some constants $\xi(\omega)$ (to see that the integral inside the exponent is of the prescribed shape, differentiate it).

Lemma 3.1. *We have*

$$\frac{\sigma(z+\omega)}{\sigma(z)} = \pm\exp\left(\eta(\omega)(z+\omega/2)\right) \qquad (3.7)$$

with the sign being $+$ *if* $\omega/2\in\Lambda$, *and* -1 *otherwise.*

Proof. We may write

$$\frac{\sigma(z+\omega)}{\sigma(z)} = C(\omega)\exp\left(\eta(\omega)(z+\omega/2)\right) \qquad (3.8)$$

for some constant $C(\omega)=\exp\left(\xi(\omega)-\eta(\omega)\omega/2\right).$ If $\omega\notin 2\Lambda$, then we may substitute $z=-\omega/2$ and get from the fact that σ is odd that $C(\omega)=-1$.

Next, calculating $\sigma(z+\omega_1+\omega_2)/\sigma(z)$ in two ways, first in one stroke, and then in two steps, we get the relation

$$C(\omega_1+\omega_2)=C(\omega_1)C(\omega_2)\exp\left(\frac{\eta(\omega_2)\omega_1-\eta(\omega_1)\omega_2}{2}\right) \qquad (3.9)$$

for any two $\omega_i\in\Lambda$. Substituting $\omega_1=\omega_2$ we get $C(2\omega)=C(\omega)^2$, which concludes the determination of the sign in the lemma (why?). □

3.2. **The Abel-Jacobi theorem revisited.** Using the Weierstrass σ function we may give an alternative proof that a divisor $D=\sum n_p[p]$ for which $\deg(D)=0$ and $s(D)=0$ is principal. Let $\tilde{p}$ be a representative for p in $\mathbb{C}$. Rewriting D with repetitions, so that all the n_p are ±1, and selecting the representatives appropriately, we may assume that in fact $\sum n_p\tilde{p}=0$ in $\mathbb{C}$, not only in $\mathbb{C}/\Lambda$. Consider then the function

$$f(z)=\prod_p\sigma(z-\tilde{p})^{n_p} \qquad (3.10)$$

whose divisor (modulo Λ) is D. As an easy consequence of the two assumptions on D we get that $f(z+\omega)=f(z)$ for every $\omega\in\Lambda$, hence f is elliptic, and D is principal.

Exercise 3.2. *Prove that there does not exist an elliptic function with one simple pole and no other poles.*

3.3. **Homology, cohomology and periods.** In this section we assume familiarity with basic notions from the theory of complex manifolds. In fact, we only need them in (complex) dimension 1. See for example the book by Farkas and Kra, or any other book on Riemann surfaces. Let ω_1 and ω_2 be an oriented basis for Λ. The straight paths from z_0 to $z_0+\omega_i$, for a fixed z_0, project in $\mathbb{C}/\Lambda$ to *closed* paths γ_i. The homology classes $[\gamma_i]$ form a basis for the homology $H_1(\mathbb{C}/\Lambda,\mathbb{Z})$, which is free of rank 2 over $\mathbb{Z}$.

A meromorphic differential form α on $\mathbb{C}/\Lambda$ is said to be *of the first kind* if it is everywhere holomorphic, and *of the second kind* if all its residues vanish. Exact forms (df for f in $\mathcal{F}(\Lambda)$) are of the second kind, and so are of course the forms of the first kind, but there are no exact forms of the first kind (except 0). The de Rham cohomology of $\mathbb{C}/\Lambda$ is by definition the quotient space

$$H^1_{dR}(\mathbb{C}/\Lambda) = \frac{\{\text{differential forms of the second kind}\}}{\{\text{exact forms}\}}. \tag{3.11}$$

The differential form $\omega = dz = d\wp/\wp'$ is of the first kind, and $\eta = \wp dz = \wp d\wp/\wp'$ is of the second kind. Their classes in H^1_{dR} are linearly independent: if $a\omega + b\eta = df$ then since ω has no poles and η has a second order pole at the origin but no others, f must have only one simple pole. However, it is a consequence of the Abel-Jacobi theorem that there are no such $f's$ (see the exercise above).

Let α be a differential form of the second kind. We denote the class of α in H^1_{dR} by $[\alpha]$. If γ and γ' are homotopic closed paths on $\mathbb{C}/\Lambda$,

$$\int_\gamma \alpha = \int_{\gamma'} \alpha. \tag{3.12}$$

In fact, as long as the homotopy does not pass through a pole of α, this is obvious. When the homotopy passes through such a pole p, the integral picks up $\pm 2\pi i \mathrm{Res}_p \alpha$, but since by definition all the residues of α vanish, it is unchanged. It follows that integration of α defines a linear functional

$$\int_\cdot \alpha : H_1(\mathbb{C}/\Lambda, \mathbb{Z}) \to \mathbb{C}.$$

The integral $\int_\gamma \alpha$ is called the *period* of α along the closed path γ (or along the homology class $[\gamma]$). If α is exact, all its periods vanish. Conversely, if the periods of α along γ_1 and γ_2 vanish, $f(z) = \int_{z_0}^z \alpha$ depends only on $p_\Lambda(z) \in \mathbb{C}/\Lambda$, is therefore in $\mathcal{F}(\Lambda)$, and $\alpha = df$ is exact. This means that we have *embedded* $H^1_{dR}(\mathbb{C}/\Lambda)$ in $Hom(H_1(\mathbb{C}/\Lambda, \mathbb{Z}), \mathbb{C})$, which is a two dimensional space. But we have seen that $[\omega]$ and $[\eta]$ are linearly independent, so the de Rham cohomology is at least two dimensional. Taken together we have the following.

Theorem 3.2. *The first de Rham cohomology $H^1_{dR}(\mathbb{C}/\Lambda)$ is two dimensional, and $[\omega]$ and $[\eta]$ form a basis for it. Integration identifies the de Rham cohomology as $Hom(H_1(\mathbb{C}/\Lambda, \mathbb{Z}), \mathbb{C})$.*

The periods of ω (resp. η) along γ_1 and γ_2 are just ω_1 and ω_2 (resp. $\eta_1 = \eta(\omega_1)$ and $\eta_2 = \eta(\omega_2)$). Linear independence of $[\omega]$ and $[\eta]$ mean that $\eta_2\omega_1 - \eta_1\omega_2 \neq 0$. We can say more.

Proposition 3.3. *We have*

$$\eta_2\omega_1 - \eta_1\omega_2 = 2\pi i. \tag{3.13}$$

Proof. Consider a point z_0 such that the (positively oriented) parallelogram whose vertices are z_0, $z_0 + \omega_2$, $z_0 + \omega_1 + \omega_2$ and $z_0 + \omega_1$ does not pass through any lattice points, and contains 0 in its interior. The quantity $\eta_2\omega_1 - \eta_1\omega_2$ is just $\int_{\partial\Pi} z\wp(z)dz$ (recall the minus sign in the equation $\eta_i = -\int_{\gamma_i} \wp(z)dz$). Since the only pole of $z\wp(z)$ in Π is at the origin, and its residue there is 1, the integral comes out to be $2\pi i$ by the residuum principle. □

Exercise 3.3. *What is the relation between this formula and the formula for $C(\omega_1+\omega_2)/C(\omega_1)C(\omega_2)$ which was obtained in the calculation of the factor of automorphy of the σ function?*

3.4. **Elliptic integrals.** Elliptic functions derive their name from the formula for the arc length of the ellipse. This is a little confusing. Elliptic curves (like $\mathbb{C}/\Lambda$), the geometric spaces on which elliptic functions live, are curves of genus 1, while ellipses (like all other conic sections) are the real points of curves of genus 0.

Consider the standard ellipse

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1, \tag{3.14}$$

where we assume $b \le a$. The arc length from $(0,b)$ to a point (x,y) in the first quadrant is given (check it!) by the formula

$$s(x)=a\int_0^{x/a}\frac{1-e^2u^2}{\sqrt{(1-u^2)(1-e^2u^2)}}du \tag{3.15}$$

where the *eccentricity*

$$e=\sqrt{1-\frac{b^2}{a^2}}. \tag{3.16}$$

Observe that for the unit circle ($a=b=1$) we get $s(x)=\arcsin(x)$. The function $s(x)$ is an example of an *elliptic integral.* In general, an elliptic integral is an indefinite function of the form

$$\int R(u,\sqrt{h(u)})du \tag{3.17}$$

where R is any rational function and h a cubic or a polynomial of degree 4. Integrals of this kind were studied intensively in the 18th century. As we know already from the case of the arcsin function, such functions are best studied as functions of a complex variable, and they are multiple-valued in general. At the beginning of the 19th cnetury, Abel and Jacobi discovered that their *inverse functions* are single valued, admit meromorphic continuation to the whole complex plane, and are doubly-periodic. These functions became known as elliptic functions. For the classical theory of elliptic integrals see the book by Whittaker and Watson.

4. Elliptic curves as plane projective curves

4.1. **The elliptic curve.** Fix a lattice Λ. The map

$$\xi:\mathbb{C}/\Lambda\to\mathbb{P}^2(\mathbb{C}) \tag{4.1}$$

where $\xi(z)=(\wp(z):\wp'(z):1)$ if $z\notin\Lambda$ and $\xi(0)=(0:1:0)=O$ is well-defined. Its image is contained in the *complex projective curve*

$$E(\mathbb{C})=\left\{(x:y:z)|\ y^2z=4x^3-g_2xz^2-g_3z^3\right\} \tag{4.2}$$

where $g_2=g_2(\Lambda)$ and $g_3=g_3(\Lambda)$. Notice that the equation defining $E(\mathbb{C})$ is homogenous, so $E(\mathbb{C})$ is well-defined, that in the affine piece

$$\mathbb{A}^2=\{(x:y:z)|\,z\neq 0\}\subset\mathbb{P}^2 \tag{4.3}$$

we may put $z=1$ and then the homogenous equation becomes just the inhomogenous Weierstrass equation in the two variables x and y, and that the only point of $E(\mathbb{C})$ where $z=0$ (the only point “at infinity”) is O.

The map ξ is continuous, since when z tends to 0, $\wp'$ has a pole of order 3 while $\wp$ has only a pole of order 2, so $\xi(z)$ approaches O.

Since the cubic $h(x) = 4x^3 - g_2x - g_3$ is separable, the affine curve given by the equation

$$F(x, y) = y^2 - 4x^3 + g_2x + g_3 = 0 \tag{4.4}$$

is non-singular: if $F(x_0, y_0) = 0$, then either $\partial F/\partial x$ or $\partial F/\partial y \neq 0$ at (x_0, y_0). In fact if x_0 is not a root of h, $y_0 \neq 0$ so $\partial F/\partial y \neq 0$ and by the implicit function theorem x is a local coordinate on E near x_0. If x_0 is a root of h then $\partial F/\partial x = -h'(x) \neq 0$ there by the separability of h.

Near O, change coordinates to $u = x/y$ and $v = z/y$. The curve E is then given in the affine piece $y \neq 0$ by the equation

$$v = 4u^3 - g_2uv^2 - g_3v^3 \tag{4.5}$$

and we see that the point O is also non-singular. Thus $E(\mathbb{C})$ is everywhere non-singular: it is a compact Riemann surface. So is $\mathbb{C}/\Lambda$, and the map ξ is analytic, so it is *open.* (Exercise: check that it is analytic near O by writing the formulas for u and v). Since $\mathbb{C}/\Lambda$ is compact, the image of ξ must be both closed and open, so it is the whole of $E(\mathbb{C})$.

Theorem 4.1. *The map ξ gives an isomorphism of Riemann surfaces between $\mathbb{C}/\Lambda$ and $E(\mathbb{C})$.*

Proof. We have seen that it is onto. One checks directly that it is a local isomorphism (i.e. that the derivative, when expressed in terms of some local parameter, is non-zero). Finally, let us see that it is one-to one. We have already seen that O is only obtained once, as $\xi(0)$. Suppose $\xi(z_1) = \xi(z_2)$ and $z_1, z_2 \in \mathbb{C}\backslash\Lambda$. Since $\wp(z_1) = \wp(z_2)$ we get that $z_1 = \pm z_2 mod\ \Lambda$. Since $\wp'(z_1) = \wp'(z_2) = -\wp'(-z_2)$ we get that if $z_1 = -z_2 mod \Lambda$ then z_1 is a zero of $\wp'$. But we have seen that in such a case z_1 is a half-period, so again $z_1 = z_2 mod \Lambda$. □

4.2. **The addition law on the cubic.** We now have two representations of our Riemann surface, as $\mathbb{C}/\Lambda$, and as the complex points of a smooth projective curve given by the homogeneous Weierstrass equation, and the map ξ is an explicit isomorphism between them.

While the representation as a smooth projective curve is useful from an algebraic point of view, it is not clear a priori how the *group structure* is reflected on $E(\mathbb{C})$. The only obvious thing is that the neutral element is the point $O = \xi(0)$.

Theorem 4.2. *Endow $E(\mathbb{C})$ with the group structure coming from $\mathbb{C}/\Lambda$ through the isomorphism ξ. Then three point P, Q and R on $E(\mathbb{C})$ are colinear if and only if*

$$P + Q + R = 0. \tag{4.6}$$

Proof. Suppose that none of the points is O, and that they are colinear. Let $l(x, y, z)$ be a linear form defining the line on which they lie. The function

$$f(z) = \frac{l}{z} \circ \xi \tag{4.7}$$

is clearly meromorphic and Λ-periodic, so it is an elliptic function. If $P = \xi(z_1)$, $Q = \xi(z_2)$ and $R = \xi(z_3)$ then f, which is a linear combination of $1, \wp$ and $\wp'$, must have a divisor

$$div(f) = [z_1] + [z_2] + [z_3] - 3[0]. \tag{4.8}$$

The Abel-Jacobi theorem then says that $z_1 + z_2 + z_3 = 0 mod \Lambda$, so $P + Q + R = 0$. Notice that if two of the points, say P and Q, coincide, then l should be a tangent to E there, and if all three coincide, then l should be a tangent which agrees with the curve to order 3 (the point is then called a *flex*).

Conversely, suppose P, Q and R are as above and their sum is 0. Then the Abel-Jacobi theorem produces for us a function $f \in \mathcal{F}(\Lambda)$ whose divisor is as above. Exercise: prove that

$$f = a\wp + b\wp' + c \tag{4.9}$$

(hint: kill the pole at 0 by subtracting a suitable linear combination of $\wp$ and $\wp'$, and use the fact that there is no elliptic function with only one simple pole). Let $l = ax + by + cz$. Then the line defined by $l = 0$ passes through P, Q and R.

The case where one of the points is O is similar and is left as an exercise. □

Corollary 4.3. *On $E(\mathbb{C})$ we have*

$$-(x_0 : y_0 : z_0) = (x_0 : -y_0 : z_0). \tag{4.10}$$

Proof. Take $R = O$, $P = (x_0 : y_0 : z_0)$ and $Q = (x_0 : -y_0 : z_0)$ and observe that they all lie on the line $z_0 x - x_0 z = 0$. □

The theorem is often called the *chord-tangent construction.* It means that geometrically, to add two points, one draws the line through them (the tangent to the curve if they coincide), looks for the third point of intersection with the cubic, and then reflects it in the x axis.

Exercise 4.1. *Derive the following explicit formulas: If*

$$P = (x_1 : y_1 : 1) \text{ and } Q = (x_2 : y_2 : 1) \tag{4.11}$$

and $P \neq \pm Q$, let

$$\lambda = \frac{y_2 - y_1}{x_2 - x_1}, \quad \mu = \frac{y_1 x_2 - y_2 x_1}{x_2 - x_1}. \tag{4.12}$$

Then $P + Q = (x_3 : y_3 : 1)$ where

$$x_3 = -x_1 - x_2 + \frac{\lambda^2}{4}, \quad y_3 = -\lambda x_3 - \mu. \tag{4.13}$$

If $P = -Q$ then $P + Q = O$ and if $P = Q \neq -Q$ then let

$$\lambda = \frac{12x_1^2 - g_2}{2y_1}, \quad \mu = y_1 - \lambda x_1 \tag{4.14}$$

and

$$x_3 = -2x_1 + \frac{\lambda^2}{4}, \quad y_3 = -\lambda x_3 - \mu. \tag{4.15}$$

Exercise 4.2. *Let $P = (2, 1) \in E$, where E is given by $y^2 = 4x^3 - 31$. Show that $2P = (140, -3313)$, and $3P = (\frac{41{,}401}{19{,}044}, \frac{4{,}175{,}605}{1{,}314{,}036})$. This should give you an idea about how complicated the points nP get when n grows.*

4.3. **Field of definition and rationality questions.** Let F be a subfield of $\mathbb{C}$. The elliptic curve E given by the equation

$$y^2z = 4x^3 - g_2xz^2 - g_3z^3 \tag{4.16}$$

is said to be *defined over F* if g_2 and g_3 belong to F. The chord-tangent construction, and more precisely the explicit formulas defined above immediately yield the following.

Proposition 4.4. *Let E be defined over F. Then $E(F)$, the set of points on E with $x, y, z \in F$, is a subgroup of $E(\mathbb{C})$.*

In particular, $E(\mathbb{Q})$ is a subgroup of $E(\mathbb{C})$ if the elliptic curve is defined over $\mathbb{Q}$. For example, for any $D \in \mathbb{Q}$, $D \neq 0$, there is a group structure on the rational solutions of the diophantine equation

$$y^2 = x^3 - D \tag{4.17}$$

(together with the point at infinity O taken as the origin).

This is how elliptic curves are connected with number theory.

There are several immediate questions now.

1) Which equations $y^2 = h(x)$ define complex elliptic curves (i.e. curves whose projective completions are smooth projective curves isomorphic, as Riemann surfaces, to $\mathbb{C}/\Lambda$ for some lattice Λ)? We shall see that h may be *any separable cubic.*

2) What is so special about the equation $y^2 = h(x)$? Is there a way to define elliptic curves intrinsically, without any reference to an embeding in $\mathbb{P}^2$ or to specific equations, and without any reference to the *uniformization* by the complex plane via the map ξ? Is it possible to give a definition that will work over any field F, even of positive characteristic?

The answer to the second question is of course positive, but to give the proper formulation we shall have to develop some terminology from algebraic geometry.

3) What can be said about the group $E(F)$ if F is a reasonably small field, e.g. a number field (a finite extension of $\mathbb{Q}$)? Here one has the celebrated Mordell-Weil theorem, which asserts that $E(F)$ is a finitely generated abelian group.

In the next section we shall answer question 1, and at the same time discuss *moduli* (algebraic families) of elliptic curves.

5. Moduli

5.1. **Homomorphisms and isomorphisms.** Suppose $\varphi : \mathbb{C}/\Lambda_1 \to \mathbb{C}/\Lambda_2$ is an analytic map of Riemann surfaces. Following it by a translation we may assume that it preserves the origin. By the homotopy lifting theorem we can lift φ to a map $\tilde{\varphi} : \mathbb{C} \to \mathbb{C}$, satisfying $\tilde{\varphi}(0) = 0$, which is entire because φ is analytic. Since for every $\omega \in \Lambda_1$, $\tilde{\varphi}(z + \omega) - \tilde{\varphi}(z) \in \Lambda_2$, differentiating we see that $\tilde{\varphi}'$ is Λ_1-periodic, hence constant. It follows that $\tilde{\varphi}(z) = \alpha z$ where $\alpha\Lambda_1 \subset \Lambda_2$. We summarize.

Proposition 5.1. *An analytic map of $\mathbb{C}/\Lambda_1$ to $\mathbb{C}/\Lambda_2$ carrying 0 to 0 is necessarily a group homomorphism, and is given by $zmod\Lambda_1 \mapsto \alpha zmod\Lambda_2$ for a constant α such that $\alpha\Lambda_1 \subset \Lambda_2$.*

Corollary 5.2. *The* group $Hom(\mathbb{C}/\Lambda_1, \mathbb{C}/\Lambda_2)$ *of elliptic curve homomorphisms (preserving both the analytic structure and the group structure, or what is the same, the origin) is isomorphic to*

$$\{\alpha \in \mathbb{C} |\ \alpha\Lambda_1 \subset \Lambda_2\} \tag{5.1}$$

as an abelian group. The isomorphism takes φ to $\varphi'(0)$.

Corollary 5.3. *The two elliptic curves $\mathbb{C}/\Lambda_i$ are isomorphic if and only if the lattices Λ_i are homothetic.*

Corollary 5.4. *The ring End$(\mathbb{C}/\Lambda)$ is isomorphic to the subring of $\mathbb{C}$ given by*

$$\{\alpha \in \mathbb{C} |\ \alpha\Lambda \subset \Lambda\}\,. \tag{5.2}$$

This ring is either $\mathbb{Z}$ or a subring of a quadratic imaginary field. The group Aut$(\mathbb{C}/\Lambda)$ is cyclic of order 2,4 or 6.

Proof. We may assume that $\Lambda = \mathbb{Z}\tau + \mathbb{Z}$ for $\tau \in \mathfrak{H}$. If the endomorphism ring is not $\mathbb{Z}$ then there exists an α such that

$$\alpha \cdot 1 = a\tau + b, \ \ \alpha \cdot \tau = c\tau + d \tag{5.3}$$

where a, b, c and d are integers. This means that τ solves a quadratic polynomial over $\mathbb{Q}$, hence $\mathbb{Q}(\tau)$ is quadratic imaginary (note that $\tau \notin \mathbb{R}$). Now $\alpha \mapsto \alpha \cdot 1 \in \mathbb{Z}[\tau]$ embeds the endomorphism ring as a subring of $\mathbb{Q}(\tau)$ (check that addition and composition go to the ring operations). The structure of the group of automorphisms comes form the structure of the group of units in rings of the form $\mathbb{Z}[\tau]$ for τ quadratic imaginary. This is elementary number theory, the exceptions being $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$ where $\omega = \exp(2\pi i/3)$. □

Definition 5.1. *If End$(\mathbb{C}/\Lambda)$ is not $\mathbb{Z}$, we say that $\mathbb{C}/\Lambda$ (or the lattice) admits* complex multiplication *(CM).*

Exercise 5.1. *Show that conversely, if τ is quadratic imaginary, $\mathbb{C}/\Lambda$ admits CM.*

5.2. **Moduli of lattices.** The classification of complex elliptic curves is therefore the classification of lattices up to homothety. Every lattice is represented by a lattice of the form $\mathbb{Z}\tau + \mathbb{Z}$ and if there is an α such that $\alpha(\mathbb{Z}\tau + \mathbb{Z}) = \mathbb{Z}\tau' + \mathbb{Z}$ then eliminating α from the equations we get

$$\tau' = \frac{a\tau + b}{c\tau + d} \tag{5.4}$$

for some matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z}). \tag{5.5}$$

The converse is also true. Now the group $SL_2(\mathbb{Z})$ *acts* on the upper half plane by the above formula (check!), so we conclude that the space of lattices is identified with the space of orbits of $SL_2(\mathbb{Z})$ in this action, denoted by

$$Y = SL_2(\mathbb{Z})\backslash\mathfrak{h}. \tag{5.6}$$

5.3. Y **as a Riemann surface and the fundamental domain.** Let Ω be the set of all complex numbers $\tau = x + iy$ where $|\tau| \geq 1$, $-1/2 \leq |x| < 1/2$ and if $|\tau| = 1$ then $-1/2 \leq x \leq 0$. Let

$$R = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \text{ and } T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}. \tag{5.7}$$

Note that $R(\tau) = -1/\tau$ is the inversion in the unit circle and $T(\tau) = \tau + 1$ is translation by one unit.

Theorem 5.5. *(1) R and T generate $SL_2(\mathbb{Z})$.*

(2) Ω is a fundamental domain for the action of $SL_2(\mathbb{Z})$ on $\mathfrak{H}$ (every orbit of $SL_2(\mathbb{Z})$ has a unique representative in Ω).

Proof. The two assertions are proved together. They sometime go under the name *reduction theory* because they are equivalent to Gauss' way of reducing a binary quadratic form by means of a change of variables.

Let Γ' be the subgroup of $\Gamma = SL_2(\mathbb{Z})$ generated by R and T.

(a) If $\tau \in \Omega$ then for $c, d \in \mathbb{Z}$, not both 0, we have $|c\tau + d| \geq 1$, with strict inequality unless $c = 0$, $d = \pm 1$,or $d = 0$, $c = \pm 1$ and $|\tau| = 1$, or $\tau = \omega = \exp(2\pi i/3)$ and $c = d = \pm 1$ (prove these claims by direct examination). From

$$Im(\frac{a\tau + b}{c\tau + d}) = \frac{Im(\tau)}{|c\tau + d|^2} \tag{5.8}$$

we get that if $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$ then $Im(g(\tau)) < Im(\tau)$ unless $c = 0$, or $d = 0$, or $\tau = \omega$ and $c = d = \pm 1$. It follows that no two points τ and τ' of Ω are in the same Γ-orbit.

(b) On the other hand let $\tau \in \mathfrak{H}$. Since for any M there are finitely many pairs (c, d) with $|c\tau + d| < M$, we can pick $\gamma \in \Gamma'$ with $Im(\gamma\tau)$ maximal. We can then follow γ by some power of T without changing $Im(\gamma\tau)$ and assume that $-1/2 \leq Re(\gamma\tau) < 1/2$. We claim that $|\gamma\tau| \geq 1$. Indeed,

$$R(\gamma\tau) = \frac{\overline{-\gamma\tau}}{|\gamma\tau|^2} \tag{5.9}$$

so if $|\gamma\tau| < 1$, $Im(R\gamma\tau) > Im(\gamma\tau)$, contradicting the choice of γ. Moreover, if $|\gamma\tau| = 1$ and $Re(\gamma\tau) > 0$, then $|R\gamma\tau| = 1$, $Im(R\gamma\tau) = Im(\gamma\tau)$, and $Re(R\gamma\tau) < 0$. This shows that $\gamma\tau \in \Omega$ for some $\gamma \in \Gamma'$.

The arguments in (a) and (b) clearly prove (2). They also prove (1): Let $\gamma \in \Gamma$, and choose $\tau \in \Omega$ different from ω and i, so that the stabilizer of τ in Γ is ± 1. Then by (b), applied to $\gamma^{-1}\tau$, there is a $\gamma' \in \Gamma'$ such that $\gamma'(\gamma^{-1}\tau) \in \Omega$. This implies that $\gamma'\gamma^{-1}\tau = \tau$, hence $\gamma'\gamma^{-1} = \pm 1$, and $\gamma \in \Gamma'$. □

We can now give $Y = \Gamma\backslash\mathfrak{H}$ the structure of a Riemann surface. Near every point τ_0 whose stabilizer is ± 1 we use $\tau - \tau_0$ as a local parameter. Near i (or points in its Γ-orbit) we use as a local parameter a function z having a zero of order 2 at i and satisfying $z \circ R = z$ (prove that such a function exists; note that the isotropy group of i in $\bar{\Gamma} = \Gamma/\pm 1$ is generated by R). Similarly near ω (or points in its Γ-orbit) we use as a local parameter a function z having a zero of order 3 at ω and satisfying $z \circ RT = z$ (note that RT is the generator of the isotropy group of ω in $\bar{\Gamma}$). It is now immediate that these are local homeomorphisms from neighborhoods of the points where they are defined onto open disks, and that change of coordinates is analytic, so we have endowed Y with a structure of a Riemann surface. *Topologically*, Y is obtained by gluing Ω along the identifications supplied by R and T on its boundary. With a little of imagination, we see that this is a once-punctured 2-sphere.

5.4. **Modular forms for $SL_2(\mathbb{Z})$ and the j-function.**

We shall now construct a certain holomorphic function j on $\mathfrak{H}$ which induces a global isomorphism of Riemann surfaces of $\Gamma\backslash\mathfrak{H}$ onto the affine line $\mathbb{C}$.

For a lattice Λ, the discriminant of the polynomial $4x^3 - g_2(\Lambda)x - g_3(\Lambda)$ is non zero. Up to an easily calculated constant, it is given by $\Delta = \Delta(\Lambda)$:

$$\Delta = g_2^3 - 27g_3^2. \tag{5.10}$$

As usual we write $\Delta(\tau)$ for $\Delta(\mathbb{Z}\tau + \mathbb{Z})$. Thus $\Delta(\tau)$ is a nowhere vanishing holomorphic function in $\mathfrak{H}$. Let

$$j = \frac{1728 g_2^3}{\Delta}. \tag{5.11}$$

Quite generally, a function f on the set of lattices is called a *modular form of weight* k $(k \in \mathbb{Z})$ if

$$f(c\Lambda) = c^{-k} f(\Lambda). \tag{5.12}$$

Taking $c = -1$ we see that if f is not identically 0, k must be even.

Exercise 5.2. *A modular form of weight which is not divisible by* 4 *must vanish at the lattice* $\mathbb{Z}i+\mathbb{Z}$. *A modular form of weight which is not divisible by* 6 *must vanish on the lattice* $\mathbb{Z}\omega + \mathbb{Z}$.

Exercise 5.3. *The sum of two modular forms of the same weight is again a modular form of the same weight. The product of modular forms of weights* k *and* l *is a modular form of weight* $k + l$.

For example, g_2 is modular of weight 4, g_3 is modular of weight 6, Δ is of weight 12 and j is of weight 0.

Lemma 5.6. *Put* $f(\tau) = f(\mathbb{Z}\tau + \mathbb{Z})$ *and assume that* f *is modular of weight* k. *Then for every* $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma = SL_2(\mathbb{Z})$

$$\begin{aligned} f(\gamma\tau) &= f(\frac{a\tau + b}{c\tau + d}) \\ &= (c\tau + d)^k f(\tau). \end{aligned} \tag{5.13}$$

We leave the easy proof as an exercise.

In particular, $f(\tau + 1) = f(\tau)$, so (under the usual conditions of smoothness) f has a Fourier expansion of the form $(\tau = x + iy)$

$$f \sim \sum_{n=-\infty}^{\infty} a_n(y) \exp(2\pi i n x). \tag{5.14}$$

If furthermore f is holomorphic, we have (prove!)

$$f = \sum_{n=-\infty}^{\infty} a_n \exp(2\pi i n \tau). \tag{5.15}$$

We call f *holomorphic at infinity* if $a_n = 0$ for $n < 0$ and a *cusp form* (or cuspidal) if $a_0 = 0$ as well. The space of holomorphic modular forms of weight k which are also holomorphic at infinity (resp. cusp forms) is denoted by $M_k(\Gamma)$ (resp. $S_k(\Gamma)$).

5.5. **Modular forms for other subgroups.** For future reference we make the following more general definition.

Definition 5.2. *Let Γ be any* subgroup *of finite index in $SL_2(\mathbb{Z})$. A function f on $\mathfrak{H}$ is called a* modular form of weight k on Γ *if the conclusion of the lemma is satisfied for $\gamma \in \Gamma$. It is said to be holomorphic, meromorphic, real analytic or smooth, if f is such as a function on $\mathfrak{H}$.*

Remark 5.1. *If $-1 \notin \Gamma$ the weight may now be odd.*

Remark 5.2. *The definition of "holomorphic at infinity" or "cuspidal" has to be modified as there are now more than one cusp in general in $\Gamma\backslash\mathfrak{H}$. We shall address this question later.*

For general Γ there does not exist a modular interpretation for the space $\Gamma\backslash\mathfrak{H}$, of the sort we had for $\Gamma = SL_2(\mathbb{Z})$, namely as the moduli space of lattices. However, as we shall see later, for a certain important class of $\Gamma's$ (the *congruence subgroups*), such a modular interpretation exists and $\Gamma\backslash\mathfrak{H}$ classifies lattices *with extra structure.*

5.6. **The j-line.** We want to prove the following theorems.

Theorem 5.7. *The function $j(\tau)$ induces an isomorphism of Riemann surfaces between $\Gamma\backslash\mathfrak{H}$ and $\mathbb{C}$.*

Theorem 5.8. *Let f be a holomorphic modular form of weight k (for $SL_2(\mathbb{Z})$) which is meromorphic at infinity. Let $ord_\tau(f) \geq 0$ be the order of (the zero) of f at τ, and $ord_\infty(f) = n$ if a_n is the lowest non-zero Fourier coefficient (thus $n \geq 0$ if f is holomorphic at infinity and $n > 0$ if it is a cusp form). Then*

$$\sum_{\tau\in\Omega,\,\tau\neq i,\omega} ord_\tau(f) + \frac{1}{2}ord_i(f) + \frac{1}{3}ord_\omega(f) + ord_\infty(f) = \frac{k}{12}. \tag{5.16}$$

Proof. Clearly $j(\tau) = j(\gamma\tau)$ for ever $\gamma \in \Gamma$. Applying the second theorem (with $k = 12$) for $\Delta(\tau)$, and noting that Δ does not vanish in $\mathfrak{H}$, we see that Δ has a simple zero at infinity. It follows that j has a simple pole at infinity, and so does $j(\tau) - j_0$ for any j_0. Applying the second theorem (for $k = 0$) for $j - j_0$, we see that $j - j_0$ has one simple (double if $j_0 = j(i) = 1728$, triple if $j_0 = j(\omega) = 0$) zero, and none else. This shows that j is one-to-one and onto, and also that $j(\tau) - j(\tau_0)$ is a local parameter near τ_0 for any τ_0 (recall how the Riemann surface structure was defined!).

To prove the first theorem integrate $df/2\pi if$ along a path which starts at $(-1/2, Y)$ and goes horizontally to $(1/2, Y)$ (for some $Y >> 0$), then follows $\partial\Omega$ clockwise except for small circular loops around zeros of f which might fall on $\partial\Omega$. For zeroes lying on the vertical lines at $x = \pm 1/2$, the two circular loops should be one the T-transform of the other (one sticking into the fundamental domain, the other out of it). For zeroes lying on $|\tau| = 1$ but different from i, ω, or $\omega + 1$ the two circular arcs should be one the R-transform of the other (again one sticking outside the unit disk, the other one into it). The circular loop around i (if needed) should start at some τ, $|\tau| = 1$, go above the unit circle, meeting it again at $-1/\tau$. The circular loops around ω and $\omega + 1$ (when needed) should also start and end at points which are related via R and T, and should both stick into Ω (leaving ω and $\omega + 1$ outside). Now the integral over the horizontal segment gives $ord_\infty(f)$, the integrals along the vertical lines cancel out each other, the integral along the small circular arcs around i and ω, $\omega+1$, give (in the limit) $\frac{1}{2}ord_i(f)+\frac{1}{3}ord_\omega(f)$, and the integral over the rest of the path (on $|\tau|=1$), when we take the fact that

$$f(\frac{-1}{\tau})=\tau^k f(\tau) \tag{5.17}$$

into account, gives $-k/12$ (the 12 coming from the angle between i and ω which is 30°). On the other hand, by Cauchy's theorem, the same integral equals

$$-\sum_{\tau\in\Omega,\,\tau\neq i,\omega} ord_\tau(f). \tag{5.18}$$

Rearranging the terms on both sides of the equation we derive the formula. □

Because of the theorem the moduli space of lattices $\Gamma\backslash\mathfrak{H}$ is sometimes called the j-line. We may compactify $\mathbb{C}$ to get $\mathbb{P}^1(\mathbb{C})$. In terms of $\Gamma\backslash\mathfrak{H}$ this means adding the point "$i\infty$" with local parameter $q=\exp(2\pi i\tau)$ around it. The notation q, like τ, is well-established and used by everybody. The Fourier expansion is therefore also called the *q-expansion* of the modular form f at infinity (it is nothing but the Taylor expansion in terms of the local parameter q on the compactified j-line).

Corollary 5.9. *Two elliptic curves given by two Weierstrass equations are isomorphic if and only if they have the same j-invariants.*

Proof. Let the two elliptic curves be given by the lattices Λ and Λ', and let j and j' be their j-invariants. We have seen that $j=j'$ if and only if they lattices are homothetic. □

Corollary 5.10. *For every separable cubic $h(x)$ the curve $y^2=h(x)$ has a non-singular projective completion, which is an elliptic curve.*

Proof. That the projective curve given by the homogenized equation is non-singular is proved exactly as before (the only fact used in the proof was the separability of h). Via a simple change of variables we may bring the equation to the form

$$y^2=4x^3-g_2x-g_3 \tag{5.19}$$

for some g_2 and $g_3\in\mathbb{C}$. What we have to prove (since our starting point in the definition of an elliptic curve was that of lattices) is that there exists a lattice Λ with $g_2=g_2(\Lambda)$ and $g_3=g_3(\Lambda)$. First we find a lattice Λ such that

$$j(\Lambda)=\frac{1728g_2^3}{g_2^3-27g_3^2}. \tag{5.20}$$

Note that the denominator does not vanish by the assumption that h is separable. The existence of Λ is guaranteed by the theorem (surjectivity of j). We now change Λ by a homothety so that $g_2(\Lambda)=g_2$. Since this does not affect the j-invariant it follows that $g_3^2(\Lambda)=g_3^2$. If we are unlucky and $g_3(\Lambda)=-g_3$, multiply Λ by a fourth root of unity (not changing $g_2(\Lambda)$) to fix $g_3(\Lambda)$. □

5.7. **More corollaries.** There are many corollaries that can be drawn from the second theorem. We leave most of the proofs as exercises.

Corollary 5.11. *If $k<0$, there are no modular forms of weight k that are also holomorphic at infinity. In other words, $M_k(\Gamma)=0$. Similarly, if $k<12$, $S_k(\Gamma)=0$.*

Corollary 5.12. *We have $M_2(\Gamma)=0$, $M_4(\Gamma)=\mathbb{C}g_2$, $M_6(\Gamma)=\mathbb{C}g_3$, $M_8(\Gamma)=\mathbb{C}g_2^2$, $M_{10}(\Gamma)=\mathbb{C}g_2g_3$.*

To prove the last corollary, note that for small k the orders at all $\tau's$ are dictated by the formula of the second theorem. If two modular forms of the same weight have the same orders everywhere (including at infinity), then their ratio must be a constant (why?).

Corollary 5.13. *We have* $S_{12}(\Gamma) = \mathbb{C}\Delta$, *and* $M_{12}(\Gamma) = \mathbb{C}\gvertneqq \bigoplus \mathbb{C}g_3^2$. *For every weight* $k \geq 12$ *we have*

$$S_k(\Gamma) = \Delta . M_{k-12}(\Gamma). \tag{5.21}$$

For the last corollary recall that Δ does not vanish anywhere in $\mathfrak{H}$ and has a simple zero at infinity so if we divide by it a given cusp form f, f/Δ is still a modular form, of weight 12 less, which is holomorphic at infinity.

Corollary 5.14. *As a graded ring (graded by the weight)* $\bigoplus_{k\geq 0} M_k(\Gamma) = \mathbb{C}[g_2, g_3]$ *and there are no relations between* g_2 *and* g_3.

That the two modular forms g_2 and g_3 generate the ring of modular forms for $SL_2(\mathbb{Z})$ follows from the previous corollary. That there are no relations is more challenging. Prove first that any relation must be *isobaric* of a given weight. Then look at orders at i and ω to get a contradiction.

Do not expect to get results of this sort for the corresponding spaces for general subgroups Γ. The spaces of modular forms and cusp forms get much richer then and the problem of constructing bases for them is difficult (although a lot of work has been done on it). However, calculating their dimensions is relatively easy (look it up in the literature).

5.8. **Eisenstein Series.** You may have asked yourself why not define modular forms of weight k by the formula

$$G_k(\Lambda) = \sideset{}{'}\sum_{\omega\in\Lambda} \omega^{-k} \tag{5.22}$$

for every even $k \geq 4$. These are easily shown to be holomorphic modular forms (even holomorphic at infinity) of weight k. They are called the (holomorphic) *Eisenstein series* for $SL_2(\mathbb{Z})$. One clearly has

$$M_k(\Gamma) = S_k(\Gamma) \oplus \mathbb{C}.G_k. \tag{5.23}$$

When the space of modular forms is one-dimensional ($k < 12$ or $k = 14$) we may deduce from this discussion that certain modular forms, e.g. G_4^2 and G_8, or G_4G_6 and G_{10}, or G_8G_6, $G_4^2G_6$ and G_{14}, must be all constant multiples of one another. These constants can be easily worked out once we find the constant terms in the Fourier expansion of the G_k. Comparing the higher Fourier coefficients (which turn out to be expressible as sums of $k-1$ powers of divisors of the integer n), one gets fascinating combinatorial identities between the functions

$$\sigma_{k-1}(n) = \sum_{d|n} d^{k-1}. \tag{5.24}$$

See for example J.-P. Serre, A Course in Arithmetic.

6. Descent and the weak Mordell-Weil theorem

6.1. **Multiplication by m on the elliptic curve.** Up till now our approach to elliptic curves has been very classical, and the methods we have been using were mainly those of complex function theory.

We now begin the study of the arithmetic of elliptic curves which are defined over a number field F. This requires some basic algebraic number theory, and to minimize the necessary background we confine ourselves to the case $F = \mathbb{Q}$. Unfortunately, even if we want to stick to the analysis of E over $\mathbb{Q}$ we have to enlarge $\mathbb{Q}$ occasionally and consider larger fields, in order to draw conclusions about $E(\mathbb{Q})$. Thus our proofs will be incomplete even for E which are defined over $\mathbb{Q}$. However, all the ideas will be present in our analysis, and the student with the necessary background in algebraic number theory will be able to generalize our results and fill in the gaps.

Consider an elliptic curve E such that E, and the four torsion points of order 2 on E are all defined over $\mathbb{Q}$. We may therefore assume that E is given by the equation

$$Y^2 = (X - \alpha)(X - \beta)(X - \gamma). \tag{6.1}$$

Here α, β, γ are three distinct rational numbers and the discriminant of the cubic

$$\Delta = (\alpha - \beta)^2(\alpha - \gamma)^2(\beta - \gamma)^2. \tag{6.2}$$

Let S be the smallest set of primes such that if $p \notin S$ then $p > 2$, α, β and γ are integral at p, and all of them are distinct modulo p. Thus S is finite, and $\Delta \in \mathbb{Z}_S^\times$ where $\mathbb{Z}_S$, the ring of S-integers, is the subring of rational numbers whose denominators are divisible only by primes from S. The group $\mathbb{Z}_S^\times$ is finitely generated.

Let $m > 1$ and denote by $[m]$ multiplication by m on $E(\mathbb{C})$.

Lemma 6.1. *$E_m = \ker([m])$ is isomorphic to $(\mathbb{Z}/m\mathbb{Z})^2$. We have a short exact sequence of $G_\mathbb{Q}$-modules where $G_\mathbb{Q} = Gal(\bar{\mathbb{Q}}/\mathbb{Q})$:*

$$0 \to E_m \to E(\bar{\mathbb{Q}}) \overset{[m]}{\to} E(\bar{\mathbb{Q}}) \to 0. \tag{6.3}$$

Proof. Everything is obvious if $\bar{\mathbb{Q}}$ is replaced by $\mathbb{C}$, using the complex uniformization. To see that $E(\bar{\mathbb{Q}})$ is m-divisible and that E_m are algebraic over $\mathbb{Q}$, use the fact that (in projective coordinates) multiplication by m is given by

$$[m](X : Y : Z) = (F_m(X, Y, Z) : G_m(X, Y, Z) : H_m(X, Y, Z)) \tag{6.4}$$

where F_m, G_m and H_m are homogeneous polynomials of the same degree with coefficients from $\mathbb{Q}$ with no common factor (use induction on the explicit addition formulas). A priori, we can not rule out the possibility that F_m, G_m and H_m all vanish at some $(x : y : z) \in E$, so that $[m]$ is not defined at such a point by the above formula.[1] However, this can happen at most at finitely many points S. This implies that for any $\sigma \in Aut(\mathbb{C})$ (field automorphism) and for any $P \in E(\mathbb{C})\backslash S$

$$\sigma([m]P) = [m](\sigma P). \tag{6.5}$$

[1] In general, a morphism of projective varieties need not be defined by a single collection of polynomials not having a common zero. Instead, there are finitely many collections of this sort, each defined in a Zariski open set, which agree where their domains of definition overlap. A more careful analysis of multiplication by m on the elliptic curve will show that one such collections suffices, but we do not go into it; see the series of papers by Cassels on the arithmetic of elliptic curves from the 1960's.

If $P \in S$ we apply this for Q and $P+Q$ such that both Q and $P+Q$ are not in S, and subtract, to get that the equation still holds for $P \in S$.

In particular if $P \in E_m$ then $\sigma P \in E_m$ as well (because σ fixes $O = (0:1:0)$). Thus a point from E_m has only finitely many conjugates under $Aut(\mathbb{C})$ so must be algebraic. Similarly if $[m]P = Q$ and $Q \in E(F)$, $[F:\mathbb{Q}] < \infty$, for any $\sigma \in Aut(\mathbb{C}/F)$, $\sigma P - P \in E_m$, so again there are only finitely many possibilities for σP, and P is algebraic. Finally we have just remarked that $[m]$ is a Galois-homomorphism. □

6.2. **Galois cohomology.** Let G be a topological group (*e.g.* $G_{\mathbb{Q}}$ with the Krull topology) and M a discrete G-module. Let

$$Z^1(G,M) = \left\{ \begin{array}{c} c : G \to M;\ c \text{ is continuous (i.e. locally constant)} \\ \text{and } c(\sigma\tau) = \sigma c(\tau) + c(\sigma) \end{array} \right\} \tag{6.6}$$

the group of "crossed homomorphisms". Notice that if G acts trivially on M then these are just the continuous homomorphisms from G to M. Let

$$B^1(G,M) = \{c;\ \text{for some } m \in M,\ c(\sigma) = \sigma m - m\}. \tag{6.7}$$

If the action is trivial, this is 0. The quotient group

$$H^1(G,M) = Z^1(G,M)/B^1(G,M) \tag{6.8}$$

is called the first cohomology group of G in M (or *with coefficients in* M).

For a G-homomorphism $f : M \to N$ we get a natural group homomorphism

$$f_* : H^1(G,M) \to H^1(G,N), \tag{6.9}$$

and $(fg)_* = f_*g_*$. In other words, $H^1(G,-)$ is a *functor* from the category of G-modules to abelian groups. For example, if $M = N$ and $f = [m]$ is multiplication by m in the G-module M, then $[m]_*$ is multiplication by m in $H^1(G,M)$ (check!). The importance of the functor H^1 stems from the fact that it appears in long exact sequences as a measure of the failure of the functor of "taking invariants" to be exact. More precisely we have the following.

Theorem 6.2. *Let $0 \to M \to N \to P \to 0$ be an exact sequence of discrete G-modules. Denote by M^G the G-invariants in M etc. For $p \in P^G$ define $\delta(p) \in H^1(G,M)$ as follows. Lift p to $n \in N$ (which need not be invariant under G anymore) and consider*

$$c(\sigma) = \sigma n - n \in M. \tag{6.10}$$

Then c is a crossed homomorphism and its cohomology class $[c] = \delta(p)$ is well defined (does not depend on the choice of the lifting n). The "long" sequence

$$0 \to M^G \to N^G \to P^G \xrightarrow{\delta} H^1(G,M) \to H^1(G,N) \to H^1(G,P) \tag{6.11}$$

is exact.

Proof. We leave te proof as a "no choice" exercise. □

Remark 6.1. *This long exact sequence can be continued indefinitely with the introduction of the higher group cohomology functors $H^i(G,-)$ for all $i \geq 1$. Thus certain elements from $H^2(G,M)$ will measure the failure of $H^1(G,N) \to H^1(G,P)$ to be surjective etc.*

6.3. **Example: Hilbert's theorem 90.** Let F be any field, $\bar{F}$ a separable closure, and m an integer relatively prime to $char.F$. Let $\mu_m \subset \bar{F}^\times$ be the group of m-th roots of unity. It is cyclic of order m, and $G = Gal(\bar{F}/F)$ acts on it via the cyclotomic character modulo m, $\chi : G \to (\mathbb{Z}/m\mathbb{Z})^\times$, defined by the equation

$$\sigma(\zeta) = \zeta^{\chi(\sigma)} \ (\zeta \in \mu_m). \tag{6.12}$$

Consider the short exact sequence

$$0 \to \mu_m \to \bar{F}^\times \overset{[m]}{\to} \bar{F}^\times \to 0 \tag{6.13}$$

where $[m](x) = x^m$. The associated long exact sequence is

$$0 \to \mu_m(F) \to F^\times \overset{[m]}{\to} F^\times \overset{\delta}{\to} H^1(G,\mu_m) \to H^1(G,\bar{F}^\times) \overset{[m]_*}{\to} H^1(G,\bar{F}^\times). \tag{6.14}$$

Here $\mu_m(F)$ is the group of m-th roots of unity in F.

Theorem 6.3. *(Hilbert's Theorem 90)* $H^1(G,\bar{F}^\times) = 0$.

Proof. We shall deduce this fundamental theorem from another fundamental theorem, Artin's theorem on the linear independence of characters of fields. See any book on Galois theory for the latter. Let $c \in Z^1(G,\bar{F}^\times)$. Since c is continuous, it factors through a finite Galois extension L/F. Let $G_{L/F}$ be its Galois group, a quotient of G by an open normal subgroup, so that $c \in Z^1(G_{L/F}, L^\times)$. Consider the equation

$$\sum_{\tau \in G_{L/F}} c(\tau)\tau(x) \neq 0. \tag{6.15}$$

By Artin's theorem, there is an $x \in L$ for which the left hand side is nonzero, and we denote its reciprocal by b. Now

$$\begin{aligned}\sigma(b)^{-1} &= \sum_\tau \sigma c(\tau)\sigma\tau(x) \\ &= c(\sigma)^{-1}\sum_\tau c(\sigma\tau)\sigma\tau(x) \\ &= c(\sigma)^{-1}b^{-1}\end{aligned} \tag{6.16}$$

yielding $c(\sigma) = \sigma(b)/b$. Thus c is a coboundary, so its class in H^1 vanishes. □

Hilbert's theorem 90 gives rise to the identification

$$H^1(G,\mu_m) \simeq F^\times/F^{\times m}. \tag{6.17}$$

There is a special case worth special attention, and this is when $\mu_m \subset F$. This is always the case when $m = 2$. In this case, the Galois action on the roots of unity is trivial, so we get

$$Hom(G,\mu_m) \simeq F^\times/F^{\times m}. \tag{6.18}$$

Tracing back the isomorphisms, and element $amodF^{\times m}$ on the right yields the homomorphism

$$\kappa_a : \sigma \mapsto \sigma(\alpha)/\alpha \tag{6.19}$$

where $\alpha^m = a$. This homomorphism indeed is independent of α. The assertion that every homomorphism from G to μ_m is a κ_a is known as *Kummer theory.* In different words it says that given a cyclic extension L/F of order m, it is obtained by extracting an m-th root from an element a of F. Moreover, any isomorphism of $Gal(L/F)$ with a subgroup of μ_m is a κ_a for some a, unique up to mth powers.

Exercise 6.1. *Prove that if the extension L/F is also obtained by extracting mth roots of b, then for some i relatively prime to m, and some $c \in F^\times$,*

$$b = a^i c^m. \tag{6.20}$$

6.4. **The descent exact sequence.** Let $G = G_{\mathbb{Q}}$ be the Galois group of $\bar{\mathbb{Q}}$ over $\mathbb{Q}$. Applying Galois cohomology to the short exact sequence of multiplication by m on E we get, since $E(\bar{\mathbb{Q}})^G = E(\mathbb{Q})$, a new (long) exact sequence

$$\begin{aligned} 0 \quad \to \quad & E_m(\mathbb{Q}) \to E(\mathbb{Q}) \overset{[m]}{\to} E(\mathbb{Q}) \overset{\delta}{\to} \\ \to \quad & H^1(G, E_m) \to H^1(G, E(\bar{\mathbb{Q}})) \overset{[m]_*}{\to} H^1(G, E(\bar{\mathbb{Q}})) \end{aligned} \tag{6.21}$$

from where we get a short exact sequence (the *descent exact sequence*)

$$0 \to E(\mathbb{Q})/mE(\mathbb{Q}) \to H^1(G, E_m) \to H^1(G, E)_m \to 0. \tag{6.22}$$

Here E is a short-hand for $E(\bar{\mathbb{Q}})$ and for any abelian group X, X_m denotes the m-torsion in X. Notice the similarity with the previous example, on the multiplicative group. However, the analogue of Hilbert's theorem 90 does not hold anymore, and this makes the elliptic curve case much more difficult.

This short exact sequence allows us to analyze $E(\mathbb{Q})/mE(\mathbb{Q})$ via the much easier module E_m with the associated Galois action. We shall now see what it is for $m = 2$, under the simplifying assumption that E_2 is defined over $\mathbb{Q}$ (as above), so that

$$H^1(G, E_2) = Hom(G, E_2). \tag{6.23}$$

If we identify E_2 with μ_2^2 (where $\mu_2 = \{\pm 1\}$) so that

$$(\alpha, 0) \mapsto (-1, 1),\ (\beta, 0) \mapsto (1, -1),\ (\gamma, 0) \mapsto (-1, -1) \tag{6.24}$$

(this is an isomorphism of G-modules since the G action on E_2 is trivial), we get from Hilbert's theorem 90

$$H^1(G, E_2) = H^1(G, \mu_2)^2 = [\mathbb{Q}^\times/(\mathbb{Q}^\times)^2]^2. \tag{6.25}$$

6.5. **Reduction modulo p.** In preparation for the Mordel-Weil theorem, we have to study elliptic curves over finite fields, and reduction modulo p.

Let us adopt the ad-hoc definition that an elliptic curve over any field of characteristic $\neq 2$ is a plane projective curve of the form $y^2 = h(x)$ where h is a separable cubic.

If we start with our rational example $E : y^2 = (x - \alpha)(x - \beta)(x - \gamma)$ and take $p \notin S$, then reading this equation "modulo p" will give us an elliptic curve $\tilde{E}$ over $\mathbb{F}_p$, the *reduced elliptic curve.* The polynomial equations defining the group law on E define also a group law on $\tilde{E}$, after we clear denominators in the equations, which is again associative and commutative, and has $(0 : 1 : 0)$ for its neutral element.

If $P = (x : y : z)$ is a point of $E(\mathbb{Q})$, clearing denominators we may assume that the coordinates are integral and not all divisible by p. Reducing modulo p we get a point $\tilde{P} \in \tilde{E}(\mathbb{F}_p)$. Reduction is a homomorphism, with respect to the group laws on E and on $\tilde{E}$:

$$(P + Q)^{\tilde{}} = \tilde{P} + \tilde{Q}. \tag{6.26}$$

This needs a proof, which can be based on the explicit equations, and which we omit.

More generally, if L is a finite extension of $\mathbb{Q}$, $\mathcal{O}_L$ its ring of algebraic integers, and $\mathfrak{p}$ a prime whose intersection with $\mathbb{Z}$ is $p\mathbb{Z}$ (we say that $\mathfrak{p}$ lies above, or divides, p), then the same procedure defines the map "reduction modulo $\mathfrak{p}$" from $E(L)$ to $\tilde{E}(\mathcal{O}_L/\mathfrak{p})$, which is again a homomorphism, and which we continue to denote by $P \mapsto \tilde{P}$.

In general, it is very much possible for two distinct points P and Q to undergo the same reduction $\tilde{P} = \tilde{Q}$. In fact, it should be intuitively clear that this is always the case if P and Q are p-adically "close" to each other, e.g. if their coordinates, after clearing denominators, are congruent modulo p. However, the fact that α, β and γ remain distinct modulo p means that *the four torsion points of order* 2 remain distinct in $\tilde{E}(\mathbb{F}_p)$, after reduction. The same is true for any m which is relatively prime to p : the torsion points E_m (once embedded in a field L, and once a prime $\mathfrak{p}$ above p has been chosen as above) reduce injectively modulo $\mathfrak{p}$. However, we shall only need this fact for $m = 2$, where it becomes obvious from our assumptions.

6.6. **The weak Mordell-Weil theorem.** The weak Mordell-Weil theorem (for $m = 2$) is the following statement.

Theorem 6.4. *The group $E(\mathbb{Q})/2E(\mathbb{Q})$ is finite.*

Proof. All proofs go by showing that the image of $E(\mathbb{Q})/2E(\mathbb{Q})$ inside $H^1(G, E_2)$, i.e. inside $[\mathbb{Q}^\times/(\mathbb{Q}^\times)^2]^2$, lies in a strictly smaller subgroup which is finite.

Let S be as above. Then $\mathbb{Z}_S^\times \subset \mathbb{Q}^\times$ is the finitely generated subgroup of rational numbers all of whose prime factors belong to S. Clearly $\mathbb{Z}_S^\times/(\mathbb{Z}_S^\times)^2$ is a finite group, of order $2^{\#S+1}$. We shall show that $E(\mathbb{Q})/2E(\mathbb{Q})$ is contained in $[\mathbb{Z}_S^\times/(\mathbb{Z}_S^\times)^2]^2$.

For that purpose we fix a decomposition group G_p of the prime p in G and consider the localization of the descent exact sequence with $m = 2$:

$$\begin{tikzcd} 0 \arrow[r] & E(\mathbb{Q})/2E(\mathbb{Q}) \arrow[r] \arrow[d] & H^1(G, E_2) \arrow[r, phantom, "="] \arrow[d] & {[\mathbb{Q}^\times/(\mathbb{Q}^\times)^2]^2} \arrow[d] \\ 0 \arrow[r] & E(\mathbb{Q}_p)/2E(\mathbb{Q}_p) \arrow[r] & H^1(G_p, E_2) \arrow[r, phantom, "="] & {[\mathbb{Q}_p^\times/(\mathbb{Q}_p^\times)^2]^2} \end{tikzcd} \tag{6.27}$$

where the rows are exact and the squares commute.

Let I_p be the inertia subgroup at p, so that G_p/I_p is (pro)cyclic, generated by the Frobenius automorphism. The kernel of the restriction homomorphism

$$H^1(G_p, E_2) \to H^1(I_p, E_2) \tag{6.28}$$

is just $H^1(G_p/I_p, E_2)$ (the *inflation-restriction* exact sequence). Since G_p/I_p is identified with the absolute Galois group of the finite field $\mathbb{F}_p$, we have, in the same way as before

$$\begin{aligned} H^1(G_p/I_p, E_2) &= [\mathbb{F}_p^\times/(\mathbb{F}_p^\times)^2]^2 \\ &= [\mathbb{Z}_p^\times/(\mathbb{Z}_p^\times)^2]^2 \subset [\mathbb{Q}_p^\times/(\mathbb{Q}_p^\times)^2]^2. \end{aligned} \tag{6.29}$$

If we show that for $P \in E(\mathbb{Q}_p)$, $\delta(P) \in H^1(G_p, E_2)$ has a trivial restriction to I_p (is *unramified*), then $\delta(P) \in H^1(G_p/I_p, E_2)$, so the mod 2 - valuation of $\delta(P)$, viewed as an element of $[\mathbb{Q}_p^\times/(\mathbb{Q}_p^\times)^2]^2$, is 0. For a global point $P \in E(\mathbb{Q})$ this implies now that $\delta(P) \in [\mathbb{Z}_S^\times/(\mathbb{Z}_S^\times)^2]^2$ as desired.

All that remains to show is the following lemma. □

Lemma 6.5. *Let $p \notin S$. Then $\delta(P)$, for any $P \in E(\mathbb{Q}_p)$, is unramified.*

Proof. Let $Q \in E(\bar{\mathbb{Q}}_p)$ satisfy $[2](Q) = P$ and pick $\sigma \in I_p$. We have to show that $\sigma Q = Q$. Let $\tilde{E}$ be the elliptic curve over $\mathbb{F}_p$ obtained by reducing the equation of E modulo p. Here we use the fact that p is not in S. Let $\tilde{Q} \in E(\bar{\mathbb{F}}_p)$ be the reduction of Q. Since σ acts trivially on $\bar{\mathbb{F}}_p$, $\sigma\tilde{Q} - \tilde{Q} = 0$. This means that the 2-torsion point $\sigma Q - Q$ reduces to 0. However, the four 2-torsion points reduce modulo p to four distinct points (here we use the fact that p does not divide Δ) and so $\sigma Q = Q$. □

7. Heights on elliptic curves and the strong Mordell-Weil theorem

7.1. **Heights on abelian groups and finite generation.** Let A be an abelian group. A *height function* on A is a function h from A to $\mathbb{R}$ satisfying

(0) $h(0) = 0$.

(1) For every r, there are only finitely many $P \in A$ with $h(P) \le r$.

(2) Let $Q \in A$. Then there is a constant $0 < C_1(Q)$ such that for any $P \in A$

$$h(P - Q) \le 2h(P) + C_1(Q). \tag{7.1}$$

(3) There is an $m \ge 2$ and an absolute constant $0 < C_2$ such that

$$h(mP) \ge m^2 h(P) - C_2. \tag{7.2}$$

Example 7.1. *Suppose there is a symmetric bilinear form $\langle.,.\rangle$ on A, whose kernel K is finite, and such that on A/K the bilinear form is positive definite. Suppose $h(P) - q(P)$, where $q(P) = \langle P, P\rangle$ is bounded on A. Then from the parallelogram law*

$$q(P - Q) + q(P + Q) = 2q(P) + 2q(Q) \tag{7.3}$$

it is evident that (2) and (3) are satisfied.

Proposition 7.1. *Suppose A/mA is finite and A admits a height function (with the same m in (3)). Then A is finitely generated.*

Proof. Let $Q_1, \dots, Q_r$ be representatives for A/mA, and let $C_1 = \max C_1(Q_i)$. For $P_0 \in A$ write inductively $(1 \le i)$

$$P_{i-1} = Q_{j_i} + mP_i. \tag{7.4}$$

By (3) and (2)

$$\begin{aligned} h(P_n) &\le m^{-2}\left(h(mP_n) + C_2\right) \\ &\le m^{-2}\left(h(P_{n-1} - Q_{j_n}) + C_2\right) \\ &\le m^{-2}\left(2h(P_{n-1}) + C\right) \end{aligned} \tag{7.5}$$

where $C = C_1 + C_2$. Now it is an easy exercise to show that there exists an r depending only on C such that for n large enough $h(P_n) \le r$. Let $R_1, \dots, R_s$ be all the points P with $h(P) \le r$. Let A' be the subgroup generated by the Q_i and the R_j. If n is such that $h(P_n) \le r$, then P_n is some R_j so belongs to A'. We now deduce by descending induction on i that $P_i \in A'$, hence $P_0 \in A'$ and $A' = A$. □

7.2. **Heights on elliptic curves and on $\mathbb{P}^1$.** If E is the elliptic curve

$$y^2 = (x-\alpha)(x-\beta)(x-\gamma) \tag{7.6}$$

where $\alpha, \beta, \gamma \in \mathbb{Q}$ we define a function $h : A = E(\mathbb{Q}) \to \mathbb{R}$ by writing the x-coordinate of a point P in reduced terms as $x(P) = u/v$ ($u, v \in \mathbb{Z}$ relatively prime) and letting

$$h(P) = \max\left(\log|u|, \log|v|\right). \tag{7.7}$$

We let $h(O) = 0$.

Theorem 7.2. *h is a height function on $A = E(\mathbb{Q})$, with $m = 2$.*

Corollary 7.3. *The group $E(\mathbb{Q})$ is finitely generated.*

Properties (0) and (1) are obvious. We have to show (2) and (3) with $m = 2$. More generally, we define the height $h(x : y)$ of any point $(x : y) \in \mathbb{P}^1(\mathbb{Q})$ as $\max(\log|x|, \log|y|)$ if x and y are both integers and relatively prime to each other. Denoting by $\wp : E \to \mathbb{P}^1$ the morphism taking $P = (x : y : z)$ to $(x : z)$ (and $O = (0 : 1 : 0)$ to $(1 : 0)$) we have $h(P) = h(\wp(P))$.

Lemma 7.4. *We have*

$$h(x : y) = \sum_{p \le \infty} \max\left(\log|x|_p, \log|y|_p\right). \tag{7.8}$$

Here p runs over all the primes and ∞, $|x|_\infty$ is the usual absolute value, and

$$|x|_p = p^{-ord_p(x)}. \tag{7.9}$$

Proof. The lemma is obvious, because if x and y are relatively prime integers, then at every $p < \infty$, $\max(\log|x|_p, \log|y|_p) = 0$, so the right hand side reduces to the definition given above. □

The formula given by the lemma has the advantage that the right hand side remains unchanged even if x and y are not relatively prime integers. In fact, if we replace them by cx and cy for some rational number c, then the summand corresponding to p is changed by $\log|c|_p$, but the *product fromula*

$$\prod_{p \le \infty} |c|_p = 1 \tag{7.10}$$

implies that the total sum is unchanged. The right hand side makes sense for every number field, where lack of unique factorization does not allow us anymore to assume that x and y are relatively prime integers. One simply replaces the set of primes and ∞ by the set of all normalized valuations of the number field. Another generalization is to higher dimensional projective spaces. Over $\mathbb{Q}$, the (logarithmic) Weil height of a point $(x_0 : \cdots : x_n) \in \mathbb{P}^n(\mathbb{Q})$ is given by

$$h(x_0 : \cdots : x_n) = \sum_{p \le \infty} \max_{0 \le i \le n} \left(\log|x_i|_p\right). \tag{7.11}$$

It is well-defined because of the product formula, as before.

7.3. **Behaviour under morphism.** A morphism φ of degree d from $\mathbb{P}^1$ to itself (defined over $\mathbb{Q}$) is given by a pair of homogenous polynomials of degree d without a common zero

$$\varphi(x:y) = (F(x,y):G(x,y)). \tag{7.12}$$

The condition of not having a common zero is equivalent to the condition that for some N, both x^N and y^N are in the ideal generated by F and G in $\mathbb{Q}[x,y]$. This is a special case of the Nullstellensatz. Clearly, we may assume that F and G are both from $\mathbb{Z}[x,y]$.

Proposition 7.5. *Let φ be a morphism of degree d from $\mathbb{P}^1$ to itself. Then there is a constant $0 < C$ depending only on φ such that for every $(x:y) \in \mathbb{P}^1(\mathbb{Q})$*

$$dh(x:y) - C \leq h(\varphi(x:y)) \leq dh(x:y) + C. \tag{7.13}$$

Proof. □

Proposition 7.6. *Let E be the elliptic curve as above. There is a morphism $\varphi : \mathbb{P}^1 \to \mathbb{P}^1$ of degree 4 such that*

$$\varphi \circ \wp(P) = \wp(2P). \tag{7.14}$$

Corollary 7.7. *If h is the height function on E, then $|h(2P) - 4h(P)|$ is bounded on $E(\mathbb{Q})$.*

This gives (3) with $m = 2$. (It seems that (3) uses only half of the corollary, but this is the significant half - the other inequality is much easier).