ELLIPTIC CURVES

R. V. Gurjar, Kirti Joshi, N. Mohan Kumar,
Kapil H. Paranjape, A. Ramanathan, T. N. Shorey,
R. R. Simha, V. Srinivas

National Board for Higher Mathematics
Instructional Conference
held at
Tata Institute of Fundamental Research
Bombay

Sept 30–Oct 18, 1991

Foreword

The present volume is the first of a series entitled *Praveshika* planned to be brought out by the National Board of Higher Mathematics (NBHM) in collaboration with Hindustan Book Agency. It constitutes essentially the notes of lectures given at an Instructional School on Elliptic Curves held in Bombay during September 30–October 18, 1991.

NBHM is an agency set up by the Government of India in 1983 and charged with the responsibility of of promotion of Higher Mathematics in the country. It was the perception of the Board from the very beginning that advanced mathematical research of high quality was possible only in a milieu of deep and wide mathematical scholarship. It was also the view of the Board that the attention paid to this aspect of building a community of scholars in India had been inadequate. Among the many initiatives the Board took to correct this imbalance are a series of Instructional Schools in diverse areas of mathematics. The Schools are essentially introductions to different mathematical topics of current interest, which are considered central and important by the mathematical community at large. Catering as they do, too students and researchers with limited background, these schools necessarily begin at somewhat elementary level but attempt in the space of three or four weeks to give the motivated scholar some glimpses into the recent development into the field. These Schools have proved to be of value to the Indian Mathematical Community and it was felt that efforts to get the notes of such schools published should be made so as to reach out to a wider circle than merely participants; and when the team led by Mohan Kumar produced the present set of excellent notes the decision to embark on publications was made with great enthusiasm.

NBHM is very grateful to the team that produced these notes: R. V. Gurjar, Kirti Joshi, N. Mohan Kumar, Kapil Paranjape, A. Ramanathan, T. N. Shorey, R. R. Simha and V. Srinivas.

We are indeed proud and happy to offer this first volume of the series to the Mathematical Community. It will be our endeavour to keep to the high standards set by this volume for future publications in this series.

A long time has elapsed since the notes were made available to NBHM by the lecturers; the delay in their publication was caused by various admin-

istrative problems for which NBHM offers its apologies to the lecturers.

It is with deep regret that we have to record here the untimely demise of two people who contributed to the production of these notes: Professor A. Ramanathan who was one of the lecturers and Shri. D. S. Moorthy who provided substantial administrative assistance.

M. S. Raghunathan,

Chairman, NBHM.

Preface

These notes are based on the Instructional Conference on Elliptic Curves held at the Tata Institute of Fundamental Research in October 1991. This Instructional Conference was meant to give an introduction to the topic of Elliptic Curves, from the viewpoints of complex analysis, algebraic geometry and number theory. The presentation is intended to be at the level of a student beginning research in mathematics. We have tried to minimise the prerequisites as much as possible, and to keep the notes reasonably self-contained.

Elliptic Curves was felt to be an appropriate subject for this conference, because of the way in which the interconnections between different areas of mathematics are brought out. The topic has diverse roots: on the analytic side, in the study of elliptic integrals, which go back to Euler, and which were 'inverted' by Gauss, Abel, Jacobi etc. to obtain elliptic functions; on the number theoretic side, Fermat's 'method of descent' for cubic and quartic Diophantine equations provided the first major advance beyond the work of the Greeks. In 'abstract' algebraic geometry, they provided the earliest non-trivial examples of what is now known as Weil's Riemann Hypothesis for varieties over finite fields. The famous conjecture of Ramanujan on the τ function, which was proved by Deligne in 1973 as a consequence of his proof of Weil's Riemann Hypothesis, has been extensively generalised in the present day theory of automorphic forms. These 'generalised Ramanujan conjectures' are almost all unproved at present, and are the subject of intensive research.

A basic property of elliptic curves, which accounts for some of their special features, is that they are 1-dimensional projective algebraic groups. Abelian varieties, which are the higher dimensional generalisation of this notion, play an important role in algebraic geometry and number theory today, involving theta functions, moduli, automorphic forms, etc. One major success of the geometric and arithmetic theory of abelian varieties is Faltings' proof in 1983

of the Mordell Conjecture. This implies, for example, that for each $n \geq 4$, the equation $X^n + Y^n = Z^n$ has only a finite number of integer solutions with X, Y, Z relatively prime, which represents the best progress to date towards a proof of the notorious 'Fermat's Last Theorem'. The Diophantine properties of elliptic curves have constituted a major area of research in number theory in the last several years, with a central theme being the conjectures of Birch and Swinnerton-Dyer. Spectacular recent progress has been made with the work of Gross-Zagier and of Kolyvagin and Rubin. The Birch-Swinnerton-Dyer conjectures themselves have been refined and vastly generalised, giving rise to the recent conjectures of Bloch and Beilinson relating algebraic K-theory and values of L-functions. L-functions themselves are related to automorphic forms, Galois representations, transcendence theory, etc. More recently, there have been connections with coding theory, and with physics. All in all, the mathematics arising out of elliptic curves makes an imposing edifice.

Coming to these notes, the reader should be aware that the results discussed here are merely the prototypes for much more general theorems and techniques. In some cases, the proofs given are a bit artificial, and in some places, are given only for a special (but typical) case. Different lecturers have prepared notes for their respective sections; we have attempted to give these notes a certain coherence, while preserving the individual flavours of the presentations. Our hope is that in these lecture notes, we have succeeded to some small extent in conveying the fascinating quality of the theory of elliptic curves, as well as the essential unity of mathematics.

The table of contents will give the reader a good idea of what material is covered in these notes. In a sense, the development of the analytic theory is most complete. The necessity to avoid the use of schemes, cohomology and deeper commutative algebra has forced us to sometimes use less transparent, though 'elementary', arguments. Due to limitations of time, certain topics were also omitted in the algebraic geometry section, notably the algebraic theory of the j-function and moduli. Since these topics are treated in the complex analysis lectures, we felt that the lacuna would not be too great. In the number theory lectures, a decision was taken to work only over the field $\mathbf{Q}$ of rationals, whenever possible, since the extensions to number fields present only technical difficulties. We also discuss asymptotic and exact formulas for the number of representations of a positive integer as a sum of $4k$ squares, by comparing the Fourier coefficients of theta functions and Eisenstein series,

though this discussion has more to do with modular forms *i.e.* the moduli of elliptic curves, rather than elliptic curves *per se*.

Apart from the 'core' lectures in these notes, a glimpse of the connections with coding theory was obtained in lectures of Hiralal Janwa; T. R. Ramadas lectured on connections with physics, in the context of the 'eight vertex model', whose solution involves elliptic curves. The R-matrix, which has led to quantum groups, etc. was first introduced in this work.

We have not always attempted to give detailed attributions throughout the text for the sources of results and proofs. Apart from the sources cited, general references for algebraic geometry are the books of Hartshorne [27], Shafarevich [51], and Griffiths and Harris [25]. An excellent source for the arithmetic of elliptic curves is Silverman's book [57]; this also contains detailed references to other sources. A good source for comprehensive information on algebraic curves over the complex numbers is [3].

We are grateful to the National Board for Higher Mathematics, and in particular, its Chairman, Prof. M. S. Raghunathan, for supporting this effort. The help of the mathematics office staff at times of great stress is much appreciated. Special thanks are due to Shri D. S. Moorthy for administrative assistance, to Shri M. K. Priyan for help with word processing and computer typesetting, to Shri K. R. Krishnakumar for keeping the accounts and to Shri D. More for other assistance.

The lecturers
TIFR, Bombay

Contents

Preface **iii**

I Analytic theory of Elliptic Curves **1**

1 Doubly periodic functions **2**
1.1 Introduction 2
1.2 Periodic and elliptic functions 5

2 Riemann surfaces **15**
2.1 Generalities on Riemann surfaces 15
2.2 Examples 16
2.3 Holomorphic 1-forms 21
2.4 Integration of 1-forms 22

3 Tori **23**
3.1 Complex tori; Embedding in $\mathbf{P}^2$ 23
3.2 Addition theorem for $\wp$; the group law on cubics 26

4 Isomorphism of tori and the j invariant **30**
4.1 Isomorphism of tori 30
4.2 The functions $g_2(\tau)$, $g_3(\tau)$, $j(\tau)$ 31
4.3 Isomorphism of plane cubics in terms of j 33

5 All smooth cubics are complex tori **35**
5.1 The first proof 35
5.2 Second proof 41
5.3 Elliptic integrals 43

6 Moduli **46**
6.1 A fundamental region for $SL(2, \mathbf{Z})$ 46
6.2 Behaviour of the j function in R^0 48
6.3 The Riemann surface $\mathcal{H}/PSL(2, \mathbf{Z})$ 49
6.4 Biholomorphic maps of $\mathbf{C}/\Gamma$ 52

II Algebraic Geometry of Elliptic Curves **54**

7 Some results from commutative algebra **55**
7.1 Noetherian rings 55
7.2 Localisation 61
7.3 Integral extensions 66
7.4 Noether normalisation, and the Nullstellensatz 68
7.5 Discrete valuation rings 73
7.6 Dedekind domains 78

8 Varieties **83**
8.1 Affine varieties 83
8.2 Projective varieties 93
8.3 Morphisms 97
8.4 Smooth varieties 99

9 Further properties of varieties **102**
9.1 Product varieties 102
9.2 Completeness of projective varieties 104
9.3 Proper morphisms 107

10 Intersection theory for plane curves **109**
10.1 Intersection multiplicities 109
10.2 Applications of Bezout's Theorem 115

11 Geometry of curves **121**
11.1 Rational maps 121
11.2 Non-singular models 127
11.3 Group actions on curves 129
11.4 Varieties over non-algebraically closed fields 130
11.5 Divisors and linear equivalence 135

12 Geometry of elliptic curves **138**
12.1 Riemann-Roch for elliptic curves . . . 138
12.2 Differentials . . . 140
12.3 The chord and tangent construction . . . 146
12.4 Curves of genus 1 . . . 153
12.4.1 Cubic curves . . . 154
12.4.2 Double covers . . . 155
12.4.3 Complete intersections . . . 158

13 Structure of endomorphisms of elliptic curves **159**
13.1 Endomorphisms of elliptic curves . . . 159
13.2 The Frobenius morphism . . . 162
13.3 Dual isogenies . . . 166
13.4 Additivity of the invariant differential . . . 170
13.5 The Riemann Hypothesis . . . 173
13.6 The structure of the endomorphism ring . . . 175

III Arithmetic of Elliptic Curves **178**

14 Rational points on curves **179**
14.1 Curves of degree ≤ 2 . . . 179
14.1.1 Linear equations . . . 180
14.1.2 Equations of degree two . . . 181
14.2 Problems of higher degree . . . 188
14.2.1 The Fermat quartic . . . 189
14.2.2 Frenicle's problem . . . 191

15 The Mordell-Weil theorem for elliptic curves over Q **195**
15.1 The Mordell-Weil Theorem . . . 195
15.2 Heights . . . 204

16 Computing the Mordell-Weil group **219**
16.1 Descent . . . 219
16.2 Local methods . . . 221
16.3 Selmer and Tate-Shafarevich groups . . . 228
16.4 Torsion points . . . 231

16.4.1 Specialization modulo a prime 231
16.4.2 Mazur's theorem 238

17 Integer points, and the theorems of Thue and Siegel 243
17.1 Siegel's theorem on integer points of elliptic curves 243
17.2 Approximations of algebraic numbers by rationals: a theorem of Thue 247
17.2.1 Introduction 247
17.2.2 Two lemmas 250
17.2.3 Proof of Theorem 17.2.1 252
17.2.4 Proof of Theorem 17.2.2 253
17.2.5 Proof of Theorem 17.2.3 262

18 Representation of numbers by squares 264
18.1 The theta function 265
18.2 Eisenstein series 270
18.3 Cusp forms 277

19 The Conjecture of Birch and Swinnerton-Dyer 289
19.1 Introduction 289
19.2 Zeta functions over finite fields 291
19.3 Reformulating the conjecture 293
19.4 Final Form 295

IV Appendices 299

A Some results from algebraic number theory 300

B The Poisson summation formula 304

Bibliography 306

List of Figures

1.1 The real solutions of a cubic polynomial 2
1.2 Arc length of the ellipse 3

6.1 The fundamental domain R^0 for $SL(2, \mathbf{Z})$ acting on $\mathcal{H}$ 48

18.1 The contour $\mathcal{C}$ 283

Part I

Analytic theory of Elliptic Curves

Chapter 1

Doubly periodic functions

1.1 Introduction

The subject of this workshop is "Elliptic Curves". We begin by briefly explaining the reason why elliptic curves are called elliptic.

As we shall learn soon in greater detail, an elliptic curve is by definition a curve of the form

$$y^2 = C(x),$$

where $C(x)$ is a cubic polynomial in x with distinct roots. Even if x, y and the coefficients of C are regarded as real numbers, *e.g.*,

$$y^2 = x^3 - x = x(x-1)(x+1),$$

this curve in $\mathbf{R}^2$ appears to have nothing in common with an ellipse (see figure 1.1). However, the problem of parametrising an ellipse by arc-length leads to elliptic curves, or more specifically the so-called elliptic integrals, as

Figure 1.1: The real solutions of a cubic polynomial

Figure 1.2: Arc length of the ellipse

follows. For the standard ellipse

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

straightforward computation yields

$$+\left(\frac{dy}{dx}\right)^2 = \left(1 - e^2\left(\frac{x}{a}\right)^2\right) / \left(1 - \left(\frac{x}{a}\right)^2\right)$$

(where $e^2 = (a^2 - b^2)/a^2$), so that (see figure 1.2)

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

Observe that, when $a = b = 1$,

$$s(x) = \int_0^x \frac{du}{\sqrt{1-u^2}} = \sin^{-1}x. \qquad (*)$$

Indefinite integrals of the above type, *i.e.*, integrals of the type $\int R(x,y)dx$, where $R(x,y)$ is a rational function of x and y and y^2 is a polynomial of degree three or four in x without multiple roots, are called elliptic integrals. They were studied in great detail in the eighteenth centrury, especially by Legendre. It was recognised quite early that everything had to be done over the complex numbers rather than the real numbers; thus one is in fact dealing with path integrals of rational differentials along paths on the complex curve (or real surface)

$$w^2 = (z - \alpha_1)\cdots(z - \alpha_4)$$

(th α_i being distinct). By a change of variables, it is in fact sufficient to consider the case

$$w^2 = (z - \beta_1)(z - \beta_2)(z - \beta_3)$$

with distinct β's, *i.e.*, when we are working on an elliptic curve as defined earlier.

However, it was only about 1825 that it was realised (by Abel and Jacobi) that, while the elliptic integrals are, by their very definition, multiple valued functions, their inverse functions can give rise to very interesting singlevalued meromorphic functions on the whole complex plane – the so-called elliptic functions. The analogy with the trigonometric (or circular) functions (which arise when the ellipse considered above degenerates to a circle, see $(*)$), or with the case of the logarithmic and exponential functions, is quite clear. The elliptic functions are *doubly periodic*, *i.e.*, they have two periods ω_1 and ω_2 whose ratio is not real.

To be specific, consider the standard elliptic integral in Weierstrass form

$$I(w) = \int_w^\infty (4w^3 - g_2 w - g_3)^{-\frac{1}{2}} dw,$$

where g_2, g_3 are complex numbers with $g_2^3 - 27g_3^2 \neq 0$ — the necessary and sufficient condition for the roots e_1, e_2, e_3 of the equation $4w^3 - gw - g_3$ to be distinct. For a given $w \neq e_i$, the values of $I(w)$ depend on the choice of a branch of the integrand near w and on the path of integration from w to ∞ – we assume that the path does not pass through any of the e_i. The inverse of this multiple-valued function of w will be a doubly periodic meromorphic function as above.

The study of these functions leads to a very rich theory, including a great simplification of the theory of elliptic integrals. Elliptic functions and integrals also appear in many classical problems of analysis and physics – like the motion of the simple pendulum when the amplitude is not assumed to be small; see *e.g.*, [19], Chapters XXXI - XXXIII.

In the complex function-theoretic part of this workshop, we shall (inverting the historical order) start with the study of elliptic functions, and see how they lead naturally to elliptic curves and facilitate the study of elliptic integrals.

1.2 Periodic and elliptic functions

Notation. For any domain (*i.e.*, connected open subset) D of $\mathbf{C}$, $\mathcal{H}(D)$ (resp. $\mathcal{M}(D)$) denotes the set of all holomorphic (resp. meromorphic) functions on D. An $\omega \in \mathbf{C}$ is said to be a period of $f \in \mathcal{M}(\mathbf{C})$ if $f(z+\omega) = f(z)$ for all $z \in \mathbf{C}$.

Proposition 1.2.1 *Let $f \in \mathcal{M}(\mathbf{C})$ be non-constant. Then the set $\Pi(f)$ of all periods of f is a discrete subgroup of the additive group $\mathbf{C}$.*

Remark: Recall that a subset D of a topological space X is said to be discrete if D is closed in X *and* consists entirely of isolated points. If *e.g.*, X is locally compact, D is discrete iff $D \cap K$ is finite for every compact $K \subset X$.

Proof: It is clear that $\Pi(f)$ is a subgroup of $\mathbf{C}$. The discreteness follows from the fact that, for any domain D in $\mathbf{C}$ and any $\varphi \not\equiv 0$ in $\mathcal{M}(D)$, the set of zeros of φ is a discrete subset of D. □

Proposition 1.2.2 *Let $\Gamma \neq \{0\}$ be a discrete subgroup of $\mathbf{C}$. Then Γ is either infinite cyclic, or it is a lattice in $\mathbf{C}$, i.e., a free abelian subgroup of rank two, generated by two complex numbers which are linearly independent over $\mathbf{R}$ (i.e., whose ratio is not real).*

Proof: Let V be the $\mathbf{R}$-subspace of $\mathbf{C}$ generated by Γ. If $\dim V = 1$, then Γ can be regarded as a discrete subgroup of $\mathbf{R}$, and hence is obviously infinite cyclic. If $\dim V = 2$, choose any two $\mathbf{R}$-independent elements α, β of Γ. If Γ' is the subgroup of Γ generated by α and β, it is clear that any $\gamma \in \Gamma$ can be written (uniquely) as

$$\gamma = \gamma' + \gamma''$$

where $\gamma' \in \Gamma'$ and $\gamma''(\in \Gamma)$ is of the form $s\alpha + t\beta$, $0 \leq s, t < 1$. Since Γ is discrete, the set of such $\gamma'' \in \Gamma$ is finite, so that Γ' is of finite index in Γ. Thus Γ is a finitely generated abelian group, and of course it is torsion-free. Hence Γ is clearly free abelian of rank two, and the proposition follows. □

Exercise 1.2.1: If Γ is not cyclic, let $w_1 \in \Gamma$ be of minimal absolute value $\mid w_1 \mid \neq 0$. Let $w_2 \in \Gamma$ be such that $w_2 \notin \mathbf{Z}w_1$, and $\mid w_2 \mid$ has minimal absolute value among elements in Γ not in $\mathbf{Z}w_1$. Then w_1, w_2 generate Γ freely. This is a result of Jacobi. Any subgroup of $\mathbf{C}$ generated by two periods w_1, w_2 which are independent over $\mathbf{R}$ is called a *lattice*.

We shall now fix a discrete subgroup $\Gamma \neq \{0\}$ of the additive group $\mathbf{C}$, and study $\mathcal{H}^\Gamma = \mathcal{H}^\Gamma(\mathbf{C}) := \{f \in \mathcal{H}(\mathbf{C}) : \Pi(f) \supset \Gamma\}$ and $\mathcal{M}^\Gamma$ (which is defined similarly). The interesting case is when Γ is a lattice (*i.e.*, has rank two). Observe that $\mathcal{H}^\Gamma$ is a $\mathbf{C}$-algebra, and $\mathcal{M}^\Gamma$ is an extension field of $\mathbf{C}$, and that both of them are stable under translations; *e.g.*, if $f \in \mathcal{H}^\Gamma$, then $f_a \in \mathcal{H}^\Gamma$ for $a \in \mathbf{C}$, where $f_a(z) := f(a+z)$. Both are also stable under differentiation.

We shall first state the results for the case when Γ is infinite cyclic.

Proposition 1.2.3 *Let Γ be infinite cyclic; let $\omega \in \Gamma$ be a generator. Then*

(i) Any $f \in \mathcal{H}^\Gamma$ has a unique representation

$$f(z) = \sum_{n \in \mathbf{Z}} a_n \exp\left(\frac{2\pi\iota n}{\omega} z\right)$$

where the series converges absolutely, uniformly on compact subsets of $\mathbf{C}$. Conversely, any such series represents an $f \in \mathcal{H}^\Gamma$

(ii) $\mathcal{M}^\Gamma$ is the quotient field of $\mathcal{H}^\Gamma$.

Proof: Consider the holomorphic map $e = e_\omega : \mathbf{C} \to \mathbf{C}^* = \mathbf{C} - \{0\}$ given by $e(z) = \exp(2\pi\iota\frac{z}{w})$. Then $f \in \mathcal{H}^\Gamma$ iff it is constant on each fibre $e^{-1}(a)$ of e. Thus, if $f \in \mathcal{H}^\Gamma$, there exists a function $g : \mathbf{C}^* \to \mathbf{C}$ such that $f = g \circ e$, and this g is holomorphic. Now (i) follows from the Laurent expansion of g. Similarly, if $f \in \mathcal{M}^\Gamma$, there exists $g \in \mathcal{M}(\mathbf{C}^*)$ such that $f = g \circ e$. Now (ii) follows from the well-known fact that $\mathcal{M}(D)$ is the quotient field of $\mathcal{H}(D)$ for any domain $D \subset \mathbf{C}$. □

Remark: Let X be a domain in $\mathbf{C}$, or more generally any Riemann surface, and Γ a properly discontinuous group of holomorphic automorphisms of X. Then the orbit-space $Y = X/\Gamma$ can be made a Riemann surface in such a way that the natural projection $p : X \to Y$ is a holomorphic map, and a function $f : Y \to \mathbf{C}$ is holomorphic iff $f \circ p$ is holomorphic on X. Thus $\mathcal{H}(Y)$ can be identified with the set $\mathcal{H}(X)^\Gamma$ of Γ-invariant holomorphic functions on X. Proposition 1.2.3 says that, when $X = \mathbf{C}$ and Γ is any (infinite) cyclic group of translations, $Y := X/\Gamma$ is always the same Riemann surface, namely $\mathbf{C}^*$. In particular the $\mathbf{C}$-algebras $\mathcal{H}^\Gamma$ and $\mathcal{M}^\Gamma$ do not depend on Γ, upto isomorphism.

Exercise 1.2.2: Formulate and prove Weierstrass' and Mittag-Leffler's theorems for $\mathcal{H}^\Gamma$ and $\mathcal{M}^\Gamma$ respectively ($\Gamma \subset \mathbf{C}$ infinite cyclic). (Observe that the set of poles of any $f \in \mathcal{M}^\Gamma$ is a discrete subset of $\mathbf{C}$ which should be Γ-stable, *i.e.*, should be a union of Γ-cosets.)

We now come to the more interesting case when Γ is a lattice in $\mathbf{C}$. First of all we have

Proposition 1.2.4 *For any lattice* Γ, $\mathcal{H}^\Gamma = \mathbf{C}$.

Proof: We have already observed that any $z \in \mathbf{C}$ can be written as $\gamma + z''$, where $\gamma \in \Gamma$ and z'' has the form $s\omega_1 + t\omega_2, 0 \leq s, t \leq 1$ (here ω_1, ω_2 denote any two $\mathbf{R}$-independent elements of Γ). Thus any $f \in \mathcal{H}^\Gamma$ assumes all its values already on the compact set $\{s\omega_1 + t\omega_2 : 0 \leq s, t \leq 1\}$. Hence $\mid f \mid$ attains its maximum, so that f must be constant. □

Remark: The same argument shows that any $f \in \mathcal{M}^\Gamma$ has only finitely many poles "modulo Γ " in the following sense. For any $f \in \mathcal{M}^\Gamma$, its set of poles $(f)_\infty$ is Γ-stable, *i.e.*, $z_0 \in (f)_\infty$ implies $z_0 + \gamma \in (f)_\infty$ for all $\gamma \in \Gamma$. Thus, $(f)_\infty$ is always infinite (unless f is constant). However, $(f)_\infty$ is a *finite* union of cosets of Γ in $\mathbf{C}$. This is because $(f)_\infty$ is discrete, *i.e.*, intersects any compact subset of $\mathbf{C}$ in only finitely many points, and we have already seen that every coset of Γ meets the compact set $\{s\omega_1 + t\omega_2 : 0 \leq s, t \leq 1\}$ mentioned in the proof of Proposition 1.2.2. The same remark applies to the zeros of any $f(\not\equiv 0)$ in $\mathcal{M}^\Gamma$.

However, non-constant elements of $\mathcal{M}^\Gamma$ can be constructed by the obvious method of constructing functions invariant under a group action : start with an arbitrary function and form the sum (or product) of all its translates under the group. Since our group Γ is infinite, there will in general by convergence problems: *cf.* the case of $\mathcal{H}^\Gamma$! However, we have:

Proposition 1.2.5 *The series*

$$\sum_{\gamma \in \Gamma} (z - \gamma)^{-3}$$

converges absolutely on $\mathbf{C} - \Gamma$, *uniformly on compact subsets of* $\mathbf{C} - \Gamma$. *The sum represents an element* f_3 *of* $\mathcal{M}^\Gamma$ *whose only poles are triple poles at points of* Γ.

Proof: The convergence statement clearly follows if we know that the series

$$\sum_{\gamma \in \Gamma^*} \mid \gamma \mid^{-3} < \infty, \text{ where}$$

$$\Gamma^* = \Gamma - \{0\};$$

the rest of the proposition is then obvious. For the convergence, choose generators ω_1, ω_2 of Γ. Since ω_1 and ω_2 are linearly independent over $\mathbf{R}$, there exists a constant $C > 0$ such that $\mid s\omega_1 + t\omega_2 \mid^2 \geq C(s^2 + t^2)$ for all $s, t \in \mathbf{R}$. Thus we may apply the following lemma □

Lemma 1.2.6

$$\sum_{m,n \in \mathbf{Z}}{}' (m^2 + n^2)^{-\alpha} < \infty \text{ if } \alpha > 1.$$

(Here $\sum'$ means that the term $m = n = 0$ is to be omitted.)

Proof: The details are left as an exercise. The fact to observe is that there exists a constant $C > 0$ such that,

$$\#\{(m, n) \in \mathbf{Z} \times \mathbf{Z} : p^2 \leq m^2 + n^2 < (p+1)^2\} \leq Cp$$

for all (sufficiently large) $p > 0$. (A similar argument also shows that $\sum'(m^2 + n^2)^{-1} = \infty$.) □

Clearly, the Laurent expansion of the meromorphic function f_3 constructed above has only one negative term (of order three) around each of its poles, hence f_3 has a primitive, which can be obtained by termwise integration of the series. For historical and other reasons, we consider the function $\wp$ defined by

$$\wp(z) = \frac{1}{z^2} + \sum_{\gamma \in \Gamma^*} \left(\frac{1}{(z-\gamma)^2} - \frac{1}{\gamma^2} \right)$$

$$= -2 \left(\int_\infty^z \frac{dz}{z^3} + \sum_{\gamma \in \Gamma^*} \int_0^z \frac{dz}{(z-\gamma)^3} \right)$$

which is thus a primitive of $-2f_3$. Observe that $\wp$ is an even function : $\wp(-z) \equiv \wp(z)$, and $\wp'$ is odd.

Proposition 1.2.7 $\wp \in \mathcal{M}^{\Gamma}$.

Proof: Since $\wp' \in \mathcal{M}^{\Gamma}$, it is clear that, for any $\gamma \in \Gamma$,

$$\wp(z+\gamma) - \wp(z) = c_{\gamma}$$

is a constant (independent of z). But putting $\gamma = \omega_i, z = -\omega_i/2$, we see that $c_{\omega_i} = 0$, since $\wp$ is an even function. Hence ω_1 and ω_2 are periods of $\wp$. $\square$

Thus $\wp \in \mathcal{M}^{\Gamma}$ has double poles precisely at the $\gamma \in \Gamma$, and the residues at the poles are still zero. Hence $\wp$ has again a primitive, which can be got by termwise integration. Again, for historical reasons, we introduce the primitive of $-\wp$, defined by

$$\zeta(z) = \frac{1}{z} + \sum_{\gamma \in \Gamma^*} \left(\frac{1}{z-\gamma} + \frac{1}{\gamma} + \frac{z}{\gamma^2} \right).$$

As in the case of $\wp, \zeta$ has the property

$$\zeta(z+\gamma) - \zeta(z) = \eta_{\gamma}$$

where $\eta_{\gamma} \in \mathbf{C}$ is a constant independent of z. However, as we shall soon see, $\zeta \notin \mathcal{M}^{\Gamma}$. Observe that ζ is odd. Since ζ has simple poles at the $\gamma \in \Gamma$ with residue one, the integral of ζ along any closed curve in $\mathbf{C}$ avoiding Γ is an integral multiple of $2\pi i$. Hence it is clear that

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

(z_0 a fixed point and z a variable point of $\mathbf{C} - \Gamma, C \neq 0$ a constant) is a nowhere-vanishing holomorphic function in $\mathbf{C} - \Gamma$, which extends to an entire function with simple zeros at the $\gamma \in \Gamma$. We normalise σ by choosing C such that $\sigma'(0) = 1$. We have $\zeta = \frac{\sigma'}{\sigma}$ by construction.

It is not hard to see that σ is again an odd function. Also, by the definition of σ, for any $\gamma \in \Gamma$,

$$\sigma(z+\gamma) = \sigma(z) \exp \left(\int_{z}^{z+\gamma} \zeta(t) dt \right).$$

But the derivative of $\int_z^{z+\gamma} \zeta(t)dt$ is obviously $\zeta(z+\gamma) - \zeta(z) = \eta_\gamma$, *i.e.*,

$$\int_z^{z+\gamma} \zeta(t)dt = \eta_\gamma z + c_\gamma$$

with c_γ a constant depending on γ. Thus

$$\sigma(z+\gamma) = \sigma(z)\ \exp\ (\eta_\gamma z + c_\gamma)$$

or, by a change of notation

$$\sigma(z+\gamma) = \sigma(z)\ \exp\ \left(\eta_\gamma(z+\frac{\gamma}{2}) + d_\gamma\right).$$

We can now determine the $\exp(d_\gamma)$: put $z = -\gamma/2$ above to get

$$\sigma(-\gamma/2) = \sigma(\gamma/2)\ \exp(d_\gamma).$$

Since σ has only simple zeros at the $\gamma \in \Gamma$, and is odd (so that σ' is even), we get $\exp(d_\gamma) = -1$ if $\gamma/2 \notin \Gamma$, and $\exp(d_\gamma) = 1$ if $\gamma/2 \in \Gamma$.

We can now prove the fact mentioned earlier that $\zeta \notin \mathcal{M}^\Gamma$. More generally, we have

Proposition 1.2.8 *There does not exist any $f \in \l M^\Gamma$ with only simple poles at the $\gamma \in \Gamma$..*

Proof: We first show that $\zeta \notin \mathcal{M}^\Gamma$. This is almost obvious: $\zeta \in \mathcal{M}^\Gamma$ iff all the η_γ vanish. But if all the η_γ vanish, it is clear that $\sigma^2 \in \mathcal{H}^\Gamma$ *i.e.*, $\sigma^2 \equiv 0$, which is absurd.

The above argument in fact shows that there is no odd $f \in \mathcal{M}^\Gamma$ with simple poles (of residue one) at the $\gamma \in \Gamma$. Observe now that $f(z) \in \mathcal{M}^\Gamma$ implies $f(-z) \in \mathcal{M}^\Gamma$, so that any $f \in \mathcal{M}^\Gamma$ is the sum of an even and an odd element of $\mathcal{M}^\Gamma$. Hence if $f \in \mathcal{M}^\Gamma$ has simple poles at the $\gamma \in \mathcal{M}^\Gamma$, its even part $\frac{1}{2}\{f(z) + f(-z)\}$ must be constant (since it has no poles), and its odd part must also vanish by the above argument. □

However, we have

Lemma 1.2.9 *Given $z_0 \notin \Gamma$, there exists $f \in \mathcal{M}^\Gamma$ whose only poles are simple poles at 0 and z_0 (and their Γ-translates), with residue $+1$ at 0 and -1 at z_0; such an f is unique upto an additive constant.*

Proof: Take $f = \zeta - \zeta_{-z_0}$ where ζ_{-z_0} is the $(-z_0)$-translate of $\zeta : \zeta_{-z_0}(z) = \zeta(z - z_0)$. □

Exercise 1.2.3: Suppose given finitely many $z_i \in \mathbf{C}$, and integers $k_i \geq 1$. Consider the vector-space $V := \{f \in \mathcal{M}^\Gamma$: the only poles of f are poles of order $\leq k_i$ at the z_i (and their Γ-translates)$\}$. Then $\dim V \leq 1 + \sum k_i$.

Theorem 1.2.1 *With the notation of Exercise 1.2.3, assume further that the z_i are incongruent mod Γ (i.e., $z_i - z_j \notin \Gamma$ if $i \neq j$). Then:*

(i) *dim $V = \sum k_i$.*

(ii) *given polynomials $p_i(T) = \sum_{j=1}^{k_i} a_{i,j}T^i$, there exists $f \in \mathcal{M}^\Gamma$ whose only poles are at the z_i (and their Γ-translates), with principal part $p_i(\frac{1}{z - z_i})$ at z_i, iff $\sum_i a_{i,1} = 0$.*

Proof: Clearly it is enough to prove (ii). By replacing the z_i by $z_i - z_1$, we may assume $z_1 = 0$. Suppose first that only one z_i given. If $k = 1$, it is clear that $V = \mathbf{C}$. If $k > 1$, it is clear that $1, \wp, \wp', \cdots, \wp^{(k-1)}$ (= derivative of $\wp$ of order $k - 1$) form a basis for V, and can be used to construct an $f \in \mathcal{M}^\Gamma$ with principal part p_1.

The general case is proved by induction. If $n(\geq 2)$ points z_i are given, then using $1, \varphi, \varphi', \cdots, \varphi^{(k_n - 1)}$ and $\zeta - \zeta_{-z_n}$, (where $\varphi = \wp_{-z_n}$), we can construct $f_1 \in \mathcal{M}^\Gamma$ which has the prescribed principal part at z_n, a simple pole of residue $-a_{n,1}$ at $z_1 = 0$ and no other singularities. Thus we are reduced to the case of $(n - 1)$ prescribed points, with $p_1(T)$ replaced by $p_1(T) + a_{n,1}T$. □

Exercise 1.2.4: Choose any $f, g \in \mathcal{M}^\Gamma$, and any integer $N \geq 1$. Let W be the vector space $\{P(S,T) \in \mathbf{C}[S,T] : \deg_S P, \deg_T P \leq N\}$. Use Exercise 1.2.3 to show that the dimension of the vector space $\{P(f,g) : P \in W\} (\subset \mathcal{M}^\Gamma)$ grows at the most linearly in N. Deduce that f and g are algebraically dependent, *i.e.*, there exists a non-zero $P \in \mathbf{C}[S,T]$ such that $P(f,g) \equiv 0$.

Exercise 1.2.5: Carry out the procedure we used for constructing f_3, $\wp$, ζ, σ in the case when Γ is infinite cyclic, $\Gamma = \mathbf{Z} \subset \mathbf{C}$ say. Observe that we can now start with $f_2 = \sum_{n \in \mathbf{Z}} (z - n)^{-2}$ since $\sum \frac{1}{n^2} < \infty$.

(In fact, $f_2 = \pi^2 \operatorname{cosec}^2 \pi z$ (see [2], Chapter 5, §2); its primitive is $-\pi \cot \pi z$ which is actually periodic. Logarithmic integration now yields $\sin \pi z$.)

We conclude this section by showing how the function σ can be used to solve the Weierstrass problem (construction of a meromorphic function with prescribed zeros and poles) inside $\mathcal{M}^\Gamma$, just as ζ was used in Theorem 1.2.1 to solve the Mittag-Leffler problem (construction of meromorphic functions with prescribed principal parts).

Theorem 1.2.2 *Suppose two finite sets of points $\{z_1, \cdots, z_r\}$ and $\{p_1, \cdots, p_s\}$, and integers $k_i, \ell_j \geq 1$, $1 \leq i \leq r$, $1 \leq j \leq s$ are given. Suppose all the z_i, p_j are mutually incongruent mod Γ. Then there exists an $f \in \mathcal{M}^\Gamma$ with zeros of order k_i at the z_i and poles of order ℓ_j at the p_j and no other zeros or poles (except for Γ-translates) if and only if $\sum k_i = \sum \ell_j$ and $\sum k_i z_i - \sum \ell_j p_j \in \Gamma$.*

Proof: The sufficiency of the stated conditions is almost obvious. Let $\sum k_i z_i - \sum \ell_j p_j = \gamma_0$. First assume that $\gamma_0 = 0$. Then it is clear that $f \in \mathcal{M}^{\mathbf{C}}$ defined by

$$f(z) = \{\prod_i \sigma(z - z_i)^{k_i}\} / \{\prod_j \sigma(z - p_j)^{\ell_j}\}$$

has the right zeros and poles. Since, for any $\alpha \in \mathbf{C}$ and any $\gamma \in \Gamma$,

$$\sigma(z + \alpha + \gamma) = \sigma(z + \alpha).\ \exp\ \{\eta_\gamma (z + \gamma/2) + \eta_\gamma \alpha + d_\gamma\}$$

it is also clear that $f(z+\gamma) \equiv f(z)$ for all $\gamma \in \Gamma$, and the sufficiency is proved in this case. If $\gamma_0 \neq 0$, we need only replace the factor $\sigma(z - p_s)^{\ell_s}$ in the definition of f by $\sigma(z - p_s)^{\ell_s - 1} \sigma(z - (p_s + \gamma_0))$.

The necessity follows from the sufficiency. Suppose indeed that the z_i and p_j are a complete set of zeros and poles mod Γ of an $f \in \mathcal{M}^\Gamma$, of orders k_i and ℓ_j respectively. If *e.g.*, $\sum k_i - \sum \ell_j = d > 0$, we can set $p_{s+1} = \frac{1}{d}(\sum k_i z_i - \sum \ell_j p_j)$ and $\ell_{s+1} = d$. Then the sufficiency part of the theorem yields a $g \in \mathcal{M}^\Gamma$ with zeros of order k_i at the z_i and poles of order ℓ_j at the p_j. Thus, $f/g \in \mathcal{M}^\Gamma$ has no poles at all, which is impossible. If $\sum k_i = \sum \ell_j$ and $\sum k_i z_i - \sum \ell_j p_j =: \alpha \notin \Gamma$, we again form $g \in \mathcal{M}^\Gamma$ by setting

$$g(z) = \{\prod_i \sigma(z - z_i)^{k_i}\} / \prod_{j<s} \sigma(z - p_j)^{\ell_j} \sigma(z - p_s)^{\ell_s - 1} \sigma(z - p_s - \alpha).$$

Then g/f has just one simple pole (at α) – a contradiction. □

Corollary 1.2.10 *Every non-constant $f \in \mathcal{M}^\Gamma$ takes on each value $\alpha \in \mathbf{C} \cup \{\infty\}$ the same number of times (mod Γ, multiplicites being taken into account); in particular it takes each value at least twice.*

Proof: For the values $\alpha = 0$ and $\alpha = \infty$, this is just a restatement of the condition $\sum k_i = \sum \ell_j$ in Theorem 1.2.2. The general case follows by considering $f - \alpha$ instead of f. □

Remark: In Theorem 1.2.1, we have seen that the sum of the residues of any $f \in \mathcal{H}^\Gamma$ at a complete set of Γ-inequivalent poles is zero. Applying this to f'/f, one can deduce Corollary 1.2.10. We shall soon see that there are various other (more natural) ways of proving these two statements. The point of Theorems 1.2.1 and 1.2.2 is the sufficiency of the conditions stated there. In fact, addition mod Γ (*i.e.*, addition in the quotient group $\mathbf{C}/\Gamma$) can be characterised in terms of $\mathcal{M}^\Gamma$ as follows. Given $z_1, z_2 \in \mathbf{C}$ (assumed distinct mod Γ for convenience), the point $z_3 = -(z_1+z_2) \mod \Gamma$ is clearly determined as follows: there is a unique linear combination $f := \wp' + a\wp + b \in \mathcal{M}^\Gamma$ with zeros at z_1 and z_2, and then z_3 is the third zero of f. We shall study these matters in greater detail in Chapter 3.

Exercise 1.2.6: (i) Show that every $f \in \mathcal{M}^\Gamma$ whose only pole (mod Γ) is one of order ≤ 2 at the origin is of the form $a\wp + b, a, b \in \mathbf{C}$; in particular, f is even.

(ii) Show that any even $f \in \mathcal{M}^\Gamma$ whose only pole (mod Γ) is one of order (say) $2k$ at the origin is a polynomial of degree k in $\wp$. Deduce that $(\wp')^2$ is a cubic in $\wp$ with leading coefficient 4.

(iii) Observe that, near $z = 0$, $\wp(z) = z^{-2} + h(z)$ where $h(z)$ is holomorphic and $h(0) = h'(0) = 0$, and deduce that $(\wp')^2 - 4\wp^3$ has a pole of order ≤ 2 at the origin. Hence the cubic of (ii) has no quadratic term.

By Exercise 1.2.6 (ii), $\wp$ and $\wp'$ are related by an equation of the form

$$(\wp')^2 = 4\wp^3 - g_2\wp - g_3 \tag{1.1}$$

where $g_2, g_3 \in \mathbf{C}$ are constants determined by Γ. We shall show that the cubic occurring on the right above, namely $4T^3 - g_2T - g_3$, has in fact distinct roots,

and hence the connection with elliptic curves is close at hand. But at this stage it is better to give up working on $\mathbf{C}$ alone, and to regard elements of $\mathcal{M}^\Gamma$ as functions on the orbit space of Γ, namely the quotient group $\mathbf{C}/\Gamma$, and this is what we proceed to do.

Chapter 2

Riemann surfaces

2.1 Generalities on Riemann surfaces

We begin with the definition of a Riemann surface.

Definition: Let X be a Hausdorff topological space. Let $\{U_\alpha\}$ be an open covering of X which satisfies the following conditions:

(1) For each α, there exists a homeomorphism $f_\alpha : U_\alpha \to V_\alpha$, where V_α is an open subset of $\mathbf{C}$

(2) If $U_\alpha \cap U_\beta \neq \phi$ for some α, β, then the map $f_\beta \circ f_\alpha^{-1} : f_\alpha(U_\alpha \cap U_\beta) \to f_\beta(U_\alpha \cap U_\beta)$ (which makes sense since each f_α is a homeomorphism) is a complex analytic map.

(3) The collection $\{(U_\alpha, f_\alpha)\}$ is maximal with respect to the properties (1) and (2) (here the collections are partially ordered by $\{(V_\beta, g_\beta)\} < \{(U_\alpha, f_\alpha)\}$ if each (V_β, g_β) equals some (U_α, f_α)).

Denoting by $\mathcal{C}$ the collection $\{(U_\alpha, f_\alpha)\}$, the pair $(X, \mathcal{C})$ is called a *complex manifold of dimension 1*. If further X is second countable, then X is called a *Riemann surface*. The pair (U_α, f_α) is called a holomorphic coordinate system in X, and f_α is called a *local holomorphic coordinate*. Given a collection $\{(V_\lambda, g_\lambda)\}$ of X satisfying only the conditions (1) and (2), where $\{V_\lambda\}$ is an open covering of X, we can construct a collection $\{(U_\alpha, f_\alpha)\}$ of X satisfying (1), (2) and (3) such that each (V_λ, g_λ) equals some (U_α, f_α), by a suitable application of Zorn's lemma; this maximal collection of pairs

is in fact unique. In practice, when we deal with a specific Riemann surface X, we will be considering suitable open coverings of X giving collections of pairs satisfying (1) and (2) only. The condition (3) is added for theoretical reasons[1].

Let $(X, \mathcal{C})$ and $(Y, \mathcal{C}')$ be Riemann surfaces and $\varphi : X \to Y$ a continuous map. We say that φ is a complex analytic (or holomorphic) map if the following condition is satisfied:

Let $x \in X$ be any point. Then there exist $(U_\alpha, f_\alpha) \in \mathcal{C}$ and $(V_\beta, g_\beta) \in \mathcal{C}'$ such that $x \in U_\alpha$, $\varphi(U_\alpha) \subset V_\beta$ such that the map $g_\beta \circ \varphi \circ f_\alpha^{-1} : f_\alpha(U) \to g_\beta(V_\beta)$ is complex analytic.

If $\varphi : (X, \mathcal{C}) \to (Y, \mathcal{C}')$ is a holomorphic map between Riemann surfaces, we say that φ is a biholomorphic map if there is a holomorphic map $\psi : (Y, \mathcal{C}') \to (X, \mathcal{C})$ such that $\psi \circ \varphi$ and $\varphi \circ \psi$ are both identity maps.

It is easy to see that φ is biholomorphic iff φ is a homeomorphism onto Y (in addition to it being a holomorphic map).

2.2 Examples

1. Let $U \subset \mathbf{C}$ be any non-empty open subset of $\mathbf{C}$. It is clear how to make U into a Riemann surface.

2. Let $X = \mathbf{C} \cup \{\infty\}$ be the 1-point compactification of $\mathbf{C}$.

   Let $U_1 = \mathbf{C}$ and $U_2 = X - \{0\}$. Define $f_1 : U_1 \to \mathbf{C}$ to be the identity map and $f_2 : U_2 \to \mathbf{C}$ as follows.

   For any $z \in U_2 - \{\infty\}$, let $f_2(z) = \frac{1}{z}$ and $f_2(\infty) = 0$.

   Clearly both f_1, f_2 are homeomorphisms. Now the map $f_2 \circ f_1^{-1} : \mathbf{C} - \{0\} \to \mathbf{C} - \{0\}$ is the map $z \to \frac{1}{z}$ which is obviously holomorphic.

   The two co-ordinate systems (U_1, f_1) and (U_2, f_2) then define a Riemann surface structure on X. Observe that X is homeomorphic to the sphere S^2 using Stereographic projection. This Riemann surface is called the complex projective space of dimension 1 and denoted by $\mathbf{CP}^1$ (or $\mathbf{P}^1_{\mathbf{C}}$).

[1]There may be many collections of pairs satisfying (1) and (2) determined by open coverings of X, which all give rise to the 'same' Riemann surface structure; if we impose the additional condition (3), then the collection is unique

3. Let X be a Riemann surface and G a group of holomorphic automorphisms of X, the action being *properly discontinuous i.e.*, for any compact sets $K_1, K_2 \subset X$, the set $\{g \in G \mid g(K_1) \cap K_2 \neq \phi\}$ is finite.

Now we will describe the important construction of making the topological space X/G, in the above example, into a Riemann surface such that the quotient map $X \to X/G$ is holomorphic. We reduce easily to the case when X is connected, and G is a group of holomorphic automorphisms of X.

First recall how X/G is made into a topological space. By definition, two points x_1, x_2 in X give the same point in X/G if $x_2 = g(x_1)$ for some $g \in G$. The set X/G is then the set of equivalence classes of points in X (each equivalence class is called an *orbit* of G in X).

We have a set theoretic map $\varphi : X \to X/G$ sending x to its orbit. Now a subset $U \subset X/G$ is open if $\varphi^{-1}(U)$ is open. This gives the quotient topology on X/G. Using the assumption that G acts properly discontinuously, we see that X/G is a Hausdorff topological space. For details, see [62], Chapter V, Prop.5.3.

In contrast to the above situation, the exercise below gives an example of a 'bad' action of $\mathbf{Z}$, acting via linear transformations.

Exercise 2.2.1: Let $X = \mathbf{C}^2 - \{(0,0)\}$, and let $G = \mathbf{Z}$ act on X by $[n] \cdot (x,y) = (2^n x, 2^{-n} y)$.

(i) Show that for any $P \in X$, the orbit $\{g(P) \mid g \in G\} \subset X$ is closed.

(ii) Show that for any $P \in X$, the stabiliser $G_P = \{g \in G \mid g(P) = P\}$ is $\{0\}$.

(iii) Show that there exist compact sets $K_1, K_2 \subset X$ such that

$$\{x \in G \mid g(K_1) \cap K_2 \neq \phi\}$$

is infinite.

(iv) Show that the quotient space X/G, with the quotient topology, is not Hausdorff.

We now return to our construction of X/G. Let $x \in X$ and $G_x = \{g \in G \mid g(x) = x\}$. Then G_x is a subgroup of G (called the *isotropy subgroup* of x). Since G acts properly discontinuously, G_x is finite.

Lemma 2.2.1 *The set of points $x \in X$ such that G_x is the trivial group is open and its complement is a discrete set in X.*

Proof: Let $x \in X$ be such that $G_x = \{e\}$. Let U be a small compact neighborhood of x such that the set $\{g \in G \mid gU \cap U \neq \phi\}$ is finite, say $\{g_1, \cdots, g_m\}$. Then $g_i(x) \neq x$ if $g_i \neq e$. We can choose a compact neighborhood $U_1 \subset U$, small enough so that $g_i(x) \notin U_1$ if $g_i \neq 1$. Then by further shrinking U_1 we can find $U_2 \subset U_1$ such that $g_i(U_2) \cap g_j(U_2) = \phi$ if $g_i \neq g_j$.

Now we claim that this neighborhood U_2 has the property that for $g \neq e, g(U_2) \cap U_2 = \phi$. For, if $g(U_2) \cap U_2 \neq \phi, g = g_i$ for some i and this is not possible.

It follows that for any $x' \in U_2, G_{x'} = \{e\}$ and the first part of the Lemma is proved.

Now suppose x_n is a sequence of elements in X such that $x_n \to x_0$ and each G_{x_n} is a non-trivial group. Let U be a neighborhood of x_0 such that $\{g \in G \mid gU \cap U \neq \phi\} = \{1, g_1, \cdots, g_m\}$ is a finite set. We can suppose that $x_n \in U$ for all n. Then clearly $G_{x_n} \subset \{1, g_1, \cdots, g_m\}$ for each n. If for some $g_i, g_i(x_0) \neq x_0$, then by using an argument similar to that used in the first part above, we can shrink U so that $g_i(U) \cap U = \phi$. Thus we can assume that U is so small that $\{1, g_1, \cdots, g_m\} = G_{x_0}$.

Now we have shown that $G_{x_n} \subset G_{x_0}$ for all n. Now consider the neighborhood $U \cap g_1(U) \cap \cdots \cap g_m(U) = V$, say, of x_0.

For any $g_i, g_i(V) = V$.

We can assume that we have a holomorphic co-ordinate system (V, f) in X such that $f(x_0) = 0 \in f(V) \subset \mathbf{C}$. Via f, elements of $G_{x_0} = H$ give holomorphic automorphisms of a neighborhood W of $0 \in \mathbf{C}$ such that W is invariant under H. If $h \in H$ induces the identity automorphism of W, then since X is connected, we see that $h = e$, the identity element, by analytic continuation. Hence we may consider H as a group of holomorphic automorphisms of W.

We have the obvious holomorphic co-ordinate z on W. Since $h_i(0) = 0$ for all $h_i \in H$, we see that as a holomorphic function $W \to W$, h_i has the expansion

$$h_i(z) = a_i z + b_i z^2 + c_i z^3 + \cdots$$

Here $a_i = h_i'(0) \neq 0$ because h_i is an automorphism. Observe that $h_j \circ h_i(z) = a_i a_j z +$ (higher degree terms).

For any holomorphic function $g : W \to W$ and $h \in H$ we can define $\rho_h(g) : W \to W$ by the formula $\rho_h(g)(z) = g(h^{-1}(z))$. Then one checks easily that $\rho_{h_i h_j}(g) = \rho_{h_i}(\rho_{h_j}(g))$ for any $h_i, h_j \in H$ *i.e.*, H acts (on the left) on the set of holomorphic functions $W \to W$. Clearly, if g_1, g_2 are holomorphic functions from $W \to W$, by restricting to a smaller open set $W' \subset W$, we can define $g_1 + g_2 : W' \to W$.

Since H is a finite group, $\underbrace{h_i \circ h_i \circ \cdots \circ h_i}_{N}$ is the identity map on W, where $N =$ order of H. It follows that $a_i^N = 1$ for each $h_i \in H$. Now consider the function

$$g = \frac{1}{N}(\sum_{h_i \in H} \frac{h_i(z)}{a_i}).$$

This is a holomorphic function from a neighborhood W_1 of 0 to W. We can choose W_1 invariant under all h_i (by using the argument for the construction of W).

We now calculate $\rho_{h_j}(g)$ for any $h_j \in H$.

$$\rho_{h_j}(g) = \frac{1}{N} \sum_{h_i \in H} \frac{\rho_{h_j}(h_i(z))}{a_i}$$

$$= \frac{1}{N} \sum_{h_i \in H} \frac{h_i(h_j^{-1}(z))}{a_i}.$$

Using $h_i \circ h_j^{-1}(z) = (a_i/a_j)z +$ (higher degree terms) and the fact that the set $\{h_i \circ h_j^{-1}\}_i = \{h_i\}_i$, we see that

$$\rho_{h_j}(g) = \frac{1}{a_j} g \text{ for each } h_j \in H.$$

Finally observe that the Taylor series for g looks like

$$g = z + \text{ (higher degree terms)}.$$

Then in a neighborhod $W_2 \subset W_1$ such that W_2 is also invariant under H, g can also be chosen as a local holomorphic co-ordinate such that the action of H is $h_i(g) = \lambda_i g$, where λ_i is a root of unit depending on h_i.

We have thus proved the following important result.

Lemma 2.2.2 *Let G be a properly discontinuous group of holomorphic automorphisms of a connected Riemann surface X and G_x the isotropy group of a point $x \in X$ (necessarily finite). Then G_x is cyclic and we can choose a co-ordinate system (W, f) such that*

(i) $x \in W$ and W is invariant under G_x

(ii) $f(x) = 0$ and via f, G_x acts as a group of automorphisms of a neighborhood of $0 \in \mathbf{C}$ such that $h(z) = \lambda_h z$ for $h \in G_x$ and λ_h is a root of unity depending on h.

(iii) the mapping $G_x \to \mathbf{C}^*$ given by $h \mapsto \lambda_h$ is an injective homomorphism.

We now go back to the proof of part 2 of Lemma 2.2.1.

It is now clear from the preceding arguments that for all points x_n sufficiently close to x_0, the isotropy group G_{x_n} is trivial which is a contradiction.

Thus Lemma 1 is completely proved.

We are now ready to make X/G into a Riemann surface.

Let S be the discrete set of points in X whose isotropy group is non-trivial. Write $Y = X - S$. Then $X - S$ is stable under G and $(X - S)/G$ is an open subset of X/G. For any $x \in (X - S)$, we can find an open neighborhood U containing x such that $U \cap gU = \phi$ for all $g \neq e$. Then the quotient map $\varphi : X \to X/G$ is a homeomorphism when restricted to U. The set of open subsets $\varphi(U)$ then give holomorphic co-ordinate systems on $(X-S)/G$ in an obvious manner. It is also clear that when $(X-S)/G$ is made a Riemann surface in this manner, the quotient map $(X - S) \to (X - S)/G$ is holomorphic (tautology !).

Suppose next $x \in S$ and G_x the isotropy group of x.

By Lemma 2.2.2, there exists a holomorphic co-ordinate system (W, f) satisfying the properties stated in Lemma 2.2.2. For simplicity, assume W is itself a G_x stable neighborhood of $0 \in \mathbf{C}$ such that $g(z) = \lambda_g z$ for any $g \in G_x$. It is enough to make W/G_x into a Riemann surface. G_x is a cyclic group of order N. Assume W is the unit disc in $\mathbf{C}$. Then W/G_x is homeomorphic to the unit disc by the map $\psi : W/G_x \to \mathbf{C}$ given by $\psi(\overline{y}) = y^N$ where $\overline{y}$ is the coset of a point $y \in W$. In view of the nature of the action of G_x on W, this is clearly well defined.

We take $(W/G_x, \psi)$ to be a holomorphic coordinate system in X/G. Together with the coordinate systems on $(X - S)/G$ constructed earlier, we

get a Riemann surface structure on X/G. We leave the verification of the condition (2) in the definition of a Riemann surface to the reader.

It is also easy to see that a function $g : X/G \to \mathbf{C}$ is holomorphic iff $g \circ \varphi : X \to \mathbf{C}$ is holomorphic. More generally, one has the following.

Exercise 2.2.2: Let $f : X \to Y$ be a surjective holomorphic map between Riemann surfaces, and $g : Y \to Z$ a mapping to another Riemann surface such that $g \circ f : X \to Z$ is holomorphic. Show that g is holomorphic.

As an important example of this construction, let X equal $\mathcal{H}$, the upper half plane of complex numbers z with $\Im z > 0$ (here $\Im z$ is the imaginary part of z). The group $G = SL(2, \mathbf{Z})$ of 2×2 matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $ad-bc = 1$ acts on X as follows. Let $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbf{Z})$ and $z \in X$ any point. Then the point $(az+b)/(cz+d)$ also lies in X and defines an action of G on X, such that $-I$ acts trivially (where I is the identity matrix). We will show later that the action of $PSL(2, \mathbf{Z}) = SL(2, \mathbf{Z})/\{\pm I\}$ is properly discontinuous and the quotient Riemann surface $X/(G/\{\pm I\})$ is biholomorphic to $\mathbf{C}$.

2.3 Holomorphic 1-forms

First let $U \subset \mathbf{C}$ be any open set. We have the natural holomorphic coordinate z on U. An expression of the form $f(z)dz$, where $f(z)$ is a holomorphic function in U, is called a holomorphic 1-form on U.

Suppose $\varphi : U \to V$ is a holomorphic function from an open set $U \subset \mathbf{C}$ to an open set $V \subset \mathbf{C}$. Let $\omega = f(z)dz$ be a holomorphic 1-form on V. We define a holomorphic 1-form $\varphi^*(\omega)$ on U as follows:

$$\varphi^*(\omega) = f(\varphi(z))\frac{d\varphi}{dz}dz.$$

It is now easy to extend these definitions to arbitrary Riemann surfaces. Let X be any Riemann surface and $(U_\alpha, \varphi_\alpha)$ a system of holomorphic coordinates for X with $\cup U_\alpha = X$. Suppose for each $\alpha, \exists$ a holomorphic 1-form ω_α on $\varphi_\alpha(U)$ such that the following compatibility holds: for $U_\alpha \cap U_\beta \neq \phi$, consider the holomorphic map $\varphi_\beta \circ \varphi_\alpha^{-1} : \varphi_\alpha(U_\alpha \cap U_\beta) \to \varphi_\beta(U_\alpha \cap U_\beta)$. Then $\omega_\alpha \mid_{\varphi_\alpha(U_\alpha \cap U_\beta)} = (\varphi_\beta \circ \varphi_\alpha^{-1})^*(\omega_\beta \mid_{\varphi_\beta(U_\alpha \cap U_\beta)})$. The collection of 1-forms ω_α give, by definition, a holomorphic 1-form ω on X.

If $f : X \to Y$ is a holomorphic map between Riemann surfaces and ω a holomorphic 1-form on Y, it is easy to define $f^*(\omega)$ using local co-ordinate system on X and Y. We leave the details to the reader.

2.4 Integration of 1-forms

Let again $U \subset \mathbf{C}$ be an open subset and $\omega = f(z)dz$ a holomorphic 1-form on U. Suppose $\gamma : [0,1] \to U$ is a piecewise-differentiable path in U. Then $\int_\gamma f(z)dz$ is defined as follows. Let $0 = t_0 < t_1 < \cdots < t_n = 1$ be a sub-division. The sums

$$\sum_{i=1}^{n} f(z(t_i))(z(t_i) - z(t_{i-1}))$$

approach a definite limit as the subdivision becomes finer and finer. This limit is denoted by $\int_\gamma \omega$. Let now $\varphi : U \to V$ be a biholomorphic map between open subsets of $\mathbf{C}$ and ω a holomorphic 1-form on V. Suppose $\gamma : [0,1] \to U$ is a piecewise differentiable path. Then $\varphi \circ \gamma$ is a piecewise differentiable path in V. We see easily that

$$\int_{\varphi\circ\gamma} \omega = \int \varphi^*(\omega). \qquad (*).$$

Let X be an arbitrary Riemann surface and ω a holomorphic 1-form on X. Let $\gamma : [0,1] \to X$ be a piecewise differentiable path. If $(U_\alpha, \varphi_\alpha)$ give a co-ordinate system on X, we can find $\alpha_1, \cdots, \alpha_r$ and a subdivision $0 = t_0 < t_1 < \cdots < t_r = 1$ of [0,1] such that $\gamma([t_{i-1}, t_i]) \subset U_{\alpha_i}$. Then we define

$$\int_\gamma \omega = \sum_{i=1}^{r} \int_{\gamma_i} \omega \mid_{U_{\alpha_i}},$$

where $\gamma : [t_{i-1}, t_i] \to U_{\alpha_i}$ is denoted by γ_i. Using (*), we check that this is well-defined.

Chapter 3

Tori

3.1 Complex tori; Embedding in $\mathbf{P}^2$

We fix a lattice Γ in $\mathbf{C}$. For certain purposes, we need also to fix a pair of generators ω_1, ω_2 for Γ. In such a case, we know that the ratio of the ω_i is not real, and the notation is assumed to be so chosen that $\Im\,(\omega_2/\omega_1) > 0$.

We consider the quotient $X = \mathbf{C}/\Gamma$. X is an abelian group, but it is also a topological space, being the orbit space for the action of Γ on $\mathbf{C}$ by translations: $\gamma \in \Gamma$ sends $z \in \mathbf{C}$ to $z + \gamma$. The action is of course fixed-point free and holomorphic; it is also properly discontinuous – this is equivalent to Γ being discrete (as a subset of $\mathbf{C}$). Thus X is in fact a Riemann surface. Recall from Chapter 2 that the natural mapping $\pi : \mathbf{C} \to X$ is holomorphic, and that a map f of an open subset U of X into any Riemann surface is holomorphic iff $f \circ \pi$ is holomorphic. Recall also that a meromorphic function f on any Riemann surface Y is just a holomorphic map $\not\equiv \infty$ of Y into the Riemann surface $\overline{\mathbf{C}} = \mathbf{C} \cup \{\infty\}$ (the Riemann sphere). Thus $\mathcal{M}^\Gamma$ (considered in §1.2) is naturally identified with the field $\mathcal{M}(X)$ of meromorphic functions on X via π.

Observe also that X is compact: π maps the compact subset $\{s\omega_1 + t\omega_2 : 0 \leq t, s \leq 1\}$ of $\mathbf{C}$ onto X. Thus $\mathcal{H}(X) = \mathbf{C}$, and we recover the fact proved earlier that $\mathcal{H}^\Gamma = \mathbf{C}$. Similarly many of the facts proved in §1.2 for $\mathcal{M}^\Gamma$ are properties of (non-constant) meromorphic functions on an arbitrary compact Riemann surface. The special property of $X = \mathbf{C}/\Gamma$ is that there exists on X a canonical holomorphic one-form without any zeros, usually denoted by dz; it is actually the one-form φ on X such that $\pi^*(\varphi) = dz$ on $\mathbf{C}$.

We note that on a general Riemann surface, one can talk about the residues only of meromorphic one-forms, and not of meromorphic functions. The statements about the residues of elements f of $\mathcal{M}^\Gamma$ in §1.2 are really statements about the one-form $f\,dz$ on X. It is a general fact that the sum of all the residues of a meromorphic form on any compact Riemann surface is zero; in particular, there does not exist a meromorphic form with just one simple pole.

We already observed that X is an abelian group. By definition of the quotient topology, it is a topological group, *i.e.*, the group operation $X \times X \to X$ sending (x_1, x_2) to $x_1 - x_2$ is continuous, since it is so on $\mathbf{C}$. For the same reason, X is a *complex Lie group*, i.e the group operations are holomorphic, *e.g.*, the maps $x \to -x$ and $x \to x + x_0$ ($x_0 \in X$ fixed) are holomorphic maps of X into itself.

Lemma 3.1.1 *As a topological group, X is isomorphic to $S^1 \times S^1$, where S^1 is the circle group, i.e., the multiplicative group of complex numbers of absolute value one.*

Proof: The map $e_1 = (1, 0) \to \omega_1$, $e_2 = (0, 1) \to \omega_2$ extends to an $\mathbf{R}$-isomorphism of $\mathbf{R}^2$ onto $\mathbf{C}$ (regarded as $\mathbf{R}^2$ in the usual way), mapping $\mathbf{Z}e_1 \oplus \mathbf{Z}e_2$ onto Γ. Thus we get an isomorphism of $\mathbf{R}^2/(\mathbf{Z}e_1 \oplus \mathbf{Z}e_2)$ onto X. Since $\mathbf{R}/\mathbf{Z}$ is isomorphic to S^1 (via $x \to \exp 2\pi i x$), our assertion follows. □

Remark: Thus, as a topological group, X is just $S^1 \times S^1$, which is called the real two-torus. But we are studying the Riemann surface structure on X defined by Γ; to emphasise this, one calls the Riemann surface $X = \mathbf{C}/\Gamma$ a *complex torus.*

Corollary 3.1.2 *X has precisely three points of order two.*

Proof: Clear, since S^1 has one point of order two. In terms of $\pi : \mathbf{C} \to X$, the points of order two in X are $\pi(\omega_1/2), \pi(\omega_2/2)$ and $\pi((\omega_1 + \omega_2)/2)$). □

Lemma 3.1.3 *Let $x_1, x_2 \in X$. Then $\wp(x_1) = \wp(x_2)$ iff $x_1 = \pm x_2$.*

Proof: Since $\wp$ is even, and takes each value (counted with multiplicity) precisely twice, what we have to show is that $\wp'(x) = 0$ if x is of order two. But since $\wp$ is even, $\wp$ cannot be one-one near an x of order two, so $\wp'(x)$ must vanish. □

Corollary 3.1.4 *The values of $\wp$ at the three points of order two in X are distinct. The three zeros of $\wp'$ are the three points of order two, hence distinct. In particular, $\wp'' \neq 0$ where $\wp' = 0$.*

Proof: If $\wp$ took the same value at two points of order two, it would take each value at least four times. □

Corollary 3.1.5 *If $(\wp')^2 = 4\wp^3 - g_2\wp - g_3$ (c.f. equation (1.1)), then*

$$\wp(\omega_1/2) =: e_1,\ \wp(\omega_1 + \omega_2)/2) =: e_2,\ \wp(\omega_2/2) =: e_3$$

are roots of the equation

$$T^3 - g_2T - g_3 = 0.$$

In particular the roots of this equation are distinct.

Proof: The e_i are roots of the equation in question since $\wp'$ vanishes at the points of order two. But the e_i are distinct by corollary 3.1.4. □

Now consider the cubic curve C in $\mathbf{P}^2$ defined by the cubic form

$$zy^2 = 4x^3 - g_2xz^2 - g_3z^3.$$

This curve has only one point on the line $z = 0$, namely $z = x = 0$ (*i.e.*, $(0, 1, 0)$) which is clearly a smooth point on C. By corollary 3.1.5, this curve is also smooth when $z \neq 0$. Thus C is a smooth curve in $\mathbf{P}^2$ and hence can be regarded as a compact Riemann surface. Now we can state:

Theorem 3.1.1 *The map $\varphi : u \to (\wp(u), \wp'(u), 1)$ of $X^* = X - \{0\}$ into $\mathbf{P}^2$ extends to an isomorphism of X onto C.*

Proof: The map φ is clearly holomorphic on X^* and maps X^* into C. To see how φ behaves near $u = 0$, recall that

$$\wp(u) = \frac{h_1(u)}{u^2}, \wp'(u) = \frac{h_2(u)}{u^3}$$

near $u = 0$, with $h_i(u)$ holomorphic near $u = 0$ and $h_i(0) \neq 0$. Thus, on X^* near $0, \varphi(u) = (uh_1(u), h_2(u), u^3)$ which is clearly holomorphic at $u = 0$ as well and maps $u = 0$ to $(0, 1, 0)$. But φ is clearly one-one on X^* since $\wp'$ is odd and vanishes only at the points of order two, while $\wp(u_1) = \wp(u_2)$ iff $u_1 = \pm u_2$. Thus, as a one-one holomorphic map of compact Riemann surfaces, φ must be biholomorphic. □

3.2 Addition theorem for $\wp$; the group law on cubics

We had already mentioned in an earlier remark that addition mod Γ in $\mathbf{C}$ could be characterised by means of $\wp$ and $\wp'$. In view of Theorem 3.1.1, this just means that the group structure on $X = \mathbf{C}/\Gamma$ can be carried over to the cubic curve C by means of the bijection φ provided by $\wp, \wp'$. We proceed to do this explicitly; observe the obvious fact that the map $u \to -u$ on X corresponds to the map $(x, y, z) \to (x, -y, z)$ on C.

Theorem 3.2.1 *Let u, v, w be three (not necessarily distinct) points of X. Then $u + v + w = 0$ iff $\varphi(u), \varphi(v)$ and $\varphi(w)$ are collinear in $\mathbf{P}^2$.*

Remark: Any line ℓ in $\mathbf{P}^2$ meets C in three points if multiplicity is taken into account. Thus, if say $u = v$ in the above theorem, the theorem means that $u + v + w = 0$ iff the tangent to C at $\varphi(u)$ meets C again $\varphi(w)$. If $u = v = w$, then theorem means that $3u = 0$ iff the tangent to C at $\varphi(u)$ meets C three times at $\varphi(u)$, *i.e.*, does not meet $C - \varphi(u)$. Recall *e.g.*, that the "line at infinity" $z = 0$ meets C only at $\varphi(0) = (0, 1, 0)$.)

Proof: We first consider the "general" case $u, v \in X^*, u \neq \pm v$. In this case $\wp(u) \neq \wp(v)$, hence there is a unique linear combination $f = \wp' + a\wp + b$ which has zeros at u and v, and we have $u + v + w = 0$, where w is the third zero of f (which could be u or v). But this means precisely that $\varphi(w)$ is the

third point where the line $\ell : ax + y + bz = 0$ through $\varphi(u)$ and $\varphi(v)$ meets C. (Observe that ℓ does not pass through $(0, 1, 0)$.)

The case $u = v \in X^*$ is similar, except that in this case, we look for the f (unique upto a constant factor) of the form $c\wp' + a\wp + b$ which has a double zero at u, *i.e.*, we want $c\wp'(u) + a\wp(u) + b = 0$ and $c\wp''(u) + a\wp'(u) = 0$. Since $\wp'$ and $\wp''$ have no common zeros (C is non-singular!), we can find such an f, and its third zero w (which could be u) satisfies $2u + w = 0$. But this clearly means that the line $\ell : ax + cy + bz = 0$ is tangent to C at $\varphi(u)$, and that $\varphi(w)$ is the third point of intersection of ℓ with C.

The case $u = -v \in X^*$ corresponds to the function $\wp + b$ ($b = -\wp(u)$) and the line $\ell : x + bz = 0$. Note that ℓ passes through $\varphi(0)$, as it should. The remaining cases are left as an exercise. □

Before deriving the explicit formula for the coordinates of the sum of two points, we note a purely algebraic geometric corollary of the above theorem. Recall that a point p on a smooth plane curve is called a *point of inflexion* of the curve if the tangent to the curve at p meets the curve at p with mutliplicity ≥ 3 (hence precisely 3 if the curve is a cubic). For the cubic $C = \varphi(X)$, Theorem 3.2.1 identifies the points of inflexion of C with points of order three together with 0 on the group X. Since X is isomorphic as a group to $S^1 \times S^1$, there are eight points of order three in X, hence C has nine distinct points of inflexion.

Theorem 3.2.2 *Let $u, v \in X^*, u \pm v \neq 0$. Let $\varphi(u) = (\wp(u), \wp'(u), 1) = (x_1, y_1, 1)$ and $\varphi(v) = (x_2, y_2, 1)$. Then $\varphi(u + v) = (\wp(u + v), \wp'(u + v), 1) = (x, y, 1)$ where*

$$x = -x_1 - x_2 + \frac{1}{4}\left(\frac{y_2 - y_1}{x_2 - x_1}\right)^2$$

$$y = -\left\{\frac{y_2 - y_1}{x_2 - x_1}\, x + \frac{x_2 y_1 - x_1 y_2}{x_2 - x_1}\right\}.$$

Proof: Let $\ell : y = ax + b$ be the line through $\varphi(u), \varphi(v)$ in the affine (x, y)-space $\mathbf{C}^2 (= \mathbf{P}^2 - \{z = 0\})$. We know that ℓ meets the curve

$$y^2 = 4x^3 - g_2 x - g_3$$

in the three points $\varphi(u), \varphi(v)$ and $\varphi(-u-v)$, *i.e.*, x_1, x_2 and x are three roots of the equation

$$(ax + b)^2 = 4x^3 - g_2 x - g_3.$$

Hence

$$x_1 + x_2 + x = \frac{a^2}{4}.$$

Since clearly

$$a = \frac{y_2 - y_1}{x_2 - x_1}$$

and

$$b = \frac{x_2 y_1 - x_2 y_2}{x_2 - x_1}$$

and $y = -ax - b$, the theorem is proved. □

Corollary 3.2.1 *For $u \in X$ with $2u \neq 0$, we have*

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

and the coordinates of $\varphi(2u)$ are given by

$$(-2x + \frac{1}{4}a^2, -a(-2x + \frac{1}{4}a^2) - b, 1)$$

where $(x, y, 1)$ are the coordinates of $\varphi(u)$ and

$$a = \frac{\wp''(u)}{\wp'(u)} = \frac{6x^2 - \frac{1}{2}g_2}{y}, \quad b = y - ax.$$

Proof: The details are left as an exercise. We have only to determine a and b so that $\wp' - a\wp - b$ has a double zero at u, *i.e.*, $\wp'(u) - a\wp(u) - b = 0$ and $\wp''(u) - a\wp'(u) = 0$. (Observe that $\wp'(u) \neq 0$). The point to note is that $\wp'' = 6\wp^2 - \frac{1}{2}g_2$, as follows by differentiating the relation $(\wp')^2 = 4\wp^3 - g_2\wp - g_3$. □

Rational functions on C induce (via φ) meromorphic functions on X. We now show that these are the only meromorphic functions on X.

Theorem 3.2.3 *Every meromorphic function on X is of the form $R_1(\wp) + \wp' R_2(\wp)$, where the $R_i(\wp)$ are rational functions of $\wp$ (i.e., quotients of polynomials in $\wp$).*

Proof: Let f be any meromorphic function on X. Since we can consider the even and odd "parts" $f(x) \pm f(-x)$ of f, it is enough to consider the two cases separately. But $\wp' f$ is even if f is odd; hence it is enough to show that any even f is a rational function of $\wp$.

Thus let f be an even meromorphic function on X. We regard f and $\wp$ as holomorphic maps $X \to \mathbf{C} \cup \{\infty\}$. Since f is even, and $\wp(u) = \wp(v)$ iff $u = \pm v$, there exists a set-theoretic map R of $\mathbf{C} \cup \{\infty\}$ into itself such that $f = R \circ \wp$. Since $\wp$ is surjective and holomorphic, and f is holomorphic, R is also a holomorphic map. Thus R is a meromorphic function on $\mathbf{C} \cup \{\infty\}$, hence a rational function. $\square$

Chapter 4

Isomorphism of tori and the j invariant

4.1 Isomorphism of tori

We now investigate the question about the isomorphism of two tori $\mathbf{C}/\Gamma$, $\mathbf{C}/\Gamma'$.

Proposition 4.1.1 *$\mathbf{C}/\Gamma$ and $\mathbf{C}/\Gamma'$ are biholomorphic to each other iff there is a non-zero constant λ such that $\lambda\Gamma = \Gamma'$.*

Proof: Let $\varphi : \mathbf{C}/\Gamma \to \mathbf{C}/\Gamma'$ be a biholomorphic map. Using a suitable translation in $\mathbf{C}/\Gamma'$, we can assume that $\varphi(0) = 0$. (Here we are using the obvious group laws on $\mathbf{C}/\Gamma$ and $\mathbf{C}/\Gamma'$, with respect to which translation is a holomorphic mapping). By the standard properties of covering spaces, we get a biholomorphic map $\tilde{\varphi} : \mathbf{C} \to \mathbf{C}$ such that we have a commutative diagram:

$$\begin{array}{ccc} \mathbf{C} & \stackrel{\tilde{\varphi}}{\to} & \mathbf{C} \\ \downarrow & & \downarrow \\ \mathbf{C}/\Gamma & \stackrel{\varphi}{\to} & \mathbf{C}/\Gamma' \end{array}$$

Both the universal covering maps $\mathbf{C} \to \mathbf{C}/\Gamma$ and $\mathbf{C} \to \mathbf{C}/\Gamma'$ are group homomorphisms. From this we see easily that $\tilde{\varphi}$ can also be chosen to be a group homomorphism (after composing with a suitable translation in $\mathbf{C}$).

Now it follows that $\tilde{\varphi}$ has the form $\tilde{\varphi}(z) = \lambda z$ for some $\lambda \neq 0$. For any $w \in \Gamma$, by the commutativity of the diagram, $\tilde{\varphi}(w) = \lambda w \in \Gamma'$. Thus $\lambda\Gamma \subset \Gamma'$. Since $\varphi, \tilde{\varphi}$ are surjective we also get $\lambda\Gamma = \Gamma'$.

Conversely, if $\lambda\Gamma = \Gamma'$ for some $\lambda \neq 0$, then the map $\widetilde{\varphi} : \mathbf{C} \to \mathbf{C}$ defined by $\widetilde{\varphi}(z) = \lambda z$ descends to a biholomorphic map $\mathbf{C}/\Gamma \to \mathbf{C}/\Gamma'$. □

4.2 The functions $g_2(\tau)$, $g_3(\tau)$, $j(\tau)$

Recall from Theorem 3.1.1 that for any torus $\mathbf{C}/\Gamma$, we have a holomorphic embedding $\varphi : \mathbf{C}/\Gamma \hookrightarrow \mathbf{P}^2$ such that the image of φ is the non-singular cubic with equation $ZY^2 = 4X^3 - g_2XZ^2 - g_3Z^3$. Then the meromorphic function X/Z is the Weierstrass $\wp$-function with period lattice Γ.

We can assume that Γ has a basis w, w' such that $\tau = w'/w = x + \iota y$ with $y > 0$. By definition,

$$g_2 = 60 \sum{}' \frac{1}{(mw + nw')^4} \quad \text{and} \quad g_3 = 140 \sum{}' \frac{1}{(mw + nw')^6}.$$

We can write

$$g_2 = g_2(w, w') = \frac{60}{w^4} \sum{}' \frac{1}{(m + n\tau)^4} = \frac{1}{w^4} g_2(1, \tau),$$

$$g_3 = g_3(w, w') = \frac{140}{w^6} \sum{}' \frac{1}{(m + n\tau)^6} = \frac{1}{w^6} g_3(1, \tau),$$

$$\Delta = g_2(w, w')^3 - 27 g_3(w, w')^2 = \frac{1}{w^{12}} \Delta(1, \tau).$$

The series for $g_2(1, \tau)$, $g_3(1, \tau)$ converge absolutely in any compact subset of the upper half-plane. Hence $g_2(1, \tau)$, $g_3(1, \tau)$ are analytic functions of τ in the upper half-plane and so also is $\Delta(1, \tau)$.

Lemma 4.2.1 *$\Delta(1, \tau) \neq 0$ for any τ in the upper half plane.*

Proof: (*cf.* Corollary 3.1.4). It suffices to show that $\Delta(w, w') \neq 0$ where w'/w has positive imaginary part.

Let $\wp$ be the Weierstrass $\wp$ function with periods w, w'. Since $\wp$ is an even function, $\wp'$ is an odd function. For any period $\Omega \in \Gamma$, $\wp'(-\Omega/2) = \wp'(\Omega/2)$ as Ω is also a period of $\wp'$. But $\wp'(-\Omega/2) = -\wp'(\Omega/2)$ as $\wp'$ is an odd function. It follows that $\wp'(z)$ is zero at the three points $w/2, w'/2$ and $(w + w')/2$ in the period parallelogram.

From the relation $\wp'^2(z) = 4\wp(z)^3 - g_2\wp(z) - g_3$, we see that the three quantities $e_1 = \wp(w/2)$, $e_2 = \wp((w+w')/2)$, $e_3 = \wp(w'/2)$ are the roots of the equation $4t^3 - g_2 t - g_3 = 0$. Now $\wp(z)$ takes each value exactly twice in the period parallelogram (counted with multiplicity). Since $\wp'(w/2) = 0, \wp(z)$ takes the value e_1 with multiplicity 2 at $w/2$, hence cannot take the value e_1 elsewhere in the parallelogram. Similar argument for e_2, e_3 shows that $e_1 \neq e_2, e_2 \neq e_3, e_1 \neq e_3$. It follows that the discriminant $\Delta = g_2^3 - 27g_3^2 \neq 0$. □

Remark: $\Delta \neq 0$ also follows from the smoothness of the cubic $\{ZY^2 = X^3 - g_2XZ^2 - g_3Z^3\}$.

Definition: The function $j(\tau) = g_2^3/\Delta$ is called the *j-function* on the upper half plane. Clearly,

$$j(\tau) = \frac{g_2(w,w')^3}{g_2(w,w')^3 - 27g_3(w,w')^2}, \text{ where } \tau = \frac{w'}{w}$$

We now prove a lemma determining the relation between two bases of the same lattice Γ. Let $\Gamma = \mathbf{Z}w + \mathbf{Z}w' = \mathbf{Z}w_1 + \mathbf{Z}w_1'$. We can assume that $\tau = w'/w$ and $\tau_1' = w_1'/w_1$ have positive imaginary parts. By elementary group theory, we have $w_1' = aw' + bw, w_1 = cw' + dw$ where a, b, c, d are integers satisfying $ad - bc = \pm 1$.

Lemma 4.2.2 *With the above notation,* $ad - bc = 1$.

Proof: We have

$$\tau_1' = \frac{a\tau + b}{c\tau + d} = \frac{(a\tau + b)(c\overline{\tau} + d)}{\mid c\tau + d \mid^2}$$

$$= \frac{(ac\tau\overline{\tau} + bd) + (ad\tau + bc\overline{\tau})}{\mid c\tau + d \mid^2}.$$

If $ad - bc = -1, bc = ad + 1$ and

$$\tau_1' = \frac{(ac\tau\overline{\tau} + bd) + ad(\tau + \overline{\tau}) + \overline{\tau}}{\mid c\tau + d \mid^2}$$

It follows that the imaginary part of τ_1' is $-y/\mid c\tau + d\tau \mid^2$, where $\tau = x + \iota y$. This is a contradiction since $y > 0$ and τ_1' has positive imaginary part. □

Now for every torus $\mathbf{C}/\Gamma$, we can construct its *j-invariant* $j(\tau)$ where $\Gamma = \mathbf{Z}w' + \mathbf{Z}w$ and $\tau = w'/w$. If w_1, w_1' is another basis of Γ with imaginary part of $\tau_1' = w_1'/w_1$ being positive, we know from Lemma 4.2.2 that $w_1' = aw' + bw$, $w_1 = cw + dw$ with $ad - bc = 1$. Since $g_2(w, w') = g_2(w_1, w_1')$ and $g_3(w, w') = g_3(w_1, w_1')$, we see that

$$j\left(\frac{a\tau + b}{c\tau + d}\right) = j(\tau) = j(\tau_1').$$

The j-function is therefore invariant under $SL(2, \mathbf{Z})$ and so any $\mathbf{C}/\Gamma$ defines a number $j(\Gamma) = j(\mathbf{C}/\Gamma) = j(\tau)$, where $\tau = w'/w$ for any basis w, w' of Γ with $\Im w'/w > 0$.

4.3 Isomorphism of plane cubics in terms of j

We have seen above that for any torus $\mathbf{C}/\Gamma$, we have a well-defined invariant $j(\Gamma) = j(\mathbf{C}/\Gamma) = j(\tau)$, where $\Gamma = \mathbf{Z}w + \mathbf{Z}w'$ and $\tau = w'/w$ with $\Im\tau > 0$. We have an embedding $\mathbf{C}/\Gamma \hookrightarrow \mathbf{P}^2$ with the homogeneous equation $ZY^2 = 4X^3 - g_2XZ^2 - g_3Z^3$ (where $g_2^3 - 27g_3^2 \neq 0$). Conversely, any equation of this type where the constants g_2, g_3 satisfy $g_2^3 - 27g_3^2 \neq 0$ defines a non-singular cubic curve in $\mathbf{P}^2$. By a suitable linear change of co-ordinates in $\mathbf{P}^2$, any non-singular cubic can be put in the above form.

We now turn our attention to the question whether any non-singular cubic with equation $ZY^2 - 4X^3 - g_2XZ^2 - g_3Z^3$ is biholomorphic to a torus $\mathbf{C}/\Gamma$ such that g_2, g_3 are the constants defined earlier for the lattice Γ. We will show presently (see Theorem 5.2.1) that this is indeed the case. But assuming this result for a moment, we will settle an easier question.

Lemma 4.3.1 *Let* $C_1 : \{ZY^2 = 4X^3 - g_2XZ^2 - g_3Z^3\}$ *and* $C_2 : \{ZY^2 = 4X^3 - g_2'XZ^2 - g_3'Z^3\}$ *be two non-singular cubics in* $\mathbf{P}^2$. *Then* C_1 *is isomorphic to* C_2 *iff* $g_2^3/(g_2^3 - 27g_3^2) = {g_2'}^3/({g_2'}^3 - 27{g_3'}^2)$.

Proof: Suppose $C_1 \cong \mathbf{C}/\Gamma_1$ and $C_2 \cong \mathbf{C}/\Gamma_2$. We have shown earlier that if $\mathbf{C}/\Gamma_1$ and $\mathbf{C}/\Gamma_2$ are biholomorphic, then $j(\Gamma_1) = j(\Gamma_2)$.

Conversely, if $j(\Gamma_1) = j(\Gamma_2)$ then either $g_2 = 0 = g_2'$, or $g_2 \neq 0$, $g_2' \neq 0$. If $g_2 = 0 = g_2'$, let $g_3 = \lambda g_3'$ for some $\lambda \in \mathbf{C}$ (observe that $g_3, g_3' \neq 0$).

Then absorbing a cube root of λ and changing variables, an equation for C_1 can be found which is same as the above equation for C_2. By similar argument, we are through if $g_3 = 0 = g_3'$. Suppose now $g_2, g_3 \neq 0, g_2', g_3' \neq 0$. Then from the equation

$$\frac{g_2^3 - 27g_3^2}{g_2^3} = \frac{{g_2'}^3 - 27{g_3'}^2}{{g_2'}^3}$$

we get

$$\frac{g_3^2}{g_2^3} = \frac{g_3'^2}{g_2'^3} \text{ or } (\frac{g_2'}{g_2})^3 = (\frac{g_3'}{g_3})^2.$$

Let $g_2' = \lambda^2 g_2$, then $(g_3'/g_3)^2 = \lambda^6$ or $g_3'/g_3 = \pm\lambda^3$. But if $g_3'/g_3 = -\lambda^3$, then $g_2' = (-\lambda)^2 g_2, g_3' = (-\lambda)^3 g_3$. It follows that in either case, we can find $\lambda \in \mathbf{C}$ such that $g_2' = \lambda^2 g_2, g_3' = \lambda^3 g_3$. Then again a suitable change of co-ordinates makes the equations of C_1, C_2 the same. $\square$

Chapter 5

All smooth cubics are complex tori

5.1 All smooth cubics are complex tori: the first proof

We saw above that every lattice Γ in $\mathbf{C}$ leads to a smooth cubic C Weierstrass form

$$zy^2 = 4x^3 - g_2xz^2 - g_3z^3$$

in $\mathbf{P}^2(\mathbf{C})$, where $g_2, g_3 \in \mathbf{C}$ can be explicitly determined from Γ (see below), and of course satisfy $g_2^3 - 27g_3^2 \neq 0$, from Lemma 4.2.1. As a Riemann surface, C is isomorphic to the complex torus $\mathbf{C}/\Gamma$.

The question naturally arises whether every smooth cubic in $\mathbf{P}^2(\mathbf{C})$ is associated to a lattice (or a torus) in this way. First of all, it is true that every smooth cubic in $\mathbf{P}^2(\mathbf{C})$ can be brought to Weierstrass form as above by a linear change of variables in $\mathbf{P}^2(\mathbf{C})$; this will be proved elsewhere in this course. Thus the question becomes: given $g_2, g_3 \in \mathbf{C}$ such that $g_2^3 - 27g_3^2 \neq 0$, can we find a lattice Γ in $\mathbf{C}$ such that the g_2, g_3 arising from the corresponding $\wp, \wp'$ are the given ones?

Let us first start from a lattice Γ in $\mathbf{C}$, and determine the corresponding g_2, g_3 explicitly. Recall that

$$\wp'(z) = -2 \sum_{\gamma \in \Gamma'} (z - \gamma)^{-3},$$

so that the Laurent expansion of $\wp'(z)$ near $z = 0$ is

$$-2z^{-3} - 2 \sum_{\gamma \in \Gamma^*} \left\{ -\frac{1}{\gamma^3}(1 + \frac{3z}{\gamma} + \frac{6z^2}{\gamma^2} + \frac{10z^3}{\gamma^3} + \cdots) \right\}$$

$$= -2z^{-3} + 2a_2 z + 4a_4 z^3 + O(z^5)$$

(recall that $\wp'$ is odd), where

$$a_2 = 3 \sum_{\gamma \in \Gamma^*} \gamma^{-4}, a_4 = 5 \sum_{\gamma \in \Gamma^*} \gamma^{-6}.$$

By integration, the Laurent expansion of $\wp$ is

$$z^{-2} + a_2 z^2 + a_4 z^4 + O(z^6)$$

Thus

$$\wp'(z) = -2z^{-3}\{1 - a_2 z^4 - 2a_4 z^6 + O(z^8)\}$$

and

$$\wp(z) = z^{-2}\{1 + a_2 z^4 + a_4 z^6 + O(z^8)\}.$$

Hence

$$\wp'(z)^2 - 4\wp(z)^3 = 4z^{-6}\{1 - 2a_2 z^4 - 4z^6 - 1 - 3a_2 z^4 - 4a_4 z^6 + O(z^8)\}$$

$$= -20a_2 z^{-2} - 28a_4 + O(z^2)$$

Thus $\wp'^2 - 4\wp^3 + 20a_2\wp + 28a_4$ is an elliptic function which has no poles at all and vanishes at $z = 0$, hence must vanish identically. We have thus proved:

Theorem 5.1.1 *The g_2, g_3 corresponding to a lattice Γ are given by*

$$g_2 = 60 \sum_{\gamma \in \Gamma^*} \gamma^{-4}, g_3 = 140 \sum_{\gamma \in \Gamma^*} \gamma^{-6}$$

Thus, to show that every smooth cubic is a complex torus, we must, given $g_2, g_3 \in \mathbf{C}$ with $g_2^3 - 27g_3^2 \neq 0$, find $\omega_1, \omega_2 \in \mathbf{C}$ with ratio not real such that

$$g_2 = 60 \sum_{(m,n) \neq (0,0)} (m\omega_1 + n\omega_2)^{-4},$$

$$g_3 = 140 \sum_{(m,n) \neq (0,0)} (m\omega_1 + n\omega_2)^{-6}.$$

Although the problem looks rather formidable when stated in this form, it can in fact be solved in this form, as shown elsewhere in this course. We only observe here that the above series for g_2, g_3 define functions of ω_1, ω_2

which are clearly homogeneous of degrees -4 and -6 respectively, hence the quotient $g_2^3/(g_2^3 - 27g_3^2)$ is homogeneous of degree zero; hence this quotient can be regarded as a function of the single complex variable $\tau = \omega_2/\omega_1$ (where $\Im\tau > 0$). This function of τ is usually denoted by j (*cf.* Section 4.2.), and our problem is equivalent to showing that j takes on every complex value. The importance of this function in the study of elliptic curves will be clear by the end of this course.

Another proof of the fact that smooth cubics in $\mathbf{P}^2(\mathbf{C})$ are biholomorphic to complex tori can be deduced from the fact that smooth cubics carry a group structure in which the group operations are algebraic. Indeed the description of the group structure on a complex torus $\mathbf{C}/\Gamma$ in terms of its embedding as a smooth cubic in Weierstrass form applies to every such cubic. The proof based on this fact will also be given elsewhere in the course (*cf.*5.2.1)

We now want to sketch a proof for the fact that smooth cubics are tori. This proof has a more topological flavour, and the considerations on which it is based are important in the study of arbitrary compact Riemann surfaces. Thus consider a smooth cubic C:

$$zy^2 = 4x^3 - g_2xz^2 - g_3z^3$$

in Weierstrass form. On the affine piece $z = 1$ of $\mathbf{P}^2(\mathbf{C})$, let us write x, y instead of x/z, y/z, and consider the corresponding affine piece C_0 of C:

$$y^2 = 4x^3 - g_2x - g_3.$$

On C_0, x and y are holomorphic functions, and we can consider the one-form $\psi = dx/y$ on C_0. The form ψ is obviously a meromorphic one-form on the whole of C:

$$\psi = \left(\frac{z}{y}\right) d\left(\frac{x}{z}\right).$$

But ψ is in fact a holomorphic form on all of C, and has no zeros. This is clear except at the points $(e_i, 0)$ on C_0 (the e_i being the three distinct roots of $4T^3 - g_2T - g_3 = 0$) and at the point at infinity $(0, 1, 0)$. However at any $(e_i, 0), y$ is a local coordinate, and we have $y^2 = (x - e_i)h$, where h is holomorphic and non-zero at $(e_i, 0)$. Thus dx has a simple zero at $(e_i, 0)$, hence $\frac{dx}{y}$ is holomorphic and non-zero there.

Similarly, at $(0,1,0)$, x/y is a local co-ordinate, and z/y has a zero of order three, hence

$$\psi = \frac{z}{y} d(\frac{x}{z})$$

is holomorphic and non-zero.

Now let us assume the following fact from topology (see the second remark after Theorem 5.1.2 below).

Lemma 5.1.1 *The fundamental group $\pi_1(C)$ of C is isomorphic to $\mathbf{Z} \oplus \mathbf{Z}$; in fact, C is homeomorphic to $S^1 \times S^1$.*

Fix generators α_1 and α_2 for $\pi_1(C)$.

Lemma 5.1.2 *Let $\omega_i = \int_{\alpha_i} \psi, i = 1, 2$. Then the ratio of the ω_i is not real.*

Proof: We first observe that ω_1 and ω_2 cannot be both zero. This is because $F(p) = \int_q^p \psi$ (q a fixed point on C) is a multiple-valued holomorphic function on C: $F(p)$ is determined modulo the periods of ψ, namely the values of $\int_\beta \psi$, β a closed path at q. But under our assumption on $\pi_1(C)$, the periods are just integral linear combinations of ω_1 and ω_2. Thus if $\omega_1 = 0 = \omega_2$, F will be a (single-valued) holomorphic function on C, hence $\equiv 0$ by the maximum principle. But this implies $\psi \equiv 0$, which is absurd.

By replacing ψ by a constant multiple and interchanging α_1 and α_2 if necessary we may assume $\omega_1 = 1$. Suppose $\omega_2 \in \mathbf{R}$. Then we can still consider the multiple-valued holomorphic function F on C defined above. If ω_1 and $\omega_2 \in \mathbf{R}$, the imaginary part of F will be single-valued on C. Being locally the imaginary part of a holomorphic function, $\Im F$ will satisfy the maximum principle, and hence reduce a constant. But this implies that $\psi \equiv 0$, which is absurd. □

Thus the function F considered above defines a holomorphic map $A : C \to \mathbf{C}/\Gamma$ where $\Gamma \subset \mathbf{C}$ is the lattice generated by ω_1 and ω_2.

Theorem 5.1.2 *$A : C \to \mathbf{C}/\Gamma$ is biholomorphic.*

Proof: Let dz denote the standard one-form on $\mathbf{C}/\Gamma$. By the definition of A, it is clear that $A^*(dz) = \psi$. Since ψ has no zeros, it follows that A is locally biholomorphic.

It is clear that $A(q) = 0$. Suppose there exists $p \neq q$ in C such that $A(p) = 0$. By the definition of A, this means that there is a path β from q to p in C such that $\int_\beta \psi = 0$. Since $\psi = A^*(dz)$, this means that $\int_{A(\beta)} dz = 0$. Now $A(\beta)$ is a closed path at 0 in $\mathbf{C}/\Gamma$, and it is clear that a closed path in $\mathbf{C}/\Gamma$ is null-homotopic iff $\int_\gamma dz = 0$. Thus we have found a null-homotopic closed path $A(\beta)$ in $\mathbf{C}/\Gamma$ whose "lift" β to C is not closed.

Now observe that $A : C \to \mathbf{C}/\Gamma$, being a local homeomorphism of compact (connected locally connected) spaces, is a covering map. It is a standard fact from covering space theory that a null-homotopic loop lifts to a (null-homotopic) loop. Thus we have a contradiction, hence we must have $A^{-1}(0) = \{q\}$.

Finally, observe that q was an arbitrary point of C; a different choice q' for q would have led to a map $A' : C \to \mathbf{C}/\Gamma$ differing from A only by a translation in $\mathbf{C}/\Gamma$. Hence A is injective. (Alternately one could use the fact that all fibres of a covering map have the same cardinality.) □

Remark: The above proof that C is a torus identifies the lattice Γ as the period lattice of the non-zero holomorphic one-form (unique upto a constant multiple) on C. It would apply to any compact Riemann surface C of genus one *i.e.*, one on which the vector space of holomorphic one-forms is one-dimensional, provided one knew that in such a case a form $\not\equiv 0$ has no zeros and (a weaker form of) Lemma 5.1.1 is true for such a C. Both these statements are in fact true, but their proofs are beyond the scope of this workshop. See however Theorem 5.1.3 below.

Remark: There are many ways of proving Lemma 5.1.1. One way is to prove that any two smooth cubics (assumed to be in Weierstrass form) are homeomorphic, since we know that at least some of them are homeomorphic to $S^1 \times S^1$. Thus, let C_i be the affine part

$$y^2 = (x - a_i)(x - b_i)(x - c_i)$$

of a smooth cubic, punctured at the three points over a_i, b_i, c_i. To produce a homeomorphism $C_1 \to C_2$, we choose any homeomorphism h of the x-space

$\mathbf{C}$ onto itself mapping a_1 to a_2, b_1 to b_2 and c_1 to c_2. Let $p_i : C_i \to \mathbf{C}$ denote the canonical projection $(x, y) \to x$. Fix points $w_i \in C_i$ such that $p_2(w_2) = h(p_1(w_1))$. For any $q_i \in C_1$, choose any path $w_1 q_1$ from w_1 to q_1 in C_1, and define $\widetilde{h}(q_1)$ as the end-point of the lift by p_2 of the path $h(p_1(w_1 q_1))$ in C_2 starting at w_2. If we check that $\widetilde{h}$ is well-defined, then $\widetilde{h}$ will be obviously a homeomorphism, and it will be easy to see that $\widetilde{h}$ extends to a homeomorphism of the corresponding full cubics.

The fact that $\widetilde{h}$ is well-defined follows easily from the lemma below.

Lemma 5.1.3 *The lift of a closed path α in $\mathbf{C} - \{a_i, b_i, c_i\}$ to C_i is a closed path in C_i if and only if the sum of the winding numbers of α around a_i, b_i, c_i is even. Further, the winding numbers of a path α in $\mathbf{C} - \{a_i, b_i, c_i\}$ around say a_1 and of $h(\alpha)$ around a_2 differ at most in sign.*

Proof: Observe that

$$\frac{dy}{y} = p_i^*\left(\frac{dz}{z - a_i} + \frac{dz}{z - b_i} + \frac{dz}{z - c_I}\right)$$

on C_i. Thus, if $\widetilde{\alpha_i}$ is a lift of α, we have

$$\int_{\widetilde{\alpha}} \frac{dy}{y} = \int_{\alpha} \left(\frac{1}{z - a_i} + \frac{1}{z - b_i} + \frac{1}{z - c_i}\right) dz.$$

Clearly $\widetilde{\alpha}$ is closed iff

$$\int_{\widetilde{\alpha}} \frac{dy}{y}$$

is an even multiple of $2\pi i$. Thus the first part of the lemma follows.

For the second part, observe that the winding number about $a \in \mathbf{C}$ of a loop γ in $\mathbf{C} - a$ is a generator of the image of $\pi_1(S^1)$ in $\pi_1(\mathbf{C} - a)$ under the map $\gamma : S^1 \to \mathbf{C} - \{a\}$ — recall that $\pi_1(S_1)$ and $\pi_1(\mathbf{C} - a)$ are infinite cyclic. $\square$

Theorem 5.1.3 *Let X be a compact Riemann surface carrying a nowhere-vanishing holomorphic 1-form. Then X is biholomorphic to a complex torus.*

Proof: We state the various steps of the proof, leaving the detailed verification to the reader. We fix a nowhere vanishing holomorphic 1-form φ on X.

1. There exists a (finite) covering of X by coordinate charts (U_i, z_i) such that $\varphi = dz_i$ on U_i.

2. There exists a finite refinement $V_\alpha \subset U_{i(\alpha)}$ of $\{U_i\}$ such that, whenever $V_\alpha \cap V_\beta \neq \emptyset$, $V_\alpha \cup V_\beta$ is contained in some U_i. (To see this, assume that X is metrisable as a topological space and take V'_αs diameter less than half the Lebesgue number of the covering $\{U_i\}$.) If f_α now denotes $z_{i(\alpha)|V_\alpha}$, then $f_\alpha - f_\beta$ is constant on $V_\alpha \cap V_\beta$ (whenever this intersection is not empty).

3. We may suppose f_α maps V_α biholomorphically onto a disc $\{|z| < r_\alpha\}$ in $\mathbf{C}$. There exist $s_\alpha < r_\alpha$ such that the sets $W_\alpha := f_\alpha^{-1}(|z| < s_\alpha)$ still cover X.

4. Let $\epsilon := \inf_\alpha(r_\alpha - s_\alpha)$. For $u \in \mathbf{C}$ with $|u| < \epsilon$, define $T_u : X \to X$ as follows: if $x \in W_\alpha$, $T_u(x)$ is the unique point in V_α such that $f_\alpha(T_u(x)) = f_\alpha(x) + u$. Then T_u is a well defined holomorphic map. If $|u|$, $|v|$ and $|u+v|$ are less than ϵ, then $T_{u+v} = T_u \circ T_v$.

5. For any $u \in \mathbf{C}$, $T_u : X \to X$ can be defined as $T^n_{u/n}$, with $n > 0$ any integer such that $|u/n| < \epsilon$. Then $T_{u+v} = T_u \circ T_v$ for all $u, v \in \mathbf{C}$.

6. For each $x \in X$, $u \mapsto T_u(x)$ is a non-constant holomorphic map $\mathbf{C} \to X$ and its image $O(x)$ is an open subset of X. Hence $O(x) = X$ for each $x \in X$.

7. X is biholomorphic to the torus $\mathbf{C}/\Gamma$, where Γ is the lattice in $\mathbf{C}$ defined by
$$\Gamma = \{u \in \mathbf{C} : T_u(x) = x\}$$
where x is an arbitrarily chosen point of X.

□

5.2 All smooth cubics are complex tori: second proof

We will now give a second proof that any non-singular cubic in $\mathbf{P}^2$ is biholomorphic to some $\mathbf{C}/\Gamma$. We need only consider cubics in Weierstrass form, since any non-singular cubic is isomorphic to one in Weierstrass form.

Theorem 5.2.1 *Let $C : \{ZY^2 = 4X^3 - g_2XZ^3 - g_3Z^3\}$ be a non-singular cubic in $\mathbf{P}^2$. Then there is a lattice Γ in $\mathbf{C}$ such that $g_2 = g_2(\Gamma)$ and $g_3 = g_3(\Gamma)$. In particular, C and $\mathbf{C}/\Gamma$ are biholomorphic.*

Proof: We have shown in the previous section that the 1-form $\eta = \dfrac{dx}{y}$ is holomorphic on C and does not vanish at any point.

Now we use the result proved in the geometric theory that C is a commutative algebraic group. Let O be the identity element of this group and w a local co-ordinate near O such that $w(O) = 0$. We can write $\eta = f(w)dw$, where f is holomorphic near O. We can find a holomorphic function $g(w)$, holomorphic near e, such that $\dfrac{dg}{dw} = f(w)$ and $g(O) = 0$.

For any $P \in C$, we have the translation isomorphism t_P of C defined by $t_P(Q) = P+Q$. Then $t_P^*(\eta)$ is also a nowhere zero holomorphic 1-form on C. We can therefore write $t_P^*(\eta) = \lambda.\eta$, where λ is a (nowhere zero) holomorphic function on C. But then λ is a constant by the maximum modulus principle (C is compact). The constant λ *a priori* depends on P, but can be easily seen to be a holomorphic function of P, hence again a constant. Thus $\lambda = 1$ and $t_P^*(\eta) = \eta$ shows that η is an invariant 1-form on C.

Next we will show that $g(P+Q) = g(P)+g(Q)$ if P, Q are close to O. For any holomorphic function $\varphi(w)$ in an open set $U \subset C$ with local co-ordinate w on U, we define the holomorphic 1-form $d\varphi$ by the equation $d\varphi = \dfrac{d\varphi}{dw}dw$. Then from the definition of t_P^*, we verify easily that for $P \in C$ and φ a holomorphic function on C near P that $t_P^*(d\varphi) = d(t_P^*\varphi)$, where $t_P^*\varphi$ is the holomorphic function in a neighborhood of O defined by $t_P^*\varphi(Q) = \varphi(P+Q)$.

Going back to our form η, we have now $\eta = dg$. Let P be very close to O, then we have $t_P^*(\eta) = \eta = t_P^*(dg) = d(t_P^*g)$, so that $d(g - t_P^*g) = 0$. Thus $g - t_P^*g$ is a constant function. Then $g - t_P^*g = g(O) - t_P^*g(O) = g(O) - g(P) = -g(P)$ since $g(O) = 0$. But then for Q close to O,

$$g(Q) - t_P^*g(Q) = g(Q) - g(P+Q) = -g(P)$$

i.e., $g(P+Q) = g(P) + g(Q)$.

Now we can quickly finish the proof of the Theorem 5.2.1. Let U be a neighborhood of O in C in which g is holomorphic. We can define a local homomorphism $\psi : U \to \mathbf{C}$ by $\psi(P) = g(P)$. Since $g'(0) \neq 0, \psi$ is a

biholomorphic map of U onto $\psi(U)$. Define $\widetilde{\psi} : \mathbf{C} \to C$ by $\widetilde{\psi}(z) = \ell\widetilde{\psi}(\frac{z}{\ell})$, where ℓ is a large integer such that $\frac{z}{\ell} \in \psi(U)$. This is easily seen to be well-defined. Now the smallest subgroup of C containing U is the whole of C, as C is connected. From this we see that $\widetilde{\psi}$ is a surjective homomorphism which is also a holomorphic map. If Γ is the kernel of $\widetilde{\psi}$, then we have a biholomorphic map $\mathbf{C}/\Gamma \to C$ (note that ψ is locally biholomorphic). Since C is compact, Γ is a lattice. By multiplication by a suitable non-zero constant, we can assume that $g_2(\Gamma) = g_2$, $g_3(\Gamma) = g_3$ (Proposition 4.1.1 and the proof of Lemma 4.3.1). The proof of the theorem is therefore complete. □

5.3 Elliptic integrals

Now that we know that every smooth cubic is a complex torus, it is easy to evaluate elliptic integrals explicitly, at least in principle.

Let us first consider elliptic integrals in Weierstrass form. Recall what this means: we are considering an indefinite integral $\int R(x, y)dx$, where R is a rational function of two complex variables, $y = y(x)$ is the (multiple-valued) function $2\{(x - e_1)(x - e_2)(x - e_3)\}^{1/2}$, with e_1, e_2, e_3 distinct, and the path of integration avoiding the e_i and the poles of $R(x, y)$. In order to make sense of the integral, we must choose a branch of y at the initial point of the path and continue it analytically along the path. Thus, practically by definition, what we are considering is the integral of the meromorphic (or rational) one-form

$$R\left(\frac{x}{z}, \frac{y}{z}\right) d\left(\frac{x}{z}\right)$$

on the cubic curve $C : zy^2 = 4\prod(x - e_i z)$ along paths on C.

Now let C be isomorphic to the complex torus $X = \mathbf{C}/\Gamma$ via the mapping $\varphi : X \to C$ given by $\varphi(u) = (\wp(u), \wp'(u), 1)$. Then

$$\varphi^*\left(R\left(\frac{x}{z}, \frac{y}{z}\right) d\left(\frac{x}{z}\right)\right) = R\left(\wp, \wp'\right) d\wp = R\left(\wp, \wp'\right) \wp' dz.$$

Thus we are looking at integrals of the form $\int f(z)dz$ along paths on $\mathbf{C}$ avoiding the poles of f, where f is an arbitrary elliptic function *i.e.*, $f \in \mathcal{M}^\Gamma$. But we know that, if (a_i) are the residues of f at its poles (z_i) (incongruent mod Γ), then

$$f(z) = \sum a_i \zeta(z - z_i) + g(z) + \text{ constant}$$

where $g(z)$ is a linear combination of translates $\wp$ and its derivatives. And we know the primitives of all the functions involved explicitly: for $\wp$ it is $-\zeta$, and for ζ it is the many-valued function $\log \sigma$. Thus we have in principle an explicit formula for the multiple-valued functions $\int^z f(t)dt$.

Example 5.3.1: We want to explain in a concrete case what the above generalities mean. Suppose we want to "determine"

$$\int_\alpha^\infty \{(z - e_1)(z - e_2)(z - e_3)\}^{-1/2} dz$$

along a given path from α to ∞ avoiding the e_i. (The upper limit is chosen as ∞ for convenience; the integral means $\lim_{t\to\infty} \int_\alpha^t$ which clearly exists.)

We are given α and a specific square root β of $\{\prod(\alpha - e_i)\}$. Thus there is a unique ξ modulo Γ (where Γ is the lattice corresponding to $zy^2 = 4\prod(x - e_i z)$) such that $\wp(\xi) = \alpha$ *and* $\wp'(\xi) = \beta$.

We can now substitute $z = \wp(u)$ in the integral and get

$$\int_\alpha^\infty \{\prod(z - e_i)\}^{-1/2} dz = \int_\xi^0 \frac{d\wp(u)}{\wp'(u)} = \int_\xi^0 du = -\xi.$$

Thus, our integral has the value $-\xi$, where ξ is the unique point mod Γ where $\wp(\xi) = \alpha$ and $\wp'(\xi) = \beta$. But recall that to determine $\wp$, we must know Γ, *i.e.*, two generators of Γ, which are in turn similar definite integrals! Thus we can say that if we know two of these definite integrals, we know all elliptic integrals in Weierstrass form (in principle).

We briefly discuss elliptic integrals of the form $\int R(x, y)dx$, where

$$y^2 = Q(x) = \prod_{i=1}^4 (x - e_i)$$

(the e_i being distinct). It can be verified that, if we make the change of variables

$$x = \frac{1}{x_1} + e_1, y = -\frac{y_1}{x_1^2},$$

then the relation $y^2 = Q(x)$ goes over into a relation $y_1^2 = C(x_1)$, where C is now a cubic with distinct roots. It is easy to bring C to Weierstrass form by a linear change of the variables. In conclusion, the equation $y^2 = Q(x)$ can be brought to Weierstrass form by a change of variables of the type

$$x = \frac{ax_1 + b}{cx_1 + d}, y = \frac{ad - bc}{(cx_1 + d)^2} y_1$$

where $ad - bc \neq 0$. Under this change of variable, the integral $\int R(x, y)dx$ goes over into an integral $\int R_1(x_1, y_1)dx_1$ in Weierstrass form.

Example 5.3.2: If $Q(x)$ is a quartic with distinct roots, there is a change of variable $x = \frac{ax_1+b}{cx_1+d}$ such that $\int Q(x)^{-1/2}dx = \int C(x_1)^{-1/2}dx_1$ for a cubic C in Weierstrass form (*i.e.*, no change of the y-variable is needed).

Chapter 6

Moduli

6.1 A fundamental region for $SL(2, \mathbf{Z})$

Given a lattice Γ we can find a basis w_1, w_1' of Γ satisfying the following conditions

(i) $\Im \frac{w_1'}{w_1} > 0$

(ii) $\mid w_1 \mid$ has the minimal value among the non-zero elements of Γ and

(iii) $\mid w_1' \mid$ has the minimal value among elements in $\Gamma - \mathbf{Z}w_1$.

We call such a pair (w_1, w_1') of elements of Γ *reduced*. We have the following basic result.

Proposition 6.1.1 *A pair (w, w') of elements of Γ is reduced iff $\tau = \frac{w'}{w} = x + \iota y$ satisfies the inequalities*

$$x^2 + y^2 \geq 1, \quad y > 0, \quad -1/2 \leq x \leq 1/2.$$

Proof: Suppose the pair (w, w') is reduced. Then $\mid w \pm w' \mid \geq \mid m' \mid \geq \mid w \mid$. Hence, $(x \pm 1)^2 + y^2 = \mid \tau \pm 1 \mid^2 \geq \mid \tau \mid^2 = x^2 + y^2 \geq 1$. From $\pm 2x + 1 \geq 0$, we get the required inequalities.

Suppose conversely that the inequalities are satisfied. Let $w_1 = mw + nw'$ be any non-zero element of Γ. If $n = 0$, then clearly $\mid w_1 \mid \geq \mid w \mid$. Let $n \neq 0$, then we have

$$\delta = \mid \frac{w_1}{w} \mid^2 - \mid \frac{w_1'}{w} \mid^2 = \mid m + n\tau \mid^2 - \mid \tau \mid^2 = (m + nx)^2 - x^2 + (n^2 - 1)y^2$$

If $n \neq \pm 1, n^2 - 1 \geq 3$ and

$$y^2 = x^2 + y^2 - x^2 \geq 1 - 1/4 = 3/4.$$

This implies that $\delta \geq -(1/4) + (9/4) = 2 > 0$, so that $\mid w_1 \mid > \mid w' \mid$.

For $n = \pm 1$ and any integer m,

$$\delta = (m \pm x)^2 - x^2 \geq 0 \text{ because } \mid x \mid \leq 1/2.$$

This shows that if w_1/w is not real, then $\mid w_1 \mid \geq \mid w' \mid$. □

Next we will sharpen the result of the previous proposition. Let $SL(2, \mathbf{Z})$ act on the upper half plane in the usual way *viz.*,
$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ sends τ to $(a\tau + b)/(c\tau + d)$. Define

$$R = \{\tau \mid \Im\tau > 0, x^2 + y^2 \geq 1, -1/2 \leq x \leq 1/2, \text{ where } \tau = x + \iota y\}.$$

Lemma 6.1.2 *For every $\tau \in R$, there is a unique $\tau_1 = \frac{a\tau+b}{c\tau+d} \in R$ with $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbf{Z})$ and satisfying the conditions:*

(i) $x_1 < 1/2$ and

(ii) if $x_1^2 + y_1^2 = 1$, then $x_1 \leq 0$,

where $\tau_1 = x_1 + \iota y_1$.

Proof: Write $\tau = x + \iota y$. If $x = 1/2$, then $\tau - 1 = (x - 1) + \iota y = -1/2 + \iota y$. Then $\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$ maps τ to $\tau - 1$, which satisfies (i) and (ii).

If $x^2 + y^2 = 1$ and $x > 0$, then $-1/\tau = -1/\tau - 0$ is the image of τ under $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

Now we prove the uniqueness of τ_1 satisfying the conditions (i), (ii). Assume therefore that $\tau \in R$ satisfies the conditions (i) and (ii). Then the pair $(1, \tau)$ is reduced for the lattice $\Gamma = \mathbf{Z} + \mathbf{Z}\tau$. Write $w_1 = c\tau + d, w_1' = a\tau + b$ where $\tau_1 = w_1'/w_1 \in R$ and τ_1 satisfies (i), (ii).

Figure 6.1: The fundamental domain R^0 for $SL(2,\mathbf{Z})$ acting on $\mathcal{H}$

By Proposition 6.1.1, (w_1, w_1') is a reduced basis for Γ. Then $\mid w_1 \mid = 1$ and $ad - bc = 1$. If $c = 0$, then $a = d = \pm + 1$ implies $\tau_1 = w_1'/w_1 = \tau + b$. By the assumption on τ and $\tau_1, b = 0$ and $\tau_1 = \tau$.

Suppose now $c \neq 0$. Again $\mid w_1 \mid = 1$,

$$\mid w_1 \mid^2 - \mid \tau \mid^2 = \mid c\tau + d \mid^2 - \mid \tau \mid^2 = (cx + d)^2 - x^2 + (c^2 - 1)y^2.$$

We claim that $c = \pm 1$. For otherwise, as in the proof of Proposition 6.1.1, $\mid w_1 \mid > \mid \tau \mid^2 \geq 1 = \mid w_1 \mid$, a contradiction. Using $\mid c \mid = 1$, we see that

$$\mid w_1 \mid^2 - \mid \tau \mid^2 = d(\pm 2x + d) \geq 0 \text{ as } \mid x \mid \leq 1/2.$$

This gives $\mid w_1 \mid = 1 = \mid \tau \mid, x \leq 0, d = 0\ \pm 1$. If $d = 0$, then $b = \pm 1, w_1 = \pm\tau, w_1' = a\tau \mp\ 1, \tau_1 = \pm a - \frac{1}{\tau}$. This either $a = 0$ and $\tau_1 = \tau = \iota$, or $a = \mp 1$ and $x = -\frac{1}{2}, \tau_1 = \tau = e^{2\pi\iota/3}$. On the other hand, if $d = \pm 1$, then $a - b = \pm 1$, $x = -1/2$, $\tau = e^{2\pi\iota/3}$, $1/t = -\tau - 1$, $\tau_1 = \pm(a\tau + b)/(\tau + 1) = \pm a + \tau$, hence $a = 0$, $\tau_1 = \tau$.

This completes the proof of the Lemma. □

Denote the subset of R of points satisfying (i) and (ii) of Lemma 6.1.2 by R^0.

The picture for the set R^0 (called a *fundamental domain* for $SL(2,\mathbf{Z})$) is the shaded area drawn in figure 6.1.

6.2 Behaviour of the j function in R^0

We have constructed above (see Lemma 6.1.2) the fundamental domain R^0 of $SL(2,\mathbf{Z})$,

$R^0 = \{\tau \in \mathbf{C} \mid \Im\tau > 0, \mid \tau \mid \geq 1, x < 1/2$ and if
$\mid \tau \mid = 1$ then $x \leq 0$, where $\tau = x + \iota y\}$.

As a consequence of previous results, we now know that every τ with $\Im\tau > 0$ has a unique $SL(2,\mathbf{Z})$ translate in R^0 and for any $\lambda \in \mathbf{C}$, there is a unique point $\tau \in R^0$ such that $j(\tau) = \lambda$ (by Theorem 5.2.1 and Lemma 4.3.1).

The subset of R^0 with $\Im\tau > K$ (K a fixed positive number > 1) is mapped in a 1-1 manner on the punctured disc $\{0 <\mid t \mid< e^{-2\pi K}\}$ by the function $\tau \to e^{2\pi\iota\tau}$. In order to study the behvaiour of j in the subset of R^0 with $\Im\tau > K$, we put $t = e^{2\pi\iota\tau}$ and express $j(\tau)$ in terms of t. We will now state some of the properties of $j(\tau)$ for $\tau \in R^0$.

1. j takes every complex value exactly once in R^0
2. $j(e^{\pi\iota/3}) = j(e^{2\pi\iota/3}) = 0$
3. $j(\iota) = 1$
4. j is real on the boundary of R_0 and on the imaginary axis. There are no further points of R at which $j(\tau)$ is real.
5. Writing $j(\tau) = u + \iota v$ in terms of real and imaginary parts, $v > 0$ if $-1/2 < x < 0$ and $x^2 + y^2 > 1$. Similarly, $v < 0$ if $0 < x < 1/2$ and $x^2 + y^2 > 1$. Here $\tau = x + \iota y$.
6. Writing $t = e^{2\pi\iota\tau}$ and expressing j as a function of t, we have

$$j(\tau) = \frac{1}{1728t} + c_0 + c_1 t + \cdots$$

   The j function thus has a simple pole at ∞.

For proofs of these results, see L. Ford's classic book, Automorphic Functions, Chapter 7 (see [22]) .

6.3 The Riemann surface $\mathcal{H}/PSL(2,\mathbf{Z})$

Let $\mathcal{H} = \{\tau \in \mathbf{C} \mid \Im\tau > 0\}$ be the upper half plane. Denote the group $SL(2,\mathbf{Z})/\{\pm I\}$ by $PSL(2,\mathbf{Z})$. We will now sketch a proof of the following theorem.

Theorem 6.3.1 *The action of $SL(2,\mathbf{Z})$ on $\mathcal{H}$ is properly discontinuous and the quotient Riemann surface is biholomorphic to the complex plane $\mathbf{C}$.*

Proof: Consider again the fundamental region R_0 of figure 6.1. We make a series of elementary observations about the action of $PSL(2,\mathbf{Z})$

1. The isotropy subgroup of the point B on the boundary of R^0 in $PSL(2,\mathbf{Z})$ has order 3 and it is generated by $U = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}$.

2. The isotropy subgroup of ι is generated by $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and has order 2 in $PSL(2,\mathbf{Z})$. This subgroup maps the arc of the unit circle $\overset{\frown}{\iota B}$ onto the arc $\overset{\frown}{\iota A}$.

3. The element $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ maps the boundary $x = -1/2$ of R^0 onto the boundary $x = 1/2$.

4. The isotropy subgroup of the point A is $T^{-1}UT$.

5. Using the transformations $T, U, \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, we see that there exist finitely many elements $g_1, g_2, \cdots, g_n$ in $PSL(2,\mathbf{Z})$ such that the union of the translates $g_1R^0, g_2R^0, \cdots, g_nR^0$ contains every point of $R = \overline{R^0}$ as an interior point.

   Now let $K \subset \mathcal{H}$ be any compact subset. We claim K is covered by finitely many $PSL(2,\mathbf{Z})$-translates of R^0. For any $\tau \in K$, there exists $g \in PSL(2,\mathbf{Z})$ such that $g\tau \in R^0$. Then one of the sets $g^{-1}g_1R^0, \ldots, g^{-1}g_nR^0$ contains τ as an interior point. By compactness of K, we can find $g_{n+1}, \cdots, g_{n+\ell}$ such that

   $$\cup_{j=1}^{n} \cup_{i=1}^{\ell} g_{n+i}^{-1} g_j R^0$$

   contains K in its interior.

   By Lemma 6.1.2, any two translates gR^0, $g'R^0$ are pairwise disjoint or equal, and the latter only if $g = g'$. Now if $K_1, K_2 \subset \mathcal{H}$ are compact sets, with

   $$K_1 \subset \cup_{j=1}^{r} g_j R^0, \quad K_2 \subset \cup_{j'=1}^{s} g_{j'} R^0,$$

and $h \in PSL(2,\mathbf{Z})$ is such that $K_1 \cap hK_2 \neq \phi$, then $h = g_j(g_{j'})^{-1}$ for some j, j'. Thus the set of such h is finite. This shows that the action is properly discontinuous.

From Lemma 6.1.2 it follows that the only points in $\mathcal{H}$ with a non-trivial isotropy group in $PSL(2,\mathbf{Z})$ are the points A, ι and their translates. The map $\mathcal{H} \to \mathcal{H}/PSL(2,\mathbf{Z})$ is a local biholomorphic map at every point of $\mathcal{H}$ with trivial isotropy group. Let S be the set of points in $\mathcal{H}$ with a non-trivial isotropy group.

Consider the map $j : \mathcal{H} - S \to \mathbf{C}$. The image misses only 2 values *viz.*, $j(A) = 0$ and $j(\iota) = 1$.

Since

$$j\left(\frac{a\tau + b}{c\tau + d}\right) = j(\tau)$$

for $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbf{Z})$ and since the analytic map $(\mathcal{H} - S) \to (\mathcal{H} - S)/PSL(2,\mathbf{Z})$ is locally biholomorphic, we get an induced map

$$\overline{j} : (\mathcal{H} - S)/PSL(2,\mathbf{Z}) \to \mathbf{C} - \{0,1\}.$$

Let $\varphi : \mathcal{H} \to \mathcal{H}/PSL(2,\mathbf{Z})$ be the quotient map. Since j is bounded in neighborhoods of A and ι, we see that $\overline{j}$ is bounded in punctured neighborhoods of $\varphi(A)$ and $\varphi(\iota)$, hence we get an extension $\mathcal{H}/PSL(2,\mathbf{Z}) \to \mathbf{C}$ which is also analytic and the map is surjective. Since $j : R^0 \to \mathbf{C}$ and $R^0 \to \mathcal{H}/PSL(2,\mathbf{Z})$ are both 1-1, we see finally that $\mathcal{H}/PSL(2,\mathbf{Z}) \to \mathbf{C}$ is a biholomorphic map.

The proof of the Theorem 6.3.1 is therefore complete. □

Remark: The above theorem states that there is a 1-1 correspondence between biholomorphic equivalence classes of tori and points of the complex plane $\mathbf{C}$. This 1-1 correspondence is given by the j-invariant. In fact, more is true: one has the notion of a family of complex tori depending holomorphically on a complex parameter; then the mapping from the parameter space to $\mathbf{C}$ given by the j-function is a *holomorphic* mapping. One expresses this fact by saying that $\mathbf{C}$ is the *moduli space* of complex tori.

6.4 Biholomorphic maps of $\mathbf{C}/\Gamma$

In this section we will describe the group of biholomorphic maps of any $\mathbf{C}/\Gamma$.

Let $\Gamma = \mathbf{Z}w_1 + \mathbf{Z}w_2$, where $\tau = w_2/w_1$ lies in $\mathcal{H}$. Let $\varphi : \mathbf{C}/\Gamma \to \mathbf{C}/\Gamma$ be any biholomorphic map. As in Proposition 4.1.1, we have a lifting of φ to the universal cover with a commutative diagram

$$\begin{array}{ccc} \mathbf{C} & \stackrel{\widetilde{\varphi}}{\to} & \mathbf{C} \\ \downarrow & & \downarrow \\ \mathbf{C}/\Gamma & \stackrel{\varphi}{\to} & \mathbf{C}/\Gamma \end{array}$$

Then $\widetilde{\varphi}(z) = \alpha z + \beta$ for some $\alpha, \beta \in \mathbf{C}$ with $\alpha \neq 0$.

It follows that $\varphi(\overline{0}) = \beta\,((\mathrm{mod}\Gamma))$ where $\overline{0}$ denotes the image of 0 in $\mathbf{C}/\Gamma$. If we define the translation $\psi : \mathbf{C}/\Gamma \to \mathbf{C}/\Gamma$ by $\psi(z\,((\mathrm{mod}\Gamma))) = z - \beta\,((\mathrm{mod}\Gamma))$ then $\psi \circ \varphi$ is also a biholomorphic map $\mathbf{C}/\Gamma \to \mathbf{C}/\Gamma$ with the lift $\widetilde{\psi} \circ \widetilde{\varphi}$ from $\mathbf{C} \to \mathbf{C}$, where $\widetilde{\psi}(z) = z - \beta$. Now

$$\widetilde{\psi} \circ \widetilde{\varphi}(z)\,(\quad (\mathrm{mod}\ \Gamma)) = \alpha z\,(\quad (\mathrm{mod}\ \Gamma)).$$

It follows that $\alpha\Gamma = \Gamma$. Then $\alpha w_2 = aw_2 + bw_1, \alpha w_1 = cw_2 + dw_1$ where a, b, c, d are integers with $ad - bc = \pm 1$. Now $\alpha\tau = a\tau + b$ and $\alpha = c\tau + d$, hence $\tau = (a\tau + b)/(c\tau + d)$. But then by Lemma 4.2.2, $ad - bc = 1$.

The equation for τ is

$$c\tau^2 + \tau(d - a) - b = 0. \qquad (*)$$

We make two cases.

Case 1. $(*)$ is satisfied by every $\tau \in \mathcal{H}$. Then $b = c = 0, a = d = \pm 1$. In this case $\widetilde{\varphi}(z) = z + \beta$ or $\widetilde{\varphi}(z) = -z + \beta$. Any such translation of $\mathbf{C}$ clearly induces a biholomorphic map of $\mathbf{C}/\Gamma$.

Case 2. If (*) is not identically satisfied by all $\tau \in \mathcal{H}$, then $c \neq 0$ since τ is not real. Then τ and $\alpha = c\tau + d$ belong to the same quadratic extension field of $\mathbf{Q}$. The minimal equation of α over $\mathbf{Q}$ is

$$\alpha^2 - (a + d)\alpha + 1 = 0.$$

Since α is not real, $(a + d)^2 < 4$. If $a + d = 0$, then $\alpha = \pm\iota$. If $a + d = 1, \alpha$ is a primitive root of the equation $T^3 + 1 = 0$. If $a + d = -1, \alpha$ is a primitive

root of $T^3 - 1 = 0$. By Lemma 6.1.2, we can assume that $\tau \in R^0$. If $\alpha = \pm\iota$, then $cx + d = 0, cy = \pm 1$, where $\tau = x + \iota y$. Then $c = \pm 1, d = 0$ and $\tau = \iota$. If $\alpha = \pm e^{2\pi\iota/3}$ or $\alpha = \pm e^{\pi\iota/3}$, then $cx + d = \pm 1/2, cy = \pm\sqrt{3}/2$ *i.e.*, $c = \pm 1, \tau = e^{2\pi\iota/3}$.

If $\tau = \iota$, then we have

$$g_3 = \sum{}' \frac{1}{(mw_1 + nw_2)^6} = \sum{}' \frac{1}{(\alpha m w_1 + \alpha n w_2)^6}$$

$$= \frac{1}{\alpha^6} g_3 = -g_3, g_3 = 0,$$

so that $j(\tau) = 1$. In the case $\tau = e^{2\pi\iota/3}$ we have similarly

$$g_2 = \sum{}' \frac{1}{(mw_1 + nw_2)^4} = \frac{1}{\alpha^4} g_2 = e^{\pm 2\pi\iota/3} g_2 = 0,$$

so that $j(\tau) = 0$.

In both cases, the birational transformation $z \to \alpha z$ gives $\wp(\alpha z) = \alpha^{-2}\wp(z)$ and $\wp'(z) = \alpha^{-3}\wp'(z)$, so that the equation $\wp'(z)^2 = 4\wp(z)^3 - g_2\wp(z) - g_3$ is invariant:

$$(\alpha^4 g_2(\Gamma) = g_2(\Gamma), \alpha^6 g_3(\Gamma) = g_3(\Gamma)).$$

Then for any $\beta \in \mathbf{C}, z \to \alpha z + \beta$ give all the biholomorphic maps $\mathbf{C}/\Gamma \to \mathbf{C}/\Gamma$.

To summarize, any self biholomorphic map of $\mathbf{C}/\Gamma$ is induced from the linear map $z \to \alpha z + \beta$ of $\mathbf{C}$, where β is arbitrary and α is a 6^{th} root of unity in case $j(\tau) = 0$, a fourth root of unity in case $j = 1$ and a second root of unity otherwise.

Part II

Algebraic Geometry of Elliptic Curves

Chapter 7

Some results from commutative algebra

7.1 Noetherian rings

We will briefly review some results from Commutative Algebra. Unless explicitly mentioned otherwise, all rings considered will be assumed to be *commutative*, with a multiplicative unit element 1; all modules M will be such that $1 \cdot m = m$ for all $m \in M$. For a more detailed developement, the reader may refer to [34].

Recall that a *principal ideal domain* (denoted PID in the sequel) is an integral domain R where all ideals are principal, *i.e.*, of the form aR, for some $a \in R$. The ring of integers $\mathbf{Z}$ is a principal ideal domain, for example. Recall that a unit in a ring R is an element which has a multiplicative inverse. An element $a \in R$ is called *irreducible* if a is not a unit, and in any factorisation $a = bc$ in R, either b or c is a unit. If $R = K[t]$, the ring of polynomials over a field K, then the irreducible elements are precisely the irreducible polynomials. Finally, we recall that for any ring R, an R-module M is called a *free module* if there exists a subset $\{x_\alpha\} \subset M$ such that any $x \in M$ is uniquely expressible as a finite R-linear combination $x = \sum_\alpha a_\alpha x_\alpha$. Equivalently, $M \cong \oplus_\alpha R x_\alpha$, where $\lambda x_\alpha = 0 \Rightarrow \lambda = 0 \forall \alpha$. The set $\{x_\alpha\}$ is called a *free basis* (or just a *basis*) of M. If M is a free R-module with a finite free basis $\{x_1, x_2, \ldots, x_n\}$ then M is said to be *free of rank* n; then M is isomorphic to the R-module R^n of columns of length n with entries in R. Then the rank of a free R-module is well defined, *i.e.*, if $R^n \cong R^m$, then $m = n$. Indeed, if $R^n \cong R^m$,then for any maximal ideal $\mathcal{M} \subset R$, $(R/\mathcal{M})^m \cong (R/\mathcal{M})^n$ as

vector spaces over $R/\mathcal{M}$, so that $m = n$.

Recall that an ideal I in a ring R is *prime* if $I \neq R$, and for any $a, b \in R$ with $ab \in I$, either $a \in I$ or $B \in I$. Equivalently, R/I is an integral domain. An ideal $I \subset R$ is *maximal* precisely when R/I is a field; in particular, maximal ideals are prime.

We start with a useful result which generalizes the classical Chinese Remainder Theorem on solutions of a system of congruences.

Theorem 7.1.1 (Chinese Remainder Theorem) *Let R be a ring, $I, J \subset R$ ideals such that the ideal generated by I and J is the unit ideal, i.e., $I+J = R$. Then there is a natural isomorphism of rings*

$$R/I \cap J \cong R/I \times R/J.$$

If $I \cap J = IJ$, then this may be re-written as $R/IJ \cong R/I \times R/J$.

Proof: Since $I + J = R$, there exist $a \in I$, $b \in J$ with $a + b = 1$. The inclusions of ideals $I \cap J \subset I$ and $I \cap J \subset J$ induce surjections

$$f : R/I \cap J \twoheadrightarrow R/I, \quad g : R/I \cap J \twoheadrightarrow R/J,$$

and hence a product map

$$f \times g : R/I \cap J \to R/I \times R/J.$$

We claim $f \times g$ is an isomorphism. Clearly it is injective, since $\ker f \times g = \ker f \cap \ker g$ (this works for arbitrary ideals I, J). On the other hand, suppose $(x + I, y + J) \in R/I \times R/J$, where $x, y \in R$ are coset representatives. Let $z = bx+ay$; then $z-x = (b-1)x+ay = a(y-x) \in I$, so $f(z+I\cap J) = z+I = x+I$, and $z-y = bx+(a-1)y = b(x-y) \in J$, so $g(z+I\cap J) = z+J = y+J$. Hence $f \times g$ is surjective. □

Exercise 7.1.1: Prove the Chinese Remainder Theorem for n ideals — if $I_1, I_2, \ldots, I_n \subset R$ are ideals in a commutative ring R (with 1), and $I_j+I_k = R$ for all $j < k$, then the natural map

$$R/\cap_j I_j \to \prod_{j=1}^{n} R/I_j$$

is an isomorphism.
(Hint: To prove the natural map is surjective, it is enough to show (justify this) that for any $x_1, \ldots, x_n \in R$, there exists $x \in R$ such that $x - x_j \in I_j$. If $y_j + z_j = 1$ where $y_j \in I_1, z_j \in I_j$, then show that $1 = \prod_{j>1}(y_j + z_j) = u_1 + v_1$ where $u_1 \in I_1$, $v_1 \in \cap_{j>1} I_j$. Similarly construct $u_j \in I_j$, $v_j \in \cap_{k \neq j} I_k$ with $u_j + v_j = 1$. Now let $x = \sum_j v_j x_j$; show that $x - x_j \in I_j$.)

Most rings that appear in Algebraic Geometry are Noetherian. So we review some of the theory of Noetherian rings here. You will find a more exhaustive treatment of the theory in any text on Algebra. *e.g.*, [34].

Definition: Let A be a ring an M an A -module. Then M is a *Noetherian module* over A if any A -submodule of M is finitely generated.

We will sometimes use the term 'finite module' to mean 'finitely generated module'.

Proposition 7.1.1 *Let A be a ring and M an A-module. Then the following are equivalent:*

1. *M is Noetherian.*
2. *Any ascendinging chain of submodules of M is a finite chain.*
3. *Any collection $\mathcal{F}$ of submodules of M has a maximal element.*

Proof: 1) $\Rightarrow$ 2). Let $M_i \subset M_{i+1}$ be a chain of submodules of M. Let $N = \cup M_i$. Then N is a submodule of M and hence finitely generated. If m_k form a finite subset generating N, then $m_k \in M_{i(k)}$ for some $i(k)$ and hence if we choose $n = \max\{i(k)\}$ then $m_k \in M_n$. Thus $N = M_n$ and hence $M_i = M_n$ for all $i \geq n$.

2) $\Rightarrow$ 3). Let $M_0 \in \mathcal{F}$. If it is not maximal, then there exists $M_1 \in \mathcal{F}$ such that M_0 is a proper submodule of M_1. We may continue this to get an ascending chain which stops by 2) and hence $\mathcal{F}$ must have a maximal element.

3) $\Rightarrow$ 1). Let $N \subset M$, a submodule. Let $m_1 \in N$. If $M_1 = Am_1 \neq N$ then we may choose $m_2 \in N$ and $\notin M_1$. If $M_2 = Am_1 + Am_2 \neq N$ then we may choose $m_3 \in N$ and $\notin M_2$. Continuing like this we get an ascending chain and this set of submodules has a maximal element say M_n. Then it is clear that N is n -generated. $\square$

Lemma 7.1.2 *If M is a Noetherian module then any submodule and quotient module of M is Noetherian.*

Proof: If $N \subset M$ is a submodule and $\mathcal{F}$ a collection of submodules of N, then it is also a collection of submodules of M and hence has a maximal element. Thus N is Noetherian. If $f : M \to N$ is a quotient and $\mathcal{F}$ is a collection of submodules of N, then the set $\{f^{-1}(K) \mid K \in \mathcal{F}\}$ is a collection of submodules of M and hence has a maximal element, say $f^{-1}(N_0)$. It is clear that N_0 is a maximal element of $\mathcal{F}$. □

Proposition 7.1.3 *Let M be any module and N a sumodule. If N and M/N are Noetherian, then so is M.*

Proof: If K is any submodule of M then $K \cap N \subset N$ and $K/(K \cap N) \cong (K+N)/N \subset M/N$ and since both these are finitely generated by hypothesis, one gets that K is finitely generated. □

Example 7.1.1:

1. Let A be a field. Then a vector space M over A is Noetherian if and only if it is finite dimensional.

2. Let $A = \mathbf{Z}$. Then any finitely generated A -module is Noetherian. This can be seen as follows: If M is finitely generated over A, then there exists a finitely generated *free* module $F = \oplus_1^n A$ which surjects onto M. So by lemma (7.1.2) it suffices to show that F is Noetherian and again by proposition (7.1.3) it suffices to show that A considered as a module over itself is Noetherian. But any submodule of A is just an ideal and hence any submodule of A is *principal* ($\mathbf{Z}$ is a *principal ideal domain*).

We will see below that the argument in case 2 above generalises.

Definition: A ring is called a *Noetherian ring* if considered as a module over itself, it is Noetherian.

So $\mathbf{Z}$ is a Noetherian ring. The argument in the above example says that if A is a Noetherian ring, then an A-module is Noetherian if and only if it is finitely generated. The followig propsition is easy from the definition:

Proposition 7.1.4 *If $A \twoheadrightarrow B$ is a surjective homomorphism of rings and A is Noetherian, then so is B.*

Theorem 7.1.2 (Hilbert basis theorem) *Let A be a Noetherian ring. Then $B = A[X]$, the polynomial ring in one variable over A, is Noetherian.*

Proof: Let $I \subset B$ be any ideal. We define an ascending chain of ideals J_k in A as follows. Let J_k be the ideal generated by leading coefficients of polynomials of degree k in I,

$$J_k = \{a \in A \mid \exists \text{ a polynomial } aX^k + a_1X^{k-1} + \ldots a_k \in I\}$$

Since A is Noetherian, there exists a d such that $J_d = J_{d+1} = \ldots$. Let a_{ij} be a finite set of generators for J_j and let $f_{ij} \in I$ be a polynomial of degree j with leading coefficient a_{ij}. We will see that the finite set f_{ij} with $j \leq d$ generate I.

Let $f \in I$ be of degree n. We will prove the claim by induction on n. If $\deg f = 0$, then $f \in J_0$ and hence f is an A-linear combination of f_{i0}'s. Assume we have proved the claim for all smaller degrees. If $n \leq d$ and the leading coefficient of f is a, then $a \in J_n$ and hence $a = \sum c_i a_{in}$. Thus $f' = f - \sum c_i f_{in} \in I$ and $\deg f' < \deg f$. So by induction we would be done. If $n > d$ and a is the leading coefficient of f, since $J_d = J_n$ we get $c_i \in A$ such that $a = \sum c_i a_{id}$. Then $f' = f - \sum c_i X^{n-d} f_{id}$ has degree less than n and $f' \in I$. Again by induction we are done. □

Example 7.1.2: If k is a field and A is a finitely generated algebra over k, then A is Noetherian. Of course we may replace k by any Noetherian ring. The Hilbert basis theorem gives us ways of constructing a large number of examples of Noetherian rings.

Now we come to an important technical concept, that of *associated primes*. Recall that if $m \in M$ where M is an A-module, then the *annihilator* of m is the ideal of A,

$$\text{Ann } m = \{x \in A \mid xm = 0\}.$$

Definition: Let M be an A-module. If $m \in M$ such that Ann $m = P$ a prime ideal of A, then we call P an *associated prime* of M. The set of associated primes of M is denoted by Ass M.

Example 7.1.3: If A is a domain and M is a torsion free module over A, then (0) is the only associated prime of M. If $A = \mathbf{Z}$ and $M = \mathbf{Z}/n\mathbf{Z}$ for some integer n, then the associated primes of M are precisely the prime ideals generated by the prime divisors of n. If A is any ring and $M = A/P$ for some prime ideal P, then Ass $M = \{P\}$.

Lemma 7.1.5 *Let A be a Noetherian ring and $M \neq 0$ a module. Then* Ass $M \neq \emptyset$. *Furthermore it contains all maximal elements of the set* $\{\text{Ann } m \mid 0 \neq m \in M\}$.

Proof: Consider the collection of ideals $\{\text{Ann } m \mid 0 \neq m \in M\}$. Since A is Noetherian this set has a maximal element, say P. We will show that P is a prime ideal and then P will be in Ass M. If not there exists $x, y \notin P$ but $xy \in P$. By choice there exists an $m \in M$ such that $P = \text{Ann } m$. Consider $m' = xm$. Since $x \notin P$, $m' \neq 0$. Clearly the annihilator of m' contains P and y and since $y \notin P$, we get a contradiction. □

Lemma 7.1.6 *Let A be any ring and M any module. If $N \subset M$ is any submodule, then* Ass $M \subset$ Ass $N \cup$ Ass (M/N).

Proof: Let $P \in$ Ass M and let $m \in M$ such that Ann $m = P$. Let $\overline{m}$ be the image of m in M/N. If Ann $\overline{m} = P$, then we are done. At any rate notice that $P \subset$ Ann $\overline{m}$. So if they are not equal there exists $x \notin P$ such that $m' = xm \in N$. Notice that $P \subset$ Ann m'. If $y \in$ Ann m' then $yx \in$ Ann $m = P$ and since $x \notin P$, $y \in P$. Thus we see that Ann $m' = P$ and hence $P \in$ Ass N. This finishes the proof. □

Lemma 7.1.7 *Let A be a Noetherian ring and M a finitely generated module. Then* Ass M *is finite.*

Proof: Consider the collection of all submodules N of M with Ass N finite. By lemma (7.1.5) there exists a submodule of M which is isomorphic to A/P for some prime ideal P and since Ass $A/P = \{P\}$, this set is non-empty and since M is Noetherian this set has a maximal element, say N. We want to show that $N = M$. If not, $M/N \neq 0$ and hence again by lemma (7.1.5), there exists an element $m' \in M/N$ with Ann $m' = P$, P a prime ideal. Let m be a lift of m' in M and let $K = N + Am$. Then $K/N = Am' \cong A/P$ and hence Ass $K/N = \{P\}$. By lemma (7.1.6) Ass $K \subset$ Ass $N \cup$ Ass K/N and hence it is finite contradicting maximality of N. □

Definition: Let M be an A -module. $x \in A$ is called a *zero divisor in* M if there exists a non-zero $m \in M$ such that $xm = 0$. The set of zero divisors in M is denoted by $Z(M)$. If $x \notin Z(M)$, it is called a non zero divisor.

Lemma 7.1.8 *Let A be a Noetherian ring and M an A-module. Then* $Z(M) = \cup\{P \mid P \in \text{Ass } M\}$.

Proof: If P is an associated prime, then there exists a non-zero $m \in M$ such that Ann $m = P$ and hence $Pm = 0$. Thus $\cup\{P \mid P \in \text{Ass } M\} \subset Z(M)$. Let $x \in Z(M)$. Then by definition there exists a non-zero $m \in M$ such that $x \in \text{Ann } m$ and by (7.1.5), $x \in P$ for some $P \in \text{Ass } M$. □

7.2 Localisation

Everyone is familiar with the concept of the quotient field of a domain. For example, that **Q**, the field of rational numbers is the quotient field of **Z**, the ring of integers. This section deals with a construction similar to the construction of quotient field, which is very useful.

Definition: Let A be a ring and $S \subset A$. S is called a *multiplicatively closed subset* of A if

(i) $0 \notin S$

(ii) $1 \in S$,

(iii) if $x, y \in S$ then so is their product xy.

Example 7.2.1:

1. Let $x \in A$ be a non-nilpotent element. Then $S = \{1, x, x^2, \ldots\}$ is a multiplicatively closed subset. This is called the *multiplicative subset generated by* x.

2. Let P be a prime ideal of A. Then $S = A - P$ is a multiplicatively closed subset.

3. Let $B = A[X]$, a polynomial ring in one variable over a ring A. Let S be the set of all monic polynomials of B (*i.e.*, all polynomials whose leading coefficient is 1). Then S is a multiplicatively closed subset.

Next we will construct the *localisation* of a ring A at a given multiplictively closed subset S. We proceed as follows: Take the set $A \times S$ and consider the relation $(a, s) \sim (b, t)$ if there exists a $u \in S$ such that $u(at - bs) = 0$. One easily checks that this is an equivalence relation. Let the set of equivalence classes be denoted by $S^{-1}A$, the *localisation* of A at S. Now define addition and multiplication on $S^{-1}A$ as follows:

We will write (a, s) interchangeably for an element in $A \times S$ as well as the equivalence class in $S^{-1}A$ containing it.

1. $(a, s) + (b, t) = (at + bs, st)$
2. $(a, s)(b, t) = (ab, st)$

One checks that these are well defined operations and these make $S^{-1}A$ into a ring. We get a natural map $f : A \to S^{-1}A$ by defining $f(a) = (a, 1)$ for any $a \in A$. One sees that it is well defined and it is an algebra homomorphism. Also elements of S go to units under the map.

Example 7.2.2:
1. Let $A = \mathbf{Z}$ and S be the multiplicative set generated by a number $n \in \mathbf{Z} - \{0\}$. Then $S^{-1}A$, usually denoted by $\mathbf{Z}_n$, is the set of all rationals of the form $\frac{a}{n^k}$ for some integer k and $a \in \mathbf{Z}$.
2. A be as before and let p be a prime. Let $S = A - pA$. Then $S^{-1}A$ is the set of all rationals a/b such that gcd $(p, b) = 1$. This is usually denoted by $\mathbf{Z}_{(p)}$.
3. What happens if we take S to be the set of all non-zero integers?

Exercise 7.2.1: Let M be an A-module and S a multiplicatively closed subset. Consider the set $M \times S$ and define a relation on it by $(m, s) \sim (n, t)$ if there exists a $u \in S$ such that $u(tm - sn) = 0$.

(i) Show that this is an equivalence relation.

Denote by $S^{-1}M$ the set of equivalence classes. As before we will denote by (m, s) an element of $M \times S$ as well as its equivalence class in $S^{-1}M$. Define operations on $S^{-1}M$ as follows:

1. $(m, s) + (n, t) = (tm + sn, st)$
2. $a(m, s) = (am, s) \quad \forall a \in A.$

(ii) Show that these operations are well defined and with these operations $S^{-1}M$ is naturally a module over $S^{-1}A$.

Lemma 7.2.1 *Let A, S be as above and f the natural homomorphism constructed above. If $g : A \to R$ is any ring homomorphism such that $g(S) \subset R^*$, where R^* is the group of units of R, then there exists a unique ring homomorphism $h : S^{-1}A \to R$ such that $h \circ f = g$.*

Proof: First let us prove the uniqueness. If h, h' are two such and $(a, s) \in S^{-1}A$, since $f(s)(a, s) = (sa, s) = (a, 1) = f(a)$ we get

$$g(a) = h(f(a)) = h(f(s))h((a, s)) = g(s)h((a, s))$$

and similarly for h'. Thus for all (a, s), we have $g(s)h((a, s)) = g(s)h'((a, s))$ and since $g(s)$ is a unit, $h((a, s)) = h'((a, s))$.

Define $h((a, s)) = g(s)^{-1}f(a)$. One can see that this is well defined and it does define a ring homomorphism as required. □

Lemma 7.2.2 *Let A, S, f be as above. Then*

$$\ker f = \{a \in A \mid sa = 0 \text{ for some } s \in S\}$$

In particular, if S contains only non-zero divisors in A, then f is an inclusion of rings.

Proof: If $sa = 0$, then $f(a) = (a, 1) = (sa, s) = (0, s) = 0$ in $S^{-1}A$. Conversely, if $f(a) = 0$ then $(a, 1) = (0, 1)$ in $S^{-1}A$ and this by defintion of the equivalence relation implies that there exists an $s \in S$ such that $sa = 0$. □

Corollary 7.2.3 *If A is a domain, then for any multiplicatively closed subset S of A, $S^{-1}A$ is naturally a subring of its quotient field.*

Exercise 7.2.2: Let A be a ring, S a multiplicatively closed subset and f the natural map from A to $S^{-1}A$.

(i) Show that if P is a prime ideal of $S^{-1}A$, then $f^{-1}(P)$ is a prime ideal of A.

(ii) if J is any ideal of $S^{-1}A$, then J is generated by $f(f^{-1}(J))$.

Recall that a ring is *reduced* if it has no nilpotents.

Corollary 7.2.4 *If A is a reduced ring then the intersection of its minimal prime ideals is zero.*

Proof: If not let $x \neq 0$ be in the intersection. Then $S = \{1, x, x^2, \ldots\}$ is a multiplicatively closed set, since x is not nilpotent. Let Q' be a minimal prime ideal of $S^{-1}A$ and $Q = f^{-1}Q$, where f is the natural map from A to $S^{-1}A$. Then by the exercise above, Q is a prime and hence must contain a minimal prime ideal and hence the element x. Then $f(x) \in Q'$, but $f(x)$ is a unit in $S^{-1}A$ and hence cannot be in a proper ideal. □

For us, a case of localisation which is of particular interest would be when $S = A - P$ for some prime ideal P.

Definition: A ring is said to be a *local ring* if it has a unique maximal ideal. A ring with finitely many maximal ideals is called a *semilocal ring.*

The following lemma tells us how to construct local rings.

Lemma 7.2.5 *Let A be a ring and P a prime ideal. Let $S = A - P$. Then $S^{-1}A$ which is usually denoted by A_P is a local ring with the maximal ideal $S^{-1}P$, which by definition is the set of equivalence classes of the form (p, s), $p \in P$, $s \in S$. (Check that this set is indeed an ideal in A_P.)*

The proof is left as an exercise.

Lemma 7.2.6 (Nakayama's lemma) *Let R be a Noetherian local ring with maximal ideal $\mathcal{M}$, and let P be a finite R-module with $\mathcal{M}P = P$. Then $P = 0$.*

Proof: By induction on the number of generators of P. If $P = Rx = \mathcal{M}x$, then $x = ax$ with $a \in \mathcal{M}$, so $(1 - a)x = 0$, where $1 - a \in R^*$, so $x = 0$. If P has generators $x_1, \ldots, x_n$, then $x_1 = \sum_i a_i x_i$ with $a_i \in \mathcal{M}$, so $(1 - a_1)x_1 = \sum_{i>1} a_i x_i$, where $(1 - a_1) \in R^*$. Hence P is generated by $x_2, \ldots, x_n$. □

Corollary 7.2.7 *Let $R, \mathcal{M}$ be as above, and $Q \subset P$ an inclusion of R-modules, where P is finitely generated. Suppose $P = Q + \mathcal{M}P$. Then $P = Q$.*

Lemma 7.2.8 *If A is Noetherian and S a multiplicatively closed subset, then $S^{-1}A$ is Noetherian.*

Proof: Let $f : A \to S^{-1}A$ be the natural map and I_k an ascending chain of ideals of $S^{-1}A$. Then $f^{-1}I_k = J_k$ is an ascending chain of ideals in A and hence must be finite. *i.e.*, $J_k = J_{k+1} = \ldots$. One easily sees that $I_k = I_{k+1} = \ldots$. (see Exercise (7.2.2)). □

Lemma 7.2.9 *Let P be a minimal prime ideal of a Noetherian ring A. Then $P \in \text{Ass } A$.*

Proof: Consider A_P the localisation. It is a local ring with the maximal ideal $S^{-1}P = \mathcal{P}$, say, where $S = A - P$. If Q is any prime ideal of A_P then $f^{-1}Q$ is a prime ideal (see exercise (7.2.2)) of A contained in P where f is the natural map from A to A_P. Since P is minimal we see immediately that $S^{-1}P$ is the *unique prime* ideal of A_P. Since A_P is Noetherian by the previous lemma, and since Ass $A_P \neq \emptyset$ by lemma (7.1.5), we see that $S^{-1}P$ is associated to A_P. Thus there exists a non-zero element $x \in A_P$ such that Ann $x = S^{-1}P$. There exists a $t \in S$ and a $y \in A$ such that $f(y) = tx$. Let $x_1, \ldots, x_n$ be a set of generators for P. Then $f(x_i y) = f(x_i)tx = 0$ for all i. Thus by (7.2.2), there exists $u_i \in S$ such that $u_i x_i y = 0$. So if we take $z = y \prod u_i$, then $x_i z = 0$ for all i and hence Ann $z \supset P$. If $a \in$ Ann z then one can easily see that $f(a) \in$ Ann x since elements of S become units in A_P. Thus $a \in P$ and hence $P \in \text{Ass } A$. □

Corollary 7.2.10 *If A is a Noetherian ring, then the set of minimal primes is a finite set.*

Corollary 7.2.11 *Let A be a Noetherian reduced ring. Then*

$$\text{Ass } A = \{P \mid P \text{ is a minimal prime}\}$$

Proof: We already know that all minimal prime ideals are associated. If $P \in \text{Ass } A$ and P is not minimal, then choose an element $0 \neq x \in A$ such that $P = \text{Ann } x$. Since $Px = 0$, $Px \subset Q$ for any minimal prime ideal Q. Since P is not contained in Q, we get that $x \in Q$. But by (7.2.4), this implies $x = 0$. □

Corollary 7.2.12 *If A is a Noetherian reduced ring, then the set of zero-divisors of A,*

$$Z(A) = \cup\{P \mid P \text{ is a minimal prime ideal}\}.$$

7.3 Integral extensions

In this section all rings will be assumed to be Noetherian unless otherwise specified. Many of the results do hold in greater generality though. We start with the following definitions.

Definition: Let $A \subset B$ be an extension of rings. An element $x \in B$ is said to be *integral over* A if x satisfies a monic polynomial with coefficients in A. The ring B is said to be *integral* over A if all elements of B are integral over A. We also say that B is an *integral extension* of A.

Example 7.3.1:

1. The extension $\mathbf{Z} \subset \mathbf{Z}[\sqrt{2}]$ is an integral extension.
2. Let $B = k[x_1, \ldots, x_n]$ be the polynomial ring in n-variables over a field k and A the subring generated by all symmetric polynomials in the x_i. Then $A \subset B$ is an integral extension.

Exercise 7.3.1: Let $A \subset B$ be an integral extension and let S be a multiplicatively closed subset of A. Show that $S^{-1}A \subset S^{-1}B$ is an integral extension.

Remark: If $A \subset B$ is an integral extension of *domains*, then the corresponding extension of their quotient fields is an algebraic extension.

We have the following easy lemma.

Lemma 7.3.1 *Let $A \subset B$ be a ring extension. An element $x \in B$ is integral over A if and only if $A[x]$, the subalgebra generated by x, is a finite A-module. Further if B is a finite A-module, then B is an integral extension of A.*

Proof: Let x satisfy the monic polynomial $X^n + a_1 X^{n-1} + \cdots + a_n = 0$, with $a_i \in A$. Then it is trivial to see that the elements $1, x, x^2, \ldots, x^{n-1}$ generate $A[x]$ as an A-module. Conversely, if $m_1, m_2, \ldots, m_k$ generate $A[x]$ as an A-module, write $m_i = \sum_{j=0}^{N} a_{ij} x^j$ with $a_{ij} \in A$. Then it is clear that $x^j, j = 0, \ldots, N$ also generate $A[x]$. So we may write $x^{N+1} = \sum_0^N r_j x^j$ with $r_j \in A$. This gives the required monic equation for x.

If B is a finite module over A and $x \in B$ then $A[x]$ is a submodule of B, and since A is Noetherian, we see that $A[x]$ is also a finite A-module. Thus by the first part of the theorem x is integral over A. $\square$

Definition: If $A \subset B$ an inclusion of rings and B is a finite module over A, we say that B is a *finite extension* of A.

Corollary 7.3.2 *Let $A \subset B$ be an inclusion of rings. Then the set of elements $R = \{x \in B \mid x \text{ is integral over } A\}$ is an A-subalgebra.*

Proof: Let $x, y \in R$ be any two elements. Since x is integral over A, by lemma (7.3.1) $A[x] = C$ is a finite A-module. Since y is integral over A, it is also integral over C. So $C[y]$ is a finite C-module. Thus $C[y]$ is also a finite A-module. So by the lemma again, $C[y] = A[x, y]$ is integral over A. □

Remark: Notice that the proof implies that if $B = A[x_1, x_2, \ldots, x_n]$ is integral over A, then B is a finite module over A.

We make one more definition.

Definition: Let $A \subset B$ be an extension of rings. Let R be the set of all elements in B which are integral over A. Then R is called the *integral closure* of A in B. If A is a domain and B is its quotient field then R is called *the* integral closure. In the above situation if $A = R$ then we say that A is *integrally closed*. In our context we may also use *normal domain* to denote the same.

One can show that, in general, R need not be Noetherian.

Example 7.3.2:

1. The integral closure of $\mathbf{Z}$ in $\mathbf{Q}(\sqrt{5})$ is $\mathbf{Z}[\frac{1+\sqrt{5}}{2}]$.
2. The integral closure of $k[t^2, t^3]$ is $k[t]$ where k is a field.

Exercise 7.3.2: Let $f(t) \in k[t]$ be a non-constant polynomial in t, where k is a field. Show that $k[f(t)] \subset k[t]$ is a finite extension.

Corollary 7.3.3 *Let $A \subset B \subset C$ be extensions of rings, and assume that B is integral over A and C is integral over B. Then C is integral over A.*

Proof: We need to show that if $x \in C$, then x is integral over A. We have an integral equation $x^n + b_1 x^{n-1} \cdots + b_n = 0$ where $b_i \in B$. Then b_i's are integral over A. So by the remark after corollary (7.3.2) $D = A[b_1, b_2, \ldots, b_n]$ is a finite module over A. Also x is integral over D and hence $D[x]$ is a finite module over D. Thus $D[x]$ is a finite A-module and hence x is integral over A. □

Lemma 7.3.4 *Let $A \subset B$ be an integral extension. Assume that A and B are domains. Then A is a field if and only if B is a field.*

Proof: Assume first that A is a field. Let $x \neq 0$ be an element of B. Write its integral equation, $x^n + a_1 x^{n-1} + \cdots + a_{n-1}x + a_n = 0$ where n is minimal. If $a_n = 0$ then since $x \neq 0$ and B is a domain, we could cancel an x in the above equation to get one with a smaller degree. So $a_n \neq 0$. Then $x(x^{n-1} + \cdots + a_{n-1}) = -a_n$. Since A is a field and $a_n \neq 0$, it is a unit, so that a multiple of x is a unit and hence so is x.

Conversely assume that B is a field. Let $x \in A$ be nonzero. Then $u = 1/x \in B$. Suppose u satisfies the integral equation $u^n + a_1 u^{n-1} + \cdots + a_n = 0$. Writing u as $1/x$ and multiplying the above equation by x^n we get $1 = x(-a_1 - \cdots - a_n x^{n-1})$ proving that x is a unit in A. □

Corollary 7.3.5 *Let $A \subset B$ be an integral extension. If $\mathcal{M} \subset B$ is a maximal ideal of B, then $\mathcal{M} \cap A$ is a maximal ideal of A.*

7.4 Noether normalisation, and the Nullstellensatz

Lemma 7.4.1 *Let $F \in k[X_1, \ldots, X_n]$ be a non-constant polynomial, where k is any field. Then by a change of variables, F can be made monic in X_n.*

Proof: We give a proof when k is infinite. The case of a finite field is left as an exercise (see any standard text in algebra, *e.g.*, [34]). In this case we will prove something stronger. We will show that for a 'generic' choice of variables $Y_i = \sum a_{ij} X_j$ where $a_{ij} \in k$ are sufficiently general, F is monic in Y_n. Write $F = F_0 + \cdots + F_d$ where F_i is homogeneous of deg i. Since changes as above preserve the decomposition into homogeneous components, we may as well assume that F is homogeneous, say of degree d. Let Y_i be indeterminates and write $X_i = Y_i + A_i Y_n$ where A_i's are indeterminates (formally). Writing $F = \sum c_{i_1 i_2 \ldots i_n} X_1^{i_1} X_2^{i_2} \ldots X_n^{i_n}$ we substitute the expression for X in terms of the Y. The coefficient of Y_n^d in this expression is $\sum c_{i_1 i_2 \ldots i_n} A_1^{i_1} A_2^{i_2} \ldots A_n^{i_n}$ which is clearly a non-zero polynomial in the A's. Since the field is infinite for 'most' values $A_i = a_i$ with $a_i \in k$, the value of this polynomial is non-zero (see Exercise (7.4.1)). So for these a_i the polynomial F is monic in Y_n. We

may of course also assume that $a_n \neq 0$ ensuring that $Y_i + a_i Y_n$ is indeed a change of variables. $\square$

Exercise 7.4.1: Let k be an infinite field, $f \in k[X_1, \ldots, X_n]$ a non-zero polynomial in the variables $X_1, \ldots, X_n$. Show that there exist infinitely many n-tuples $(a_1, \ldots, a_n) \in k^n$ such that $f(a_1, \ldots, a_n) \neq 0$.

Definition: A finitely generated algebra over a field is called an *affine algebra.*

Theorem 7.4.1 (Noether Normalization lemma) *Let A be an affine algebra over a field k. Then there exists elements $t_1, t_2, \ldots, t_d \in A$ which are algebraically independent over k such that A is integral over $R = k[t_1, t_2, \ldots, t_d]$.*

Proof: Let $A = k[x_1, x_2, \ldots x_n]$. If the x_i's are algebraically independent over k, then we may take $d = n$ and $t_i = x_i$. So assume that they are dependent. Then there exists a non-constant polynomial $F(X_1, \ldots, X_n)$ over k such that $F(x_1, \ldots x_n) = 0$. By the previous lemma, we may change coordinates so that F becomes monic in X_n. In other words we may assume that the relation between the x_i's is of the form $x_n^r + a_1 x_n^{r-1} + \cdots + a_r = 0$ where $a_i \in B = k[x_1, \ldots, x_{n-1}]$. So A is integral over B. By an obvious induction on n we may assume that there exists $t_1, \ldots t_d \in B$ such that they are algebraically independent and B is integral over $R = k[t_1, \ldots, t_d]$. But x_n is integral over B, which by Corollary (7.3.3) implies that A is integral over R. $\square$

Lemma 7.4.2 *Let $R_n = k[x_1, \ldots, x_n]$ be a polynomial ring over an algebraically closed field k. Then any maximal ideal is of the form $(x_1 - a_1, \ldots, x_n - a_n)$ for some $a_i \in k$.*

Proof: The proof is by induction on n. If $n = 1$, the fact that k is algebraically closed trivially implies the result, since any maximal ideal is generated by an irreducible polynomial.

So assume that $n > 1$ and let $\mathcal{M}$ be any maximal ideal. If we show that $R/\mathcal{M}$ is equal to k then one can easily see that we can take a_i's to be the images of the x_i's in this quotient field. Now we may change variables and assume that $\mathcal{M}$ contains a monic polynomial in x_n by lemma (7.4.1). If we denote by $\mathcal{N} = \mathcal{M} \cap k[x_1, \ldots, x_{n-1}]$, then $R_{n-1}/\mathcal{N} \subset R_n/\mathcal{M}$ is integral and

since the second one is a field, so is the first by lemma (7.3.4). Thus by induction we see that $R_{n-1}/\mathcal{N} = k$, and now by the one variable argument, so is $R_n/\mathcal{M}$. □

Next we want to prove a theorem which is crucial in Algebra and Geometry. We first start with a useful result on field extensions.

Theorem 7.4.2 *Let k be a perfect field and K a function field over k of transcendence degree d. Then there exists $x_1, \ldots, x_d \in K$, algebraically independent over k and K is a finite separable extension of $L = k(x_1, \ldots, x_d)$.*

Proof: Write $K = k(x_1, \ldots, x_n)$. If $n = d$, there is nothing to prove. So assume that $n > d$. Thus there is a polynomial f in n variables X_i, over k such that $f(x_i) = 0$. Choose such an f of smallest degree. Clearly f is irreducible. I claim that not all exponents of all variables occurring in f is divisible by p, the characteristic of the field. If it were divisible, since k is perfect, we may write $f = g^p$ for some polynomial g and thus f is not irreducible. Without loss of generality, we may assume that there is a term containing X_n^l in f and p does not divide l. Writing $f = a_0 X_n^N + a_1 X_n^{N-1} + \ldots + a_N$, with $a_i \in k[X_1, \ldots, X_{n-1}]$, we see easily that x_n is separable over $k(x_1, \ldots, x_{n-1})$. Now an easy induction finishes the proof. □

Theorem 7.4.3 *Let A be an affine domain over k, where k is a perfect field and let K be the quotient field of A. Let L be a finite extension of K and B the integral closure of A in L. Then B is a finite A-module. In particular B is an affine algebra.*

Proof: By Noether Normalisation lemma (7.4.1) , we may choose a polynomial ring $A' \subset A$ so that A is integral over A'. By lemma (7.3.3) the integral closure of A' in L is also B. So if we show that B is a finite A'-module we will be done (Notice that we have used Hilbert basis theorem (7.1.2).) So we may assume that A is a polynomial ring.

We will first prove the result when L is separable over K. Since L is finite over K, we may find $x_1, \ldots, x_n \in B$ which form a K-basis for L (Hint: clear denominators for algebraic equations over K). We have the K-linear trace map, $T : L \to K$. Thus we get a K-linear map

$$f : L \to K^n \quad x \mapsto (T(xx_1), \ldots, T(xx_n))$$

Using separability, one checks that this map is an isomorphism of vector spaces as follows. We only need to check that f is injective since both sides have dimension n over K. Let $x \neq 0$ be an element of L. Separability implies that there exists $y \in L$ such that $T(xy) \neq 0$ [Theorem 9,pp 211 [34]]. Write

$$y = \sum q_i x_i \quad q_i \in K$$

Thus

$$T(xy) = \sum q_i T(xx_i)$$

This implies that at least one of the $T(xx_i) \neq 0$.

Now let $x \in B$. Then $T(xx_i) \in K$ and it is integral over A. But since A is a UFD, it is integrally closed. So $T(xx_i) \in A$. Thus we see that $B \subset A^n$ and hence B is a finite A-module.

We are left with the case when L is not necessarily separable over K. We may prove the theorem using theorem (7.4.2), but we will give a slightly different proof below. Now let L be an arbitrary finite extension of K and assume it is not separable. Then the characteristic must be positive, say p. Let L_0 be the separable closure of K in L, so that $L_0 \subset L$ is purely inseparable, and let its degree be p^e. We prove the result by induction on e.

Let $\overline{L}$ be an algebraic closure of L and let $K' = K^{1/p} \subset \overline{L}$. Since $K = k(x_1, \ldots, x_n)$ where x_i's are algebraically independent and k is perfect, one sees that $K' = k(x_1^{1/p}, \ldots, x_n^{1/p})$. Let $L' = L \cdot K' \subset \overline{L}$ be the subfield of $\overline{L}$ generated by L and K'. Let $L'_0 \subset L'$ be the separable closure of K' in L'.
Claim: $[L' : L'_0] < [L : L_0] = p^e$.

Granting the claim for the moment, we are done by induction since the integral closure B' of A in L' is the same as the integral closure of $A' = k[x_1^{1/p}, \ldots, x_n^{1/p}]$ in L', and it is a finite extension of A' from the induction hypothesis. But A' is clearly a finite extension of A by the remark after corollary (7.3.2), and hence by corollary (7.3.3), B' is a finite module over A. Since B is an A-submodule of B' and A is Noetherian, B is a finite A-module.

To prove the claim we first prove the following:

Lemma 7.4.3 *Let k, K be as in the Theorem. Let L a finite separable extension of K and let $K' = k(x_1^{1/p}, \ldots, x_n^{1/p})$. Then $K' \cdot L = L^{1/p}$.*

Proof: Since $K' = k(x_1^{1/p}, \ldots, x_n^{1/p}) = K^{1/p}$, $K'L \subset L^{1/p}$. To prove the reverse inclusion, let us start with an element $a \in L$ and let $f(X) = \sum_{i=0}^{d} a_i X^i$

be its minimal equation with $a_i \in K$ and $a_d = 1$. Then f is separable by hypothesis. Let $b_i = a_i^{1/p} \in K'$ and consider $g(Y) = \sum b_i Y^i$. Then $g(a^{1/p})^p = f(a) = 0$. Also $g(X)^p = f(X^p)$. We claim that this implies g is separable over K'. If g had a multiple root, say α, in an algebraic closure of K', then $(X-\alpha)^2$ divides $g(X)$ and hence $(X-\alpha)^{2p} = (X^p - \alpha^p)^2$ divides $f(X^p)$ and thus $(X-\alpha^p)^2$ divides $f(X)$, and this contradicts the separability of f. Thus $a^{1/p}$ is separable over K'. Thus $a^{1/p}$ is separable as well as purely inseparable over $K' \cdot L$ and hence must be in $K' \cdot L$. Since this is true for all $a \in L$, we are done. □

Now to finish the proof of the theorem, notice that $L_0' \supset L_0^{1/p}$ by the above lemma. Let $x \in L - L_0$ such that $x^p \in L_0$. Then we can choose a basis of L over L_0 with $1, x$ as two of the basis elements. Of course the same set will generate L' over L_0'. So if we show that $1, x$ are linearly dependent over L_0', we would have proved the claim. But $x \in L_0^{1/p} \subset L_0'$. □

Corollary 7.4.4 (to lemma (7.4.3)) *Let L be a finite separable extension of $k(t)$ where k is a perfect field and t a transcendental element. Let L' be a p purely inseparable extension of L of degree p (as usual, p denotes the characteristic). Then $L' = L^{1/p}$, and in particular there is a unique inseparable extension of degree p.*

Proposition 7.4.5 *Let $A \subset B$ be an finite extension. Then given any prime ideal P of A there exists a prime ideal Q of B such that $Q \cap A = P$.*

Proof: By the exercise (7.3.1), it is easy to reduce to the case when A is local with maximal ideal P. If $PB \neq B$ then PB is a proper ideal and hence contained in a maximal ideal Q. One can see that $Q \cap A = P$. To show that $PB \neq B$ one just appeals to Nakayama's lemma. (7.2.6) □

Corollary 7.4.6 *Let A be an affine domain, K its quotient field and L a finite extension of K. Let B be the integral closure of A in L. Then for every prime ideal of A there is a prime ideal of B lying over it. In other words, given a prime ideal P of A there exists a prime ideal Q of B such that $Q \cap A = P$.*

Proof: This is just an application of the previous proposition and theorem (7.4.3) □

7.5 Discrete valuation rings

If R is an integral domain, M an R- module, K is the quotient field of R, then the localisation $M_K = \{x/a \mid x \in M, a \in R - \{0\}\}$ (where $x/a = x'/a'$ $\Leftrightarrow a'x - xa' = 0$ in M) (see exercise (7.2.1)) is a vector space over K; if $M \cong R^n$, then $n = \dim_K M_K$. Recall that a Noetherian integral domain R is a *unique factorisation domain*, denoted UFD, if intersection of two principal ideals is again principal. This definition is equivalent to the following more familiar definition: there exists a set $P \subset R$ such that any element $0 \neq x \in R$ can be uniquely expressed as

$$x = u \prod_{p \in P} p^{v(p)} \quad v(p) \geq 0 \quad v(p) = 0 \text{for all but finitely many} p$$

and u is a unit in R.

Theorem 7.5.1 *Let R be a principal ideal domain. Then:*

(i) R is a unique factorisation domain. In particular, R is integrally closed in its quotient field.

(ii) Any finitely generated R-module M can be written as a direct sum of cyclic modules,

$$M \cong R/a_1R \oplus R/a_2R \oplus \cdots \oplus R/a_kR.$$

(iii) If N is a submodule of a finitely generated R-module M, then N is also a finitely generated R-module.

(iv) The non-zero prime ideals in R are precisely the ideals generated by irreducible elements in R, and these are maximal ideals.

Proof: We first prove (i). Since R is a PID, any two elements $a, b \in R$ have a greatest common divisor, namely any generator of the ideal $aR+bR$, which is well defined upto multiplication by a unit. If a is a non-unit, either a is irreducible, or $a = a_1b_1$ where a_1, b_1 are non-units. Repeat this with a_1 and b_1; etc. Either this process stops after a finite number of steps (in which case a has been written as a product of irreducible elements), or else we can find an infinite sequence of non-units $x_1, x_2, \ldots$ such that $x_{i-1} = x_iy_i$ where y_i is

a non-unit. Consider the ideal I generated by all the x_i. Since R is a PID, $I = xR$ for some $x \in I$; then x is expressible as a finite sum of multiples of the x_i (I is precisely the set of all such sums). If n is the largest index such that x_n appears with a non-zero coefficient in x, then all the terms in the expression for x are multiples of x_n; hence x is a multiple of x_n. Since $x_{n+1} \in I$, x_{n+1} and x_n are multiples of each other; this implies y_{n+1} is a unit, a contradiction. Hence any non-unit in R is expressible as a product of irreducible elements. It is standard that any UFD is integrally closed: this follows easily from Gauss' lemma, that the polynomial ring $A[X]$ over a UFD A is itself a UFD.

Now if $p \in R$ is irreducible, and $a, b \in R$ with $p|ab$, we claim that either $p|a$ or $p|b$ — indeed, if $a \notin pR$, then the greatest common divisor of a, p is 1; thus $1 = ac + pd$ for some $c, d \in R$; hence $b = abc + pbd \in pR$, since $p|ab$. This implies that if p divides a product of irreducible elements, it divides at least one of them, say q; so $q = pa$ with $a \in R$, which (by irreducibility of q) means a is a unit. This easily implies the claimed uniqueness, upto multiplication by units, of the factorisations into irreducible elements (prove it by induction on the number of irreducible factors).

The statement (iii) of the Theorem follows from example (7.1.1). If pR is a prime ideal with $p \neq 0$, then p is a non-unit, and for any $a, b \in R$ such that $p|ab$, either $p|a$ or $p|b$. Hence if $p = ab$, then $p|a$ or $p|b$; if $p|a$, clearly b is a unit, and if $p|b$, clearly a is a unit. Hence p is irreducible. Then pR is clearly maximal, since any ideal containing pR is generated by a divisor of p; this proves (iv).

For a proof of (ii), see Lang, *Algebra* [34] or almost any book on abstract algebra (a proof is outlined in Exercise (7.5.1) at the end of this section). As a consequence of (ii), note that any finitely generated torsion-free R-module is free. □

Next, recall that a *discrete valuation ring* is a PID R which is *local i.e.*, R has a unique non-zero maximal ideal, say $\mathcal{M}$; thus, any element of R is either contained in $\mathcal{M}$, or is a unit. An element of R is thus a unit if and only if its image in the *residue field* $R/\mathcal{M}$ is non-zero. Since $\mathcal{M}$ is a principal ideal, $\mathcal{M} = \pi R$ for some $\pi \in R$; any such element π is called a *uniformising parameter* of R. Some examples of discrete valuation rings are:
(a) $R = \mathbf{Z}_{(p)} = \{a/b \mid a, b \in \mathbf{Z},\ b \neq 0,\ b \text{ not divisible by } p\}$, where p is a prime number *i.e.*, R is the ring of rational numbers whose denominators are

not multiples of p. Here p is a uniformising parameter.
(b) K is a field, t a variable over K (*i.e.*, $K(t)$ is a transcendental extension of K), and $R = \{\frac{f(t)}{g(t)} \mid \ f(t), g(t) \in K[t], \text{ and } g(0) \neq 0\}$. Here t is a uniformising parameter.

A common generalisation of these examples is given by the following result.

Theorem 7.5.2 *Let A be a Noetherian integral domain, $\mathcal{P}$ a prime ideal. Let $R = A_{\mathcal{P}} = \{a/b \mid a, b \in A, \ \ b \notin \mathcal{P}\}$, the* localisation *of A at $\mathcal{P}$. Suppose the $R/\mathcal{P}R$-module $\mathcal{P}R/(\mathcal{P}R)^2$ is non-zero, and generated by 1 element. Then R is a discrete valuation ring.*

Proof: By construction, R is a local integral domain with maximal ideal $\mathcal{M} = \mathcal{P}R$. Since R is a localisation of A, R is Noetherian by lemma (7.2.8). Now suppose the $R/\mathcal{M}$-vector space $\mathcal{M}/\mathcal{M}^2$ has dimension 1. Then $\mathcal{M}$ is principal, by Nakayama's lemma applied to $N = \mathcal{M}/\pi R$ where $\pi \in \mathcal{M} - \mathcal{M}^2$ is arbitrary.

We claim that any non-zero $x \in R$ can be written as $x = \pi^r u$ for some integer $r \geq 0$ and a unit $u \in R$, where $\mathcal{M} = \pi R$. Indeed, either x is a unit, or $x \in \mathcal{M}$ *i.e.*, $x = \pi x_1$ with $x_1 \in R$. Again, either x_1 is a unit, or $x_1 = \pi x_2$ with $x_2 \in R$. Repeating the argument with x_2, etc. we see that either $x = \pi^r u$ for some unit u, or else we can find an infinite sequence of elements $x_n \in R$ with $x_{n-1} = \pi x_n$. Let I be the ideal generated by all the x_n; then I is finitely generated, say by $y_1, \ldots, y_r$. Each y_j is a finite linear combination of x_i with coefficients in R; hence for sufficiently large n, $y_j \in x_1 R + x_2 R + \cdots + x_n R = x_n R$ for all j, so that $I \subset x_n R$. Since $x_{n+1} \in I$, $x_{n+1} = a x_n$, with $a \in R$, so that $x_{n+1} = a\pi x_{n+1}$ *i.e.*, π is a unit, a contradiction. Since any element of R has the form $\pi^r u$ for some unit u, any ideal in R equals $\pi^r R$ for some $r \geq 0$. Hence R is a discrete valuation ring. □

If R is a discrete valuation ring, K its quotient field, and π is a uniformising parameter, then any non-zero element $x \in K$ can be uniquely written as $x = \pi^n u$ for some $n \in \mathbf{Z}$, and a unit $u \in R$. This is because R is a unique factorisation domain with only one maximal ideal πR, so any element of R can be uniquely written as $\pi^n u$ with $n \geq 0$, and u a unit. If $K^* = K - \{0\}$ is

the multiplicative group of K, we can define a homomorphism $v_R : K^* \to \mathbf{Z}$ by $v(\pi^n u) = n$, where u is any unit. Define $v_R(0) = \infty$, which is defined to be larger than any positive integer. Then

(i) $v_R(xy) = v_R(x) + v_R(y)$ for all $x, y \in K$, and $v_R(x) = \infty \Leftrightarrow x = 0$;

(ii) $v_R(x + y) \geq \min\{v_R(x), v_R(y)\}$ for all $x, y \in K$; and

(iii) $R = \{x \in K \mid v_R(x) \geq 0\}$.

A mapping $v : K \to \mathbf{Z} \cup \{\infty\}$ satisfying (i), (ii), and such that $v(x) \neq 0$ for some $x \in K^*$, is called a *discrete valuation* of the field K. One verifies that the expression in (iii) defines a subring R of K, which is a discrete valuation ring; the maximal ideal is $\{x \in K \mid v(x) > 0\}$. From (i), $v(K^*) = d\mathbf{Z}$ for some $d \geq 0$; then $d > 0$ since $v(x) > 0$ for some $x \in K^*$. Then we see easily that $v(x) = dv_R(x)$. The mapping v_R is called the *normalised* discrete valuation associated to v. The conditions (i) and (ii) easily imply that if $x \in K$ satisfies a monic polynomial with coefficients in R, then $x \in R$, *i.e.*, any discrete valuation ring R is integrally closed in K.

Let K be a field with a discrete valuation v. Define an *absolute value* associated to v by

$$|x|_v = \alpha^{v(x)}$$

where α is a fixed real number with $0 < \alpha < 1$. Here $|0|_v$ is defined to be 0. Then

(i) $|x|_v \geq 0$, and $|x|_v = 0 \Leftrightarrow x = 0$;

(ii) $|xy|_v = |x|_v |y|_v$ for all $x, y \in K$; and

(iii) $|x + y|_v \leq \max\{|x|_v, |y|_v\}$.

(iii) clearly implies the triangle inequality for $|\ |_v$. Hence K becomes a metric space with respect to $d(x, y) = |x - y|_v$. It is not difficult to show that if $v : K \to \mathbf{R}$ is any function satisfying (i), (ii), (iii) such that v is non-constant on K^* (equivalently, d is not the discrete metric on K), then $R = \{x \in K \mid |x|_v \leq 1\}$ is a discrete valuation ring with quotient field K, and for some $\alpha < 1$, v is the absolute value associated to v_R.

Exercise 7.5.1: This problem outlines the proof of the 'basis theorem for finitely generated modules over a PID', which states that if R is a PID, M a finitely generated R-module, then

$$M \cong R/a_1R \oplus R/a_2R \oplus \cdots \oplus R/a_kR$$

for some $a_i \in R$.

1. Show that if F is a finitely generated free R-module of rank n, and $F' \subset F$ is a submodule, then F' is free of rank $\leq n$.
   (Hint: Induction on $n = \text{rank}F$. If $n = 1$, F' is an ideal, which is principal, hence is 0 or $\cong R$. If $n > 1$, let $\pi : F \twoheadrightarrow F_1$ be a homomorphism onto a free module F_1 of rank $n-1$, such that $\ker\pi \cong R$; then $\pi(F')$ is free of rank $\leq n-1$, say with free basis $y_1, \ldots, y_r$. Choose arbitrary $x_1, \ldots, x_r \in F'$ with $\pi(x_i) = y_i$; show that $F' \cong Rx_1 \oplus Rx_2 \oplus \cdots \oplus Rx_r \oplus ((\ker\pi) \cap F')$, where the submodule generated by the x_i is free, and $(\ker\pi) \cap F'$ is either 0, or free of rank 1).

2. Let M be an R-module generated by n elements, $N \subset M$ an R- submodule. Show that N is generated by $\leq n$ elements.
   (Hint: Write M as a quotient of a free module of rank n, and use (a).)

3. Let M be a finitely generated *torsion-free* R-module (*i.e.*, if $m \in M - \{0\}$, $a \in R - \{0\}$, then $am \neq 0$). Show that M is a free R-module of finite rank.
   (Hint: Let n be the maximum of the ranks of free submodules of M, and let $N \subset M$ be free of rank n. For any $m \in M$, show that there exists $a \in R - \{0\}$ such that $am \in N$ (use maximality of n); deduce that there exists $a \in R - \{0\}$ such that $aM \subset N$. But then $M \cong aM$ is free of finite rank.)

4. Let $T(M) = \{m \in M \mid \exists a \in R - \{0\}$ such that $am = 0\}$ be the torsion submodule of M. Show that $M \cong T(M) \oplus F$ where $F \cong M/T(M)$ is a free R-module of finite rank.
   (Hint: $M/T(M)$ is finitely generated and torsion-free, hence free; let $y_1, \ldots, y_r$ be a basis for $M/T(M)$, and let $x_i \in M$ be an arbitrary preimage of y_i, for each i. Let $F = \sum Rx_i$. Show that F is free with basis x_i, and $M = T(M) \oplus F$.)

5. Let M be a finitely generated torsion R-module *i.e.*, $M = T(M)$. Show that there exists $a \in R - \{0\}$ such that $aM = 0$. Choose such an a, and suppose $a = up_1^{n_1} \cdots p_k^{n_k}$ is its prime factorisation, where u is a unit, and the p_i are distinct primes. Let $b_i = \prod_{j \neq i} p_j^{n_j}$, and $M_j = b_j M$. Show that $p_j^{n_j} M_j = 0$, and $M \cong \oplus_{j=1}^{k} M_j$.
(Hint: Clearly $b_1, \ldots, b_k$ have greatest common divisor 1; let $c_j \in R$ such that $\sum_j b_j c_j = 1$. Then for any $m \in M$, $m = \sum_j c_j(b_j m)$ where $b_j m \in M_j$. Hence $M = \sum_j M_j$. Now show that $M_j \cap \sum_{k \neq j} M_k = 0$ for all j.)

6. Let $p \in R$ be a prime (*i.e.*, an irreducible element), M a finitely generated R-module such that $p^n M = 0$ for some $n > 0$. Show that M is a direct sum of cyclic R-modules,

$$M \cong R/p^{a_1}R \oplus \cdots \oplus R/p^{a_k}R, \quad a_1 \geq a_2 \geq \cdots \geq a_k \geq 0,$$

for unique choices of the exponents a_i.
(Hint: Work by induction on n; for $n = 1$, M is a vector space over R/p, so the result holds with $a_i = p$ for all i. Now suppose $n > 0$; then $p^{n-1}M$ is a vector space over R/pR. Let $x_1, \ldots, x_s \in M$ such that $p^{n-1}x_j$ are a basis for $p^{n-1}M$. Let $M_1 = \sum Rx_j$. Then $M_1 \cong (R/p^nR)^s$, and if $M_2 = M/M_1$, then $p^{n-1}M_2 = 0$; hence M_2 has a direct sum decomposition of the desired form. Show that if $y \in M_2$ with $p^t y = 0$, $p^{t-1}y \neq 0$, then there exists $x \in M$ with $x \longmapsto y$ and $p^t x = 0$ (first choose any preimage x; if $p^t x = \sum_{i=1}^{s} p^{t_i} c_i x_i$, where $p \not| \; c_i$, show that $t \geq t_i$; now modify x suitably). Use this to complete the proof of the inductive step.)

7.6 Dedekind domains

A *Dedekind domain* R is a Noetherian domain which is integrally closed in its quotient field, such that any non-zero prime ideal of R is maximal.

The basic examples of Dedekind domains are the following.

(i) Any principal ideal domain is a Dedekind domain.

(ii) The ring of algebraic integers in an algebraic number field is a Dedekind domain (see the appendix on algebraic number theory).

(iii) Let k be a field, $R = k[t]$ the polynomial ring in 1 variable over k. Let $K = k(t)$ be the quotient field of R, and let L be a finite algebraic extension of K; let S be the integral closure of R in L *i.e.*, S is the subring of L of elements which are integral over R. Then S is a Dedekind domain.

The proof that the ring S in (iii) is a Dedekind domain is as follows. Clearly S in integrally closed in L, its quotient field. By theorem (7.4.3), S is Noetherian. If $\mathcal{P} \subset S$ is a non-zero prime ideal, and $x \in \mathcal{P}$ a non-zero element, then x satisfies a monic polynomial $x^n + a_1 x^{n-1} + \cdots + a_{n-1} + a_n = 0$, where $a_n \neq 0$. Then $a_n \in \mathcal{P} \cap R$ and $\mathcal{P} \cap R$ is a non-zero prime ideal in the polynomial ring, which is maximal, since R is a PID. Hence $S/\mathcal{P}$ is an integral extension of the field $R/\mathcal{P} \cap R$, and is hence a field by lemma (7.3.4).

The next result gives an important property of Dedekind domains. Before stating it, we recall that if I, J are ideals in a commutative ring R, then their product IJ is the set of all finite sums $\sum a_i b_i$ with $a_i \in I$, $b_i \in J$ for all i. If $I = aR$, $J = bR$ then clearly $IJ = abR$. We recall that a characterisation of prime ideals is the following: I is prime if for any product of ideals $I_1 I_2 \subset I$, either $I_1 \subset I$ or $I_2 \subset I$. This statement for principal ideals I_1, I_2 is the usual definition of primality, but this easily implies the characterisation.

Theorem 7.6.1 (Unique factorisation theorem for ideals) *Any proper ideal (one not equal to* 0 *or the whole ring) in a Dedekind domain A is uniquely expressible (upto the order of the factors) as a product of non-zero prime ideals. Given any non-zero ideals $I \subset J$, there exists a unique ideal J' such that $JJ' = I$. For any pair of ideals I and J which have no common prime factor, $IJ = I \cap J$.*

Proof: We first show that any non-zero ideal contains a non-trivial finite product of prime ideals. We prove this by 'Noetherian induction' — if this is false, let $\mathcal{C}$ be the collection of ideals which do not contain any non-trivial finite product of primes; then $\mathcal{C}$ has a maximal element, say I. Since A has maximal ideals, clearly $I \neq A$; also I is not itself prime. Hence there exist $a, b \in A$ with $ab \in I$, but $a \notin I$, $b \notin I$. Then $I + aA$ and $I + bA$ are strictly larger ideals than I, so by maximality, $I + aA$ and $I + bA$ do not lie in $\mathcal{C}$ *i.e.*, they each contain products of primes. But then $(I + aA)(I + bA) \subset I$ contains a product of primes.

Now let $\mathcal{P}$ be any maximal ideal of A; we claim that there exists $\lambda \in K - A$ such that $\lambda\mathcal{P} \subset A$. Let $a \in \mathcal{P}$; choose a product $\mathcal{P}_1\mathcal{P}_2\cdots\mathcal{P}_r \subset aA$, where $\mathcal{P}_i$ are primes, and r is minimal. Then $\mathcal{P}$ contains $\mathcal{P}_1\cdots\mathcal{P}_r$; since $\mathcal{P}$ is prime, $\mathcal{P}$ must contain one of the $\mathcal{P}_i$; say $\mathcal{P}_1 \subset \mathcal{P}$. Since $\mathcal{P}$, $\mathcal{P}_1$ are in fact maximal, $\mathcal{P} = \mathcal{P}_1$. Now $\mathcal{P}_2\cdots\mathcal{P}_r \not\subset aA$ since r is minimal, so there exists $b \in \mathcal{P}_2\cdots\mathcal{P}_r - aA$. Then $b\mathcal{P} \subset \mathcal{P}_1\cdots\mathcal{P}_r \subset aA$, so $(b/a)\mathcal{P} \subset A$ even though $(b/a) \notin A$ (since $b \notin aA$ by choice); hence we may take $\lambda = b/a$. Note that $\lambda I \not\subset I$ for any ideal I of A — indeed, λ is not in A, and hence is not integral over A; hence there does not exist any finite A-submodule M of K such that $\lambda M \subset M$. In particular, $\mathcal{P} + \lambda\mathcal{P}$ is an ideal of A which is strictly larger than $\mathcal{P}$ *i.e.*, is equal to A. Thus there exist $p_1, p_2 \in \mathcal{P}$ with $p_1 + \lambda p_2 = 1$.

We now prove unique factorisation for ideals I, by Noetherian induction. Let I be a maximal counterexample. Then I is not itself maximal. Suppose I is contained in a maximal ideal $\mathcal{P}$. Choose $\lambda \in K - A$ such that $\lambda\mathcal{P} \subset A$. As above, there exist $p_1, p_2 \in \mathcal{P}$ with $p_1 + \lambda p_2 = 1$. Now let $J = I + \lambda I$. This is strictly larger than I, since $\lambda I \not\subset I$. We claim $I = \mathcal{P}J$. Indeed, $\mathcal{P}J = \mathcal{P}I + \lambda\mathcal{P}I \subset I$, and $I \subset p_1 I + \lambda p_2 I \subset \mathcal{P}(I + \lambda I) = \mathcal{P}J$. Thus $J \neq A$, and since I is a maximal counterexample, J has a unique factorisation into primes, $J = \mathcal{Q}_1\cdots\mathcal{Q}_s$. Then $I = \mathcal{P}J = \mathcal{P}\mathcal{Q}_1\mathcal{Q}_2\cdots\mathcal{Q}_s$. This proves I is a product of primes.

To prove uniqueness, note that since $I \subset \mathcal{P}$, any product of primes contained in I must have $\mathcal{P}$ as one of the factors in the product. Suppose $I = \mathcal{P}\mathcal{P}_1\mathcal{P}_2\cdots\mathcal{P}_r = \mathcal{P}J'$, where $J' = \mathcal{P}_1\mathcal{P}_2\cdots\mathcal{P}_r$. Then $\mathcal{P}J = \mathcal{P}J' \Rightarrow J = \lambda I + I = J'$ (use $1 = p_1 + \lambda p_2$); by uniqueness for J, $r = s$ and the $\mathcal{P}$'s are a permutation of the $\mathcal{Q}$'s.

In the course of the above argument, we have shown that if I is a non-zero ideal, and $I \subset \mathcal{P}$ for some prime ideal $\mathcal{P}$, then $I = \mathcal{P}J$ for a unique ideal J. This implies that if $I \subset I'$ are non-zero ideals, then $I = I'J'$ for a unique ideal J'; the proof is by induction on the number of prime factors of I'. In particular, for any prime $\mathcal{P}$, the power of $\mathcal{P}$ dividing I is the highest power of $\mathcal{P}$ containing I.

Suppose now that I, J are non-zero ideals which have no common prime factors (I, J are relatively prime). Then $I \cap J = IJ$. Indeed, if $\mathcal{P}$ is a prime containing I, and $\mathcal{P}^n$ is the power of $\mathcal{P}$ dividing I, then (since $\mathcal{P}$ does not divide J) $\mathcal{P}^n$ is the power of $\mathcal{P}$ dividing IJ. Hence the power of $\mathcal{P}$ dividing $I \cap J$ is also $\mathcal{P}^n$, from $IJ \subset I \cap J \subset I$. A similar argument with the primes dividing J proves that $I \cap J$ and IJ have the same prime factorisation. $\square$

Theorem (7.6.1) may be used to deduce the following characterisation of Dedekind domains.

Theorem 7.6.2 *Let A be a Noetherian integral domain with quotient field K. Then A is a Dedekind domain if and only if for any maximal ideal $\mathcal{P} \subset A$, the localisation*

$$A_{\mathcal{P}} = \{\frac{a}{b} \mid a, b \in A,\ b \notin \mathcal{P}\}$$

is a discrete valuation ring. Further,

$$A = \bigcap_{\mathcal{P}} A_{\mathcal{P}}$$

where $\mathcal{P}$ runs over all maximal ideals, and the intersection takes place inside K.

Proof: First assume A is a Dedekind domain. Associated to each maximal ideal $\mathcal{P} \subset A$, there is a mapping

$$v_{\mathcal{P}} : A - \{0\} \to \mathbf{Z},$$

$$v_{\mathcal{P}}(x) = \text{ power of } \mathcal{P} \text{ in the factorisation of the ideal } xA\ .$$

Then from unique factorisation, $v_{\mathcal{P}}$ is well defined, and satisfies

(i) $v_{\mathcal{P}}(x) \geq 0$,

(ii) $v_{\mathcal{P}}(xy) = v_{\mathcal{P}}(x) + v_{\mathcal{P}}(y)$, and

(iii) $v_{\mathcal{P}}(x+y) \geq \min\{v_{\mathcal{P}}(x), v_{\mathcal{P}}(y)\}$.

Let K be the quotient field of A. From (ii), $v_{\mathcal{P}}$ extends uniquely to a homomorphism $K^* \to \mathbf{Z}$, by the formula $v_{\mathcal{P}}(a/b) = v_{\mathcal{P}}(a) - v_{\mathcal{P}}(b)$ for all $a, b \in A - \{0\}$. For any $x, y \in K^*$, there exists $z \in A - \{0\}$ such that $xz, yz \in A$; using this, one sees that (iii) holds for all $x, y \in K^*$. Thus, $v_{\mathcal{P}}$ yields a well defined discrete valuation of K, which is normalised — since $\mathcal{P} \neq \mathcal{P}^2$ by unique factorisation, there exists $x \in \mathcal{P} - \mathcal{P}^2$, and this clearly satisfies $v_{\mathcal{P}}(x) = 1$. This is called the $\mathcal{P}$-adic valuation of K. The discrete valuation ring $\mathcal{O}_{\mathcal{P}}$ consists of all fractions a/b with $a, b \in A$ and $v_{\mathcal{P}}(a) \geq v_{\mathcal{P}}(b)$.

Suppose $x \in K^*$, and $x = a/b$ is any expression with $a, b \in A$. Let $aA = I\mathcal{P}_1\mathcal{P}_2 \cdots \mathcal{P}_r$, $b = I\mathcal{Q}_1\mathcal{Q}_2 \cdots \mathcal{Q}_s$, where the $\mathcal{P}_i$ are distinct from all the

$\mathcal{Q}_j$, and I is a non-zero ideal. Then $xI\mathcal{Q}_1\mathcal{Q}_2\cdots\mathcal{Q}_s = I\mathcal{P}_1\mathcal{P}_2\cdots\mathcal{P}_r$. If $c \in I$ is non-zero, we may write $cA = I'I$ for some non-zero ideal I', by unique factorisation; hence $xI'I\mathcal{Q}_1\mathcal{Q}_2\cdots\mathcal{Q}_s = I'I\mathcal{P}_1\mathcal{P}_2\cdots\mathcal{P}_r$ *i.e.*, $xc\mathcal{Q}_1\mathcal{Q}_2\cdots\mathcal{Q}_s = c\mathcal{P}_1\mathcal{P}_2\cdots\mathcal{P}_r$. This implies $x\mathcal{Q}_1\mathcal{Q}_2\cdots\mathcal{Q}_s = \mathcal{P}_1\mathcal{P}_2\cdots\mathcal{P}_r$. The $\mathcal{Q}_j$ are precisely the primes corresponding to the ($\mathcal{P}$-adic) discrete valuations where $v(x) < 0$, while the $\mathcal{P}_i$ correspond to those where $v(x) > 0$.

We deduce first that if $v_{\mathcal{P}}(x) \geq 0$ for all $\mathcal{P}$, then $x \in A$, since in this case, the set of $\mathcal{Q}_j$ must be empty. This implies that

$$A = \bigcap_{\mathcal{P}} \mathcal{O}_{\mathcal{P}}.$$

Next, if $\mathcal{P}$ is any prime distinct from each of the $\mathcal{Q}_j$, we can write $x = a/b$ with $b \notin \mathcal{P}$. Indeed, choose any $b \in \mathcal{Q}_1\mathcal{Q}_2\cdots\mathcal{Q}_s - \mathcal{P}$, and let $a = xb$. Thus, any element in the discrete valuation ring $\mathcal{O}_{\mathcal{P}}$ is of the form a/b with $a, b \in A$ and $b \notin \mathcal{P}$. Hence $\mathcal{O}_{\mathcal{P}}$ is the localisation of A at the maximal ideal $\mathcal{P}$ *i.e.*, $\mathcal{O}_{\mathcal{P}} = A_{\mathcal{P}}$. Hence

$$A = \bigcap_{\mathcal{P}} A_{\mathcal{P}}.$$

Conversely, suppose A is a Noetherian integral domain, such that for any maximal ideal $\mathcal{P} \subset A$, the localisation $A_{\mathcal{P}}$ is a discrete valuation ring. If $\mathcal{Q}$ is any non-zero prime ideal, then there exists a maximal ideal $\mathcal{P}$ containing $\mathcal{Q}$. Then $\mathcal{Q}A_{\mathcal{P}}$ is a non-zero prime ideal in the discrete valuation ring $A_{\mathcal{P}}$ *i.e.*, $\mathcal{Q}A_{\mathcal{P}} = \mathcal{P}A_{\mathcal{P}}$, which implies $\mathcal{Q} = \mathcal{Q}A_{\mathcal{P}} \cap A = \mathcal{P}A_{\mathcal{P}} \cap A = \mathcal{P}$. Hence any non-zero prime ideal of A is maximal. It remains to show that A is integrally closed in its quotient field K. Suppose $x \in K$ is integral over A. Let $D_A(x) = \{y \in A \mid xy \in A\}$ be the ideal of denominators of x; it suffices to show that $D_A(x) = A$. If not, $D_A(x) \subset \mathcal{P}$ for some maximal ideal $\mathcal{P}$. Hence $D_{A_{\mathcal{P}}}(x) = D_A(x)A_{\mathcal{P}} \subset \mathcal{P}A_{\mathcal{P}}$. Hence $x \in K - A_{\mathcal{P}}$ is integral over $A_{\mathcal{P}}$ *i.e.*, the discrete valuation ring $A_{\mathcal{P}}$ is not integrally closed, which is absurd. □

Exercise 7.6.1: Let A be a Dedekind domain with quotient field K, and let R be a discrete valuation ring contained in K such that $A \subset R$. Show that $R = A_{\mathcal{P}}$ for some maximal ideal $\mathcal{P}$ of R.

Chapter 8

Varieties

8.1 Affine Varieties

Let k be an algebraically closed field. The n-fold product of k with itself is called the *affine n-space over k* and denoted by $\mathbf{A}^n$ (or $\mathbf{A}^n_k$, if we wish to call attention to the underlying field k). The projections from $\mathbf{A}^n$ to k are called the coordinate functions and denoted by $x_1, \cdots, x_n$. A k-valued function on $\mathbf{A}^n$ is called an algebraic function or a *regular function* if it is given by a polynomial in $x_1, \cdots, x_n$.

Let A be the k-algebra of all regular functions on $\mathbf{A}^n$. We call A the *coordinate ring* of $\mathbf{A}^n$. The natural map from the polynomial ring $k[X_1, \cdots, X_n]$ in n independent variables $X_1, X_2, \cdots X_n$ to A given by $X_i \mapsto x_i$ is an isomorphism, since k contains infinitely many elements (see Exercise (7.4.1)) So A is the polynomial ring $k[x_1, \cdots, x_n]$.

Let $\{f_\alpha\}_{\alpha \in J}$ be a set of polynomials in $x_1, \cdots, x_n$. Then we have the set of common zeros of the f_α,

$$V(\{f_\alpha\}_{\alpha \in J}) = \{(a_1, \cdots, a_n) \in \mathbf{A}^n \mid f_\alpha(a_1, \cdots, a_n) = 0 \text{ for all } \alpha \in J\}.$$

Note that if I is the ideal generated by $\{f_\alpha\}_{\alpha \in J}$ in A, then X is also equal to

$$V(I) = \{(a_1, \cdots, a_n) \mid f(a_1, \cdots, a_n) = 0 \quad \forall f \in I\}.$$

Moreover, since $k[x_1, \ldots, x_n]$ is Noetherian, any ideal in $k[x_1, \cdots x_n]$ is finitely generated; hence the set of common zeros of an ideal I can also be defined as the set of common zeros of a *finite* set of polynomials, say $f_1, \ldots, f_r$, in I.

We call such a set an *affine algebraic set* or an *affine algebraic variety* (or simply affine variety) defined by $f_1, \cdots f_r$ (or equivalently, by I).

The k-valued functions an affine variety $X \subset \mathbf{A}^n$, obtained by restricting the regular functions on $\mathbf{A}^n$ to X, are called *regular functions* on X. These form an algebra which can identified in the obvious manner with $A/I(X)$, where $I(X)$ is the ideal in A of all regular functions vanishing on X. We call this the *coordinate ring* of X. We denote it by $\Gamma(X)$.

There is a natural identification $\mathbf{A}^r \times \mathbf{A}^s = \mathbf{A}^{r+s}$, so that if $x_1, \ldots, x_r$ and $y_1, \ldots, y_s$ are the coordinate functions on $\mathbf{A}^r$, $\mathbf{A}^s$ respectively, then the $r+s$ functions $x_1, \ldots, x_r, y_1, \ldots, y_s$ are the coordinate functions on $\mathbf{A}^{r+s}$. If $X \subset \mathbf{A}^r$ and $Y \subset A^s$ are defined by ideals $I \subset k[x_1, \ldots, x_r]$ and $J \subset k[y_1, \ldots, y_s]$, then $X \times Y \subset \mathbf{A}^{r+s}$ is the subset defined by the ideal generated by $I \cup J$. In more abstract terms, we may identify the coordinate ring of $\mathbf{A}^r \times \mathbf{A}^s$ with the tensor product (as k-algebras) of the coordinate rings of the factors; now we check easily that the coordinate ring of $X \times Y$ is just the tensor product $\Gamma(X) \otimes_k \Gamma(Y)$ of the coordinate rings of X and Y.

Example 8.1.1: If $P = (a_1, \ldots, a_n) \in A^n$, then $X = \{P\}$ is the algebraic set defined by $x_1 - a_1 = \ldots = x_n - a_n = 0$, the ideal $I(X)$ is the maximal ideal $(x_1 - a_1, \ldots, x_n - a_n) \subset k[x_1, \ldots, x_n]$, the coordinate ring $\Gamma(X) \cong k$. From lemma (7.4.1), it follows that there are 1-1 correspondence between the points of $\mathbf{A}^n$, the maximal ideals of $k[x_1, \ldots, x_n]$ and k-algebra homomorphisms $k[x_1, \ldots, x_n] \to k$.

Given the affine variety $X \subset \mathbf{A}^n$ defined by an ideal I, consider all the functions in $k[x_1, \cdots, x_n]$ which vanish on X. They form the ideal $I(X)$. Clearly $I \subset I(X)$, but $I(X)$ could in general be larger. For example, $I(X)$ is clearly a radical ideal, since if f^n vanishes on X for some $n > 0$, then f vanishes on X. Thus $\Gamma(X) = A/I(X)$ is always a *reduced* ring (*i.e.*, it has no non-zero nilpotent elements). The precise relation between the ideals I and $I(X)$ is given by Hilbert's Nullstellensatz (Theorem (8.1.1)), which we now prove.

We begin with a special case.

Lemma 8.1.1 *Let $R = k[x_1, \ldots, x_n]$ be the polynomial ring and I an ideal of R. Then $V(I) = \phi$ if and only if $I = R$.*

Proof: If $I = R$ clearly $V(I) = \phi$. So assume I is a proper ideal of R. Then it is contained in a maximal ideal $\mathcal{M}$ of R. By lemma (7.4.2), $\mathcal{M} = (x_1 - a_1, \ldots, x_n - a_n)$. Then it is clear that $(a_1, \ldots, a_n) \in V(I)$. □

Corollary 8.1.2 *Let X be an affine variety over an algebraically closed field and let A be its coordinate ring. Let I be an ideal in A. Then $V(I) = \phi$ if and only if $I = A$.*

Theorem 8.1.1 (Hilbert's Nullstellensatz) *Let X be an affine variety over an algebraically closed field and let $\Gamma(X)$ be its coordinate ring. Let I be an ideal of $\Gamma(X)$ and Y the subvariety of X defined by I. Let J be the set of functions in $\Gamma(X)$ vanishing on Y. Then $J = \sqrt{I}$.*

Proof: It is clear that $\sqrt{I} \subset J$. Now let $f \in J$. Consider the ring $R = \Gamma(X)[t] = \Gamma(X \times \mathbf{A}^1)$ and the ideal $\mathcal{P} = (I, 1 - tf)$. Then one easily checks that $V(\mathcal{P}) = \phi$. Thus by the previous corollary $\mathcal{P} = R$. This is same as saying that $1 - tf$ is a unit in the ring $(\Gamma(X)/I)[t]$. Then the following exercise finishes the proof. □

Exercise 8.1.1: Let $R = A[t]$ a polynomial ring in one variable over a ring A. Let $a_0 + a_1 t + a_2 t^2 + \cdots + a_n t^n$ be a unit in R. Then show that a_0 is a unit in A and $a_i, \quad i > 0$ are all nilpotent in A.

Corollary 8.1.3 *For any affine variety X, there is a 1-1 correspondence between algebraic subsets in X, radical ideals in $\Gamma(X)$ and reduced quotient rings of $\Gamma(X)$.*

If $\{X_\alpha\}$ is a collection of algebraic sets in $\mathbf{A}^n$, where X_α is defined by an ideal I_α, then the reader can check that $\cap X_\alpha$ is the algebraic set defined by the ideal $\sum I_\alpha$. Similarly, the union $X_\alpha \cup X_\beta$ is defined by the product ideal $I_\alpha I_\beta$. Hence we get a topology on $\mathbf{A}^n$ by taking the algebraic sets to be closed sets. This topology is called the *Zariski topology* on $\mathbf{A}^n$.

If X is an affine algebraic set in $\mathbf{A}^n$, we have the induced subspace topology on X, also called the Zariski topology on X. Thus the Zariski open subsets of an affine variety $X \subset \mathbf{A}^n$ are all of the form $X - Y$ where Y is another affine variety in $\mathbf{A}^n$.

Exercise 8.1.2: The non-empty Zariski open subsets of $\mathbf{A}^1$ are the complements of finite sets of points; thus the Zariski topology on $\mathbf{A}^1$ is the cofinite topology.

Exercise 8.1.3: Let X be an affine variety. Show that X, with its Zariski topology, is compact, *i.e.*, any open cover of X has a finite subcover.
(Hint: Show that if $\{X_\alpha\}$ is any family of closed subsets with $\cap_\alpha X_\alpha = \phi$, then some finite intersection $X_{\alpha_1} \cap X_{\alpha_2} \cap \cdots \cap X_{\alpha_n} = \phi$.)

Exercise 8.1.4: Let X be an affine variety in $\mathbf{A}^n$ with coordinate ring $\Gamma(X)$. Show that a basis for the open sets in the Zariski topology of X is given by subsets of the form $D(f) = \{x \in X \mid f(x) \neq 0\}$, where $f \in \Gamma(X)$.

Note that the open set $D(f)$ in the above exercise has a natural structure as an affine variety in $\mathbf{A}^{n+1} = \mathbf{A}^n \times \mathbf{A}^1$ given by the ideal $(I(X), 1 - tf) \subset A[t]$, where t is the coordinate function on the factor $\mathbf{A}^1$. This fact was already used in the proof of the Nullstellensatz.

Definition: A topological space is called *irreducible* if it is not the union of two proper closed subsets.

Note that this is equivalent to saying that any two nonempty open subsets have nonempty intersection.

Since the ring $k[X_0, \cdots X_n]$ is Noetherian, so is the coordinate ring $\Gamma(X)$ of an affine variety $X \subset \mathbf{A}^n$. We had noted that there is a 1-1 correspondence between subvarieties $Y \subset X$ and radical ideals $I \subset \Gamma(X)$. We deduce that if X is an affine variety in $\mathbf{A}^n$, then as a topological space, X is Noetherian, *i.e.*, any descending chain $X_1 \supset X_2 \supset \cdots \supset \cdots$ of Zariski closed subsets of X stabilises (*i.e.*, $\exists N$ such that $X_N = X_{N+1} = \cdots$). From this it follows easily that X can be written uniquely as a finite union $X_1 \cup \cdots \cup X_r$ of closed irreducible subsets such that $X_i \not\subseteq X_j$ for $i \neq j$. The X_i are called the *irreducible components* of X.

Exercise 8.1.5: Let $X \subset \mathbf{A}^n$ be an affine variety and $\Gamma(X)$ its coordinate ring.

1. Show that X is irreducible $\Leftrightarrow \Gamma(X)$ is a domain $\Leftrightarrow I(X) \subset k[x_1, \ldots, x_n]$ is a prime ideal.

2. If $X = X_1 \cup \cdots \cup X_r$ is the decomposition of X into irreducible components, then $I(X) = \cap_i I(X_i)$ where $I(X_i)$ are the minimal prime ideals in $k[x_1, \ldots, x_n]$ lying over $I(X)$ (thus, by (7.2.12), $f \in \Gamma(X)$ is a zero divisor $\Leftrightarrow$ f vanishes identically on some X_i).

We have the notion of rational functions on $\mathbf{A}^n$ given by the elements of the field $k(x_1, \cdots, x_n)$ of fractions of the polynomial ring $k[x_1, \cdots x_n]$. Also, if $P \in \mathbf{A}^n$ is a point, then one has the subring of rational functions which are regular at P, which are ratios of polynomials $f(x_1, \ldots, x_n)/g(x_1, \ldots, x_n)$ with $g(P) \neq 0$. If P corresponds to the maximal ideal $\mathcal{M} \subset k[x_1, \ldots, x_n]$, then rational functions regular at P are just elements of the ring $k[x_1, \ldots, x_n]_{\mathcal{M}}$ obtained by localising the polynomial ring at $\mathcal{M}$.

The algebraic versions of the definitions allow us to extend these function theoretic notions to arbitrary affine varieties. If $X \subset \mathbf{A}^n$ with coordinate ring $\Gamma(X)$, we define a *rational function* on X to be an element of the total quotient ring $k(X)$ of $\Gamma(X)$. By definition, this is the localisation of $\Gamma(X)$ with respect to the multiplicative set of non zero divisors. Thus rational functions are expressible as ratios f/g where $f, g \in \Gamma(X)$ and g is a non zero divisor (*i.e.*, from the above exercise, g does not vanish identically on any irreducible component of X). However, two different ratios may represent the same rational function on X, as seen from the example below.

Example 8.1.2: Consider the variety $X \subset \mathbf{A}^4$ defined by $xy - zw = 0$, where x, y, z, w are the coordinates on $\mathbf{A}^4$. Then the elements x/z and w/y of $k(X)$ are equal, *i.e.*, define the same rational function on X. This phenomenon occurs on X because the coordinate ring $\Gamma(X)$ is not a unique factorisation domain, unlike the polynomial ring.

If X is irreducible, then $\Gamma(X)$ is a domain, and $k(X)$ is its quotient field. If $X = \cup_i X_i$ is the decomposition into irreducible components, then $k(X) = \prod_i k(X_i)$ is a product of fields.

Any ratio f/g of regular functions defines a k-valued function on the open subset of X where $g \neq 0$. Using this, one then sees that a rational function φ determines a k-valued function on a certain open set in X. If $\Gamma(X)$ is the coordinate ring of X, $k(X)$ its total quotient ring, then $\Gamma(X) \subset k(X)$, and $\varphi \in k(X)$. Let $Den(\varphi) = \{g \in \Gamma(X) \mid g\varphi \in \Gamma(X)\}$ be the 'ideal of denominators' of φ. Then φ determines a k-valued function on $X - Y$, where

$Y \subset X$ is the subvariety defined by $Den(\varphi)$. Indeed, by the Nullstellensatz, if $P \in X - Y$, then there exists $g \in Den(\varphi)$ with $g(P) \neq 0$, and $g\varphi = f \in \Gamma(X)$, so that we may define $\varphi(P) = f(P)/g(P)$. One sees at once that this is independent of the choice of g.

Conversely, the k-valued function uniquely determines the rational function φ. Hence the set of rational functions may be identified with a certain set of k-valued functions, but these do not all have the same domain, in general; thus it is a little clumsy to rigorously define the ring operations on rational functions in terms of the associated k-valued functions. Nevertheless, it is useful to consider the above k-valued functions, and in particular, justifies the use of the word 'function' in the term 'rational function'.

If $P \in X$ is any point, then P corresponds to a maximal ideal $\mathcal{M}_P$ in $\Gamma(X)$. Conversely, all maximal ideals in $\Gamma(X)$ are obtained in this way, since this is true when $X = \mathbf{A}^n$, as seen earlier. The *local ring of* P is defined to be the local ring $\Gamma(X)_{\mathcal{M}_P}$ obtained by localising R at P. If $\mathcal{M} \subset A$ is the maximal ideal of P in $\mathbf{A}^n$, then $\mathcal{M}_P = \mathcal{M}/I(X)$, and $\Gamma(X)_{\mathcal{M}_P} = A_{\mathcal{M}}/(I(X)A_{\mathcal{M}})$.

From a function theoretic point of view, this local ring is constructed as follows. We first recall the notion of the *germ* of a k-valued function determined at a point P of a topological space X (where for the moment, k could be any set). We consider k-valued functions $f : U \to k$ where U is an open neighbourhood of P; two such pairs (U, f) and (V, g) are said to be equivalent if $f = g$ on some smaller neighbourhood $W \subset U \cap V$ of P. An equivalence class of such pairs is called a germ of a function; given a pair (U, f), the corresponding equivalence class is called the germ of the function f at P. Clearly, if k is a ring, the set of germs of k-valued functions at P forms a ring in a natural way. From the definition, if (U, f) and (V, g) have the same germ at P (*i.e.*, are equivalent, as defined above) then $f(P) = g(P)$. Hence if f is a germ at P, it makes sense to speak of the value of $f(P)$ of the germ at P. f k is a field, the ring of germs of k-valued functions at P has a distinguished maximal ideal consisting of the subset of germs vanishing at P. However, even if X is an affine variety, this ring takes into account only the topology of X; this ring is also not local, in general.

Now let $X \subset \mathbf{A}^n$ be an affine variety and let $P \in X$. If f, g are regular functions on X with $g(P) \neq 0$, then the ratio f/g defines a k-valued function on the (Zariski open) neighbourhood $U = \{Q \in X \mid g(Q) \neq 0\}$ of P, as remarked above. This k-valued function is said to be *regular at* P. If φ_1

and φ_2 are two such functions, obtained from ratios of regular functions on X, and which are both regular at P, they determine the same germ at P precisely if $\varphi_1(Q) = \varphi_2(Q)$ for all Q in a neighbourhood of P. The germs of such k-valued functions on X, obtained from ratios of regular functions, which are regular at P, form a local ring with the maximal ideal consisting of those germs vanishing at P; it is a subring of the ring of germs of arbitrary k-values functions considered in the preceeding paragraph. This ring is local, because for any ratio φ of regular functions on X which is regular and non-zero at P, $1/\varphi$ is also such a function. This local ring, denoted by $\mathcal{O}_{X,P}$, is also called the *local ring of X at P*.

Clearly, regular functions on an affine variety X determine germs of rational functions regular at P, for any point $P \in X$. Thus there is a natural homomorphism $\Gamma(X) \to \mathcal{O}_{X,P}$. If $g \in \Gamma(X)$ does not vanish at P, the germ of g is a unit in $\mathcal{O}_{X,P}$. This implies that elements of the multiplicative set $\Gamma(X) - \mathcal{M}_P$ map to units in $\mathcal{O}_{X,P}$, so that from the universal property of localisation, there is an induced homomorphism $\rho : \Gamma(X)_{\mathcal{M}_P} \to \mathcal{O}_{X,P}$. The Nullstellensatz implies that this is an isomorphism; the proof is sketched in the exercise below.

Exercise 8.1.6: Let $X \subset \mathbf{A}^n$ be an affine variety with coordinate ring A, and let $P \in X$. Show that the above homomorphism $\rho : \Gamma(X)_{\mathcal{M}_P} \to \mathcal{O}_{X,P}$ is an isomorphism.
(Hint: Clearly ρ is surjective. If $\rho(f/g) = 0$ where $f, g \in \Gamma(X)$ and $g \notin \mathcal{M}_P$, then the closed set $\{Q \in X \mid f(Q) = 0\}$ contains a neighbourhood of P. Hence f vanishes on each irreducible component of X which contains P. If Y is the union of the irreducible components of X not containing P, and $J \subset R$ is its ideal, then $J + \mathcal{M}_P = \Gamma(X)$, from the Nullstellensatz. Deduce that there exists $h \in J - \mathcal{M}_P$, and show that $hf = 0$ in $\Gamma(X)$; hence $f/g = 0$ in $\Gamma(X)_{\mathcal{M}_P}$.)

Exercise 8.1.7: Let $X \subset \mathbf{A}^n$ be an affine variety with coordinate ring $\Gamma(X)$, $f \in \Gamma(X)$, and $P \in D(f)$, so that $f(P) \neq 0$. Then there is a natural isomorphism $\mathcal{O}_{X,P} \cong \mathcal{O}_{D(f),P}$.
(Hint: Show that $D(f) \subset \mathbf{A}^{n+1}$ has coordinate ring $\Gamma(X)_f$. Then the exercise follows from the previous one.)

Exercise 8.1.8: Let X be an affine variety with coordinate ring $\Gamma(X)$. If f is a rational function on X which is regular at all points of X then f belongs

to $\Gamma(X)$.
(Hint: Let $Den(f) = \{g \in \Gamma(X) \mid \ gf \in \Gamma(X)\}$. Show that $Den(f)$ is not contained in any maximal ideal.)

Exercise 8.1.9: Let $X \subset \mathbf{A}^n$ be an affine variety with coordinate ring $\Gamma(X)$. Let $f_1, f_2, g_1, g_2, h \in \Gamma(X)$ such that f_1/g_1 and f_2/g_2 determine the same germ at each point of $D(g_1g_2)$. Show that $g_1g_2(f_1g_2 - f_2g_1) = 0$ in $\Gamma(X)$.
(Hint: Since $(f_1/g_1)(P) = (f_2/g_2)(P)$ for all $P \in D(g_1g_2)$, the function $g_1g_2(f_1g_2 - f_2g_1)$ vanishes on $D(g_1g_2)$. But it clearly vanishes on the complement too.)

Proposition 8.1.4 *Let $X \subset \mathbf{A}^n$ be an affine variety with coordinate ring $\Gamma(X) = A/I(X)$. Suppose that X has an open covering $\{U_i\}$, and we are given regular functions $f_i, g_i \in \Gamma(X)$ such that*

(i) $g_i(P) \neq 0$ for all $P \in U_i$, so that f_i/g_i is a k-values function, obtained from a ratio of regular functions on X, which is regular at each point of U_i

(ii) for each i, j and each $P \in U_i \cap U_j$, f_i/g_i and f_j/g_j determine the same germ in $\mathcal{O}_{X,P}$.

Then there is a unique element $f \in \Gamma(X)$ such that for each $P \in U_i$, f and f_i/g_i determine the same germ at P.

Proof: A basis for the Zariski topology is given by sets $D(h)$ with $h \in \Gamma(X)$; hence by refining the above open cover, we may assume $U_i = D(h_i)$ for some $h_i \in \Gamma(X)$. Then $D(h_i) \subset D(g_i)$, so by the Nullstellensatz, $h_i \in \sqrt{(g_i)}$. If $h_i^n = g_ik_i$, then $k_i \neq 0$ on U_i, so f_ik_i/h_i^N represents the same germ as f_i/g_i at each point of U_i. Since $D(h_i) = D(h_i^N)$, we reduce to the case when $h_i = g_i$ for each i. Now exercise 8.1.9 implies that $g_ig_j(f_ig_j - f_jg_i) = 0$ in $\Gamma(X)$, for all i, j. Since $\cup U_i = X$, the Nullstellensatz implies that the g_i^2 generate the unit ideal in $\Gamma(X)$. If $\sum a_ig_i^2 = 1$, take $f = \sum a_ig_if_i$. Then $fg_j^2 = \sum_i a_ig_ig_j^2f_i = \sum_i a_ig_i^2g_jf_j = g_jf_j$ in $\Gamma(X)$, so on $D(g_j)$, f and f_j/g_j have the same germ at each point. □

Corollary 8.1.5 *Let X be an irreducible affine variety with function field $k(X)$. Then*

$$\Gamma(X) = \cap_{P \in X} \mathcal{O}_{X,P},$$

where the intersection takes place in $k(X)$.

The content of this proposition is that a k-valued function on X, which is locally equal to the restriction of a ratio of regular functions, is in fact a regular function on X. Thus the property of a k-valued function on X of being regular is a *local property* for the Zariski topology, like the notion of a continuous function on a topological space.

If X, Y are affine varieties with coordinate rings $\Gamma(X)$, $\Gamma(Y)$, a map $f : X \to Y$ is called a *morphism* of affine varieties if the association $h \mapsto h \circ f$ gives a ring homomorphism $f^* : \Gamma(Y) \to \Gamma(X)$, *i.e.*, if regular functions on Y pull back under f to regular functions on X. In particular, an *isomorphism* of affine varieties $f : X \to Y$ is a morphism which is bijective, such that $f^* : \Gamma(Y) \to \Gamma(X)$ is an isomorphism. Clearly, a morphism of affine varieties is continuous for the Zariski topology.

Exercise 8.1.10: If φ is any k-algebra homomorphism $\Gamma(Y) \to \Gamma(X)$, then $\varphi = f^*$ for a unique morphism $f : X \to Y$.
(Hint: If $h : \Gamma(X) \to k$ is any k-algebra homomorphism, we associate to it the homomorphism $h \circ \varphi : \Gamma(Y) \to k$. This gives a map of sets $f : X \to Y$; show that f is a morphism with $f^* = \varphi$.)

Exercise 8.1.11: Let $f : \mathbf{A}^1 \to \mathbf{A}^2$ be the map $t \mapsto (t^2, t^3)$.

1. Show that f is a morphism, whose image is the subvariety $C \subset \mathbf{A}^2$ defined by $Y^2 - X^3 = 0$.

2. Show that f is a homeomorphism $\mathbf{A}^1 \to C$.

3. Show that the inverse homeomorphism $C \to \mathbf{A}^1$ is *not* a morphism.

Thus for affine varieties, morphisms are precisely the same as k-algebra homomorphisms between their coordinate rings (in the other direction!). Thus, any reduced affine k-algebra uniquely determines an affine algebraic variety, upto isomorphism. This variety is denoted by $\operatorname{Spec} A$.

If $f : X \to Y$ is a morphism of varieties, then for each $P \in X$, there is an induced homomorphism of local rings $f^* : \mathcal{O}_{Y,f(P)} \to \mathcal{O}_{X,P}$. In fact this property characterizes morphisms among arbitrary continuous maps, as seen from the exercise below. Again, the exercise may be interpreted as showing that for a set theoretic map between varieties, the property of being a morphism is *local* on X.

Exercise 8.1.12: Let X, Y be affine varieties with coordinate rings $\Gamma(X)$, $\Gamma(Y)$ respectively.

1. Let $f : X \to Y$ be a set-theoretic map, such that for any $g \in \Gamma(Y)$, and any $P \in X$, the germ at P of the k-valued function $g \circ f : X \to k$ lies in $\mathcal{O}_{X,P}$. Then show that f is a morphism.

2. Let $f : X \to Y$ be a continuous map (for the respective Zariski topologies), such that for any $P \in X$, and any $g \in \mathcal{O}_{Y,f(P)}$, the germ $g \circ f$, which is the germ at P of a k-valued function, is actually a germ in $\mathcal{O}_{X,P}$. Then show that f is a morphism.

(Hint: For the first part, note that from proposition 8.1.4, $g \circ f \in \Gamma(X)$. The second part follows from the first.)

We now define the important notion of the *dimension* of an affine variety X, as follows. If X is irreducible, then $k(X)$ is a finitely generated extension field of the algebraically closed field k, and hence has a finite transcendence degree, say d. We define the dimension of X to be d. This is a reasonable notion, since

(i) $\mathbf{A}^n$ has dimension n

(ii) if X is irreducible of dimension n, then by the Noether normalisation theorem (see Theorem (7.4.1)), its coordinate ring $\Gamma(X)$ is an integral extension of a polynomial subring S in n variables. The inclusion of rings $S \subset \Gamma(X)$ corresponds to a morphism $f : X \to \mathbf{A}^n$. Since $\Gamma(X)$ is integral over S, any prime ideal of $\Gamma(X)$ lying over a maximal ideal $\mathcal{M}$ of S is a maximal ideal $\mathcal{M}'$ of R, by corollary (7.3.5); further, for any $\mathcal{M}$, there is at least one $\mathcal{M}'$, and since $\Gamma(X)/\mathcal{M}\Gamma(X)$ is a finite dimensional vector space over $\Gamma(X)/\mathcal{M}$, there are only finitely many $\mathcal{M}'$. Hence for any $P \in \mathbf{A}^n$, all fibres $f^{-1}(P) \subset X$ are finite and non-empty.

For a reducible variety X, we define its dimension to be the maximum of those of its irreducible components. If $P \in X$, we define the *dimension of X at P* to be the dimension of the union of the irreducible components containing P. We denote this by $\dim_P X$. We use $\dim X$ to denote the dimension of X.

In particular, an affine *curve* is an affine variety all of whose irreducible components have dimension 1.

8.2 Projective varieties

The n dimensional projective space over k, denoted by $\mathbf{P}_k^n$ (or just $\mathbf{P}^n$), is defined as the set of equivalence classes of elements of $k^{n+1} - 0$ under the equivalence relation

$$(a_0, a_1, \cdots a_n) \sim (b_0, b_1, \cdots, b_n) \Leftrightarrow \exists \lambda \neq 0 \text{ such that } a_i = \lambda b_i, i = 0, \cdots n.$$

The equivalence class of $(a_0, \cdots, a_n)$, where some $a_i \neq 0$, is denoted by $[a_0, \cdots, a_n]$ (the notation $(a_0 : a_1 : \cdots : a_n)$ is also used). Clearly, $\mathbf{P}^n$ can also be thought of as the set of lines (through the origin) in k^{n+1}, since an equivalence class of elements in $k^{n+1} - 0$ is just the set of non-zero elements of such a line.

Consider the polynomial ring $k[X_0, \cdots, X_n]$ in $(n+1)$ variables. Let F be a homogeneous polynomial in $X_0, \cdots X_n$ of degree d. Even though F does *not* define a function on $\mathbf{P}^n$ we can still talk about the zeros of F in $\mathbf{P}^n$ because, by the homogeneity of F, $F(\lambda a_0, \cdots, \lambda a_n) = \lambda^d F(a_0, \cdots a_n)$.

An ideal $I \subset k[X_0, \cdots, X_n]$ is called homogeneous if for every $F \in I$ all the homogeneous components[1] of F also lie in I. This is equivalent to I being generated by homogeneous elements. If I is a homogeneous ideal in $k[X_0, \cdots, X_n]$ we define the set

$$V(I) = \{[a_0, \cdots, a_n] \in \mathbf{P}^n \mid f(a_0, \cdots a_n) = 0 \quad \forall f \in I\}$$

to be the algebraic set (or *projective variety*) corresponding to I.

Let J be the homogeneous ideal generated by all the homogeneous polynomials vanishing on a projective variety $Z \subset \mathbf{P}^n$. We call the quotient $k[X_0, \cdots X_n]/J$ the *homogeneous coordinate ring* of Z. Since J is a homogenoeus ideal, $R = k[X_0, \ldots, X_n]/J$ is a *graded ring*. The elements of the homogeneous coordinate ring *do not* give functions on Z.

We have the following relationship between projective varieties and affine varieties. Let U_i be the open subset

$$\{[a_0, \cdots a_n] \mid a_i \neq 0\}$$

[1] If $F = F_0 + F_1 + \ldots + F_r$ where F_i is homogeneous of degree i, then F_i are the homogeneous components of F; the set of non-zero homogeneous components of F is well defined.

of $\mathbf{P}^n$. Then it can be seen that the map $U_i \to \mathbf{A}^n$ given by

$$[a_0, \cdots a_n] \mapsto \left(\frac{a_0}{a_i}, \cdots, \frac{a_n}{a_i}\right)$$

is a bijection, with inverse

$$(b_1, \ldots, b_n) \mapsto (b_1, \ldots, b_{i-1}, 1, b_i, \ldots, b_n).$$

which carries regular functions on $\mathbf{A}^n$ to rational functions on $\mathbf{P}^n$ regular on the whole of U_i. Moreover $\mathbf{P}^n$ is the union of $U_i, i = 0, \cdots n$.

If $F(X_0, \ldots, X_n) \in k[X_0, \ldots, X_n]$ is a homogenoeous polynomial, then we see at once that $\{F = 0\} \cap U_i$ is the affine variety in $U_i \cong \mathbf{A}^n$ defined by $\{F(X_0, \ldots, X_{i-1}, 1, X_{i+1}, \ldots, X_n) = 0\}$, where we have set $X_i = 1$ to obtain a polynomial (which need not be homogeneous) in n variables $X_0, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ which we regard as the coordinates on $\mathbf{A}^n$. Hence if $Z \subset \mathbf{P}^n$ is any projective variety, then applying the above remark to each homogeneous polynomial vanishing on Z, we see that the intersections $Z \cap U_i \subset U_i \cong \mathbf{A}^n$ give affine varieties $Z \cap U_i$ in $\mathbf{A}^n$. Thus Z is the union of the affine varieties $Z \cap U_i$.

If we fix one of these $U_i \cong \mathbf{A}^n$ (traditionally one takes U_0 or U_n) then we may regard $\mathbf{P}^n$ as obtained from $\mathbf{A}^n$ by adjoining the 'hyperplane at infinity' $\mathbf{P}^n - U_i = \{a_i = 0\} \cong \mathbf{P}^{n-1}$. In particular, if $n = 2$, $\mathbf{P}^2 - U_i$ is a line, which we call the 'line at infinity'. Points on the hyperplane at infinity are traditionally referred to as "points at infinity", etc.

The collection of algebraic sets, taken as the collection of closed sets, gives a topology on $\mathbf{P}^n$, and hence a topology (the subspace topology) on every projective variety $Z \subset \mathbf{P}^n$. We call this the *Zariski topology* on Z. Clearly each U_i is open in $\mathbf{P}^n$, and its subspace topology is just the Zariski topology obtained by identifying U_i with $\mathbf{A}^n$.

We can still define the notion of a rational function on a projective variety Z: it is an element of the total quotient ring of the homogeneous coordinate ring R which has the form F/G where $F, G \in R$ are *homogeneous of the same degree*. Note that a given element of the total quotient ring may have more than one such expression as a ratio of homogeneous elements. The example given earlier for affine varieties involves ratios of homogeneous polynomials of the same degree, so is easily adapted to give such an example on a projective variety.

A rational function does give a well defined k-valued function on a certain Zariski open subset U of Z. In fact, for each way of writing the rational function as a ratio of two homogeneous elements of equal degree, we obtain (see below) an open set on which this ratio gives a well-defined function; U is then the union of all these open sets, as one checks that the different 'locally defined' functions patch together to give a well-defined function on U. We could also define this open set directly as the complement of the subvariety defined by a suitable homogeneous ideal; we leave this to the reader as an exercise.

To see how such a ratio determines a function, and to determine the open set corresponding to a ratio, consider the following situation: let $P \in Z$ and suppose that an element of the quotient field $k(X_0, \cdots X_n)$ can be written in the form f/g with f, g both homogeneous polynomials of the same degree and $g(P) \neq 0$. Then we get a k-valued function on the neighbourhood $U = \{Q \in Z \mid g(Q) \neq 0\}$ of P in Z defined by

$$Q = [a_0, \cdots, a_n] \mapsto f(a_0, \cdots a_n)/g(a_0, \cdots a_n).$$

Note that this is well defined, because if we take a different representative for Q say $(\lambda a_0, \cdots \lambda a_n)$ then

$$\begin{aligned} f(\lambda a_0, \cdots, \lambda a_n) &= \lambda^d f(a_0, \cdots a_n) \\ g(\lambda a_n, \cdots, \lambda a_n) &= \lambda^d g(a_0, \cdots, a_n) \end{aligned}$$

where d is the common degree of f and g.

As in the case of affine varieties we also have the notion of the local ring of Z at P given by the local ring of germs of k-values functions, obtained as above by restricting rational functions on $\mathbf{P}^n$, which are regular at P. The (equivalent) purely algebraic definition in terms of localisation is as follows: let $\mathcal{P} \subset R$ be the ideal in the homogeneous coordinate ring R of all homogeneous elements which vanish at P (*i.e.*, the images in R of homogeneous polynomials vanishing at P). Then the local ring consists of all ratios F/G in the local ring $R_{\mathcal{P}}$ such that F and G are homogeneous of the same degree.

Definition: An open subset of an affine variety is called a *quasi-affine variety*, an open subset of a projective variety is called a *quasi-projective variety* By a *variety* (or an algebraic variety) we mean a quasi-projective variety.

Note that an affine variety is quasi-affine, and affine, quasi-affine and projective varieties are special cases of quasi-projective varieties.

One can show that if X is any variety, then as a topological space, X is Noetherian, *i.e.*, any descending chain $X_1 \supset X_2 \supset \cdots \supset \cdots$ of Zariski closed subsets of X stabilises. This follows easily from the case of affine varieties discussed earlier. It follows that any variety X can be written uniquely as a finite union $X_1 \cup \cdots \cup X_r$ of closed irreducible subsets such that $X_i \not\subseteq X_j$ for $i \neq j$. The X_i are called the *irreducible components* of X.

From the definitions, it follows easily that if X is either an affine or a projective variety, then the rational functions on X are naturally identified with the rational functions on any dense affine open subset. This motivates the definition of *rational functions* on an arbitrary variety: they are equivalence classes of rational functions defined on dense affine open subsets of X, where rational functions φ_U on U and φ_V on V are equivalent if they define the same rational function on a dense affine open subset of $U \cap V$.

If X is an irreducible variety the set of rational functions form a field which we call the *function field* of X, denoted by $k(X)$. Clearly the function field of X equals that of any non-empty open subset of X.

Similarly, for any point P in a variety X, we may define the local ring of X at P, denoted $\mathcal{O}_{X,P}$, to be the local ring of U at P for any affine open neighbourhood U of P in X; this is clearly independent of the neighbourhood.

Next, we note that if X is an affine or projective variety, and U is an affine open dense subset, then $\dim X = \dim U$. Similarly $\dim_P X = \dim_P U$ if $P \in U$. Hence we may define $\dim X, \dim_P X$ for any variety by defining it to be $\dim U$ or $\dim_P U$, respectivwly, where U is an affine open dense subset (which contains P, in the latter case).

Definition: An *algebraic curve* (or just *curve*) is an irreducible variety of dimension 1. We refer to a projective algebraic curve which is embedded in $\mathbf{P}^2$ as a *plane curve* .

We define the notion of a *regular function* on X: it is a k-valued function which is regular on each affine open subset. This reduces to the usual notion for affine varieties. The ring of regular functions on X is denoted by $\Gamma(X)$. This notion has less importance for arbitrary varieties than in the case of affine varieties, as seen from the exercise below.

Exercise 8.2.1: Show that the regular functions on $\mathbf{P}^n$ are the constant functions.

More generally, we will show later (see (9.2.1)) that the regular functions on any irreducible projective variety are the constant functions. (Contrast this with the case of affine varieties).

8.3 Morphisms

Let X and Y be varieties and $\pi : X \to Y$ a map, continuous for the Zariski topology on X and Y. Let $P \in X$ and $Q = \pi(P)$, and A (respectively B) be the ring of germs of arbitrary k-valued functions (*i.e.*, just set-theoretic functions) at P (respectively at Q). Then π induces a homomorphism $\pi^* : B \to A$ by composition. We call π a *morphism of varieties* if for every P in X, $Q = \pi(P)$, we have $\pi^*(\mathcal{O}_{Y,Q}) \subset \mathcal{O}_{X,P}$. Clearly, if $f : X \to Y$ is a morphism, then for any regular function φ on Y, $f \circ \varphi$ is a regular fucntion on X (here φ is regarded as a k-valued

It is easy to see that this is equivalent to the following: there exists a covering $\{U_i\}_i$ of Y by affine open subsets and a covering $\{V_{ij}\}_j$ of each $\pi^{-1}(U_i)$, again by affine open subsets of X, such that $\pi \mid_{V_{ij}} : V_{ij} \to U_i$ takes algebraic regular functions on the affine variety U_i to those on V_{ij} for all i, j.

Some easy ways to construct morphisms are as follows.

Example 8.3.1:

1. Let Y be an algebraic variety, and X any open subset of Y. Then the inclusion $i : X \to Y$ is a morphism.

2. Let Y be an algebraic variety, X a closed subvariety. Then the inclusion $i : X \to Y$ is a morphism.

3. Let $f : X \to Y$, $g : Y \to Z$ be morphisms. Then $g \circ f : X \to Z$ is a morphism.

4. Let $f : X \to Y$ be a morphism, and $Z \subset Y$ a subvariety such that $f(X) \subset Z$. Then the induced map $X \to Z$ is a morphism.

5. Let $f : X \to Y$ be a map between varieties, such that for some open covering $\{U_i\}$ of Y, $f \mid_{U_i} : U_i \to Y$ is a morphism for each i. Then f is a morphism. In practice, to check a map is a morphism, this reduces one to checking it on affine open sets.

Exercise 8.3.1: For any variety X, show that a morphism $f : X \to \mathbf{A}^1$ is just a regular function on X (namely, the pullback of the coordinate function). More generally, show that for any affine variety Y, morphisms $f : X \to Y$ are in 1-1 correspondence with homomorphisms $\Gamma(Y) \to \Gamma(X)$.

Some further examples of morphisms are given in the following exercises.

Exercise 8.3.2: (i) Let $X \subset \mathbf{P}^n$, and let $X_0, \ldots, X_n$ be the homogeneous coordinates on $\mathbf{P}^n$. Let $F_0, \ldots, F_m \in k[X_0, \ldots, X_n]$ be homogeneous polynomials of degree d, such that $V(F_0, \ldots, F_n) \cap X = \phi$ (*i.e.*, the F_i have no common zeroes on X). Then $f : X \to \mathbf{P}^m$ defined by

$$f([x_0, \ldots, x_n]) = [F_0(x_0, \ldots, x_n), \ldots, F_m(x_0, \ldots, x_n)]$$

is a morphism.

(ii) If $X = \mathbf{P}^n$, then every morphism $f : X \to \mathbf{P}^m$ is as in (i).

Exercise 8.3.3: Let $C \subset \mathbf{P}^2_k$ be defined by the homogeneous polynomial equation $f(X, Y, Z) = Y^2Z - (X^3 + aX^2Z + bXZ^2 + cZ^3) = 0$ (*i.e.*, C is the projective variety corresponding to the principal ideal $f(X, Y, Z)k[X, Y, Z]$.

1. Suppose $c \neq 0$. Then the mapping $\pi : C \to \mathbf{P}^1_k$ given by $[X, Y, Z] \mapsto [X, Y]$ is a morphism. The same formula in fact defines a morphism $\mathbf{P}^2_k - \{[0, 0, 1]\} \to \mathbf{P}^1$, called *projection*.

2. The mapping $\wp : C \to \mathbf{P}^1$ defined by

$$[X, Y, Z] \mapsto \begin{cases} [X, Z] & \text{if } Z \neq 0 \\ [0, 1] & \text{if } Z = 0 \end{cases}$$

is a morphism. This is an important morphism in the study of elliptic curves (it is essentially the Weierstrass $\wp$ function considered earlier in the complex analytic theory).

3. Show that $\wp$ does not have a description as in (i) of exercise 8.3.2.

Definition: A morphism $\pi : X \to Y$ is called an *isomorphism* if there exists a morphism $\varphi : Y \to X$ such that the composite $\varphi\pi$ is the identity map of X and the composite $\pi \circ \varphi$ is the identity map of Y.

Lemma 8.3.1 *Let $f : X \to Y$ be a morphism. Then f is an isomorphism $\Leftrightarrow$ f is bijective and $f^* : \mathcal{O}_{Y,f(P)} \to \mathcal{O}_{X,P}$ is an isomorphism for all $P \in X$.*

Proof: Let us first assume that f is an isomorphism. Then since there is an inverse $g = f^{-1} : Y \to X$ by definition such that the two composites are identities, we immediately see that f is bijective. If $P \in X$ and $Q = f(P)$ one has $g(Q) = P$. Also one has the induced maps f^* and g^* whose composites are identities. Thus the map $f^* : \mathcal{O}_{Y,Q} \to \mathcal{O}_{X,P}$ is an isomorphism.

Conversely if f is bijective, then we can define the set map $g = f^{-1} : Y \to X$ as the inverse of f. We need to check that this is indeed a morphism. By definition of a morphism it suffices to check that g is continuous for the Zariski topology. Which of course will follow if we can check that f is a closed map. So let $Z \subset X$ be a closed subset and let Z' be the closure of $f(Z)$. If $Z' \neq f(Z)$, let $Q \in Z' - f(Z)$. By hypothesis there exists a point $P \in X$ such that $f(P) = Q$. $P \notin Z$. Since Z is closed, there exists an element $\varphi \in \mathcal{O}_{X,P}$ such that $\varphi(P) \neq 0$ and $\varphi(z) = 0$ for all $z \in Z$ where φ is regular. By assumption $\varphi \in \mathcal{O}_{Y,Q}$. Thus φ is regular at Q and does not vanish there. So there exists a Zariski open neighbourhood U of Q where φ is regular and it does not vanish at any point in U. Since $Q \in Z'$ we see that $U \cap f(Z) \neq \emptyset$. But φ vanishes on Z implies that it vanishes on $U \cap f(Z)$. This is a contradiction. □

Two varieties are said to be *isomorphic* if there is an isomorphism between them.

8.4 Smooth varieties

We first define the notion of a smooth point on a variety. If X is a variety, $P \in X$, and $\mathcal{M}_P$ the maximal ideal of the local ring $\mathcal{O}_{X,P}$, then P is a *smooth point* of X (equivalently, X is *smooth* or *non-singular* at P) if $\dim_k \mathcal{M}_P/\mathcal{M}_P^2 = \dim_P X$. By Nakayama's lemma, $\dim_k \mathcal{M}_P/\mathcal{M}_P^2$ is the minimal number of generators[2] of the ideal $\mathcal{M}_P$. If $P \in X$ is a smooth point and $\dim_P X = n$, then there exists a set of elements $t_1, \cdots, t_n \in \mathcal{M}_P$ which minimally generate

[2]Hence X is smooth at $P \Leftrightarrow$ the local ring $\mathcal{O}_{X,P}$ is a *regular local ring* in the sense of commutative algebra, since one can show that $\dim_P X$ equals the Krull dimension of the local ring $\mathcal{O}_{X,P}$.

$\mathcal{O}_{X,P}$; such a set $\{t_1, \ldots, t_n\}$ is called a *regular system of parameters* at P. These are the algebraic analogue of *local coordinates* on X near P.

The following result is an algebraic analogue of the *implicit function theorem*; in fact, it implies (using the usual implicit function theorem) that over the complex number field $\mathbf{C}$, an affine variety is smooth at a point precisely if it is a *complex sub-manifold*[3] of $\mathbf{C}^n$ in a neighbourhood of that point.

Theorem 8.4.1 *(Jacobian Criterion for affine varieties). Let $X \subset \mathbf{A}^n$ be an affine variety whose ideal $I(X) \subset A = k[x_1, \ldots, x_n]$ is generated by the polynomials $f_1, \cdots, f_r$. Then X is nonsingular at P if and only if the Jacobian matrix*

$$J = \left(\frac{\partial f_i}{\partial x_j}(P) \right)$$

has rank equal to $n - \dim_P X$, i.e., the codimension at P of X in $\mathbf{A}^n$.

Proof: Let $P \in \mathbf{A}^n$ have coordinates $(a_1, \cdots, a_n)$, and let $\mathcal{M} = (x_1 - a_1, \ldots, x_n - a_n)$ be the maximal ideal of P in the polynomial ring $k[x_1, \ldots, x_n]$. Define a linear transformation $T : k[x_1, \ldots, x_n] \to k^n$ by

$$T(f) = \left(\frac{\partial f}{\partial x_1}(P), \cdots, \frac{\partial f}{\partial x_n}(P) \right).$$

The elements $T(x_i - a_i) \in k^n$ form a basis, and $T(\mathcal{M}^2) = 0$. Thus T induces an isomorphism $\mathcal{M}/\mathcal{M}^2 \cong k^n$.

Let I be the ideal of X in $k[x_1, \ldots, x_n]$, and let $f_1, \ldots, f_r$ generate I. Then the rank of the Jacobian matrix J is the dimension of $T(I)$ as a subspace of k^n. Using the isomorphism $\mathcal{M}/\mathcal{M}^2 \cong k^n$ induced by T, this is the same as the dimension of $(I + \mathcal{M}^2)/\mathcal{M}^2$. On the other hand, the local ring of X at P is obtained by localising $k[x_1, \ldots, x_n]/I$ at the image of $\mathcal{M}$. Hence if $\mathcal{M}_P$ is the maximal ideal of $\mathcal{O}_P$, then $\mathcal{M}_P/\mathcal{M}_P^2 \cong \mathcal{M}/I + \mathcal{M}^2$. Hence the rank of the Jacobian is equal to $n - \dim_k \mathcal{M}_P/\mathcal{M}_P^2$. This equals the codimension at P of X if and only if $\dim \mathcal{M}_P/\mathcal{M}_P^2 = \dim_P X$. □

[3] A complex manifold is the analogue in arbitrary dimension of the notion of a Riemann surface, considered in Chapter 2.

Exercise 8.4.1: (Jacobian criterion for a projective variety) Let $X \subset \mathbf{P}^n$ be a projective variety with $\dim_P(X) = r$. Let $f_1, \ldots, f_t \in k[X_0, \ldots, X_n]$ be homogeneous polynomials which generate the ideal of X. Let $P = [a_0, \ldots, a_n] \in X$. Show that X is smooth at $P \Leftrightarrow$ the homogeneous Jacobian matrix

$$J = \left(\frac{\partial f_i}{\partial x_j}(a_0, \ldots, a_n) \right)$$

is of rank $n - r$. In particular, if X is the hypersurface with ideal generated by f, then X is smooth at $P \Leftrightarrow$ some partial derivative $f_{X_i}(a_0, \ldots, a_n) \neq 0$. (Hint: Show using Euler's formula that this Jacobian criterion is equivalent to the affine Jacobian criterion on each of the affine open subsets $U_i \cap X$.)

Definition: Let $P \in X$ be a point of an algebraic variety. The *Zariski tangent space* of X at P is defined to be the dual vector space of $\mathcal{M}_P/\mathcal{M}_P^2$, where $\mathcal{M}_P$ is the maximal ideal of the local ring $\mathcal{O}_{X,P}$ of X at P.

Thus a variety X is smooth at a point P precisely when its Zariski tangent space at P has dimension $\dim_P X$.

Exercise 8.4.2:

1. The Zariski tangent space of $\mathbf{A}^n$ can be identified with k^n. In particular, $\mathbf{A}^n$ is non-singular.

2. What is the Zariski tangent space at the origin of $X \subset \mathbf{A}^2$ defined by $xy = 0$? Is X non-singular at the origin?

One can show that the set of smooth points of an irreducible variety form a nonempty open subset. This is clear if X is affine, since the set of non-smooth points is defined by the vanishing of determinants of regular functions, and is hence Zariski closed. The general case follows by covering X by affine open sets.

Theorem 8.4.2 *An affine algebraic curve is smooth at a point if and only if its local ring at that point is a discrete valuation ring.*

Proof: This follows from Theorem (7.5.2), and from the Jacobian criterion. □

Corollary 8.4.1 *Let C be an irreducible affine curve with coordinate ring $\Gamma(C)$. Then C is smooth if and only if $\Gamma(C)$ is a Dedekind domain.*

Proof: This follows from the Theorem, and the characterisation of Dedekind domains in Theorem (7.6.2). □

Chapter 9

Further properties of varieties

9.1 Product varieties

Consider the product $\mathbf{P}^m \times \mathbf{P}^n$. This can be considered as a projective variety as follows.

The tensor product $k^{m+1} \otimes k^{n+1}$ is isomorphic to $k^{(m+1)(n+1)}$ under the natural map

$$(a_0, \ldots, a_m) \otimes (b_0, \ldots, b_n) \mapsto (a_0 b_0, \ldots, a_i b_j, \ldots, a_m b_n).$$

This clearly induces a map $S_{m,n} : \mathbf{P}^m \times \mathbf{P}^n \to \mathbf{P}^{(m+1)(n+1)-1}$. Consider the affine open subsets $U_i = \{X_i \neq 0\} \subset \mathbf{P}^m$, $V_j = \{Y_j \neq 0\} \subset \mathbf{P}^n$. Then $U_i \cong \mathbf{A}^m$, $V_j \cong \mathbf{A}^n$. Then $S_{mn}(U_i \times V_j) =$ Image $(S_{m,n}) \cap \{Z_{ij} \neq 0\} = W_{ij}$, say.

Proposition 9.1.1 *$S_{m,n}$ is injective, and its image is a Zariski closed subset of $\mathbf{P}^{(m+1)(n+1)-1}$; hence W_{ij} defined above is an affine variety. Further, S_{mn} restricts to an isomorphism of affine varieties $U_i \times V_j \to W_{ij}$.*

Proof: Suppose $P = [a_0, \ldots, a_m]$, $P' = [a'_0, \ldots, a'_m]$ are points in $\mathbf{P}^m$, and $Q = [b_0, \ldots, b_n]$, $Q' = [b'_0, \ldots, b'_n]$ are points in $\mathbf{P}^n$ such that $S_{m,n}(P, Q) = S_{m,n}(P', Q')$. Then there exists a non-zero $\lambda \in k$ such that for all i, j we have $a_i b_j = \lambda a'_i b'_j$. Replacing a'_i by $\lambda a'_i$ for all i (which doesn't change the point P') we may assume $\lambda = 1$. Suppose $a_i \neq 0$ and $b_j \neq 0$ (such a pair of indices exist, since $(a_0, \ldots, a_m)$ and $(b_0, \ldots, b_m)$ represent points in projective spaces). Then $a'_i b'_j \neq 0$. Choose homogeneous coordinates for

P, Q, P', Q' with $a_i = b_j = a'_i = b'_j = 1$. Then for all l, $a_l b_j = a'_l b'_j \Rightarrow a_l = a'_l$ for all l; similarly $a_i b_l = a'_i b'_l \Rightarrow b_l = b'_l$ for all l. Hence $P = P'$, $Q = Q'$. Thus $S_{m,n}$ is injective.

Let Z_{ij}, $0 \leq i \leq m$, $0 \leq j \leq n$ be homogeneous coordinates on $\mathbf{P}^{(m+1)(n+1)-1}$. We claim that the image of $S_{m,n}$ is the subvariety defined by the homogeneous equations

$$Z_{ij} Z_{kl} - Z_{il} Z_{kj} = 0 \ \ \forall 0 \leq i, k \leq m, \ \ 0 \leq j, l \leq n.$$

Clearly any point in the image of $S_{m,n}$ satisfies these equations. Suppose $c_{ij}, 0 \leq i \leq m, \ 0 \leq j \leq n$ is a solution of the system of equations, corresponding to a point in $\mathbf{P}^{(m+1)(n+1)-1}$. Regarding $[c_{ij}]$ as a matrix, the above equations express the condition that the matrix has rank ≤ 1, since the 2×2 minors vanish. Since $[c_{00}, \ldots, c_{mn}]$ represents a point in projective space, the rank is exactly 1. Suppose the i^{th} row is not identically 0. Let the j^{rmth} row equal the product of the i^{th} row by a_j. Let $P = [a_0, \ldots, a_m]$ where $a_i = 1$, and let $Q = [c_{i0}, \ldots, c_{in}]$. One verifies at once that $S_{mn}(P, Q) = [c_{00}, \ldots, c_{mn}]$.

It remains to prove that $U_i \times V_j \to W_{ij}$ is an isomorphism. With respect to the affine coordinates obtained by setting $X_i = 1$, $Y_j = 1$, $Z_{ij} = 1$, the map $S_{m,n}$ restricts to a map $\mathbf{A}^m \times \mathbf{A}^n \to \mathbf{A}^{mn+m+n}$ given (after permuting the coordinates) by

$$((x_1, \ldots, x_m), (y_1, \ldots, y_n)) \mapsto (x_1, \ldots, x_n, y_1, \ldots, y_n, x_1 y_1, \ldots, x_i y_j, \ldots, x_m y_n).$$

This map is clearly an isomorphism onto its image W_{ij}, with inverse given by projection onto the first mn coordinates. □

The mapping $S_{m,n}$ is called the *Segre embedding* of $\mathbf{P}^m \times \mathbf{P}^n$ in $\mathbf{P}^{(m+1)(n+1)-1}$. We use this mapping to give $\mathbf{P}^m \times \mathbf{P}^n$ a structure as a projective variety. From the proposition, it is clear that this is the unique variety structure on $\mathbf{P}^m \times \mathbf{P}^n$ which restricts to the earlier variety structure on $U_i \times V_j$ for all i, j. Clearly, the projections $p_1 : \mathbf{P}^m \times \mathbf{P}^n \to \mathbf{P}^m$ and $p_2 : \mathbf{P}^m \times \mathbf{P}^n \to \mathbf{P}^n$ are morphisms, since they are known to be morphisms when restricted to $U_i \times V_j$ for all i, j.

Exercise 9.1.1: If $X \subset \mathbf{P}^m$ and $Y \subset \mathbf{P}^n$ are Zariski closed then $X \times Y$ is closed in the topology induced from $\mathbf{P}^{(m+1)(n+1)-1}$ under the above map.

From the above exercise, if X, Y are any two quasi-projective varieties, then $X \times Y$ is a locally closed subset of the projective variety $\mathbf{P}^m \times \mathbf{P}^n$, and hence can be given the structure of a quasi-projective variety.

We have the following universal property for the product variety:
if $Z \to X, Z \to Y$ are morphisms then there exists a unique morphism $Z \to X \times Y$ such that the diagram

$$\begin{array}{ccc} Z & \to & X \\ \downarrow & \searrow & \uparrow \\ Y & \leftarrow & X \times Y \end{array}$$

commutes. Here the morphisms $X \times Y \to X$ and $X \times Y \to Y$ are the projections $(x, y) \to x$ and $(x, y) \mapsto y$ respectively. In fact, there exists a unique such set theoretic map $Z \to X \times Y$, which is seen to be a morphism by restricting to suitable affine open subsets of Z, X, Y.

9.2 Completeness of Projective Varieties

An important property of projective varieties is that they are complete (defined below). This is a property akin to compactness in Topology. The usefulness of this property cannot be overstated. We will start with the definition.

Definition: A variety X is said to be *complete* if for any variety Y, the projection $p : X \times Y \to Y$ is a closed map *i.e.*, image of a closed set is closed.

Notice that affine varieties are in general not complete. For example if we took X and Y to be $\mathbf{A}^1$ then the hyperbola $xy = 1$ is closed in $\mathbf{A}^1 \times \mathbf{A}^1 = \mathbf{A}^2$, while its image under any of the projections is not closed.

We will start with some useful properties of complete varieties.

Proposition 9.2.1 *Let X be a complete variety and Y any variety. If $f : X \to Y$ any morphism, then f is a closed map.*

Proof: Let $i = (Id, f) : X \to X \times Y$ be the graph. Then i is a closed immersion. If p is the second projection, then $f = p \circ i$. If Z is any closed subvariety of X, then $i(Z)$ is a closed subvariety of $X \times Y$ and by the definition, $p(i(Z))$ is closed in Y. But $p \circ i = f$ and so $f(Z)$ is closed. □

Corollary 9.2.2 *Let X be an irreducible complete variety. Then X has no non-constant regular functions.*

Proof: If f is a regular function, it gives a morphism $f : X \to \mathbf{A}^1$. Since X is irreducible and the image of f is closed by the Proposition, the image of f must be either a single point or the whole of $\mathbf{A}^1$. If f is not constant then the image of f must be the whole of $\mathbf{A}^1$. Consider the inclusion $i : \mathbf{A}^1 \to \mathbf{P}^1$. Then $i \circ f$ is a morphism from X to $\mathbf{P}^1$. So again the image must be closed. But the image is $\mathbf{A}^1$, which is not closed in $\mathbf{P}^1$. This is a contradiction. □

Corollary 9.2.3 *If X is a complete irreducible variety, then any morphism from X to an affine variety must be constant.*

Proof: If $f : X \to Y$ is any morphism, where Y is affine, we may embed Y in $\mathbf{A}^n$. Thus composing with this embedding we get a morphism to $\mathbf{A}^n$. Suffices to show that this morphism is constant. Further we have projections $p : \mathbf{A}^n \to \mathbf{A}^1$. Composing again we see that it suffices to show that a morphism from X to $\mathbf{A}^1$ is constant. But this is exactly the content of the prvious corollary. □

Proposition 9.2.4 (Rigidity Lemma) *Let X be a complete irreducible variety, Y, Z any varieties. Let $f : X \times Y \to Z$ be a morphism such that for some $y_0 \in Y$, $f(X \times \{y_0\})$ is a single point $z_0 \in Z$. Then there is a morphism $g : Y \to Z$ such that $f = g \circ p$ where p is the second projection $p : X \times Y \to Y$.*

Proof: Choose a point x_0 in X and define g to be $g(y) = f(x_0, y)$. We will show that f and $g \circ p$ coincide in a nonempty open set, which by virtue of the fact that X is an irreducible variety will imply the result [Exercise]. Let U be an affine open neighbourhood of $z_0 \in Z$. Let $F = Z - U$ and $G = p(f^{-1}(F))$; then G is closed in Y since X is complete. Further $y_0 \notin G$ by our hypothesis. Therefore $Y - G = V$ is a non-empty open set of Y. For each $y \in V$, the complete variety $X \times \{y\}$ gets mapped by f into the affine open set U and hence by the previous corollary must map to a point. So for any $y \in V$ one has $f(X \times \{y\}) = f(x_0, y) = g(y)$. This proves the result. □

Theorem 9.2.1 *Projective varieties are complete.*

Proof: We first reduce to a more tractable situation. So let X be a projective variety and let Y be any variety. Let $p : X \times Y \to Y$ be the second projection. Since X is a closed subvariety of a projective space, it is clear that we may replace X by $\mathbf{P}^n$ for some n. Also to check a set in Y is closed we need only check this after restricting to an open cover. Thus we may replace Y by an affine variety. But then Y is a closed subvariety of some affine space and it is clear that we may replace Y by this affine space. So we are finally reduced to the following situation: $p : \mathbf{P}^n \times \mathbf{A}^m \to \mathbf{A}^m$ is the second projection and Z a closed subvariety of $\mathbf{P}^n \times \mathbf{A}^m$. We want to prove that $p(Z)$ is closed.

Let Z be defined by finitely many polynomials $f_1(u, v), f_2(u, v), \ldots, f_k(u, v)$ where the components of u are the coordinates in $\mathbf{P}^n$ and v the coordinates in $\mathbf{A}^m$. Then a point v_0 is in $p(Z)$ if and only if the $f_i(u, v_0)$'s have a common nonzero solution. Let I_s be the vector space of homogeneous polynomials in u of degree s. By Hilbert's Nullstellensatz, the $f_i(u, v_0)$'s have a common nonzero solution if and only if the ideal generated by f_i's does not contain I_s for any s. Let T_s be the set of points v_0 in $\mathbf{A}^m$ such that the ideal generated by $f_i(u, v_0)$'s does not contain I_s. Then $p(Z) = \cap T_s$. We will show that each T_s is closed. This will finish the proof.

Let $k_i = s - \deg_u f_i$. Let N_α be any monomial in u of degree k_i. Then $g_{i\alpha} = N_\alpha f_i$ is a degree s polynomial. Let $x \in I_s$ which is in the ideal generated by the $f_i(u, v_0)$'s for some $v_0 \in \mathbf{A}^m$. Write $x = \sum p_i(u) f_i(u, v_0)$. Let q_i be the homogeneous part of degree k_i in p_i. Then by equating homogeneous parts it is clear that $x = \sum q_i(u) f_i(u, v_0)$. Since q_i is homogeneous of degree k_i, we may write $q_i = \sum r_{i\alpha} N_\alpha$, where $r_{i\alpha}$ belong to the base field. Thus we see that $x = \sum r_{i\alpha} g_{i\alpha}(u, v_0)$. So we see that, if I_s is not contained in the ideal generated by $f_i(u, v_0)$'s, then I_s is not contained in the vector space generated by $g_{i\alpha}(u, v_0)$. But the g's have all degree s in u. So if m_α form a vector space basis for I_s then we can write $g_{i\alpha} = \sum a_{i\alpha}(v) m_\alpha$. So for a $v_0 \in \mathbf{A}^m$, I_s is not contained in the ideal generated by $g_{i\alpha}(u, v_0)$ is equivalent to saying that the matrix $(a_{i\alpha}(v_0))$ has rank less than d, the dimension of I_s. In other words such v_0's are precisely defined by the vanishing of all $d \times d$ minors of the matrix $(a_{i\alpha}(v))$. So T_s is a closed set. □

Thus the results that we have proved for complete varieties apply to projective varieties.

9.3 Proper morphisms

In this section we study proper morphisms. These morphisms have many useful properties.

Let $X \stackrel{\psi}{\to} Z$ and $Y \stackrel{\pi}{\to} Z$ be morphisms of varieties. Consider the subset

$$\{(x, y) \in X \times Y \mid \psi(x) = \pi(y)\}$$

This is closed in $X \times Y$ (Exercise: Prove this). We define $X \times_Z Y$, the product of X, Y over Z, to be this closed subvariety. Note that we have the commutative diagram:

$$\begin{array}{ccc} X \times_Z Y & \to & X \\ \downarrow & & \downarrow \psi \\ Y & \stackrel{\to}{\pi} & Z \end{array}$$

Definition: A morphism $X \stackrel{\psi}{\to} Z$ is called *proper* if for every morphism $Y \stackrel{\pi}{\to} Z$ the induced morphism $X \times_Z Y \to Y$ is a closed map.

Let us write down a few obvious properties of proper maps.

Proposition 9.3.1 *1. Let $f : X \to Y$ be a morphism and let U_i be any open cover of Y. If $f : f^{-1}(U_i) \to U_i$ is proper for all i, then f is proper.*

2. *If X is a closed subvariety of Y then the inclusion $i : X \hookrightarrow Y$ is proper.*

3. *If $f : X \to Y$ is proper and $g : Z \to Y$ is any morphism, then the* base change,
$$f \times Id : X \times_Y Z \to Z$$
is proper.

4. *If $f : X \to Y$ and $g : Y \to Z$ are proper, then $g \circ f : X \to Z$ is proper.*

A not so obvious property of proper maps is the following:

Proposition 9.3.2 *Let $f : X \to Y$ and $g : Y \to Z$ be morphisms such that $h = g \circ f$ is proper. Then f is proper.*

Proof: Since h is proper, one gets by the previous proposition that the base change $h' = h \times Id : X \times_Z Y \to Y$ is proper. We have the identification of X with the graph of f, $\Gamma_f \subset X \times_Z Y$, and thus by previous proposition the composite $X \to \Gamma_f \to Y$ is proper. This composite is clearly f. □

Theorem 9.3.1 *If X, Y are projective varieties then any morphism $\pi : X \to Y$ is proper.*

Proof: This follows from Theorem (9.2.1). □

Exercise 9.3.1: Let $A \subset B$ be affine algebras and assume that B is integral over A. Let $X = \operatorname{Spec} B$ and $Y = \operatorname{Spec} A$ and f the induced morphism from X to Y. Show that the image of a closed set is closed. Deduce that f is proper.

Chapter 10

Intersection theory for plane curves

10.1 Intersection multiplicities

Let $R = k[x, y, z]$ a polynomial ring in three variables and $A = k[x, y]$ a polynomial ring in two variables. We first define maps $D : R \to A$ and $h : A \to R$ as follows. Let

$$D(F(x, y, z)) = F(x, y, 1)$$

If $f \in A$ is written as $f = f_d + f_{d-1} + \cdots + f_0$ where f_i is homogeneous of degree i and $f_d \neq 0$ then we define

$$h(f(x, y)) = f_d + zf_{d-1} + \cdots + z^d f_0.$$

Here h stands for homogenisation and D stands for dehomogenisation.

These maps have the following properties:

1. Let $F \in R$ and assume that F is homogeneous and $F = z^k G$ with $G(x, y, 0) \neq 0$. Then $h(D(F)) = G$.
2. Let $f \in A$. Then $D(h(f)) = f$.
3. $D(FG) = D(F)D(G)$; $h(fg) = h(f)h(g)$.
4. $D(F+G) = D(F)+D(G)$; given $f, g \in k[x, y]$ there exists non-negative integers l, m, n not all zero such that $z^l h(f) + z^m h(g) = z^n h(f + g)$.

We will first state an elementary lemma. The proof is left as an exercise.

Lemma 10.1.1 *Let $V_d \subset k[x_0, x_1, \ldots, x_n]$ be the vector space of degree d homogeneous polynomials. Then* $\dim V_d = \binom{n+d}{d}$.

Now we embark on our aim of defining intersection multiplicities. For this let us first consider two elements $f, g \in A$, which have no common factor. For convenience we will also assume our field k is algebraically closed.

Lemma 10.1.2 *Let f, g be as above. Then the ring $B = A/(f, g)$ is a finite dimensional vector space (over k).*

Proof: First observe that if $f = f_1 f_2$, then the kernel of the natural surjective map $B \to A/(f_2, g)$ is isomorphic to $A/(f_1, g)$. This is succinctly expressed by saying that

$$0 \to A/(f_1, g) \xrightarrow{f_2} B \to A/(f_2, g) \to 0 \tag{10.1}$$

is *exact*. So to prove the lemma it suffices to prove it for the two rings $A/(f_1, g)$ and $A/(f_2, g)$. So we may as well assume that f is irreducible. Now by lemma (7.4.1) we may assume that f is monic in y. Let $S = A/(f)$. Then S is a finite module over $k[x]$. Let $t^n + a_1 t^{n-1} + \cdots + a_n = 0$ be the minimal degree integral equation for the image of g in S over $k[x]$. Then the ideal generated by g in S contains a_n. Since B is a quotient of $S/(a_n)$, it suffices to show that this ring is finite dimensional. But it is a finite module over $k[x]/(a_n)$. So it suffices to show that $k[x]/(a_n)$ is a finite dimensional vector space. Now $a_n \neq 0$ since S is a domain and $g \neq 0$ in S. Thus, if a_n is a polynomial of degree d in x it is easy to see that $k[x]/(a_n)$ is a d-dimensional vector space. □

Exercise 10.1.1: Let R be a k-algebra which is finite dimensional. Show that $R \cong R_1 \times R_2 \times \cdots \times R_k$, where R_i's are k-algebras with a unique maximal ideal. (Hint: Chinese Remainder theorem).

With these preliminaries we can begin the definition of intersection multiplicities. Let $f(x, y) = 0$ and $g(x, y) = 0$ define two curves C, D in $\mathbf{A}^2$. Assume that $f(0, 0) = g(0, 0) = 0$. We wish to define an intersection multiplicity of C and D at $P = (0, 0)$. Loosely speaking, we want to define an integer which will reflect how tangential the two curves are at P. For instance if both the curves are non-singular at P and have distinct tangents

at P, then we should get 1 as their intersection number. If they were simply tangent to each other, then we should get 2, etc.

Let us make this precise. First we may write $f = pf_1$ and $g = pg_1$ where p is the greatest common factor of f and g. If $p(0,0) = 0$ then intersection multiplicity is not defined at P. If not, then intersection multiplicity of C and D at P is the same as that of the two curves defined by $f_1 = 0$ and $g_1 = 0$. So for the definition we may as well assume that f, g have no common factors.
a) If one of f, g does not vanish at $P = (0,0)$, then we define their intersection multiplicity at P to be zero.

Consider $B = k[x,y]/(f,g)$ and the maximal ideal $\mathcal{M} = (x,y)$. Then $\mathcal{M}^n \subset \mathcal{M}^{n-1}$ is a decreasing chain of vector spaces in B, which is a finite dimensional vector space by lemma (10.1.2). Thus there exists an N such that $\mathcal{M}^n = \mathcal{M}^{n+1}$ for all $n \geq N$. Let $B' = B/\mathcal{M}^N$.
b) In general we define the intersection multiplicity of C and D at P as

$$I(P, C \cap D) = \dim_k B'.$$

Sometimes we will also write this as $I(P, f, g)$.

Example 10.1.1: If $f = x$ and $g = y$ then $I(0, x, y) = 1$ as is suggested intuitively. On the other hand if $g = x + y^n$ then one can easily see that $I(0, x, g) = n$. Let us compute a slightly less trivial example. Let $f = x^2 + y^3$ and $g = x^3 + y^2$. Then the ideal $(f, g) = (f, -xy^3 + y^2)$. Now by ((10.1)) we see that

$$I(0, f, g) = I(0, f, y^2) + I(0, f, -xy + 1)$$

But the second term is zero since $-xy + 1$ does not vanish at 0. The ideal $(f, y^2) = (x^2, y^2)$. Now an easy computation will show that $k[x,y]/(x^2, y^2)$ has a unique maximal ideal $\mathcal{M}$ corresponding to $(0,0)$ and $\mathcal{M}^3 = 0$. Thus B' in the above definition is $k[x,y]/(x^2, y^2)$ and its vector space dimension can easily seen to be 4. Thus $I(0, f, g) = 4$.

Example 10.1.2: Let us compute a different but general example. Let $f = y + h$ where h has order at least two at $P = (0,0)$. That is to say $h = ax^2 + bxy + cy^2$ for some polynomials a, b, c. Let g be any polynomial. We wish to compute $I(0, f, g)$. If $g(0,0) \neq 0$ then clearly the intersection multiplicity is zero. So assume that $g(0,0) = 0$.

Let us compute first $I(P, f, x)$. We may write $f = yq(y) + x\tilde{f}$ where $q(0) = 1$. Thus $I(P, f, x) = 1$. From this it is easy to see, using the exact sequence we wrote down, that $I(P, f, x^d) = d$ for any positive integer d.

Now we claim that $I(P, f, g) \geq d$ if and only if $q^d g \in (f, x^d)$. We prove this by induction on d. Notice that if $d = 1$, then since $(f, x) = (yq(y), x)$ and since $g = a(y)y + bx$ one has $qg \in (f, x)$. Assume we have proved the result for $d - 1$. Thus $q^{d-1}g \in (f, x^{d-1})$. Writing this explicitly we have

$$q^{d-1}g = \lambda f + \mu x^{d-1}.$$

Thus $d = I(P, f, g) = I(P, f, q^{d-1}g)$, since $q(0) \neq 0$. This in turn is equal to

$$I(P, f, \lambda f + \mu x^{d-1}) = I(P, f, \mu x^{d-1}) = d - 1 + I(P, f, \mu).$$

Thus $I(P, f, \mu) \geq 1$ and now applying the first part we get that $q\mu \in (f, x)$. Putting all these together, one gets the first implication.

Conversely if $q^d g \in (f, x^d)$, then since $I(P, f, g) = I(P, f, q^d g)$, $I(P, f, x^d) = d$ and $(f, q^d g) \subset (f, x^d)$, so that $k[x, y]/(f, q^d g)$ surjects onto $k[x, y]/(f, x^d)$, we get the result.

This claim proves that $(P, f, g) = d$ if and only if $q^d g \in (f, x^d)$ and $q^{d+1}g \notin (f, x^{d+1})$. This will be a very useful fact to compute intersection multiplicities. In particular we note that $I_P(f, g) = 1$ if and only if the function g gives a *uniformising parameter* (*c.f.* 7.5.2) at the point P on the curve $f = 0$.

Exercise 10.1.2: Let f be as above and t be any polynomial with $I(P, f, t) = 1$. Then there exists a u with $u(P) \neq 0$ such that for any g, $I(P, f, g) = d$ if and only if $u^d g \in (f, t^d)$ and $u^{d+1}g \notin (f, t^{d+1})$.

Exercise 10.1.3: Let notation be as in the previous exercise. Let $I(P, f, g) = d$. Then show that there exists a v with $v(P) \neq 0$ so that for any h, $I(P, f, h) \geq e$ if and only if $v^d t^n h \in (f, g)$ where $n = \max(0, d - e)$.

Let now C and D be two curves in $\mathbf{P}^2$ defined by $F = 0$ and $G = 0$. Further assume that they have no common components. This is same as saying that F, G have no common factor. Then first notice that they have only finitely many common zeroes. This is because of the following. First let us look at points whose z-coordinates are zero where as usual we will

take x, y, z as the homogeneous coordinates in $\mathbf{P}^2$. Since F and G have no common factor at least one of F or G must be nonzero modulo z, say F. Then $F = F_0 + zF_1$ where F_0 is a non-zero homogeneous polynomial in x and y. But then F_0 has only finitely many zeroes upto scalars. Since all common zeroes of F and G with z-coordinate zero must have one of these points as its (x, y)-component, there are only finitely many such. So we need to show that the rest of the common zeroes are finite. These have z-coordinate 1 and hence we are looking at the common zeroes of $f = D(F)$ and $g = D(G)$. Since F and G have no common factors neither do f and g. Now by lemma (10.1.2) we know that $k[x, y]/(f, g)$ is finite dimensional. This says that they have only finitely many common zeroes.

Since intersection multiplicity should depend only at the points of intersection of C and D we may rigorously define it now as follows. Since there are only finitely many intersection points choose coordinates so that $z = 0$ does not pass through any of them. If $f, g \in k[x, y]$ are the new dehomogenised polynomials, then we may as before define intersection multiplicity for any point in their intersection. So finally define

$$I(C, D) = \sum_{P \in C \cap D} I(P, C \cap D).$$

We may also use the notation $I(P, F, G)$ to mean $I(P, C \cap D)$.

Exercise 10.1.4: Using the previous exercise and with notation as in the previous paragraph, prove that $I(C, D) = \dim_k k[x, y]/(f, g)$.

Using this definition we prove a classical theorem.

Theorem 10.1.1 (Bezout's Theorem) *Let C and D be as above. Let $m = \deg C = \deg F$ and $n = \deg D = \deg G$. Then*

$$I(C, D) = mn$$

Proof: Since $C \cap D$ is finite we may after a change of variables assume that all the common zeroes have non-zero z-coordinates. So by exercise (10.1.1) one has

$$I(C, D) = \dim_k k[x, y]/(D(F), D(G))$$

Let $f = D(F)$ and $g = D(G)$. Let $A = k[x,y]/(f,g)$, $R = k[x,y,z]$ and $B = R/(F,G)$. Let R_d be the vector space of homogeneous polynomials of degree d in R and B_d its image in B.
Step 1. $\dim B_d = mn$ if $d \geq m+n$.

Consider the map $R \oplus R \xrightarrow{\pi} R$ given by $\pi(a,b) = aF + bG$. Also consider the map $R \xrightarrow{i} R \oplus R$, given by $i(a) = (aG, -aF)$. Then $\pi \circ i = 0$. Also if $(a,b) \in R \oplus R$ such that $\pi(a,b) = 0$, then using the fact that F and G have no common factors (and R is a UFD) one sees that there exists $c \in R$ such that $(a,b) = (cG, -cF)$. What all this means is that the following is an exact sequence

$$\rightarrow R \xrightarrow{i} R \oplus R \xrightarrow{\pi} R \rightarrow B \rightarrow 0.$$

Restricting to homogeneous components, one gets an exact sequence of vector spaces

$$\rightarrow R_{d-m-n} \rightarrow R_{d-m} \oplus R_{d-n} \rightarrow R_d \rightarrow B_d \rightarrow 0.$$

The ranks of the vector spaces R_k are given by Lemma 10.1.1 The alternating sum of these ranks then gives required rank for B_d.
Step 2. The map from B_d to B_{d+1} given by multiplication by z is injective and hence an isomorphism if $d \geq m+n$.

We have to show that if $H \in R$ is homogeneous and such that $zH \in (F,G)$ then $H \in (F,G)$. So write $zH = aF + bG$ where everything is homogeneous. Write $a = a_0 + za_1$ where $a_0 \in k[x,y]$, $a_1 \in k[x,y,z]$. Similarly for b, F, G. The fact that C and D have no common zeroes along $z = 0$ implies that F_0 and G_0 have no common factors. Since $a_0F_0 + b_0G_0 = 0$ we get $a_0 = pG_0$ and $b_0 = -pF_0$. Thus $a = pG + zq$ and $b = -pF + zr$. But then $zH = zqF + zrG$ which implies that $H \in (F,G)$.
Step 3. Let $A_i \in R_d, \quad i = 1, \ldots, mn$ whose images in B_d form a basis for it for $d \geq m+n$. Then $D(A_i) \in A$ form a basis for A.

Let $a_i = D(A_i)$. First we show that they are linearly independent in A. Assume that $\sum k_i a_i = 0$ in B for some $k_i \in k$. So $\sum k_i a_i = qf + rg$ in the polynomial ring $k[x,y]$. This in turn implies, by Property 4 of the maps h, D stated in the beginning of this section, that $z^a \sum k_i A_i = z^b h(q)F + z^c G$. If the degree of these polynomials is $d' \geq d$ then since the images of $z^a A_i$ in $R_{d'}$ also form a basis of B_d by step 2, we get $k_i = 0$ for all i.

Finally let $s \in A$ and let $s' \in k[x,y]$ be any lift of s. Then we may write the image of $z^l h(s')$ in B as a linear combination of A_i's, if l is large enough so that the degree of $z^l h(s')$ is bigger than $m+n$. That means

$z^l h(s') = \sum k_i z^a A_i + pF + qG$. Putting $z = 1$ (which is just applying D), we get $s' = \sum k_i a_i + D(p)f + D(q)g$ and hence $s = \sum k_i a_i$ in A. □

10.2 Applications of Bezout's Theorem

Bezout's theorem is one of the most basic theorems in the theory of plane curves. It has many interesting corollaries. We state some of the easier ones. The theory of intersection multiplicities which had its beginning in Bezout's theorem is an imposing subject today.

Corollary 10.2.1 *Let C and D be two curves of degrees m and n respectively. If they meet in exactly mn distinct points, then these points are smooth on both the curves.*

Corollary 10.2.2 *Let the notation be as before. If the curves intersects in more than mn points then the curves have a common component.*

Let C be a plane curve and $P \in C$. Let L be the tangent line at P, *i.e.*, a line L such that $I_P(C, L)$ is maximal among all lines through P. We say that P is a *flex* (short form for *inflexion point*) of C if $I(P, C \cap L) > 2$.

Example 10.2.1: Let C be a non-singular conic. Then by the above corollary it cannot have a flex. Let C be a cubic defined by the equation $y^2 z - f(x, z) = 0$. Then one easily checks that $z = 0$ is a flex of the point $P = (0, 1, 0)$.

Let F be a homogeneous polynomial in x_0, x_1, x_2 of degree n. Let

$$F_{ij} = \frac{\partial^2 F}{\partial x_i \partial x_j}.$$

Then the F_{ij}'s are homogeneous of degree $n - 2$. Consider the determinant $H = \det(F_{ij})$. H is called the *Hessian* of F.

Proposition 10.2.3 *If $F = 0$ is a non-singular curve over a field of characteristic zero, then the points of intersection of $F = 0$ and $H = 0$ are precisely the flexes of $F = 0$.*

Remark: The hypothesis of zero characteristic is necessary. For instance if we took $F = x^y + y^2z + z^2x$, then in characteristic two, it is nonsingular, the point $[0 : 1 : 0]$ is not a flex, but all the second partial derivatives are zero and hence the hessian $H = 0$.

Proof: First one checks that the curve D defined by $H = 0$ does not depend on the coordinates, though its definition *a priori* does. So make a change of coordinates so that the point P we are interested in is $(1, 0, 0)$ and $x_2 = 0$ is the tangent there. Let $f = D(F) = F(1, x_1, x_2)$ and $h = D(H) = H(1, x_1, x_2)$. We have the Euler equations

$$\sum x_i F_{ij} = (n-1)F_j$$

$$\sum x_i F_i = nF$$

Using this one easily computes the Hessian to be

$$h = f_{x_1}^2 f_{x_2x_2} - 2f_{x_1}f_{x_2}f_{x_1x_2} + f_{x_2}^2 f_{x_1x_1} + nf$$

Thus

$$I(P, F, H) = I(P, f, h) = I(P, f, g)$$

where

$$g = f_{x_1}^2 f_{x_2x_2} - 2f_{x_1}f_{x_2}f_{x_1x_2} + f_{x_2}^2 f_{x_1x_1}.$$

By our hypothesis $f = x_2 + ax_1^2 + bx_1x_2 + cx_2^2 + s$, where s has only terms of degree bigger than or equal to three. Using this one can compute g and see that $g = 2a + s'$ where $s'(0, 0) = 0$. Thus $I(P, F, H) > 0$ if and only if $a = 0$. But this is also equivalent to the statement $I(P, f, x_2) > 2$ proving our contention. □

Example 10.2.2: Let C be a non-singular cubic (say in characteristic zero). Then the degree of its Hessian H is also three. So by Bezout's theorem we see that C has nine flexes (check that the Hessian and C are distinct curves, since characteristic is zero, so that they meet at finitely many points). Of course we have not proved that the number of flexes is exactly nine since we do not know that H meets C in distinct points. However, we have

Exercise 10.2.1: If H meets C at P with multiplicity greater than one, then the tangent line at P meets C with multiplicity greater than 3 at P.

Now the tangent line cannot meet the cubic at the flex more than thrice by Bezout's theorem. Thus, each point of intersection of H and C has intersection multiplicity 1 and so there are exactly nine distinct points of intersection.

Here is something to mull over. Let C be a non-singular curve defined by $F = 0$. Then by Euler's theorem the polynomials F_x, F_y, F_z have no non-trivial zeroes. So it gives a morphism $\phi : \mathbf{P}^2 \to \mathbf{P}^2$ by

$$\phi(a, b, c) = (F_x(a, b, c), F_y(a, b, c), F_z(a, b, c))$$

Then $\phi(C) = D$ is a curve (if $\deg C > 1$), and it is called the *dual curve*.

Exercise 10.2.2: Show that the image of a flex of C is a cusp of D and if a line is tangent to two different points $P, Q \in C$ then $\phi(P) = \phi(Q)$. It is harder to see the fact that for 'most' points $Q \in D$, $\phi^{-1}(Q)$ is a single point (if $\deg C \geq 2$).

Example 10.2.3: Let us again look at the case of a non-singular cubic. Now we know that it has a flex. So without loss of generality let us assume that $P = (0, 1, 0)$ is the flex and $z = 0$ the tangent line at P. Write $F = \sum_0^3 a_i z^i$ where a_i is homogeneous in x, y of degree $3 - i$. The fact that $z = 0$ is a flex at P implies that a_0 is x^3 upto a constant. This constant must be non-zero since our curve is irreducible. So by scaling we may assume that it is 1.

Now write the equation in terms of y. Since there are no degree three terms other than x^3 and the ones which involve z we see that $F = b_1 z y^2 + zly + b_3$ where $b_1 \in k$, $l \in k[x, z]$ homogeneous of degree 1 and $b_3 \in k[x, z]$ is homogeneous of degree 3. If $b_1 = 0$ one easily sees that C is singular at P. So by scaling y we may assume that $b_1 = 1$. Now if characteristic is not 2 then by changing y to $y + \frac{1}{2}l$ we may further assume that the linear term in y is zero. Now putting $z = 1$ we get the *Weierstrass form* of a cubic to be

$$y^2 = f(x) = x^3 + g_1 x^2 + g_2 x + g_3$$

Also non-singularity will force f to have distinct roots. If the characteristic of k is not 3, then we may also make g_1 to be zero by changing x to $x + \frac{1}{3}g_1$.

Now let us make a slightly more complicated computation. We know that the curve has nine flexes. Since one flex is the point P, 'the point at infinity', we should find the remaining eight of them using our Weierstrass

form. So let us start with $y^2 = x^3 + g_2x + g_3$. First one sees that no line of the form $x = constant$ can be the tangent at a flex. This is easy since when we substitute $x = constant$ we get an equation of the form $y^2 = constant$ which clearly has at most a double root. To get a flex we must have a triple root. Now substitute a general equation of a line $y = ax + b$ in the Weierstrass form. We get the polynomial

$$x^3 + g_2x + g_3 - (ax + b)^2.$$

We want to know for what values of a, b this polynomial has a triple root. Now for a polynomial g to have triple roots, we know by elementary calculus that we should find a common root of g, g', g''. Start with g''. This is linear in x. Substitute this in the quadratic polynomial g'. This will give b in terms of a. Then substituting this in g we will get a degree eight polynomial in a which has distinct roots. These will give the remaining eight flexes. The details are left to the reader.

Exercise 10.2.3: Let C be an irreducible *singular* cubic defined by $F = 0$. Show that after a change of variables it is one of the following: a) $y^2z = x^3$; b) $y^2z = x^2(x + z)$.

Now we come to yet another fundamental theorem. We will only prove it in a simpler case since the general version requires more concepts. We might tacitly use the general version though, in some applications.

Theorem 10.2.1 (Max Noether's Theorem) *Let $C : F(X, Y, Z) = 0$ be a smooth plane curve. Let G and H be two homogeneous polynomials in X, Y, Z. Assume that G is not divisible by F. Then there exists A and B such that $H = AF + BG$ if and only if for every point $P \in C$,*

$$I(P, F, H) \geq I(P, F, G).$$

Proof: If $H = AF + BG$ then

$$I(P, F, H) = I(P, F, AF + BG) = I(P, F, BG) \geq I(P, F, G).$$

Let us now prove the converse. As usual we may choose the coordinates so that $z = 0$ does not contain any of the common zeroes of F, G. Now we may dehomogenise and call $D(F) = f$ etc. Under the hypothesis we must

show that $h \in (f,g)$. Let P_i be the points of intersection of f,g and let $d_i = I(P_i,f,g)$. Then by Exercise (10.1.3), we have polynomials v_i such that $v_i(P_i) \neq 0$ and $v_i^{d_i} h \in (f,g)$. Now the ideal $(v_1,\ldots,v_n,f,g)$ is the whole ring. An elementary calculation now shows that $h \in (f,g)$. □

Remark: We really did not use the fact that C is smooth. If we knew that G intersected C in smooth points the proof would have still worked. In fact the theorem is valid more generally, even when the curves meet at singular points. But then our hypothesis should involve intersection multiplicities at all "infinitely near points" and we will not go into it here.

We will derive some corollaries of this now.

Corollary 10.2.4 *Let $F=0$ be a smooth cubic. Assume that a cubic $G=0$ meets $F=0$ at nine distinct points. Then any cubic $H=0$ passing through eight of these points passes also through the ninth.*

Proof: Let us call P this ninth point. Let $L=0$ be a line through P not containing any of the other eight points. Let $L=0$ meet $F=0$ in two more points Q,R. Now by the theorem $LH \in (F,G)$. Writing this as

$$LH = AF + BG$$

with A,B linear we see that since $G(Q)$ and $G(R)$ are not zero, $B(Q) = B(R) = 0$. But then $B = bL$ for some constant b. But then L divides AF and since F is smooth L does not divide F. So $A = aL$ for some constant a and hence $H = aF + bG$. □

Corollary 10.2.5 *If a line cuts a smooth cubic $F=0$ at three distinct points, the residual intersections of the tangents at these points are collinear.*

Proof: Let $G=0$ define the product of the three tangents at the points of intersections and $H = L^2M$ where $L=0$ define the original line and $M=0$ define the line joining two of the residual intersections and apply the theorem. □

Corollary 10.2.6 *The line joining two flexes of a smooth cubic passes through a third flex.*

Proof: Take the line in the above corollary to be the line joining two of the flexes. □

Corollary 10.2.7 *If a conic is tangent to a smooth cubic at three distinct points the residual intersections of the tangents at these points are collinear.*

The following corollaries need the remark we made after the proof of Noether's Theorem.

Corollary 10.2.8 (Pascal) *If a hexagon is inscribed in an irreducible conic, then the opposite sides meet in collinear points.*

Corollary 10.2.9 (Pappus) *Let L_1 and L_2 be two lines and $P_i \in L_1$ and $Q_i \in L_2$ for $i = 1, 2, 3$ be six distinct points . Assume that none of these is the intersection point of L_1 and L_2. Let L_{ij} be the line joining P_i and Q_j. Let $R_k = L_{ij} \cap L_{ji}$. Then the R_k's are collinear.*

Chapter 11

Geometry of curves

11.1 Rational maps

Let X be an irreducible variety and Y any variety. A morphism $U \to Y$, where U is a nonempty open subset of X is called *a rational map* from X to Y. A rational map from a curve is defined except possibly at a finite number of points, since complement of a non-empty Zariski open subset of an irreducible curve is a finite set of points. If a rational map from X to Y has dense image then it induces a homomorphism $k(Y) \to k(X)$ between the function fields.

Example 11.1.1: Let X be an irreducible variety and let f be a rational function on X. Then f gives a rational map from X to $\mathbf{P}^1$. If f is regular at a point P, then at P, f is a morphism and $f(P) \in \mathbf{A}^1 = \mathbf{P}^1 - \{\infty\}$.

Proposition 11.1.1 *Let X be a curve and Y a projective variety. Let $X \stackrel{\varphi}{\to} Y$ be a rational map where φ is a morphism from a nonempty open subset U of X to Y and U contains all singular points of X. Then φ can be extended to a morphism from the whole of X.*

Proof: Without loss of generality we can assume that $Y = \mathbf{P}^n$. We can also assume that $\varphi(U)$ is not contained in any of the coordinate planes $x_i = 0$. Let $P \in X - U$. The rational functions x_i/x_j on $\mathbf{P}^n$ by composition with φ give rational functions on X. By Theorem (8.4.2), the local ring of X at P is a discrete valuation ring; let val_P be the associated discrete valuation. Let $\mathrm{val}_P(x_i/x_0)$ be minimum at $i = k$. Then $\mathrm{val}_P(x_i/x_k) = \mathrm{val}_P(x_i/x_0) - \mathrm{val}_P(x_k/x_0) \geq 0$. Then the map

$$Q \mapsto [\frac{x_0}{x_k}(Q), \cdots, \frac{x_n}{x_k}(Q)]$$

is defined in a neighbourhood of P and extends φ to P. □

Example 11.1.2: Let X be a non-singular curve and let f be a rational function on X. Then f defines a *morphism* from X to $\mathbf{P}^1$. Further if f is not constant and X is projective then by completeness, this morphism is onto.

It is not true that a rational function on a singular curve gives a morphism to $\mathbf{P}^1$ as the following example shows.

Example 11.1.3: Consider the affine plane curve $y^2 - x^3 = 0$ in $\mathbf{A}^2$ with coordinates x, y. Then $f(x, y) = (y/x, 1) \in \mathbf{P}^1$ can easily seen to be a rational map; but it does not extend to the origin.

However, the situation is not as bad as it seems, as the following corollary shows.

Corollary 11.1.2 *Let X be any irreducible curve and $f \in k(X)$ a rational function which is regular at all singular points of X. Then f defines a morphism of X to $\mathbf{P}^1$.*

Corollary 11.1.3 *Let X be an irreducible curve. Then there exists a non-constant morphism $f : X \to \mathbf{P}^1$.*

Proof: Let $U \subset X$ be an affine open set of X containing the finitely many singular points of X. Since $k(X) = k(U)$ has transcendence degree 1 over k and $k(U)$ is the quotient field of $\Gamma(U)$ we have a non-constant regular function $f \in \Gamma(U)$. f is a rational function regular at all singular points of X and hence gives a morphism to $\mathbf{P}^1$ by the previous corollary. It is clearly non-constant. □

Definition: A rational map X to Y is called *birational* if it induces an isomorphism on the function fields.

Proposition 11.1.4 *Let X be a curve, Y a non-singular curve and $f : X \to Y$ a morphism which is birational. Then $f(X)$ is a (Zariski) open subset of Y and $f : X \to f(X)$ is an isomorphism.*

Proof: First we will show that $f(X)$ contains a non-empty Zariski open subset of Y. For this we may clearly assume that X and Y are affine, and let $A = \Gamma(X)$ and $B = \Gamma(Y)$. So we have an inclusion of rings $f^* : B \to A$ which is an isomorphism when we go to their quotient fields. Since A is finitely generated, say by $x_1, \ldots, x_n$, we may find a $b \in B$, non-zero such that $bx_i \in B$ for all i. But then $A_b = B_b$ and hence the affine open set $b \neq 0$ is in the image of f.

Since $f(X)$ contains a Zariski open set and complement of such a set is finite, we see that $Y - f(X)$ is also a finite set and thus $f(X)$ is a Zariski open set. Thus $f(X)$ is a non-singular curve. Thus replacing Y by $f(X)$ we may assume that f is surjective. Now we will show that f is an isomorphism. Let us first show that it is injective. If $f(P) = f(Q) = y$, then we get an inclusion of local rings, $\mathcal{O}_y \subset \mathcal{O}_P$ and $\mathcal{O}_y \subset \mathcal{O}_Q$. But $\mathcal{O}_y$ is a DVR since Y is non-singular. Thus these inclusions are actually equalities and then $P = Q$. Also $f^* : \mathcal{O}_y \to \mathcal{O}_P$ is an isomorphism for all $P \in X$ and $y = f(P)$ imply that f is an isomorphism by lemma (8.3.1) □

Corollary 11.1.5 *Let X be a projective curve and Y a non-singular curve. Let $f : X \to Y$ be a morphism which is birational. Then f is an isomorphism.*

Proof: By the previous proposition, f is an isomorphism to the image $f(X)$ and $f(X)$ contains a non-empty Zariski open set of Y. But since X is projective it is complete and hence $f(X)$ is closed by Theorem (9.2.1). Thus $f(X) = Y$. □

Let X, Y be irreducible projective algebraic curves and $\pi : X \to Y$ be a non constant morphism. Then $\pi(X)$ is a closed subvariety of Y and is of dimension 1. Since Y is irreducible, $\pi(X) = Y$. Moreover if $Q \in Y$ then $\pi^{-1}(Q)$ is a proper closed subset of X and hence is a finite set of points. So any such π is a surjective map with finite fibres.

Definition: A morphism $\pi : X \to Y$ is called *affine* if Y can be covered by open subsets U_i which are affine varieties and $\pi^{-1}(U_i)$ are also affine varieties.

Example 11.1.4: If $f : X \to Y$ is a morphism between affine varieties, clearly it is an affine morphism.

Example 11.1.5: In this example we study morphisms $f : \mathbf{P}^1 \to \mathbf{P}^1$. By exercise (8.3.2) we know that such morphisms are given by

$$(x, y) \mapsto (F(x, y), G(x, y))$$

where F, G are homogeneous polynomials in x, y of some degree d with no non-trivial common zeroes. If either F or G is zero, then since they have no common non-trivial zeroes the other must be a non-zero constant and thus the map must be constant. Let us assume that f is not constant. Then neither F nor G is zero. If u, v are the coordinates of the range $\mathbf{P}^1$, then the inverse image of $\mathbf{A}^1 = \{v \neq 0$ say is the open subset $F \neq 0$ and thus it is affine. This says that any such morphism is affine. We will prove below a much more general result.

Proposition 11.1.6 *Any non constant morphism $\pi : X \to Y$ between irreducible projective algebraic curves is an affine morphism.*

Proof: 1) For any projective curve there exist nonconstant morphisms $\mathbf{P}^1$ (corollary (11.1.3)).

2) It is enough to prove that $f : X \to \mathbf{P}^1$ is affine, if nonconstant.
For: $X \xrightarrow{\pi} Y \xrightarrow{g} \mathbf{P}^1$. Assume both g and $g \circ \pi$ affine. Cover $\mathbf{P}^1$ by U_i (resp. V_j) such that $g^{-1}(U_i)$ (resp. $(g \circ \pi)^{-1}(V_j)$) are affine. In $\mathbf{P}^1$, $U_i \cap V_j$ are still affine (Prove this) and the covering $g^{-1}(U_i \cap V_j)$ of Y will do for π.

3) To prove $f : X \to \mathbf{P}^1$ is affine: We must show that given $P \in \mathbf{P}^1$, there exists an affine neighbourhood U of P in $\mathbf{P}^1$ such that $f^{-1}(U)$ is affine. Let $Q \in \mathbf{P}^1$, different from P. Then $f^{-1}(P)$ and $f^{-1}(Q)$ are two disjoint finite subsets of X. Since X is projective, take $X \subset \mathbf{P}^n$ for some n. Then there exists a hypersurface $F = 0$ in $\mathbf{P}^n$ (of sufficiently large degree) such that $F(\xi) = 0 \quad \forall \xi \in f^{-1}(Q)$ and $F(\xi) \neq 0 \quad \forall \xi \in f^{-1}(P)$. Let $V = X \cap \{F \neq 0\}$. Since $\{F \neq 0\}$ is an affine open subset of $\mathbf{P}^n, V$ is an affine open set of X. Let $F = X - V$. $F \supset f^{-1}(Q)$ and $F \cap f^{-1}(P) = \phi$. Let $G = f(F), U = \mathbf{P}^1 - G$ and let $W = f^{-1}(U)$. Then $f^{-1}(P) \subset W \subset V$.

We have $V \xrightarrow{f} \mathbf{A}^1 = \mathbf{P}^1 - \{Q\}$ and $W = f^{-1}(\mathbf{A}^1 - G)$. Since the co-ordinate ring of $\mathbf{A}^1$ is $k[t]$, which is a principal ideal domain, and G

a proper closed subset of $\mathbf{A}^1$, G is defined by the vanishing of a single polynomial $0 \neq p(t) \in k[t]$. So W in V is defined by the non-vanishing of a single regular function, the composite of f and $p(t)$. But then W is affine.

□

Definition: A morphism $\pi : X \to Y$ is called *finite* if there is an affine cover U_i of Y such that $\pi^{-1}(U_i)$ is affine and the coordinate ring of $\pi^{-1}(U_i)$ is integral over that of U_i.

Theorem 11.1.1 *Let $f : X \to Y$ be a morphism between affine curves. Then f is proper if and only if f is finite.*

Remark: The theorem is valid for any affine variety, not just curves. We have enough machinery to prove this here with very little effort. But we shall not do so.

Proof: Let $A = \Gamma(X)$ and $B = \Gamma(Y)$. First let us assume that f is finite. One sees immediately that A is a finite B-module under the induced map f^*. By property 1) of proposition (9.3.1) to show that f is proper, it suffices to show that f is closed (justify this). Again to show this it suffices to show that under the hypothesis, f is surjective. Now proposition (7.4.5) finishes the proof.

Conversely assume that f is proper. If A' is the integral closure of A in its quotient field, then from the first part Spec $A' \to X$ is proper and thus by propostion (9.3.1), the composite Spec $A' \to Y$ is proper. Also if we knew that $B \to A'$ is integral it would follow that $B \to A$ is integral. So we may assume that A is integrally closed. Similar argument would say that we may assume B to be normal too. Let C be the the set of all elements of A integral over B. Thus $B \to C$ is integral and the quotient field of $C$is the same as that of A. Let $Z =$ Spec C. Then Z is an affine variety by theorem (7.4.3). The composite $X \to Z \to Y$ is proper and thus by propostion (9.3.1), $X \to Z$ is proper. Since A is normal so is C and thus Z is non-singular. Since $X \to Z$ is proper $f(X) = Z$. Z is a non-singular curve by proposition (11.1.4) and so $X \to Z$ is an isomorphism. □

Theorem 11.1.2 *Any nonconstant morphism $\pi : X \to Y$ between two irreducible projective algebraic curves is a finite morphism*

Proof: We have already seen that π is affine, *i.e.*, we can cover Y by affine U such that $\pi^{-1}(U)$ is affine. Also $\pi : \pi^{-1}(U_i) \to U_i$ is proper by proposition (9.3.1) and hence by the previous theorem, it is finite. □

Another 'good' property of finite morphisms between projective smooth curves is the following.

Proposition 11.1.7 *Let $f : X \to Y$ be a non-constant separable morphism of curves over an algebraically closed field k. Then there exists an integer $d > 0$, called the* degree*, such that for all but a finite set of points of Y, the fibre $f^{-1}(P)$ consists of exactly d points.*

Proof: Let K be the function field of Y, and L the function field of X. Then L/K is a finite separable extension. Let $U \subset Y$ be a non-empty affine open subset, $V = f^{-1}(U)$. Then V is also affine, since f is a finite morphism, by Theorem (11.1.2). Also $X - V$, $Y - U$ are finite sets of points.

Let A be the ring of regular functions on U, and B the ring of regular functions on B. Then the inclusion $K \subset L$ induces an inclusion $A \subset B$. Since L/K is separable, we can find a primitive element $x \in B$ such that $L = K(x)$ (choose an arbitrary primitive element, then multiply it by a suitable element of A). Since B is a finitely generated A-module (since f is finite), and $L = K[x]$, there exists a non-zero $a \in A$ such that $aB \subset A[x]$. Then $A[x, a^{-1}] = A_a[x] = B_a$, where A_a is the ring of regular functions on $U - \{a = 0\}$. So we may assume without loss of generality that $B = A[x]$. Now x satisfies an irreducible polynomial $f(t)$ of degree $d = [L : K]$ with coefficients in A. Let $a_0 \in A$ be the leading coefficient of this polynomial, and $a_1 \in K$ its discriminant. Then $a_2 = a_0^{d(d-1)} a_1 \in A$. Consider $U - \{a_0 a_2 = 0\}$. We claim that for $y \in U$, $f^{-1}(y)$ consists of precisely d points. Indeed, when we evaluate the coefficients of $f(t)$ at y, we obtain a monic polynomial of degree d over k which has non-zero discriminant, hence has d distinct roots. Since $B = A[x]$, this means the homomorphism $A \twoheadrightarrow k$ corresponding to y extends to a homomorphism $B \twoheadrightarrow k$ in exactly d ways. These extensions correspond precisely to the preimages of y under f. □

Corollary 11.1.8 *Let $f : X \to Y$ be a non-constant morphism between smooth curves. Let d be the separable degree of the function field extension $k(Y) \subset k(X)$. Then for all but a finite number of points x of Y, $f^{-1}(x)$ consists of d points.*

In fact a more precise result is true regarding inverse images of points under a finite map. We need to discuss first the correct way to interpret the *fibre of a point*. So let $f : X \to Y$ be a morphism between non-singular curves and let us assume that it is not constant. The we know that for any point $P \in Y$ there are only finitely many points in the inverse image. Let $P \in Y$ and let $f^{-1}(P) = \{Q_1, \ldots, Q_n \in X\}$. Let $Q = Q_i$ for some i. Then we have the induced map $f^* : \mathcal{O}_P \to \mathcal{O}_Q$. Both these are discrete valuation rings and if t_P, t_Q are the respective uniformising parameters, then $f^*(t_P) = ut_Q^{e_Q}$ for some unit $u \in \mathcal{O}_Q$ and a well defined positive integer e_Q. We define the *scheme theoretic inverse image of* P to be the formal sum $\sum e_{Q_i} Q_i$ and denote it by $f^*(P)$. Now we are ready for the general result on fibers of morphisms.

Theorem 11.1.3 *Let $f : X \to Y$ be a non-constant morphism between non-singular projective curves. let d be the degree of the morphism. Then for any point $P \in Y$, if $f^*(P) = \sum e_{Q_i} Q_i$, then $\sum e_{Q_i} = d$.*

Proof: Let $S = f^{-1}(P) = \{Q_1, \ldots, Q_n\}$ and let A be the semilocal ring of functions on X regular at all the Q_i. We have the induced map $\mathcal{O}_P \to A$ and since f is finite, this is a finite extension. Since A is a finitely generated torsion free module over $\mathcal{O}_P$, it is a free module of some rank say d' (see exercise (7.5.1)). Then the quotient field of $A = k(X)$ is a vector space of rank d' over the quotient field of $\mathcal{O}_P = k(Y)$. But this rank is precisely the degree of f. Thus $d = d'$.

Let t_P be the uniformising parameter of $\mathcal{O}_P$. Now we know that $A/t_P A$ is a d-dimensional vector space over $k = \mathcal{O}_P/(t_P)$. By definition of e_{Q_i}, we also know that $A_i = \mathcal{O}_{Q_i}/(t_P)$ is a e_{Q_i}-dimensional vector space. So if we show that the natural map $A/(t_P) \to \prod A_i$ is an isomorphism we would be done. Let $\mathcal{M}_{Q_i}$ be the maximal ideal of $\mathcal{O}_{Q_i}$ and let $\mathcal{M}_i$ the maximal ideal of A corresponding to Q_i. Let $J = \cap \mathcal{M}_i^{e_{Q_i}}$. Then by localising we easily see that $J = (t_P)$. Now by Nakayama's lemma (7.2.6), $A/J = \prod A/\mathcal{M}_i^{e_{Q_i}}$ and since $A/\mathcal{M}_i^{e_{Q_i}} = A_{\mathcal{M}_i}/\mathcal{M}_i^{e_{Q_i}}$, we are done. □

11.2 Non-singular models

The aim of this section is to prove the existence of non-singular models for curves. We will work over an algebraically field k for convenience.

Theorem 11.2.1 *Let K be a function field in one variable over k. Then there exists a unique (upto isomorphism) non-singular projective curve C such that $k(C)$, the field of rational functions on C is isomorphic to K.*

Proof: We may find a $t \in K$ such that K is a finite separable extension of $k(t)$ [(7.4.2)]. Then K is a simple extension of $k(t)$ and hence we can find $u \in K$ such that $K = k(t, u)$. Since u is algebraic over $k(t)$ we get a relation $F(x, y) \in k[x, y]$ such that $F(t, u) = 0$. Clearly we may assume that F is irreducible since K is an integral domain. Consider the affine curve $D' \subset \mathbf{A}^2$ defined by the vanishing of F. Let D be its closure in $\mathbf{P}^2$. Then the rational functions on D is isomorphic to K. Unfortunately this curve may not be non-singular. So we will have to work a little harder.

The irreducibility implies that the curve D has only finitely many singular points (by Jacobian Criterion (8.4.1)). Let $L = 0$ be a line not passing through any of these singular points and $H = 0$ be a curve not passing through any of the finitely many intersection points of L and D but passing through all the singular points. Let $U = D \cap \{L \neq 0\}$ and $V = D \cap \{L' \neq 0\}$. Then U, V are affine open subsets of D covering D and V is non-singular. Let A be the coordinate ring of U and B its integral closure. By corollary (8.4.1) we know that B is an affine ring of an affine non-singular curve X'. Let X be a projective curve containing X' as an open subset.

We have a proper morphism $p : X' \to U$ by theorem (11.1.1) and since the function fields of X' and U are both K there exists an open subset $W \subset U$ where $p : p^{-1}(W) \to W$ is an isomorphism. We may shrink W further if necessary so that $W \subset U \cap V$. Thus we get a morphism $W \to X \times D$ the first is got by using the isomorphism of W with $p^{-1}(W)$ which in turn is included in X. The second is the natural inclusion of $W \subset D$. Let C be the closure of its image. Since $X \times D$ is projective, C is projective. Also since W is open in C, the rational functions on W is K.

Now we will show that C is non-singular. This will finish the construction. Some details will be left to the reader. Let π_1 and π_2 be the two projections from C to X and D. First notice that both these are birational morphisms. Also $\pi_1^{-1}(X') = U_1$ and $\pi_2^{-1}(V) = V_1$ cover C. [Check this.] Since π_i's are proper we see that they are finite morphisms and in particular affine. So $U_1 \to X'$ is a finite map. But it is also birational and since B is integrally closed, we see that π_1 is an isomorphism. So U_1 is non-singular. By similar argument one sees that V_1 is non-singular since V is. Thus C is non-singular.

Now the uniqueness follows from (11.1.1). □

Corollary 11.2.1 *Let K and L be function fields in one variable. Let C and D be the corresponding non-singular projective curves. Then giving a non-constant morphism $f : C \to D$ is equivalent to giving an inclusion of fields $L \subset K$.*

11.3 Group actions on curves

Let A be any ring and let G be a group of ring automorphisms of A.. Then we have the *ring of invariants*

$$A^G = \{x \in A \mid gx = x \ \forall g \in G\}$$

We will be mainly interested in the case of finite groups.

Lemma 11.3.1 *Let A, G be as above and let G be finite. Then $A^G \subset A$ is an integral extension.*

Proof: For any $x \in A$ consider the polynomial

$$P(T) = \prod_{g \in G} (T - gx).$$

Then one can easily see that the coefficients of the monic polynomial $P(T)$ actually lie in A^G and of course $P(x) = 0$. □

Let X be any variety. Then we can define the set End X to be the set of all *morphisms* from X to X. This is in a natural way a monoid with respect to composition. We have a subgroup Aut X of this, namely the set of all isomorphisms of X. We say that a group G *acts* on X if we are given a group homomorphism $G \to$ Aut X. If X is irreducible any automorphism of X gives a k-automorphism of $k(X)$, the function field of X. If X is a non-singular curve, then by the results of the previous section, we see that given a k-automorphism of $k(X)$ we naturally get an automorphism of X. We have the following definition, using the results of the previous section.

Definition: Let X be a non-singular curve and let X/G be the non-singular curve corresponding to the function field $k(X)^G$ and let $\pi : X \to X/G$ be the natural morphism. Then we call X/G the *quotient of X by G* and π the *quotient map.*

The main result we want to prove is the following:

Proposition 11.3.2 *Let $f : X \to Y$ be a morphism of non-singular curves and let G be a finite group acting on X. Assume that $f^{-1}(y)$ is invariant under G for all $y \in Y$. Then there exists a unique morphism $g : X/G \to Y$ such that $f = g \circ \pi$.*

Proof: Clearly we may assume that f is not constant and then f has finite fibres. f induces a map $f^* : k(Y) \hookrightarrow k(X)$. We only need to show that this inclusion factors through $k(X)^G$ by results in the previous section. Let $\varphi \in k(Y) \subset k(X)$. Let S be the finite subset of Y outside which φ is regular. Then φ as a rational function on X is regular outside the finite set $T = f^{-1}(S)$. Let $x \in X - T$. Then notice that by hypothesis $y = gx \in X - T$ and $f(y) = f(x)$. We have

$$\varphi(x) = \varphi(f(x)) = \varphi(f(y)) = \varphi(y).$$

Since $g(\varphi)(x) = \varphi(g^{-1}x) = \varphi(x)$ for x in the open set $X - T$, $g(\varphi) = \varphi$. Thus φ is invariant under the group action and hence it belongs to $k(X)^G$ proving the result. □

11.4 Varieties over non-algebraically closed fields

Let k be a field, not necessarily algebraically closed. For simplicity we assume that k is a perfect field.

Let $\overline{k}$ be an algebraically closed field containing k. Let X be an affine variety in $\mathbf{A}^n_{\overline{k}}$. If the ideal $I \subset \overline{k}[x_1, \cdots, x_n]$ of functions vanishing on X has generators $f_1, \cdots, f_r$ which are polynomials with coefficients in k, then we say that X is *defined over k.*

If X is defined over k and $I(X) \subset \overline{k}[x_1, \cdots, x_n]$ is the ideal of functions vanishing on X, then $I(X) = (I(X) \cap k[x_1, \ldots, x_n]) \otimes_k \overline{k}$, and the coordinate

ring $\Gamma(X)$ of X is given by $\Gamma_k(X) \otimes_k \overline{k}$ where $\Gamma_k(X) = k[x_1, \cdots, x_n]/(I(X) \cap k[x_1, \ldots, x_n])$. (We identify $\overline{k}[x_1, \cdots, x_n]$ with $k[x_1, \cdots x_n] \otimes_k \overline{k}$ in the natural way). We call $\Gamma_k(X)$ the *coordinate ring of X over k.*

If an affine variety is defined over k, we can talk of its k *rational* (or *k-valued*) points. A point $(a_1, \cdots, a_n)$ with all the $a_i \in k$ which is a zero of all polynomials in $I(X)$ is a k-rational point of X.

We can interpret this in terms of the coordinate ring over k of X, and thus get a more intrinsic characterisation of a point being k-rational. First recall that if we have any point $P = (a_1, \cdots, a_n)$ of X, then evaluation at P gives a homomorphism $\Gamma(X)\overline{k}$ defined by $f \mapsto f(a_1, \cdots, a_n)$. Conversely, given a $\overline{k}$-algebra homomorphism of $\Gamma(X) \to \overline{k}$, the Nullstellensatz implies that it comes from a unique point $(a_1, \cdots, a_n)$ of X (see corollary to (8.1.1)) Thus a point of X is equivalent to a $\overline{k}$-algebra homomorphism of its coordinate ring into $\overline{k}$. The maximal ideals of $\Gamma(X)$ are precisely the kernels of such homomorphisms to $\overline{k}$.

A k-rational point of X then corresponds to a homomorphism $\overline{\varphi} : \Gamma(X) \to \overline{k}$ for which there exists a k-algebra homomorphism $\varphi : \Gamma_k(X) \to k$ such that $\overline{\varphi} = \varphi \otimes id_{\overline{k}}$. The maximal ideals of $\Gamma_k(X)$ *with residue field k* are precisely the kernels of such homomorphisms. There may in general be other maximal ideals of $\Gamma_k(X)$, which are contractions to $\Gamma_k(X)$ of maximal ideals of $\Gamma(X)$ which correspond to points which are not k-rational.

Example 11.4.1: Let $\mathbf{R}$ denote the field of real numbers.

(i) The curve X defined by $x^2 + y^2 - 1 = 0$ in $\mathbf{A}^2_{\mathbf{C}}$ is defined over $\mathbf{R}$ (and even over $\mathbf{Q}$), and the $\mathbf{R}$-rational points are just the points on the unit circle in $\mathbf{R}^2$. The ring $\Gamma_{\mathbf{R}}(X) = \mathbf{R}[x, y]/(x^2 + y^2 - 1)$; the maximal ideal $(x - \sqrt{2}, y^2 + 1)$ with residue field $\mathbf{C}$ does not correspond to a $\mathbf{R}$-rational point.

(ii) The curve $i(x^2 + y^2 - 1) = 0$ in $\mathbf{A}^2_{\mathbf{C}}$ is defined over $\mathbf{R}$, since it can also be defined by an equation with real coefficients.

(iii) The curve $x^2 + y^2 = 0$ in $\mathbf{A}^2_{\mathbf{C}}$ is defined over $\mathbf{R}$, and has exactly one $\mathbf{R}$-rational point, the origin $(0, 0)$.

(iv) The curve $x^2 + y^2 + 1 = 0$ in $\mathbf{A}^2_{\mathbf{C}}$ is defined over $\mathbf{R}$ and has no $\mathbf{R}$-rational points.

(v) The line $(2i)x + 3 + (4i)y - (5 + 4i) = 0$ in $\mathbf{A}^2_{\mathbf{C}}$ is *not* defined over $\mathbf{R}$, though the point $(1, 1)$ lies on it.

(vi) The curve $(2i)x^2 - (3+i)y^2 + (3-i)xy - (2+2i) - (1+i)y + 2 = 0$ in $\mathbf{A}^2_{\mathbf{C}}$ is not defined over $\mathbf{R}$, though it contains the line $x - y - 1 = 0$, which is defined over $\mathbf{R}$.

Example 11.4.2: Fermat's Last Theorem is the assertion that the Fermat curve $x^n + y^n = 1$ in $\mathbf{A}^2_{\mathbf{C}}$, which is defined over $\mathbf{Q}$, has no rational points with $xy \neq 0$, if $n \geq 3$.

A projective variety $X \subset \mathbf{P}^n_{\overline{k}}$ is said to be defined over k if the homogeneous ideal $I(X)$ in $\overline{k}[x_0, \cdots, x_n]$ of homogeneous polynomials vanishing on X is generated by homogeneous polynomials with coefficients in k, *i.e.*, by elements from $k[x_0, \cdots, x_n]$.

Let $U_i = \{[a_0, \cdots a_n] \in \mathbf{P}^n_k \mid a_i \neq 0\}$. Then $U_i \simeq \mathbf{A}^n_k$ and $X_i = X \cap U_i$ are affine varieties. If $X \subset \mathbf{P}^n_{\overline{k}}$ is defined over k then so is $X_i \subset \mathbf{A}^n_{\overline{k}}$, since the ideal of X_i is obtained from $I(X)$ by setting $x_i = 1$.

If U is an open subvariety of a variety X defined over k, we say that U is defined over k if the complement $X - U$ is defined over k. Thus, for an arbitrary quasi-projective variety in $\mathbf{P}^n$, the notion of being *defined over* k is meaningful.

Let X be a variety defined over k. Then any point of X has an affine neighbourhood $U \subset \mathbf{A}^n$ which is defined over k. If a $\overline{k}$-valued function on some open subset V of X can be written in the neighbourhood of a point $P \in U \cap V$ in the form f/g with $g(P) \neq 0$ and $f, g \in k[x_1, \cdots, x_n]$, we say that the germ at P of this function (see Chapter (8)) is a k-rational germ which is regular at P. The k-rational germs regular at P form a local k-subalgebra of the local ring $\mathcal{O}_{X,P}$ of germs regular at P. If $\mathcal{M}$ is the maximal ideal in $\Gamma_k(X)$ (with residue field k) corresponding to P, then the localisation $\Gamma_k(X)_{\mathcal{M}}$ is naturally isomorphic to the ring of k-rational germs regular at P. Similarly, we can define the notion of k-rational functions on a variety X which is defined over k; for affine X, these are elements of the total quotient ring of $\Gamma_k(X)$. The ring of k-rational functions is denoted $k(X)$.

The notion of irreducibility is a bit subtle over an arbitrary field k. We define a k-variety X to be *k-irreducible* (or *irreducible over k*) if it cannot be expressed as a union of two proper closed subvarieties each of which is

defined over k. If X is affine, this is equivalent to $\Gamma_k(X)$ being a domain. The notion of the function field of a k-irreducible variety may be defined as follows: if X is an affine k-irreducible variety, the function field $k(X)$ is the quotient field of $\Gamma_k(X)$. In general, we replace X by a non-empty affine open subset defined over k. A k-regular function is a k-rational function which is a regular function over $\overline{k}$. The ring of k-regular functions on a k-variety X is denoted by $\Gamma_k(X)$.

However, an irreducible k-variety $X \subset \mathbf{A}^n_{\overline{k}}$ *need not* be an irreducible $\overline{k}$-variety: for example, the variety in Example 11.4.1 (iii) is $\mathbf{R}$-irreducible, but not $\mathbf{C}$ irreducible. This corresponds to the algebraic fact that even if a finitely generated k-algebra R is a domain, $R \otimes_k \overline{k}$ may not be one. However, $R \otimes_k k'$ is a domain for any extension field k' of k provided k is algebraically closed in R. Hence, we may define a k-variety to be *geometrically irreducible* over k if k is algebraically closed in $\Gamma_k(X)$, or, what is the same thing, in $k(X)$.

Next, we consider morphisms between varieties defined over k. Let X, Y be varieties defined over k and $\pi : X \to Y$ a morphism. We say that π is defined over k if it takes k-rational germs on Y to k-rational germs on X. If X, Y are affine, this is equivalent to giving a k-algebra homomorphism $\pi^* : \Gamma_k(Y) \to \Gamma_k(X)$.

Example 11.4.3: For an affine variety $X \subset \mathbf{A}^n_{\overline{k}}$, the notion of being defined over k depends on the embedding, in general. For example, let $k = \mathbf{Q}$; then the curves in $\mathbf{A}^2_{\mathbf{C}}$ given by

$$x^2 + y^2 - 2 = 0,$$

$$x^2 + y^2 - 3 = 0$$

are isomorphic, are both defined over $\mathbf{Q}$, but we cannot find an isomorphism which is defined over $\mathbf{Q}$,since the first curve has a $\mathbf{Q}$-rational point $(1, 1)$ while the second has no $\mathbf{Q}$-rational points (since 3 is not a sum of 2 rational squares – see Theorem 14.1.1). Thus, as $\mathbf{Q}$-curves, these curves are to be considered as different curves.

Exercise 11.4.1: For any variety X defined over k, show that a k-morphism $f : X \to \mathbf{A}^1$ is just a k-regular function on X (namely, the pullback of the coordinate function). More generally, show that for any affine k-variety Y,

k-morphisms $f : X \to Y$ are in 1-1 correspondence with homomorphisms $\Gamma_k(Y) \to \Gamma_k(X)$.

Exercise 11.4.2: (i) Let $X \subset \mathbf{P}^n_{\overline{k}}$ be defined over k, and let $X_0, \ldots, X_n$ be the homogeneous coordinates on $\mathbf{P}^n_{\overline{k}}$. Let $F_0, \ldots, F_m \in k[X_0, \ldots, X_n]$ be homogeneous polynomials of degree d, such that $V(F_0, \ldots, F_n) \cap X = \phi$ (*i.e.*, the F_i have no common zeroes on X). Then $f : X \to \mathbf{P}^m_{\overline{k}}$ defined by

$$f([x_0, \ldots, x_n]) = [F_0(x_0, \ldots, x_n), \ldots, F_m(x_0, \ldots, x_n)]$$

is a k-morphism.

(ii) If $X = \mathbf{P}^n_{\overline{k}}$, then every morphism $f : X \to \mathbf{P}^m$ is as in (i).

Exercise 11.4.3: If $C \subset \mathbf{P}^2_{\overline{k}}$ is defined over k by the homogenoeus cubic equation $f(X, Y, Z) = Y^2Z - (X^3 + aX^2Z + bXZ^2 + cZ^3) = 0$, where $a, b, c \in k$, then the mapping $\wp : C \to \mathbf{P}^1$ defined in Exercise 8.3.3 by

$$[X, Y, Z] \mapsto \begin{cases} [X, Z] & \text{if } Z \neq 0 \\ [0, 1] & \text{if } Z = 0 \end{cases}$$

is a k-morphism.

If two varieties defined over k are k'-isomorphic for some Galois extension field k', then each is called a *Galois twist* (or just a *twist*) of the other. One of the results we will prove later about elliptic curves over finite fields is Lang's theorem, that an elliptic curve defined over a finite field has no non-trivial (*i.e.*, , not k-isomorphic) twist.

For any field k, the Segre embedding $S_{m,n} : \mathbf{P}^m_{\overline{k}} \times \mathbf{P}^n_{\overline{k}} \to \mathbf{P}^{(m+1)(n+1)-1}_{\overline{k}}$ is defined over k. Hence if X, Y are defined over k, then so is $X \times Y$.

Finally, we consider non-singular projective geometrically irreducible curves over $\overline{k}$ which are defined over k. As for algebraically closed fields, these are determined by their function fields over k, which are finitely generated field extensions of k of transce dence degree 1 in which k is algebraically closed. Conversely, any such field has a non-singular projective model over k, *i.e.*, is the function field of k-rational functions of a non-singular projective geometrically irreducible curve defined over k. There is also a 1-1 correspondence between k-morphisms $f : X \to Y$ between such curves X, Y and k-algebra

homomorphisms $k(Y) \to k(X)$. In particular, the following generalisation of proposition (11.3.2) holds; we leave the proof (which is a slight modification of that of proposition (11.3.2)) to the reader.

Lemma 11.4.1 *Let $f : X \to Y$ be a k-morphism between projective non-singular geometrically irreducible curves defined over k, and let G be a finite group of k-automorphisms of X, such that each element of G preserves the fibres (over $\overline{k}$) of f. Then the quotient morphism $\pi : X \to X/G$ is defined over k, and there is a unique k-morphism $h : X/G \to Y$ such that $f = h \circ \pi$.*

11.5 Divisors and linear equivalence

Let C be a projective non-singular curve over the algebraically closed field k. Let K be the function field of C.

Definition: A *divisor* on C is a finite formal linear combination $\sum n_i P_i$ of points of C with integer coefficients.

We can define the sum of two divisors in the obvious way: $D_1 = \sum m_i P_i$, $D_2 = \sum n_j P_j$ then $D_1 + D_2 = \sum (m_i + n_i) P_i$. The divisors on C thus form an abelian group. In fact it is the free abelian group generated by the points of C.

Definition: The *degree* of a divisor , $\deg D$, is defined to be the integer $\sum n_i$ where $D = \sum n_i P_i$.

Let f be a non-zero rational function. It gives rise to a divisor $\operatorname{div} f$, the *divisor of zeros and poles* of f, as follows. Let $P \in C$. Then $\mathcal{O}_P$, the local ring of C at P is a discrete valuation ring and its quotient field is K. Let $t \in \mathcal{O}_P$ be a parameter of $\mathcal{O}_P$, ie a generator of the maximal ideal of $\mathcal{O}_P$. We then have the valuation on K with respect to $\mathcal{O}_P$: $\operatorname{val}_P(f) = n$ where $f = ut^n$, u unit in $\mathcal{O}_P$. This n is called the *order* of zero or pole of f at P. We define the *divisor of the function* f as $\operatorname{div} f = \sum_P \operatorname{val}_P(f) P$.

Note that since f has only finitely many zeros (and poles) (see (11.1.3)) the above definition of $\operatorname{div} f$ gives a divisor. It is easy to see that $\operatorname{div} fg = \operatorname{div} f + \operatorname{div} g$ and $\operatorname{div}(1/f) = -\operatorname{div} f$.

One can show that $\deg \operatorname{div} f = 0$ for any C. We will prove it here only for plane curves. So we assume $C \subset \mathbf{P}^2$, defined by a homogeneous polynomial

$F(X,Y,Z)$. Let f be a rational function on C. Then, by definition, f coincides with G/H, where G, H are homogeneous polynomials of the same degree, on the complement of a finite set of points and hence on C. It can be seen that for $P \in C, \mathrm{val}_P(f)$ is the difference $m_P - m'_P$ where $m_P (resp. m'_P)$ is the multiplicity of the intersection of the curve C with the curve defined by $G = 0$ (resp. $H = 0$) at P. Therefore $\mathrm{div}\, f = \sum m_P - \sum m'_P$. By Bezout's theorem $\sum m_P$ is md where m is the degree of G and d that of F. Since G and H have the same degree $\sum m'_P$ is also md. Thus $\deg \mathrm{div}\, f = 0$.

Definition: A divisor D is called effective if it is of the form $D = \sum m_i P_i$ with $m_i \geq 0$ for all i. We define $D_1 \geq D_2$ if $D_1 - D_2$ is effective.

Definition: For a divisor D, we define $H^0(D)$ to be $= \{f \in K \mid f = 0 \text{ or } \mathrm{div}\, f + D \geq 0\}$.

Exercise 11.5.1: i) $H^0(D)$ is a k-vector subspace of K.

ii) $D_1 \geq D_2 \Rightarrow H^0(D_1) \supseteq H^0(D_2)$.

Proposition 11.5.1 *$H^0(D)$ is a finite dimensional vector space over k. In fact* $\dim H^0(D) \leq \deg D + 1$.

Proof: If $0 \neq f \in H^0(D), \mathrm{div}\, f + D \geq 0$ and therefore $\deg \mathrm{div}\, f + \deg D \geq 0$. Thus in the case $\deg D < 0, H^0(D) = 0$.

It is now enough to prove that $\dim H^0(D+P) \leq \dim H^0(D) + 1$, because we can use induction on $\deg D$ to complete the proof.

Let t be a parameter in $\mathcal{O}_P$. Let P occur with multiplicity n in D. Then for any $f \in H^0(D+P), t^{n+1} f$ is regular at P and hence we get a linear map $H^0(D+P) \to k, f \mapsto (t^{n+1} f)(P)$. Clearly kernel of this map is $H^0(D)$. Therefore $\dim H^0(D+P) \leq \dim H^0(D) + 1$. □

We denote by $h^0(D)$ the dimension of $H^0(D)$.

Definition: Two divisors D_1 and D_2 are said to be *linearly equivalent* if there is a non-zero rational function f such that $D_2 = \mathrm{div}\, f + D_1$.

We use the notation $D_1 \sim D_2$ to denote that D_1 is linearly equivalent to D_2.

If $D_1 \sim D_1'$ and $D_2 \sim D_2'$ then clearly $D_1 + D_2 \sim D_1' + D_2'$. Therefore the linear equivalence classes form an abelian group with the class of $\operatorname{div} f$ being the zero element. Moreover if $D_1 \sim D_1'$ then $\deg D_1 = \deg D_1'$. Therefore deg goes down to a homomorphism from the group of linear equivalence classes of divisors to the integers.

If $D_2 \sim D_1$ with $D_2 = \operatorname{div} f + D_1$ the linear map $H^0(D_2) \to H^0(D_1)$ given by $g \mapsto fg$ gives an isomorphism. In particular $h^0(D_1) = h^0(D_2)$.

Chapter 12

Geometry of elliptic curves

12.1 Riemann-Roch for elliptic curves

Theorem 12.1.1 *Let C be an elliptic curve. Then for any divisor D on C*

$$h^0(D) - h^0(-D) = \deg D.$$

Proof: We will prove several claims on existence of rational functions with prescribed poles on C.

Claim: 1. Given $P, Q, R \in C$ with R distinct from P, Q, (P may be equal to Q), we can find a rational function f on C such that

$$\operatorname{div} f = P + Q - R - S$$

where S is also distinct from P, Q.

2. Let $n \geq 2$. Given $P \in C$ there exists a rational function f such that it is regular outside P and has a pole of order exactly n at P.

Let us finish the proof of the theorem assuming the claims.

Since an effective divisor has non-negative degree we see that $h^0(D) > 0$ implies that $\deg D \geq 0$. Next we prove a partial converse to this.

Claim: If $\deg D > 0$ then $H^0(D) \neq 0$.

Since $H^0(D) \subset H^0(D+P)$ for any point P, we see that it is enough to prove the claim when $\deg D = 1$. So write D as

$$\sum_{i=1}^{n}(P_i - Q_i) + R,$$

with n minimal. Then $R \neq Q_i$ for any i and $P_i \neq Q_j$ for any i, j. We will do this by induction on n. If $n = 0$ then $D = R$ which is effective proving the claim. So assume that $n > 0$. By our earlier claim (2) we have a rational function f such that

$$\operatorname{div} f = Q_n + S - P_n - R$$

where $S \in C$. Since D and $D + \operatorname{div} f$ are linearly equivalent we may use the latter as a representative and then

$$D + \operatorname{div} f = \sum_{i=1}^{n-1}(P_i - Q_i) + S.$$

Now by induction we are through.

To prove the theorem we need only consider the case of $\deg D \geq 0$. If $\deg D = 0$ and $h^0(D) \neq 0$ then $D = 0$ and hence the theorem is clear. If $h^0(-D) \neq 0$ the same argument does it. If both $h^0(D) = h^0(-D) = 0$ of course there is nothing to prove.

Thus we are reduced to the case of divisors with $\deg D > 0$ and hence by the claim, to the case of effective divisors. If $\deg D = 1$ then $D = P$ a single point. If there was non-constant function $f \in H^0(P)$ then f gives a degree 1 morphism to $\mathbf{P}^1$ and hence our curve would be rational which is not the case. So theorem is proved in this case too.

So now let us assume that $\deg D \geq 2$ and it is effective. First notice that for any divisor D and any point P

$$h^0(D) \leq h^0(D+P) \leq h^0(D) + 1.$$

(See the proof of Proposition 11.1.1.) Write $D = G + nP$ for some point P and G effective with no P in its support. If $n \geq 2$ then by claim (1) there exists an f with pole of order exactly n at P and regular everywhere else. So $f \in H^0(D)$ and it is not in $H^0(D-P)$. By an obvious induction on degree of the divisor we are done. So let us assume that all such points occur with multiplicity one in D. Since degree is bigger or equal to 2, we may write

$D = G + P + Q$ with P not in the support of G. By claim (1), there exists a function f which has simple poles at P, Q and no other poles. Then this function clearly belongs to $H^0(D)$ and does not belong to $H^0(D - P)$ and again the induction does the job.

Now let us prove the claims. The main technique is Bezout's theorem.

If P, Q are two points we will denote by $L(P, Q)$ the line joining P, Q if they are distinct and the tangent at P if they are equal. We will denote by the same letter the equation defining the line.

Take $L(P, Q)$ and let A be the third point of intersection. Then $L(A, R) \neq L(P, Q)$ and let S be the third point of intersection of $L(A, R)$. Then take $f = L(P, Q)/L(A, R)$ to prove the first claim.

For proving the second claim we will show that there exists functions with poles of order 2 and three at P and then taking suitable product of these functions we will be done. By taking $P = Q$ in the first claim and R any other point and taking the inverse of the function we get a rational function which has only poles at P and it is of order 2 at P. Next if P is a point of inflexion, take $L(P, P)$ and take M to be any line not passing through P. Then $M/L(P, P)$ has a pole of order 3 at P and no other poles. If P is not a point of inflexion take $L(P, P)$ and M any other line passing through P. Let A be the residual intersection of L and B, C the residual intersection of M. Take $N = L(A, B) \cdot L(A, C)$. Then $N/L(P, P) \cdot M$ has the required property. □

12.2 Differentials

Let A be a k-algebra and M an A-module. A derivation of A into M is a k-linear map $D : A \to M$ such that $D(ab) = aD(b) + bD(a)$. The definition immediately implies that $D(a) = 0$ for any $a \in k$. If D, E are derivations of A into M, then clearly $D + E$ defined by $D + E(a) = D(a) + E(a)$ is also derivation. Similarly if $x \in A$ then xD defined by $(xD)(a) = xD(a)$ is also a derivation. Thus the set of all derivations form an A-module, denoted by Der (A, M).

Lemma 12.2.1 *Let $D : A \to M$ be a derivation. Then* $\ker D$ *is a k-subalgebra of A.*

Proof: Let $x, y \in \ker D$. Then clearly $x + y \in \ker D$. Also by the definition, $xy \in \ker D$. □

Example 12.2.1: Let $A = k[x_1, \ldots, x_n]$, the polynomial ring in n variables. Then the map $D_i : A \to A$ given by

$$D_i(f) = \frac{\partial f}{\partial x_i}$$

is a derivation of A into A, by the Leibnitz rule. If D is any derivation of A into A, let $f_i = D(x_i)$. Then $E = D - \sum f_i D_i$ is also a derivation. By choice $x_i \in \ker E$, and thus by the lemma, we get $E = 0$. In other words, Der (A, A) is generated as an A-module by D_i's. If $\sum f_i D_i = 0$ for some $f_i \in A$, then applying it to x_i we see that $f_i = 0$. Thus Der (A, A) is a free A-module over A with the D_i as a basis.

Proposition 12.2.2 *For any k-algebra A there exists an A-module $\Omega_k(A)$ and a derivation $d : A \to \Omega_k(A)$ such that $\Omega_k(A)$ is generated, as an A-module, by the image of A under d, and such that if $D : A \to M$ is any derivation, then there exists a unique A-linear map $\varphi : \Omega_k(A) \to M$ such that $D = \varphi \circ d$. Moreover such an $\Omega_k(A)$ is unique upto isomorphism.*

Proof: Let F be the free A-module generated by the elements of the set A. We denote an element a of A, when considered as a generator of F, by $[a]$. Let N be the submodule of F generated by the elements of the form $[a+b] - [a] - [b], [\lambda\, a] - \lambda[a], \lambda \in k$, and $[ab] - a[b] - b[a]$. Let $\Omega_k(A) = F/N$ and $d : A \to \Omega_k(A)$ be defined by $da =$ class of $[a]$ in F/N. We leave the properties of $\Omega_k(A)$ to be checked as exercise. □

Exercise 12.2.1:

(i) If A is generated as a k-algebra by $x_1, \ldots, x_n$ then $\Omega_k(A)$ is generated by $dx_1, \ldots, dx_n$ as an A-module.

(ii) If $A = k[x_1, \ldots, x_n]$, a polynomial ring in n variables then $\Omega_k(A)$ is a free module with basis $dx_1, \ldots, dx_n$.

(iii) If A is a k-algebra and S a multiplicatively closed subset of A, then $\Omega_k(S^{-1}A) = S^{-1}\Omega_k(A)$.

(iv) If A is a field $k(x_1, \ldots, x_n)$ then $\Omega_k(A)$ is generated by $dx_1, \ldots, dx_n$ as an A-vector space.

(v) If $f : A \to B$ is a homomorphism of k-algebras, there is an induced homomorphism $\Omega_k(A) \to \Omega_k(B)$, given by $da \mapsto df(a)$. If f is surjective, so is $\Omega_k(A) \to \Omega_k(B)$.

(vi) If $A = k[x_1, \ldots, x_n]$ is the polynomial ring, and R is the localisation of A at the maximal ideal $\mathcal{M} = (x_1, \ldots, x_n)$, then $\Omega_k(R)$ is a free R-module on the basis $dx_1, \ldots, dx_n$.

(vii) If X is an affine variety over k, $P \in X$ a point with local ring $\mathcal{O}_P$, then $\Omega_k(\mathcal{O}_P)$ is a finite $\mathcal{O}_P$-module.

Proposition 12.2.3 *Let K be the function field of an irreducible variety of dimension n over an algebraically closed field k. Then $\Omega_k(K)$ is a K-vector space of dimension n.*

Proof: By the definition of dimension of a variety, K has transcendence degree n over k. Since k is algebraically closed, we can find a separating transcendence basis for K over k (Theorem (7.4.2)) *i.e.*, there exist n elements $x_1, \cdots, x_n \in K$ which are algebraically independent over k, such that K is a separable algebraic extension of $L = k(x_1, \cdots, x_n)$. From Exercise (12.2.1)(ii), $\Omega_k(L)$ is a vector space of dimension n over L with basis $dx_1, \ldots, dx_n$. Since K is separable over L, the primitive element theorem (see [34]) implies that $K = L(y)$ where the minimal polynomial for y over L is separable.

If the minimal polynomial of y is $f(t) = t^d + a_1 t^{d-1} + \cdots + a_{d-1}t + a_d$, then

$$= d(f(y)) = f'(y)dy + y^{d-1}da_1 + y^{d-2}da_2 + \cdots + da_n$$

where $f'(t)$ is the derivative. Since $f(t)$ is a separable polynomial, $f'(y) \neq 0$. Hence dy lies in the K-subspace of $\Omega_k(K)$ generated by the image of $\Omega_k(L)$. This immediately implies that the image of $\Omega_k(L)$ generates $\Omega_k(K)$. Since dx_i generate $\Omega_k(L)$, their images in $\Omega_k(K)$ generate $\Omega_k(K)$.

It remains to show that dx_i are linearly independent over K. Since dx_i are an L-basis for $\Omega_k(L)$, this is equivalent to the assertion that any L-linear homomorphism $\Omega_k(L) \to K$ extends to a K-linear homomorphism $\Omega_k(K) \to K$ *i.e.*, any derivation $D : L \to K$ extends to a derivation $\widetilde{D} : K \to K$. Any element of K is uniquely expressible as a polynomial in y of degree $< d$. One

verifies that the formula

$$\widetilde{D}(b_0 + b_1 y + \cdots + b_{d-1} y^{d-1}) = \sum_i D(b_i) y^i + \sum_i b_i i y^{i-1} \frac{-1}{f'(y)} (\sum_i y^{d-i} da_i)$$

determines an extension $\widetilde{D}$. □

Theorem 12.2.1 *Let X be a non-singular variety of dimension n over an algebraically closed field k, and $P \in X$ a point. Let $t_1, \cdots, t_n \in \mathcal{O}_P$ be a regular system of parameters. Then $\Omega_k(\mathcal{O}_P)$ is a free $\mathcal{O}_P$-module with basis $dt_1, \ldots, dt_n$. The natural mapping $\Omega_k(\mathcal{O}_P) \to \Omega_k(K)$, where K is the function field, is injective.*

Proof: Let $\mathcal{M} \subset \mathcal{O}_P$ be the maximal ideal. Since $t_1, \cdots, t_n$ is a regular system of parameters, $t_i \in \mathcal{M}$ for all i, and $\mathcal{M}$ is generated by the t_i. Thus if $x \in \mathcal{O}_P$, then we can write $x = a + \sum_i b_i t_i$, where $a \in k$ and $b_i \in \mathcal{O}_P$. Then $dx = \sum_i b_i \, dt_i + \sum_i t_i \, db_i$. Since $\Omega_k(\mathcal{O}_P)$ is a finitely generated $\mathcal{O}_P$-module, Nakayama's lemma implies that $dt_1, \ldots, dt_n$ generate $\Omega_k(\mathcal{O}_P)$.

We claim the images of the dt_i generate $\Omega_k(K)$. Indeed, if $y \in K$, and $y = z/w$ with $z, w \in \mathcal{O}_P$, then $dy = (1/w)dz + (-z/w^2)dw$; hence the image of $\Omega_k(\mathcal{O}_P) \to \Omega_k(K)$ generates $\Omega_k(K)$. This proves the claim. Since $\Omega_k(K)$ has dimension n over K and $dt_1, \ldots, dt_n$ generate it, the dt_i must form a K-basis for $\Omega_k(K)$. In particular, $dt_i \in \Omega_k(\mathcal{O}_P)$ are $\mathcal{O}_P$-linearly independent. Hence $\Omega_k(\mathcal{O}_P)$ is a free $\mathcal{O}_P$-module with basis dt_i, and $\Omega_k(\mathcal{O}_P) \to \Omega_k(K)$ is injective. □

We call an element of $\Omega_k(K)$ a *differential form* (or a 1-form) on X. If for some $P \in X$, ω lies in the image of $\Omega_k(\mathcal{O}_P) \to \Omega_k(K)$, then we say ω is *regular* at P. We say ω is regular on X if it is regular at each $P \in X$. Equivalently, ω is regular on X if and only if

$$\omega \in \bigcap_{P \in X} \Omega_k(\mathcal{O}_P).$$

Here we have identified $\Omega_k(\mathcal{O}_P)$ with its image under the natural inclusion into $\Omega_k(K)$.

Exercise 12.2.2: Let A be the coordinate ring of an irreducible affine variety X and K its quotient field. Show that a 1-form is regular everywhere on X if and only if it is in the image of $\Omega_k(A)$. (Hint: see exercise (8.1.8)).

Suppose now that K is the function field of a non-singular curve C. If $\omega \in \Omega_k(K)$ and $P \in C$ we define the *order* of ω at P, denoted $\mathrm{ord}_P(w)$, as $\mathrm{val}_P(f)$ where $\omega = f dt$, with t a parameter of $\mathcal{O}_P$ and $f \in K$. Since $\Omega_k(\mathcal{O}_P)$ is free of rank 1 generated by dt, which also is a K-vector space generator of $\Omega_k(K)$, we may also define $\mathrm{ord}_P(\omega)$ to be the largest integer n such that $t^{-n}\omega \in \Omega_k(\mathcal{O}_P)$. This shows that ord_P is well defined *i.e.*, does not depend on the choice of the parameter t. Clearly $\mathrm{ord}_P(\omega) \geq 0$ precisely when ω is regular at P.

Lemma 12.2.4 $\mathbf{P}^1$ *has no 1-form which is regular everywhere.*

Proof: We may cover $\mathbf{P}^1$ by two affine open sets $U_i \cong \mathbf{A}^1$, $i = 1, 2$. Then $\Gamma(U_1) = k[t]$ and $\Gamma(U_2) = k[t^{-1}]$, both naturally subrings of $k(t)$, the rational functions on $\mathbf{P}^1$. By the exercises (12.2.1) and (12.2.2) the regular 1-form on U_1 is of the form $f(t)dt$ with $f(t) \in k[t]$ and on U_2 is of the form $g(t^{-1})dt^{-1}$ with $g(t^{-1}) \in k[t^{-1}]$. If $\mathbf{P}^1$ had a regular 1-form then we would have f, g as above with $fdt = gdt^{-1}$. If $\deg g = l$ then $g(t^{-1})dt^{-1} = -t^{-l-2}h(t)dt$ where $h(t) \in k[t]$ and $h(0) \neq 0$. But then $t^{-l-2}h(t)$ is not in $k[t]$ for any non-negative value of l and hence there can be no solutions for the above relation. □

Lemma 12.2.5 *Let C be the elliptic curve defined by $Y^2Z = F(X, Z)$. Then dx/y is a 1-form on C regular and non-vanishing (i.e., $\mathrm{ord}_P w = 0$) at all points of the curve C. where as usual $x = X/Z, y = Y/Z$ denote the dehomogenised coordinate functions on the elliptic curve.(x, y are rational functions on C).*

Proof: This has been verified in the first part of these notes on the analytic theory (see Chapter 5, section 1). We just go through the same proof and show that the proof did not really depend on the base field.

The plan is to write this 1-form as a multiple of dt for a local parameter t at any point then find its zeroes or poles as in the discussion above. For this, first we have to find a local parameter. By results in the chapter on Intersection theory (Chapter (10)) for plane curves, if X is an affine plane curve and L is a line through $P \in X$ then the function defining L is a local parameter at P if and only if $I(C, L, P) = 1$. Let us first look at a point on the affine open set $Z \neq 0$. Then x, y are regular functions at these points

and our curve is defined by the affine equation, $y^2 = f(x) = F(X/Z, 1)$. Remember that f has distinct roots. If $P = (x_0, y_0)$ is a point on the curve, and $f(x_0) \neq 0$, then $t = x - x_0$ is a regular parameter at P. Thus $dx/y = dt/y$ and since y does not vanish at P, we are done. Now let us look at a point P where $f(x_0) = 0$. Then of course $y_0 = 0$. One sees easily that $t = y$ is a regular parameter at P. Since we have the relation $2y\ dy + f'(x)\ dx = 0$ which implies $dx/y = -2dy/f'(x)$ and since $f'(x_0) \neq 0$, we get the desired result.

Finally we are left with the 'points at infinity', the points on C with vanishing Z coordinate. There is only one such point, $P = (0, 1, 0)$. Let $u = X/Y$ and $v = Z/Y$. Then the curve in the affine coordinates u, v is defined by $v = F(u, v)$ and the point at infinity P is the origin in the (u, v)-plane. Also, u is a local parameter. Notice that $u = x/y$ and $v = 1/y$. Thus the 1-form dx/y in the new coordinates is $-2\ du/(1 - F_v)$ where F_v is the partial derivative of F with respect to v. Thus the denominator does not vanish at the origin and hence the form is regular and non-vanishing. □

If $\omega \in \Omega_k(K)$ is a non-zero 1-form on C, an arbitrary projective nonsingular curve, we get a divisor $\sum_P \mathrm{ord}_P\omega$.

Exercise 12.2.3: Show that $\mathrm{ord}_P w = 0$ for all except finitely many P.

Since $\Omega_k(K)$ is of dimension 1 over K any other 1-form is of the form $f\omega$ and hence its divisor is linearly equivalent to that of ω. Thus we get a well defined class of divisors on C. This linear equivalence class is called the *canonical divisor class* of C. Let us denote this class by K_C. If C is elliptic then $K_C = 0$. The general Riemann-Roch theorem states the following:

Theorem 12.2.2 *Let C be a non-singular projective curve and let D be any divisor on C. Then*

$$h^0(D) - h^0(K_C - D) = \mathrm{deg}D + 1 - g$$

where $g = h^0(K_C)$.

We will not prove this here. (See Fulton's book [23], Chapter 8, for example).

Definition: The number g in the above theorem is called the *genus* of the curve C.

Remark: For $\mathbf{P}^1$, $g = 0$ and for an elliptic curve $g = 1$ by the lemmas above. This proves that an elliptic curve cannot be isomorphic to $\mathbf{P}^1$.

12.3 The chord and tangent construction

Let K be a field (of characteristic $\neq 2$) and E/K an elliptic curve over K, given by the equation

$$E: \quad Y^2 = X^3 + aX^2 + bX + c \tag{12.1}$$

or in the homogeneous form

$$E: \quad Y^2Z = X^3 + aX^2Z + bXZ^2 + cZ^3$$

where the polynomial X^3+aX^2+bX+c has coefficients in K and has distinct roots in some algebraic closure of K.

Note that the point O with homogeneous coordinates $O = (0 : 1 : 0)$, is a flex on the curve E, the tangent line at this point being given by the equation $Z = 0$. Moreover, by example (10.2.3) any cubic with a K-rational point can be transformed into the above form. Let $E(K)$ the set of all K-rational points of E. The main aim of this section is to set up a group structure on the set $E(K)$ and show in fact that the variety E is a variety with a group structure *i.e.*, in which the group operations are morphisms of varieties. First we give a geometric description of the group operation:

Proposition 12.3.1 *Let $P, Q \in E(K)$ be two K-rational points. Then if $P \neq Q$, we join P, Q by a line L; or if $P = Q$ then we take the tangent L at P. In either case the line L intersects E in a third point which is a* K-rational point. *We denote the third intersection point by PQ. By joining PQ and $O = (0 : 1 : 0)$ we have yet another line; this line intersects the curve once again, say in, R.* Define

$$P \oplus Q = R.$$

Then $E(K)$ is an abelian group with O as the identity and the group operation $\oplus$ as defined above. In this group law three point $P_1, P_2, P_3 \in E(K)$ "add" up to O iff the points P_1, P_2, P_3 are collinear.

Proof: First let us check that the line joining any two K-rational points on E intersects E at a third point which is again K-rational. This follows from the fact that any cubic polynomial in one variable with coefficients in K and with two roots in K, then its third root is also in K.

We first show that $P \oplus O = P$ *i.e.*, O is the identity element of the group. The line joining P and O intersects the curve in some S, but then line joining S and O intersects the curve again only at P. So $P \oplus O = P$. It is also clear that $O \oplus P = P$ and $P \oplus Q = Q \oplus P$, from the symmetric nature of construction. Hence we see that this operation is commutative and hence if the set forms a group under this operation then it is an abelian group. Now the inverse of P is not very hard to get. We draw the tangent to the curve at O. This tangent meets the curve again O (as O is flex). Join P and O. The third intersection of this line is $-P$. Let us check this by adding P and $-P$. The third intersection of the line joining P and $-P$ is O, and the line joining O and O is tangent at O, so the third intersection is also O. Hence $P \oplus (-P) = O$.

All that remains to be checked now is the associativity. So suppose we have points P_1, P_2, P_3 and we want to show that $P_1 \oplus (P_2 \oplus P_3) = (P_1 \oplus P_2) \oplus P_3$. Let us write:

$$P_1 \oplus P_2 = P_4 \tag{12.2}$$
$$P_4 \oplus P_3 = P_6 \tag{12.3}$$

and

$$P_2 \oplus P_3 = P_5 \tag{12.4}$$
$$P_1 \oplus P_5 = P_7 \tag{12.5}$$

and the two identities:

$$P_5 \oplus O = P_5 \tag{12.6}$$
$$P_4 \oplus O = P_4 \tag{12.7}$$

We want to show that $P_6 = P_7$. By the definition of the group law the equations (12.2), (12.3), (12.6) give us three lines L_1, L_2, L_3. Hence we have a cubic $C = L_1 L_2 L_3$ which meets E at the nine points P_1, P_2, $-P_4$, P_4, P_3, $-P_6$, P_5, O, $-P_5$. Similarly the remaining three equations give us three lines L_1', L_2', L_3' and so a cubic $C' = L_1' L_2' L_3'$ which meets E in the following nine

points P_2, P_3, $-P_5$, P_1, P_5, $-P_7$, P_4, O, $-P_4$. Now the two cubics C, C' have eight points in common and then by Max Noether's theorem (10.2.1), we see at once that $-P_6 = -P_7$ which implies that $P_6 = P_7$ as was required. □

The following two propositions describe the group law explicitly in terms of coordinates.

Proposition 12.3.2 *Assume that our elliptic curve is given by* eq. *(12.1). Suppose $P = (x_0, y_0)$ is a point on E. Then*

$$-P = (x_0, -y_0).$$

Proof: Following the procedure for constructing $-P$, we join P, O by a line. The homogeneous equation of this line (one checks) is given by

$$X - x_0 Z = 0.$$

To find the third point of the intersection, substituting this equation into eq. (12.1), we have a quadratic polynomial in y: $y^2 - (x_0^3 + ax_0^2 + bx_0 + c)$ and this has one root y_0. Moreover, the other root is $-y_0$. Thus it is the y-coordinate of $-P$. □

Proposition 12.3.3 *Suppose $P_1 = (x_1, y_1), P_2 = (x_2, y_2)$ are two points on E. Assume $P_1 \neq -P_2$. Define $P_3 = P_1 \oplus P_2$. Then we have the following formulae for the coordinates of $P_3 = (x_3, y_3)$:*

$$\begin{aligned} x_3 &= \lambda^2 - a - x_1 - x_2, \\ y_3 &= -\lambda x_3 - \nu, \end{aligned}$$

where λ, ν are given by the following formulae: if $x_1 \neq x_2$ then

$$\begin{aligned} \lambda &= (y_2 - y_1)/(x_2 - x_1) \\ \nu &= (y_1 x_2 - y_2 x_1)/(x_2 - x_1); \end{aligned}$$

if $x_1 = x_2$ then

$$\begin{aligned} \lambda &= \frac{3x_1^2 + ax_1 + b}{2y_1} \\ \nu &= \frac{-x_1^3 + bx_1 + 2c}{2y_1} \end{aligned}$$

Proof: Note that if $x_1 = x_2$ and if $y_1 \neq -y_2$ then $P_1 = P_2$. So in this case the line to be used to obtain the third point is the tangent at P_1.

Firstly assume P_1, P_2 are two distinct points. Let the line joining these two points be $y = \lambda x + \nu$. Since then the slope is $\lambda = (y_2 - y_1)/(x_2 - x_1)$. The y-intercept ν is given by $\nu = (y_1 x_2 - y_2 x_1)/(x_2 - x_1)$. Now to find the coordinates of the third point, substitute the equation of the line in the equation of the elliptic curve. We get the following cubic polynomial in x on doing this:

$$(\lambda x + \nu)^2 = x^3 + ax^2 + bx + c.$$

Moreover, x_1, x_2, x_3 are its three roots. So it must factor as:

$$(\lambda x + \nu)^2 - (x^3 + ax^2 + bx + c) = -(x - x_1)(x - x_2)(x - x_3).$$

Then the coefficients of x^2 on both sides must be equal and this gives:

$$\lambda^2 - a = x_1 + x_2 + x_3.$$

This gives us x_3. Now to get y_3 note that P_3 lies on the line $y = \lambda x + \nu$. Then letting $x = x_3$ we get y_3. This proves the first part of the claim.

The proof for the second part of the claim is no different except that we now consider the intersection of the curve with the tangent line $(y = \lambda x + \nu)$ at P_1. Elementary calculus and coordinate geometry tell us that the slope of the tangent at P_1 is

$$\lambda = \frac{3x_1^2 + ax_1 + b}{2y_1}.$$

And the y-intercept is given by

$$\nu = \frac{-x_1^3 + bx_1 + 2c}{2y_1}.$$

The coordinates of the third point can be computed quite easily as earlier. The formula for coordinates of $2P$ is called the *duplication formula.* □

Exercise 12.3.1: Let $E : Y^2 = X^3 + X + 9$. Then the points $P_1 = (0, 3), P_2 = (8, 23)$ are on the curve. Show that $P_1 + P_2 = (-7/4, 11/8)$.

Remark: The formulae in proposition (12.3.2) and proposition (12.3.3) show that the maps from $E \longrightarrow E$ which sends the point $P \mapsto -P$ and for a fixed P_0 the map $P \mapsto P + P_0$ are in fact rational maps from $E \longrightarrow E$. By proposition , we see that these rational maps extend in a unique way to morphisms $E \longrightarrow E$. However, it is not clear what the extensions are.

Lemma 12.3.4 *Let E/K be any elliptic curve over a field K. Then we have a natural isomorphism of groups*

$$f : E \longrightarrow \mathrm{Pic}^0(E)$$

where $f(P) = P - O$.

Proof: To show that the map is a morphism of groups we have to show that if three points P, Q, R satisfy

$$P \oplus Q \oplus R = O$$

then the divisor $P + Q + R - 3O$ is linearly equivalent to 0 in $\mathrm{Pic}^0(E)$. In other words, if three points add up to zero in the group law then we have to show that there is a rational function on E whose divisor of zeros and poles is the divisor $P + Q + R - 3O$. Since P, Q, R add up to O in the group law, they are collinear and hence lie on a line given by the equation $L = 0$, then it is easy to see that the rational function L/Z is a rational function on the curve E. Moreover its divisor of zeros and poles is clearly $P + Q + R - 3O$. So this proves that the map is a homomorphism of groups.

To prove surjectivity we use Riemann-Roch theorem. Let $D \in \mathrm{Pic}^0(E)$ be any divisor of degree zero. Then, the divisor $D + O$ is a divisor of degree 1 and so by Riemann-Roch theorem $h^0(D + O) = 1$. Then clearly $\mathrm{div}(f) = P - D - O$, and so $D \sim P - O$ and so the map is surjective.

Now we prove injectivity. Suppose $f(P) = 0$ for some P in E. Then there is a non constant rational function $f \in H^0(P)$, but by Riemann-Roch $H^0(P)$ is one dimensional and consists of constant functions. Thus we are done. □

Remark: We can now use the two additions $\oplus$ (addition in the group law) and $+$ (addition of divisors) interchangeably.

Theorem 12.3.1 *Let E be an elliptic curve over a field K. Then:*

1. *The map $i : E \to E$ given by $P \mapsto -P$ is an automorphism which is different from Identity.*

2. *The map $m : E \times E \to E$ given by $(P, Q) \mapsto P + Q$ is a morphism of varieties*

Lemma 12.3.5 *Let E be a cubic curve and let $P, Q \in E$ two not necessarily distinct points. Then there exists an automorphism of E, σ such that for any point $R \in E$, $P + Q \sim R + \sigma(R)$. Moreover σ is not the identity automorphism.*

Proof: By the Riemann-Roch Theorem (cf.), we see $h^0(P + Q) = 2$ and hence it contains a non-constant function f. This gives a non-constant morphism $f : E \to \mathbf{P}^1$ by Proposition (11.1.1). Since $E \ncong \mathbf{P}^1$ (cf. (12.2)), this morphism is separable by the following exercise.

Exercise 12.3.2: Let $K = k(t)$ be the function field of $\mathbf{P}^1$ and let $L = K(z)$ with $z^p = f \in K$ where p is the characteristic of k. Then show that $L = k(t^{1/p})$.

Also since $f^{-1}(\infty) = \{P, Q\}$, we see that it has degree 2. Thus the field extension $k(\mathbf{P}^1) \subset k(E)$ is separable of degree 2. So it is Galois. Thus we get an automorphism of $k(E)$. This induces an automorphism of E by Proposition (11.1.1) which we call σ. Now it is easy to check the lemma using the fact that $\mathrm{div}(f - \text{constant}) = 0 \in Pic^0 E$. □

Lemma 12.3.6 *Let $P_0 \in E$. Then the map given by $P \mapsto P + P_0$ is a morphism.*

Proof: Now apply Lemma (12.3.5) to the points O and P_0. As before we get a morphism σ such that

$$O + P_0 \sim P + \sigma(P).$$

Composing this with the inverse map we can easily verify the lemma. □

Proof: (of the theorem) The first assertion of the theorem now follows by applying the lemma (12.3.5) with $P = Q = O$. Then the lemma gives us a morphism $i : E \longrightarrow E$ such that for any $R \in E$,

$$R + i(R) = O + O = O,$$

so that $i(R) = -R$. Thus inversion is a non-trivial automorphism.

We now prove the second assertion. By Lemma (12.3.6), we know that translation by a fixed point on the elliptic curve is a morphism. Let $\Delta \longrightarrow$

$E \times E$ denote the diagonal. It is a closed subvariety of $E \times E$. Now by Proposition (12.3.3) it is immediate that m is a morphism on $E \times E - \Delta$, where Δ is the diagonal. We want to show that m is a morphism near the diagonal. Take any point (P, O) where $P \neq O$. Consider the following commutative diagram :

$$E \times EmE+(P,O)+PE \times EmE \tag{12.8}$$

where the downward maps are respectively translation by (P, O) and P respectively. By the previous lemma both the vertical arrows are isomorphisms (as varieties). If we denote by U the inverse image of $E \times E - \Delta$ under the left vertical map we get a commutative diagram:

$$U+(P,O)E \times E - \Delta mmE+PE \tag{12.9}$$

Now $m : E \times E - \Delta \to E$ is a morphism. Since the vertical arrows are isomorphisms one gets that $m : U \to E$ is also a morphism. But U is an open neighbourhood of Δ and hence m is a morphism everywhere. □

Definition: Let V/K be a projective variety over a field K together with a K-rational point O. We say that V is a *group variety* over K if there exists morphisms $m : V \times V \longrightarrow V$ (" the group multiplication" and a morphism $i : V \longrightarrow V$ "inverse" such that V becomes a group under the operation m with O as the identity and i as the inverse operation. In formal language this means that the morphisms m, i satisfy the following conditions:

1. m is associative *i.e.*,
$$m(x, m(y, z)) = m(m(x, y), z), \forall x, y, z \in V.$$
2. $m(x, O) = x = m(O, x) \forall x \in V$,
3. and
$$m(i(x), x) = O = m(x, i(x)).$$

Then Theorem (12.3.1) can be restated as :

Proposition 12.3.7 *Any elliptic curve over a field K is a group variety over K.*

Remark: If V_1, V_2 are two group varieties over K then so is $V_1 \times V_2$. In particular, product of a finite number of elliptic curves is again a group variety.

12.4 Curves of genus 1

We have defined an elliptic curve as a projective plane curve defined by the homogeneous equation

$$Y^2Z + aXYZ + bYZ^2 = X^3 + cX^2Z + dXZ^2 + eZ^3,$$

where the constants a, b, c, d and e are so chosen that the resulting curve is smooth. In characteristic different from 2 and 3 this form can be simplified further to yield the *Weierstrass normal form*

$$Y^2Z = X^3 + g_2XZ^2 + g_3Z^3.$$

By abuse of notation we shall refer to both these forms as the Weierstrass normal forms.

We may reformulate this in geometric terms as follows. Let C be a smooth projective plane cubic (a cubic curve defined by a cubic equation) with a k-rational point O which is a point of inflection. It was shown in section 10.2.3 that by making the tangent line at this point the line at infinity and suitably choosing the remaining coordinates we can write the equation of the curve in Weierstrass normal form; *i.e.*, C is an elliptic curve. Further, in section 10.2.2 it was shown that any smooth plane cubic has a point of inflexion over an algebraically closed field. Thus we see

Lemma 12.4.1 *Over an algebraically closed field any smooth plane cubic is an elliptic curve.*

In section 12.2 we defined a curve of genus 1 as curve with $h^0(K) = 1$. Moreover one sees easily that plane cubics are curves of genus 1. This suggests that we can generalise the above statement as follows,

Lemma 12.4.2 *Let C be a curve of genus 1 over an algebraically closed field. There is a morphism $C \to \mathbf{P}^2$ which embeds C as a smooth plane cubic (in Weierstrass normal form).*

This lemma immediately follows from the following more general

Proposition 12.4.3 *Let C be a curve of genus 1 over field k and let O be a k-rational point on C. Then there is a map $C \to \mathbf{P}^2$ which embeds C as a smooth plane cubic and the point O goes to a point of inflection on the plane curve.*

In particular, by applying the previous lemma we see that such a curve is also an elliptic curve.

Sketch of Proof: We cannot give a detailed proof of this fact here but we indicate in the next paragraph a sketch of the argument.

We have proved the Riemann-Roch theorem for elliptic curves C. For any point O on such a curve there is a non-constant function f with a pole of order exactly 2 at O and no other poles. Moreover any other such function is of the form $af + b$ for some constants a and b. Similarly, there is a non-constant function g with a pole of order exactly 3 at O and no other poles. The main unproved assertion is the Riemann-Roch theorem for curves of genus 1 as stated in Theorem (12.2.2), which will allow us to construct such functions for any curve of genus 1.

The functions f^3 and g^2 both have poles of order exactly 6 at O and no other poles. Thus there are non-zero constants a and b such that $ag^2 - bf^3$ has a pole of order at most 5. Thus there are constants c, d and e such that $ag^2 - bf^3 - cfg - df^2 - eg$ has a pole of order at most 2 at O, and by choice of f and g it has no other poles. Thus it is a constant linear combination of f and the constant function (by the "uniqueness" of f). By suitably scaling the functions we can see that we have the equation in Weierstrass form. □

Since we require explicit formulae for obtaining the equations in Weierstrass normal form, we will prove explicit versions of the above theorem in some special cases. In the arithmetical applications of later sections only these cases will be used.

12.4.1 Cubic curves

Let C be *any* smooth projective plane *cubic* curve over k, and let O be a k-rational point on C. If O is a *point of inflexion* on C, then we have seen above that C is an elliptic curve. We show the following result by means of explicit computation.

Proposition 12.4.4 *Let C be a smooth projective plane cubic curve which has a k-rational point O. Then there is a pair of rational functions (u, v) on C such that the morphism $C \to \mathbf{P}^2$ given by $(u, v, 1)$ makes C isomorphic to a smooth plane cubic curve in* Weierstrass form. *Moreover, the point O goes to the point at infinity.*

Proof: Let L denote the line in $\mathbf{P}^2$ which is tangent to C at the point O. If C and L do not meet except at O then O is a point of inflection and so by the results of section 10.2.3 mentioned above, we are done. If not, then by 10.2.2 we have another k-rational point P which lies in the intersection of C and L. Let us choose homogeneous coordinates in $\mathbf{P}^2$ such that L is given by $Z = 0$ and the points O and P are $(0, 1, 0)$ and $(1, 0, 0)$ respectively. The equation for C then takes the form

$$X^2Y + ZQ(X, Y, Z) = 0$$

where $Q(X, Y, Z)$ is a homogeneous quadratic form in (X, Y, Z).

Now, we can look upon this equations as a quadratic equation for X with coefficients as polynomials in Y and Z, *i.e.*, we have

$$(Y + a_0Z)X^2 + Za_1(Y, Z)X + Za_2(Y, Z) = 0$$

where a_0 is a constant, a_1 and a_2 are linear and quadratic forms respectively, in Y and Z. Dehomogenising with respect to Z, we wish to solve this for $x = X/Z$ and $y = Y/Z$. But this is the same as asking for a solution for a quadratic equation

$$Ax^2 + Bx + c = 0$$

given A, B and C. Such a solution can be obtained once we are given a square root of $B^2 - 4AC$. Thus, the problem of finding a point on C is equivalent to solving

$$t^2 = a_1(y, 1)^2 - 4(y + a_0)a_2(y, 1).$$

The explicit transformation is

$$2(y + a_0)x = t - a_1(y, 1)$$

Which allows us to produce x given t and vice versa. This gives the required birational isomorphism. □

12.4.2 Double covers

The equation $Y^2 = X^3 + g_2X + g_3$ expresses C as a double cover of $\mathbf{P}^1$ which is *ramified* over the zeroes of $X^3 + g_2X + g_3$ and also at the point at infinity,

i.e., if X is considered as a coordinate on $\mathbf{P}^1$, then there are two points of C for each point of $\mathbf{P}^1$ except for the zeroes as above. At these values of X the two values for Y coincide. This gives us four points of ramification. Another way of getting a double cover of $\mathbf{P}^1$ ramified at four points is to study the curve defined by $Y^2 = F(X)$, where $F(X)$ is a polynomial of degree 4. One can check that this double cover is ramified only at the zeroes of $F(X)$ and not at infinity. Over an algebraically closed field one can choose one of the four roots of $F(X)$ and take it as the point at infinity by a suitable choice of coordinates on $\mathbf{P}^1$. The equation for the double cover then takes the familiar Weierstrass normal form. More generally, we will show

Proposition 12.4.5 *Let C be the double cover of $\mathbf{P}^1$ defined by an equation of the form*

$$Y^2 = aX^4 + bX^3 + cX^2 + dX + e$$

where a, b, c, d and e are constants such that the right hand side has distinct roots, i.e., the discriminant of this degree 4 polynomial is non-zero. Moreover assume that there is a k-rational solution $P = (x, y)$ of this equation.

Then there is a pair of rational functions (u, v) on C such that the morphism $C \to \mathbf{P}^2$ given by $(u, v, 1)$ makes C isomorphic to a smooth plane cubic curve in Weierstrass form. *Moreover, the point O goes to the point at infinity.*

Proof: Suppose we have a smooth plane cubic passing through the point $(0, 1, 0)$ in $\mathbf{P}^2$. If the curve is defined by the homogeneous equation $F(U, V, W) = 0$, then we can write

$$F(U, V, W) = V^2 F_1(U, W) + V F_2(U, W) + F_3(U, W),$$

where $F_k(U, W)$ is a homogeneous polynomial of degree k in U and W. If we try to solve this equation by "completing the squares", we must find the square root Y of the discriminant

$$Y^2 = F_2(U, W)^2 - 4F_1(U, W)F_3(U, W)$$

and then we have

$$V = \frac{-F_2(U, W) \pm Y}{2F_1(U, W)}.$$

The former equation is of the same type as that for C and so this suggests a technique for the proof. We will first write a plane cubic C' which is isomorphic to our curve and then use the construction of Proposition 12.4.4 to give the Weierstrass form.

By replacing X by $X + x$ for suitable x, we may assume that the point P is of the form $(0, y)$. Thus, we see that $y^2 = e$. Now if $e = 0$, then we have a ramification point o which is k-rational and by the remarks preceding this proof we are done. Thus we may replace Y by yY and hence assume that the equation is of the form

$$Y^2 = aX^4 + bX^3 + cX^2 + dX + 1.$$

We now use the following result

Lemma 12.4.6 *Let $g(X) = aX^4 + bX^3 + cX^2 + dX + 1$ be a polynomial of degree 4. There are polynomials $f_k(X)$ for $k = 1, 2, 3$ of degree k such that*

$$g(X) = f_2(X)^2 - 4f_1(X)f_3(X).$$

Proof: Let $f_2(X) = eX^2 + fX + 1$. Then we have

$$f_2(X)^2 = e^2X^4 + 2efX^3 + (2e + f^2)X^2 + 2fX + 1.$$

Let $f = d/2$ and $e = (c - f^2)/2$, then we have

$$g(X) - f_2(X)^2 = (a - e^2)X^4 + (b - 2ef)X^3.$$

Thus, putting $f_3(X) = X^3/4$ and $f_1(X) = (e^2 - a)X + (2ef - b)$ we have the required equation. □

Now the curve C is isomorphic to the plane cubic C' given by the homogeneous equation

$$V^2((a - e^2)U + (b - 2ef)W) + V(eU^2 + fUW + W^2) + \frac{1}{4}U^3 = 0$$

where e and f are expressed in terms of c and d as above. This curve C' has a natural point $(U, V, W) = (0, 1, 0)$. In fact, the line $V = 0$ is an inflexional tangent at the point $(0, 0, 1)$. Thus we have the curve C' in Weierstrass form. The morphism from C to C' is given by sending (X, Y) to $(U, V, W) = (X, (-f_2(X) - Y)/(2f_1(X)), 1)$. The inverse morphism is given by sending (U, V, W) to the point $(X, Y) = (U/W, -(2f_1(X)V/W + f_2(X)))$. □

12.4.3 Complete intersections

In section 16.1 we will encounter curves of genus 1 which are embedded in $\mathbf{P}^3$ as curves of degree 4. Such a curve C is defined as the zero locus of a pair of homogeneous quadratic forms. As before let us assume that the curve has a k-rational point. We wish to explicitly write down a plane cubic defined by an equation in Weierstrass form such that the space curve is isomorphic to it.

Let us assume that for a suitable choice of homogeneous coordinates (W, X, Y, Z) on $\mathbf{P}^3$, the k-rational point on C is $(0, 0, 0, 1)$. The two quadratic equations for the curve then have the form

$$\begin{aligned} A(W, X, Y)Z + B(W, X, Y) &= 0 \\ C(W, X, Y)Z + D(W, X, Y) &= 0 \end{aligned}$$

Given any point (W, X, Y) in the projective plane defined by $Z = 0$, the above equations have a common solution if and only if the determinant $F(W, X, Y) = A(W, X, Y)D(W, X, Y) - B(W, X, Y)C(W, X, Y)$ is zero. In case this is zero we have $Z = -B(W, X, Y)/A(W, X, Y)$. Thus, for a general solution of $F(W, X, Y) = 0$ we can find a value of Z which satisfies the original quadratic equations. Conversely, given a point (W, X, Y, Z) which satisfies the quadratic equations, we must have $F(W, X, Y) = 0$.

Thus we have a plane cubic C' defined by $F(W, X, Y)$ which is isomorphic to C and the curve C' has a natural point given by the intersection of the two lines $A(W, X, Y) = 0$ and $C(W, X, Y) = 0$. By an application of Proposition 12.4.4 we can find an isomorphism of C' with a plane cubic C'' which is in Weierstrass form. Thus we have shown,

Proposition 12.4.7 *Let C be smooth curve in $\mathbf{P}^3$ which is defined by a pair of homogeneous quadratic forms. Assume that C has a k-rational point. Then there is a pair of rational functions (u, v) on C such that the morphism $C \to \mathbf{P}^2$ given by $(u, v, 1)$ makes C isomorphic to a smooth plane cubic curve in* Weierstrass form. *Moreover, the point O goes to the point at infinity.*

Chapter 13

Structure of endomorphisms of elliptic curves

13.1 Endomorphisms of elliptic curves

In this section we undertake in more detail, the study of morphisms between elliptic curves. This will give us very useful information about the torsion points of a given elliptic curve. This will also be a basic tool required to study the number of rational points of an elliptic curve defined over a finite field. We will be working over an arbitrary field k, though sometimes for convenience, we will assume that the characteristic of the field is different from 2 and 3. The reader may fill in the necessary extra details on such occassions. We will usually not specify the identity element of the elliptic curve, unless there is a possibility of confusion.

Let E be an elliptic curve defined over a field k. Let K be any extension of k. We denote by $E(K)$ the set of points of E which are defined over K.

Lemma 13.1.1 *Let $f : C \to D$ be a morphism between two elliptic curves such that $f(O) = O$, where O denotes the identity. Then f is a group homomorphism.*

Proof: Define $\phi : C \times C \to D$ as $\phi(x, y) = f(x+y) - f(x) - f(y)$. Then ϕ is a morphism. By hypothesis $\phi(x, O) = O$ for all $x \in C$. Thus by the rigidity lemma (9.2.4), there exists a morphism $g : C \to D$ such that $\phi(x, y) = g(y)$ for all $x, y \in C$. But $g(y) = \phi(O, y) = O$ for all $y \in C$. Thus $g \equiv 0$. So $f(x + y) = f(x) + f(y)$ for all $x, y \in C$. □

Corollary 13.1.2 *Let $f : C \to D$ be a morphism between two elliptic curves. Then $\phi(x) = f(x) - f(O)$ is a group homomorphism. In other words, upto translation, all morphisms are group homomorphisms.*

Definition: A non-constant morphism between elliptic curves, which preserves the identity is called an *isogeny.*

Example 13.1.1: Let $a, b \in k$ be two elements such that $c = a^2 - 4b \neq 0$, and $b \neq 0$.Consider the two elliptic curves

$$\begin{aligned} E_1 &: y^2 = x^3 + ax^2 + bx \\ E_2 &: v^2 = u^3 - 2au^2 + cu \end{aligned}$$

We can define isogenies of degree two as follows:

$$\begin{array}{cc} f : E_1 \to E_2 & \widehat{f} : E_2 \to E_1 \\ (x, y) \mapsto (\frac{y^2}{x^2}, \frac{y(b - x^2)}{x^2}) & (u, v) \mapsto (\frac{v^2}{4u^2}, \frac{v(c - u^2)}{8u^2}) \end{array}$$

With some effort one may also check that $f \circ \widehat{f} = \widehat{f} \circ f =$ multiplication by 2 on the elliptic curves.

Denote by $\mathrm{End}(E)$ the ring of endomorphisms, which preserve the identity, of an elliptic curve E. If $f \in \mathrm{End}(E)$ and $f \neq 0$ then by ((11.1.2)) f is finite and hence has a degree. Defining the degree of the zero element to be zero, we have a map $d : \mathrm{End}(E) \to \mathbf{Z}$. Notice that for $f, g \in \mathrm{End}(E)$, $d(f \circ g) = d(f)d(g)$. This proves the following:

Lemma 13.1.3 *The ring of endomorphisms of an elliptic curve has no zero divisors. More generally, if $f : E_1 \to E_2$ is a non-constant morphism of elliptic curves and $g : E_2 \to E_3$ another morphism of elliptic curves then $g \circ f$ is constant implies that g is constant.*

At present the only elements in $\mathrm{End}(E)$ that we know are multiplication by integers m, which we denote by $[m]$. Next we will show that when $m \neq 0$, $[m]$ is an isogeny.

Lemma 13.1.4 $[m] \neq 0$ *in* $\mathrm{End}(E)$ *if* $m \neq 0$.

Proof: By theorem 12.3.1 the morphism $P \mapsto -P$ is not the identity automorphism. Thus there is a point $P \in E(K)$ such that $[2]P \neq 0$. If we write $m = m_1 m_2$ then to show that $[m] \neq 0$ it suffices (by lemma 13.1.3), to show that $[m_1]$ and $[m_2]$ are not zero. Thus we only need to prove the lemma for odd values of m. We will do this when characteristic is not 2 (in 13.4.2 we will give a different proof for the case of characteristic 2). We may assume that the base field is algebraically closed. We may write the elliptic curve in the form

$$y^2 = x^3 + ax^2 + bx + c.$$

For the three distinct roots x_0 of the above cubic, if $P = (x_0, 0)$ then $P = -P$. Thus there are three non-trivial 2-torsion points. If m is odd then $[m] = 0$ will imply that $mP = 0$ for the three 2-torsion points also. But $mP = 0$ and $2P = 0$ implies that $P = 0$, which is absurd. □

Corollary 13.1.5 *Let E be an elliptic curve defined over k and K an algebraically closed field containing k. Then $E(K)$ is a divisible group.*

Proof: By the previous lemma for any non-zero integer m, multiplication by m is not zero. Since the image of $[m]$ is closed and E is irreducible we see that $[m]$ is onto the set $E(K)$. This means that the group $E(K)$ is divisible. □

Corollary 13.1.6 *We have an injective ring homomorphism*

$$[\ \ ] : \mathbf{Z} \to \mathrm{End}(E)$$

'Usually' an elliptic curve has no more endomorphisms.

Definition: If the above ring homomorphism is not surjective, we say that the elliptic curve has *complex multiplication.*

Example 13.1.2: Let E be an elliptic curve over k, of characteristic not equal to 2. Consider the elliptic curve given in Weierstrass form

$$y^2 = x^3 - x.$$

Then E has an endomorphism which we denote by $[\iota]$, which is not multiplication by an integer. It is defined as follows:

$$[\iota] : (x, y) \mapsto (-x, iy)$$

Here i is the square root of -1. Since $[\iota] \circ [\iota] = [-1]$ this endomorphism is not in the image of $\mathbf{Z}$.

13.2 The Frobenius morphism

For this section we will assume that we are working over a field of non-zero characteristic p. We will define Frobenius morphisms for varieties over such a field. These are very important for the study of elliptic curves over finite fields. In fact they provide a very powerful tool for studying varieties over any field of positive characteristic. Here we will work over a finite field or an algebraically closed field or more generally a *perfect field* for simplicity. The general case is technically somewhat harder. We begin by defining the Frobenius morphism for function fields. By (11.2.1) this will enable us to study this morphism for curves.

Given a field K of characteristic $p > 0$, we have a natural inclusion of fields $K \subset K^{1/p}$. This is because in characteristic p, any element has a *unique* p^{th} root and the corresponding map is a map of fields. Assume that K is a function field over a perfect field k (*i.e.*, K is a finitely generated field over k). Since k is perfect every element has a p^{th} root. So $k^{1/p} = k$. Thus $k \subset K \subset K^{1/p}$ is an inclusion of fields over k. We can express this better in the following diagram:

$$\begin{array}{ccc} k & \longrightarrow & K \\ \cong \downarrow & & \downarrow \\ k^{1/p} & \longrightarrow & K^{1/p} \end{array}$$

So what we are really doing is viewing K as a field over $k^{1/p}$, which is of course same as k via the inverse of the isomorphism above, $k \to k^{1/p}$. Then the field inclusion $K \to K^{1/p}$ becomes a map of fields over $k^{1/p}$. We hope that the following examples will clarify the situation.

Example 13.2.1: Let $K = k(t)$ where k is algebraically closed and t an indeterminate. Then the above inclusion is nothing but $k \subset k(t) \subset k(t^{1/p})$. In this case the two fields $k(t)$ and $k(t^{1/p})$ are isomorphic as fields over k. This is not true in general, as seen in the following example.

Example 13.2.2: Let k be an algebraically closed field of positive characteristic not equal to 2 or 3. Let $K = k(x, y)$ with $y^2 = x^3 + ax + b$ where $a, b \in k$. Then $K^{1/p}$ as a field over $k^{1/p} = k$ is isomorphic to $k(u, v)$ with $v^2 = u^3 + a^{1/p}u + bu$. The inclusion of fields being $x \mapsto u^p$ and $y \mapsto v^p$. We may choose a, b suitably so that two fields are not isomorphic.

Now let us define the Frobenius for curves over any field of non-zero characteristic p. Let the curve C be defined by the polynomial $f(x,y) = 0$. Let us denote by $g(x,y)$ the polynomial obtained by raising the coefficients of f to the p^{th} power. Let $C^{[p]}$ be the curve defined by $g = 0$. Then we have a morphism

$$\phi : C \to C^{[p]} \;\; \phi(x,y) = (x^p, y^p).$$

If k is perfect and K is the function field of the curve $C^{[p]}$ then the function field of C is $K^{1/p}$. Thus given a curve D over such a field k we can find a curve C such that $D = C^{[p]}$; we can then write $D^{[1/p]} = C$. If we iterate the construction then for every integer n we get curves $C^{[q]}$, $q = p^n$ defined by the functions g_n got by replacing the coefficients of f by its q^{th} roots.

Now let us assume that the base field is finite say of $q = p^n$ elements. This means that the equation f is defined over k. So the coefficients of f are invariant under the map $x \to x^q$. So $\phi^n = \phi_q$ defines a morphism from $C \to C$. Thus we get an interesting morphism of the curve to itself.

Remark: Notice that the Frobenius map is purely inseparable and has degree p.

Exercise 13.2.1: If C is a curve defined over a finite field k of q elements, then the fixed points of ϕ_q are precisely the k-rational points of the curve.

Lemma 13.2.1 *Let $f : C \to D$ be a k-morphism of curves where k is a perfect field of characteristic p, and assume that f is not separable. Then*

1. *There exists a separable morphism $g : C^{[p]} \to D$ such that $f = g \circ \phi^n$.*
2. *There exists a separable morphism $h : C \to D^{[1/p]}$ such that $f = \phi^n \circ h$.*

Proof: We have $f^* : k(D) \subset k(C)$, which is a finite field extension. This can be factorised in two ways. In the first case we replace $k(D)$ by its separable closure in $k(C)$. In the second case, we replace $k(C)$ by the maximal subfield which is purely inseparable over $k(D)$ (this is obtained by adjoining to $k(D)$ all $\alpha \in k(C)$ such that $\alpha^{p^n} \in k(D)$ for some n). Thus we can assume that f is purely inseparable.

If we have a finite purely inseparable extension of fields $K \subset L$ of transcendence degree 1, then we will show that $K = L^{p^n}$ for some n. Take $u \in L - K$ such that $u^p \in K$. We need to check that $K(u)^p = K$. Then by

an easy induction we would be done. So let us assume that $L = K(u)$ with $u \notin K$ and $u^p = v \in K$. Since the fields have transcendence degree 1, u and v are transcendental and the fields are finite extensions of $k(u)$ and $k(v)$. It is clear that $L^p \subset K$. So if we show that for fields K of transcendence degree 1, $\deg[K : K^p] = p$, we would be done.

Let $u \in K$ a separating transcendental basis. Consider $L = K^p(u) \subset K$. Then every element of K is separable as well as purely inseparable over L. So $L = K$. But $u^p \in K^p$ and so the degree of the extension is at most p.

Let $[k(C) : k(D)] = q = p^n$. Then $k(C)^q \subset k(D)$. Since k is perfect and the transcendence degree of $k(C)$ is $\dim C = 1$ we have seen that $[k(C) : k(C)^q] = q$. Hence $k(D) = k(C)^q$ and this implies that $f = \phi^n$, thereby proving the lemma. □

Corollary 13.2.2 *Let $E \subset \mathbf{P}^2$ be a smooth cubic curve defined over a finite field k. Then it has a k-rational point. In particular it is an elliptic curve.*

Proof: Let $\#(k) = q$. Then we have the Frobenius morphism ϕ_q defined as above. E is of course an elliptic curve over the algebraic closure of k. So we have a morphism

$$f : E \to E, f(P) = \phi_q(P) - P.$$

Since E is complete and irreducible $f(E)$ is closed in E. So the image must either be a single point or the whole of E. If it is the whole of E, then there exists a point P such that $f(P) = O$, which implies that $\phi_q(P) = P$ which in turn means that P is a k-rational point. So we need to show that f is not constant. This will we shown in lemma 13.4.4. Until then we can argue as follows.

Assume $f(P) = P_0 \forall P \in E$. So $\phi_q(P) = P + P_0$. So $\phi_q^n(P) = P + nP_0$ for all positive integers n. Let K be a finite extension of degree n, over which E has a rational point say Q_0. Then

$$\phi_q^n(Q_0) = \phi_{q^n}(Q_0) = Q_0$$

So $nP_0 = O$. But then $\phi_{q^n} = Id$ and hence all points of E are defined over K. Since K is finite, $\#(\mathbf{P}^2(K))$ is finite and this means that the curve E has only finitely many points defined over the algebraic closure. This is absurd. □

This proof is due to S. Lang. You must contrast this with the following well known theorem due to Chevalley and Warning.

Theorem 13.2.1 (Chevalley-Warning[50]) *Let $f(x_0, x_1, \ldots, x_n)$ be a homogeneous polynomial of degree d over a finite field k. Assume that $d \leq n$. Then f has a non-trivial zero. i.e., there exists $(\alpha_0, \alpha_1, \ldots, \alpha_n)$ $\alpha_i \in k$, not all of them zero and $f(\alpha_0, \alpha_1, \ldots, \alpha_n) = 0$.*

This theorem will not give the above corollary, since $d = 3$ and $n = 2$.

Remark: In the previous corollary, we have not really used the fact that our curve is a cubic curve. What we really proved is that a curve over a finite field which over the algebraic closure becomes an elliptic curve is actually an elliptic curve.

Definition: Let E be an elliptic curve defined over a field k and m any integer. We denote by

$$E[m] = \{P \in E(k) \mid mP = O\}$$

the m-torsion points of E.

Example 13.2.3: Let E be an elliptic curve given by the Weierstrass equation $y^2 = f(x) = x^3 + ax^2 + bx + c$. We have already seen in lemma (13.1.4) that the three points (x_i, y_i) on the elliptic curve where x_i are the three distinct roots of $f(x)$ give three non-trivial 2-torsion points of E. We will later show that there are no more 2-torsion points on an elliptic curve. In fact we will later see that

$$E[m] \cong \mathbf{Z}/m\mathbf{Z} \times \mathbf{Z}/m\mathbf{Z}$$

when m is prime to the characteristic of the base field. If $m = p^n$ where p is the characteristic then it is also true that

$$E[m] \cong \mathbf{Z}/m\mathbf{Z} \text{ or } 0.$$

Exercise 13.2.2: Show that the nine flexes of E are the 3-torsion points.

13.3 Dual isogenies

In this section we will construct what is known as a dual isogeny. Given an endomorphism ϕ of an elliptic curve (or more generally a morphism $\phi : E_1 \to E_2$ from an elliptic curve to another elliptic curve) we will construct another endomorphism $\widehat{\phi}$ (a morphism $\widehat{\phi} : E_2 \to E_1$) such that $\phi \circ \widehat{\phi} = [\deg \phi]$. We will call such a $\widehat{\phi}$ as a *dual isogeny* of ϕ.

Theorem 13.3.1 *Let $f : E_1 \to E_2$ be an isogeny between two elliptic curves. Then*

1. $\#(\ker f) = \deg_s f$, *where* $\deg_s$ *denotes the separable degree.*
2. *We have a natural map* $\ker f \to \mathrm{Aut}[K(E_1)/f^*K(E_2)]$ *which is an isomorphism.*
3. *Assume that the ground field k is perfect. Then the fixed field $K(E_1)^{\ker f}$ is $K(E_2)^{1/q}$, where q is a positive power of the characteristic.*

Proof: 1. One knows that for a general point $Q \in E_2$, $\#f^{-1}(Q) = \deg_s f$. Since f is a group homomorphism this must be same for all points.
2. Let $P \in \ker f$. P induces an automorphism of E_1 by translation.

Exercise 13.3.1: Since the group law is induced by a morphism $\mu : E \times E \to E$, show that for any finite subgroup $A \subset E(K)$, we have an induced homomorphism $A \to Aut(K(E)/K)$.

This gives a homomorphism $\ker f \to \mathrm{Aut}(K(E_1)/K)$. If $x \in f^*K(E_2)$ since $x(Q) = x(P+Q)$ for any $Q \in E_1$ we see that these automorphisms fix $f^*K(E_2)$. Thus we have a well defined map as in the theorem. Since $\#\ker f = \deg_s f$ and $\#\,\mathrm{Aut}[K(E_1)/f^*K(E_2)] \leq \deg_s f$, if we show that the above map is injective, we will be done. So let $P \in \ker f$ and assume that it acts trivially on $K(E_1)$. But then for any rational function $x \in K(E_1)$, $x(P) = x(O)$. But this implies that $P = O$ proving the injectivity.
3. By Lemma (13.2.1) we can factorise $f = \phi^n \circ g$, where $g : E_1 \to E_2^{[1/q]}$ ($q = p^n$) is a separable morphism. By Part 1 we have the equation

$$\#(\ker g) = \deg_s g = \deg_s f = \#(\ker f)$$

and clearly $\ker g \subset \ker f$, thus $\ker g = \ker f$. Now by Part 2 we are done. □

Lemma 13.3.1 *Let E_1, E_2, E_3 be elliptic curves. Let $f : E_1 \to E_2$ and $g : E_1 \to E_3$ be two isogenies. Assume that $\ker f \subset \ker g$ and that f is separable. Then there exists an $h : E_2 \to E_3$ such that $h \circ f = g$.*

Proof: By the previous theorem, we know that $f^* : K(E_2) \to K(E_1)$ and is a Galois extension with Galois groups $\ker f$. Moreover, $K(E_3)$ is contained in the fixed field of the action of $\ker g$ on $K(E_1)$. The inclusion $\ker f \subset \ker g$ then implies that there exists an inclusion $h^* : K(E_3) \to K(E_2)$ so that $f^* \circ h^* = g^*$. This map h induces a morphism $h : E_2 \to E_3$ with the required properties. (11.1.1) □

Now we come to the main theorem of this section.

Theorem 13.3.2 *Let E_1, E_2 be elliptic curves and $f : E_1 \to E_2$ a morphism with $f(O) = O$. Let $\deg f = m$. Then there exists a morphism $\widehat{f} : E_2 \to E_1$ such that $f \circ \widehat{f} = [m]$. Further if f is not constant then this element is unique.*

Proof: If $f \equiv O$ then we define $\widehat{f} \equiv O$. So we will assume that f is not constant and its degree is $m \neq 0$. First let us prove the uniqueness. If $\widehat{f}$ and $\widehat{g}$ are two such, then $f \circ (\widehat{f} - \widehat{g}) = 0$. Since $f \neq O$, this implies that $\widehat{f} = \widehat{g}$ (by lemma (13.1.3)).

Next assume that $f = g \circ h$ and assume that we have found $\widehat{g}$ and $\widehat{h}$. We then claim that $\widehat{f} = \widehat{h} \circ \widehat{g}$. Let $a = \deg g$ and $b = \deg h$

$$g \circ h \circ \widehat{h} \circ \widehat{g} = g \circ [b] \circ \widehat{g} = [b] \circ g \circ \widehat{g} = [b] \circ [a] = [b \cdot a] = [\deg f].$$

By Theorem 13.3.1 any endomorphism can be factored into separable and inseparable parts. Thus we may treat these cases separately.
Separable case. In this case since f has degree m, $\#\ker f = m$. Thus $\ker f \subset \ker[m]$. By lemma (13.3.1), we get an element $\widehat{f} : E_2 \to E_1$ such that $\widehat{f} \circ f = [m]$. One can easily check that $f \circ \widehat{f}$ also is $[m]$.
Inseparable case. Now assume that f is purely inseparable. Then by lemma (13.2.1), we see that $f = \phi^n$ for some positive integer n, where as usual ϕ denotes the Frobenius. So by the argument in the beginning, we have to prove the theorem only for $f = \phi$. Now $[p]$ is not separable. So by lemma (13.2.1) above there exists a $\widehat{\phi}$ such that $[p] = \phi \circ \widehat{\phi}$. Since the degree of ϕ is p, we are done. □

Now we come to a crucial property of the dual map, which is rather deep. People who are familiar with more Algebraic Geometry will be able to recognise a line of thought which is crucial in many deeper arguments in the subject.

We will first give a less mysterious interpretation of the dual.

Lemma 13.3.2 *Let $f : E_1 \to E_2$ be an isogeny. Then we have the following description of $\widehat{f}$. If $Q \in E_2$, $P \in f^{-1}(Q)$ and $d = \deg_i f$, the inseparable degree of f, then*

$$\widehat{f}(Q) = d(\sum_{R \in f^{-1}(Q)} R - \sum_{S \in \ker f} S) = [\deg f]P$$

where the addition and subtraction are done using the group law.

Proof: The first equality is obvious by first applying f to both sides and then using the fact that $\#f^{-1}(Q) = \deg_s f$. Notice that if $\eta \in \ker f$ then $f(P + \eta) = Q$ and all the inverse images of Q are of this form. So

$$\begin{aligned} \widehat{f}(Q) &= d(\sum_{\eta \in \ker f} (P + \eta) - \sum_{\eta \in \ker f} \eta) \\ &= d(\sum_{\eta \in \ker f} (P)) = [\deg f]P \end{aligned}$$

□

Theorem 13.3.3 *Let $f, g : E_1 \to E_2$ two isogenies. Then $\widehat{f + g} = \widehat{f} + \widehat{g}$.*

Proof: *Step 1.* Let K_i = the rational functions on E_i $i = 1, 2$. Giving a morphism f from E_1 to E_2 is just giving a K_1-valued point of E_2.*i.e.*, an element of $E_2(K_1)$. This can be seen as follows. Let $G_1(x, y) = 0$ define E_1 and $G_2(u, v) = 0$ define E_2. Then f induces a field inclusion of $K_2 \subset K_1$. Thus we can express $u = p(x, y)$, $v = q(x, y)$ where $p, q \in K_1$. So we have found $p, q \in K_1$ satisfying $G_2(p, q) = 0$, and this gives a point in $E_2(K_1)$. Conversely, assume that we are given $P \in E_2(K_1)$. This is the same as giving two elements $p, q \in K_1$ such that $G_2(p, q) = 0$. Thus we get a rational map $E_1 \to E_2$ by sending $(x, y) \mapsto (p(x, y), q(x, y))$. But a rational map is a morphism since E_1 is smooth and E_2 is projective. (11.1.1)

Let us denote this point by $t(f)$.

Step 2. The definition of $f+g$ implies that $t(f+g)-t(f)-t(g)=O$ in the group $E_2(K_1)$. So let us consider the divisor $D=t(f+g)-t(f)-t(g)+O$. Then its degree is zero and it is O in the group law. So it is linearly equivalent to zero. Thus there exists a rational function $R \in K(E_2(K_1)) = K_1K_2$ whose divisor is the above divisor. Notice that K_1K_2 is the set of rational functions on $E_1 \times E_2$.
Step 3. Let Γ_f denote the graph of the morphism f in $E_1 \times E_2$.The above discussion shows that if we look at the zeroes and poles of R on $E_1 \times E_2$, then it has zeroes along Γ_{f+g} and $E_1 \times \{O\}$. It has poles along Γ_f and Γ_g. And all the remaining zeroes and poles are along curves of the form $\{P\} \times E_2$ for finitely many points P. Now take a general point $Q \in E_2$ and restrict the function to a rational function on $E_1 \times \{Q\}$. We know that the divisor corresponding to this rational function must add up to O in the group law. But we can compute this divisor in another fashion, since we know the zeroes and poles of R on $E_1 \times E_2$.

For this first let us see what $\Gamma_f \cap (E_1 \times \{Q\})$ is. If you think about it for a moment it is nothing but $f^{-1}(Q)$. Now by the previous lemma this divisor adds up to $\widehat{f}(Q)$. Also $(\{P\} \times E_1) \cap (E_1 \times \{Q\}) = P \times Q$. Putting these together we see that $\widehat{f+g}(Q) - \widehat{f}(Q) - \widehat{g}(Q)$ is independent of Q for a general point Q. But it is a morphism and since it maps most points to a single point, it must be constant. Since it is also a group homomorphism, it must be identically equal to O. So $\widehat{f+g} = \widehat{f} + \widehat{g}$. □

Now we are ready to derive a number of useful facts.

Theorem 13.3.4 *1. If $[m]$ is the multiplication by m of an elliptic curve E then $\widehat{[m]} = [m]$.*

2. $\deg [m] = m^2$.

3. If $f : E_1 \to E_2$ is an isogeny then $\widehat{(\widehat{f})} = f$ and $\deg f = \deg \widehat{f}$.

Proof: 1. We will prove it by induction. For $m = 0$ or 1 it is obvious. Now

$$\widehat{[m+1]} = \widehat{[m]} + \widehat{[1]} = [m] + [1] = [m+1].$$

2. By definition we know that $[m] \circ \widehat{[m]} = [\deg [m]]$. By the first step $[m] \circ \widehat{[m]} = [m] \circ [m] = [m^2]$. Thus $\deg [m] = m^2$.

3. Let $m = \deg f$. Then

$$[m^2] = [\deg [m]] = [\deg(f \circ \widehat{f})] = [\deg f \cdot \deg \widehat{f}] = [m \deg \widehat{f}]$$

which clearly implies that $\deg \widehat{f} = m$.

$$f \circ \widehat{f} = [m] = \widehat{[m]} = \widehat{f \circ \widehat{f}} = \widehat{(\widehat{f})} \circ \widehat{f}$$

and this clearly implies that $f = \widehat{\widehat{f}}$. □

Corollary 13.3.3 *Let E be an elliptic curve (over an algebraically closed field). Then*

1. *if m is prime to the characteristic then $E[m] \cong \mathbf{Z}/m\mathbf{Z} \times \mathbf{Z}/m\mathbf{Z}$.*
2. *if $m = p^n$ where p is the characteristic, then $E[m] \cong \mathbf{Z}/m\mathbf{Z}$ or 0.*

Proof: If m is prime to the characteristic, then multiplication by m is a separable map. So $\#\ker[m] = \deg[m] = m^2$. Since this is true for any divisor of m also an elementary argument in group theory will finish the proof of the first part.

Since multiplication is purely inseparable, we see that $\#\ker[p] = p$ or 0. Then an easy induction will finish the proof. □

13.4 Additivity of the invariant differential

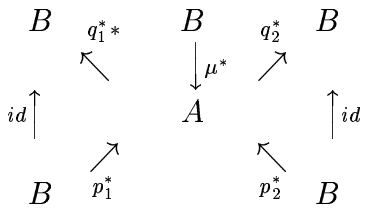
In this section we want to analyse the behaviour of the 1-form $\omega = dx/y$ of an elliptic curve given in Weierstrass form $y^2 = x^3 + ax + b$ under addition. We will prove the following theorem.

Theorem 13.4.1 *Let $\mu : E \times E \to E$ be the group law. Let x_i, y_i denote the respective Weierstrass coordinates and $\omega_i = dx_i/y_i$. Then $\mu^*\omega_3 = \omega_1 + \omega_2$.*

Proof: We know that ω_3 is a regular 1-form and it has no zeroes on E. We also know that ω_3 is a meromorphic 1-form implies $\mu^*\omega_3$ is a meromorphic 1-form on $E \times E$ since μ is dominant. Also it is easy to see that any meromorphic 1-form on $E \times E$ is of the form $f\omega_1 + g\omega_2$, where f, g are rational functions on $E \times E$. So we may write $\mu^*\omega_3$ in this form. Since ω_1 and ω_2

have no zeroes one sees that they generate freely Ω_P as an $\mathcal{O}_P$ module for any point $P \in E \times E$. Since μ is a morphism we see that μ^* takes $\mathcal{O}_{\mu(P)}$ to $\mathcal{O}_P$ for any point $P \in E \times E$. This implies that $\mu^*\omega_3$ is also regular at P for any point P. So f, g are regular on $E \times E$ and hence must be constants.

To prove that these constants are equal to 1, we proceed as follows. If O denotes the identity element of E we have the following commutative diagram, where $A = \mathcal{O}_{(O,O)}$ denotes the local ring of regular functions at $(O, O) \in E \times E$ and $B = \mathcal{O}_O$, the local ring at O of E, and $q_1^* = (id \times O)^* : A \to B$, $q_2^* = (O \times id)^* : A \to B$.



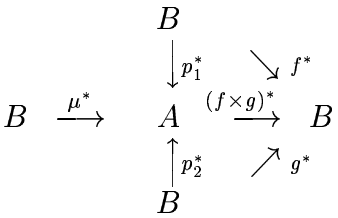
Now $q_1^*\mu^* = q_1^*p_1^* = id$ and this will easily imply that $f = 1$. Similarly we get $g = 1$, proving the theorem. □

Corollary 13.4.1 *Let f, g be two morphisms from E to E which take O to itself. Then*

$$(f + G)^*\omega = f^*\omega + g^*\omega,$$

where as before ω is the regular 1-form on E.

Proof: Let A, B be as in the earlier proof. We have the commutative diagram



We have

$$(f + g)^*\omega = (f \times g)^*\mu^*\omega.$$

But by the previous theorem,

$$\mu^*\omega = p_1^*\omega + p_2^*\omega.$$

Thus

$$(f \times g)^* \mu^* \omega = (f \times g)^* p_1^* \omega + (f \times g)^* p_2^* \omega = f^* \omega + g^* \omega.$$

Thus the required equality is true of the 1-forms in $\Omega(A)$. But since $\Omega(A) \subset \Omega(k(E))$ we have the corollary. □

Corollary 13.4.2 *Let m be any integer. Then $[m]^* \omega = m\omega$.*

Proof: For $m = 1$ this is obvious. Using the above corollary we have $[m + 1]^* \omega = [m]^* \omega + \omega$. Now an induction will finish the proof. □

Remark: The above corollary gives another proof of the fact that $[m] \neq 0$ at least when m is relatively prime to the characteristic.

Corollary 13.4.3 *Let m be any integer. Then if m is relatively prime to $p =$ the characteristic of the base field, then $[m]$ is separable. $[p]$ is inseparable.*

Proof: If m is relatively prime to p then $0 \neq m\omega = m^* \omega$ and hence the map $[m]$ is separable. If $m = p$ then since $p\omega = 0$, this map is inseparable. (These field theoretic statements are described in the following exercise). □

Exercise 13.4.1: Let K be a function field in one variable over an algebraically closed field k. Let L be a finite extension. Then show that L is separable over K if and only if the natural map $\Omega(K) \to \Omega(L)$ is non-zero.

Corollary 13.4.4 *Let E be an elliptic curve defined over a field of positive characteristic and let ϕ be the Frobenius. Then $1 - \phi$ is separable.*

Proof: We know that ϕ is purely inseparable and hence $\phi^* \omega = 0$. But then by the theorem, one has

$$(1 - \phi)^* \omega = \omega$$

and hence it is not zero. Thus $1 - \phi$ is separable. □

13.5 The Riemann Hypothesis

We will compute the number of rational points on an elliptic curve defined over a finite field of q elements. Before this let us compute the number of rational points on certain other varieties which are easier to compute.

Lemma 13.5.1

$$\#\mathbf{P}^n(k) = q^n + q^{n-1} + \cdots + q + 1$$

where $q = \#k$.

Proof: The points in $\mathbf{P}^n$ are tuples of the form $(x_0, \ldots, x_n)$ where $x_i \in k$ and all of them are non-zero. So there are $q^{n+1} - 1$ such tuples. Also we have a relation and under this relation $q - 1 = \#k^*$ elements are identified. So the number of points in $\mathbf{P}^n$ is $(q^{n+1} - 1)/(q - 1) = q^n + \ldots + q + 1$. □

Since conics in $\mathbf{P}^2$ are isomorphic to $\mathbf{P}^1$ (we are working over finite fields, and hence this is implied by Chevalley-Warning theorem) we see that they have $q + 1$ points. A priori an elliptic curve may contain all the $q^2 + q + 1$ points of $\mathbf{P}^2$. But looking at the case of conics it seems that there might be better bounds for cubics too. This is indeed what we will prove in this section.

Towards this goal we will first study the degree map on the endomorphism ring of an elliptic curve.

Let $d = \deg : \mathrm{End}(E) \to \mathbf{Z}$ be the degree map where E is any elliptic curve.

Lemma 13.5.2 *d gives a positive definite quadratic form on* $\mathrm{End}(E)$.

Proof: Let $A(f, g) = d(f + g) - d(f) - d(g)$. We will first check that A is bilinear. Since $[d(f)] = f \circ \widehat{f}$ we can compute $A(f, g)$ as follows.

$$[A(f,g)] = [d(f+g)] - [d(f)] - [d(g)] = (f+g) \circ \widehat{(f+g)} - f \circ \widehat{f} - g \circ \widehat{g} = f \circ \widehat{g} + g \circ \widehat{f}$$

Now it is easy to see the bilinearity. Since $A(f, f) = 2d(f)$, we see that d is a quadratic form. Degree of a non-constant map being positive, we get the positive definiteness. □

Next we prove a version of the Cauchy-Schwarz inequality.

Lemma 13.5.3 *Let $d : G \to \mathbf{Z}$ be a positive definite quadratic form, where G is an abelian group. Then for all $x, y \in G$*

$$\mid d(x+y) - d(x) - d(y) \mid \leq 2\sqrt{d(x)d(y)}.$$

Proof: Let A be as in the previous proof. Then

$$\leq d(ax+by) = \frac{1}{2}A(ax+by, ax+by) = a^2 d(x) + abA(x,y) + b^2 d(y)$$

for all integers a and b. Taking $a = -A(x,y)$ and $b = 2d(x)$ we get

$$\leq d(x)[4d(x)d(y) - A(x,y)^2].$$

Now the lemma is clear. □

The following theorem was originally conjectured by E. Artin and proved by H. Hasse.

Theorem 13.5.1 (Riemann Hypothesis) *Let E be an elliptic curve defined over a finite field k of q elements. Then*

$$\mid \#E(k) - q - 1 \mid \leq 2\sqrt{q}.$$

Proof: Let $\phi = \phi_q$ be the Frobenius as defined earlier. Then we want to know the fixed points of ϕ *i.e.*, kernel of the endomorphism $1 - \phi$. Since $1 - \phi$ is separable, we have $\#\ker\phi = \deg(1-\phi)$. Now apply the previous lemma with $G = \mathrm{End}(E)$, $d =$ the degree map, $x = 1$ and $y = -\phi$ to get the desired result. □

Example 13.5.1: Let $x^3 + y^3 + z^3 = 0$ be the Fermat cubic over the field k of four elements. If $a \neq 0 \in k$, then $a^3 = 1$. So one sees that any solution of the above cubic must have at least one vanishing coordinate. If $z = 0$ then all solutions of the cubic are of the form $(a, 1, 0)$ where a is any nonzero element of k. Thus the cubic has exactly 9 rational points, the maximum allowed by Riemann Hypothesis.

13.6 The structure of the endomorphism ring

We will finish these lectures by describing the endomorphism ring of an elliptic curve. We know already by lemma (13.1.3), that $\operatorname{End} E$ of an elliptic curve E has no zero divisors. We also have an anti-involution $f \mapsto \widehat{f}$ on $\operatorname{End} E$. We will use it to describe the endomorphism ring more precisely.

Lemma 13.6.1 *Let $f \in \operatorname{End} E$. Define trace of f, $T(f) = f + \widehat{f}$. Then $T(f) \in \mathbf{Z}$.*

Proof: Define norm of f to be $N(f) = f\widehat{f}$. We know that $N(f)$ is an integer by the definition of $\widehat{f}$. Now

$$T(f) = N(f+1) - N(f) - 1$$

and hence it is an integer. □

Lemma 13.6.2 *$f \in \operatorname{End} E$ satisfies a quadratic polynomial over $\mathbf{Z}$.*

Proof:

$$P(X) = (X - f)(X - \widehat{f})$$

is satisfied by f. □

Let me recall some standard definitions.

Definition: Let $\mathcal{K}$ be a $\mathbf{Q}$-algebra (not necessarily commutative), which is finite dimensional as a vector space over $\mathbf{Q}$. An *order* $\mathcal{R}$ of $\mathcal{K}$ is a subring, which is finitely generated as a $\mathbf{Z}$-module.

Definition: A (definite) *quaternion algebra* over $\mathbf{Q}$ is an algebra of the form

$$\mathcal{K} = \mathbf{Q} + \mathbf{Q}\alpha + \mathbf{Q}\beta + \mathbf{Q}\alpha\beta$$

satisfying

$$\alpha^2, \beta^2 \in \mathbf{Q}, \quad \alpha^2 < 0, \quad \beta^2 < 0, \quad \alpha\beta = -\beta\alpha.$$

Now we will prove the structure theorem for endomorphism rings of elliptic curves.

Theorem 13.6.1 *Let $\mathcal{R} = \operatorname{End} E$ where E is an elliptic curve. Then $\mathcal{R}$ is one of the following.*

1. $\mathcal{R} \cong \mathbf{Z}$.
2. $\mathcal{R}$ *is an order in an imaginary quadratic extension of* $\mathbf{Q}$.
3. $\mathcal{R}$ *is an order in a quaternion algebra over* $\mathbf{Q}$.

Proof: Let $\mathcal{K} = \mathcal{R} \otimes \mathbf{Q}$. First we analyse the structure of $\mathcal{K}$. If $\mathcal{K} \neq \mathbf{Q}$ choose $\alpha \in \mathcal{K} - \mathbf{Q}$. Then changing α to $\alpha - \frac{1}{2}T(\alpha)$ if necessary, we may assume that $T(\alpha) = 0$. So $\alpha^2 = -N(\alpha) < 0$ by lemma (13.5.2). Thus $\mathbf{Q}(\alpha)$ is an imaginary quadratic extension of $\mathbf{Q}$. If $\mathcal{K} \neq \mathbf{Q}(\alpha)$ then pick an element $\beta \in \mathcal{K}$ but not in $\mathbf{Q}(\alpha)$. Then replacing β by

$$\beta - \frac{1}{2}T(\beta) - \frac{1}{2}(T(\alpha\beta)/\alpha^2)\alpha$$

we may further assume that

$$T(\alpha) = T(\beta) = T(\alpha\beta) = 0.$$

Using these we see that $\alpha\beta = -\beta\alpha$ and $\beta^2 < 0$. Thus $\mathcal{K}' = \mathbf{Q}[\alpha, \beta] = \mathbf{Q} + \mathbf{Q}\alpha + \mathbf{Q}\beta + \mathbf{Q}\alpha\beta$ is a quaternion algebra.

We want to show that $\mathcal{K} = \mathcal{K}'$. If not take $\gamma \in \mathcal{K}$ which is not in $\mathcal{K}'$. Replacing γ by

$$\gamma - \frac{1}{2}(T(\gamma) + (T(\alpha\gamma)/\alpha^2)\alpha + (T(\beta\gamma)/\alpha^2)\beta$$

we may assume as before that

$$T(\gamma) = T(\alpha\gamma) = T(\beta\gamma) = 0.$$

Thus we get

$$\alpha\gamma = -\gamma\alpha \quad \beta\gamma = -\gamma\beta.$$

Using all these relations it is easy to see that $\beta\gamma$ commutes with α. But the algebra generated by $\beta\gamma$ and α is then a field (by the positive definiteness, every non-zero element has an inverse and easy to see that the relevant inverses belong to the subalgebra). If $\beta\gamma$ belonged to $\mathbf{Q}(\alpha)$, then $\gamma \in \mathbf{Q}[\alpha, \beta]$

which is contrary to our hypothesis. So we get a field extension of degree 4 in $\mathcal{K}$ and thus an element which satisfies a degree 4 equation over $\mathbf{Q}$ and no smaller degree. This is clearly a contradiction.

Finally we have to show that $\mathcal{R}$ is an order *i.e.*, we must show that it is finitely generated as a $\mathbf{Z}$-module. Pick $a_1, \ldots, a_n \in \mathcal{R}$ a $\mathbf{Q}$-basis for $\mathcal{K}$. Consider the map

$$\mathcal{K} \to \mathbf{Q}^n \quad x \mapsto (T(xa_1), \ldots, T(xa_n))$$

One easily sees that it is an isomorphism of $\mathbf{Q}$-vector spaces and it takes $\mathcal{R}$ into $\mathbf{Z}^n$. Thus $\mathcal{R}$ is finitely generated. □

Part III

Arithmetic of Elliptic Curves

Chapter 14

Rational points on curves

14.1 Curves of degree ≤ 2

A *Diophantine equation* is a polynomial equation

$$f(X_1, \ldots, X_n) = 0$$

for which we seek solutions $(x_1, \ldots, x_n)$ where the x_i are integers, or rational numbers. Usually, f is a polynomial with rational or integer coefficients. One may also consider integer or rational solutions of systems of equations, etc.

The term 'Diophantine equation' refers to Diophantus of Alexandria, in the 3rd century AD, who systematically studied solutions of algebraic equations in rational numbers or integers. This followed earlier work by the Pythagoreans (who, for example, were interested in positive integer solutions of $x^2 + y^2 = z^2$), and Euclid. The work of Diophantus and his school was published in 13 books, six of which have been preserved. An account of Diophantus' work may be found in [30], [29]. Ancient Indian mathematicians were also interested in problems of finding integer solutions; for an account of their contributions to such problems, see [13], [20], [11], [53]. The results on linear equations, and some cases of the results on quadratic equations discussed below, were known to these mathematicians.

The 'modern' study of Diophantine equations began with the work of the French mathematician Fermat (1601-1665), who first systematically studied equations of higher degree, and developed his famous *method of descent*, which will play a central role later in the lectures on the so called *Mordell-Weil group*. For further historical information, the reader may consult [16], [61].

14.1.1 Linear equations

An important theme in the lectures on the number theory of elliptic curves is the study of the set of rational and integer solutions of Diophantine equations

$$f(X, Y) = 0$$

where $f(X, Y)$ is a polynomial of degree 3 with rational number coefficients. But before we begin this study, it is appropriate to discuss the cases of curves of degrees 1 and 2.

In geometric language, a plane algebraic curve, or just a *curve*, is a set

$$C = \{(x, y) \in \mathbf{C}^2 \mid f(x, y) = 0\},$$

where $f(X, Y) \in \mathbf{C}[X, Y]$ is a non-zero polynomial. For number theory, it is of primary interest to consider the curves C defined by polynomials f with rational coefficients, $f(X, Y) \in \mathbf{Q}[X, Y]$. Such a curve C is said to be *defined over* $\mathbf{Q}$. In this case, we may consider the set of points $P \in C$ with rational coordinates. Such a point P is called a *rational point* of C. The set of rational points of C is denoted by $C(\mathbf{Q})$. Thus, our Diophantine problem of finding the rational solutions to the polynomial equation $f(X, Y) = 0$ is equivalent to finding all rational points of the curve C defined by f.

A measure of the 'complexity' of a curve is the degree of the defining polynomial $f(X, Y)$; we call this the *degree* of the curve. From this point of view, the 'simplest' curves are those defined by equations of degree 1, which are just lines. Let

$$f(X, Y) = aX + bY + c = 0, \quad a, b, c \in \mathbf{Q},$$

be the equation of a line which is defined over $\mathbf{Q}$. If $a \neq 0$, the equation $f(X, Y) = 0$ may be solved for X in terms of Y,

$$X = \frac{-bY - c}{a},$$

so that a complete set of rational points is found by giving arbitrary rational values to Y, and using the above formula to determine the corresponding value of X, which is also rational. A similar argument works if $b \neq 0$, by solving for Y in terms of X. If a, b both vanish, the 'curve' is the empty set, unless $c = 0$, in which case f is the zero polynomial. So the situation for curves of degree 1 is quite trivial.

14.1.2 Equations of degree two

Before making a general study of curves of degree 2, we consider the special case

$$f(X,Y) = X^2 + Y^2 - c = 0, \quad c \in \mathbf{Q}.$$

If $c < 0$, there are no solutions in real numbers, hence no rational solutions. Similarly, if $c = 0$, the only real solution is $(0,0)$, which is a rational point. So the interesting case is when $c > 0$. Using unique prime factorisation, write

$$c = c_0^2 c_1,$$

where $c_0 \in \mathbf{Q}$, and c_1 is a positive integer which is *square free* (not divisible by any square > 1). Then

$$\begin{aligned} & X^2 + Y^2 - c = 0 \\ \Leftrightarrow \quad & \left(\frac{X}{c_0}\right)^2 + \left(\frac{Y}{c_0}\right)^2 = c_1. \end{aligned}$$

Hence we are reduced to considering the case when c is a square free, positive integer. Note that the real points of C form a circle of radius $\sqrt{c}$.

Suppose now that the curve C defined by $f(X,Y) = 0$ has a rational point $P = (a,b)$. Then there is a simple procedure to determine the set $C(\mathbf{Q})$ of all rational points of C, as follows. Let $R \in C(\mathbf{Q})$ be another rational point. Then the line $\overline{PR}$ through P and R is either vertical (*i.e.*, P and R have the same X coordinate, so $R = (a,-b)$), or $\overline{PR}$ has rational slope, say m, and hence is given by

$$(Y - b) - m(X - a) = 0,$$

i.e., is defined over $\mathbf{Q}$. Clearly the only points of intersection of $\overline{PR}$ with C are P and R. Conversely, consider any line L through P with arbitrary rational slope m, defined by

$$(Y - b) - m(X - a) = 0.$$

We compute the points of intersection of L with C, by substituting for Y; we get

$$X^2 + (b + m(X - a))^2 = c = a^2 + b^2$$

which we may rewrite as

$$(X^2 - a^2) + 2bm(X - a) + m^2(X - a)^2 = 0$$

$$i.e., \quad (X-a)(X+a+m^2(X-a)+2bm)=0.$$

The second solution is

$$X=\frac{am^2-2bm-a}{1+m^2}, \quad Y=\frac{-bm^2-2am+b}{1+m^2}.$$

Thus, we have obtained another rational point, unless the line L is tangent to the circle C.

Example 14.1.1: Let C be the unit circle $X^2+Y^2=1$. Then C has the rational point $P=(1,0)$, where the vertical line $X=1$ is the tangent to C at P. Thus all other rational points are of the form

$$X=\frac{m^2-1}{m^2+1}, \quad Y=\frac{2m}{m^2+1},$$

from the above general formulas (we have replaced m by $-m$). If $m=p/q$ in lowest terms, then we may rewrite this as

$$X=\frac{p^2-q^2}{p^2+q^2}, \quad Y=\frac{2pq}{p^2+q^2}.$$

The fractions for X, Y are in lowest terms unless p, q are both odd. But in this case, let $r=(p+q)/2$, $s=(p-q)/2$. Then r, s are in lowest terms, and $r^2+s^2=(p^2+q^2)/2$, so that

$$X=\frac{2rs}{r^2+s^2}, \quad Y=\frac{r^2-s^2}{r^2+s^2}.$$

If r, s are again both odd, we may repeat this process. Hence, we see that any solution of $X^2+Y^2=Z^2$, where X,Y,Z are relatively prime, is of the form $X=p^2-q^2$, $Y=2pq$, $Z=p^2+q^2$ where p,q are relatively prime non-zero integers of opposite parity, or is of a similar form with X and Y interchanged.

The principle used above is valid for any curve C of degree 2, which is defined over $\mathbf{Q}$, and whose set of real points is a conic. If C has a rational point P, then it has infinitely many, and any other rational point (with 1 possible exception) is determined as the second point of intersection of C with a line of rational slope. Thus, we obtain a parametric representation of

(almost) all rational points of C. The underlying idea is that if a line and a conic are defined over $\mathbf{Q}$, and one point in their intersection is a rational point, then so is the other. This is because if a quadratic equation with coefficients in $\mathbf{Q}$ has 1 rational solution, its second solution is also rational.

Incidentally, for cubic curves, we may apply the following variation of this argument: given 2 rational points P_1, P_2, the line joining them in general meets the cubic curve in a 3^{rd} rational point. A second variation: given 1 rational point P, at which the cubic has a tangent line L, the tangent meets C in 1 other rational point, unless P is an inflexion point. These constructions, called the *chord and tangent constructions*, will be the starting point of our study of rational points on cubics.

Returning to the circle C above, the remaining problem is to determine when there exists at least 1 rational point. A similar problem remains for the general conic.

If C is now any conic defined over $\mathbf{Q}$, then it is defined by an equation

$$aX^2 + bY^2 + 2hXY + 2gX + 2fY + c = 0, \quad a, b, c, f, g, h \in \mathbf{Q}.$$

If $aX^2 + 2hXY + bY^2$ is the product of 2 linear polynomails with rational coefficients, then after a linear change of variables (with rational coefficients), we are reduced to considering either the rectangular hyperbola $XY = 1$, the parabola $Y^2 = X$ or a curve $Y^2 = \alpha$ for some rational α. In these case, it is trivial to decide if there are any rational points (they always exist for the hyperbola and parabola, and they exist for the curve $Y^2 = \alpha$ precisely when a is a square).

Hence we may assume $aX^2 + 2hXY + bY^2$ is irreducible over $\mathbf{Q}$. In particular, we may assume at a, b are both non-zero. Completing the square then gives an equation

$$a(X - \alpha_1 Y - \alpha_2)^2 + b'(Y - \beta)^2 + c' = 0,$$

where $b' \neq 0$ (if $b' = 0$, one sees that $aX^2 + 2hXY + bY^2 = a(X - \alpha_1 Y)^2$). Hence it suffices to consider conics

$$aX^2 + bY^2 + c = 0.$$

We may assume further that a, b, c are square free integers, which have no common factor. If they are all of the same sign, or one of them is 0, the

'conic' is empty, or a single point, or a union of lines. So the interesting case is when exactly two of a, b, c have the same sign, which we may take to be positive. If $a, c > 0$ and $b < 0$, make the substitution $X' = X/Y$, $Y' = 1/Y$; hence we may assume that $a, b > 0$ and $c < 0$.

Now finding rational solutions to the above equation is equivalent to finding integer solutions to

$$aX^2 + bY^2 + cZ^2 = 0.$$

If a prime p divides a and b, then it does not divide c, so it must divide Z; hence we are reduced to considering

$$aX^2 + bY^2 + cp^2(Z')^2 = 0,$$

$$i.e., \quad (a/p)X^2 + (b/p)Y^2 + c(Z')^2 = 0.$$

Hence we may assume a, b are relatively prime; similarly we may assume b, c and c, a are relatively prime.

Theorem 14.1.1 (Legendre) *Let a, b, c be non-zero square free integers which are pairwise relatively prime, and not all of the same sign. Then the equation*

$$aX^2 + bY^2 + cZ^2 = 0$$

has an integer solution $(X, Y, Z) \neq (0, 0, 0)$ if and only if the congruences

$$\begin{aligned} u^2 &\equiv -ab \pmod{c}, \\ v^2 &\equiv -bc \pmod{a}, \\ w^2 &\equiv -ca \pmod{b} \end{aligned}$$

all have solutions.

Proof: The Theorem is trivial if $|abc| = 1$. Hence we may assume at least one of a, b, c is divisible by some prime.

First we show that the condition is necessary. Suppose the equation has a solution $(x, y, z) \neq (0, 0, 0)$. If $z \neq 0$, the conic $aX^2 + bY^2 + c = 0$ has a rational point, hence has infinitely many; hence it has a rational point both of whose coordinates are non-zero. Thus there is a solution where x, y, z are all non-zero integers. A similar argument works if the first solution has $x \neq 0$

or $y \neq 0$. We may further assume x, y, z are relatively prime. If p is a prime dividing a, we claim p doesn't divide y or z; indeed, if p divides y, then it divides cz^2, hence divides z; then p^2 divides ax^2, and since a is square-free, p divides x, a contradiction. Hence we may assume: x is relatively prime to b and c, y is relatively prime to a and c, and z is relatively prime to a and b.

Now $ax^2 + by^2 \equiv 0 (\bmod c)$, where a, b, x, y are relatively prime to c; hence

$$\begin{aligned} \left(\frac{ax}{y}\right)^2 &\equiv -ab \pmod{c} \\ i.e.,\ ,\ u^2 &\equiv -ab \pmod{c} \end{aligned}$$

has a solution. Similarly the other two congruences modulo a and b have solutions.

Conversely, suppose the congruences have solutions. If

$$u^2 \equiv -ab \pmod{c},$$

the quadratic form $aX^2 + bY^2$ over the ring $\mathbf{Z}/c\mathbf{Z}$ has discriminant $-4ab$ which is a square, and a is a unit modulo c; hence $aX^2 + bY^2 \equiv aX^2 + bY^2 + cZ^2$ factorizes as a product of 2 linear factors over $\mathbf{Z}/c\mathbf{Z}$ *i.e.*, there exist linear homogeneous polynomials

$$\begin{aligned} L_c(X, Y, X) &= \alpha(c)X + \beta(c)Y + \gamma(c)Z, \\ M_c(X, Y, Z) &= \alpha'(c)X + \beta'(c)Y + \gamma'(c)Z \end{aligned}$$

with integer coefficients, such that

$$aX^2 + bY^2 + cZ^2 \equiv L_c(X, Y, Z)M_c(X, Y, Z) \pmod{c}.$$

Similarly there exist homogeneous linear polynomials with integer coefficients

$$\begin{aligned} L_a &= \alpha(a)X + \beta(a)Y + \gamma(a)Z, \\ M_a &= \alpha'(a)X + \beta'(a)Y + \gamma'(a)Z, \\ L_b &= \alpha(b)X + \beta(b)Y + \gamma(b)Z, \\ M_b &= \alpha'(b)X + \beta'(b)Y + \gamma'(b)Z, \end{aligned}$$

such that

$$\begin{aligned} aX^2 + bY^2 + cZ^2 &\equiv L_aM_a \pmod{a} \\ aX^2 + bY^2 + cZ^2 &\equiv L_bM_b \pmod{b}. \end{aligned}$$

Since a, b, c are pairwise relatively prime, the Chinese Remainder theorem implies that we can solve the systems of congruences

$$\begin{array}{rclrclrcl} \alpha & \equiv & \alpha(a) \pmod{a}, & \alpha & \equiv & \alpha(b) \pmod{b}, & \alpha & \equiv & \alpha(c) \pmod{c}, \\ \beta & \equiv & \beta(a) \pmod{a}, & \beta & \equiv & \beta(b) \pmod{b}, & \beta & \equiv & \beta(c) \pmod{c}, \\ \gamma & \equiv & \gamma(a) \pmod{a}, & \gamma & \equiv & \gamma(b) \pmod{b}, & \gamma & \equiv & \gamma(c) \pmod{c}, \end{array}$$

Then $L(X, Y, Z) = \alpha X + \beta Y + \gamma Z$ is a linear form with integer coefficients which satisfies the congruences

$$L \equiv L_a \pmod{a} \;\; L \equiv L_b \pmod{b}, \;\; L \equiv L_c \pmod{c}.$$

Similarly, we can find a linear form with integer coefficients $M(X, Y, Z)$ such that

$$M \equiv M_a \pmod{a}, \;\; M \equiv M_b \pmod{b}, \;\; M \equiv M_c \pmod{c}.$$

Then

$$aX^2 + bY^2 + cZ^2 \equiv LM \pmod{abc},$$

since a, b, c are pairwise relatively prime.

Now consider all triples (x, y, z) of integers with

$$\leq x < \sqrt{|bc|}, \;\; 0 \leq y < \sqrt{|ca|}, \;\; 0 \leq z < \sqrt{|ab|}.$$

Since $|abc| \neq 1$, and a, b, c are pairwise relatively prime and square-free, at least 2 of the square roots are irrational. Hence the number of such triples is strictly greater than $\sqrt{|ab|}\sqrt{|bc|}\sqrt{|ca|} = |abc|$, which is the cardinality of $\mathbf{Z}/abc\mathbf{Z}$. Hence there exist a pair of distinct triples $(x_1, y_1, z_1) \neq (x_2, y_2, z_2)$ such that

$$L(x_1, y_1, z_1) \equiv L(x_2, y_2, z_2) \pmod{abc}.$$

Thus if $x_0 = x_1 - x_2$, $y_0 = y_1 - y_2$, $z_0 = z_1 - z_2$ then $(x_0, y_0, z_0) \neq (0, 0, 0)$ and

$$ax_0^2 + by_0^2 + cz_0^2 \equiv 0 \pmod{abc}.$$

But $|x_0| < \sqrt{|bc|}$, $|y_0| < \sqrt{|ca|}$, $|z_0| < \sqrt{|ab|}$. Further, the signs of a, b, c may be chosen so that exactly 2 are positive. Then

$$-|abc| < ax_0^2 + by_0^2 + cz_0^2 < 2|abc|.$$

Since $ax_0^2 + by_0^2 + cz_0^2$ is divisible by abc, we must have

$$ax_0^2 + by_0^2 + cz_0^2 = 0 \text{ or } ax_0^2 + by_0^2 + cz_0^2 = |abc| = -abc.$$

In the former case, we have found a solution. In the latter case, we compute that

$$a(x_0z_0 + by_0)^2 + b(y_0z_0 - ax_0)^2 + c(z_0^2 + ab)^2 = 0.$$

This is a non-trivial solution of $aX^2 + bY^2 + cZ^2 = 0$, since $z_0^2 + ab > 0$, as $a, b > 0$. □

For further results on quadratic forms over the rationals, see the books of Serre [50] and Borevich and Shafarevich [8].

As stated earlier, the study of cubic equations is essentially the study of elliptic curves. I will briefly mention the most important result about rational points on higher degree curves — the *Mordell conjecture*, proved in 1984 by Gerd Faltings. For a plane curve C defined by $f(X, Y) = 0$, where f has degree d, we can associate the homogeneous equation $F(X, Y, Z)$ of degree d, uniquely defined by the property that $F(X, Y, 1) = f(X, Y)$. In the algebraic geometry lectures, it will be shown that this defines a curve $\overline{C}$ in the *projective plane* $\mathbf{P}^2$, which contains the original curve C as an open subset. The Mordell conjecture implies that if $\overline{C}$ is non-singular of degree ≥ 4, then $\overline{C}$ (and hence C) has only a finite number of rational points. The condition that $\overline{C}$ is non-singular is equivalent to the algebraic condition that only common zero of the partial derivatives

$$\frac{\partial F}{\partial X}, \quad \frac{\partial F}{\partial Y}, \quad \frac{\partial F}{\partial Z}$$

is $(0, 0, 0)$. This holds for the Fermat curve of degree ≥ 4, defined by $f(X, Y) = X^n + Y^n - 1$, or equivalently by $F(X, Y, Z) = X^n + Y^n - Z^n$ in the projective plane. Thus Faltings' result implies that the Fermat curve has only a finite number of rational points, for each value of n. This is still a far cry from Fermat's Last Theorem, which is (in our geometric language) the statement that the only rational points on the Fermat curve of degree ≥ 3 are those where one of the coordinates vanishes; however, Faltings' result is the best result known which is valid for all the Fermat curves. For a discussion of Falting's results, along with the needed background material and related results, see [12].

Here are some exercises for the reader.

Exercise 14.1.1: Find all rational numbers which can be expressed as a sum of 2 squares.
(Hint: Apply Legendre's theorem to determine whether the circle of radius a has rational points, for any given $a \in \mathbf{Q}$).

Exercise 14.1.2: Let C be the curve defined by $X^2-2Y^2-2X-4Y-1=0$.

1. Show that C has infinitely many real points (points with real coordinates).

2. Determine all rational points on C.

14.2 Problems of higher degree

In the previous section we have seen how to solve linear and quadratic 'Diophantine problems'. We now go on to study equations of higher degree. The most famous *unsolved* 'Diophantine problem' is called "Fermat's Last Theorem".

Conjecture 1 (Fermat's Last Theorem) *The only solutions over rational numbers of the equation*

$$X^n + Y^n = Z^n$$

where $n \geq 3$ *are on the locus* $XYZ = 0$.

Fermat claimed that he had a proof of this but no one has been able to discover this or any other proof.

We immediately recognise the above equation as defining a projective plane curve. An easy computation shows that the curve is smooth. The problem then is that of finding rational points on a smooth plane curve. This point of view has been one of the most fruitful in dealing with Diophantine problems.

We will now see how two Diophantine problems considered by Fermat can be solved using the techniques of algebraic geometry. The method used is derived from that used by Fermat.

14.2.1 The Fermat quartic

We first consider the special case of Conjecture 1 where $n = 4$. The assertion which Fermat proved is

Proposition 14.2.1 *The equations $X^4 + Y^4 = Z^4$ has no solutions in integers other than those with $XY = 0$.*

We will prove the result is several steps. The first step is to note that it is enough to prove that the equation

$$X^4 + Y^4 = Z^2 \tag{14.1}$$

has no integer solutions other than those with $XY = 0$. Using the variables $U = X^2$, $V = Y^2$ and $W = XY$, we can re-write the equation (14.1) in the variables (U, V, W, Z):

$$UV = W^2 \tag{14.2}$$

$$U^2 + V^2 = Z^2 \tag{14.3}$$

Note that solutions to these two equations do not neccesarily give solutions to the original problem. But it is clearly enough to show that the only solutions to (14.2) and (14.3) are those with $W = 0$.

The equations (14.2) and (14.3) are homogeneous quadratic equations in the variables (U, V, W, Z). As we have seen in (12.4.7) the simultaneous solutions of the two equations (14.2) and (14.3) describe a *curve of genus 1*. Moreover, we have the "obvious" solution $(U, V, W, Z) = (1, 0, 0, 1)$, and using the methods of (12.4.7) we see that this curve is isomorphic to an elliptic curve E which has the equation

$$2w^2 = v^3 - v \tag{14.4}$$

In section (12.3.1) we have seen that the points of E form a group. In particular, we can multiply by 2 and hence obtain more points given one point. Since we have the point $(v, w) = (0, 0)$ which is of order two in the group we can do this doubling in two stages.

First we take the "quotient" by the group generated by this point. This can be calculated (11.3.2) to be the morphism

$$(U, V, W, Z) \mapsto (A, B, C, D) = (U^2 - V^2, 2UV, Z^2, 2WZ).$$

Here (A, B, C, D) satisfy the equations

$$A^2 = C^2 - B^2 \tag{14.5}$$
$$D^2 = 2BC \tag{14.6}$$

This also gives us a curve defined by 2 quadratic equations in four variables which has a natural point $(A, B, C, D) = (1, 0, 1, 0)$. Hence, this is an elliptic curve (12.4.7) E' which has the equation

$$b^2 = d(d^2 - 3) \tag{14.7}$$

This has a point $(b, d) = (0, 0)$ which is of order 2. As before we take the "quotient" by the group generated by this point. This is the morphism

$$(A, B, C, D) \mapsto (U, V, W, Z) = (A^2, D^2, AD, (B^2 + C^2)).$$

The resulting (U, V, W, Z) satisfy (14.2) and (14.3) as we expect, since the composite of the two quotients is the multiplication by 2 on the elliptic curve E.

Thus we have an explicit procedure for obtaining another point. Since some squaring process is involved in the construction we can expect the coordinates of the new point to be larger than those of the original point. Fermat's idea (now called "Fermat's method of infinite descent") was to show that this doubling can be *reversed* in this case. Thus we can get points with smaller and smaller coordinates. Since the coordinates are to be integers this is clearly impossible. The same idea will be used in the proof of the Mordell-Weil theorem (15.1.1) for elliptic curves.

In order to reverse the procedure we first need to obtain (A, B, C, D) such that $(U, V, W, Z) = (A^2, D^2, AD, (B^2+C^2))$ when $W \neq 0$. Since any solution can be multiplied by a non-zero scalar, we may assume that (U, V, W, Z) are positive, coprime integers. The equation (14.2) then says that U and V are squares, hence we may write $U = A^2$ and $V = D^2$. We may also assume that U is odd and V is even, since they are coprime. The theorem on pythagorean triples can be applied to (14.3) to give us integers B and C such that

$$U = C^2 - B^2$$
$$V = 2BC$$
$$Z = C^2 + B^2$$

Clearly, (A, B, C, D) satisfy the equations (14.5) and (14.6). This gives the first step. To obtain the second step, we need to find (U_1, V_1, W_1, Z_1) such that

$$(A, B, C, D) = (U_1^2 - V_1^2, 2U_1V_1, Z_1^2, 2W_1Z_1).$$

As before we can assume that (A, B, C, D) are positive, coprime integers. Moreover, since U is odd, we see that A is odd and so B is even. By the theorem on pythagorean triples applied to (14.5) we have U_1 and V_1 such that

$$\begin{aligned} A &= U_1^2 - V_1^2 \\ B &= 2U_1V_1 \end{aligned}$$

By using (14.6) and the fact that B is even, we see that there are integers W_1 and Z_1 so that $B = 2W_1^2$, $C = Z_1^2$ and $D = 2W_1Z_1$. This gives us the required 'undoubled" point corresponding to (U, V, W, Z). Moreover we see by the steps outlined above that $U_1 < U$. This gives the proof by contradiction, since positive integers cannot descend indefinitely.

14.2.2 Frenicle's problem

While the possibility of "undoubling" points is restricted to some special elliptic curves, we can use the doubling technique to construct new solutions in a great many cases. One such which Fermat solved by this technique is

Frenicle's problem Find a right angled triangle with integer sides such that the sum of the legs $(= x + y)$ is a square and the hypotenuse $(= z)$ is also a square.

Written algebraically (as a matter of historical interest Fermat *did not* have algebraic notation available), the equations are

$$\begin{aligned} z^2 &= x^2 + y^2 \\ x + y &= a^2 \\ z &= b^2 \end{aligned}$$

These equations have an "obvious" solution with $x = 1 = z = a = b$ and $y = 0$. However, it is clear that in Frenicle's problem we are asking for

solutions in *positive integers*. We find such solutions by relating this problem to elliptic curves and "doubling" to obtain such a solution. When we study *heights* it will be clear why this procedure works.

Introducing the new variables $w = ab$ we can re-write the above equation in the form

$$z^2 = x^2 + y^2 \tag{14.8}$$
$$w^2 = (x+y)z \tag{14.9}$$

This is a pair of homogeneous quadratic equations in the four variables (w, x, y, z) and so we can apply (12.4.7) to write it as an equation in Weierstrass normal form. Since this computation is a bit complicated we do it here. First of all we have the "obvious" solution $(1, 1, 0, 1)$ to (14.8) and (14.9) as before. The tangent planes to (14.8) and (14.9) are given by $u = x - z = 0$ and $v = x + y + z - 2w = 0$ respectively. We re-write these equations in the coordinates (u, v, y, z) to obtain

$$2uz = -u^2 - y^2 \tag{14.10}$$
$$4vz = (u + y - v)^2 \tag{14.11}$$

Now the "obvious" point has become $(u, v, y, z) = (0, 0, 0, 1)$. Thus by the procedure described in (12.4.7) the simultaneous solutions of (14.10) and (14.11) are in 1-1 correspondence with the solutions of the cubic equation

$$2v(-u^2 - y^2) - u(u + y - v)^2 = 0 \tag{14.12}$$

Now we follow the method of proof of (12.4.4) to convert this to the equation of an elliptic curve in Weierstrass form. We dehomogenise with respect to u and put $s = v/u$ and $t = y/u$. The above equation becomes a quadratic equation for t with coefficients as polynomials in s,

$$(-2s - 1)t^2 + 2(s - 1)t - ((s - 1)^2 - 2s) = 0.$$

Given s this has a solution in t if and only if we have a square root r of the discrimant (which is a polynomial in s). In other words, we must solve

$$r^2 = -2s(s^2 + 2) \tag{14.13}$$

This is an equation in Weierstrass normal form. Note that the transformation between (r, s) and (t, s) is given by $(-2s - 1)t = r - (s - 1)$.

The equations (14.8) and (14.9) also have the solution $(w, x, y, z) = (-1, 0, 1, 1)$. Under the above tranformation we see that $u = -1$ and $v = 4$. Thus $s = -4$ and $r = 12$ is a solution of (14.13). Moreover, from (12.3.1) we see that it is not a 2 torsion point in the group law on this elliptic curve (since $r \neq 0$). Thus we can double this point. As in (12.3.1) this is done by finding the intersection of the curve (14.13) with tangent line at this point which is given by $6(r-12) + 25(s+4) = 0$. The intersection point can be found by writing the equation (14.13) in the "variables" $(r-12)$ and $(s+4)$. Then we can simultaneously solve (14.13) and the tangent equation, to obtain the new solution $(r, s) = (791/432, -49/72)$. We do the above computation in reverse to obtain x, y and z corrsponding to this point. We obtain $b = 1343$ and $u = 2750257$. But then $b^2 = 1803649 < u$ so we need to grind out yet another solution by the above procedure. But we can adopt an easier method as follows.

Another way to obtain a solution is to "add" the two solutions $(12, -4)$ and $(791/432, -49/72)$ in the group law on the curve (see 12.3.1). This is done by finding the third point of intersection of the line joining these two point with the curve. We can do this computation and obtain Fermat's solutions

$$\begin{aligned} a &= 2165017 \\ b &= 2372159 \\ b^2 &= 5627138321281 \\ u &= 3503833734241 \\ x &= 4565486027761 \\ y &= 1061652293520 \\ z &= 4687298610289 \end{aligned}$$

In fact it turns out that this is the smallest solution, but the proof of this a bit too elaborate to go into here.

Exercise 14.2.1: Let $n > 1$ be a natural number. We say n *is a congruent number* if there is a right angled triangle with rational sides x, y, z and of area equal to n.

1. Show that n is congruent iff there is a rational number u such that the numbers $u, u+n, u-n$ are all squares of rational numbers.

2. Thus it follows that a congruent number gives rise to a rational solution of $V^2 = U^3 - n^2U$. When does a solution of this elliptic curve determine a solution to the congruent number problem?

3. Show that 1 is not a congruent number, determine first few congruent numbers.

Chapter 15

The Mordell-Weil theorem for elliptic curves over Q

15.1 The Mordell-Weil Theorem

Let $E/\mathbf{Q}$ be an elliptic curve defined by an equation:

$$E: Y^2 = X^3 + aX^2 + bX + c.$$

We have seen in Chapter (12) that the set of points $E(\mathbf{Q})$ is a group under with the element $O = (0 : 1 : 0)$ as the identity element, and the group operation being defined by the chord and the tangent construction. In this section we will analyse the group $E(\mathbf{Q})$ in detail. We begin with the following:

Theorem 15.1.1 (Mordell-Weil) *Let $E/\mathbf{Q}$ be an elliptic curve. Then the group of rational points, $E(\mathbf{Q})$, is a finitely generated abelian group.*

Remark: This was proved by Mordell in [40]. Subsequently, Weil [59] proved that the theorem is also valid for elliptic curves[1] defined over any number field (*i.e.*, a finite extension of $\mathbf{Q}$).

To simplify matters we will assume that the elliptic curve E has the following form, where $\alpha, \beta, \gamma \in \mathbf{Z}$.

$$E: \quad Y^2 = (X - \alpha)(X - \beta)(X - \gamma).$$

[1]He also proved that it is valid for Abelian varieties defined over number fields.

The proof in the general case is similar. The difficulties are technical, the general case is treated in [41]. We will begin by proving several results which are of independent interest.

Recall that the discriminant Δ of the curve E is defined as

$$\Delta = (\alpha - \beta)^2(\beta - \gamma)^2(\alpha - \gamma)^2.$$

Proposition 15.1.1 *Let $P = (X, Y)$ the map from $f_\alpha : E(\mathbf{Q}) \to \mathbf{Q}^*/\mathbf{Q}^{*2}$ given by*

$$P = (X, Y) \longmapsto (X - \alpha) \pmod{\mathbf{Q}^{*2}} P = (\alpha, 0) \longmapsto (\alpha - \beta)(\alpha - \gamma) \pmod{\mathbf{Q}^{*2}} O = (\infty) \longmapsto 1 \tag{15.1}$$

is a homomorphism of groups.

Remark: We can similarly define maps f_β, f_γ. We leave it as an exercise to check that these maps satisfy the condition

$$f_\alpha = f_\beta f_\gamma.$$

Proof: We have to show that if

$$P_1 + P_2 + P_3 = O$$

then

$$f_\alpha(P_1) f_\alpha(P_2) f_\alpha(P_3) = 1 \pmod{\mathbf{Q}^{*2}}.$$

The proof breaks up in to several cases: (1) $P_i \neq O, (\alpha, 0)$ for $i = 1, 2, 3$, (2) $P_1 = O$, (3) $P_1 = (\alpha, 0)$. We will prove the case (1). The rest are similar and left as exercises. So suppose that $P_i \neq O, (\alpha, 0)$ for $1 \leq i \leq 3$. Then we have to show that:

$$(X_1 - \alpha)(X_2 - \alpha)(X_3 - \alpha) = \text{perfect square}.$$

Now by the group law $P_1 + P_2 + P_3 = O$ if and only if P_1, P_2, P_3 are collinear. We may assume that $P_i = (X_i, Y_i) \quad (1 \leq i \leq 3)$ lie on a straight line

$$Y = \lambda X + \mu, \qquad \lambda, \mu \in \mathbf{Q}.$$

Each P_i lies on the curve and also on the line $Y = \lambda X + \mu$; hence the cubic polynomial

$$(X - \alpha)(X - \beta)(X - \gamma) - (\lambda X + \mu)^2$$

has three roots X_1, X_2, X_3. Then, since this polynomial is monic we have

$$(X-\alpha)(X-\beta)(X-\gamma) - (\lambda X+\mu)^2 = (X-X_1)(X-X_2)(X-X_3).$$

Then putting $X = \alpha$ we get

$$-(\lambda\alpha+\mu)^2 = (\alpha-X_1)(\alpha-X_2)(\alpha-X_3)$$

and so the proposition is proved. □

The following lemma is very useful in practice and will be used in the proof of the next proposition.

Lemma 15.1.2 *Suppose P, P' are two distinct points in $E(\mathbf{Q})$. Then the following statements are equivalent:*

1. $P - P' \equiv O(\mathrm{mod} 2E(\mathbf{Q}))$.
2. *The following hold:*

$$(X'-\alpha) \equiv (X-\alpha) \pmod{\mathbf{Q}^{*2}} \quad (X'-\beta) \equiv (X-\beta) \pmod{\mathbf{Q}^{*2}} \quad (X'-\gamma) \equiv (X-\gamma) \pmod{\mathbf{Q}^{*2}} \tag{15.2}$$

Proof: Let $P-P' = Q$, then it suffices to check that $f_\alpha(2Q) = 1, f_\beta(2Q) = 1$. A computation using the "duplication formula" (see Proposition (12.3.3)) shows that if $Q = (X, Y)$, $2Q = (X', Y')$ then the duplication formula yields

$$X' - \alpha = \frac{(X^2 - 2\alpha X + \beta\gamma - \alpha(\beta+\gamma))^2}{4Y^2},$$

and similar formulae for $X'-\beta, X'-\gamma$, and which are all squares as claimed. □

Using α, β, we get two maps $E(\mathbf{Q})\mathbf{Q}^*/\mathbf{Q}^{*2}$. The product of these maps gives us a map $E(\mathbf{Q})\mathbf{Q}^*/\mathbf{Q}^{*2} \times \mathbf{Q}^*/\mathbf{Q}^{*2}$. Using Lemma (15.1.2), we can identify the kernel of this product map:

Proposition 15.1.3 *Let $E/\mathbf{Q}: \quad Y^2 = (X-\alpha)(X-\beta)(X-\gamma)$. Then the kernel of the map*

$$f_\alpha \times f_\beta : E(\mathbf{Q}) \longrightarrow \mathbf{Q}^*/\mathbf{Q}^{*2} \times \mathbf{Q}^*/\mathbf{Q}^{*2} \tag{15.3}$$

(which is explicitly given by)

$$O = (\infty) \longmapsto (1,1) \pmod{\mathbf{Q}^{*2}} P = (X,Y) \longmapsto (X-\alpha, X-\beta) \pmod{\mathbf{Q}^{*2}} P = (\alpha,0) \longmapsto (\alpha - \tag{15.4}$$

is equal to $2E(\mathbf{Q})$.

Remark: Thus we get an injective homomorphism:

$$E(\mathbf{Q})/2E(\mathbf{Q}) \hookrightarrow \mathbf{Q}^*/\mathbf{Q}^{*2} \times \mathbf{Q}^*/\mathbf{Q}^{*2}.$$

Proof: Follows immediately from the lemma, on noting that for any point $(x_0, y_0) \in E(\mathbf{Q})$, if any two of $x_0 - \alpha, x_0 - \beta, x_0 - \gamma$ are squares then so is the third. □

Notation 1 *If x is a rational number and p is a prime then writing $x = p^n a/b$ where $n \in \mathbf{Z}, (a,p) = (b,p) = 1$, we define*[2] $\|0|_p = 0, \|x|_p = p^{-n}$, $\mathrm{ord}_p(0) = -\infty$, $\mathrm{ord}_p(x) = n$, *and we also use the notation $p^n \parallel x$. Thus p^n is the highest power of p in x. We set $\|0|_p = 0$. Note that one has the easy inequality: $\|x+y|_p \leq \max(\|x|_p, \|y|_p)$ valid for all rational x, y and all primes p. For $x \neq 0$, we define $\mathrm{ord}_\infty(x) = x/\|x|$. Thus $\mathrm{ord}_\infty(x)$ is simply the sign of x.*

It is then easy to see that we have an identification of abelian groups

$$\mathbf{Q}^*/\mathbf{Q}^{*2} \times \mathbf{Q}^*/\mathbf{Q}^{*2} = \prod_{\infty,\, p} (\mathbf{Z}/(2) \times \mathbf{Z}/(2)),$$

where on the right hand product there is one factor $\mathbf{Z}/2 \times \mathbf{Z}/2$ for each prime p and also one for the signs (*i.e.*, the "factor at ∞"), and the isomorphism is given by

$$(x,y) \longmapsto (\mathrm{ord}_p(x) \bmod 2, \mathrm{ord}_p(y) \bmod 2)_{p,\infty}.$$

Now we are ready to prove the following theorem, which is usually considered as the first step in proving the Mordell-Weil theorem.

Theorem 15.1.2 (Weak Mordell-Weil) *Let $E/\mathbf{Q}$ be an elliptic curve given by $Y^2 = (X-\alpha)(X-\beta)(X-\gamma)$. Then image of the map $f_\alpha \times f_\beta$ constructed above is finite. In fact, we have an injection*

$$E(\mathbf{Q})/2E(\mathbf{Q}) \hookrightarrow \prod_{\infty,\, p|\Delta} (\mathbf{Z}/2 \times \mathbf{Z}/2).$$

[2]This defines a "p-adic valuation" on $\mathbf{Q}$. For more on this see Chapter (15.2)

Proof: Suppose $P = (x_0, y_0) \in E(\mathbf{Q})$. Note that as $\alpha, \beta, \gamma \in \mathbf{Z}$, for any prime p, $\|\alpha|_p, \|\beta|_p, \|\gamma|_p$ are all less then or equal to 1.

For a prime p, let $p^a \parallel x_0 - \alpha$, $p^b \parallel x_0 - \beta$, $p^c \parallel x_0 - \gamma$. We must have $a+b+c \equiv 0(\bmod 2)$ because the product $(x_0-\alpha)(x_0-\beta)(x_0-\gamma)$ is a square. If $p^a \parallel x_0 - \alpha$ and if $a < 0$ then $x_0 - \alpha$ has a factor of $p^{|a|}$ in the denominator. Then it follows very easily that $p^a \parallel x_0-\alpha, x_0-\beta, x_0-\gamma$. For, we know that $\alpha \in \mathbf{Z}$ and $|x_0 - \alpha|_p = p^{|a|}$ and then

$$|x_0 - \alpha|_p \leq \max\{|x|_p, |\alpha|_p\}$$

and $|\alpha|_p \leq 1$ so $|x_0 - \alpha|_p \leq |x_0|_p$. And this means that $|x_0|_p \geq p^{|a|}$ so x_0 has a factor of p^a in the denominator and then so do $x_0 - \beta, x_0 - \gamma$, moreover p^a is the exact power of p dividing all three of them. Then as $a+b+c \equiv 0(\bmod 2)$, gives $a = b = c \equiv 0(\mathrm{mod}2)$. Now let us assume then that

$$p^a \parallel x_0 - \alpha, \qquad a > 0.$$

If $p|\alpha - \beta$ then $p|x_0 - \beta$ and so if $p \not| \Delta$ then $b = c = 0$.

Hence we conclude that if $p \not| \Delta$ then $a = b = c \leq 0$ and $a \equiv 0(\mathrm{mod}2)$ or $a > 0$ and $b = 0, c = 0$, and further even in this case $a \equiv 0(\mathrm{mod}2)$. So we see that if $p \not| \Delta$ then the image of $((x_0 - \alpha), (x_0 - \beta))$ is $(0,0)$ in the p^{th} component of

$$\mathbf{Q}^*/\mathbf{Q}^{*2} \times \mathbf{Q}^*/\mathbf{Q}^{*2} = \prod_{\infty, p} (\mathbf{Z}/2 \times \mathbf{Z}/2)\,.$$

This finishes the proof of the weak Mordell-Weil theorem. □

In fact we can be even more careful and get a better idea of the image. Before we do that let us introduce some terminology[3]. We say a prime p is *fairly bad* if p divides exactly one of $\alpha - \beta, \alpha - \gamma, \beta - \gamma$. We say p is *very bad* if p divides atleast two of $\alpha - \beta$,$\beta - \gamma$,$\alpha - \gamma$. We say p is *good* if $p \not| \Delta$. Then calculation above (details are left to the reader) shows that if p is a very bad prime (so $p|\alpha - \beta$ and $p|\beta - \gamma$, say) then the image lands inside $\mathbf{Z}/2 \times \mathbf{Z}/2$. Now if $p|\alpha - \beta$ but not the rest then image lands inside $\mathbf{Z}/(2)$. Thus we have proved the following:

[3]The highbrow reader will realise the connection of the terminology with the standard terminology in Algebraic Geometry of stable, semistable and unstable reduction.

Proposition 15.1.4 *The image of the above map is contained in*

$$\mathbf{Z}/(2) \times \prod_{p \text{ fairly bad}} \mathbf{Z}/2 \times \prod_{p \text{ very bad}} (\mathbf{Z}/2 \times \mathbf{Z}/2)\,.$$

In particular we have the following corollary:

Corollary 15.1.5 *With notations as above,*

$$\begin{aligned} n_1 &= \#\text{ fairly bad primes,} \\ n_2 &= \#\text{ very bad primes} \end{aligned}$$

then we get:

$$|E(\mathbf{Q})/2E(\mathbf{Q})| \leq 2^{n_1+2n_2+1}.$$

We now have to show that $E(\mathbf{Q})$ is finitely generated. For this we state the following very general lemma.

Lemma 15.1.6 *Suppose A be an abelian group and suppose there is a function $h : A \to \mathbf{R}$ with the following properties :*

1. *$h(O) = 0$*
2. *Let $Q \in A$. There is a constant C_1 depending on A and Q so that for all $P \in A$,*
$$h(P+Q) \leq 2h(P) + C_1.$$
3. *There is an integer $m \geq 2$ and a constant C_2 depending on A, so that for all $P \in A$*
$$h(mP) \geq m^2 h(P) - C_2.$$
4. *For every constant C_3, the set*
$$\{P \in A \mid h(P) \leq C_3\}$$
is a finite set.

And suppose that for m as in (3), the group A/mA is finite. Then A is finitely generated.

Proof: Choose element $Q_1, \ldots, Q_r \in A$ which are coset representatives for A/mA. Now let $P \in A$. The idea is to show that by subtracting appropriate linear combination of $Q_1, \ldots, Q_r$ from P, we shall make the height of the resulting point less than a constant independent of P. Then the points $Q_1, \ldots, Q_r$ and a finite set of points will generate A. Write

$$P = mP_1 + Q_{i_1} \qquad \text{for } 1 \leq i_1 \leq r$$

$$P_1 = mP_2 + Q_{i_2} \qquad \text{for } 1 \leq i_2 \leq r$$

etc.

$$P_{n-1} = mP_n + Q_{i_n} \qquad \text{for } 1 \leq i_n \leq r.$$

Now for any j, we have

$$\begin{aligned} h(P_j) &\leq m^{-2}(h(mP_j) + C_2) \\ &\leq m^{-2}(h(P_{j-1} - Q_{i_j}) + C_2) \\ &\leq m^{-2}(h(P_{j-1}) + C_1' + C_2) \end{aligned}$$

where we have chosen $C_1' =$ largest of constants from Q_i and hence working backwards from P_n we have

$$\begin{aligned} h(P_n) &\leq (2m^{-2})^n h(P) + (m^{-2} + 2m^{-4} + \cdots + 2^{n-1}m^{-2n})\,(C_1' + C_2) \\ &< (2m^{-2})^n h(P) + \tfrac{C_1'+C_2}{m^2-2} \\ &\leq 2^{-n}h(P) + (1/2)(C_1' + C_2) \qquad (\text{as } m \geq 2) \end{aligned}$$

So it follows by taking n very large that

$$h(P_n) \leq 1 + \frac{C_1' + C_2}{2}$$

and also

$$P = m^n P_n + \sum_{j=1}^{n} m^{j-1} Q_{i_j}$$

it follows that every $P \in A$ is a linear combination of the points in the set

$$\{Q_1, \ldots, Q_r\} \cup \{Q \in A \mid h(Q) \leq 1 + (1/2)(C_1' + C_2)\}\,.$$

□

Definition: Any function h on an abelian group A satisfying properties (1)-(4) of Lemma (15.1.6) is called an *height* function on A. For $P \in A$, $h(P)$ is called the height of P.

The Mordell-Weil theorem follows from Lemma (15.1.6) (provided one constructs a function on $E(\mathbf{Q})$ satisfying the conditions of the lemma) and from the Weak Mordell-Weil theorem (15.1.2). Thus the problem boils down to constructing an height function on $E(\mathbf{Q})$. This is accomplished by the following.

Theorem 15.1.3 *Let $h : E(\mathbf{Q}) \to \mathbf{R}$ defined as follows : If $P = O$ then $h(O) = 0$; if $P \neq O$ and $P = (x, y), x = p_1/q_1$ then*

$$h(P) = \max(\log |p_1|, \log |q_1|)$$

Then h above satisfies the condition of the Lemma (15.1.6) with $m = 2$. In particular $E(\mathbf{Q})$ is finitely generated.

This theorem will be proved in Chapter (15.2). The function h as above is called the *height function* on the elliptic curve or sometimes also called the *naive* height on an elliptic curve.

Remark: Recall that the natural logarithm of a natural number is roughly twice the number of digits in its base 10 expansion ($\log(10) \log_{10}(n) = \log(n)$). So it is reasonable to think of h as a measure of how "complicated" the solution is. A solution with large height has a very large numerator or denominator in its x-coordinate and so is more "complicated". Thus the height should be thought of as a measure of "complexity" of the solution.

We will see in Chapter (cf. (15.2)) that $h(2P)$ is roughly 4-times $h(P)$, so that this agrees with our original idea of "ascent" where we constructed larger solutions by the "chord and tangent" procedure. Descent can be rephrased to mean constructing from a given solution, another solution with a smaller height. The finite generation of the Mordell-Weil group and the fact that the height function is non-negative shows that this procedure cannot continue beyond a finite number of steps. Thus Mordell-Weil theorem is the embodiment of the principle of "infinite descent" as envisaged by Fermat.

Definition: The Mordell-Weil theorem and structure theorem for finitely generated groups implies that we have an isomorphism of abelian groups

$$E(\mathbf{Q}) = \mathbf{Z}^r + \mathrm{Tor}(E(\mathbf{Q}))$$

. For some integer $r \geq 0$. Then r is called the *rank* of the elliptic curve E, more precisely the $\mathbf{Q}$-rank of E. We will denote the rank of E by r_E and if there is no confusion then simply by r.

Corollary 15.1.7 *Let r_E be the rank of $E(Q)$, in the notations of Corollary (15.1.5) we have:*

$$r_E \leq n_1 + 2n_2 - 1. \tag{15.5}$$

Proof: Follows at once from Corollary (15.1.5) and from following fact: if $A = \mathbf{Z}^r + \mathrm{Tor}\, A$ is a finitely generated abelian group and if $\mathrm{Tor}_2(A)$ is the 2-torsion subgroup of A, then

$$\|A/2A| = 2^r \| \mathrm{Tor}_2(A)|.$$

□

Exercise 15.1.1: For $y^2 = x^3 + x + 9$. Show that $r(E(\mathbf{Q})) \geq 2$. (In fact the rank is exactly 2, but this is much harder to prove). Hint: prove that the points $(8, 23), (0, 3)$ are "linearly independent" in the Mordell-Weil group. It can be shown the curve $y^2 = x^3 + x + d^2$ has rank at least one for all odd integers d.

Lemma 15.1.8 *Let E_1, E_2 be two elliptic curves over $\mathbf{Q}$. Suppose $f : E_1 \longrightarrow E_2$ is an isogeny between them defined over $\mathbf{Q}$. Then*

$$r_{E_1} = r_{E_2}.$$

Proof: The isogeny f induces induces a homomorphism of Mordell-Weil groups $E_1(\mathbf{Q}) \longrightarrow E_2(\mathbf{Q})$. More over as f is an isogeny, the kernel of this iduced map is finite. Consequently image of any point $P \in E_1(\mathbf{Q})$ which is not of finite order is not of finite order in $E_2(\mathbf{Q})$. So that the free abelian part of $E_1(\mathbf{Q})$ injects into $E_2(\mathbf{Q})$. So that $r_{E_1} \leq r_{E_2}$. Now to prove the inequality in the reverse direction, use the dual isogeny (see Section (13.3) of f. □

The rank of an elliptic curve is its most important arithmetical "invariant". Elliptic curves of small rank are easy to come by; but curves of large rank are very rare. In fact it is conjectured that there exist elliptic curves over $\mathbf{Q}$ of arbitrarily large rank over $\mathbf{Q}$. The largest rank known today is 12 (computed by Mestre[39]). There is a similar conjecture for elliptic curves over function fields of transcendence degree one (over finite fields), and this conjecture has been proved by Tate and Shafarevich [52]. From the number theorist's point of view it is important that the rank is computable. Unfortunately, the proof of the Mordell-Weil theorem does not give an effective set of generators for the Mordell-Weil group. We will see in Chapter (16.1) that the bound in Corollary (15.5) is a very crude one.

15.2 Heights

In this section, we construct the height function needed to complete the proof of the Mordell-Weil theorem (see (15.1.3)). We begin by defining the height of a rational point in the projective space $\mathbf{P}^n$. Now $P \in \mathbf{P}^n(\mathbf{Q})$ precisely when P has homogeneous coordinates $P = (a_0 : a_1 : \cdots : a_n)$ where $a_i \in \mathbf{Q}$. We may then find homogeneous coordinates for P where (i) a_i is an integer for each i, and (ii) the greatest common divisor of the a_i is 1. Such a choice of homogeneous coordinates is unique, upto multiplication by -1; we call them *normalised* homogeneous coordinates of a rational point.

Let $P = (a_0 : a_1 : \cdots : a_n)$ be normalised homogeneous coordinates for P. We define the *height* of P as

$$h(P) = \sup_{0 \le i \le n} \log |a_i| .$$

Here if some $a_i = 0$, we define $\log |a_i| = -\infty$, which is defined to be smaller than any real number. Note that $h(P)$ is independent of the choice of normalised coordinates. (It is usual to refer to $h(P)$ as the *logarithmic height*, and $H(P) = \exp(h(P))$ is what is usually called the height of the point P.)

It is convenient to have an expression for the height in terms of a given set of (rational) homogeneous coordinates. If p is a prime number, then any non-zero rational number x can be expressed in the form

$$x = p^{v_p(x)} \frac{a}{b}$$

where $v_p(x)$ is an integer, and a, b are non-zero integers which are relatively prime to p; the expression is unique if we also take $b > 0$. We define the *p-adic absolute value* $|x|_p$ of x by

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

Then $|\ |_p : \mathbf{Q} \to \mathbf{R}$ satisfies

(i) $|xy|_p = |x|_p + |y|_p$

(ii) $|x + y|_p \leq \max\left(|x|_p, |y|_p\right)$, and equality holds if $|x|_p \neq |y|_p$

We may compare these with the following properties of the usual absolute value:

(i) $|xy| = |x|\,|y|$

(ii) $|x + y| \leq 2\max\{|x|, |y|\}$,

where (ii) follows immediately from the triangle inequality.

We remark that for any non-zero rational x, $|x|_p = 1$ for all but a finite number of p. This is because only finitely many primes divide the numerator or denominator of x. We deduce that for any finite set S of rationals, not all 0, we have that $\sup_{x \in S} |x|_p = 1$ for all but a finite number of primes p.

Lemma 15.2.1 *Let $P = (a_0 : a_1 : \cdots : a_n) \in \mathbf{P}^n(\mathbf{Q})$. Then the height $h(P)$ is given by*

$$h(P) = \sup_{0 \leq i \leq n} \log |a_i| + \sum_{p \text{ prime}} \sup_{0 \leq i \leq n} \log |a_i|_p \qquad \cdots \quad (*)$$

In particular, the expression on the right side of $(*)$ depends only on P, and not on the particular set of homogeneous coordinates chosen.

Proof: We first note that if $x \in \mathbf{Q}^*$ is a non-zero rational, then

$$\log |x| + \sum_{p \text{ prime}} \log |x|_p = 0.$$

Indeed, this is equivalent to

$$|x| = \prod_{p \text{ prime}} p^{v_p(x)},$$

which is a restatement of the unique factorisation of rationals into products of primes.

Next, we note that the expression on the right side of the formula $(*)$ is independent of the particular homogeneous coordinates chosen. This is because if we replace $(a_0 : a_1 : \cdots : a_n)$ with $(xa_0 : xa_1 : \cdots : xa_n)$, then the right side of $(*)$ becomes

$$\begin{aligned}
&\sup_{0\le i\le n} \log|xa_i| + \sum_{p \text{ prime}} \sup_{0\le i\le n} \log|xa_i|_p \\
&= \sup_{0\le i\le n} (\log|x| + \log|a_i|) + \sum_{p \text{ prime}} \sup_{0\le i\le n} (\log|x|_p + \log|a_i|_p) \\
&= \log|x| + \sum_{p \text{ prime}} \log|x|_p + \sup_{0\le i\le n} \log|a_i| + \sum_{p \text{ prime}} \sup_{0\le i\le n} \log|a_i|_p \\
&= \sup_{0\le i\le n} \log|a_i| + \sum_{p \text{ prime}} \sup_{0\le i\le n} \log|a_i|_p .
\end{aligned}$$

Hence to prove the lemma, we may assume $(a_0 : a_1 : \cdots : a_n)$ are normalised coordinates. In this case, if p is any prime number, then $v_p(a_i) \ge 0$ for all i, while $v_p(a_i) = 0$ for some i, since the greatest common divisor of the a_i is 1, so that some a_i is not divisible by p. Hence

$$\sup_{0\le i\le n} \log|a_i|_p = 0$$

for each prime p. Thus the expression on the right in the formula for $h(P)$ reduces to the earlier definition for $h(P)$. □

We now define heights for elliptic curves. Let E be an elliptic curve over $\mathbf{Q}$, defined in $\mathbf{P}^2$ by a homogeneous equation

$$Y^2Z = X^3 + aXZ^2 + bZ^3,$$

where a, b are integers. Recall that any point of E other than $O = (0 : 1 : 0)$ has its Z-coordinate non-zero, so may be described using the affine coordinates

$$x = X/Z, \quad y = Y/Z,$$

in terms of which $E - \{O\}$ is defined by

$$y^2 = x^3 + ax + b.$$

Such a point of E is in $E(\mathbf{Q})$ precisely if $x, y \in \mathbf{Q}$.

There is a natural mapping

$$\wp : E \to \mathbf{P}^1$$

(essentially the Weierstrass $\wp$-function defined in Chapter 1), defined by

$$\wp(O) = (1 : 0),$$

$$\wp(x, y) = \wp(x : y : 1) = (x : 1).$$

This is defined on the set of points of E with complex coordinates, and is a morphism of degree 2 of algebraic curves (see Example 8.3.2). From a geometric point of view, $\wp$ is induced by the projection from $\mathbf{P}^2$ to $\mathbf{P}^1$ with center $O = (0 : 1 : 0)$ (thus $\wp$ is not defined at O itself, to begin with, but may be defined there by continuity along E). Clearly $\wp$ yields a map $E(\mathbf{Q}) \to \mathbf{P}^1(\mathbf{Q})$.

The *height function* $h_E : E(\mathbf{Q}) \to \mathbf{R}$ on the group $E(\mathbf{Q})$ of rational points is defined using the function $\wp$: let $h_E(P) = h(\wp(P))$, where $h(\wp(P))$ is the height defined above of the point $\wp(P) \in \mathbf{P}^1(\mathbf{Q})$.

Proposition 15.2.2 *The height function $h_E : E(\mathbf{Q}) \to \mathbf{R}$ has the following properties.*

(i) $h_E \geq 0$.

(ii) For any $C > 0$, the set $\{P \in E(\mathbf{Q}) \mid h_E(P) \leq C\}$ is finite.

(iii) For each $n > 1$, there exists constants C_1, C_2 depending only on E, n, such that for all points $P \in E(\mathbf{Q})$,

$$C_1 \leq h_E(nP) - n^2 h_E(P) \leq C_2.$$

Proof: (i) is clear from the definition. To prove (ii), it suffices to prove that

$$\left\{x \in \mathbf{P}^1(\mathbf{Q}) \mid h(x) \leq C\right\}$$

is finite, since the mapping $\wp : E \to \mathbf{P}^1$ has finite fibres (each fibre is of the form $\pm P$). If $x = (q : r)$ are normalised homogeneous coordinates, then q, r are relatively prime integers with $\sup\{|q|, |r|\} \leq e^C$. There are only finitely many such pairs of integers q, r.

The proof of (iii) involves the formulas describing multiplication by n in the group law of $E(\mathbf{Q})$. Specifically, we make use of the following fact, which is proved in Section (12) elliptic curves: there exist a pair of polynomials $F_n(X, Y)$, $G_n(X, Y)$ with coefficients in $\mathbf{Q}$, with the following properties –

(i) F_n and G_n are homogeneous of degree n^2

(ii) for some integer m satisfying $m > n^2$, X^m and Y^m lie in the ideal generated by F_n and G_n in the polynomial ring $\mathbf{Q}[X, Y]$

(iii) there is a commutative diagram of mappings (where $[n] : E \to E$ denotes multiplication by n on E)

$$\begin{array}{ccc} E & \xrightarrow{[n]} & E \\ \wp \downarrow & & \downarrow \wp \\ \mathbf{P}^1 & \xrightarrow{(F_n : G_n)} & \mathbf{P}^1 \end{array}$$

Note that $(F_n : G_n)$ does give a well defined mapping $\mathbf{P}^1 \to \mathbf{P}^1$. This is because by (ii), we can write

$$X^m = AF_n + BG_n, \quad Y^m = CF_n + DG_n$$

for suitable polynomials A, B, C, D; hence the only solution of $F_n = G_n = 0$ is $X = Y = 0$ *i.e.*, does not correspond to a point of $\mathbf{P}^1$. After replacing A, B, C, D by the sums of their homogeneous terms of degree $m - n^2$, we may assume A, B, C, D are all homogeneous of degree $m - n^2$.

This can be made quite explicit for $n = 2$. The duplication formula states that $2(x, y) = (x_2, y_2)$, where for $y \neq 0$, x_2 satisfies

$$x_2 = \frac{x^4 - 2ax^2 - 8bx + a^2}{4(x^3 + ax + b)}.$$

Hence one has $F_2(X, Y) = X^4 - 2aX^2Y^2 - 8bXY^3 + a^2Y^4$, $G_2 = 4X^3Y + 4aXY^3 + 4bY^4$. Now we may verify the following by direct computation:

Sublemma 15.2.3 *Let F_2, G_2 be as above, and let*

$$\begin{aligned} A(X,Y) &= \frac{1}{4\Delta}\Big(4(4a^3+27b^2)X^3 - 4a^2bX^2Y + \\ &\quad 4a(3a^3+22b^2)XY^2 + 12b(a^3+8b^2)Y^3\Big) \\ B(X,Y) &= \frac{1}{4\Delta}\Big(a^2bX^3 + a(5a^3+32b^2)X^2Y + \\ &\quad 2b(13a^3+96b^2)XY^2 - 3a^2(a^3+8b^2)\Big) \\ C(X,Y) &= \frac{1}{4\Delta}(12X^2Y + 16AY^3) \\ D(X,Y) &= \frac{-1}{4\Delta}(3X^3 - 5aXY^2 - 27bY^3) \end{aligned}$$

where $\Delta = 4a^3+27b^2 \neq 0$ is the discriminant of the cubic x^3+ax+b. Then

$$AF_2 + BG_2 = X^7, \;\; CF_2 + DG_2 = Y^7.$$

The commutative diagram in (iii) above immediately implies that for any $P \in E(\mathbf{Q})$, we have $h_E(nP) = h((F_n : G_n)(\wp(P)))$. So it suffices to show that there exist constants C_1, C_2 such that for any point $(x:y) \in \mathbf{P}^1(\mathbf{Q})$, we have

$$C_1 \leq h(F_n(x,y) : G_n(x,y)) - n^2 h(x:y) \leq C_2.$$

For any polynomial $F \in \mathbf{Q}[X,Y]$, with

$$F = \sum_{i,j} a_{ij} X^i Y^j,$$

and any prime number p, let

$$|F|_p = \sup_{i,j} |a_{ij}|_p .$$

If F is homogeneous of degree d, then for any $x, y \in \mathbf{Q}$ and any prime number p, we have

$$|F(x,y)|_p \leq |F|_p \left(\sup\left\{|x|_p, |y|_p\right\}\right)^d .$$

This follows from the properties (i), (ii) of the function $|\;|_p$ stated earlier, since any monomial M in x, y of degree d has $|M|_p \leq \left(\sup\left\{|x|_p, |y|_p\right\}\right)^d$.

Similarly, we define

$$|F| = \sup_{i,j} |a_{ij}| .$$

We then have

$$|F(x,y)| \leq (d+1)\,|F|\,(\sup\{|x|,|y|\})^d .$$

This follows from the triangle inequality

$$|x+y| \leq |x| + |y| ,$$

and the fact that there are $d+1$ distinct monomials in x, y of degree d.

Now let p be a prime number, and $(x:y) \in \mathbf{P}^1(\mathbf{Q})$. Then we have

$$\begin{aligned} &\sup\left\{\log|F_n(x,y)|_p, \log|G_n(x,y)|_p\right\} \\ &\quad\leq \sup\left\{\log|F_n|_p, \log|G_n|_p\right\} + n^2 \sup\left\{\log|x|_p, \log|y|_p\right\}. \end{aligned}$$

Similarly,

$$\begin{aligned} &\sup\{\log|F_n(x,y)|, \log|G_n(x,y)|\} \\ &\quad\leq n^2 \sup\{\log|x|, \log|y|\} + \log(d+1) + \sup\{\log|F_n|, \log|G_n|\} \end{aligned}$$

Thus

$$h(F_n(x,y) : G_n(x,y)) \leq n^2 h(x:y) + \log C_3,$$

where

$$C_3 = (d+1)(\sup\{|F_n|, |G_n|\}) \prod_{p \text{ prime}} \sup\left\{|F_n|_p, |G_n|_p\right\}.$$

Here $\sup\left\{|F_n|_p, |G_n|_p\right\}$ is a positive constant (*i.e.*, it does not depend on the ratio $(x:y)$), which equals 1 for all but a finite number of p. Hence C_3 is a positive constant.

On the other hand,

$$|x|_p^m = |x^m|_p \leq \sup\left\{|A(x,y)|_p, |B(x,y)|_p\right\} \sup\left\{|F_n(x,y)|_p, |G_n(x,y)|_p\right\},$$

so that

$$\begin{aligned} m\log|x|_p \quad\leq\quad &\sup\left\{\log|A(x,y)|_p, \log|B(x,y)|_p\right\} + \\ &\sup\left\{\log|F_n(x,y)|_p, \log|G_n(x,y)|_p\right\}. \end{aligned}$$

Also

$$|x|^m \leq 2 \sup\{|A(x,y)|, |B(x,y)|\} \sup\{|F_n(x,y)|, |G_n(x,y)|\},$$

so that

$$\begin{aligned} m \log|x| &\leq \log 2 + \sup\{\log|A(x,y)|, \log|B(x,y)|\} + \\ &\quad \sup\{\log|F_n(x,y)|, \log|G_n(x,y)|\}. \end{aligned}$$

This gives

$$\begin{aligned} m \log|x|_p &\leq \sup\left\{\log|A|_p, \log|B|_p\right\} + (m-n^2)\left(\sup\left\{\log|x|_p, \log|y|_p\right\}\right) + \\ &\quad \sup\left\{\log|F_n(x,y)|_p, \log|G_n(x,y)|_p\right\}, \end{aligned}$$

for any prime p, and

$$\begin{aligned} &m \log|x| \leq \log(2(m-n^2+1)) + \sup\{\log|A|, \log|B|\} + \\ &\quad (m-n^2)(\sup\{\log|x|, \log|y|\}) + \sup\{\log|F_n(x,y)|, \log|G_n(x,y)|\}. \end{aligned}$$

Similarly, we see that

$$\begin{aligned} &m \log|y|_p \leq \sup\left\{\log|C|_p, \log|D|_p\right\} + \\ &(m-n^2)\left(\sup\left\{\log|x|_p, \log|y|_p\right\}\right) + \left(\sup\left\{\log|F_n(x,y)|_p, \log|G_n(x,y)|_p\right\}\right), \end{aligned}$$

for any prime p, and

$$\begin{aligned} &m \log|y| \leq \log(2(m-n^2+1)) + \sup\{\log|C|, \log|D|\} + \\ &\quad (m-n^2)(\sup\{\log|x|, \log|y|\})(\sup\{\log|F_n(x,y)|, \log|G_n(x,y)|\}). \end{aligned}$$

Hence

$$\begin{aligned} &m \sup\left\{\log|x|_p, \log|y|_p\right\} \leq \log C_p + \\ &\sup\left\{\log|F_n(x,y)|_p, \log|G_n(x,y)|_p\right\} + (m-n^2)\sup\left\{\log|x|_p, \log|y|_p\right\}, \end{aligned}$$

where

$$C_p = \sup\left\{|A|_p, |B|_p, |C|_p, |D|_p\right\},$$

and

$$m(\sup\{\log|x|,\log|y|\}) \le \log C_\infty + \\ \sup\{\log|F_n(x,y)|,\log|G_n(x,y)|\} + (m-n^2)\sup\{\log|x|,\log|y|\}$$

where

$$C_\infty = 2(m-n^2+1)\sup\{|A|,|B|,|C|,|D|\}.$$

Note that $C_p = 1$ for all but finitely many p. Hence

$$n^2 \sup\left\{\log|x|_p,\log|y|_p\right\} \le \log C_p \sup\left\{\log|F_n(x,y)|_p,\log|G_n(x,y)|_p\right\},$$

and

$$n^2\sup\{\log|x|,\log|y|\} \le \log C_\infty \sup\{\log|F_n(x,y)|,\log|G_n(x,y)|\},$$

which yields

$$h(F_n(x,y):G_n(x,y)) - n^2 h(x:y) \ge -\log C_4,$$

where

$$C_4 = 2(m-n^2+1)\sup\{|A|,|B|,|C|,|D|\}\prod_{p \text{ prime}}\sup\left\{|A|_p,|B|_p,|C|_p,|D|_p\right\}$$

is a positive constant.

This completes the proof of the proposition. □

We also need the following lemma.

Lemma 15.2.4 *Let $P_0 \in E(\mathbf{Q})$. Then there is a constant C, depending only on E and P_0, such that for all $P \in E(\mathbf{Q})$,*

$$h_E(P+P_0) \le 2h_E(P) + C.$$

Proof: We may assume $P \neq 0, \pm P_0$ (where $0 \in E$ is the origin), and that $P_0 \neq 0$. Then P, P_0 are represented by affine coordinates $P = (x,y)$, $P_0 = (x_0,y_0)$ for some $x, y, x_0, y_0 \in K$ with $x \neq x_0$. The addition formula yields $P + P_0 = (x', y')$ where

$$x' = -x - x_0 + \frac{(y-y_0)^2}{(x-x_0)^2}.$$

We may rewrite this using the fact that (x, y), (x_0, y_0) satisfy

$$y^2 = x^3 + ax + b, \quad y_0^2 = x_0^3 + ax_0 + b$$

as

$$x' = \frac{xx_0(x + x_0) - 2yy_0}{(x - x_0)^2},$$

so that

$$\wp(P + P_0) = (xx_0(x + x_0) - 2yy_0 : (x - x_0)^2) \in \mathbf{P}^1(\mathbf{Q}),$$

and

$$h_E(P + P_0) = h\left((xx_0(x + x_0) - 2yy_0 : (x - x_0)^2)\right).$$

Now for any prime number p,

$$\left|(x - x_0)^2\right|_p = |x - x_0|_p^2 \leq \sup\left\{|x|_p, |x_0|_p\right\}^2$$

$$\leq C_p \sup\left\{|x|_p, 1\right\}^2,$$

where

$$C_p = \sup\left\{|x_0|_p, 1\right\}^2;$$

also

$$\left|(x - x_0)^2\right| \leq C_\infty \left(\sup\left\{|x|, 1\right\}\right)^2,$$

where

$$C_\infty = 4(\sup\left\{|x_0|, 1\right\})^2.$$

Similarly,

$$|xx_0(x + x_0) - 2yy_0|_p \leq C_p' \sup\left\{|x|_p^2, |y|_p, 1\right\}$$

where

$$C_p' = \sup\left\{|x_0|_p^2, |2y_0|_p, 1\right\},$$

and

$$|xx_0(x + x_0) - 2yy_0| \leq C_\infty' \sup\left\{|x|^2, |y|, 1\right\}$$

where

$$C_\infty' = 3 \sup\left\{|x_0|^2, |2y_0|, 1\right\}$$

(the factor of 3 is because there are 3 terms). From the cubic equation

$$y^2 = x^3 + ax + b$$

we see that

$$|y|_p^2 \leq \sup\left\{1, |a|_p, |b|_p\right\}\left(\sup\left\{|x|_p, 1\right\}\right)^{3/2},$$

so that

$$|xx_0(x+x_0) - 2yy_0|_p \leq C_p''\left(\sup\left\{|x|_p, 1\right\}\right)^2$$

where

$$C_p'' = \sup\left\{1, |a|_p, |b|_p\right\} C_p'$$

further,

$$|y|^2 \leq \sup\{1, |a|, |b|\}\left(\sup\{|x|, 1\}\right)^{3/2},$$

so that

$$|xx_0(x+x_0) - 2yy_0| \leq C''(\sup\{|x|, 1\})^2$$

where

$$C'' = \sup\{1, |a|, |b|\} C'.$$

Note that $C_p'' = 1$ for all but finitely many primes p. Now

$$h_E(P) = h(\wp(P)) = h(x:1) = \sup\{\log|x|, 0\} + \sum_{p \text{ prime}} \sup\left\{\log|x|_p, 0\right\}.$$

Hence

$$h_E(P + P_0) \leq 2h_E(P) + C$$

where

$$C = \log 4 + (\sup\{0, 2\log|x_0|, \log|2y_0|\} + \sup\{0, \log|a|, \log|b|\}) + \\ + \sum_{p \text{ prime}} \left(\sup\left\{0, 2\log|x_0|_p, \log|2y_0|_p\right\} + \sup\left\{0, \log|a|_p, \log|b|_p\right\}\right).$$

□

One sees immediately that Theorem (15.1.3) follows from Proposition (15.2.2) and Lemma (15.2.4).

Proposition (15.2.2) may be viewed as saying that the height function $h_E : E(\mathbf{Q}) \to \mathbf{R}$ is 'approximately' a "quadratic form", in that it takes on non-negative values, and satisfies $h_E(nP) = n^2 h_E(P)$ 'approximately'. One

may in fact use the function h_E to define a genuine positive definite quadratic form $\mathbf{h}_E : E_{\mathbf{R}} = E(\mathbf{Q}) \otimes_{\mathbf{Z}} \mathbf{R} \to \mathbf{R}$, called the *canonical height* or *Néron-Tate height*; the associated bilinear form $\mathbf{h}_E : E(\mathbf{Q})_{\mathbf{R}} \otimes_{\mathbf{R}} E(\mathbf{Q})_{\mathbf{R}} \to \mathbf{R}$ is called the *height pairing*.

This is done as follows: the inequality (from Proposition (15.2.2)(iii))

$$|h_E(2P) - 4h_E(P)| < C,$$

where C depends only on E, gives

$$\left|h_E(2^n P) - 4h_E(2^{n-1}P)\right| < C,$$

which we may rewrite as

$$\left|\frac{h_E(2^n P)}{4^n} - \frac{h_E(2^{n-1}P)}{4^{n-1}}\right| < \frac{C}{4^n}.$$

Hence if

$$a_n = \frac{h_E(2^n P)}{4^n},$$

then for $n > m \geq 0$,

$$|a_n - a_m| < C\left(\frac{1}{4^{m+1} + \cdots + \frac{1}{4^n}}\right) = \frac{C}{3}\left(\frac{1}{4^m} - \frac{1}{4^n}\right). \tag{15.6}$$

Hence $\{a_n\}$ is a Cauchy sequence. Define

$$\mathbf{h}_E(P) = \lim_{n\to\infty} a_n = \lim_{n\to\infty} \frac{h_E(2^n P)}{4^n}.$$

This defines a function $\mathbf{h}_E : E(\mathbf{Q}) \to \mathbf{R}$. Note that eq. (15.6) with $m = 0$ implies that

$$|a_n - h_E(P)| < \frac{C}{3}\left(1 - \frac{1}{4^n}\right),$$

so that

$$|\mathbf{h}_E(P) - h_E(P)| \leq \frac{C}{3}. \tag{15.7}$$

We state a result that gives the properties of $\mathbf{h}_E$.

Proposition 15.2.5 *The Neron-Tate height* $\mathbf{h}_E : E(\mathbf{Q}) \to \mathbf{R}$ *satisfies the following properties.*

(i) For all $P, Q \in E(\mathbf{Q})$,

$$\mathbf{h}_E(P+Q) + \mathbf{h}_E(P-Q) = 2\mathbf{h}_E(P) + 2\mathbf{h}_E(Q)$$

This is known as the parallelogram law.

(ii) For all $P \in E(\mathbf{Q})$,

$$\mathbf{h}_E(nP) = n^2 \mathbf{h}_E(P).$$

(iii) $\mathbf{h}_E : E(\mathbf{Q}) \to \mathbf{R}$ *is a quadratic form i.e., the pairing*

$$< P, Q >= \mathbf{h}_E(P+Q) - \mathbf{h}_E(P) - h_E(Q)$$

is symmetric and bilinear.

(iv) Let $P \in E(\mathbf{Q})$. *Then* $\mathbf{h}_E(P) = 0 \Leftrightarrow P$ *is a torsion point.*

(v) $P \longmapsto \mathbf{h}_E(P) - h_E(P)$ *is a bounded function on* $E(\mathbf{Q})$.

Proof: We have already proved (v) (see eq. (15.7)). It is a standard result that (i) and (iii) are equivalent; indeed, if $< , >$ is bilinear, then

$$\begin{aligned}
&\mathbf{h}_E(P+Q) + \mathbf{h}_E(P-Q) \\
&\quad = \quad 2\mathbf{h}_E(P) + 2\mathbf{h}_E(Q) + < P, Q > + < P, -Q > \\
&\quad = \quad 2\mathbf{h}_E(P) + 2\mathbf{h}_E(Q) + < P, 0 > \\
&\quad = \quad 2\mathbf{h}_E(P) + 2\mathbf{h}_E(Q).
\end{aligned}$$

Conversely, suppose the parallelogram law holds. We compute that

$$\begin{aligned}
&< P+R, Q > - < P, Q > - < R, Q > \\
&\quad = \quad (\mathbf{h}_E(P+Q+R) - \mathbf{h}_E(P+R) - \mathbf{h}_E(Q)) \\
&\qquad -(\mathbf{h}_E(P+Q) - \mathbf{h}_E(P) - \mathbf{h}_E(Q)) \\
&\qquad -(\mathbf{h}_E(Q+R) - \mathbf{h}_E(Q) - \mathbf{h}_E(R)) \\
&\quad = \quad \mathbf{h}_E(P+Q+R) - \mathbf{h}_E(P+Q) - \mathbf{h}_E(Q+R) \\
&\qquad -\mathbf{h}_E(R+P) + \mathbf{h}_E(P) + \mathbf{h}_E(Q) + \mathbf{h}_E(R). \qquad (15.8)
\end{aligned}$$

Now use the formulas

$$\begin{aligned}
\mathbf{h}_E(P+Q+R)+\mathbf{h}_E(P+Q-R) &= 2\mathbf{h}_E(P+Q)+2\mathbf{h}_E(R) \quad (15.9)\\
\mathbf{h}_E(P+Q+R)+\mathbf{h}_E(P-Q+R) &= 2\mathbf{h}_E(P+R)+2\mathbf{h}_E(Q) \quad (15.10)\\
\mathbf{h}_E(P+R-Q)+\mathbf{h}_E(P+Q-R) &= 2\mathbf{h}_E(P)+2\mathbf{h}_E(Q-R) \quad (15.11)\\
\mathbf{h}_E(Q+R)+\mathbf{h}_E(Q-R) &= 2\mathbf{h}_E(Q)+2\mathbf{h}_E(R) \quad (15.12)
\end{aligned}$$

Subtracting eq. (15.11) from the sum of eq. (15.9) and eq. (15.10) yields

$$\begin{aligned}
&\mathbf{h}_E(P+Q+R) = \\
&\quad \mathbf{h}_E(P+Q)+\mathbf{h}_E(P+R)+\mathbf{h}_E(Q)+\mathbf{h}_E(R) \\
&\quad -\mathbf{h}_E(P)-\mathbf{h}_E(Q-R). \qquad (15.13)
\end{aligned}$$

Now subtracting eq. (15.12) from eq. (15.13), we obtain the desired formula.

We do not give the proof of eq. (15.11) here; for a proof, see Silverman's book [57], Ch. VIII, Theorem 9.3, or the appendix to Mumford's book [42] on the Mordell-Weil theorem for abelian varietes, where the expression on the right side of eq. (15.8) is shown to vanish, using the so called 'theorem of the cube' (this is less elementary than Silverman's proof, but gives a conceptual reason for the validity of eq. (15.11).

To prove the formula $\mathbf{h}_E(nP) = n^2\mathbf{h}_E(P)$, note that it holds by definition of $\mathbf{h}_E$ for $n = 2^k$. Now for any given n, we know by Proposition (15.2.2) that $|h_E(nQ) - n^2 h_E(Q)| < C$ for some constant C, for all $Q \in E(\mathbf{Q})$; since $|\mathbf{h}_E(Q) - h_E(Q)|$ is also bounded, we see that $|\mathbf{h}_E(nQ) - n^2\mathbf{h}_E(Q)| < C_1$ for some constant C_1 (depending on n). Applying this to $Q = 2^m P$, we get $|\mathbf{h}_E(nP) - n^2\mathbf{h}_E(P)| < C_1 4^{-m}$, where m is arbitrarily large *i.e.*, $\mathbf{h}_E(nP) = n^2\mathbf{h}_E(P)$.

Lastly, we claim that $\mathbf{h}_E(P) = 0 \Leftrightarrow P$ is a torsion point. If P is a torsion point with $nP = 0$, then $0 = \mathbf{h}_E(nP) = n^2\mathbf{h}_E(P)$, so $\mathbf{h}_E(P) = 0$. If $\mathbf{h}_E(P) = 0$, then $\mathbf{h}_E(nP) = 0$ for all n. Hence $h_E(nP)$ is bounded as a function of n; by Proposition (15.2.2), this implies the set of multiples of P is finite, so that P is a torsion point. □

Exercise 15.2.1: Let $F : \mathbf{P}^n \to \mathbf{P}^m$ be a morphism of complex varieties which is defined over $\mathbf{Q}$ by polynomials of degree d. Show that there is a constant $C > 0$ such that for any $P \in \mathbf{P}^n(\mathbf{Q})$,

$$|h(F(P)) - dh(P)| \leq C.$$

(Hint: F is given by

$$(X_0 : \cdots : X_n) \longmapsto \\ (F_0(X_0, \ldots, X_n) : F_1(X_0, \ldots, X_n) : \cdots : F_m(X_0, \ldots, X_n))$$

where $F_i \in \mathbf{Q}[X_0, \ldots, X_n]$ are homogeneous of degree d. Since F defines a morphism on $\mathbf{P}^n(\mathbf{C})$, the polynomials F_i have no common zeroes in $\mathbf{P}^n$. Use the nullstellensatz to show that for some $e > 0$, $X_i^e \in \sum_j F_j \mathbf{Q}[X_0, \ldots, X_n]$ for all i. Then imitate the proof of Proposition (15.2.2).)

Exercise 15.2.2: Let $f : \mathbf{P}^n \times \mathbf{P}^m \to \mathbf{P}^{mn+m+n}$ be the *Segre embedding*, given by

$$((X_0 : \cdots : X_n), (Y_0 : \cdots : Y_m)) \longmapsto \\ (X_0Y_0 : X_0Y_1 : \cdots : X_iY_j : \cdots : X_nY_m).$$

Show that for any $P \in \mathbf{P}^n(\mathbf{Q})$, $Q \in \mathbf{P}^m(\mathbf{Q})$ we have

$$h(f(P, Q)) = h(P) + h(Q).$$

Chapter 16

Computing the Mordell-Weil group

In this chapter we discuss the problem of computing the Mordell-Weil group of elliptic curves over $\mathbf{Q}$. In principle the proof of the Mordell-Weil theorem, gives us a procedure to give a bound on the rank of the Mordell-Weil group of an elliptic curve.

But before we go into the general principles involved let us consider the following example. Let E_n be the elliptic curve given by $y^2 = x^3 - n^2x = x(x-n)(x+n)$ (cf. Exercise (14.2.1)). The discriminant of E_n is $\Delta_n = (-n-n)^2(n-0)^2(-n-0)^2 = 4n^6$. Hence if n is even, every prime dividing Δ_n is a very bad prime and if n is odd then 2 is fairly bad and any prime dividing n is very bad. Hence in particular, if $\nu(n)$ denotes the number of distinct prime divisors of n, then by Corollary 15.5, we see that the rank of $E_n(\mathbf{Q})$ satisfies:

$$\text{rank } E_n(\mathbf{Q}) \leq \begin{cases} 2\nu(n) - 1 & \text{if } n \text{ is even;} \\ 2\nu(n) & \text{if } n \text{ is odd.} \end{cases} \tag{16.1}$$

Unfortunately, this bound is not a very good one. For instance, for $n = 3$, the curve is $E_3 : y^2 = x^3 - 3^2x$ and this has rank 0 (we will prove this soon) but the bound we get is that the rank of $E_3 \leq 2$. We would therefore like to improve upon this crude bound on the rank. So what we are going to do now is to examine the proof of the Mordell-Weil theorem in detail.

16.1 Descent

Let us now derive a corollary to the Mordell-Weil theorem, which is central to descent computations. Though the procedure is very explicit, in practice

it is often more convenient to do these computations on a computer. Let

$$E : Y^2 = (X - \alpha)(X - \beta)(X - \gamma)$$

be an elliptic curve with α, β, γ in integers. Let $P = (x, y)$ be a rational point on E. Then we can write:

$$\begin{aligned} x - \alpha &= au^2, \\ x - \beta &= bv^2, \\ x - \gamma &= abw^2. \end{aligned}$$

where $a, b, ab \in \mathbf{Z}$ are squarefree and divide the discriminant and u, v, w are non-zero rational numbers. Eliminating x between these three equations we have two equations:

$$\begin{aligned} au^2 - bv^2 &= \beta - \alpha \\ bv^2 - abw^2 &= \gamma - \beta. \end{aligned}$$

In other words the triple (u, v, w) is a rational solution to the pair of quadratic equations:

$$aU^2 - bV^2 = \beta - \alpha \tag{16.2}$$

$$bV^2 - abW^2 = \gamma - \beta. \tag{16.3}$$

Conversely, if for any pair of numbers (a, b) we can find a solution to eq. (16.2) and eq. (16.3) then we can get a rational point on our original elliptic curve by setting $x = au^2 + \alpha = bv^2 + \beta = abw^2 + \gamma, y = abuvw$. We have a pair of quadratic equations eq. (16.2), eq. (16.3) for each pair of numbers (a, b) and there are only finitely many pairs of equations with a, b both dividing the discriminant,. The set of these pairs is a group (under coordinate-wise multiplication and reduction modulo $\mathbf{Q}^{*2}$) which is isomorphic to the group $\prod_{p|\Delta,\infty} (\mathbf{Z}/2 \times \mathbf{Z}/2)$. Now what we have checked so far can be formally stated as follows:

Proposition 16.1.1 *Let $E : Y^2 = (X - \alpha)(X - \beta)(X - \gamma)$ be an elliptic curve as above. Then the image of the map (15.3) from*

$$E(\mathbf{Q})/2E(\mathbf{Q})\mathbf{Q}^*/\mathbf{Q}^{*2} \times \mathbf{Q}^*/\mathbf{Q}^{*2}$$

is contained in $\prod_{p|\Delta,\infty} \mathbf{Z}/2 \times \mathbf{Z}/2$. A pair (a, b) is in the image of the map iff the pair of corresponding quadrics eq. *(16.2)and* eq. *(16.3) have a rational solution.*

Remark: From Section (12.4.3) we know that the pair of quadrics eq. (16.2), eq. (16.3) are defining equations of a curve C in $\mathbf{P}^3$ and, C has genus 1. If the pair eq. (16.2), eq. (16.3) have a common rational solution *i.e.*, , C has a rational point then C is in fact an elliptic curve over $\mathbf{Q}$. Further, C is in fact isomorphic to E.

So the problem is reduced to solving, in rational numbers, a pair of quadratic equations. Let us assume that the numbers α, β, γ are in ascending order. Then it follows that $\beta - \alpha, \gamma - \beta$ are both positive. Hence it follows that if a pair (a, b) satisfies $a < 0, b > 0$ then eq. (16.2) has no rational solution because there are no real solutions. And similarly if both $a, b < 0$ then eq. (16.3) has no rational solutions because there are no real solutions. Thus we see at once that we have eliminated a fair number of pairs (a, b). This fact is of practical value: for instance consider the curve E_p where p is an odd prime. Then the only primes dividing Δ_p are $2, p$. And there are 64 pairs of equations! But the above remarks reduce the number to 32. We are now going to introduce an important concept which will help us to generalize this observation. We digress from the main theme and recall a few definitions for those who are not familiar with p-adic numbers.

16.2 Local methods

Let p be a prime number. In Chapter (15.2) we have introduced an absolute value ("the p-adic absolute value") on the field of rational numbers. Briefly it is the following function: write every nonzero rational number a/b as $a/b = p^n(a_1/b_1)$ where $n \in \mathbf{Z}$ and $p \not| \ a_1, b_1$. Then the *p-adic absolute value* is defined to be $\|a/b|_p = p^{-n}$. One checks easily that this satisfies the property that $\|\alpha + \beta|_p \leq \max(\|\alpha|_p, \|\beta|_p)$. Thus this absolute value, usually known as non-archimedean valuation, satisfies a slightly stronger form of the triangle inequality. One also sees at once that the space of rational numbers becomes a metric space with respect to the metric[1] given by the p-adic absolute value. This is quite analogous to the usual absolute value. In fact we sometimes think of the usual absolute value corresponding to the "infinite prime" of $\mathbf{Q}$. This of course is only a symbolic aid.

[1]The metric is $d(\alpha, \beta) = \|\alpha - \beta|_p$

This metric space is not complete and by the standard procedure of completing metric spaces we obtain a complete metric space which we denote by $\mathbf{Q}_p$. This is a field (a "topological field") and the field of rational numbers is naturally a subfield. For the usual absolute value on $\mathbf{Q}$, the completion is of course $\mathbf{R}$ which we sometimes denote by $\mathbf{Q}_\infty$. We call elements of $\mathbf{Q}_p$ "p-adic" numbers and the field $\mathbf{Q}_p$ is called the field of p-adic numbers. Every p-adic number has a unique convergent series expansion of the form

$$\sum_{n \geq -N}^{\infty} a_n p^n \qquad \text{for some } N > 0$$

where $0 \leq a_n \leq p-1 \ \forall n \geq -N$. The absolute value $|\ |_p$ on $\mathbf{Q}$ extends naturally to an absolute value on $\mathbf{Q}_p$, again denoted by $|\ |_p$. The subset $\mathbf{Z}_p = \left\{x \in \mathbf{Q}_p | \|x|_p \leq 1\right\}$ is a ("closed") subring, which is just the closure of the ring $\mathbf{Z}$ in $\mathbf{Q}_p$; $\mathbf{Z}_p$ is called the ring of *p-adic integers*. It is compact and the set $p\mathbf{Z}_p = \{x \in \mathbf{Z}_p | \|x|_p < 1\}$ is a maximal ideal and in fact it is the unique maximal ideal of $\mathbf{Z}_p$. Moreover there is a natural isomorphism $\mathbf{Z}_p/p\mathbf{Z}_p \xleftarrow{\sim} \mathbf{Z}/p\mathbf{Z}$, which gives rise to a homomorphism $\mathbf{Z}_p\mathbf{Z}/p\mathbf{Z}$ which is called *reduction modulo p*.

Exercise 16.2.1: If $x, y \in \mathbf{Q}_p$ and if $\|x|_p > \|y|_p$, then $\|x+y|_p = \|x|_p$.

One property which distinguishes $\mathbf{Q}_\infty$ from $\mathbf{Q}_p$ is the following important property which was discovered by Hensel (who discovered the field of p-adic numbers). We will state it in the simplest case to avoid more notation.

Lemma 16.2.1 (Hensel's Lemma) *Let p be a prime number and let $f(X)$ be a polynomial with integer coefficients. Suppose that α is a solution of the congruence*

$$f(X) \equiv 0 \pmod{p}$$

such that $f'(\alpha) \not\equiv 0 (\mathrm{mod} p)$ (where f' is the derivative of f). Then there is a p-adic number $\tilde{\alpha} \in \mathbf{Z}_p$ such that $f(\tilde{\alpha}) = 0$ and $\tilde{\alpha} \equiv \alpha (\mathrm{mod} p)$.

Proof: The proof is really the well known "Newton's Method" of constructing approximate roots of a polynomial. So assume α is a root modulo p. We will construct a sequence of integers, $\alpha_1, \alpha_2, \ldots$ converging in $\mathbf{Z}_p$. Let

$\alpha_1 \equiv \alpha (\mathrm{mod} p)$ be any integer. Then for any β we have the following expansion: (the "Taylor expansion" of f)

$$f(\alpha_1 + p\beta) = f(\alpha_1) + p\beta f'(\alpha_1) + \text{higher order terms}$$

The higher order terms are all divisble by p^2. Then we have

$$f(\alpha_1 + p\beta) \equiv f(\alpha_1) + p\beta f'(\alpha_1) \pmod{p^2}.$$

Thus if we arrange to have

$$f(\alpha_1) + p\beta f'(\alpha_1) \equiv 0 \pmod{p^2}$$

then we will get a solution of $f(X) \equiv 0 (\mathrm{mod} p^2)$. So let us see if we can acheive this. Since $f(\alpha_1) \equiv 0 (\mod p)$, we can write $f(\alpha_1) = pr$ then we see at once that if $r + \beta f'(\alpha_1) \equiv 0 (\mod p)$ then we are done. And so far we have not chosen β, so we choose β now so that the congruence $\beta f'(\alpha_1) \equiv -r (\mathrm{mod} p)$ is solvable. This is certainly possible as $f'(\alpha_1) \not\equiv 0 (\mathrm{mod} p)$. Thus we can by a similar inductive argument construct a sequence of integers $\{\alpha_n\}$ with the property that $f(\alpha_n) \equiv 0 (\mathrm{mod} p^n)$ and such that $\alpha_n \equiv \alpha_{n-1} (\mathrm{mod} p^{n-1})$. This sequence converges in the field of p-adic numbers to give a root of $f(X) = 0$. $\square$

Exercise 16.2.2: Prove the following general form of Hensel's Lemma. Suppose $f(X_1, \ldots, X_n)$ is a polynomial with integer coefficients. Suppose there is natural number $m \geq 1$ such that there is a solution $(x_1, \ldots, x_n)$ to the congruence

$$f(X_1, \ldots, X_n) \equiv 0 \pmod{p^m}$$

and suppose further that

$$\partial f / \partial X_i (x_1, \ldots, x_n) \not\equiv 0 \pmod{p^m} \qquad \forall 1 \leq i \leq n.$$

Then the equation $f(X_1, \ldots, X_n) = 0$ has p-adic solutions.

Remark: More generally Hensel's Lemma can be applied in the following context. Let V be a variety defined by equations with integer coefficients (even p-adic integer coefficients). If the variety obtained by reducing all the equations modulo a fixed prime p has a smooth point defined over $\mathbf{Z}/p\mathbf{Z}$ then this point can be "lifted" to a point with coordinates in $\mathbf{Z}_p$ which reduces to the original one under reduction modulo p. *Moreover, for a variety to have rational solutions, it is clearly neccessary to have p-adic solutions for every p and also real solutions.*

Remark: If $V/\mathbf{Q}_p$ is a variety defined by polynomials with coefficients in $\mathbf{Z}_p$. Denote by V_p the variety defined by reducing the polynomials defining V modulo p. Let $V_p(\mathbf{Z}/p)_{\text{non-sing}}$ denote the set of points of V_p defined over $\mathbf{Z}_p$ which are smooth points of V_p. Then by Remark (16.2) the map of sets $V(\mathbf{Q}_p) \longrightarrow V_p(\mathbf{Z}/p)_{\text{non-sing}}$ is surjective.

Remark: The theorem of Legendre which was proved in Section (14.1.1) can be restated in the following equivalent form:

Theorem 16.2.1 *Let a, b, c be integers. The quadratic form $ax^2+by^2+cz^2 = 0$ has solutions in $\mathbf{Q}$ if and only if it has solutions in $\mathbf{Q}_p$ for every p and also in $\mathbf{Q}_\infty$.*

It is not very hard to check that this form of the theorem is equivalent to the form in which it is proved in Section (14.1.1). Nowadays this version of the theorem is also known as "Hasse principle" for ternary quadratic forms. Sometimes it is also known as the "Local–Global" principle. More generally if V is any variety over the field of rationals we say that the Hasse principle holds for V if the existence of $\mathbf{Q}_p$-rational points (*i.e.*, "local solutions") on V is sufficient to ensure the existence of rational points (*i.e.*, "global solutions"). Fields like $\mathbf{Q}$ are called "global" fields and fields like $\mathbf{Q}_p$ are called "local" fields..

Now coming back to our problem of a pair of quadratic equations, by Remark (16.2), we can further eliminate those which don't have solutions in p-adic numbers for some prime p. In fact we have already done this for $\mathbf{Q}_\infty$. The following proposition shows that we need in fact consider only those primes that divide the discriminant.

Proposition 16.2.2 *Let p be an odd prime and a, b, c, d, e, f be integers all coprime to p. Then the pair of equations*

$$\begin{aligned} aU^2 - bV^2 &= cT^2 \\ dV^2 - eW^2 &= fT^2 \end{aligned}$$

has a solution in the field of p-adic numbers.

Proof: The idea of the proof is to use Hensel's Lemma. The condition that p does not divide $abcdef$ ensures that the reduced equations define a variety which is smooth over $\mathbf{Z}/p\mathbf{Z}$. So we will be done if we can find a solution modulo p. Here is a subtle point: the two equations define a curve of genus 1 over the algebraic closure of $\mathbf{Z}/p\mathbf{Z}$ (cf. Section (12.4.3)). Hence by the Theorem of Lang (13.2.2) there is a rational point modulo p. So there is always a p-adic solution by Hensel's Lemma and we are done. □

So we now have a finite set of equations and a finite set of primes (all of them must divide the discriminant); and we have to check if pairs of equations have solutions in the corresponding p-adic fields. Moreover, as the set of pairs (a, b) forms a group, and the map from $E/2E$ is a homomorphism into this group, we can use the group property to reduce the number of computations further: it follows from the group law that if two pairs of equations corresponding to $(a, b), (a', b')$ have solutions in a given p-adic field then so does the pair (aa', bb'). And if precisely one of them has no p-adic solutions then neither does the product.

Let us now do this for the curve $E_1 : y^2 = (x-1)x(x+1)$. We know the rank is 0 by corollary (15.5) ; but we will verify it using the p-adic methods now. There are eight pairs of equations to consider. Firstly note that the points $(-1, 0)$, $(0, 0)$, $(1, 0)$ and O on E_1 are mapped onto pairs $(2, -1), (1, -1)$, $(2, 1)$ and $(1, 1)$ respectively. The remaining four pairs can be obtained from these four pairs by multiplying them by the pair $(1, 2)$. Then by remarks in the preceeding paragraph it suffices to show that the pair of equations corresponding to $(1, 2)$ has no 2-adic solution.

Lemma 16.2.3 *The equations corresponding to the pair* $(1, 2)$*,*

$$U^2 - 2V^2 = 1, 2V^2 - 2W^2 = 1$$

do not have 2*-adic solutions.*

Proof: For, suppose they do have a 2-adic solution (u, v, w). Then we first note that u, v, w must be 2-adic integers. For, the first equation forces u, v be 2-adic integers as the two terms $u^2, 2v^2$ are divisible by powers of two of opposite parity and their difference is an integer. Then $w^2 = v^2 - 1/2$. Hence on taking absolute values we see that $\|w^2|_2 \leq \max(\|v^2|_2, \|1/2|_2) = \max(1, 2)$. Hence $\|w^2|_2 \leq 2$. Now write $w = 2^m w_1$ where $w_1 \notin 2\mathbf{Z}_2$. Then $\|w^2|_2 =$

$2^{-2m} \leq 2 \Rightarrow m \geq 0$. So all of them are 2-adic integers and hence reducing the second equation modulo 2 gives us $0 \equiv 1 (\bmod 2)$ which is impossible, and we are done. □

Hence the order of $E_1/2E_1$ is 4. Now by the Mordell-Weil theorem we know that $E_1(\mathbf{Q}) = \mathbf{Z}^r \oplus$ Torsion subgroup. And torsion in this case contains a subgroup of the form $\mathbf{Z}/2\mathbf{Z} \oplus \mathbf{Z}/2\mathbf{Z}$ corresponding to the four points of order 2. These points are $O, (0,0), (-1,0), (1,0)$. Thus we see at once that the order of $E_1/2E_1$ is $2^{\text{rank}(E_1)+2}$. Thus $E_1/2E_1$ consists only of the 2-torsion points of E_1, and so the rank of E_1 is zero.

Let us discuss one more example of descent calculations to get familiar with the techniques involved. The curve this time is $E_p : y^2 = x^3 - p^2x$, where p is an odd prime. There are only two primes dividing the discriminant and so there are 32 pairs (a, b) in all to deal with. We will only sketch the arguments. To give them in detail here would be too tedious.

Lemma 16.2.4 *The equations corresponding to the following 16 pairs*

$$(1,-2), (2,-2), (p,-2), (2p,-2),$$
$$(1,-2p), (2,-2p), (p,-2p), (2p,-2p)$$
$$(1,2), (2,2), (p,2), (2p,2)$$
$$(1,2p), (2,2p), (p,2p), (2p,2p)$$

do not have 2-adic solutions.

Proof: We will treat one typical case. The rest are proved similarly. The case $(1,2)$: the equations involved are

$$U^2 - 2V^2 = p, 2V^2 - 2W^2 = p.$$

Consider the second equation. Since the left hand side is a 2-adic unit, we claim that this forces any solution (u_0, v_0, w_0) to be 2-adic integral. To see this let us take 2-adic absolute values: note that $\|u_0|_2^2$ is an even power of 2 and $\|2v_0^2|_2$ is an odd power of 2. But the right hand side of the equality is a 2-adic unit. So that both u_0, v_0 must be 2-adic integers. Then it follows that w_0 is also a 2-adic integer. For $w_0^2 = v_0^2 - p/2$. Now taking absolute values we see that $\|w_0|_2^2 \leq \max(\|v_0|^2, \|p/2|_2)$. But then as $\|v_0|_2 \leq 1$, $\|w_0^2|_2 = 2$. Now

write $w_0 = 2^m w_1$ where $2 \not| w_1$. Then $\|w_0^2|_2 = 2^{-2m} \leq 2$, a contradiction. So that the number of pairs drops by 16! So in particular the rank is now ≤ 2 which is still what was predicted by the crude bound. Of the rest of the 16 pairs of equations the four pairs $(1,1),(p,1),(2,p),(2p,p)$ are images of the two-torsion points of E_p so that we don't have to bother about these. $\square$

Hence we are left with 12 pairs to consider and these have to be taken up case by case. Here is an example of the real power of the method:

Lemma 16.2.5 *Let $p \equiv 3(\mathrm{mod}8)$. Then the equations corresponding to the* 12 *pairs*

$$(1,-1),(2,-1),(2p,-1),(1,-p),$$
$$(2,-p),(p,-p),(2p,-p),(2,1),$$
$$(p,1),(2p,1),(1,p),(p,p)$$

do not have p-adic solutions.

Proof: Consider the first pair $(1,-1)$. Then the equations are

$$U^2 + V^2 = p, -V^2 + W^2 = p.$$

Suppose u, v, w is a solution of the above pair of equations. Then let us take the absolute value of the first equality: $\|p|_p = \|u^2 + v^2|_p$. Then $\|u|_p = \|v|_p$. For, if not then assume without loss of generality that $\|u|_p > \|v|_p$. Then by Exercise (16.2.1) we see that $1/p = \|u^2 + v^2|_p = \|u^2|_p$, which cannot happen because the right hand side is always an even power of p. Then let $u = p^m u_1, v = p^m v_1$ where u_1, v_1 are p-adic units. Then substituting this in the first equation, we see that $p^{2m}(u_1^2 + v_1^2) = p$. Thus $u_1^2 + v_1^2 = 1/p^{2m-1}$. Now taking p-adic absolute values we get $p^{2m-1} = \|u_1^2 + v_1^2|_p \leq 1$. But then this forces $m \leq 0$. Thus we can reduce the equation $u_1^2 + v_1^2 = 1/p^{2m-1}$ to see that $u_1^2 + v_1^2 \equiv 0(\mathrm{mod}p)$. Or equivalently that $(u_1/v_1)^2 \equiv -1(\mathrm{mod}p)$. But this means that -1 is a quadratic residue (*i.e.*, a square) modulo p. By the theory of quadratic residues (cf. [26]), we know that this cannot happen if $p \equiv 3(\mathrm{mod}4)$. And any prime which satisfies $p \equiv 3(\mathrm{mod}8)$ also satisfies $p \equiv 3(\mathrm{mod}4)$. So we are done. So this pair is ruled out. A similar

argument with the next pair leads us to a congruence $x^2 \equiv 2(\mathrm{mod}p)$. But again by invoking the theory of quadratic residues and the determination of the quadratic residue symbol of $2(\mathrm{mod}p)$, we see that this cannot happen as $p \equiv 3(\mathrm{mod}8)$. The rest of the cases are handled similarly and are left as exercises. □

Thus we have proved:

Proposition 16.2.6 *If p is a prime of the form $8n+3$ then the curve E_p*

$$E_p : y^2 = x^3 - p^2 x$$

has rank 0.

Remark: Using the terminology of Exercise (14.2.1) we can restate the above proposition as: if $p \equiv 3(\mathrm{mod}8)$ then p is not a congruent number.

The analysis carried out above can in fact be pushed further to prove the following.

Proposition 16.2.7

$$\mathrm{rank}(E_p) \leq \begin{cases} 2 & \text{if } p \equiv 1 \pmod 8, \\ 1 & \text{if } p \equiv 5, 7 \pmod 8. \end{cases}$$

Thus in principle we can use this method of quadrics to compute the rank. However, the fundamental problem with the procedure is that after eliminating all pairs of equations which do not have p-adic solutions, we have no way of deciding whether the remaining pairs of equations have rational solutions. This brings us to an important concept in the theory of elliptic curves.

16.3 Selmer and Tate-Shafarevich groups

It is reasonably easy to decide using Hensel's lemma, and considerations similar to the ones given above, whether a given pair of quadric equations has a solution in every $\mathbf{Q}_p$. And clearly the set of pairs which actually have a rational solution is contained in the set of pairs having solutions at every $\mathbf{Q}_p$.

We may ask if the converse is true: given a pair of equations as above having solutions in all $\mathbf{Q}_p$, do they they neccessarily have a rational solution? The answer in general is no. Lind and Reichardt (around 1942) found examples where this happens. Lind's example is the curve $2w^2 = 1 - 17z^4$ which (one checks) has solutions in $\mathbf{Q}_p$ for all p. Selmer [49] (1951) shortly after that made an extensive study of the curves $x^3 + y^3 = m$ and found more examples of this phenomenon. One of the simplest examples due to Selmer is the curve $3X^3 + 4Y^3 + 5Z^3 = 0$, which has p-adic solutions for every p but no rational solutions.

But before we proceed let us formalize what we have done so far:

Proposition 16.3.1 *The set of pairs of quadrics which have a rational solution in all $\mathbf{Q}_p$ is a group, which we denote by $S^{(2)}(E)$. This is a finite group in which every element is of order 2. Moreover this group contains $E/2E$ as a subgroup. The order of the group $S^{(2)}(E)$ can be computed in a finite number of steps.*

The group $S^{(2)}(E)$ is called *Selmer* 2-group.

Thus we have a map $E/2E \hookrightarrow S^{(2)}(E)$, and the cokernel of this map is denoted[2] by $Ш^{(2)}(E)$. So we have an exact sequence of abelian groups,

$$0 E/2E S^{(2)}(E) Ш^{(2)}(E) 0.$$

The group $Ш^{(2)}(E)$ thus consists of those pairs of quadrics which have solutions in every $\mathbf{Q}_p$ but no rational solution. This happens often enough in practice. We will soon give examples where it is non zero. If this group was zero then of course we could get the exact order of the groups in question and hence the rank. But most of the time we have to do some more work. In most of the classical examples including those studied by Fermat and others, the group $Ш^{(2)}$ usually turned out to be zero.

Remark: The calculations we carried out in the previous section for the curve $E_p : y^2 = x^3 - p^2 x$ where p is an odd prime prove that $Ш^{(2)}(E_p) = 0$ for $p \equiv 3 (\mathrm{mod} 8)$.

[2] Ш (pronounced 'sha') is the Russian letter 'Ш' the first letter in Shafarevich's name.

Remark: The curve $E_{17} : y^2 = x^3 - 17^2x$ has rank 0. Moreover it can be shown by above considerations that $S^{(2)}(E_{17})$ has order 16. Thus $Ш^{(2)}(E_{17})$ is of order 4. It is in fact isomorphic to $\mathbf{Z}/2 \times \mathbf{Z}/2$. This gives us examples of pairs of quadrics which have solutions in every $\mathbf{Q}_p$ but not in $\mathbf{Q}$. One typical pair is $U^2 + V^2 = 17, W^2 - V^2 = 17$. By Proposition (16.2.2), we know that it has solutions in every $\mathbf{Q}_p$ except possibly $\mathbf{Q}_{17}$. One checks easily that it has solutions in $\mathbf{Q}_{17}$. To see that it has no solutions in $\mathbf{Q}$, we proceed as follows: suppose $u^2 + v^2 = 17, w^2 - v^2 = 17$ is a solution in rational numbers. Then writing $u = x/t, v = y/t, w = z/t$ we see that $(x/t)^2 + (y/t)^2 = (z/t)^2 - (y/t)^2$ and hence $2(y/t)^2 = (z^2 - x^2)/t^2$. Hence, $2y^2 = (z - x)(z + x)$. So it follows that $x \equiv y \equiv z \equiv 0(\text{mod}2)$. But then t is odd and so we cannot have $x^2 + y^2 = 17t^2$.

Remark: The group $Ш^{(2)}(E)$ is in fact the 2-torsion subgroup of a bigger group denoted by $Ш(E)$ and called the *Tate-Shafarevich group* of E. We will not define it as its definition is beyond the scope of this book, the interested reader may consult [57].

The Tate-Shafarevich group is an important "arithmetical invariant" of an elliptic curve. If $Ш(E) = 0$ then the E satisfies the Hasse principle. Our calculations show that if $Ш^{(2)}(E) = 0$ then we can compute the rank of E the method of "quadratic equations" outlined above. Cassels has shown that if $Ш(E)$ is finite then its order must be a perfect square. *It is conjectured that $Ш(E)$ is finite for any elliptic curve over* $\mathbf{Q}$. Until recently there was no evidence to support this conjecture. The recent work of Rubin [47] and Kolyvagin [32] has led to the proof of this conjecture for certain special classes of elliptic curves. For instance Rubin's theorem leads (as very special cases) to the following examples (for more examples see Chapter 19):

1. Let $E : y^2 = x^3 - x$. Then $Ш(E) = 0$.
2. Let $E : y^2 = x^3 + 17x$. Then $Ш(E) = \mathbf{Z}/2 \times \mathbf{Z}/2$.
3. Let $E : y^2 = x^3 - (2^4 3^2 5)^2 x$, then $Ш(E) = \mathbf{Z}/3 \times \mathbf{Z}/3$.

The precise statement of Rubin's theorem is given in Chapter 19. Kolyvagin's work is rather technical and to even state the results is beyond the scope of this book.

16.4 Torsion points

16.4.1 Specialization modulo a prime

In this section, we indicate how the finite part of the Mordell-Weil group can be computed. This rests on a property of the torsion subgroup $E(\mathbf{Q})_{tor}$, which can be proved in an elementary way, and helps in explicitly computing the torsion subgroup of a given elliptic curve. Let E be an elliptic curve defined over $\mathbf{Q}$ with a Weierstrass equation

$$Y^2Z = X^3 + aXZ^2 + bZ^3,$$

where a, b are integers (after a rational linear change of variables, we may assume this). Let Δ be the discriminant of the cubic $x^3 + ax + b$; then Δ is an integer, and for any prime p not dividing Δ, we may consider the elliptic curve $\overline{E}$ defined over $\mathbf{F}_p = \mathbf{Z}/p\mathbf{Z}$, the finite field, by

$$Y^2Z = X^3 + \overline{a}XZ^2 + \overline{b}Z^3,$$

where the overbar denotes the image in $\mathbf{F}_p$.

There is a mapping (called *specialization*)

$$S_p : \mathbf{P}^2(\mathbf{Q}) \to \mathbf{P}^2(\mathbf{F}_p)$$

from the set of points of the projective plane with coordinates in $\mathbf{Q}$ to the set of points of the projective plane with coordinates in the finite field $\mathbf{F}_p$. To define it, recall that any point $P \in \mathbf{P}^2(\mathbf{Q})$ has a set of reduced homogeneous coordinates $(X : Y : Z)$ representing it, where X, Y, Z are integers which are relatively prime; this is unique upto multiplication by ± 1.

Given a reduced triplet $(X : Y : Z)$, define $S_p(X : Y : Z) = (\overline{X} : \overline{Y} : \overline{Z})$ where $\overline{X}$ denotes the image of X in $\mathbf{F}_p = \mathbf{Z}/p\mathbf{Z}$. If we multiply X, Y Z by -1, then $\overline{X}$, $\overline{Y}$, $\overline{Z}$ get multiplied by $\overline{-1}$, a non-zero element of $\mathbf{F}_p$; hence $(\overline{X} : \overline{Y} : \overline{Z})$ is a well defined point of $\mathbf{P}^2(\mathbf{F}_p)$. More generally, we may multiply X, Y, Z by any integer m which is not divisible by p; then $(\overline{mX} : \overline{mY} : \overline{mZ}) = (\overline{X} : \overline{Y} : \overline{Z})$. Hence if $P = (X : Y : Z) \in \mathbf{P}^2(\mathbf{Q})$ where $X, Y, Z \in \mathbf{Z}$ and at least one of X, Y, Z is not divisible by p, then $S_p(P) = (\overline{X} : \overline{Y} : \overline{Z})$.

Note that specialization modulo a non-zero prime ideal may be defined more generally for the projective space $\mathbf{P}^n$ over the quotient field of any

principal ideal ring. An important special case is specialization from $\mathbf{P}^n$ over the quotient field of a discrete valuation ring, to $\mathbf{P}^n$ over the residue field; *e.g.*, for the p-adic numbers $\mathbf{Q}_p$, one can define specialization $\mathbf{P}^n(\mathbf{Q}_p) \to \mathbf{P}^n(\mathbf{Z}/p\mathbf{Z})$.

Clearly, if $(X : Y : Z)$ are reduced homogeneous coordinates of a point P in $\mathbf{P}^2(\mathbf{Q})$ such that X, Y, Z satisfy the cubic equation

$$Y^2Z = X^3 + aX^2Z + bXZ^2 + cZ^3,$$

then $\overline{X}$, $\overline{Y}$, $\overline{Z}$ satisfy

$$\overline{Y}^2\overline{Z} = \overline{X}^3 + \overline{a}\overline{X}^2\overline{Z} + \overline{bXZ}^2 + \overline{c}\overline{Z}^2.$$

Hence specialization defines a mapping

$$S_p : E(\mathbf{Q}) \to \overline{E}(\mathbf{F}_p).$$

We state and prove the result below over $\mathbf{Q}$, for simplicity; it is valid more generally for specialisation for an elliptic curve 'with good reduction'[3] over a discrete valuation ring with residue characterictic p.

Theorem 16.4.1 *Let p be a prime number not dividing Δ. Then the specialisation mapping $S_p : E(\mathbf{Q}) \to \overline{E}(\mathbf{F}_p)$ is a* group homomorphism.

Proof: This will be proved by comparing the chord-and-tangent construction over $\mathbf{Q}$ and over $\mathbf{F}_p$, once we have an appropriate technique for specializing lines from $\mathbf{P}^2(\mathbf{Q})$ to $\mathbf{P}^2(\mathbf{F}_p)$.

If A is a line in $\mathbf{P}^2$ over $\mathbf{Q}$, then A may be defined by a homogeneous linear equation

$$dX + eY + fZ = 0$$

where $d, e, f \in \mathbf{Z}$ are relatively prime; such an equation is well defined upto multiplication by ± 1. If $\overline{d}$, $\overline{e}$, $\overline{f}$ are the images of d, e, f in $\mathbf{F}_p$, then

$$\overline{d}X + \overline{e}Y + \overline{f}Z = 0$$

defines a line $\overline{A}$ in $\mathbf{P}^2$ over $\mathbf{F}_p$, which we call the specialization of A. If $P \in \mathbf{P}^2(\mathbf{Q})$ lies on A, then clearly $S_p(P)$ lies on $\overline{A}$. Hence if $P, Q \in E(\mathbf{Q})$

[3]This means that the plane cubic obtained over the residue field is non-singular.

and $S_p(P) \neq S_p(Q)$, then the line through P and Q specializes to the line through $S_p(P)$ and $S_p(Q)$. Similarly, one checks that the tangent line to the cubic E at P specializes to the tangent line to $\overline{E}$ at $S_p(P)$.

Finally, we remark that if $P, Q \in E(\mathbf{Q})$ are distinct, but $S_p(P) = S_p(Q)$, then the line through P and Q specializes to the tangent to $\overline{E}$ at $S_p(P) = S_p(Q)$. To see this, let $P = (x : y : z)$, $Q = (x', y', z')$ where $x, y, z, x', y', z' \in \mathbf{Z}$, and x, y, z and x', y', z' are each relatively prime. Let an overbar denote the image in $\mathbf{F}_p$; then $(\overline{x} : \overline{y} : \overline{z}) = (\overline{x'} : \overline{y'} : \overline{z'})$. Hence

$$\overline{xy' - x'y} = \overline{yz' - y'z} = \overline{xz' - x'z} = 0 \text{ in } \mathbf{F}_p. \qquad (*)$$

Now $(x : y : z) \neq (x' : y' : z')$, so at least one of $\alpha = xy' - x'y$, $\beta = yz' - y'z$, $\gamma = xz' - x'z$ is non-zero. Let p^r be the largest power of p dividing all of α, β, γ; then $r > 0$ by $(*)$.

At least one of $\overline{x}$, $\overline{y}$, $\overline{z}$ is non-zero. Suppose $\overline{x} \neq 0$ (the argument in the other cases is similar). Then also $\overline{x'} \neq 0$. Hence $P = (x'x : x'y : x'z)$, $Q = (xx' : xy' : xz')$ where $xy' = x'y + p^r y''$, and $xz' = x'z + p^r z''$. Then $R = (0 : y'' : z'')$ lies on the line through P and Q. We compute that

$$x(yz' - y'z) = y(x'z + p^r z'') - z(x'y + p^r y'') = p^r(yz'' - y''z).$$

Hence from the definition of r, at least one of y'', z'', $yz'' - y''z$ is not divisible by p, *i.e.*, at least one of y'', z'' is not divisible by p. Then $S_p(R) = (0 : \overline{y''} : \overline{z''})$. Since $\overline{x} \neq 0$, $S_p(R) \neq S_p(P)$. Hence, to prove that the line through P, Q specialises to the tangent to $\overline{E}$ at $S_p(P)$, it suffices to show that $S_p(R)$ lies on this tangent.

Let $F(X, Y, Z) = Y^2Z - X^3 - aXZ^2 - bZ^3$, so that $F = 0$ is the homogeneous equation for E. Then $F(x'x, x'y, x'z) = F(xx', xy', xz') = 0$. Hence by Taylor's theorem,

$$0 = p^r y'' F_Y(x'x, x'y, x'z) + p^r z'' F_Z(x'x, x'y, x'z) + \\ p^{2r}(y''^2 F_{YY}(x'x, x'y, x'z) + 2y''z'' F_{YZ}(x'x, x'y, x'z) + \cdots)$$

where F_Y, etc. denote the respective partial derivative polynomials. Cancelling a factor of p^r, we get

$$y'' F_Y(x'x, x'y, x'z) + z'' F_Z(x'x, x'y, x'z) = p^r m$$

for some $m \in \mathbf{Z}$. Hence on passage to the residue field $\mathbf{F}_p$, and using the fact that F_Y, F_Z are homogeneous of degree 2, we get an equation in $\mathbf{F}_p$ (since $r > 0$)

$$\overline{y''}(\overline{x'})^2\,\overline{F}_Y(\overline{x},\overline{y},\overline{z}) + \overline{z''}(\overline{x'})^2\,\overline{F}_Z(\overline{x},\overline{y},\overline{z}) = 0,$$

where $\overline{F}$ is the cubic obtained from F by reducing the coefficients of F modulo p (so that $\overline{F} = 0$ is a homogeneous equation for the elliptic curve $\overline{E}$). But the line defined by the homogeneous linear equation

$$X\overline{F}_X(\overline{x},\overline{y},\overline{z}) + Y\overline{F}_Y(\overline{x},\overline{y},\overline{z}) + Z\overline{F}_Z(\overline{x},\overline{y},\overline{z}) = 0$$

is the tangent line to $\overline{E}$ at $(\overline{x} : \overline{y} : \overline{z})$; hence $S_p(R) = (0 : (\overline{x'})^2\overline{y''} : (\overline{x'})^2\overline{z''}) = (0 : \overline{y''} : \overline{z''})$ lies on the tangent line, as desired.

Since the group operations on E and $\overline{E}$ are defined in terms of chords and tangent lines, the above remarks easily imply that $S_p : E(\mathbf{Q}) \to \overline{E}(\mathbf{F}_p)$ is a homomorphism of groups (the details are left to the reader). □

The following theorem is the key to computing the torsion subgroup of an elliptic curve over $\mathbf{Q}$.

Theorem 16.4.2 *Let $E/\mathbf{Q}$ be an elliptic curve, suppose p is a prime not dividing the discriminant. If n is not divisible by p, then $S_p : E(\mathbf{Q}) \to \overline{E}(\mathbf{F}_p)$ is injective on the subgroup ${}_nE$.*

Proof: To do this, we study the structure of the kernel of the specialization map $S_p : E(\mathbf{Q}) \to \overline{E}(\mathbf{F}_p)$. The identity element $(0 : 1 : 0) \in E(\mathbf{Q})$ specializes to the identity of $\overline{E}(\mathbf{F}_p)$. Hence the kernel of specialization, which we will denote by G, consists of the subgroup of $E(\mathbf{Q})$ consisiting of points which have the same specialization as $(0 : 1 : 0)$. In particular, any point of G has the form $(x : y : z)$ where $x, y, z \in \mathbf{Z}$, y is not divisible by p, and $x, z \in p\mathbf{Z}$. Hence all these points lie in the affine open set in $\mathbf{P}^2(\mathbf{Q})$ given by $Y \neq 0$, on which we may choose affine coordinates $x = X/Y$, $z = Z/Y$; the points of G consist of those points of $E(\mathbf{Q})$ whose coordinates x, z lie in $p\mathbf{Z}_{(p)}$, where

$$\mathbf{Z}_{(p)} = \{a/b \in \mathbf{Q} \mid p \not| \ b\}.$$

Note that $\mathbf{Z}_{(p)}$ is a discrete valuation ring with maximal ideal $p\mathbf{Z}_{(p)}$, and residue field $\mathbf{F}_p$.

Let $G_r \subset G$ be the subset of points whose affine coordinates x, z lie in $p^r \mathbf{Z}_{(p)}$. Now x, z satisfy the cubic polynomial

$$z = x^3 + axz^2 + bz^3 \qquad (**)$$

If $x, z \in p^r \mathbf{Z}_{(p)}$, and r is the largest such integer, the right side of the equation (**) lies in $p^{3r}\mathbf{Z}_{(p)}$, so that the equation implies that $z \in p^{3r}\mathbf{Z}_{(p)}$; thus $x = p^r x'$, $z = p^{3r} z'$ where x', z' are units in $\mathbf{Z}_{(p)}$. In particular, for any positive integer r, we have that $(x, z) \in G_r \Leftrightarrow x \in p^r \mathbf{Z}_{(p)}$.

In terms of the affine coordinates x, z the group law of the elliptic curve is described as follows: given distinct points (x_1, z_1) and (x_2, z_2) satisfying the cubic (**), the line joining them is $(z_2 - z_1)(x - x_1) = (x_2 - x_1)(z - z_1)$. Assuming $x_2 - x_1 \neq 0$, we may write this as

$$z = mx + m'$$

where

$$m = \frac{z_2 - z_1}{x_2 - x_1}, \quad m' = \frac{z_1 x_2 - z_2 x_1}{x_2 - x_1}.$$

Substituting for z in (**), we get

$$mx + m' = x^3 + ax(mx + m')^2 + b(mx + m')^3$$

which we may rewrite as

$$x^3 + \frac{2amm' + 3bmm'^2}{1 + am^2 + bm^3}x^2 + \quad (\text{terms of degree} \leq 1 \text{ in } x) = 0,$$

so that the third root of this cubic is

$$-x_1 - x_2 - \frac{2amm' + 3bmm'^2}{1 + am^2 + bm^3}.$$

Hence if $(x_1, z_1) + (x_2, z_2) = (x_3, z_3)$ and $x_1 \neq x_2$, then

$$x_3 = x_1 + x_2 + \frac{2amm' + 3bmm'^2}{1 + am^2 + bm^3},$$

where m, m' are as above. In the case $x_1 = x_2$, with $(x_i, z_i) \in G$, we will see below that $z_1 = z_2$. One can work out that if $(x_3, z_3) = 2(x_1, z_1)$ then

$$x_3 = \left(2x_1 + a\frac{z_1 f_z(x_1, z_1) + x_1 f_x(x_1, z_1)}{f_z(x_1, z_1)}\right) \times$$

$$\left(1 + a\frac{-f_x(x_1,z_1)}{f_z(x_1,z_1)} + b\frac{f_x(x_1,z_1)^2}{f_z(x_1,z_1)^2} + c\frac{-f_x(x_1,z_1)^3}{f_z(x_1,z_1)^3}\right)^{-1}$$

where $f_x = 3x^2 + az^2$, $f_z = -1 + 2axz + 3bz^2$ are the partial derivatives of $f = -z + x^3 + axz^2 + bz^3$.

We now claim that $G_r \subset G$ is a subgroup for each $m > 0$. Since $(x,z) \in G$ lies in $G_r \Leftrightarrow x \in p^r\mathbf{Z}_{(p)}$, it suffices to show that if $(x_1,z_1),(x_2,z_2) \in G_r$, then in the above formulas for x_3, we have $x_3 \in p^r\mathbf{Z}_{(p)}$.

From the equations

$$z_i = x_i^3 + ax_iz_i^2 + bz_i^3, \quad i = 1,2$$

we obtain (after subtraction and rearrangement)

$$(z_2 - z_1)\left(1 - b(z_1^2 + z_1z_2 + z_3^2) - ax_1(z_2 + z_1)\right) = \\ (x_2 - x_1)\left((x_1^2 + x_1x_2 + x_2^2) + az_2^2\right).$$

Since $x_i \in p^r\mathbf{Z}_{(p)}$, and $z_i \in p^{3r}\mathbf{Z}_{(p)}$, where $r > 0$, the coefficient of $(z_2 - z_1)$ on the left is a unit, while the coefficient of $(x_2 - x_1)$ on the right lies in $p^{2r}\mathbf{Z}_{(p)}$. Hence we first see that if $x_1 = x_2$, then $z_1 = z_2$. Next, if $x_1 \neq x_2$, then

$$m = \frac{z_2 - z_1}{x_2 - x_1} \in p^{2r}\mathbf{Z}_{(p)}.$$

From

$$z_1 = x_1m + m'$$

we then see that

$$m' \in p^{3r}\mathbf{Z}_{(p)}.$$

Finally, we see that $1 + am^2 + bm^3$ is a unit, so that

$$\frac{2amm' + 3bmm'^2}{1 + am^2 + bm^3} \in p^{5r}\mathbf{Z}_{(p)}.$$

These remarks imply that $x_3 \in p^r\mathbf{Z}_{(p)}$, and

$$x_3 \equiv x_1 + x_2 \pmod{p^{r+1}\mathbf{Z}_{(p)}}.$$

Similar considerations show that if $(x,z) = 2(x_1,z_1)$ with $(x_1,z_1) \in G_r$, then $x \in p^r\mathbf{Z}_{(p)}$, and

$$x \equiv 2x_1 \pmod{p^{r+1}\mathbf{Z}_{(p)}}.$$

In particular, since $\mathbf{F}_p$ has characteristic p, we have $p(x_1, z_1) = (x, y)$ with $x \equiv px_1 (\mathrm{mod} p^{r+1}\mathbf{Z}_{(p)})$, which gives $x \in p^{r+1}\mathbf{Z}_{(p)}$, since $p \in p\mathbf{Z}_{(p)}$. Hence for any $R \in G_r$, $pR \in G_{r+1}$. On the other hand, clearly

$$\bigcap_{r \geq 0} G_r = \{0\},$$

the group consisting of only the identity. Thus, the only possible torsion present in G is p-primary torsion. In particular, since n is relatively prime to p, G has no n-torsion, *i.e.*, the specialization map S_p is injective on ${}_nE$. This completes the proof of the Theorem. □

A further study of the filtration G_r leads to the notion of the *formal group* associated to the elliptic curve; see [57] for details.

Exercise 16.4.1: Let R be a discrete valuation ring with quotient field K and residue field k.

(i) Show that there is a well defined specialization mapping $S_R : \mathbf{P}^n(K) \to \mathbf{P}^n(k)$.

(ii) Let $E \subset \mathbf{P}^2(K)$ be an elliptic curve defined by an equation $Y^2Z - X^3 - aXZ^2 - bZ^3 = 0$ with $a, b \in R$, such that the discriminant of the cubic $x^3 + ax + b$ is a unit in R. Let $\overline{E}$ be the elliptic curve over k defined by reducing the above equation modulo the maximal ideal of R. Show that $S_R : \mathbf{P}^2(K) \to \mathbf{P}^2(k)$ induces a mapping $S_R : E(K) \to E(k)$ which is a homomorphism.

(iii) In the situation of (ii), suppose k has characteristic 0. Show that S_R is injective on the torsion subgroup of $E(K)$.

(iv) In the situation of (ii), suppose k has characteristic $p > 0$. Show that S_R is injective on the subgroup of $E(K)$ of points of finite order relatively prime to p.

(Hint: Let πR be the maximal ideal of R, where $\pi \in R$ is a local parameter. In (ii), (iii), (iv), the role played by the prime number p in the proof of Theorem 16.4.1 is now played by the local parameter π. For example, to prove (ii), we must show that if $P \neq Q$ and $S_R(P) = S_R(Q)$, then the chord PQ specializes to the tangent at $S_R(P)$; to do this, one makes an expansion

of a value of the cubic polynomial f in powers of π, using Taylor's theorem. For (iii), (iv), the filtration $\{G_r\}$ of $G = \ker S_R : E(K) \to E(k)$ is defined as above in terms of the power of π dividing x; one shows that the map $G_r/G_{r+1} \to k$ given by $(x,z)(\mathrm{mod} G_{r+1}) \longmapsto x(\mathrm{mod}\pi R)$ is an injective map of abelian groups.)

16.4.2 Mazur's theorem

Let E be an elliptic curve defined over $\mathbf{Q}$. According to the Mordell-Weil theorem, the torsion subgroup $E(\mathbf{Q})_{tor}$ of the Mordell-Weil group $E(\mathbf{Q})$ is a finite group. Suppose $E(\mathbf{Q})_{tor}$ has exponent n *i.e.*, every element has order dividing n. Then $E(\mathbf{Q})_{tor}$ is a subgroup of the group ${}_nE(\mathbf{C})$ of n-torsion points of E over the complex numbers, and we know (from the geometry lectures) that ${}_nE(\mathbf{C}) \cong \mathbf{Z}/n\mathbf{Z} \times \mathbf{Z}/n\mathbf{Z}$. Thus

$$E(\mathbf{Q})_{tor} \cong \mathbf{Z}/a\mathbf{Z} \times \mathbf{Z}/b\mathbf{Z}$$

for some positive integers a, b.

The main result on the group $E(\mathbf{Q})_{tor}$ is the following important theorem of B. Mazur (see [37], [38]).

Theorem 16.4.3 *Let E be an elliptic curve defined over $\mathbf{Q}$. Then $E(\mathbf{Q})_{tor}$ satisfies:*

$$E(\mathbf{Q})_{tor} = \begin{cases} \mathbf{Z}/n\mathbf{Z} & \text{for } 1 \leq n \leq 10 \text{ or } n = 12; \\ \mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2n\mathbf{Z}, & \text{for } 1 \leq n \leq 4. \end{cases}$$

Further, each of these fifteen possible groups occur as the torsion subgroup of some elliptic curve E over $\mathbf{Q}$.

Silverman's book [57] contains the following examples, showing that all fifteen possible groups occur (see [57], Exercise 8.12).

(1) $y^2 = x^3 - 2$

(2) $y^2 = x^3 + 8$

(3) $y^2 = x^3 + 4$

(4) $y^2 = x^3 + 4x$

(5) $y^2 - y = x^3 - x^2$

(6) $y^2 = x^3 + 1$

(7) $y^2 + y = x^3 - x + 137$

(8) $y^2 + 7xy = x^3 + 16x$

(9) $y^2 + xy + y = x^3 - x^2 - 14x + 29$

(10) $y^2 + xy = x^3 - 45x + 81$

(11) $y^2 + 43xy - 210y = x^3 - 210x^2$

(12) $y^2 = x^3 - 4x$

(13) $y^2 + xy - 5y = x^3 - 5x^2$

(14) $y^2 + 5xy - 6y = x^3 - 3x^2$

(15) $y^2 + 17xy - 120y = x^3 - 60x^2$

As an exercise, the reader may find it amusing to compute the torsion subgroups in some of these cases.

The proof of Mazur's theorem requires tools from algebraic geometry and number theory which are far beyond the scope of this book. One needs to show that if $p \geq 11$ is a prime, then there does not exist any elliptic curve E defined over $\mathbf{Q}$ which has a $\mathbf{Q}$-rational point of order p. In Chapter (6), it was proved that the isomorphism class of an elliptic curve over $\mathbf{C}$ is determined by its *j-invariant*; further, if an elliptic curve is defined over a subfield $K \subset \mathbf{C}$, then its j-invariant lies in K. In particular, an elliptic curve defined over $\mathbf{Q}$ has a rational j-invariant. We also know that all possible values of the j-function occur as j-invariants of elliptic curves. Thus, the j-invariant allows us to consider the complex plane $\mathbf{C}$ as the parameter space for all isomorphism classes of elliptic curves. In fact, if we regard $\mathbf{C}$ as the set of complex points of the affine line $\mathbf{A}^1$, then we can regard $\mathbf{A}^1$ as a parameter *variety*, parametrizing all elliptic curves over $\mathbf{C}$ *i.e.*, $\mathbf{A}^1$ is what is called the *moduli space*[4]. In fact, this turns out to be true in the following strong sense:

[4]Actually, "coarse moduli space".

given an 'algebraic family of elliptic curves' parametrized[5] by an algebraic variety T, say $\{E_t\}_{t\in T}$, then the mapping $T \to \mathbf{A}^1$ given by $t \longmapsto j(E_t)$ is a morphism of algebraic varieties.

In a similar way, it is possible to construct a moduli variety parametrizing pairs (E, P) where E is an elliptic curve (with a given origin) and $P \in E(\mathbf{C})$ is a point of order n. This parameter variety maps to $\mathbf{A}^1$, the moduli of elliptic curves, by the map 'forget the point of order n' *i.e.*, the isomorphism class of (E, P) maps to that of E. The fibres are clearly finite, of cardinality equal to the number of elements of order n in $E(\mathbf{C})$ *i.e.*, in $\mathbf{Z}/n\mathbf{Z} \times \mathbf{Z}/n\mathbf{Z}$ (if $n = p$, a prime, this is p^2-1). Thus this moduli space is an affine curve, called a *modular curve*, and is denoted by $Y_1(n)$. It turns out that $Y_1(n)$ is defined over $\mathbf{Q}$, and is non-singular; the corresponding non-singular projective curve is denoted by $X_1(n)$, and the map $Y_1(n) \to \mathbf{A}^1$ given by the j-invariant extends to a morphism $X_1(n) \to \mathbf{P}^1$. As a Riemann surface, $Y_1(n)$ is the quotient $\mathcal{H}/\Gamma_1(n)$, where $\mathcal{H}$ is the upper half plane, and $\Gamma_1(n) \subset SL(2, \mathbf{Z})$ is the congruence subgroup

$$\Gamma_1(n) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mid c \equiv 0 \pmod{n},\ a \equiv d \equiv 1 \pmod{n} \right\}.$$

It turns out that if (E, P) is a pair consisting of an elliptic curve E defined over $\mathbf{Q}$, and $P \in {}_nE(\mathbf{Q})$, then the corresponding point of $Y_1(n)$ is a $\mathbf{Q}$-rational point. Mazur's result is that for any prime $p \geq 11$, $Y_1(p)$ has no $\mathbf{Q}$-rational points. Equivalently, the only $\mathbf{Q}$-rational points of $X_1(p)$ are the 'cusps' $X_1(p) - Y_1(p)$ lying over $\infty \in \mathbf{P}^1$.

Mazur proves this by considering another modular curve $Y_0(p)$, intermediate between $Y_1(p)$ and $Y_1(1) = \mathbf{A}^1$. The curve $Y_0(n)$ parametrizes pairs (E, G) where E is an elliptic curve over $\mathbf{C}$, and $G \subset E$ is a cyclic group isomorphic to $\mathbf{Z}/n\mathbf{Z}$. Again, $Y_0(n)$ is defined over $\mathbf{Q}$. If E is defined over $\mathbf{Q}$, and the Galois group $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ leaves G invariant (so that the 'subgroup variety' G is defined over $\mathbf{Q}$), then the pair (E, G) gives a rational point of $Y_0(n)$. Over $\mathbf{C}$, the Riemann surface $Y_0(n)$ is isomorphic to $\mathcal{H}/\Gamma_0(n)$, where

$$\Gamma_0(n) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mid c \equiv 0 \pmod{n} \right\}.$$

[5]We mean a morphism of algebraic varieties $f : X \to T$, with a section $s : T \to X$, such that each fibre $(E_t, s(t))$ is an elliptic curve with origin $s(t)$, such that the group operation 'varies algebraically' *i.e.*, is given by a morphism of varieties $X \times_T X \to X$.

There is an obvious mapping $Y_1(n) \to Y_0(n)$, given by $(E, P) \longmapsto (E, < P >)$ where $< P >$ is the cyclic subgroup generated by P. In particular, rational points of $Y_1(p)$ yield rational points of $Y_0(p)$ as well.

Another important geometric construction related to modular curves is the following. For any positive integer n, we have a natural mapping $f_n : Y_0(n) \to Y_0(1) = \mathbf{A}^1$, given by $(E, G) \longmapsto E$. There is another morphism $g_n : Y_0(n) \to Y_0(1) = \mathbf{A}^1$, given by $(E, G) \longmapsto E/G$, where E/G is the quotient elliptic curve ($E' = E/G$ is characterised upto isomorphism by the fact that there is an isogeny $E \to E'$ whose kernel is G). The morphisms f_n and g_n extend uniquely to morphisms of projective non-singular curves (denoted by the same symbols) $f_n : X_0(n) \to \mathbf{P}^1$, $g_n : X_0(n) \to \mathbf{P}^1$. Then the pair (f_n, g_n) yield a mapping $X_0(n) \to \mathbf{P}^1 \times \mathbf{P}^1$, whose image we denote by Γ_n. Then Γ_n is an irreducible projective curve on the surface $\mathbf{P}^1 \times \mathbf{P}^1$. We may think of Γ_n as the graph of a 'multivalued' algebraic function $\mathbf{P}^1 \to \mathbf{P}^1$; such a 'multivalued' function is called a *correspondence* from $\mathbf{P}^1$ to itself. Similarly, if p is a prime not dividing n, one may construct a correspondence $\Gamma_n(p) \subset X_0(p) \times X_0(p)$. This is called the n^{th} *Hecke correspondence*. These correspondences, for varying n, yield certain operators on modular forms, called *Hecke operators*, which we discuss below a bit further. The properties of these operators are an important ingredient in Mazur's proof.

For any projective non-singular curve C over an algebraically closed field k, there is an associated variety called the *Jacobian variety* $J(C)$ of the curve, which is an *abelian variety* (*i.e.*, a projective algebraic torus) of dimension equal to the genus g of C; if C is defined over a field k, this torus is also defined over k. The Jacobian variety has the property that its group of $\mathbf{C}$-valued points is just the group $\text{Pic}^0(C)$ of linear equivalence classes of divisors of degree 0 on the complex projective curve C.

A general property of any correspondence Γ between curves C_1 and C_2 is that that it induces a homomorphism of abelian varieties $[\Gamma]_* : J(C_1) \to J(C_2)$, where $J(C_i)$ is the Jacobian variety of C_i. Thinking of points of $J(C_1)$ as divisor classes, $[\Gamma]_*$ is loosely described as follows: if $p_1 : C_1 \times C_2 \to C_1$, $p_2 : C_1 \times C_2 \to C_2$ are the projections, then $[\Gamma]_*(D) = p_2(\Gamma \cap p_1^*(D))$. This is made more precise using *intersection theory* on the surface $C_1 \times C_2$.

In particular, there are endomorphisms $[\Gamma_n]_*$ of the Jacobian variety of $X_0(p)$ corresponding to the Hecke correspondences. One may identify the vector space of regular differentials $\Omega(J(X_0(p)))$ with the vector space of *cusp forms* of weight 2 for $\Gamma_0(p)$. Thus $[\Gamma_n]_*$ induces an endomorphism T_n of

the space of cusp forms of weight 2, called the n^{th} *Hecke operator*.

There are analogous Hecke operators acting on the spaces of cusp forms and Eisenstein series of arbitrary weight for $\Gamma_0(p)$. Some examples of cusp forms and Eisenstein series can be found in Chapter (18) The actions of these Hecke operators may be described explicitly using matrices (see Ogg's book [44]). However, we wished to emphasize here the geometric construction of these operators, which illustrates how they arise naturally from the point of view of moduli.

Mazur defines a certain quotient of the Jacobian variety of $X_0(p)$; the quotient torus, which he calls the *Eisenstein quotient*, is defined using a suitable space of cusp forms of weight 2 (*i.e.*, regular differentials on $J(X_0(p))$) whose Fourier coefficients are congruent to those of certain Eisenstein series.

The Mordell-Weil theorem, proved earlier for elliptic curves, can be generalised to show that if A is any abelian variety defined over $\mathbf{Q}$, then the group of $\mathbf{Q}$-rational points $A(\mathbf{Q})$ of the variety is a finitely generated abelian group; we call it the Mordell-Weil group of A. The main step in the proof of Mazur's theorem is the assertion that the Mordell-Weil group of the Eisenstein quotient is *finite*. Mazur's proof of this uses the full machinery of 'abstract' algebraic geometry, including group schemes, flat cohomology, p-adic representations etc.

Chapter 17

Integer points, and the theorems of Thue and Siegel

17.1 Siegel's theorem on integer points of elliptic curves

We wish to consider the problem of finding integer solutions to a cubic Diophantine equation

$$Y^2 = aX^3 + bX^2 + cX + d,$$

where a, b, c, d are integers with $a \neq 0$, such that the cubic polynomial has distinct roots. Equivalently, we wish to determine the set of *integral points* (*i.e.*, points with integer coordinates) on the affine plane curve E defined by the above cubic equation.

We will discuss the following theorem, which is a special case of an important result of C. L. Siegel, proved in a paper in 1929 (see [56]).

Theorem 17.1.1 *Let E be defined as above. Then the set*

$$E(\mathbf{Z}) = \left\{(x, y) \in \mathbf{Z}^2 \mid y^2 = ax^3 + bx^2 + cx + d\right\}$$

is finite.

It will not be possible for us to give a full proof of this result in this book, since the proof requires more background in algebraic number theory than we have assumed. However, we will sketch the proof of the reduction of Siegel's theorem to the solution of *Thue's equation* over certain algebraic number fields.

Recall that an *algebraic integer* is a complex number x which is integral over the ring $\mathbf{Z}$ of integers, *i.e.*, x is a root of a monic polynomial with integer coefficients. If K is a number field, *i.e.*, a finite algebraic extension of $\mathbf{Q}$, then the subset of K consisting of algebraic integers forms a subring $\mathcal{O}_K$, called the ring of algebraic integers of K. The result on the Thue equation which is needed for the proof of Siegel's theorem is the following.

Theorem 17.1.2 *Let a, b, c be algebraic integers in an algebraic number field K, with $c \neq 0$. Then the equation*

$$aU^3 + bV^3 = c$$

has only finitely many algebraic integer solutions $U, V \in K$.

A proof of this result in the above generality would involve more background in number theory than we have assumed. However, the proof of this theorem for the case $K = \mathbf{Q}$ is more elementary, and contains the essential new ideas; this will be proved in Section (17.2.1).

Our goal here is to explain how Theorem (17.1.1) may be deduced from Theorem 17.1.2. For simplicity, we will further assume that the cubic

$$f(X) = aX^3 + bX^2 + cX + d$$

factorises as

$$f(X) = a(X - u)(X - v)(X - w)$$

for some (distinct) integers u, v, w. Now suppose $(x, y) \in \mathbf{Z}^2$ satisfies

$$y^2 = a(x - u)(x - v)(x - w).$$

Any common factor of $x - u$ and $x - v$ must divide $u - v$, etc. Hence we may write

$$x - u = u_1 u_2^2, \quad x - v = v_1 v_2^2, \quad x - w = w_1 w_2^2$$

for some integers u_i, v_i, w_i, where u_1, v_1, w_1 are all square free, and divide

$$a(u - v)(v - w)(w - u).$$

This is similar to the 'descent' step in the proof of the Weak Mordell-Weil Theorem; making the 'descent' (*i.e.*, finding pre-images of (x, y) under an appropriate isogeny) amounts to finding square roots of u_1, v_1, w_1 (see (15.1.2)).

Let $K = \mathbf{Q}(\sqrt{u_1}, \sqrt{v_1}, \sqrt{w_1})$. Let $\mathcal{O}_K$ be the ring of algebraic integers in K. Then we may write

$$u_1 = u_3^2, \;\; v_1 = v_3^2, \;\; w_1 = w_3^2$$

for some algebraic integers $u_3, v_3, w_3 \in \mathcal{O}_K$. We may then write

$$v - u = u_2^2 u_3^2 - v_2^2 v_3^2 = (u_2 u_3 - v_2 v_3)(u_2 u_3 + v_2 v_3),$$

$$w - v = v_2^2 v_3^2 - w_2^2 w_3^2 = (v_2 v_3 - w_2 w_3)(v_2 v_3 + w_2 w_3),$$

$$u - w = w_2^2 w_3^2 - u_2^2 u_3^2 = (w_2 w_3 - u_2 u_3)(w_2 w_3 + u_2 u_3).$$

Let

$$\theta_1 = u_2 u_3 - v_2 v_3, \;\; \theta_2 = v_2 v_3 - w_2 w_3, \;\; \theta_3 = w_2 w_3 - u_2 u_3.$$

Then clearly $\theta_1 + \theta_2 + \theta_3 = 0$, and $\theta_1, \theta_2, \theta_3$ all divide $(u-v)(v-w)(w-u)$ in $\mathcal{O}_K$.

We now make use of the fact, proved in the Appendix (A), that $\mathcal{O}_K$ is a *Dedekind domain*. A basic property of Dedekind domains is the unique factorisation of ideals into products of prime ideals. We apply this to the θ_i, as follows. Let $J = (u-v)(v-w)(w-u)\mathcal{O}_K$, and let $J = \mathcal{P}_1^{r_1}\mathcal{P}_2^{r_2}\cdots\mathcal{P}_m^{r_m}$ be its prime factorisation. Consider the finite set of factor ideals of J which are principal; fix a generator for each of these principal ideals. This determines a finite set $T \subset \mathcal{O}_K$ such that

$$\theta_i = x_i \varepsilon_i, \;\; i = 1, 2, 3$$

where $x_i \in T$, and $\varepsilon_i \in \mathcal{O}_K^*$ are units.

Another basic result from number theory is the theorem that the group $\mathcal{O}_K^*$ of units is finitely generated; in fact, the *Dirichlet's unit theorem* gives a formula for the rank of this group. In Appendix (A), we sketch a proof of the finite generation, which is all we need.

This implies that we can find another finite set T' so that $\theta_i = x_i \varepsilon_i^3$, where $x_i \in T'$, and $\varepsilon_i \in \mathcal{O}_K^*$. This is because the finite generation of $\mathcal{O}_K^*$ implies that $\mathcal{O}_K^*/\mathcal{O}_K^{*3} \cong \mathcal{O}_K^* \otimes_{\mathbf{Z}} \mathbf{Z}/3\mathbf{Z}$ is finite; we take T' to be the set of all products xy, where $x \in T$, and y runs over a (finite) chosen set of coset representatives for $\mathcal{O}_K^* \otimes \mathbf{Z}/3\mathbf{Z}$.

Now the relation $\theta_1 + \theta_2 + \theta_3 = 0$ becomes

$$x_1 \varepsilon_1^3 + x_2 \varepsilon_2^3 + x_3 \varepsilon_3^3 = 0,$$

which we may rewrite as

$$AU^3 + BV^3 = C,$$

where

$$A = x_1, \;\; B = x_2, \;\; U = \varepsilon_1\varepsilon_3^{-1}, \;\; V = \varepsilon_2\varepsilon_3^{-1}, \;\; C = -x_3.$$

Here (A, B, C) ranges over a finite set, and U, V are units, so in particular, are algebraic integers. By Theorem 17.1.2, there are only finitely many algebraic integer solutions (U, V) for each choice of (A, B, C). Thus there are only finitely many possibilities for the ratios $\varepsilon_1\varepsilon_3^{-1}, \varepsilon_2\varepsilon_3^{-1}$, and hence for the ratios $\theta_1/\theta_3, \theta_2/\theta_3$.

In a similar vein, let

$$\theta_1' = u_2u_3 + v_2v_3, \;\; \theta_2' = v_2v_3 + w_2w_3.$$

Since $\theta_3 = w_2w_3 - u_2u_3$, we see that

$$\theta_1' - \theta_2' + \theta_3 = 0;$$

also θ_1', θ_2' are divisors of $(u-v)(v-w)(w-u)$. As above, we can then write

$$\theta_1' = x_1'\varepsilon_1'^3, \;\; \theta_2' = x_2'\varepsilon_2'^3$$

where $x_i' \in T'$, and ε_i' are units. Then

$$x_1'(\varepsilon_1')^3 - x_2'(\varepsilon_2')^3 + x_3\varepsilon_3^3 = 0,$$

so that we obtain a solution of

$$AU^3 + BV^3 = C,$$

with

$$A = x_1', B = -x_2', C = -x_3, U = \varepsilon_1'\varepsilon_3^{-1}, V = \varepsilon_2'\varepsilon_3^{-1}.$$

Again, we deduce that there are only finitely many possibilities for the ratios $\theta_1'/\theta_3, \theta_2'/\theta_3$.

If we fix values for $\theta_1/\theta_3, \theta_1'/\theta_3$, then θ_1/θ_1' is fixed; since $\theta_1\theta_1' = v - u$, we have fixed the value of θ_1^2. Hence there are only finitely many possibilites for θ_1, θ_1' and hence for $(\theta_1 + \theta_1')^2 = 4u_1u_2^2 = 4(x-u)$, and so there are only finitely many possibilities for x. This completes the proof of Siegel's theorem.

Exercise 17.1.1: Assuming Theorem 17.1.2 on the Thue equation, prove the following generalisation of Theorem 17.1.1: if $f(x) \in \mathbf{Z}[x]$ such that f has at least 3 distinct integer roots occuring with odd multiplicities, then $y^2 = f(x)$ has only a finite number of integer solutions.

17.2 Approximations of algebraic numbers by rationals: a theorem of Thue

17.2.1 Introduction

Let α be an algebraic number. Then, we observe that α satisfies the polynomial of minimal degree such that it has relatively prime integer coefficients and the leading coefficient positive. This is the *minimal polynomial* of α. The degree of this polynomial is called the *degree of* α. The maximum of the absolute values of the coefficients of this polynomial is called the *height of* α. We write $\nu = \nu(\alpha)$ for the least positive integer such that $\nu\alpha$ is an algebraic integer ($\nu\alpha$ satisfies a monic polynomial with integer coefficients). The integer ν exists and it is called the *denominator of* α.

We start with the following theorem of Liouville [35] on the approximations of algebraic numbers by rationals.

Theorem 17.2.1 *Let α be an algebraic number of degree $n \geq 2$. There exists a constant $c = c(\alpha) > 0$ such that*

$$\left|\alpha - \frac{p}{q}\right| > \frac{c}{q^n} \quad (p, q \in \mathbf{Z} \ \ with \ \ q > 0) \tag{17.1}$$

The inequality (17.1) will be referred to as Liouville's inequality. Any non-trivial sharpening of (17.1) implies immediately that certain equations have only finitely many integral solutions. For instance, let $x > 0$ and $y > 0$ be integers satisfying

$$x^3 - 2y^3 = 1 \tag{17.2}$$

We write

$$\rho = 2^{1/3} \in \mathbf{R}.$$

We observe that

$$= (x - \rho y)(x^2 + \rho^2 y^2 + \rho xy) > (x - \rho y)\rho^2 y^2 > (x - \rho y)y^2.$$

Therefore

$$x - \rho y < \frac{1}{y^2}$$

which implies that

$$\left|\rho - \frac{x}{y}\right| < \frac{1}{y^3}.$$

Let $F(q)$ be a positive real valued function tending to infinity with $q > 0$ such that

$$\left|\rho - \frac{p}{q}\right| > \frac{F(q)}{q^3} \qquad (p, q \in \mathbf{Z} \text{ with } q > 0).$$

This is what we meant by any non-trivial improvement of (17.1) with $\alpha = 2^{1/3}$. Then

$$\frac{F(y)}{y^3} < \left|\rho - \frac{x}{y}\right| < \frac{1}{y^3}.$$

Thus $F(y) < 1$ which implies that y is bounded. Now, we observe from (17.2) that x is bounded. *Hence, equation*(17.2) *has only finitely many solutions in integers* $x > 0$ *and* $y > 0$. On the other hand, we recall that $x^2 - 2y^2 = 1$ has infinitely many solutions in integers $x > 0$ and $y > 0$.

Thue [58] introduced a fundamental method to establish the first non-trivial sharpening of Liouville's inequality.

Theorem 17.2.2 *Let* α *be an algebraic number of degree* $n \geq 3$. *Let* $\varepsilon > 0$ *and* $\kappa = \frac{n}{2} + 1 + \varepsilon$. *There exists a constant* $c_1 = c_1(\alpha, \varepsilon) > 0$ *such that*

$$\left|\alpha - \frac{p}{q}\right| > \frac{c_1}{q^\kappa} \qquad (p, q \in \mathbf{Z} \text{ with } q > 0) \tag{17.3}$$

We observe that $\kappa < n$ whenever $0 < \varepsilon < 1/2$ and then, $F(q) = c_1 q^{n-\kappa}$ tends to infinity with $q > 0$. Further, we may rewrite (17.3) as

$$\left|\alpha - \frac{p}{q}\right| > \frac{F(q)}{q^n} \qquad (p, q \in \mathbf{Z}, \text{with } q > 0)$$

to observe that (17.3) is a non-trivial improvement of (17.1). The inequality (17.3) has been improved as follows:

$\kappa = 2\sqrt{n} + \varepsilon$ Siegel [55]

$\kappa = \sqrt{2n} + \varepsilon$, { Dyson [18] / Gel'fond [24] } independently, and

$\kappa = 2 + \varepsilon$. Roth [46]

Roth's inequality is not far off from the best possible in view of the following well-known result of Dirichlet:
Let $\alpha \notin \mathbf{Q}$ be a real number. There are infinitely many pairs $(p, q) \in \mathbf{Z}^2$ with $q > 0$ and $\gcd(p, q) = 1$ such that

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^2}.$$

This also shows that Liouville's inequality (17.1) with $n = 2$ is the best possible apart from the constant c. For proof, see Schmidt [48], page 2. On the other hand, it is an open problem to replace ε in Roth's inequality by a function of q tending to zero as $q \to \infty$.

Finally, as above, we derive from Theorem 17.2.2 the following extension of the result that (17.2) has only finitely many solutions in integers $x > 0$ and $y > 0$.

Theorem 17.2.3 *Let $f(X, Y) \in \mathbf{Z}[X, Y]$ be an irreducible binary form (homogeneous polynomial) of degree ≥ 3 and let $m \in \mathbf{Z}$. The equation*

$$f(x, y) = m \quad in \quad x, y \in \mathbf{Z} \tag{17.4}$$

has only finitely many solutions.

The equation (17.4) is known as *Thue's equation.* For an integer μ with $|\mu| > 1$, we write $P(\mu)$ for the greatest prime factor of μ and we put $P(\pm 1) = 1$. Theorem 17.2.3 states that

$$|f(x, y)| \to \infty \text{ as } \max(|x|, |y|) \to \infty \text{ with } x, y \in \mathbf{Z}.$$

Mahler [36] extended this result by proving that

$$P(f(x, y)) \to \infty \text{ as } \max(|x|, |y|) \to \infty \text{ with } x, y \in \mathbf{Z} \text{ and } \gcd(x, y) = 1.$$

The under-lying equation for this assertion is known as *Thue-Mahler equation.* For the connection between Thue and Thue-Mahler equations and elliptic curves, see Baker [4], Baker [5], Coates [9], Aggarwal *et al.* [1],. Finally, we remark that Theorem (17.2.3) is basic for the study of integeral points on curves; see Shorey and Tijdeman [54], Chapters 5-8.

17.2.2 Two lemmas

In this section, we prove two lemmas for the proof of Theorem 17.2.2. The first one is the well-known Thue-Siegel lemma and it provides a basic construction essential in Diophantine approximations and Transcendence. The proof depends on the Box principle.

Lemma 17.2.1 *Let $a_{i,j} \in \mathbf{Z}$ with $1 \le i \le q, 1 \le j \le p$ and $A \ge \max_{i,j} |a_{i,j}|$. If $q > p$, the system*

$$\sum_{i=1}^{q} a_{i,j} x_i = 0, \qquad 1 \le j \le p, \tag{17.5}$$

has a non-trivial solution in $x_1, \cdots, x_q \in \mathbf{Z}$ satisfying

$$|x_i| \le (2qA)^{p/(q-p)} \quad for \quad 1 \le i \le q \tag{17.6}$$

For $\delta > 0$, we shall apply Lemma 17.2.1 with $q > (1+\delta)p$. Then $p/(q-p) < 1/\delta$ and we derive

Corollary 17.2.2 *Let $\delta > 0$. If $q > (1+\delta)p$, the system (17.5) has a non-trivial solution in $x_1, \cdots, x_q \in \mathbf{Z}$ satisfying*

$$|x_i| \le (2qA)^{1/\delta} \quad \text{for} \quad 1 \le i \le q.$$

Proof of Lemma 17.2.1: For any positive integer X, we put

$$A = \{(x_1, \cdots, x_q) \in \mathbf{Z}^q \mid 0 \le x_i \le X\}$$

and

$$B = \left\{ (y_1, \cdots, y_p) \in \mathbf{Z}^p \mid \ y_j = \sum_{i=1}^{q} a_{i,j} x_i \text{ for } 1 \le j \le p \text{ and } (x_1, \cdots, x_q) \in A \right\}.$$

We have

$$f : A \longrightarrow B$$

given by

$$f((x_1, \cdots, x_q)) = (y_1, \cdots, y_p).$$

Every element $(y_1, \cdots, y_p)$ of B satisfies

$$|y_j| \le qAX \quad \text{for} \quad 1 \le j \le p.$$

Thus

$$|B| \leq (2qAX+1)^p$$

and

$$|A| = (X+1)^q$$

where $|B|$ and $|A|$ denote the number of elements in the sets B and A, respectively. We fix

$$X = \left[(2qA)^{p/(q-p)}\right],$$

the greatest integer not exceeding $(2qA)^{p/(q-p)}$. Then

$$X+1 > (2qA)^{p/(q-p)}$$

i.e.,

$$(X+1)^{q-p} > (2qA)^p$$

and hence

$$|A| = (X+1)^p(X+1)^{q-p} > (X+1)^p(2qA)^p > (2qAX+1)^p \geq |B|\,.$$

Thus f is not injective. Therefore, there exist distinct elements $(x_1', \cdots, x_q')$ and $(x_1'', \cdots, x_q'')$ of A such that

$$\sum_{i=1}^{q} a_{i,j}x_i' = \sum_{i=1}^{q} a_{i,j}x_i'' \quad \text{for} \quad 1 \leq j \leq p.$$

Finally, we take

$$x_i = x_i' - x_i'' \quad \text{for} \quad 1 \leq i \leq q$$

to conclude that $x_1, \cdots, x_q$ is a non-trivial integral solution of (17.5) satisfying (17.6). □

Let α be an algebraic integer of degree $n \geq 2$ and height H. For $s = 0, 1, 2, \cdots$, we write

$$\alpha^s = \sum_{j=0}^{n-1} b_{j,s}\alpha^j \tag{17.7}$$

where $b_{j,s} \in \mathbf{Z}$. The coefficients $b_{j,s}$ are estimated in the next lemma.

Lemma 17.2.3

$$\max_{0 \leq j < n} |b_{j,s}| \leq (2H)^s \quad \textit{for} \quad s = 0, 1, 2, \cdots$$

Proof: Let

$$x^n + A_1 x^{n-1} + \cdots + A_n$$

be the minimal polynomial of α. We prove the lemma by induction on s. If $s = 0$, we observe from (17.7) that $b_{0,0} = 1, b_{j,0} = 0$ for $j > 0$ and the estimate of the lemma follows. By (17.7), we have

$$\alpha^{s+1} = \alpha \sum_{j=0}^{n-1} b_{j,s}\alpha^j = \sum_{j=1}^{n} b_{j-1,s}\alpha^j.$$

Further

$$b_{n-1,s}\alpha^n = -b_{n-1,s} \sum_{j=0}^{n-1} A_{n-j}\alpha^j.$$

Therefore

$$\alpha^{s+1} = \sum_{j=0}^{n-1} b_{j,s+1}\alpha^j$$

where

$$b_{0,s+1} = -b_{n-1,s}A_n$$

and

$$b_{j,s+1} = b_{j-1,s} - A_{n-j}b_{n-1,s} \quad \text{for} \quad 1 \leq j < n.$$

Thus

$$\max_{0\leq j<n} |b_{j,s+1}| \leq 2H \max_{0\leq j<n} |b_{j,s}| \quad \text{for} \quad s = 0, 1, 2, \cdots$$

Now, we apply the induction hypothesis to the right hand side of the above inequality to conclude that

$$\max_{0\leq j<n} |b_{j,s+1}| \leq (2H)^{s+1} \text{ for } s = 0, 1, 2, \ldots$$

□

17.2.3 Proof of Theorem 17.2.1

We may assume that α is a real number otherwise (17.2.1) follows immediately. For later reference in the proof of Theorem 17.2.2, we split the proof of Theorem 17.2.1 in four steps.

Step I Take the minimal polynomial $f(x)$ of α.

Step II We may assume that

$$\left|\alpha - \frac{p}{q}\right| < 1.$$

Then

$$\left|f\left(\frac{p}{q}\right)\right| = \left|f(\alpha) - f\left(\frac{p}{q}\right)\right| \leq c_2 \left|\alpha - \frac{p}{q}\right| \tag{17.8}$$

where $c_2 = c_2(\alpha) > 0$ is a constant.

Step III We observe that $f(p/q) \neq 0$, since α has degree ≥ 2. Further, we notice that $f(p/q)$ is a non-zero rational number with denominator not exceeding q^n. Therefore

$$\left|f\left(\frac{p}{q}\right)\right| \geq \frac{1}{q^n} \tag{17.9}$$

Step IV We combine (17.8) and (17.9) to derive (17.1) with $c = \min(1, c_2^{-1})$. □

17.2.4 Proof of Theorem17.2.2

As in the proof of Theorem 17.2.1, we may assume that α is a real number. We write

$$\left|\alpha - \frac{p}{q}\right| = \frac{1}{\nu}\left|\nu\alpha - \frac{\nu p}{q}\right|$$

to observe that we may suppose that α is an algebraic integer. Further, it suffices to prove (17.3) for all reduced fractions p/q with $q > 0$. The proof of Theorem 17.2.1 depends on the minimal polynomial of α. Thue constructed a polynomial in two variables and extended the arguments of Theorem 17.2.1 to this polynomial. In the subsequent work of Siegel, Dyson and Gel'fond, more general polynomials in two variables are considered. Finally, Roth developed Thue's method for polynomials in several variables.

Step I This is Thue's basic construction for a suitable polynomial and the construction depends on Lemma 17.2.1. For a polynomial $R(x, y)$, we write

$$R_m(x, y) = \frac{1}{m!}\frac{\partial^m}{\partial x^m} R(x, y) \tag{17.10}$$

Lemma 17.2.4 *Let $\delta > 0$. Let L and k be positive integers satisfying*

$$2(L+1) > (1+\delta)nk \tag{17.11}$$

There exists a non-zero polynomial

$$R(x,y) = \sum_{\lambda_1=0}^{L} \sum_{\lambda_2=0}^{1} p(\lambda_1,\lambda_2) x^{\lambda_1} y^{\lambda_2} \in \mathbf{Z}[x,y] \tag{17.12}$$

satisfying

$$R_m(\alpha,\alpha) = 0 \quad for \quad 0 \le m < k \tag{17.13}$$

and

$$\max_{(\lambda_1,\lambda_2)} |p(\lambda_1,\lambda_2)| \le u^L \tag{17.14}$$

where $u = u(\alpha,\delta) > 0$ is a constant.

Proof: For $R(x,y)$ given by (17.12), we derive from Lemma 17.2.3 that

$$\begin{aligned} R_m(\alpha,\alpha) &= \sum_{\lambda_1=0}^{L} \sum_{\lambda_2=0}^{1} p(\lambda_1,\lambda_2) \binom{\lambda_1}{m} \alpha^{\lambda_1+\lambda_2-m} \\ &= \sum_{\lambda_1=0}^{L} \sum_{\lambda_2=0}^{1} p(\lambda_1,\lambda_2) \binom{\lambda_1}{m} \sum_{j=0}^{n-1} b_{j,\lambda_1+\lambda_2-m} \alpha^j \\ &= \sum_{j=0}^{n-1} \alpha^j \sum_{\lambda_1=0}^{L} \sum_{\lambda_2=0}^{1} p(\lambda_1,\lambda_2) \binom{\lambda_1}{m} b_{j,\lambda_1+\lambda_2-m} \end{aligned}$$

and

$$\max_{(\lambda_1,\lambda_2)} \left| \binom{\lambda_1}{m} b_{j,\lambda_1+\lambda_2-m} \right| \le 2^L (2H)^{L+1} \le (4H)^{2L} \quad \text{for} \quad m = 0,1,2,\cdots,$$

where H denotes the height of α. Now, since α is of degree n, the system (17.13) is equivalent to

$$\sum_{\lambda_1=0}^{L} \sum_{\lambda_2=0}^{1} p(\lambda_1,\lambda_2) \binom{\lambda_1}{m} b_{j,\lambda_1+\lambda_2-m} = 0 \quad (0 \le m < k, 0 \le j < n) \tag{17.15}$$

We apply Corollary 17.2.2 with $q = 2(L+1), p = nk$ and $A = (4H)^{2L}$ to the system of linear equations (17.15) in $p(\lambda_1,\lambda_2)$. We conclude from eq. (17.11) that there exist $p(\lambda_1,\lambda_2) \in \mathbf{Z}$ satisfying

$$0 < \max_{(\lambda_1,\lambda_2)} |p(\lambda_1,\lambda_2)| \le \left(4(L+1)(4H)^{2L}\right)^{1/\delta} \le (4H)^{4L/\delta} =: u^L.$$

□

Step II For two approximations p_1/q_1 and p_2/q_2 to α, we apply Taylor expansion to $R_m(x,y)$ at (α,α) for estimating $\left|R_m\left(\frac{p_1}{q_1},\frac{p_2}{q_2}\right)\right|$ from above.

Lemma 17.2.5 *Let δ, L and k be as in Lemma 17.2.4 satisfying* (17.11). *Let p_1/q_1 and p_2/q_2 be distinct reduced fractions with $q_1 > 0$ and $q_2 > 0$ satisfying*

$$\left|\alpha - \frac{p_i}{q_i}\right| < 1 \quad \text{for} \quad i = 1, 2 \tag{17.16}$$

There exists a constant $u_1 = u_1(\alpha,\delta) > 0$ such that

$$\left|R_m\left(\frac{p_1}{q_1},\frac{p_2}{q_2}\right)\right| \le \left(\left|\alpha - \frac{p_1}{q_1}\right|^{k-m} + \left|\alpha - \frac{p_2}{q_2}\right|\right) u_1^L \quad for \quad 0 \le m < k \tag{17.17}$$

Proof: We write u_2, u_3 and u_4 for positive constants depending only on α and δ. By (17.12), we have

$$R(x,y) = P(x) - yQ(x) \tag{17.18}$$

where

$$P(x) = \sum_{\lambda_1=0}^{L} p(\lambda_1, 0)x^{\lambda_1}, \quad Q(x) = -\sum_{\lambda_1=0}^{L} p(\lambda_1, 1)x^{\lambda_1} \tag{17.19}$$

Then, we observe from (17.18), (17.10) and (17.19) that

$$R_m(x,y) = P_m(x) - yQ_m(x) \tag{17.20}$$

where

$$P_m(x) = \frac{1}{m!}P^{(m)}(x), \quad Q_m(x) = \frac{1}{m!}Q^{(m)}(x) \tag{17.21}$$

are given by

$$P_m(x) = \sum_{\lambda_1=0}^{L} p(\lambda_1, 0)\binom{\lambda_1}{m}x^{\lambda_1-m}, \quad Q_m(x) = \sum_{\lambda_1=0}^{L} p(\lambda_1, 1)\binom{\lambda_1}{m}x^{\lambda_1-m} \tag{17.22}$$

Now, we apply (17.14) to estimate

$$|P_m(x)| \le (L+1)u^L 2^L(1+|x|)^L \le (u_2(1+|x|))^L \tag{17.23}$$

Similarly

$$|Q_m(x)| \leq (u_2(1+|x|))^L \tag{17.24}$$

Therefore, we derive from (17.20), (17.23) and (17.24) that

$$|R_m(x,y)| \leq |P_m(x)| + |y|\,|Q_m(x)| \leq (u_2(1+|x|))^L(1+|y|) \tag{17.25}$$

Further, we rewrite (17.20) with $x = \frac{p_1}{q_1}, y = \frac{p_2}{q_2}$ as

$$R_m\left(\frac{p_1}{q_1}, \frac{p_2}{q_2}\right) = P_m\left(\frac{p_1}{q_1}\right) - \alpha Q_m\left(\frac{p_1}{q_1}\right) - \left(\frac{p_2}{q_2} - \alpha\right) Q_m\left(\frac{p_1}{q_1}\right).$$

Thus

$$\left|R_m\left(\frac{p_1}{q_1},\ \frac{p_2}{q_2}\right)\right| \leq \left|R_m\left(\frac{p_1}{q_1}, \alpha\right)\right| + \left|\alpha - \frac{p_2}{q_2}\right| \left|Q_m\left(\frac{p_1}{q_1}\right)\right| \tag{17.26}$$

By (17.24) and (17.16), we observe that the second summand on the right hand side of (17.26) does not exceed

$$\left|\alpha - \frac{p_2}{q_2}\right| u_3^L \tag{17.27}$$

Now, we estimate the first summand on the right hand side of (17.26). Let $0 \leq m < k$. We put

$$S(x) = R_m(x,\alpha)\ ,\ S_\nu(\alpha) = \frac{1}{\nu!} S^{(\nu)}(\alpha).$$

By (17.10), we observe that

$$S_\nu(\alpha) = \binom{m+\nu}{\nu} R_{m+\nu}(\alpha,\alpha).$$

Then, we refer to (17.13) for deriving that

$$S_\nu(\alpha) = 0 \ \text{ for } \ \nu < k - m \tag{17.28}$$

Next, we apply Taylor expansion to obtain

$$S\left(\frac{p_1}{q_1}\right) = S\left(\frac{p_1}{q_1} - \alpha + \alpha\right) = \sum_{\nu=0}^{\infty} S_\nu(\alpha)\left(\frac{p_1}{q_1} - \alpha\right)^\nu.$$

By (17.28), we derive that

$$R_m\left(\frac{p_1}{q_1},\alpha\right) = \sum_{\nu=k-m}^{L-m} \binom{m+\nu}{\nu} R_{m+\nu}(\alpha,\alpha)\left(\frac{p_1}{q_1}-\alpha\right)^{\nu}$$

which, together with (17.25) and (17.16), implies that

$$\left|R_m\left(\frac{p_1}{q_1},\alpha\right)\right| \le u_4^L \left|\alpha-\frac{p_1}{q_1}\right|^{k-m} \tag{17.29}$$

Finally, we combine (17.26) , (17.27) and (17.29) to conclude (17.17). □

We observe from (17.17) that it is desirable to take L as small as possible. On the other hand, the inequality (17.11) must be satisfied. Therefore, from now onward, we always understand that L and k are positive integers given by

$$L = \left[\frac{1}{2}(1+\delta)nk\right] \tag{17.30}$$

so that (17.11) is satisfied. Consequently, the assertions of Lemmas 17.2.4 and 17.2.5 are valid.

Step III We prove the following lemma:

Lemma 17.2.6 *Let $0 < \delta < 1$ and L, k be positive integers given by* (17.30). *Let p_1/q_1 and p_2/q_2 with $q_1 > 1$ and $q_2 \ge 1$ be distinct reduced fractions. There exist $u_5 = u_5(\alpha,\delta) > 0$ and m satisfying*

$$\le m \le \left[\frac{u_5 k}{\log\, q_1}\right] + 2$$

such that

$$R_m\left(\frac{p_1}{q_1},\frac{p_2}{q_2}\right) \ne 0.$$

Proof: Let t be a positive integer. We suppose that

$$R_m\left(\frac{p_1}{q_1},\frac{p_2}{q_2}\right) = 0 \quad \text{for} \quad 0 \le m \le t \tag{17.31}$$

For $0 \leq m \leq t$ and $0 \leq m' \leq t$, we observe from (17.31), (17.20) and (17.21) that

$$P^{(m)}\left(\frac{p_1}{q_1}\right) - \frac{p_2}{q_2}Q^{(m)}\left(\frac{p_1}{q_1}\right) = 0$$

and

$$P^{(m')}\left(\frac{p_1}{q_1}\right) - \frac{p_2}{q_2}Q^{(m')}\left(\frac{p_1}{q_1}\right) = 0.$$

Eliminating p_2/q_2 in these equations, we obtain

$$P^{(m)}\left(\frac{p_1}{q_1}\right)Q^{(m')}\left(\frac{p_1}{q_1}\right) - P^{(m')}\left(\frac{p_1}{q_1}\right)Q^{(m)}\left(\frac{p_1}{q_1}\right) = 0 (0 \leq m \leq t, 0 \leq m' \leq t) \tag{17.32}$$

We put

$$W(x) = P(x)Q^{(1)}(x) - P^{(1)}(x)Q(x) \tag{17.33}$$

Then, we see from (17.33) and (17.32) that

$$W^{(\mu)}\left(\frac{p_1}{q_1}\right) = 0 \quad \text{for} \quad 0 \leq \mu < t$$

Therefore

$$W(x) = (q_1x - p_1)^t H(x) \tag{17.34}$$

where

$$H(x) \in \mathbf{Z}[x],$$

by Gauss Lemma and $\gcd(p_1, q_1) = 1$. We write w for the maximum of the absolute values of the coefficients of $W(x)$. Then, we derive from (17.33), (17.19), (17.22), (17.21) and (17.14) that

$$w \leq u_6^L \tag{17.35}$$

where $u_6 = u_6(\alpha, \delta)$ is a constant.

First, we prove the lemma under the assumption that $W(x) \not\equiv 0$. Now, we observe from (17.34) that $H(x) \not\equiv 0$ and we compare the coefficients of the highest power of x on both the sides of (17.34) to derive from (17.35) and (17.30) that

$$q_1^t \leq w \leq u_7^k$$

where $u_7 = u_7(\alpha, \delta) > 1$ is a constant. Therefore

$$t \leq \left[\frac{k \ \log u_7}{\log \ q_1} \right] + 1.$$

Hence, the assertion of Lemma 17.2.6 is valid with $u_5 = \log u_7$.

It remains to show that $W(x) \not\equiv 0$. By Lemma 17.2.4 and (17.18), we observe that either $P(x)$ or $Q(x)$ is not identically zero. Let $Q(x) \equiv 0$. Then $P(x) \not\equiv 0$ and we derive from (17.18) and (17.13) that

$$P^{(m)}(\alpha) = 0 \ \text{ for } \ 0 \leq m < k.$$

Consequently

$$P^{(m)}(\alpha^{(j)}) = 0 \ \text{ for } \ 0 \leq m < k, 0 \leq j < n$$

where $\alpha^{(1)} = \alpha, \cdots, \alpha^{(n)}$ denote the conjugates of α. Then

$$nk \leq \deg P \leq L < nk,$$

by $P(x) \not\equiv 0$ and (17.30) with $0 < \delta < 1$. This is a contradiction.

Thus, we conclude that $Q(x) \not\equiv 0$. Next, we show that $P(x)/Q(x)$ is not a constant function. We suppose that $P(x)/Q(x) = \lambda$ is a constant function. First, we notice that $\lambda \neq \alpha$ since $\alpha \notin \mathbf{Q}$. Further, we observe from (17.18) that

$$R(x, \alpha) = (\lambda - \alpha)Q(x)$$

Thus, we derive from (17.13) that $Q^{(m)}(\alpha^{(j)}) = 0$ for $0 \leq m < k$ and $0 \leq j < n$. Again, we observe $nk \leq \deg Q \leq L < nk$ to derive a contradiction.

This proves that $P(x)/Q(x)$ is not a constant function. This implies that

$$(P(x)/Q(x))^{(1)} = -W(x)/Q^2(x)$$

is not identically zero. This completes the proof of $W(x) \not\equiv 0$. □

Proof of Theorem 17.2.2: (Continued) Let $0 < \varepsilon < 1$ and $\kappa = \frac{n}{2} + 1 + \varepsilon$. If there are only finitely many reduced fractions p/q with $q > 0$ satisfying

$$\left| \alpha - \frac{p}{q} \right| < \frac{1}{q^\kappa}, \tag{17.36}$$

then we take $c_1' = \min \mid \alpha - \frac{p}{q} \mid$ where the minimum is taken over all reduced fractions p/q with $q > 0$ satisfying (17.36) for deriving (17.3) with $c_1 = \min(1, c_1')$ for all reduced fractions p/q with $q > 0$. Thus, we may assume that there are infinitely many reduced fractions p/q with $q > 0$ satisfying (17.36). Out of these infinitely many p/q, we shall choose suitably two p_1/q_1 and p_2/q_2 satisfying $q_2 > q_1 > 1$. We put

$$k = \left[\frac{\log\, q_2}{\log\, q_1}\right].$$

Thus

$$q_1^k \leq q_2 < q_1^{k+1} \tag{17.37}$$

Let $0 < \delta < \frac{1}{2}$ and L be given by (17.30), so that the assertions of Lemmas 17.2.5 and 17.2.6 are valid. Let

$$q_1 \geq e^{u_5/\delta} \tag{17.38}$$

Then

$$\left[\frac{u_5 k}{\log\, q_1}\right] + 2 \leq k\delta + 2.$$

By Lemma 17.2.6, we find m with

$$0 \leq m \leq k\delta + 2 \tag{17.39}$$

such that

$$R_m\left(\frac{p_1}{q_1}, \frac{p_2}{q_2}\right) \neq 0.$$

Consequently, we derive from (17.20) and (17.22) that

$$\left|R_m\left(\frac{p_1}{q_1}, \frac{p_2}{q_2}\right)\right| \geq q_1^{-L} q_2^{-1} \tag{17.40}$$

Now, we combine (17.40), (17.30) and (17.37) to obtain

$$\left|R_m\left(\frac{p_1}{q_1}, \frac{p_2}{q_2}\right)\right| \geq q_1^{-\frac{1}{2}(1+\delta)nk-k-1} \tag{17.41}$$

On the other hand, we refer to Lemma 17.2.5, (17.36), (17.39) and (17.30) for obtaining

$$\left|R_m\left(\frac{p_1}{q_1},\frac{p_2}{q_2}\right)\right| \le \max\left(q_1^{-\kappa(k(1-\delta)-2)}, q_2^{-\kappa}\right) u_8^k \tag{17.42}$$

where $u_8 = u_8(\alpha,\delta) > 0$. We can take a reduced fraction p_1/q_1 satisfying (17.36) and

$$q_1 > \max\left(e^{u_5/\delta}, u_8^{1/\delta}\right) \tag{17.43}$$

so that (17.38) is satisfied. Now, we derive from (17.42), (17.43) and (17.37) that

$$\left|R_m\left(\frac{p_1}{q_1},\frac{p_2}{q_2}\right)\right| \le q_1^{-\kappa(k(1-\delta)-2)+\delta k} \tag{17.44}$$

We combine (17.41) and (17.44) to observe that

$$\kappa k(1-\delta) - 2\kappa - \delta k \le \frac{1}{2}(1+\delta)nk + k + 1 \tag{17.45}$$

Let P_1/Q_1 and P_2/Q_2 be distinct reduced fractions satisfying $Q_2 > Q_1 > 1$ and (17.36). Then

$$\frac{1}{Q_1Q_2} \le \left|\frac{P_1}{Q_1} - \frac{P_2}{Q_2}\right| \le \left|\alpha - \frac{P_1}{Q_1}\right| + \left|\alpha - \frac{P_2}{Q_2}\right| < \frac{2}{Q_1^\kappa}$$

which implies that

$$Q_2 > Q_1^{\kappa-1}/2.$$

In view of the above *gap principle*, we can choose a reduced fraction p_2/q_2 satisfying $q_2 > 0$ and (17.36) such that

$$k \ge \delta^{-1} \tag{17.46}$$

Now we can combine (17.45), (17.46) and $\kappa \le (n+4)/2$ to derive that

$$\kappa \le \frac{n}{2} + 1 + \delta v$$

where $v \ge 1$ is a number depending only on n. Finally, we take $\delta = \varepsilon/2v$ to observe that $0 < \delta < 1/2$ and

$$\frac{n}{2} + 1 + \varepsilon = \kappa \le \frac{n}{2} + 1 + \frac{\varepsilon}{2}.$$

This is a contradiction. □

17.2.5 Proof of Theorem 17.2.3

We write

$$f(X,Y) = a_0(X-\alpha_1 Y)\cdots(X-\alpha_n Y)$$

where $0 \neq a_0 \in \mathbf{Z}$ and $\alpha_1, \cdots, \alpha_n$ are distinct algebraic numbers. We put

$$u_9 = \min_{i\neq j} |\alpha_i - \alpha_j| \ , \ u_{10} = u_9/2.$$

Let $x, y \in \mathbf{Z}$ with $|x| \leq |y|$ satisfy (17.4). Let $\alpha = \alpha(x,y) \in \{\alpha_1, \cdots, \alpha_n\}$ such that

$$\left|\alpha - \frac{x}{y}\right| = \min_{1\leq i\leq n} \left|\alpha_i - \frac{x}{y}\right|.$$

We re-write (17.4) as

$$a_0(x-\alpha_1 y)\cdots(x-\alpha_n y) = m.$$

Therefore

$$\prod_{i=1}^{n} \left|\alpha_i - \frac{x}{y}\right| = \left|\frac{m}{a_0 y^n}\right| \tag{17.47}$$

We may assume that

$$\left|\alpha - \frac{x}{y}\right| < u_{10}.$$

Then, for $\alpha_i \neq \alpha$, we have

$$\left|\alpha_i - \frac{x}{y}\right| \geq |\alpha - \alpha_i| - \left|\alpha - \frac{x}{y}\right| > u_{10}.$$

Therefore, we derive from (17.47) that

$$\left|\alpha - \frac{x}{y}\right| < \frac{u_{11}}{|y|^n} \tag{17.48}$$

where

$$u_{11} = |m| \,/\, |a_0|\, u_{10}^{n-1}.$$

By Theorem 17.2.2 with $\varepsilon = 1/4$ and (17.48), we conclude that

$$\frac{c_1}{|y|^\kappa} < \frac{u_{11}}{|y|^n}.$$

Therefore

$$|y|^{n-\kappa} < u_{11}c_1^{-1}$$

which implies that there are only finitely may possibilities for y and hence for x also, since $|x| \leq |y|$. This assertion follows from (17.47) whenever $|\alpha - x/y| \geq u_{10}$. The case for $|x| \geq |y|$ is similar. □

The constant c in Theorem 17.2.1 is effectively computable. This means that the method of proof of Theorem 17.2.1 allows to determine c explicitly in terms of α. This is not the case with the subsequent improvements of Thue, Siegel, Dyson, Gel'fond and Roth. The first effective non-trivial improvement of Liouville's inequality was established by Baker [4]. The best known result is due to Fel'dman [21] with $\kappa = n - \varepsilon$. It is clear from the proof of Theorem 17.2.3 that Baker's effective improvement of Liouville's inequality establishes an effective version of Theorem 17.2.3 *i.e., equation* (17.4) *implies that* $\max(|x|, |y|)$ *is bounded by an effectively computable number depending only on* f *and* m.

Chapter 18

Representation of numbers by squares

Let n be a natural number. Define

$$r_k(n) = \#\{(n_1, n_2, \ldots, n_k) \mid n_i \in \mathbf{Z},\ \sum_i n_i^2 = n\}$$

$$= \text{ number of ways of representing } n \text{ as a sum of } k \text{ squares.}$$

We wish to study the function $r_{4k}(n)$ for $k = 1, 2, \ldots$ using the *theta function*. Our goal is to prove the following theorem.

Theorem 18.0.4 *We have the asymptotic formula*

$$r_{4k}(n) = A_k \sigma^*_{2k-1}(n) + O(n^k),$$

where

$$A_k = \frac{(-1)^{k+1} 4k}{(2^{2k} - 1) B_{2k}},$$

$$\sigma^*_{2k-1}(n) = \sum_{d|n,\ d \text{ odd}} (n/d)^{2k-1} + \sum_{d|n,\ d \text{ even}} (-1)^{k+(n/d)} (n/d)^{2k-1}.$$

For $k = 1, 2$, we get the exact formulas

$$r_4(n) = 8\Big(\sum_{d|n,\ d \text{ odd}} (n/d) - \sum_{d|n,\ d \text{ even}} (-1)^{n/d}(n/d)\Big)$$

$$= \begin{cases} 8 \sum_{d|n} d & \text{if } n \text{ is odd;} \\ 24 \sum_{d|n,\ d \text{odd}} d & \text{if } n \text{ is even.} \end{cases}$$

and

$$r_8(n) = 16(\sum_{d|n,\ d \text{ odd}} (n/d)^3 + \sum_{d|n,\ d \text{ even}} (-1)^{n/d}(n/d)^3).$$

Here B_n is the n^{th} Bernoulli number, defined by the Taylor series

$$\frac{x}{1-e^{-x}} = \sum_{n=0}^{\infty} B_n \frac{x^n}{n!}.$$

We will also use the following notation: for a complex number z, the symbols $\Re z$ and $\Im z$ will denote its real and imaginary parts.

The results on representations by 4 and 8 squares are due to Jacobi. The proof given below is adapted from the proof for $k = 1$ given in [43], Ch. 1, §15.

18.1 The theta function

Let

$$f(t) = \sum_{n \in \mathbf{Z}} t^{n^2} = 1 + 2\sum_{n=1}^{\infty} t^{n^2}.$$

Then clearly

$$(f(t))^k = \sum_{n=0}^{\infty} r_k(n) t^n.$$

The technique we will use is to study the associated function

$$\theta(\tau) = \sum_{n \in \mathbf{Z}} e^{\pi \iota n^2 \tau},$$

where $\iota = \sqrt{-1}$. This is just $f(e^{\pi\iota\tau})$. Then $r_k(n)$ is interpreted as a Fourier coefficient of θ.

It is convenient to study the function

$$\theta(\tau, z) = \sum_{n \in \mathbf{Z}} e^{\pi \iota n^2 \tau + 2\pi\iota n z}.$$

Proposition 18.1.1 *The series*

$$\theta(\tau, z) = \sum_{n \in \mathbf{Z}} e^{\pi \iota n^2 \tau + 2\pi\iota n z}$$

converges absolutely and uniformly on compact subsets of $\mathcal{H} \times \mathbf{C}$, *where* $\mathcal{H} = \{\tau \in \mathbf{C} \mid \Im\tau > 0\}$ *is the upper half plane, and defines an analytic function*[1] *on* $\mathcal{H}$ *for each* $z \in \mathbf{C}$.

$\theta(\tau, z)$ *satisfies the transformation formulas*

(i) $\theta(\tau + 2, z) = \theta(\tau, z)$

(ii) $\theta(\tau + 1, z) = \theta(\tau, z + 1/2)$

(iii) $\theta(-1/\tau, z/\tau) = \sqrt{\frac{\tau}{\iota}} e^{\pi\iota z^2/\tau}\theta(\tau, z)$, *where* $\sqrt{\frac{\tau}{\iota}}$ *is the branch of the square root which is positive on the positive imaginary axis* $\{iy \in \mathbf{C} \mid y > 0\} \subset \mathcal{H}$.

Proof: The function $e^{\pi\iota\tau}$ gives an analytic mapping $e(\tau) : \mathcal{H} \to \Delta^*$, where $\Delta^* = \{t \in \mathbf{C} \mid 0 < |t| < 1\}$ is the punctured open unit disc. Hence for any analytic function $h(t)$ on Δ^*, the composition $(h \circ e)(\tau) = h(e(\tau))$ is analytic on $\mathcal{H}$. Apply this to

$$h(t) = 1 + \sum_{n=1}^{\infty} (e^{2\pi\iota nz} + e^{-2\pi\iota nz})t^{n^2} = 1 + \sum_{n=1}^{\infty} (2\cos 2\pi nz)t^{n^2}.$$

Since $|2\cos 2\pi nz| \leq 2e^{-2\pi ny}$, where $y = \Im z$, we see that $|\cos 2\pi nz|^{1/n^2} \leq 2^{1/n^2}e^{-2\pi y/n} \to 1$ as $n \to \infty$. Hence the radius of convergence of the power series $h(t)$ is ≥ 1, and it defines an analytic function on the open unit disc $\Delta = \{t \in \mathbf{C} \mid |t| < 1\}$.

The estimate

$$\left|e^{\pi\iota n^2\tau \pm 2\pi\iota nz}\right|^{1/n} = (e^{-\pi(\Im\tau)n \pm 2\pi y})$$

for $n > 0$ shows that the series defining $\theta(\tau, z)$ converges absolutely and uniformly on compact subsets of $\mathcal{H} \times \mathbf{C}$.

The transformation formula $\theta(\tau + 2, z) = \theta(\tau, z)$ is clear, since $e^{\pi\iota n^2\tau}$ is periodic with period 2 for each n. The formula $\theta(\tau + 1, z) = \theta(\tau, z + 1/2)$ follows by substitution in the series, since $e^{\pi\iota n^2} = (-1)^{n^2} = (-1)^n = e^{\pi\iota n}$.

The third transformation formula for θ is proved using the *Poisson summation formula*, which states (in our context) that if $g(t)$ is a C^∞ complex

[1] In fact $\theta(\tau, z)$ is an analytic function of two variables on $\mathcal{H} \times \mathbf{C}$, but we do not need this.

valued function on $\mathbf{R}$ such that g and all its derivatives are rapidly decreasing[2] as $|t| \to \infty$, then

$$\sum_{n \in \mathbf{Z}} g(n) = \sum_{n \in \mathbf{Z}} \widehat{g}(n),$$

where

$$\widehat{g}(s) = \int_{-\infty}^{\infty} g(t) e^{-2\pi \iota s t} dt$$

is the *Fourier transform* of g (the integral converges absolutely because g is rapidly decreasing at infinity). This formula is proved in the Appendix (B) (part of the proof is that the series on the right converges absolutely).

We apply this to the function

$$g(t) = e^{\pi \iota \tau t^2 + 2\pi \iota z t}$$

where $(\tau, z) \in \mathcal{H} \times \mathbf{C}$, $t \in \mathbf{R}$. Note that

$$|g(t)| = e^{-\pi (\Im \tau) t^2 - 2\pi y t}$$

which (for a fixed $(\tau, z) \in \mathcal{H} \times \mathbf{C}$) is rapidly decreasing as $|t| \to \infty$. A similar estimate works for any (partial) derivative with respect to t. The formula

$$\int_{-\infty}^{\infty} e^{-\pi a t^2} dt = \frac{1}{\sqrt{a}}$$

for positive real a yields

$$\int_{-\infty}^{\infty} e^{\pi \iota \tau t^2} dt = \sqrt{\frac{\iota}{\tau}}$$

where the square root is the branch which is positive on the positive imaginary axis; this yields

$$\widehat{g}(s) = \int_{-\infty}^{\infty} g(t) e^{-2\pi \iota s t} dt = \sqrt{\frac{\iota}{\tau}} e^{-\pi \iota (z-s)^2 / \tau}.$$

Hence the Poisson summation formula yields

$$\theta(\tau, z) = \sum_{n \in \mathbf{Z}} g(n) = \sum_{n \in \mathbf{Z}} \widehat{g}(n) = \sqrt{\frac{\iota}{\tau}} \sum_{n \in \mathbf{Z}} e^{-\pi \iota (z-n)^2 / \tau}$$

$$= \sqrt{\frac{\iota}{\tau}} e^{-\pi \iota z^2 / \tau} \sum_{n \in \mathbf{Z}} e^{\pi \iota n^2 (-1/\tau) + 2\pi \iota n (z/\tau)} = \sqrt{\frac{\iota}{\tau}} e^{-\pi \iota z^2 / \tau} \theta(-1/\tau, z/\tau)$$

which is the third transformation formula for θ. □

[2]Recall that a function $g(t)$ on $\mathbf{R}$ is *rapidly decreasing at infinity* if $|t^n g(t)| \to 0$ as $|t| \to \infty$, for any $n \geq 0$. Then $\int_{-\infty}^{\infty} t^n g(t) dt$ is absolutely convergent for any $n \geq 0$.

Remark: It is obvious that $\theta(\tau, z)$ satisfies $\theta(\tau, z+1) = \theta(\tau, z)$; it is also easy to see that $\theta(\tau, z+\tau) = e^{-2\pi\iota z - \pi i \tau}\theta(\tau, z)$. This implies that if $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbf{C}$ with $a_1 + \cdots + a_n = b_1 + \cdots + b_n$, then for fixed τ,

$$f(z) = \prod_{j=1}^{n} \frac{\theta(\tau, z - a_j)}{\theta(\tau, z - b_j)}$$

is meromorphic on $\mathbf{C}$, and is doubly periodic with periods $1, \tau$ *i.e.*, is an elliptic function on the torus $E_\tau = \mathbf{C}/\mathbf{Z} + \tau\mathbf{Z}$; further, one can show that all elliptic functions on E_τ are of this form, upto a constant multiple. In fact $\theta(\tau, z)$ is closely related to the function $\sigma(\tau, z)$ (the σ-function associated to the lattice generated by 1 and τ) constructed in Chapter (1) (for more details see Exercise (18.3.2)).

In another direction, one can show that the second derivative of $\log \theta(\tau, z)$ is a translate of the $\wp$ function, upto an additive constant. For proofs of these facts, see Mumford's book [43], Ch. 1, §6, or Lang's book [33]. The function $\theta(\tau) = \theta(\tau, 0)$ is thus a sort of function on the space of lattices *i.e.*, the parameter space for elliptic curves.

As seen in the proof of the proposition, the function $\theta(\tau)$ corresponds to the analytic function $f(z) = 1 + 2\sum_{n=1}^{\infty} z^{n^2}$ on Δ^*, which extends to an analytic function on Δ, with $f(0) = 1$. This implies that for τ in the vertical strip $|\Re\tau| \leq 1$, we have

$$\lim_{\Im\tau \to \infty} \theta(\tau) = 1.$$

In a similar fashion, consider the function $\theta(1 - 1/\tau)$. Again, if τ lies in the vertical strip $|\Re\tau| \leq 1$, then as $\Im\tau \to \infty$, $1 - 1/\tau \to 1$, and we have

$$\theta(1 - 1/\tau) = \sum_{n \in \mathbf{Z}} (-1)^n e^{-\pi\iota n^2/\tau}$$

$$= \theta(-1/\tau, 1/2) = \sqrt{\frac{\tau}{\iota}} e^{\pi\iota(\tau/2)^2/\tau}\theta(\tau, \tau/2)$$

$$= \sqrt{\frac{\tau}{\iota}} e^{\pi\iota\tau/4} \sum_{n \in \mathbf{Z}} e^{\pi\iota n^2\tau + \pi\iota n\tau}.$$

Thus if

$$k(t) = \sum_{n \in \mathbf{Z}} t^{n^2+n},$$

then $k(t)$ is analytic in Δ, with $k(0) = 2$ (as $n^2 + n = 0$ for $n = 0, -1$) and

$$\sqrt{\frac{\iota}{\tau}} e^{-\pi\iota\tau/4}\theta(1 - 1/\tau) = k \circ e(\tau).$$

Hence for τ in the vertical strip $|\Re\tau| \leq 1$,

$$\lim_{\Im\tau\to\infty} \sqrt{\frac{\iota}{\tau}} e^{-\pi\iota\tau/4}\theta(1 - 1/\tau) = 1.$$

Since

$$\left|\sqrt{\frac{\tau}{\iota}} e^{\pi\iota\tau/4}\right| = |\tau|^{1/2}\, e^{-\pi\Im\tau/4} \to 0 \text{ as } \Im\tau \to \infty,$$

we deduce that for τ in the vertical strip,

$$\lim_{\Im\tau\to\infty} \theta(1 - 1/\tau) = 0.$$

We thus have:

Proposition 18.1.2 *For any $k > 0$, $\theta(z)^{4k}$ satisfies*

(i) $\theta(\tau + 2)^{4k} = \theta(\tau)^{4k}$

(ii) $\theta(-1/\tau)^{4k} = (-1)^k \tau^{2k}\theta(\tau)^{4k}$

(iii) $\theta(\tau)^{4k}$ *has the Fourier expansion*

$$\theta(\tau)^{4k} = 1 + \sum_{n=1}^{\infty} r_{4k}(n) e^{\pi\iota n\tau}.$$

Thus, for τ lying in the vertical strip $|\Re\tau| \leq 1$,

$$\lim_{\Im\tau\to\infty} \theta(\tau)^{4k} = 1.$$

(iv) $\tau^{-2k}\theta(1 - 1/\tau)^{4k}$ *has a Fourier expansion*

$$\tau^{-2k}\theta(1 - 1/\tau)^{4k} = \sum_{n=1}^{\infty} a_n e^{\pi\iota n\tau}$$

for some constants a_n. Thus, for τ lying in a vertical strip $|\Re\tau| \leq 1$,

$$\lim_{\Im\tau\to\infty} \theta(1 - 1/\tau)^{4k} = 0.$$

18.2 Eisenstein series

As noted above, $r_{4k}(n)$ is the n^{th} Fourier coefficient of $\theta(\tau)^{4k}$, which is expanded in powers of $e^{\pi\iota\tau}$. The idea is to consider a certain other function $E^*_{2k}(\tau)$, an *Eisenstein series*, which has transformation properties similar to that of $\theta(\tau)^{4k}$, and similar behaviour as $\Im\tau \to \infty$ and near $\tau = 1$. The Fourier series for E^*_{2k} will be explicitly computed in terms of a divisor function. We will then estimate the Fourier coefficients of $E^*_{2k}(\tau) - \theta(\tau)^{4k}$ to have smaller order of magnitude than those of $E^*_{2k}(\tau)$. Hence the Fourier coefficients of $\theta(\tau)^{4k}$ must be asymptotically equal to those of $E^*_{2k}(\tau)$ *i.e.*, we get an *asymptotic formula* for $r_{4k}(n)$ and $r_8(n)$. This will turn out to be an exact formula for $r_4(n)$.

We define

$$E^*_{2k}(\tau) = C \sum_{m,n\in\mathbf{Z}} \left(\frac{1}{(2m\tau + 2n + 1)^{2k}} + \frac{(-1)^k}{((2m+1)\tau + 2n)^{2k}} \right).$$

Here $C = ((2 - 2^{1-2k})\zeta(2k))^{-1}$ where $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$ is the Riemann zeta function. We note that for $k \geq 1$, the series for $E^*_{2k}(\tau)$ converges absolutely and uniformly for τ lying in compact subsets of $\mathcal{H}$ (in fact, for $k > 1$, each of the two double series $\sum_{m,n} (2m\tau + 2n + 1)^{-2k}$ and $\sum_{m,n} ((2m+1)\tau + 2n)^{-2k}$ has the covregence property cf. Proposition (1.2.5), and the proof of (a) in the proposition below makes use of this fact; for the case $k = 1$, see Exercise (18.3.1) below.) We may rewrite the series as

$$\sum_{m\in\mathbf{Z}} \left(\sum_{n\in\mathbf{Z}} \frac{1}{(2m\tau + 2n + 1)^{2k}} + \sum_{n\in\mathbf{Z}} \frac{(-1)^k}{((2m+1)\tau + 2n)^{2k}} \right),$$

since for fixed m, the two series in n converge absolutely.

Proposition 18.2.1 *(a) $E^*_{2k}(\tau)$ satisfies the transformation formulas*

(i) $E^*_{2k}(\tau + 2) = E^*_{2k}(\tau)$

(ii) $E^*_{2k}(-1/\tau) = (-1)^k \tau^{2k} E^*_{2k}(\tau)$.

(b) $E^*_{2k}(\tau)$ has the Fourier expansion

$$E^*_{2k}(\tau) = 1 + \frac{\pi^{2k}}{(1 - 2^{-2k})\zeta(2k)(2k-1)!} \sum_{n=1}^{\infty} \sigma^*_{2k-1}(n) e^{\pi\iota n\tau},$$

where

$$\sigma^*_{2k-1}(n) = \sum_{d|n,\ d \text{ even}} (-1)^{k+(n/d)}(n/d)^{2k-1} + \sum_{d|n,\ d \text{ odd}} (n/d)^{2k-1}.$$

Hence for τ in the vertical strip $|\Re\tau| \le 1$,

$$\lim_{\Im\tau\to\infty} E^*_{2k}(\tau) = 1.$$

(c) $\tau^{-2k}E^*_{2k}(1-1/\tau)$ is periodic with period 1, and has a Fourier expansion

$$\tau^{-2k}E^*_{2k}(1-1/\tau) = \sum_{n=1}^{\infty} b_n e^{\pi\imath n\tau}$$

for some constants b_n. Hence for τ in the vertical strip $|\Re\tau| \le 1$,

$$\lim_{\Im\tau\to\infty} E^*_{2k}(1-1/\tau) = 0.$$

Proof: (a) Clearly $E^*_{2k}(\tau)$ is periodic with period 2. We compute that

$$\begin{aligned}
&E^*_{2k}(-1/\tau)\\
&= C\sum_{m,n\in\mathbf{Z}}\left(\frac{1}{(2m(-1/\tau)+2n+1)^{2k}} + \frac{(-1)^k}{((2m+1)(-1/\tau)+2n)^{2k}}\right)\\
&= C\sum_{m,n\in\mathbf{Z}}\left(\frac{\tau^{2k}}{((2n+1)\tau+2m)^{2k}} + \frac{(-1)^k\tau^{2k}}{(2n\tau+2m+1)^{2k}}\right)\\
&= C(-1)^k\tau^{2k}\sum_{m,n\in\mathbf{Z}}\left(\frac{1}{(2n\tau+2m+1)^{2k}} + \frac{(-1)^k}{((2n+1)\tau+2m)^{2k}}\right)\\
&= (-1)^k\tau^{2k}E^*_{2k}(\tau).
\end{aligned}$$

For (b) and (c), we make use of the Fourier expansion of

$$f_r(x) = \sum_{n\in\mathbf{Z}} \frac{1}{(x+n)^r}$$

where $r \ge 2$ is an integer (the series converges absolutely and uniformly on compact sets for $x \in \mathbf{C}-\mathbf{Z}$, and defines a periodic function of x with period 1). We begin with the well known series (see Ahlfors' book [2] Ch. 5, §2.1)

$$\pi\cot\pi x = \frac{1}{x} + \sum_{n=1}^{\infty}\frac{2x}{x^2-n^2}.$$

Differentiating $r-1$ times, and noting that

$$\frac{2x}{x^2-n^2} = \frac{1}{x-n} + \frac{1}{x+n},$$

we see that

$$\pi\left(\frac{d}{d\,x}\right)^{r-1} \cot \pi x = (-1)^{r-1}(r-1)!\left(\frac{1}{x^r} + \sum_{n=1}^{\infty} \frac{1}{(x+n)^r} + \frac{1}{(x-n)^r}\right),$$

so that

$$f_r(x) = \frac{(-1)^{r-1}\pi}{(r-1)!}\left(\frac{d}{d\,x}\right)^{r-1} \cot \pi x.$$

However, if $x \in \mathcal{H}$, the upper half plane,

$$\begin{aligned} \cot \pi x &= \iota \frac{e^{\pi \iota x} + e^{-\pi \iota x}}{e^{\pi \iota x} - e^{-\pi \iota x}} = -\iota \frac{1 + e^{2\pi \iota x}}{1 - e^{2\pi \iota x}} \\ &= -\iota(1 + e^{2\pi \iota x}) \sum_{n=0}^{\infty} e^{2\pi \iota n x} = -\iota(1 + 2\sum_{n=1}^{\infty} e^{2\pi \iota n x}) \end{aligned}$$

(note that since $\Im x > 0$, $|e^{2\pi \iota x}| = e^{-\pi \Im x} < 1$). Differentiating $r-1$ times, we get (since $r \geq 2$)

$$\pi\left(\frac{d}{d\,x}\right)^{r-1} \cot \pi x = -(2\pi\iota)^r \sum_{n=1}^{\infty} n^{r-1} e^{2\pi \iota n x},$$

so that

$$f_r(x) = \frac{(-2\pi\iota)^r}{(r-1)!} \sum_{n=1}^{\infty} n^{r-1} e^{2\pi \iota n x}.$$

If x lies in the lower half plane, then $-x \in \mathcal{H}$, and we use $f_r(-x) = (-1)^r f_r(x)$ to obtain the Fourier expansion

$$f_r(x) = \frac{(2\pi\iota)^r}{(r-1)!} \sum_{n=1}^{\infty} n^{r-1} e^{-2\pi \iota n x}.$$

We apply this as follows. We have

$$C^{-1} E_{2k}^*(\tau)$$

$$
\begin{aligned}
&= \sum_{n\in\mathbf{Z}} \frac{1}{(2n+1)^{2k}} + \sum_{m\neq 0}\Big(\sum_{n\in\mathbf{Z}} \frac{1}{(2m\tau+2n+1)^{2k}}\Big) + \\
&\quad (-1)^k \sum_{m\in\mathbf{Z}}\Big(\sum_{n\in\mathbf{Z}} \frac{1}{((2m+1)\tau+2n)^{2k}}\Big) \\
&= 2\sum_{n=0}^{\infty}(2n+1)^{-2k} + \sum_{m\neq 0} 2^{-2k} f_{2k}(m\tau+1/2) + \\
&\quad (-1)^k \sum_{m\in\mathbf{Z}} 2^{-2k} f_{2k}((m+1/2)\tau) \\
&= 2\sum_{n=0}^{\infty}(2n+1)^{-2k} + 2^{1-2k}\sum_{m=1}^{\infty} f_{2k}(m\tau+1/2) + \\
&\quad 2^{1-2k}(-1)^k \sum_{m=0}^{\infty} f_{2k}((m+1/2)\tau),
\end{aligned}
$$

where we have used $f_{2k}(-x) = f_{2k}(x)$, and $f_{2k}(x+1) = f_{2k}(x)$. Now

$$
\sum_{n=0}^{\infty}(2n+1)^{-2k} = \sum_{n=1}^{\infty} n^{-2k} - \sum_{n=1}^{\infty}(2n)^{-2k} = (1-2^{-2k})\zeta(2k),
$$

where $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$ is the Riemann zeta function. Hence

$$
\begin{aligned}
&C^{-1}E_{2k}^*(\tau) \\
&\quad = 2(1-2^{-2k})\zeta(2k) + 2\frac{\pi^{2k}}{(2k-1)!} \times \\
&\qquad \left((-1)^k \sum_{m=1}^{\infty}\Big(\sum_{n=1}^{\infty}(-1)^n n^{2k-1} e^{\pi\iota(2m)n\tau}\Big) + \sum_{m=0}^{\infty}\Big(\sum_{n=1}^{\infty} n^{2k-1} e^{\pi\iota(2m+1)n\tau}\Big)\right).
\end{aligned}
$$

Let

$$
\sigma_{2k-1}^*(n) = \sum_{d|n,\ d \text{ even}} (-1)^{k+(n/d)}(n/d)^{2k-1} + \sum_{d|n,\ d \text{ odd}} (n/d)^{2k-1}.
$$

Then we have

$$
E_{2k}^*(\tau) = C(2-2^{1-2k})\zeta(2k) + \frac{2C\pi^{2k}}{(2k-1)!}\sum_{n=1}^{\infty}\sigma_{2k-1}^*(n)e^{\pi\iota n\tau}.
$$

By our choice of C, we get

$$
E_{2k}^*(\tau) = 1 + \frac{\pi^{2k}}{(1-2^{-2k})\zeta(2k)(2k-1)!}\sum_{n=1}^{\infty}\sigma_{2k-1}^*(n)e^{\pi\iota n\tau}.
$$

This proves (c). Finally, consider

$$\begin{aligned}
&C^{-1}E_{2k}^*(1-1/\tau)\\
&= \sum_{m,n\in\mathbf{Z}}\left(\frac{1}{(2m-(2m/\tau)+2n+1)^{2k}}+\frac{(-1)^k}{(2m+1-((2m+1)/\tau)+2n)^{2k}}\right)\\
&= \tau^{2k}\sum_{m,n\in\mathbf{Z}}\left(\frac{1}{((2m+2n+1)\tau-2m)^{2k}}+\frac{(-1)^k}{((2m+2n+1)\tau-2m-1)^{2k}}\right)\\
&= \tau^{2k}2\sum_{p=0}^{\infty}\sum_{q\in\mathbf{Z}}\left(\frac{1}{((2p+1)\tau+2q)^{2k}}+\frac{(-1)^k}{(2p+1)\tau+2q+1)^{2k}}\right)\\
&= \tau^{2k}2^{1-2k}\sum_{p=0}^{\infty}\left(f_{2k}((p+1/2)\tau)+(-1)^k f_{2k}((p+1/2)\tau+1/2)\right).
\end{aligned}$$

Hence $\tau^{-2k}E_{2k}^*(1-1/\tau)$ has a Fourier expansion of the desired form, proving (d). □

Thus, $F(\tau)=\theta(\tau)^{4k}-E_{2k}^*(\tau)$ has the following properties:

(i) F is analytic in $\mathcal{H}$, and satisfies the transformation formulas

$$F(\tau+2)=F(\tau),\quad F(-1/\tau)=(-1)^k\tau^{2k}F(\tau).$$

(ii) F has a Fourier expansion ('at infinity')

$$F(\tau)=\sum_{n=1}^{\infty}\alpha_n e^{\pi\imath n\tau},$$

where

$$\alpha_n=\left(r_{4k}(n)-\frac{\pi^{2k}\sigma_{2k-1}^*(n)}{(1-2^{-2k})\zeta(2k)(2k-1)!}\right).$$

(iii) $\tau^{-2k}F(1-1/\tau)$ has a Fourier expansion

$$\tau^{-2k}F(1-1/\tau)=\sum_{n=1}^{\infty}\beta_n e^{\pi\imath n\tau}.$$

An analytic function $G(\tau)$ on $\mathcal{H}$ which satisfies the above two transformation formulas, and has Fourier expansions

$$G(\tau) = \sum_{n=1}^{\infty} \gamma_n e^{\pi \imath n \tau},$$

$$\tau^{-2k} G(1 - 1/\tau) = \sum_{n=1}^{\infty} \gamma'_n e^{\pi \imath n \tau}$$

will be called[3] a *cusp form of weight* $2k$.

Theorem 18.2.1 *(i) Let $G(\tau)$ be a cusp form of weight $2k$, with Fourier expansion*

$$G(\tau) = \sum_{n=1}^{\infty} a_n e^{\pi \imath n \tau}.$$

Then there is a constant $A > 0$ such that $|a_n| \leq An^k$.

(ii) Let $G(\tau)$ be a cusp form of weight 2 or 4. Then $G = 0$.

Granting this, we have our main result on the number of representations of a number by $4k$ squares

Theorem 18.2.2 *We have the asymptotic formula*

$$r_{4k}(n) = A_k \sigma^*_{2k-1}(n) + O(n^k),$$

where

$$A_k = \frac{\pi^{2k}}{(1 - 2^{-2k})\zeta(2k)(2k-1)!},$$

$$\sigma^*_{2k-1}(n) = \sum_{d|n,\ d \text{ odd}} (n/d)^{2k-1} + \sum_{d|n,\ d \text{ even}} (-1)^{k+(n/d)} (n/d)^{2k-1}.$$

For $k = 1, 2$, we get the exact formulas

$$r_4(n) = 8\Big(\sum_{d|n,\ d \text{ odd}} d - \sum_{d|n,\ d \text{even}} (-1)^{n/d} d\Big)$$

and

$$r_8(n) = 16\Big(\sum_{d|n,\ d \text{ odd}} (n/d)^3 + \sum_{d|n,\ d \text{ even}} (-1)^{n/d} (n/d)^3.$$

[3]Actually, such a function G is a cusp form of weight $2k$, with character, for the subgroup of $SL_2(\mathbf{Z})$ generated by $\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$.

Theorem 18.0.4 follows immediately from this, using Euler's formula for $\zeta(2k)$ in terms of Bernoulli numbers,

$$\zeta(2k) = \frac{(-1)^{k+1}(2\pi)^{2k}B_{2k}}{2(2k)!}.$$

Here B_n is defined by

$$\frac{x}{1-e^{-x}} = \sum_{n=0}^{\infty} B_n \frac{x^n}{n!},$$

so that B_n may be computed inductively using $B_0 = 1$, $B_1 = 1/2$, $B_n = 0$ if n is odd and > 1, and the recurrence formula

$$\begin{pmatrix} n \\ 0 \end{pmatrix} \frac{B_0}{n+1} - \begin{pmatrix} n \\ 1 \end{pmatrix} \frac{B_1}{1} + \begin{pmatrix} n \\ 2 \end{pmatrix} \frac{B_2}{n-1} - \cdots$$
$$+(-1)^r \begin{pmatrix} n \\ r \end{pmatrix} \frac{B_r}{n+1-r} + \cdots + (-1)^n B_n = 0.$$

Euler's formula is proved easily from the formula (which we cited earlier)

$$\pi \cot \pi x = \frac{1}{x} + \sum_{n=1}^{\infty} \frac{2x}{x^2 - n^2}.$$

Thus for $|x| < 1$,

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

since the double sum converges absolutely. On the other hand,

$$\pi x \cot \pi x = \pi \iota x \frac{e^{\pi \iota x} + e^{-\pi \iota x}}{e^{\pi \iota x} - e^{-\pi \iota x}}$$

$$\begin{aligned} &= -\pi\iota x + \frac{2\pi\iota x}{1-e^{-2\pi\iota x}} \\ &= -\pi\iota x + \sum_{n=0}^{\infty} B_n \frac{(2\pi\iota x)^n}{n!}. \end{aligned}$$

Comparing coefficients of powers of x yields Euler's formula (and also the statement that $B_n = 0$ for odd $n > 1$).

18.3 Cusp forms

It remains to prove Proposition 18.2.1. Our discussion is loosely based on that in Ogg's book [44].

Let $\Gamma \subset SL_2(\mathbf{Z})$ be the subgroup generated by the matrices $\gamma_1 = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$ and $\gamma_2 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. Then Γ acts on $\mathcal{H}$ through analytic automorphisms by

$$\gamma\tau = \frac{a\tau + b}{c\tau + d} \text{ for all } \gamma = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \Gamma.$$

We observe that

$$\gamma \longmapsto J_\gamma(\tau) = c\tau + d$$

satisfies the *cocycle condition*

$$J_{\gamma\gamma'}(\tau) = J_\gamma(\gamma'\tau) J_{\gamma'}(\tau)$$

(computation left to the reader). Note that

$$\Im(\gamma\tau) = |J_\gamma(\tau)|^{-2} \Im\tau.$$

Define a homomorphism $\chi : \Gamma \to \{\pm 1\}$ by

$$\chi(\gamma_1) = 1, \quad \chi(\gamma_2) = -1.$$

To see that this is well defined, one observes that Γ is contained in[4] the subgroup $\Gamma' \subset SL_2(\mathbf{Z})$ generated by γ_2 and the *congruence subgroup* $\Gamma(2)$

[4] Actually $\Gamma = \Gamma'$.

consisting of matrices in $SL_2(\mathbf{Z})$ which are congruent to the identity matrix modulo 2. Now $\gamma_2^2 \in \Gamma(2)$, and $\Gamma(2) \subset SL_2(\mathbf{Z})$ is a normal subgroup (it is the kernel of the natural homomorphism $SL_2(\mathbf{Z}) \to SL_2(\mathbf{Z}/2\mathbf{Z})$ given by reduction modulo 2). This implies that $\Gamma(2) \subset \Gamma'$ is a normal subgroup of index 2; hence there is a natural character of order 2 of Γ' which has kernel $\Gamma(2)$. The character χ on Γ is the restriction of the one on Γ'.

We need one more fact about the action of Γ on $\mathcal{H}$.

Lemma 18.3.1 *Let*

$$\mathcal{F} = \{\tau \in \mathcal{H} \mid |\Re\tau| \leq 1, \ \textit{and} \ |\tau| \geq 1\}.$$

Then

$$\bigcup_{\gamma \in \Gamma} \gamma\mathcal{F} = \mathcal{H}.$$

Proof: If $\tau \in \mathcal{H}$, $\gamma \in \Gamma$, then

$$\Im(\gamma\tau) = \frac{\Im\tau}{|c\tau + d|^2},$$

where

$$\gamma = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Hence for a given τ, if $\gamma \in \Gamma$ with $\Im(\gamma\tau) > \Im\tau$, then $|c\tau + d| < 1$. Since c, d are integers, we see that for a given τ, there are only finitely many possibilities for c, d. Thus

$$\{\Im(\gamma\tau) \mid \gamma \in \Gamma\}$$

has maximal elements. Applying a suitable translation, we can then find $\gamma \in \Gamma$ such that $\Im(\gamma\tau)$ is maximal, and $|\Re\tau| \leq 1$. Then $\tau' = \gamma\tau \in \mathcal{F}$ – if not, then $|\tau'| < 1$, and $\Im(-1/\tau') = \Im\tau'/|\tau'|^2 > \Im\tau'$ contradicting maximality. □

Remark: In fact, one can show that $\mathcal{F}$ is a *fundamental region* (see Chapter (6)) for the action of Γ on $\mathcal{H}$, in the sense that if $\gamma_1 x = \gamma_2 x$ for some $\gamma_i \in \Gamma$, where $x \in \mathcal{F}$ is an *interior* point, then $\gamma_1 = \gamma_2$. Hence, the quotient space $\mathcal{H}/\Gamma$ may be constructed from $\mathcal{F}$ by making suitable identifications along the boundary. For a proof that $\mathcal{F}$ is a fundamental region, see Ogg's book [44], page 32.

We are now ready to begin the proof of the proposition. We claim first that a cusp form $G(\tau)$ of weight $2k$ satisfies

$$G(\gamma\tau) = \chi(\gamma)J_\gamma(\tau)^{2k}G(\tau) \ \ \forall\gamma \in \Gamma.$$

This is true when γ is one of the generators, by the definition of a cusp form. This implies it in general, since the set of $\gamma \in \Gamma$ for which the formula holds is a subgroup of Γ, from the cocycle condition for $J_\gamma(\tau)$. Hence

$$h(\tau) = (\Im\tau)^k \, |G(\tau)|$$

is *invariant* under Γ.

Proposition 18.3.2 *$h(\tau) = (\Im\tau)^k \, |G(\tau)|$ is bounded on $\mathcal{H}$.*

Proof: Since h is Γ-invariant, it suffices to show that $h \mid_{\mathcal{F}}$ is bounded, where $\mathcal{F} \subset \mathcal{H}$ is a subset such that *displaystyle* $\cup_{\gamma\in\Gamma} \gamma\mathcal{F} = \mathcal{H}$.

Now G has a Fourier expansion

$$G(\tau) = \sum_{n=1}^{\infty} a_n e^{\pi\imath n\tau}$$

for some constants a_n. Hence for $|\Re\tau| \leq 1$,

$$\lim_{\Im\tau\to\infty} e^{-\pi\imath\tau}G(\tau) = a_1.$$

Since $|e^{\pi\imath\tau}| = e^{-\pi\Im\tau}$, this implies that for $|\Re\tau| \leq 1$,

$$\lim_{\Im\tau\to\infty} h(\tau) = \lim_{\Im\tau\to\infty} (\Im\tau)^k \, |G(\tau)| = a_1 \lim_{\Im\tau\to\infty} (\Im\tau)^k e^{-\Im\tau} = 0.$$

Hence $h(\tau)$ is bounded in the region $\mathcal{F}(\varepsilon) = \mathcal{F} \cap \{\Im\tau \geq \varepsilon\}$, for any $\varepsilon > 0$.

Similarly, the Fourier expansion

$$\tau^{-2k}G(1 - 1/\tau) = \sum_{n=1}^{\infty} b_n e^{\pi\imath n\tau}$$

for some constants b_n implies that for $|\Re\tau| \leq 1$,

$$\lim_{\Im\tau\to\infty} h(1 - 1/\tau) = b_1 \lim_{\Im\tau\to\infty} e^{-\pi\Im\tau} = 0.$$

Hence $h(\tau)$ is bounded in the region $\mathcal{F}_1(C) = \mathcal{F} \cap \{1 - 1/\tau \mid \Im\tau > C,\ |\Re\tau| \leq 1\}$.

We determine more explicitly what the region $\mathcal{F}_1(C)$ is. First, we compute the image of the horizontal line $\Im\tau = C$ under $\tau \longmapsto 1 - 1/\tau$. It $\tau = s + \iota C$ lies on this line, then

$$-1/\tau = \frac{s - 1 + \iota C}{s + \iota C} = \frac{s^2 - s + C^2 + \iota C}{s^2 + C^2},$$

so that the image of the line is the curve given parametrically by

$$x = 1 - \frac{s}{s^2 + C^2}, \quad y = \frac{C}{s^2 + C^2};$$

hence $(1 - x)/y = s/C$, and we get

$$(x - 1)^2 + (y - 1/2C)^2 = 1/4C^2,$$

the circle of radius $1/2C$ centred at $1 + \iota 1/2C$ *i.e.*, the circle of radius $1/2C$ tangent to the real axis at 1; the image of the horizontal line $\Im\tau = C$ is the complement in the circle of the point 1 of tangency. This circle is called the *horocircle* at 1 with radius $1/2C$. We see at once that the image of $\Im\tau \geq C$ is a union of horocircles *i.e.*, is the closed disc of radius $1/2C$ tangent to the real axis at 1, with the point of tangency removed. Next, consider the images on the horocircle of radius $1/2C$ of the points $\pm 1 + \iota C$ on the line $\Im\tau = C$. Substituting in the above formulas, we see that

$$1 + \iota C \longmapsto \frac{C^2}{1 + C^2} + \iota \frac{C}{1 + C^2}, \quad -1 + \iota C \longmapsto \frac{2 + C^2}{1 + C^2} + \iota \frac{C}{1 + C^2}.$$

The first image point lies within the unit disc, while the second image point has real part > 1. Hence $\mathcal{F}_1(C)$ consists of the intersection of $\mathcal{F}$ with the closed disc bounded by the horocircle. In particular, for some $\varepsilon > 0$, $\mathcal{F} \cap \{\Im\tau < \varepsilon\} \subset \mathcal{F}_1(C)$. Thus $\mathcal{F} = \mathcal{F}_1(C) \cup \mathcal{F}(\varepsilon)$. Since $h(\tau)$ is bounded on each of the sets $\mathcal{F}_1(C)$ and $\mathcal{F}(\varepsilon)$, $h(\tau)$ is bounded on $\mathcal{F}$, and hence on $\mathcal{H}$, as claimed. This proves the proposition. □

We now estimate the Fourier coefficients a_n of $G(\tau)$ as follows. Let $k(x) = G(x + \iota y)$ where $y > 0$ is fixed. Then

$$k(x) = \sum_{n=1}^{\infty} a_n(y) e^{\pi \iota n x},$$

where

$$a_n(y) = a_n e^{-\pi n y}.$$

Now the Fourier coefficients of $k(x)$ satisfy

$$a_n(y) = 1/2 \int_0^2 k(x) e^{-\pi \iota n x} dx.$$

Hence

$$\begin{aligned} 2\,|a_n(y)| &\leq \int_0^1 |k(x)|\,dx = \int_0^2 |G(x+\iota y)|\,dx \\ &= y^{-k} \int_0^2 y^k\,|G(x+\iota y)|\,dx = y^{-k} \int_0^2 h(x+\iota y) dx \leq C y^{-k} \end{aligned}$$

for some constant C independent of y. Thus

$$a_n \leq C y^{-k} e^{\pi n y} \quad \forall\ y > 0.$$

Taking $y = 1/n$ on the right, we get $a_n = O(n^k)$, as desired.

Remark: In fact the *Ramanujan-Petersson conjecture*, proved by P. Deligne as a consequence of his proof of Weil's Riemann hypothesis for varieties over finite fields, implies that for any cusp form $G(\tau)$ of weight k, the Fourier coefficients a_n satisfy $a_n = O(n^{\frac{k-1}{2}+\epsilon})$.

We are left with showing that if $G(\tau)$ is a cusp form of weight 2 or 4, then $G = 0$. More generally, let $G(\tau)$ be a non-zero cusp form of weight $2k$. We will prove a formula relating the weight $2k$ and the number of zeroes of G in $\mathcal{F}$ and at the 'cusps' 1 and infinity (the zeroes have to be 'counted properly'). This formula will show there are no non-zero forms of weight 2 or 4 as stated above.

We say that $G(\tau)$ has a zero of order $n_\infty > 0$ at infinity if G has a Fourier expansion

$$G(\tau) = \sum_{n \geq n_\infty} a_n e^{\pi \iota n \tau}$$

where $a_{n_\infty} \neq 0$. Similarly, $G(\tau)$ is said to have a zero of order $n_1 > 0$ at 1 if

$$\tau^{-2k} G(1 - 1/\tau) = \sum_{n \geq n_1} a_n e^{\pi \iota n \tau}$$

where $a_{n_1} \neq 0$. A cusp form has a zero of order > 0 at both 1 and infinity, by definition. At any point of $\mathcal{F}$, G is an analytic function, so it makes sense to speak of the order of its zero at any point of $\mathcal{F}$.

The Fourier expansion of G shows that it is obtained as $G(\tau) = h(e(\tau))$ for some analytic function h on the unit disc Δ, and the origin is an isolated zero of h; hence G has no zeroes in $\mathcal{F} \cap \{\Im\tau > C\}$ for sufficiently large $C > 0$. Similarly, the Fourier expansion of $\tau^{-2k}G(1 - 1/\tau)$ shows that this function is obtained from an analytic function on a disc centred at 0, and 0 is an isolated zero of this function; this implies that for sufficiently small $\varepsilon > 0$, G has no zeroes in $\mathcal{F} \cap \{\Im\tau < \varepsilon\}$.

Consider the following contour $\mathcal{C}$ in $\mathcal{H}$. It consist of the vertical line segment from $1 + \iota\varepsilon$ to $1 + \iota C$, the horizontal line segment from $1 + \iota C$ to $-1 + \iota C$, the vertical line segment from $-1 + \iota C$ to $-1 + \iota\varepsilon$, a small arc from this vertical line to the semicircle $\{\tau \mid |\tau| = 1, \Im\tau > 0\}$, the semicircle itself with small arcs removed at $\tau = \pm 1$, and finally a small arc joining the end of the semicircle near $\tau = 1$ to the vertical line segment beginning at $1 + \iota\varepsilon$.

This contour is indented along the vertical lines and the semicircle around any zero of G. The indentation is always to the right of the zero on each of the vertical lines, above the zeros at ι and $(1 + \sqrt{-3})/2$, and below the zero at $(-1 + \sqrt{-3})/2$. The picture is as in Figure 18.1.

Let N be the number of zeroes of G within the contour. Let n_ι denote the order of the zero (if any; else $n_\iota = 0$) at $\tau = \iota$.

Proposition 18.3.3 *If G is a non-zero cusp form of weight $2k$, then*

$$N + n_\infty + \frac{n_1}{2} + \frac{n_\iota}{4} = \frac{k}{2}.$$

In particular, $k \geq 3$.

Proof: First note that the formula implies $k \geq 3$, since $n_\infty, n_1 \geq 1$.

To prove the formula, we compute the integral of $\dfrac{\mathrm{dlog}\, G}{2\pi\iota}$ around the contour $\mathcal{C}$. From the residue theorem,

$$\frac{1}{2\pi\iota}\int_{\mathcal{C}} \mathrm{dlog}\, G = N.$$

Since G is periodic with period 2 and the (possibly indented) vertical line segments differ by a translation by 2, and are oriented in opposite directions in $\mathcal{C}$, the integrals along these segments cancel.

Figure 18.1: The contour $\mathcal{C}$

Consider the integral along the horizontal segment joining $1+\iota C$ to $-1+\iota C$. From the Fourier expansion of G,

$$G = e^{\pi\iota n_\infty \tau} G_1$$

where G_1 has a non-zero finite limit at infinity, say α, we see that

$$\frac{1}{2\pi\iota}\,\mathrm{dlog}\, G = \frac{n_\infty}{2} + \frac{1}{2\pi\iota}\,\mathrm{dlog}\, G_1.$$

Now $G_1 = \alpha + e^{\pi\iota\tau} G_2$ from the Fourier expansion of G_1, where G_2 is bounded as $\Im\tau \to \infty$. Hence

$$\mathrm{dlog}\, G_1 = \frac{1}{G_1} d\,(G_1 - \alpha) = \frac{e^{\pi\iota\tau}}{G_1}(\pi\iota G_2 d\,\tau + d\,G_2).$$

Now the integral of $\mathrm{dlog}\, G_1$ tends to 0 as $C \to \infty$, since $|e^{\pi\iota\tau}| = e^{-\pi\Im\tau} \to 0$. Hence, as $C \to \infty$, the integral along the horizontal segment tends to $-n_\infty$.

Similarly, we estimate the integral along the small arc joining the point $-1+\iota\varepsilon$ to a nearby point on the semicircle given by $|\tau| = 1$. We make a change of variables $\tau' = 1/(1-\tau)$, so that $\tau = 1 - 1/\tau'$. Hence

$$G(\tau) = \tau'^{-2}(\tau'^2 G(1 - 1/\tau')) = \tau'^{-2} e^{\pi\iota n_1 \tau'} G_1(\tau')$$

where G_1 has a nonzero limit as $\Im\tau' \to \infty$. Now $\Re\tau = 1 \Leftrightarrow \Re\tau' = 0$, so that $\tau \to \tau'$ maps the vertical line $\Re\tau = 1$ to the vertical line $\Re\tau' = 0$. Similarly, if $|\tau| = 1$, then

$$\Re\tau' = \Re\left(\frac{1}{1-\tau}\right) = \Re\left(\frac{1-\overline{\tau}}{(1-\tau)(1-\overline{\tau})}\right) = \frac{1-\Re\tau}{1-2\Re\tau+|\tau|^2} = \frac{1}{2}.$$

Thus $\tau \longmapsto \tau'$ maps the semicircle $|\tau| = 1$ onto the vertical line $\Re\tau' = 1/2$. We choose the small arc near 1 in $\mathcal{C}$ to be the transform under $\tau' \longmapsto \tau$ of a horizontal line segment joining the vertical line $\Re\tau' = 1/2$ to the line $\Re\tau = 0$, at a height $1/\varepsilon$. As in the case of n_∞, we see that the integral over this arc of $(2\pi\iota)^{-1}\,\mathrm{dlog}\,G$ tends to $-n_1/2$ as $\varepsilon \to 0$. Similarly, the other small arc near -1 contributes $-n_1/2$ in the limit as $\varepsilon \to 0$.

If $n_\iota > 0$, then $\mathcal{C}$ has an indentation at $\tau = \iota$, which is approximately a small semicircle around ι. Now $\tau \longmapsto -1/\tau$ maps the indentation to a curve within the unit disc which joins the end points of the indentation, giving a closed contour around $\tau = \iota$; the integral around this contour of $(2\pi\iota)^{-1}G$ is n_ι. Since $G(-1/\tau) = (-1)^k\tau^{2k}G(\tau)$, we see that in the limit as the radius of the indentation tends to 0, the integral around the indentation tends to $n_\iota/2$.

Finally it remains to compute the integral around the semicircle $\mathcal{H} \cap \{|\tau| = 1\}$. We divide this into the two quarter circles given by $\Re\tau > 0$ and $\Re\tau < 0$. Again using $G(-1/\tau) = (-1)^k\tau^{2k}G(\tau)$, and the fact that $\tau \longmapsto -1/\tau$ interchanges the two arcs, reversing orientation, the sum of the integrals around the two arcs is the integral

$$\frac{1}{2\pi\iota}\int_1^\iota \mathrm{dlog}\,\tau^{2k}$$

(the integral is from 1 to ι since it is obtained by transforming the integral from -1 to ι). This integral equals $(2k)/4 = k/2$ since we are integrating over a quarter circle.

Thus we have shown that

$$N = -n_\infty - n_1 - n_\iota/2 + k/2,$$

which is equivalent to the claimed formula. □

Corollary 18.3.4 *(i) Let G_1, G_2 be cusp forms of weight $2k$ with Fourier expansions*

$$G_1 = \sum_{n=1}^{\infty} a_{1,n} e^{\pi \imath n \tau}, \quad G_2 = \sum_{n=1}^{\infty} a_{2,n} e^{\pi \imath n \tau}.$$

Suppose $a_{1,n} = a_{2,n}$ for $1 \leq n \leq (k/2) - 1$. Then $G_1 = G_2$.

(ii) The cusp forms of weight $2k$ form a $\mathbf{C}$-vector space of dimension $\leq (k/2) - 1$.

We end with several exercises for the reader.

Exercise 18.3.1: In this exercise we show that the function $E_2^*(\tau)$ has the required transformation properties and defines a holomorphic function on the upper half plane.

1. Recall that the

$$\wp(z, \omega_1, \omega_2) = z^{-2} + \sum_{m,n}{}' \left(\frac{1}{(z - m\omega_1 - n\omega_2)^2} - \sum_{m,n}{}' \frac{1}{(m\omega_1 + n\omega_2)^2} \right).$$

   Show that

$$\forall \lambda \in \mathbf{C}^*, \wp(\lambda z, \lambda\omega_1, \lambda\omega_2) = \lambda^{-2} \wp(z, \omega_1, \omega_2).$$

2. Show that the function $f(\tau) = \wp(1/2, 1, \tau) - \wp(\tau/2, 1, \tau)$ satisfies the properties $f(\tau + 2) = f(\tau)$ and $f(-1/\tau) = -\tau^{-2} f(\tau)$.

3. Show that there is a constant $C \neq 0$ such that $E_2^*(\tau) = C f(\tau)$.

Exercise 18.3.2: 1. Verify that $\theta(\tau, 1/2 + \tau/2) = 0$.

2. Verify that the function $\phi(\tau, z) = e^{az^2 + bz} \sigma(\tau, z)$ satisfy $\phi(\tau, z + 1) = \phi(\tau, z)$ if $a = \eta_1/2$ and $b = \pi i$ (for the definition of η_1 see Chapter (1)). For these a, b, verify that $\phi(\tau, z + \tau) = -e^{2\pi i z} \phi(\tau, z)$ using the so called Legendre relation $\eta_1 \tau - \eta_\tau = 2\pi i$ (cf. Ahlfors, Ch. 7, Section 3.2).

3. Let $\psi(\tau, z) = \phi(\tau, z + 1/2 + \tau/2)$. Verify that $\psi(\tau, z + \tau = e^{-2\pi i z - \pi i \tau} \psi(\tau, z)$ and deduce that $\theta(\tau, z)/\psi(\tau, z)$ is a doubly periodic function of z and has no poles and hence is a constant.

Exercise 18.3.3: Using Euler's formula and the recurrence formula for Bernoulli numbers, compute $\zeta(2k)$ for $k = 1, 2, 3, 4, 5$.

Exercise 18.3.4: Let

$$f(t,x) = \frac{te^{(1+x)t}}{e^t - 1} = \sum_{n=0}^{\infty} B_n(x)\frac{t^n}{n!}.$$

1. Show that $f(t,x) = f(t,0)e^{tx}$. Deduce that

$$B_n(x) = \sum_{i=0}^{n} \left(\begin{array}{c} n \\ i \end{array} \right) B_i x^{n-i}$$

where B_n is the n^{th} Bernoulli number. In particular, $B_n(0) = B_n$. $B_n(x)$ is called the n^{th} *Bernoulli polynomial.*

2. For any integer $k \geq 0$, let

$$S_k(n) = \sum_{a=1}^{n} a^k,$$

the sum of the first n k^{th} powers. Show that

$$S_k(n) = \frac{1}{k+1}\left(B_{k+1}(n) - B_{k+1}(0)\right).$$

(Hint: Show first that $f(t,x) - f(t,x-1) = te^{tx}$. Deduce that $B_n(x) - B_n(x-1) = nx^{n-1}$. Now replace n by $n+1$, and give suitable values to x ....)

Exercise 18.3.5: The Eisenstein series introduced in the lectures are satisfy transformation formulas associated to elements of the subgroup $\Gamma \subset SL(2, \mathbf{Z})$, generated by the matrices

$$\gamma_1 = \left[\begin{array}{cc} 1 & 2 \\ 0 & 1 \end{array} \right], \quad \gamma_2 = \left[\begin{array}{cc} 0 & -1 \\ 1 & 0 \end{array} \right].$$

One has Eisenstein series ('the' Eisenstein series) associated to the full group $SL(2, \mathbf{Z})$, defined as follows:

$$E_{2k}(\tau) = \sum_{\substack{m,\, n \in \mathbf{Z} \\ (m,n) \neq (0,0)}} \frac{1}{(m\tau + n)^{2k}}, \quad k \geq 1.$$

1. Show that the series for $E_{2k}(\tau)$ converges absolutely and uniformly on compact subsets of $\mathcal{H}$ for $k \geq 2$. For $k = 1$, show that it converges uniformly on compact subsets of $\mathcal{H}$ if we sum first over n and then over m.

2. Show that the Weierstarss $\wp$ function $\wp(z, \tau)$, considered as a function on $\mathbf{C} \times \mathcal{H}$, may be expanded in a Laurent series (about $z = 0$)

$$\wp(z, \tau) = \frac{1}{z^2} + \sum_{k=1}^{\infty} (2k+1) E_{2k+2}(\tau) z^{2k}.$$

3. Show that $g_2(\tau) = 60E_4(\tau)$, $g_3(\tau) = 140E_6(\tau)$; find an expression for Δ in terms of E_4 and E_6.

4. Show that $E_{2k}(\tau)$ satisfies the transformation formula

$$E_{2k}\left(\frac{a\tau + b}{c\tau + d}\right) = (c\tau + d)^{2k} E_{2k}(\tau),$$

for all $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbf{Z})$, provided $k \geq 2$.

5. Show that $E_{2k}(\tau)$ has a Fourier expansion ("at infinity")

$$E_{2k}(\tau) = 2\zeta(2k) + \frac{2(2\pi\iota)^{2k}}{(2k-1)!} \sum_{n=1}^{\infty} \sigma_{2k-1}(n) e^{2\pi\iota n\tau},$$

where

$$\sigma_r(n) = \sum_{d|n} d^r$$

is the r^{th} divisor function.
The transformation formula under elements of $SL(2, \mathbf{Z})$ and the Fourier expansion imply that $E_{2k}(\tau)$ is, by definition, a *modular form* of weight $2k$ for $SL(2, \mathbf{Z})$. Hence g_2, g_3 are modular forms of weights 4, 6 respectively.

6. Show that $\Delta(\tau)$ is a modular form of weight 12 for $SL(2, \mathbf{Z})$ with Fourier expansion

$$\Delta(\tau) = (2\pi)^{12} e^{2\pi\iota\tau} \left(! + \sum_{n=1}^{\infty} \tau(n) e^{2\pi\iota n\tau} \right),$$

where the $\tau(n)$ are *integers*.
Since Δ is a modular form for $SL(2,\mathbf{Z})$ whose zeroeth Fourier coefficient (at infinity) vanishes, Δ is by definition a *cusp form* for $SL(2,\mathbf{Z})$. The notation $\tau(n)$ for the coefficients is due to Ramanujan; the *Ramanujan Conjecture* (which is generalised to the Ramanujan-Petersson Conjecture, proved by Deligne) is that $\tau(n) = O(n^{11/2+\epsilon})$ for any $\epsilon > 0$.
(Hint: Substitute the known values for $\zeta(4)$, $\zeta(6)$ in the Fourier expansions for E_4, E_6.)

7. Show that the normalised j-function

$$J(\tau) = 12^3 \frac{g_2^3}{\Delta}$$

has a Fourier expansion

$$J(\tau) = \frac{1}{q} + \sum_{n=0}^{\infty} a_n q^n, \text{ with } q = e^{2\pi\iota\tau},$$

where the a_n are integers.

Exercise 18.3.6: Let $\mathcal{F}_0$ be the subset of $\mathcal{H}$ defined by

$$\mathcal{F}_0 = \{\tau \in \mathcal{H} \mid -1 \leq \Re\tau < 1 \text{ and } |\tau| \geq 1, \text{ where if } |\tau| = 1, \text{ then } \Re\tau \leq 0\}.$$

Let $\Gamma \subset SL(2,\mathbf{Z})$ be the subgroup defined by the matrices

$$\gamma_1 = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \quad \gamma_2 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Show that if $x, y \in \mathcal{F}_0$ and $\gamma \in \Gamma$ such that $\gamma x = y$, then $x = y$. What are the possibilities for γ?
(Hint: Reason as in Gurjar's lectures giving the fundamental domain R^0 for the action of $SL(2,\mathbf{Z})$ on $\mathcal{H}$).

Exercise 18.3.7: Let $\Gamma \subset SL(2,\mathbf{Z})$ be the subgroup defined by the matrices

$$\gamma_1 = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \quad \gamma_2 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Show that Γ acts properly discontinuously on $\mathcal{H}$, and show that the quotient Riemann surface is biholomorphic to $\mathbf{C}^*$.

Chapter 19

The Conjecture of Birch and Swinnerton-Dyer

In this chapter we want to motivate certain conjectures made by Birch and Swinnerton-Dyer, concerning the rank of the Mordell-Weil group of an elliptic curve. This chapter is intended as an introduction to an area of intense research today. We will briefly touch upon the recent developments in the subject towards the end of this chapter. Because of the lack of adequate machinery from algebraic number theory, we will restrict ourselves to stating most of the theorems over $\mathbf{Q}$. We have also kept the style of the chapter less formal than before.

19.1 Introduction

Let $E : y^2z = x^3 + axz^2 + bz^3$ be an elliptic curve over $\mathbf{Q}$ with $a, b \in \mathbf{Z}$. Fix a prime number p. Then we can reduce the equation of E modulo p. This gives us a cubic polynomial defined over the finite field $\mathbf{Z}/p\mathbf{Z}$. By Section (16.4) we know that the curve defined by this polynomial is an elliptic curve iff the discriminant Δ_E of E is not divisible by p. If p divides the discriminant of E then the reduced polynomial gives rise to singular cubic. The results of Chapter (10), tell us that any singular cubic can be transformed to an equation of the form (cusp) or a (node). In any case we shall use the notation E_p to denote the reduced curve.

Moreover, if $P = (a_0 : a_1 : a_2)$ are the normalised homogenous coordinates of P (*i.e.*, a_0, a_1, a_2 are coprime integers), then we can reduce a_0, a_1, a_2 modulo

p. This gives us a point on the reduced curve. In fact this gives a map of sets $E(\mathbf{Q})E_p(\mathbf{Z}/p\mathbf{Z})$, which we call the *reduction modulo p* or *specialisation*; in the case when $p \not| \Delta$, it was proved in Section (16.4) that it is in fact a homomorphism of groups.

For each p we define a natural number $N_p(E_p)$ as follows:

$$N_p(E_p) = \begin{cases} \#E_p(\mathbf{F}_p) & \text{if } p \not| \Delta \\ p & \text{if } E_p \text{ is a node and two tangents defined over } \mathbf{F}_p \\ p+2 & \text{if } E_p \text{ has is a node with tangents defined over } \mathbf{F}_{p^2}, \\ p+1 & \text{if } E_p \text{ is a cusp.} \end{cases} \tag{19.1}$$

In Section (13.5) it was shown that for $p \not| \Delta$ the $N_p(E_p)$ satsify the inequality

$$\|N_p - p - 1| \leq 2\sqrt{p}.$$

Now suppose E had lots of rational points on it. Then for most p, the number of points modulo p, N_p would also tend to be very large. Write $N_p = p + 1 - a_p$, where the bound on N_p says that $\|a_p| \leq 2\sqrt{p}$. If N_p was large for almost all primes then one might expect that N_p/p is bigger than 1 most of the time. So that we can expect the product $\prod_{p \leq X} N_p/p$ to tend to infinity with X. This however is only heuristic and is impossible to justify. However, Birch and Swinnerton-Dyer (see [6]), inspired by some work of Siegel on quadratic forms, looked at the behaviour of the product

$$f(X) = \prod_{p \leq X} N_p/p$$

as a function of X. In fact they computed the values of this function for several hundred elliptic curves and found that this product tends to be significantly large when the curve had infinitely many solutions (apparently in agreement with our naive heuristic). In fact based on their computations they made the following conjecture about the behaviour of $f(X)$ as a function of X and the rank of E.

Conjecture 2 (Birch–Swinnerton-Dyer (I)) *Let E be an elliptic curve over $\mathbf{Q}$ and r be its rank. Then there is positive constant $C(E)$ depending on E such that*

$$f(X) = \prod_{p \leq X} N_p/p \sim C(E)(\log X)^r \qquad \text{as } X \longrightarrow \infty. \tag{19.2}$$

This is not the most natural form of the conjecture as we shall soon see; nor is it a practical one. For, the product grows far too slowly to be useful in computing the rank. So one has to look for a natural setting for this conjecture. To set up this formulation we need some machinery.

19.2 Zeta functions over finite fields

Let E be an elliptic curve defined over a finite field $\mathbf{F}_q$. Let us define the ζ-function of E ("zeta function of E"). We can think of E as a curve defined over any finite extension of $\mathbf{F}_q$ and count the number of points on E for all extensions $\mathbf{F}_{q^n}$ of $\mathbf{F}_q$. This furnishes us with a sequence of natural numbers $N_{q,n}(E) = \#E(\mathbf{F}_{q^n})$. Note that $N_{q,1} = N_q(E)$.

Definition: The Zeta-function of $E/\mathbf{F}_q$ is the "formal" power series in T given by

$$\log \zeta(E/\mathbf{F}_q, T) = \sum_{n \geq 1} N_{q,n}(E) T^n/n.$$

Though this definition looks very artificial, it should be thought of as a sort of generating series for the numbers $N_{q,n}(E)$.

Remark: More generally, if V is any variety defined over $\mathbf{F}_q$, then we can similarly define $N_{q,n}(V) = \#V(\mathbf{F}_{q^n})$. And as above, we define:

$$\log \zeta(V/\mathbf{F}_q, T) = \log \zeta(V, T) = \sum_{n \geq 1} N_{q,n}(V) T^n/n;$$

we call $\zeta(V/\mathbf{F}_q, T)$ the zeta function of V.

As a simple example let us calculate the ζ-function of $\mathbf{P}^1$ over $\mathbf{F}_q$. A simple argument shows that $\#N_{q,n}(\mathbf{P}^1) = q^n + 1$. Then $\log \zeta(\mathbf{P}^1/\mathbf{F}_q, T) = \sum_{n \geq 1}(q^n + 1)T^n/n = \sum_n (qT)^n/n + \sum_n T^n/n = -\log(1 - qT) - \log(1 - T)$. Thus

$$\zeta(\mathbf{P}^1/\mathbf{F}_q, T) = 1/(1 - T)(1 - qT).$$

Notice that because of the 'log' in the definition of zeta function the final form of the answer is simple! The ζ-function is a rational function of T.

Exercise 19.2.1: Show that $\zeta(\mathbf{P}^n/\mathbf{F}_q, T) = \prod_{0 \leq i \leq n}(1 - q^i T)^{-1}$.

Exercise 19.2.2: Show that $\zeta(\mathbf{A}^n/\mathbf{F}_q, T) = (1 - q^n T)^{-1}$.

It turns out that for any elliptic curve defined over a finite field $\mathbf{F}_q$, $\zeta(E/\mathbf{F}_q, T)$ is of the form:

$$\zeta(E/\mathbf{F}_q, T) = (1 - a_q T + qT^2)/(1 - T)(1 - qT), \tag{19.3}$$

where a_q depends on E. Moreover, we also have $N_q = q+1-2a_q$. Note that if we substitute $T = q^{-s}$ in the polynomial $1-a_qT+qT^2$ we get $1-a_qq^{-s}+q^{1-2s}$. In particular we see that if we put $s = 1$ in this expression then we get N_q/q.

Remark: Let E, E' be elliptic curves over $\mathbf{F}_q$. Suppose $E \longrightarrow E'$ is an isogeny defined over $\mathbf{F}_q$. Then

$$\zeta(E/\mathbf{F}_q, T) = \zeta(E'/\mathbf{F}_q, T).$$

Proof: This follows easily from Lemma (13.2.1). □

Remark: [Historical Aside]

The Zeta function of an elliptic curve dates back to Emil Artin's doctoral thesis. He computed it in special cases for a large number of elliptic curves and other curves. He also conjectured the rationality and the general form of the Zeta function of an elliptic curve. These conjectures were proved by Hasse [28].

Later Weil [60] considered the Zeta function of an arbitrary variety over a finite field. He also computed it for a large class of varieties in this paper. In the same paper he conjectured that the zeta function of any variety over $\mathbf{F}_q$ is a rational function; and the Zeta function satisfies the "Riemann Hypothesis" (this is only a part of the conjecture). More precisely, the conjecture can be stated (in a special case) as follows:

Theorem 19.2.1 *Let $V/\mathbf{F}_q$ be a smooth projective variety over $\mathbf{F}_q$. Then*

1. *$\zeta(V/\mathbf{F}_q, T)$ is a rational function of T, with integer coefficients.*
2. *If we write $\zeta(V/\mathbf{F}_q, T) = P(T)/Q(T)$, where P, Q are polynomials in T with integer coefficients. Then for any root α of P or Q we have $\|\alpha|^2 = q^n$, for some integer n satisfying $0 \leq n \leq 2\dim X$. This is known as the "Riemann Hypothesis".*

The last statement is analogous to the famous Riemann Hypothesis [8] for the Classical Zeta function $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$, which states that all the "non-trivial" zeros of the $\zeta(s)$ satsify $\Re(s) = 1/2$. This analogy can be seen by substituting $T = q^{-s}$. Then the statement about zeros of P, Q reads as zeros of $P(q^{-s}), Q(q^{-s})$ are on the line $\Re(s) = n/2$ for some $0 \leq n \leq 2 \dim X$.

We have now stated the conjectures as theorems because they have now been proved in full generality. The rationality of the zeta function was proved by Dwork [17] using p-adic methods and also by Grothendieck using his theory of étale cohomology. The "Riemann Hypothesis" was proved by Deligne (see [14], [15]).

19.3 Reformulating the conjecture

Coming back to our discussion of the Birch–Swinnerton-Dyer conjecture, we are now ready to cast it in a more interesting form. From the results of Section (19.2) we can conclude that the product $f(X) = \prod_{p \leq X} N_p/p$ is the value at $s = 1$ of the product

$$g(X, s) = \prod_{p \leq X} \left(1 - a_p p^{-s} + p^{1-2s}\right).$$

So the conjecture of Birch and Swinnerton-Dyer can be restates as :

$$\text{as } X \longrightarrow \infty, \qquad 1/g(X,1) \begin{cases} \text{constant}, & \text{if } E \text{ has rank } 0, \\ 0 & \text{if } E \text{ has rank } \geq 1. \end{cases}$$

In fact since we are interested in dealing with all the primes p, we introduce the function

$$L(E, s) = \prod_{p|\Delta} (1 - a_p p^{-s})^{-1} \prod_{p \not| \Delta} (1 - a_p p^{-s} + p^{1-2s})^{-1}. \tag{19.4}$$

where a_p is defined by $a_p = p + 1 - N_p$. This is called the *L-function* of E or sometimes the "global zeta function" of E.

As a function of s we first note that the product makes sense in a suitable range:

Proposition 19.3.1 *The product $L(E, s)$ converges absolutely for $\Re(s) > 3/2$.*

Proof: We just have to note that as $\|a_p| \leq \sqrt{q}$,

$$\|a_p p^{-s} + p^{1-2s}| \leq \|a_p p^{-s}| + \|p^{1-2s}| \leq p^{-\Re(s)+1/2} + p^{1-2\Re(s)}. \tag{19.5}$$

Then it is easy to see that the series

$$\sum_p p^{-\Re(s)+1/2}, \sum_p p^{1-2\Re(s)}$$

converge for $\Re(s) > 3/2$. Then, as the sum of these two series dominates $\sum_p (a_p p^{-s} + p^{1-2s})$, the latter converges whenever $\Re(s) > 3/2$. And hence the product converges in that range. □

The function $L(E, s)$ is in fact a complex analytic function of s in that range. One would however like to define (or analytically/meromorphically continue) $L(E, s)$ in the entire complex plane. But first note that the conjectures of Birch and Swinnerton-Dyer suggests that this function, as a function of s, should vanish at $s = 1$ if and only if the rank of E is at least one! But unfortunately we do not know (in general) how to define the function in the region left of $\Re(s) = 3/2$. So now one is in a bit of a mess: we want to conjecture that the value of a function is zero at a point where it is not even defined!

So the first question is: Does the L-function of any $E/\mathbf{Q}$ have an analytic (or a meromorphic) continuation to all of $\mathbf{C}$?. The following conjecture is a very precise formulation of this problem:

Conjecture 3 *Let $E/\mathbf{Q}$ be an elliptic curve. Then there is a positive integer N which divides the discriminant, such that the complex valued function $\Lambda(s) = N^{s/2}(2\pi)^{-s}\Gamma(s)L(E/\mathbf{Q}, s)$ is holomorphic in the entire complex plane and satisfies the functional equation*

$$\Lambda(s) = \pm\Lambda(2 - s) \tag{19.6}$$

Remark: This conjecture is known in the case when $E/\mathbf{Q}$ is an elliptic curve with complex multiplication and was proved by Deuring. Thus for instance as the curve $E_n : y^2 = x^3 - n^2x$ is a curve with complex multiplication and so that the $L(E_n, s)$ has analytic continuation to all of $\mathbf{C}$ and it also has a functional equation with $N = 32n^2$ if n is odd and $N = 16n^2$ if n is even

(cf [31]). The conjecture is also known for a special class of elliptic curves called "modular elliptic curves".

One hopes optimistically that the behaviour of the $L(E, s)$ of an arbitrary elliptic curve is not far different from that of an elliptic curve with complex multiplication – at least as far as analyticity properties are concerned. The number N which appears is usually called the "conductor" of E.

Armed with this information and assuming hence forth that $L(E/\mathbf{Q}, s)$ satisfies the properties as above, we can now state a slightly more refined version of conjecture (2) which brings out the precise inter-relationship between the rank of E and the vanishing of its L-function (see [6]):

Conjecture 4 (Birch–Swinnerton-Dyer (II)) *Let E be an elliptic curve over $\mathbf{Q}$. Let r be its rank. Then the L-function of E, $L(E, s)$ has a zero of order r at $s = 1$.*

This form of the conjecture says that the the function $L(E, s)$ vanishes at $s = 1$ if and only if the rank is positive and the order of the zero at this point is exactly equal to the rank of E. In practice one can compute the L-function at $s = 1$ (*i.e.*, whenever the L-function has analytic continuation and functional equation), with considerable ease using a computer and with algorithms which facilitate rapid and accurate computation of this value.

Now as a function of s we can write

$$L(E, s) = a_r(s-1)^r + \text{higher order terms in } (\mathrm{s}-1),$$

where $r = \operatorname{rank} E$. Then one question comes up naturally: *what is the arithmetical significance of the constant a_r?* It turns out, at least conjecturally, that this constant carries a lot of information about the elliptic curve. This will then lead us to the final form of the Birch–Swinnerton-Dyer conjecture. The whole gamut of invariants associated with elliptic curves is involved in this final form.

19.4 Final Form

The statement of the final form of the conjecture will require the definition of some more invariants of elliptic curves. We will now spend some time defining them.

The elliptic curve $E/\mathbf{Q}$ comes equipped with a (regular) differential form ω, when thought of as a curve over complex numbers. The choice of ω is unique upto a scalar. However, over the rational numbers, it is possible to choose an equation of an elliptic curve, so that it is discriminant is as small as possible[1], which is then essentially unique. For such an equation, the choice of the form ω is ambigous only upto a sign.

The set of real valued points on E is compact and hence we can integerate this form over the set of real points to get a number, denoted ω_∞:

$$\omega_\infty = \int_{E(\mathbf{R})} \omega.$$

Thus by Section (5.3) ω_∞ is either the real period E or twice the real period depending on whether $E(\mathbf{R})$ has one or two connected components.

For any prime p, we can consider the set of points in $E(\mathbf{Q}_p)$ whose image in E_p is a non singular point of E_p (recall that E_p is the reduction of E modulo p). The set of such points of $E(\mathbf{Q}_p)$ is usually denoted by $E_0(\mathbf{Q}_p)$. It can be shown using methods of Chapter (16.4) that $E_0(\mathbf{Q}_p)$ *is a subgroup of finite index in* $E(\mathbf{Q}_p)$. *Denote this index by* c_p. *Note that if* $p \not| \Delta$, *then* $c_p = 1$.

We now define another invariant of the Mordell-Weil group called the *Regulator* of E, denoted R_E. Firstly, if rank of $E = 0$ then we set $R_E = 1$. Otherwise let $P_1, P_2, \ldots, P_r$ be generators for the Mordell-Weil group of E. In Chapter (15.2) we defined a natural height function on E, denoted by $_E$. Recall that this is a real valued function $_E : E(\mathbf{Q})\mathbf{R}$ with the property that the function $E(\mathbf{Q}) \times E(\mathbf{Q})\mathbf{R}$ (called the *height pairing*) defined by

$$< P, Q >=_E (P + Q) -_E (P) -_E (Q)$$

is bilinear as a function of P, Q. This gives us a matrix whose entries are $< P_i, P_j >$ for $1 \leq i, j \leq r$. This matrix is called the *regulator matrix* of E. We define $\mathbf{R}_E$ to be determinant of this matrix. One checks easily that this is independent of the choice of generators.

Suppose now that $Ш(E)$ is a finite group, let $[Ш]$, $[\mathrm{Tor}\, E(\mathbf{Q})]$ denote the orders of the Tate-Shafarevich group and the torsion subgroup of $E(\mathbf{Q})$ respectively.

We are now ready for the final form of the Birch Swinnerton-Dyer conjecture (see [7]).

[1]This is what is called the global minimal model of E.

Conjecture 5 *Let $E/\mathbf{Q}$ be an elliptic curve over $\mathbf{Q}$. Then in the notations as above we have:*

$$\lim_{s1} L(E,s)/(s-1)^{r_E} = \frac{[Ш]}{[\mathrm{Tor}\, E(\mathbf{Q})]^2} R_E \omega_\infty \prod_{p|\Delta} c_p \tag{19.7}$$

Note that the order of Ш has made an appearence in the final form of the conjecture. The whole conjecture was built out of the extensive numerical evidence gathered by Birch and Swinnerton-Dyer and others. the role of computers in this area has been quite significant. Most of the calculations – like descent can be easily programmed and the value of the L-function can also be computed with desired accuracy.

Note that in the final form of the conjecture one had to assume that the Ш was finite. But with the work of Kolyvagin [32] and Rubin[47], there is renewed hope that Ш is finite for any elliptic curve. Rubin's theorem (for elliptic curves over $\mathbf{Q}$) can be stated as follows.

Theorem 19.4.1 *Let $E/\mathbf{Q}$ be an elliptic curve with complex multiplication. If $L(E/\mathbf{Q},1) \neq 0$, then $Ш(E/\mathbf{Q})$ is finite.*

Kolyvagin's work however is in a slighly different direction. Unfortunately, a description of his work would take us far beyond the scope of this book.

Thus we see that the conjecture of Birch and Swinnerton-Dyer not only gives us a way of computing the rank of E but also the order of the Tate-Shafarevich group Ш. Moreover, the geometry over complex number and over finite fields and over the p-adic fields is intimately linked with the arithmetic of E. The order of vanishing of the L-function at $s=1$ is the rank of E. This function thus captures an enormous ammount of information about the elliptic curve.

In the direction of the Birch–Swinnerton-Dyer conjecture the earliest progress came in the work of Coates and Wiles [10]. They proved the following:

Theorem 19.4.2 *Let $E/\mathbf{Q}$ be an elliptic curve over $\mathbf{Q}$ with complex multiplication by an imaginary quadratic field K. Suppose that the class number of K is one (i.e., the ring of integers of K is a UFD) and suppose that $r_E \geq 1$. Then*

$$L(E/\mathbf{Q},1) = 0.$$

So for instance, this covers certain cases like $y^2 = x^3 - dx$ which has complex multiplication by $\mathbf{Q}(i)$ (and as is well known, $\mathbf{Q}(i)$ has class number 1). The ranks of these curves were computed by Birch and Swinnerton-Dyer in a large number of cases.

Recently Rubin [47] has proved the following strengthening of the Coates-Wiles theorem:

Theorem 19.4.3 *Let E be an elliptic curve over $\mathbf{Q}$, with complex multiplication. If $r_E \geq 2$ then the order of vanishing of $L(E/\mathbf{Q}, s)$ at $s = 1$ is at least 2.*

Moreover now the formulation (4) of the Birch Swinnerton-Dyer conjecture has been proved by Rubin in the following case:

Theorem 19.4.4 *Let $E/\mathbf{Q}$ be an elliptic curve with complex multiplication. Suppose further that $\mathrm{ord}_{s=1} L(E, s) \leq 1$. Then*

$$r_E = \mathrm{ord}_{s=1} L(E/\mathbf{Q}, s).$$

Proposition 19.4.1 *The let E_p be the elliptic curve given by $y^2 = x^3 - p^2 x$, where $p \equiv 3 (\mathrm{mod} 8)$. Then the Tate-Shafarevich group of E_p is finite.*

Proof: Follows at once from Rubin's theorem (as E_p has complex multiplication by $\mathbf{Q}(i)$, provided the $L(E_p/\mathbf{Q}, 1) \neq 0$. In Chapter 16.1 it was shown that this curve has rank 0. By the conjectures of Birch and Swinnerton-Dyer, we expect that its L-function is non-zero at $s = 1$. This has been shown in this case by M. Razar [45]. □

The Birch–Swinnerton-Dyer conjecture has now been generalised by S. Bloch, A. A. Beilinson and others. Though the whole program is conjectural, there is much better understanding of the information carried by L-functions. These conjectures are subjects of intense research.

Part IV

Appendices

Appendix A

Some results from algebraic number theory

We first prove: Let K be a number field, $\mathcal{O}_K$ the ring of algebraic integers in K. Then $\mathcal{O}_K$ is a Dedekind domain.

Proof: $\mathcal{O}_K$ is certainly an integrally closed domain. If $\mathcal{P} \subset \mathcal{O}_K$ is a non-zero prime ideal, then $\mathcal{P} \cap \mathbf{Z}$ is also a non-zero prime ideal; indeed, if $x \in \mathcal{P} - \{0\}$ satisfies a monic polynomial of minimal degree

$$x^n + a_1 x^{n-1} + \cdots + a_n = 0,$$

then $a_n \in x\mathcal{O}_K \subset \mathcal{P}$ is non-zero. Hence $\mathcal{O}_K/\mathcal{P}$ is an integral extension of $\mathbf{Z}/\mathcal{P} \cap \mathbf{Z}$, which is a field, so that $\mathcal{O}_K/\mathcal{P}$ is a field *i.e.*, $\mathcal{P}$ is maximal. Hence any non-zero prime ideal of $\mathcal{O}_K$ is maximal. Finally, to see that $\mathcal{O}_K$ is Noetherian, it suffices to see that it is a finite $\mathbf{Z}$-module *i.e.*, a finitely generated abelian group.

If $x \in \mathcal{O}_K$ satisfies a monic polynomial of minimal degree n with rational coefficients, then this polynomial divides some monic polynomial with integer coefficients which is also satisfied by x; hence by the Gauss lemma, the minimal monic polynomial for x itself has integer coefficients. In particular, the trace $\mathrm{Tr}(x)$ of x is an integer.

Since K is the quotient field of $\mathcal{O}_K$, there exists a $\mathbf{Q}$-basis S of K consisting of elements of $\mathcal{O}_K$. If $S = \{x_1, \ldots, x_n\}$, where $n = [K : \mathbf{Q}]$, then the map $f : K \to \mathbf{Q}^n$ defined by

$$f(x) = (\mathrm{Tr}(xx_1), \ldots, \mathrm{Tr}(xx_n))$$

is a $\mathbf{Q}$-linear transformation. We claim f is injective — if $x^{-1} = \sum a_i x_i$ with $a_i \in \mathbf{Q}$, then

$$n = \mathrm{Tr}(1) = \mathrm{Tr}(xx^{-1}) = \sum_i a_i \,\mathrm{Tr}(xx_i),$$

so $\mathrm{Tr}(xx_i) \neq 0$ for some i. Since $f(\mathcal{O}_K) \subset \mathbf{Z}^n$, $\mathcal{O}_K$ is isomorphic to a subgroup of a finitely generated abelian group, hence is finitely generated. $\square$

Next, we prove that the unit group $\mathcal{O}_K^*$ is finitely generated. Let K be a number field. Then the group $\mathcal{O}_K^*$ of units of $\mathcal{O}_K$ is finitely generated.

Proof: From the primitive element theorem in field theory, $K = \mathbf{Q}(x)$ for some x. Let $n = [K : \mathbf{Q}]$, and let $x_1, \ldots, x_n$ be the complex roots of the minimal polynomial of x over $\mathbf{Q}$; then each x_i determines an embedding $f_i : K \hookrightarrow \mathbf{C}$. For any $y \in K$, let $|y|_i = |f_i(y)|$. For any $y \in K$, clearly the distinct $f_i(y)$ are precisely the complex roots of the minimal polynomial of y.

Consider the map

$$f : \mathcal{O}_K^* \to \mathbf{R}^n,$$
$$f(x) = (\log |y|_1, \ldots, \log |y|_n).$$

We claim that f has the following properties:

(i) f is a homomorphism of groups, where $\mathbf{R}^n$ is regarded as an additive group

(ii) the set

$$\{y \in \mathcal{O}_K^* \mid \| f(y) \| \leq C\}$$

is finite, for any $C > 0$. Here $\| \quad \|$ denotes the Euclidean norm on $\mathbf{R}^n$.

The first statement is clear, since $|xy|_i = |x|_i \, |y|_i$ for each i. To see the second statement, we note that if $\| f(y) \| \leq C$, then have $|f_i(y)| \leq e^C$ for all i. Hence the degree and the coefficients of the monic minimal polynomial of y over $\mathbf{Q}$ are bounded; there are thus only finitely many possibilities for the monic minimal polynomial of y over $\mathbf{Q}$, and hence only finitely many possibilities for y itself.

Thus, the kernel of f is finite, hence consists of the group of roots of unity in K (since any root of unity is clearly in the kernel), and the image, which

is an additive subgroup of $\mathbf{R}^n$, has finite intersection with the ball of radius C around the origin. Hence $f(\mathcal{O}_K^*) \subset \mathbf{R}^n$ has finite intersection with the ball of radius C about each of its points *i.e.*, is a discrete subgroup of $\mathbf{R}^n$ (it has no limit points).

So it suffices to prove that any discrete subgroup $\Gamma \subset \mathbf{R}^n$ is free abelian of rank $\leq n$. We do this by induction on n; the case $n = 1$ is easy, since either $\Gamma = 0$, or Γ is generated by its smallest positive element. Now suppose $n > 1$, and let $y \in \Gamma$ be an element of minimal length. Then $\mathbf{R}y$, the real line through y, satisfies $\Gamma \cap \mathbf{R}y = \mathbf{Z}y$. We may write $\mathbf{R}^n = \mathbf{R}y \oplus \mathbf{R}^{n-1}$; let $\pi : \mathbf{R}^n \to \mathbf{R}^{n-1}$ be the projection onto the second factor, so that $\ker \pi = \mathbf{R}y$. Let $\Gamma' = \pi(\Gamma)$. Then any for any $z \in \Gamma'$, we can find a unique $ay + z \in \Gamma$ with $-1/2 < a \leq 1/2$. If z is non-zero, then so is $ay + z$, so the length of $ay + z$ is at least that of y, by our choice of y; by the triangle inequality, this implies that z has length at least $1/2$ that of y. Hence Γ' is discrete; by induction, Γ' is free abelian of rank at most $n - 1$. If $z_1, z_2, \ldots, z_r$ is a free basis for Γ', and $y_i \in \pi^{-1}(z_i) \cap \Gamma$ are arbitrary, then $y, y_1, \ldots, y_r$ is easily seen to be a free basis for Γ. □

Remark: If $K = \mathbf{Q}(x)$, then the roots $x = x_1, \ldots, x_n$ of the minimal polynomial of x over $\mathbf{Q}$ can be divided into two sets: the set S_1 of real roots, and its complement, the set S_2 of non-real roots. Then S_2 consists of pairs of complex conjugate numbers. Let $\#(S_1) = r_1$, $\#(S_2) = 2r_2$; then the Dirichlet unit theorem states that

$$\mathcal{O}_K^* \cong \mu_K \times \mathbf{Z}^{r_1 + r_2 - 1}.$$

Here μ_K is the group of roots of unity in K. In fact, if x_i, x_j are complex conjugate roots, then for any $y \in K$, $f_i(y)$ and $f_j(y)$ are complex conjugate, and so $|y|_i = |y|_j$. Thus there are at most $r_1 + r_2$ distinct functions $|y|_i$; in fact, these functions are known to be all distinct. If y is a unit, then $\prod_i f_i(y) = N(y) = \pm 1$, where $N(y)$ is the norm of y. Hence the $r_1 + r_2$ functions $\log |y|_i$ are linearly dependent over $\mathbf{Q}$. The Dirichlet unit theorem is proved by showing that this is the only $\mathbf{Q}$-linear relation satisfied by these functions.

Let $K = \mathbf{Q}(\sqrt{d})$ be a quadratic extension of $\mathbf{Q}$, where $d \neq 1$ is a square-free integer. Show that $\mathcal{O}_K = \mathbf{Z} + \mathbf{Z}\alpha$, where

$$\alpha = \begin{cases} \sqrt{d} & \text{if } d \not\equiv 1 \pmod 4 \\ \dfrac{1+\sqrt{d}}{2} & \text{if } d \equiv 1 \pmod 4 \end{cases}$$

(Hint: Find the conditions on pairs of rationals (a, b) such that the minimal polynomial of $a + b\sqrt{d}$ has integer coefficients).

Let $K = \mathbf{Q}(\sqrt{-d})$, where d is a square-free positive integer (such a field K is called an *imaginary quadratic field*). Compute the unit group $\mathcal{O}_K^*$.
(Hint: Let α be as in Exercise A. If $u = m + \alpha n$ is a unit, where $m, n \in \mathbf{Z}$, then the norm $Nu = (m + \alpha n)(m + \overline{\alpha} n) = \pm 1$. Use this to solve for possible m, n.)

Appendix B

The Poisson summation formula

Let $f : \mathbf{R} \to \mathbf{C}$ be a C^∞ function such that f and each of its derivatives is rapidly decreasing at infinity *i.e.*, for any $n \geq 0$,

$$\lim_{|t| \to \infty} \left| t^n f^{(m)}(t) \right| = 0.$$

Equivalently, $\left| t^n f^{(m)}(t) \right|$ is bounded for all $m, n \geq 0$. Then we may define the *Fourier transform* of f by

$$\widehat{f}(s) = \int_{-\infty}^{\infty} f(t) e^{-2\pi \iota s t} d\, t,$$

for $s \in \mathbf{R}$. Since f is rapidly decreasing at infinity, the integral defining $\widehat{f}(s)$ converges absolutely, and uniformly in s; hence $\widehat{f}$ is also continuous. In fact, by differentiation under the integral sign, with respect to the variable s, we get

$$\widehat{f}^{(m)}(s) = \int_{-\infty}^{\infty} (-2\pi \iota t)^m f(t) e^{-2\pi \iota s t} d\, t.$$

This is valid because the integral converges absolutely and uniformly in s. Hence we see that $\widehat{f}$ is C^∞. We claim that all derivatives $\widehat{f}^{(m)}(s)$ are rapidly decreasing as $|s| \to \infty$. This follows from the formula, obtained by repeated integration by parts,

$$s^k \widehat{f}^{(m)}(s) = \int_{-\infty}^{\infty} (-2\pi \iota)^{m-k} (-1)^k \frac{d^k}{d\, t^k} (t^m f(t)) d\, t,$$

where the integral converges absolutely and uniformly in s, and defines a bounded, continuous function of s.

The *Poisson summation formula* states that for f as above,

$$\sum_{n\in\mathbf{Z}} f(n) = \sum_{n\in\mathbf{Z}} \widehat{f}(n).$$

Since $f, \widehat{f}$ are both rapidly decreasing at infinity, both series are absolutely convergent.

To prove the formula, consider the auxiliary function

$$h(t) = \sum_{n\in\mathbf{Z}} f(t+n).$$

Since f is rapidly decreasing at infinity, the series for h converges absolutely and uniformly for t in any finite interval. Hence h is well defined, and is a continuous function which is periodic with period 1. In fact, the series

$$\sum_{n\in\mathbf{Z}} f^{(m)}(t)$$

also converges absolutely and uniformly on finite intervals, for any $m > 0$, so that h is C^∞ with derivatives obtained by differentiating term by term.

Under these conditions on h, it is known that h can be expanded in a Fourier series

$$h(t) = \sum_{n\in\mathbf{Z}} a_n e^{2\pi\imath nt}$$

where a_n is given by the formula

$$a_n = \int_0^1 h(t)e^{-2\pi\imath nt}\,dt = \int_0^1 \sum_{m\in\mathbf{Z}} f(t+m)e^{-2\pi\imath nt}\,dt$$

$$= \sum_{m\in\mathbf{Z}} \int_0^1 f(t+m)e^{-2\pi\imath nt}\,dt = \sum_{m\in\mathbf{Z}} \int_m^{m+1} f(t)e^{-2\pi\imath nt}\,dt$$

$$= \int_{-\infty}^{\infty} f(t)e^{-2\pi\imath nt}\,dt = \widehat{f}(n).$$

Hence

$$\sum_{n\in\mathbf{Z}} f(n) = h(0) = \sum_{n\in\mathbf{Z}} a_n = \sum_{n\in\mathbf{Z}} \widehat{f}(n).$$

Bibliography

[1] M. K. Aggarwal, J. H. Coates, D. C. Hunt, and A. J. van der Poorten. Elliptic curves of conductor 11. *Math. Comp.*, 35:991–1002, 1980.

[2] L. Ahlfors. *Complex Analysis.* McGraw-Hill, 2nd edition, 1966.

[3] E. Arbarello, M. Cornalba, P. A. Griffiths, and J. Harris. *Geometry of Algebraic Curves*, volume 267 of *Grundlehren I.* Springer Verlag, New York, 1985.

[4] A. Baker. Contributions to the theory of diophantine equations I, On the representation of integers by binary forms. II, The diophantine equation $y^2 = x^3 + k$. *Phil. Trans. Royal Soc. London*, 263(A):173–208, 1968.

[5] A. Baker. Bounds for the solutions of the hyperelliptic equation. *Proc. Camb. Phil. Soc.*, 65:439–444, 1969.

[6] B. Birch and H. P. F. Swinnerton-Dyer. Notes on elliptic curves (I). *J. Reine. Math.*, 212:7–25, 1963.

[7] B. Birch and H. P. F. Swinnerton-Dyer. Notes on elliptic curves (II). *J. Reine. Math.*, 218:79–108, 1965.

[8] Z. Borevich and I. R. Shafarevich. *Number Theory*, volume 20 of *Pure Appl. Math.* Academic Press, 1966.

[9] J. Coates. An effective p-adic analogue of a theorem of Thue III: The diophantine equation $y^2 = x^3 + k$. *Acta Arith.*, 16:425–435, 1970.

[10] J. Coates and A. Wiles. On the conjecture of Birch and Swinnerton-Dyer. *Invent. Math.*, 39:223–251, 1977.

[11] H. T. Colebrook. *Algebra with Arithmetic and mensuration, from the Sanskrit of Brahmagupta and Bhaskara.* Dr. Martin Sändig OHG, 1973. London (1817) reprinted.

[12] G. Cornell and J. H. Silverman, editors. *Arithmetic Geometry.* Springer-Verlag, 1986.

[13] B. Datta and A. N. Singh. *History of Hindu Mathematics: A source book.* Asia Publishing House, Bombay, 1962. (2 Volumes).

[14] P. Deligne. La conjecture de Weil (I). *Publ. Math. I.H.E.S.*, 43:273–307, 1974.

[15] P. Deligne. La conjecture de Weil (II). *Publ. Math. I.H.E.S.*, 52:137–252, 1980.

[16] L. E. Dickson. *History of the Theory of Numbers.* Chelsea Publ. Co., 1966. (3 volumes).

[17] B. Dwork. On the rationality of the zeta function of an algebraic variety. *Amer. J. Math.*, 82:631–648, 1960.

[18] F. J. Dyson. The approximation to algebraic numbers by rationals. *Acta Math.*, 79:225–240, 1947.

[19] J. Edwards. *A Treatise on the Integral Calculus*, volume II. Chelsea Publ. Co., 1954.

[20] K. Elfering. *Die Mathematik des Aryabhatta, Übesetzung aus dem Sanskrit und Kommentar*, volume I. W. Fink Verlag, Munchen, 1975.

[21] N. I. Fel'dman. An effective sharpening of the exponent in Liouville's theorem. *Izv. Akad. Nauk SSSR Ser. mat.*, 35:973–990, 1971.

[22] L. Ford. *Automorphic Functions.* Chelsea Publ. Co., 1951.

[23] W. Fulton. *Algebraic Curves.* Benjamin, New York, 1969.

[24] A. O. Gel'fond. *Transcendental and algebraic numbers.* Dover, New York, (1952) russian edition, 1960.

[25] P. A. Griffiths and J. Harris. *Principles of Algebraic Geometry.* John Wiley & Sons, New York, 1978.

[26] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers.* Oxford University Press, fourth edition edition, 1975.

[27] R. Hartshorne. *Algebraic Geometry*, volume 52 of *Grad. Text Math.* Springer-Verlag, New York, 1977.

[28] H. Hasse. Beweis des Analogons der Riemannschen Vermuting für die Artinsche und F. K. Smidttschen kongruenzzetafunktionen in gewissen elliptische fällen. *Ges. d. Wiss. Nachrichten Gottingen, Math. Phys. Klasse*, 3:253–262, 1933.

[29] Sir Thomas L. Heath. *A history of Greek mathematics*, volume Vol. I and II. Oxford Univ. Press, 1960.

[30] Sir Thomas L. Heath. *Diophantus of Alexandria – A study in history of Greek Algebra.* Dover, second edition, 1964.

[31] N. Koblitz. *Introduction to Elliptic Curves and Modular Forms.* Springer-Verlag, 1984.

[32] V. A. Kolyvagin. Finiteness of $E(\mathbf{Q})$ and $Ш(E/\mathbf{Q})$ for a class of Weil curves. *Izv. Akad. Nauk SSSR Ser. Math.*, 52:522–540, 1988.

[33] S. Lang. *Elliptic Functions.* Addison-Wesley, 1973.

[34] S. Lang. *Algebra.* Addison-Wesley, 2nd edition, 1984.

[35] J. Liouville. Sur des classes trés étendues de quantitiés dont la valeur n'est ni algébrique, ni méme réductible á des irrationelles álgebriques. *C.r. hebd. Séane. Acad. Sci. Paris*, 18:883–885, 1844.

[36] K. Mahler. Zur Approximation algebraischer Zahlen, I : Ueber den grössten Primteiler binärer Formen. *Math. Ann.*, 107:691–730, 1933.

[37] B. Mazur. Modular curves and the Eisenstein ideal. *Publ. Math. I.H.E.S.*, 47:33–186, 1977.

[38] B. Mazur. Rational isogenies of prime degree. *Invent. Math.*, 44:129–162, 1978.

[39] J. F. Mestre. Construction of an elliptic curve of rank ≥ 12. *Comptes Rendus Acad. Sci. Paris*, 295:643–644, 1982.

[40] L. J. Mordell. On the rational solutions of the indeterminate equations of the 3rd and the 4th degrees. *Proc. Camb. Phil. Soc.*, 21:179–192, 1922.

[41] L. J. Mordell. *Diophantine Equations.* Academic Press, 1969.

[42] D. Mumford. *Abelian Varieties.* Oxford Univ. Press, 1974.

[43] D. Mumford. *Tata Lectures on Theta I*, volume 28 of *Prog. Math.* Birkhäuser, 1983.

[44] A. Ogg. *Modular Forms and Dirichlet Series.* Benjamin, 1969.

[45] M. Razar. The non-vanishing of $L(1)$ of certain elliptic curves with no first descents. *Amer. J. Math.*, 96:104–126, 1974.

[46] K. F. Roth. Rational approximations to algebraic numbers. *Mathematika*, 2:1–20, 1955.

[47] K. Rubin. Tate-Shafarevich groups and L-functions of elliptic curves with complex multiplication. *Invent. Math.*, 89:527–560, 1987.

[48] W. M. Schmidt. *Diophantine Approximation*, volume 785 of *Lecture Notes in Mathematics.* Springer-Verlag, 1980.

[49] E. Selmer. The diophantine equation $ax^3 + by^3 + cz^3 = 0$. *Acta. Math.*, 85:203–362, 1951.

[50] J.-P. Serre. *A Course in Arithmetic*, volume 7 of *Grad. Text Math.* Springer-Verlag, 1973.

[51] I. R. Shafarevich. *Basic Algebraic Geometry.* Springer-Verlag, 1977.

[52] I. R. Shafarevich and J. Tate. The rank of elliptic curves. *Dokl. Akad. Nauk. USSR.*, 175:770–773, 1967.

[53] Ram Swarup Sharma, editor. *Brahma Sphuta Siddhanta, with Vasana, Vijnana and Hindi commentaries*, volume I, II. Indian Inst. of Astronomical and Sanskrit Research, 1966.

[54] T. N. Shorey and R. Tijdemann. *Exponential Diophhantine equations.* Cambridge University Press, 1986.

[55] C. L. Siegel. Approximation algebraischer Zahlen. *Math. Z.*, 10:173–213, 1921.

[56] C. L. Siegel. Über einige Anwendungen Diophantinischer Approximationen. *Abh. Preuss. Akad. Phys. Math. Klasse*, pages 41–69, 1929. = Ges. Abh. Bd. I, Springer-Verlag, pp. 209-266.

[57] J. H. Silverman. *The Arithmetic of Elliptic Curves*, volume 106 of *Grad. Texts Math.* Springer-Verlag, 1986.

[58] A. Thue. Ueber annäherungswerte algebraischer zahlen. *J. reine angew. Math.*, 135:284–305, 1909.

[59] A. Weil. Sur un théorème de Mordell. *Bull. Sci. Math.*, 54:182–191, 1930.

[60] A. Weil. Number of solutions of equations over finite fields. *Bull. Amer. Math. Soc.*, 55:497–508, 1949.

[61] A. Weil. *Number Theory: An approach through history from Hammurapi to Legendre.* Birkhäuser, 1983.

[62] R. O. Wells. *Differential Analysis on Complex Manifolds*, volume 65 of *Grad. Texts Math.* Springer-Verlag, 1980.

Index

$H^0(D)$, 136
$J_\gamma(\tau)$, 277
L-function
    analytic continuation of, 294
    convergence of, 293
    of an elliptic curve, 293
    order of vanishing at $s = 1$ of, 295
$PSL(2, \mathbf{Z})$, 21, 49
$\Delta(\tau)$, 31
Δ^*, the open unit disc, 266
$\Gamma_0(n)$, 240
$\Gamma_1(n)$, 240
$\Im$, for the imaginary part of a complex number, 21, 33, 46, 47, 49, 265
$\Re$, for the real part of a complex number, 265
$\perp\!\!\!\perp$, 229
$\mathcal{H}$, for the upper half plane, 21, 49, 266, 272
ω_∞, 296
ζ-function
    of a variety, 291
    of an elliptic curve, 291
    rationality of, 292
$f_r(x)$, 271
$g_2(\tau)$, $g_3(\tau)$, 31
$h^0(D)$, 136
j-function, 32
    properties of, 49
j-invariant, 239
    of a torus, 33
k-rational, 131
p-adic
    absolute value, 205, 221
    integers, 222
    methods, 293
    metric, 221
    numbers, 222
$r_k(n)$, 264
$r_{4k}(n)$
    as Fourier coefficient, 270
    asymptotic formula for, 264

action
    of $SL(2, \mathbf{Z})$ on $\mathcal{H}$, 47
affine
    algebra, 69
    algebraic variety, 84
    morphism, 123
    space, 83
Aggarwal, M. K., 249
algebraic
    group, 42
    integer, 244
    set, 84, 93
    variety, 95
algebraic function, 83

algebraically independent, 69
annihilator, 59
Artin, E., 174
associated prime, 59
asymptotic formula, for $r_{4k}(n)$, 270

Baker, A., 249, 263
Beilinson, A. A., 298
Bernoulli numbers, 265, 276
    recurrence formula for, 276
Bezout's theorem, 113
biholomorphic map
    between Riemann surfaces, 16
birational map, 122
Birch–Swinnerton-Dyer conjecture, 290
Bloch, S., 298

canonical divisor class, 145
canonical height, 215
Chevalley-Warning theorem, 165
Chinese Remainder theorem, 56, 186
chord and tangent constructions for cubics, 183
Coates, J., 249
cocycle condition, 277, 279
cohomology, étale, 293
complete variety, 104
complex analytic map
    between Riemann surfaces, 16
complex manifold, 100
    of dimension 1, 15
complex multiplication, 161
complex torus, 24
congruence subgroup, 240, 277
congruent number problem, 193
coordinate ring, 83, 84
    homogeneous, 93
covering space, 30
    universal, 30
cubic curves, 183, 187
curve
    defined over $\mathbf{Q}$, 180
    of genus 1, 153
    plane algebraic, 180
cusp form, 275, 279, 281, 282, 285
    of weight $2k$, dimension of space of, 285

Dedekind domain, 78, 245, 300
defined over k, 130
degree
    of a morphism of curves, 126
    of a plane curve, 180
    of an algebraic number, 247
Deligne, P., 281, 293
denominator, of an algebraic number, 247
derivation, 140
descent, 179, 219
differential form, 143
Diophantine
    equation, 180
      linear, 180
    problems, 188
Diophatine
    equation
      of degree two, 181
Dirichlet unit theorem, 245, 302
Dirichlet, P. G. L., 249
discrete subgroup, of $\mathbf{C}$, 5
discriminant, 289
divisible group, 161
divisor, 135

degree of a, 135
effective, 136
linear equivalence, 136
of a function, 135
of zeros and poles, 135
doubly periodic, 4
dual curve, 117
dual isogeny, 166
duplication formula, 208
Dwork, B., 293
Dyson, F. J., 248, 253, 263

effectively computable, 263
Eisenstein
quotient, 242
series, 242, 270
elliptic curve, 187
equation of, 146
elliptic function, 4
in terms of theta functions, 268
elliptic integrals, 2
in Weierstrass form, 4
embedding of a torus, as a plane cubic, 31

factorisation
of ideals in a Dedekind domain, 245
Faltings, G., 187
Fel'dman, N. I., 263
Fermat curve, 187
Fermat's Last Theorem, 187, 188
for $n = 4$, 189
Fermat, P., 229
finite extension, 67
flex, 115
formal group, 237
Fourier coefficients, 265
of a cusp form, 280, 281
Fourier expansions
for a cusp form, 275
Fourier series
for $\theta(\tau)$, 269
for $f_r(x)$, 271
for Eisenstein series, 270, 271
Fourier transform, 267, 304
Frenicle's problem, 191
function field, 96
fundamental domain
for $SL(2, \mathbf{Z})$ acting on $\mathcal{H}$, 48

Gel'fond, A. O., 248, 253, 263
genus of a curve, 145
germ of a function, 88
global fields, 224
global solutions, 224
Grothendieck, A., 293
group action, 129
group of divisors, 135
group varieties, 152

Hasse principle, 224
Hasse, H., 174
Hecke operators, 241
height
logarithmic, 204
of a rational point in $\mathbf{P}^n$, 204, 205
of an algebraic number, 247
height function, 202
on an elliptic curve, 207
height pairing, 215, 296
Hensel's Lemma, 222
Hessian, 115

Hilbert basis theorem, 59
Hilbert's Nullstellensatz, 85
holomorphic
    1-form, 21
    co-ordinate system, 15
    coordinate, 15
    map between Riemann surfaces, 16
homeomorphism, 15
homogeneous coordinates, normalised, 204, 289
horocircle, 280
hyperplane at infinity, 94

imaginary quadratic field, 303
infinite prime, 221
inflexion point, 115
integers
    square free, 181
integral
    closure, 67
    element, 66
    extension, 66
integral of a 1-form, along a path in a Riemann surface, 22
integral points, 243
integrally closed, 67
intersection multiplicities, 109
irreducible
    components, 86, 96
    topological space, 86
isogeny, 160
isomorphism
    of tori, 30
    of varieties, 98
isotropy subgroup, 17

Jacobian variety, 241

Kolyvagin, V. A., 297

Lang's theorem, 225
Lang, S., 165
lattice, 5
Legendre, A. M., 184
Lind, C.-E., 229
line at infinity, 94
Liouville's inequality, 247, 248, 263
Liouville, J., 247
local fields, 224
local methods, 221
local ring, 64
    of a point on an affine variety, 88
    of a variety at a point, 89, 95
local solutions, 224
Local–Global principle, 224
localisation, 62

Mahler, K., 249
Max Noether's theorem, 118
maximum modulus principle, 42
metric space, completion of, 222
modular curve, 240
moduli space, 51
    for elliptic curves, 239
Mordell's conjecture, 187
Mordell-Weil group
    computation of, 219
    rank of, 226
Mordell-Weil theorem, 195
morphism
    defined over k, 133
    of algebraic curves, 207
    of varieties, 97
    proper, 107

multiplicatively closed, 61

Néron-Tate height, 215
Nakayama's lemma, 64
Newton's method, 222
Noether normalization lemma, 69
Noetherian
    module, 57
    ring, 58
normal domain, 67

orbit, 17
order, of zero or pole
    of a differential, 144
    of a function, 135

Pappus theorem, 120
Pascal's theorem, 120
period
    of a function, 5
    of an integral, 38
piecewise differentiable path, 22
plane curve, 96
point of inflexion, 27, 154
point, k-valued, 131
Poisson summation formula, 266, 267, 305
projective
    plane, 187
    space, 93
    variety, 93
properly discontinuous action, 17, 21
    of $PSL(2, \mathbf{Z})$ on $\mathcal{H}$, 50, 51

quadrics, 158
quasi-affine variety, 95
quasi-projective variety., 95
quotient
    by a properly discontinuous action, 17
    topology, 17
    by a group, 130
    map, 130
    of a Riemann surface, 17
    Riemann surface obtained from $\mathcal{H}$, 50

Ramanujan-Petersson conjecture, 281
rapidly decreasing at infinity, 267, 304
rational point, 180, 187
reduced basis
    of a lattice in $\mathbf{C}$, 46
reduced ring, 64
reduction modulo p, 222
regular function, 83, 105
regular system of parameters, 143
regulator matrix, 296
Reichardt, H., 229
Riemann hypothesis, 174
Riemann surface, 15
Rigidity lemma, 105
ring of endomorphisms, 160
Roth's inequality, 249
Roth, K. F., 248, 253, 263
Rubin, K., 297

Schmidt, W. M., 249
second countable, 15
Segre embedding, 103
Selmer group, 228
Selmer, E., 228
Serre, J.P., 165
Shorey, T. N., 249

Siegel's theorem, 243
Siegel, C. L., 243, 248, 253, 263
specialisation, 290
    modulo p, 231

Tate-Shafarevich group, 228
    order of, 296
theta function, 264, 265
Thue's equation, 243, 249
Thue, A., 248, 253, 263
Thue-Mahler equation, 249
Thue-Siegel lemma, 250
Tijdemann, R., 249
topological field, 222
torsion points, 165, 290
torsion subgroup, 296
transformation formulas
    for a cusp form, 275
    for Eisenstein series, 270
    for theta, 266

Weak Mordell-Weil theorem, 244
Weierstrass $\wp$ function, 207
Weierstrass normal form, 117, 153
    of the equation of a cubic, 41
weight, of a cusp form, 275
Weil conjectures, 292
Weil's Riemann Hypothesis, 281
Weil, A., 195

Zariski tangent space, 101
Zariski topology, 85
zero divisor, 61
zeta function
    of an elliptic curve, 291
    over finite fields, 291
    Riemann, 270, 273
    values at positive even integers, 276