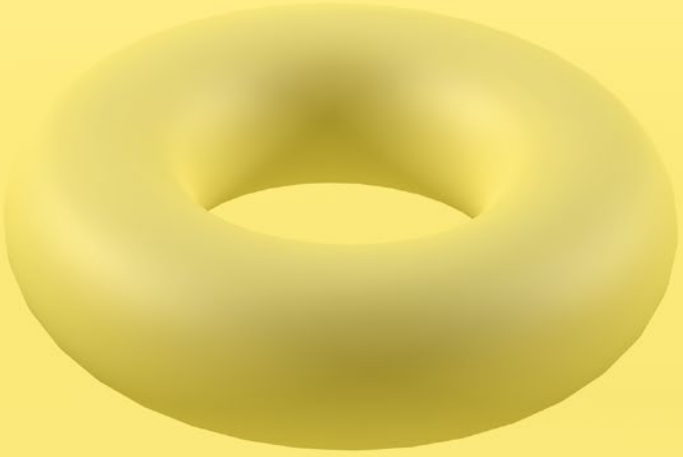
Anil Nerode · Noam Greenberg

# Algebraic Curves and Riemann Surfaces for Undergraduates

## The Theory of the Donut



Springer

# Algebraic Curves and Riemann Surfaces for Undergraduates

Anil Nerode • Noam Greenberg

# Algebraic Curves and Riemann Surfaces for Undergraduates

## The Theory of the Donut

Springer

Anil Nerode
Department of Mathematics
Cornell University
Ithaca, NY, USA

Noam Greenberg
School of Mathematics, Statistics
and Operations Research
Victoria University of Wellington
Wellington, New Zealand

This work was supported by Royal Society Te Apārangi, Marsden Fund grant, and Rutherford Discovery Fellowship

ISBN 978-3-031-11615-5 ISBN 978-3-031-11616-2 (eBook)
https://doi.org/10.1007/978-3-031-11616-2

Mathematics Subject Classification: 51-01, 14-01, 30-01, 30Fxx, 14Hxx, 14H52, 33E05



This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

The Sumerians and Babylonians (2000–700 BCE) and the Greeks (700 BCE–150 CE) made periodic measurements of the locations of the sun and the known planets. They concluded that these bodies travel in circles with the earth as their center, with small retrograde deviations. Their tables and simple astronomical instruments were used to guide travelers on camels over the silk roads of Asia and captains of ships on the high seas. These methods were codified in the great Almagest of Ptolemy (100–170 CE). This book was used for navigation for well over a thousand years. Copernicus (1473–1543) re-did the calculations based on circular orbits with the sun as center. Tycho Brahe (1542–1601) improved the observations, resulting in Kepler's three laws. They imply that the orbits around the sun are ellipses which are not circles. Mathematicians introduced elliptic analogs of the circular functions. It turned out that the elliptic analogues of the arcsine and arccosine are doubly periodic functions of a complex variable. This led to a unified theory of algebraic curves and Riemann surfaces, incorporating algebra, analysis, geometry, and topology. The purpose of this book is to make the unity of mathematics apparent to undergraduates. It is intended to serve as a textbook for a capstone course in undergraduate mathematics. This book originates in a semester-long such course taught by Prof. Nerode at Cornell University.

The book consists of three parts. The first deals with algebraic curves. It focuses on the projective plane, tangents, and intersection multiplicities and culminates in a proof of the associativity of the "chord-and-tangent" group operation on nonsingular cubic curves. It can be used on its own for a short course. The second works toward Riemann surfaces; it takes time to build the required tools from topology, calculus, and complex analysis. The third ties together the first two parts, explaining how complex curves can be given the structure of a Riemann surface. It ends with the isomorphism theorem for complex tori and elliptic curves, and with the use of analytic parameterisations of curves to redefine intersection multiplicities.

The book's style and content are a mixture of modern and older mathematics. We live in the modern mathematical world: at our disposal are set theory and logic. The arguments we give conform to modern standards of rigor. And we allow a certain level of abstraction: for example, we give axiomatic definitions of groups and rings; we define path homotopy and simple connectedness. We do, however, try to keep this kind of abstraction to a minimum, and overall, to give the reader some of the

flavor of mid-nineteenth-century mathematics, prior to the abstract turn championed by Dedekind in the 1870s. We define and study algebraic curves without introducing the notion of an ideal in a ring. We do not give an axiomatic definition of topological spaces: rather, we restrict ourselves to topological subspaces of manifolds. Most prominently, we make heavy use of Kronecker's elimination theory, which uses purely computational methods to define, for example, intersection multiplicities of curves.

We attempt to be self-contained. We review the required background material in some detail. We expect, however, that a reader will have already studied some linear algebra, groups, and some multivariable calculus, and so these topics are discussed a bit more briefly.

Greenberg would like to thank Moshe Zadka and Alex Usvyatsov, for their support during the writing of an early version of the book; and his colleagues Joe Miller, Denis Hirschfeldt, Dan Turetsky, Rod Downey, and Rob Goldblatt, for their support over the years. We would like to thank VUW students Lennox Leary, Giovanna Le Gros, Jim Paterson, Tim Caldwell, Jayden Mudge, Eli Gadsby, and Antonia King, who have worked through various iterations of the book.

Ithaca, NY, USA — Anil Nerode
Wellington, New Zealand — Noam Greenberg

# Contents

# List of Symbols

| | |
|---|---|
| $(a_0 : a_1 : \cdots : a_n)$ | point in $\mathbb{P}^n$ with homogeneous coordinates $(a_0, \ldots, a_n)$. 76 |
| $\mathbb{A}^n(\mathbb{K})$ | $n$-dimensional affine space over the field $\mathbb{K}$. 56 |
| $A =_\sim B$ | the multisets $A$ and $B$ are equivalent up to association. 30 |
| $A \subseteq_\sim B$ | $A$ is a subset of $B$ up to association. 30 |
| $a \mid b$ | ring element $a$ divides $b$. 25 |
| $\langle A \rangle_G$ | subgroup of $G$ generated by $A$. 37 |
| $\boldsymbol{\alpha}^*$ | the change of variable induced by $\boldsymbol{\alpha}$. 87 |
| $\lvert \boldsymbol{a} \rvert$ | norm of a vector in $\mathbb{R}^n$. 192 |
| $\lVert A \rVert$ | operator norm of a matrix $A$. 232 |
| $a \sim b$ | ring elements $a$ and $b$ are associates. 25 |
| $[a]_\sim$ | association class of $a$. 26 |
| $\langle \mathsf{a} \rangle$ | linear span of $\mathsf{a}$. 42 |
| $A^\intercal$ | transpose of the matrix $A$. 46 |
| $\lfloor A \rfloor$ | underlying set of the multiset $A$. 29 |
| | |
| $[b_1, b_2, \ldots, b_n]$ | multiset whose elements are $b_1, \ldots, b_n$. 29 |
| $B(\boldsymbol{a}, r)$ | open ball in $\mathbb{R}^n$. 192 |
| $B^*_\Sigma(0, r)$ | punctured neighbourhood of 0 in $\Sigma$. 402 |
| $B^*_{\Sigma/n}(0, r)$ | punctured neighbourhood of 0 in $\Sigma/n$. 403 |
| | |
| $C \cdot D$ | intersection multiset of $C$ and $D$. 147 |
| $\mathrm{char}(R)$ | characteristic of an integral domain $R$. 40 |
| $C_n$ | cyclic group of order $n$. 39 |
| $\cos z$ | complex cosine. 300 |
| | |
| $d(\boldsymbol{a}, \boldsymbol{b})$ | Euclidean distance in $\mathbb{R}^n$. 192 |
| $\deg f$ | degree of the polynomial $f$. 23 |
| $\partial g$ | general tangent operator. 117 |
| $\Delta_I \gamma$ | the vector $\gamma(t) - \gamma(s)$ where $I = [s, t]$. 257 |
| $\det(A)$ | determinant of the matrix $A$. 46 |

| | |
|---|---|
| $Df$ | formal / total derivative of $f$. 111, 231 |
| $dg$ | differential of meromorphic function. 337 |
| $D^i f$ | partial derivative of $f$ in direction $x_i$. 235 |
| $\mathrm{disc}_x(f)$ | discriminant of $f$ with respect to $x$. 113 |
| $ds$ | generalised form for path length. 256 |
| $D^*$ | the collection of nonsingular points on the curve $D$. 352 |
| $D^\sharp$ | the projective closure of $D$. 84 |
| $D^x f$ | formal / partial derivative of $f$ with respect to $x$. 110, 235 |
| $dx_i$ | component linear form. 262 |
| $dz$ | complex linear form. 285 |
| | |
| $E_\Gamma$ | elliptic curve isomorphic to $T_\Gamma$. 382 |
| $\ell_\infty$ | line at infinity. 85 |
| $\exp z$ | complex exponential. 299 |
| $e^z$ | complex exponential. 300 |
| | |
| $\mathfrak{f}$ | analytic function defined by the formal power series $f$. 398 |
| $\dot{f}$ | derivative of a function of a single variable. 232 |
| $f^\flat$ | dehomogenisation of $f$ with respect to $x_0$. 83 |
| $f^{\flat x}$ | dehomogenisation of $f$ with respect to $x$. 82 |
| $F \cdot dr$ | work form of a vector field. 266 |
| $f_P$ | implicit definition of the place $P$. 416 |
| $f^\sharp$ | homogenisation of $f$ with respect to $x_0$. 83 |
| $f^{\sharp x}$ | homogenisation of $f$ with respect to $x$. 82 |
| $f_{\boldsymbol{u},\boldsymbol{v}}$ | general intersection polynomial. 123, 144 |
| $F_{\mathtt{wind}}$ | winding number vector field. 269 |
| $f^*\omega$ | pull-back of $\omega$ by $f$. 332, 335 |
| $F(\boldsymbol{x})$ | field of formal rational functions with coefficients in $F$. 34 |
| | |
| $\gamma_2$ | coefficient associated with the lattice $\Gamma$ (similarly, $\gamma_3$). 381 |
| $\mathbb{G}_d$ | the space of curves of degree $d$. 92 |
| $G/H$ | The quotient of the group $G$ by $H$. 39 |
| $\mathrm{GL}_n(R)$ | general linear group over $R$. 42 |
| | |
| $\mathcal{H}_C$ | Hessian curve of $C$. 169 |
| $H_\infty$ | hyperplane at infinity. 81 |
| | |
| $i(D, Q)$ | intersection multiplicity of a curve and a place. 421 |
| $\lvert I \rvert$ | length of the interval $I$. 257 |
| $\int_a^b f\,dt$ | real / complex Riemann integral. 262, 285 |

| | |
|---|---|
| $\int_\gamma f\,dz$ | complex path integral. 285 |
| $\int_\gamma \omega$ | integral of $\omega$ along $\gamma$. 257, 338 |
| $\iota$ | bijection between $\mathbb{P}^2$ and $\check{\mathbb{P}}^2$. 93 |
| $\iota_d$ | isomorphism between $\mathbb{P}^k$ and $\mathbb{G}_d$. 92 |
| $i_p(C, D)$ | intersection multiplicity at $p$ of the curves $C$ and $D$. 146 |
| $i_p(C, \ell)$ | intersection multiplicity at $p$ of $C$ with the line $\ell$. 121 |
| $i_p(f, g)$ | the intersection multiplicity $i_p\left(V_{\mathbb{P}^2}(f), V_{\mathbb{P}^2}(g)\right)$. 156 |
| $i(P, Q)$ | intersection multiplicity of places. 419 |
| | |
| $\ell(\gamma)$ | length of the path $\gamma$. 259 |
| $\limsup_n r_n$ | limit superior. 296 |
| $\ell^k C$ | general $k$th-order tangent. 117 |
| $\ell_p^k C$ | $k$th -order tangent to $C$ at $p$. 114 |
| Log | principal branch of the complex logarithm. 307 |
| $L_p$ | set of pairs $(q, r)$ such that $\{p, q, r\}$ are collinear. 143 |
| $\ell(P)$ | tangent of a place. 425 |
| $\ell_p C$ | tangent to $C$ at $p$. 115 |
| | |
| $m_a(A)$ | multiplicity of $a$ in the multiset $A$. 29 |
| $M_c$ | matrix for complex multiplication. 282 |
| $M^{d,e}(f, g)$ | the $d, e$-Sylvester matrix of $f$ and $g$. 59 |
| $M_n(R)$ | ring of $n \times n$ matrices with entries from $R$. 41 |
| | |
| $o$ | the origin. 58 |
| $o(P)$ | order of a place. 424 |
| $o_p(C)$ | order of $p$ on $C$. 115 |
| $\mathrm{ord}_a(f)$ | order of meromorphic form or function. 314 |
| $\mathrm{ord}(f)$ | order of formal power / Laurent / fractional series $f$. 28, 34, 401 |
| | |
| $\pi_G$ | quotient map from $\mathbb{R}^n$ to $\mathbb{R}^n/G$. 214 |
| $\pi_n$ | quotient map from $\mathbb{K}^{n+1} \setminus \{\mathbf{0}\}$ to $\mathbb{P}^n(\mathbb{K})$. 76 |
| $p_\infty$ | point at infinity. 315 |
| $\pi_\Sigma$ | projection from Riemann surface for the logarithm. 326 |
| $\pi_{\Sigma/n}$ | projection from Riemann surface for the $n$th root. 329 |
| $\check{\mathbb{P}}^2$ | dual projective plane. 93 |
| $\mathbb{P}^n(\mathbb{K})$ | $n$-dimensional projective space over the field $\mathbb{K}$. 76 |
| $\overline{p\,q}$ | the line passing through $p$ and $q$. 79 |
| $p +_C q$ | addition of points on the curve $C$. 172 |
| $p * q$ | third point of intersection of $\overline{p\,q}$ and $C$. 171 |

| | |
|---|---|
| $\wp$ | Weierstrass's elliptic function. 376 |
| $\mathrm{pwr}_n$ | inverse of $\mathrm{rt}_n$. 330 |
| | |
| $\mathfrak{R}(D)$ | collection of ramification points of the curve $D$. 357 |
| $\mathrm{res}^{d,e}(f,g)$ | the $d,e$-resultant of $f$ and $g$. 60 |
| $\mathrm{res}(f,g)$ | the resultant of $f$ and $g$. 60 |
| $\mathrm{res}_{x,y}(f,g)$ | homogeneous resultant of $f$ and $g$. 140 |
| $R(f)$ | radius of convergence of the formal power series $f$. 398 |
| $R_{f,g}$ | homogeneous resultant of general intersection polynomials. 144 |
| $\mathrm{rsd}_a(f)$ | residue of $f$ at $a$. 317 |
| $R^*$ | the group of units of $R$. 26, 36 |
| $\mathrm{rt}_n$ | $n$th root on $\Sigma$ or $\Sigma/n$. 328, 329 |
| $R[\boldsymbol{x}]$ | ring of polynomials with coefficients from $R$. 23 |
| $R[\![x]\!]$ | ring of formal power series with coefficients from $R$. 18 |
| $R(\!(x)\!)$ | ring of formal Laurent series with coefficients from $R$. 34 |
| $R[\![x^{1/n}]\!]$ | ring of fractional power series with exponent $1/n$. 401 |
| $R[\![\![x]\!]\!]$ | ring of fractional power series. 401 |
| | |
| $S$ | the unit circle. 1 |
| $\mathrm{sgn}(\sigma)$ | sign of the permutation $\sigma$. 40 |
| sh | Shift map on $\Sigma$, $\Sigma/n$ or $\mathbb{C}[\![x]\!]$. 330, 404 |
| $\Sigma$ | Riemann surface for the logarithm. 326 |
| $\Sigma/n$ | Riemann surface for the $n$th root. 328 |
| $\sin z$ | complex sine. 300 |
| $S_n$ | symmetric group on $n$ elements. 36, 40 |
| $S_P(\omega,\gamma)$ | partial sum. 257 |
| | |
| $T_\Gamma$ | the torus $\mathbb{C}/\Gamma$. 216 |
| | |
| $V_{\mathbb{A}^n(\mathbb{K})}(f)$ | affine hypersurface defined by the polynomial $f$. 56 |
| $V_{\mathbb{P}^n(\mathbb{K})}(f)$ | projective hypersurface defined by the polynomial $f$. 77 |
| | |
| $\mathrm{wnd}_\gamma(p)$ | winding number of $\gamma$ around $p$. 289 |


# Introduction

# 1

An *algebraic curve* is the collection of points satisfying a polynomial equation in two variables, for example, the circle $x^2 + y^2 = 1$, or the parabola $y = x^2$. A *Riemann surface* is an object that locally looks like the complex plane, and on which one can perform complex differentiation. Examples are the Riemann sphere, or a complex torus. In this book we will present the theory of both algebraic curves and Riemann surfaces, and then show that these two concepts are in fact connected. Our presentation culminates in the theory of elliptic functions, and the isomorphism theorem for elliptic curves which, roughly, says that complex tori are the same as nonsingular cubic curves.

## 1.1 The Theory of the Circle

Elliptic functions and curves are in some sense a cubic analogue of the trigonometric (circular) functions and conic curves. As an illustration of the development presented in this book, we consider the conic case first, by examining the circle.

### 1.1.1 Pythagorean Triples

One motivation for what follows is the problem of finding all *Pythagorean triples*: integers $a, b, c$ satisfying $a^2+b^2 = c^2$. In other words, finding the triples of integers that form the lengths of right triangles. We rephrase the problem by replacing the triple $(a, b, c)$ by the pair $(a/c, b/c)$; the task is then to find all points on the unit circle that have rational coordinates. In this book we denote the unit circle by $S$.

Here is an approach to solving this problem. Fixing the rational point $(0, 1)$ on the unit circle, for any point $(t, 0)$ on the $x$-axis we draw the line $\ell_t$ that passes through the points $(0, 1)$ and $(t, 0)$. The line $\ell_t$ intersects $S$ at another point, which



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_1

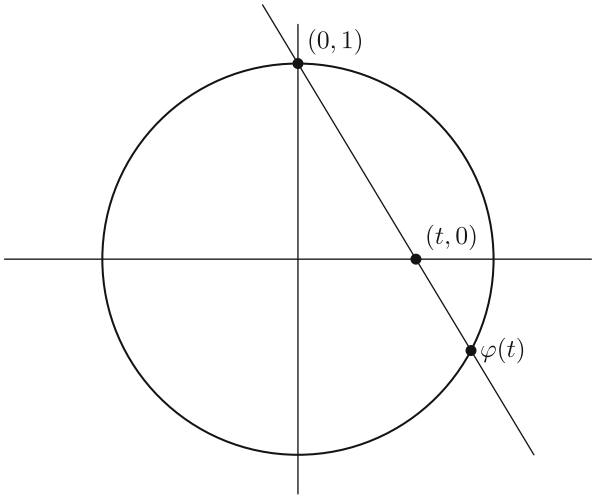



**Fig. 1.1** The rational parameterisation of the unit circle

we name $\varphi(t)$, whose coordinates can be computed to be

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

See Fig. 1.1.

The function $\varphi$, from the real line $\mathbb{R}$ to the unit circle $S$, is injective, and its range is $S \setminus \{(0, 1)\}$. Moreover, the two coordinate functions of which $\varphi$ is comprised are *rational functions* of $t$, that is, quotients of polynomials. So if $t$ is a rational number, then $\varphi(t)$ is a rational point on $S$. Further examination allows us to use $\varphi$ to obtain a full solution of the problem of finding Pythagorean triples; see Exercise 4.82.

We note that our technique has a flaw: the original point $(0, 1)$ is not obtained as $\varphi(t)$ for any $t$. The family of lines $\ell_t$ which we used to parameterise the circle is missing one line passing through $(0, 1)$, namely the line $y = 1$, the *tangent* to the circle at the point $(0, 1)$. The fact that the line $y = 1$ does not intersect the $x$-axis indicates that there is a "missing point" on the $x$-axis: we would like to let $(0, 1) = \varphi(\infty)$. This missing point lies on the *projective line*. Our study of algebraic curves will make extensive use of tangents, and we shall see that the *projective plane* is the correct arena for algebraic curves.

### 1.1.2 The Circular Functions

The number $2\pi$ is the circumference of the unit circle. We can measure $\pi$, as Archimedes did, by approximating the circle by regular $n$-gons inscribed in the circle. More generally, if $\theta$ is the length of the arc on the unit circle from $(1, 0)$ to a point $(x, y)$ on the circle, then the coordinates of the point satisfy $x = \cos\theta$ and

$y = \sin\theta$. This is usually taken as the definition of the circular functions sin and cos (often not mentioning that the size of an angle is defined to be that arc-length). The definition is rarely presented in this way, but this approach in fact first defines the functions arcsin and arccos, and then defines sin and cos as the inverses of these. That is, given $x$ between $-1$ and 1, we define $\arccos x$ to be the length of the arc from $(x, y)$ to $(1, 0)$, where $y \geqslant 0$ is chosen so that $(x, y)$ lies on the unit circle (that is, $y = \sqrt{1-x^2}$). Using calculus, we can express this value in terms of an arc-length integral:

$$\arccos x = \int_x^1 \frac{dt}{\sqrt{1-t^2}}.$$

Historically, this was the way that *elliptic functions* were discovered. These are "higher degree" analogues of the circular functions sin and cos. They were first defined as the inverses of *elliptic integrals* such as $\int dt/\sqrt{1-t^4}$, some of which arise as arc-lengths of curves other than the circle, in particular, of ellipses, whence the name. Elliptic functions have *addition formulas*, similar to the familiar formulas for $\sin(\theta+\rho)$ and $\cos(\theta+\rho)$. They are also *periodic*, similar to the fact that $\cos(\theta + 2\pi) = \cos(\theta)$, and the same for sin. One of Abel and Jacobi's main contributions was the revelation that the elliptic functions should be extended to the complex plane, where they are *doubly periodic*, meaning there is a two-dimensional lattice of periods (see Fig. 8.2).

Our historical survey (Chap. 16) provides further details of this development, including the method of deriving the addition formulas for sin and cos using the original definition of cos as the inverse of $\int dt/\sqrt{1-t^2}$. The more modern approach to elliptic functions was developed by Weierstrass and Riemann. They define elliptic functions directly as *analytic* functions, the sums of converging power series. This is the approach we present in this book. We now briefly present an analogous development of the circular functions and the number $\pi$. We will return to this argument later in the book once we have developed the necessary machinery; see p. 299.

Define, for all $t \in \mathbb{R}$,

$$e^{it} = 1 + it + \frac{(it)^2}{2!} + \frac{(it)^3}{3!} + \frac{(it)^4}{4!} + \cdots$$

(where $i$ is the imaginary root of $-1$). By the Weierstrass $M$-test, this series converges for all $t$ and gives a function $f : \mathbb{R} \to \mathbb{C}$ satisfying $f' = if$, where by the derivative of $g + ih$ we mean $g' + ih'$. This is done by differentiating term by term and comparing power series. Indeed, this map is the unique solution for $f' = if$ and $f(0) = 1$.

Now fixing $s \in \mathbb{R}$ and differentiating $t \mapsto e^{i(s+t)}$, we see that this function satisfies the same differential equation, and so equals $e^{is} \cdot e^{it}$. That is, we obtain the addition formula $e^{i(t+s)} = e^{it} \cdot e^{is}$, justifying the exponential notation.

*Define*, for $t \in \mathbb{R}$, $\cos t$ to be the real part of $e^{it}$ and $\sin t$ to be the imaginary part of $e^{it}$. That is,

$$e^{it} = \cos t + i \sin t.$$

Separating the real and imaginary parts gives power series representations for sin and cos, yielding $\sin' = \cos$, $\cos' = -\sin$, $\sin(-x) = -\sin x$ and $\cos(-x) = \cos x$. That is, $e^{-it} = \overline{e^{it}}$ is the complex conjugate of $e^{it}$. The addition theorem shows that for all $t$,

$$|e^{it}|^2 = e^{it} \cdot \overline{e^{it}} = e^{it} e^{-it} = e^{i(t-t)} = 1,$$

that is, the range of $t \mapsto e^{it}$ is contained in the unit circle. This gives the trigonometric formula $\cos^2 t + \sin^2 t = 1$. The addition formula for $e^{it}$ also gives the addition formulas for sin and cos.

The power series representation of the cosine function starts with $\cos t = 1 - t^2/2! + t^4/4! - \cdots$. When we plug in $t = 2$ we get $\cos 2 = 1 - 2 + 2/3 - \cdots$, where the rest is an alternating series. It follows that $\cos 2 < 0$. Since $\cos 0 = 1$, by the intermediate value theorem, there is a number $\rho \in (0, 2)$ such that $\cos(\rho) = 0$. A similar use of the Taylor series shows that $\sin \rho > 0$, and so $\sin \rho = 1$, that is, $e^{i\rho} = i$. Hence $e^{4i\rho} = i^4 = 1$. The addition formula shows that $4\rho$ is a *period* of $e^{it}$, and so of both sin and cos: for all $t$, $e^{i(t+4\rho)} = e^{it}$. We *define* $\pi$ to be $2\rho$. Euler's formula $e^{i\pi} = i^2 = -1$ follows from this definition.

The fact that $|(e^{it})'| = |e^{it}| = 1$ for all $t$ shows that $t \mapsto (\cos t, \sin t)$ is an *arc-length parameterisation* of the circle, and this proves that indeed, $2\pi$ is the circumference of the unit circle. So the two ways of defining $\pi$ are equivalent.

### The Isomorphism Theorem for the Circle

As with the rational parameterisation of the unit circle discussed above, the development we just presented has several ingredients (such as analytic functions) that are used in this book. Overall, we obtain an isomorphism theorem for the circle. The map $t \mapsto e^{it}$ induces a bijection $t + 2\pi\mathbb{Z} \mapsto e^{it}$ between the quotient $\mathbb{R}/2\pi\mathbb{Z}$ and the unit circle $S$ which is an isomorphism of manifolds and abelian groups.

In the relevant chapters of the book, we shall, of course, provide precise definitions of these notions. Informally, the map $t + 2\pi\mathbb{Z} \mapsto e^{it}$ preserves addition and proximity. We can add points in the quotient group $\mathbb{R}/2\pi\mathbb{Z}$; this is induced by usual addition of real numbers. We can also "add" points on the unit circle, by regarding them as complex numbers and using complex multiplication (rather than addition). The addition formula above shows that $t + 2\pi\mathbb{Z} \mapsto e^{it}$ translates addition in $\mathbb{R}/2\pi\mathbb{Z}$ to complex multiplication in $S$, that is, it is a group isomorphism.

The real line is not only a group, but also has topological and differential structure. We can measure the distance between points, and have a notion of closeness based on open intervals; this gives rise to the notions of continuous and differentiable functions. The unit circle *locally* resembles the real line. Informally,

we say that when magnified, small portions of the circle look like a line (the curvature becomes small). Slightly less informally, we can assign *local coordinates* to points on the circle. One way to do this is to project onto lines (often, tangent lines). For example, near the point $(0, 1)$, the unit circle is close to the tangent $y = 1$. The projection $(x, y) \mapsto (x, 1)$ from the circle near $(0, 1)$ to this line indicates that we can assign the coordinate $x$ to points $(x, y)$ on the unit circle close to $(0, 1)$. In fact, we can do this for all such points satisfying $y > 0$, and we obtain a bijection between part of the unit circle and the open interval $(-1, 1)$. Based on this, we can now define continuous and differentiable functions on that part of the unit circle, by imagining that it *is* the interval $(-1, 1)$.

By using other lines, we can cover the entire unit circle by "patches" that each look like an open interval. For example, near the point $(1, 0)$ we use the line $x = 1$ and assign the coordinate $y$ to a point $(x, y)$. The same point can be an element of more than one patch, and so be assigned two different coordinates: for example, if both $x$ and $y$ are positive, the point $(x, y)$ can be assigned either $x$ or $y$ as a coordinate, depending on whether we use the projection to the line $y = 1$ or the line $x = 1$. This could lead to ambiguity as to what we mean by a continuous or differentiable function on the unit circle. What we ensure is that the *translation* between coordinates is itself differentiable. For example, the translation from $x$ to $y$ on the unit circle is the map $x \mapsto \sqrt{1 - x^2}$, which is indeed differentiable. So in fact, it makes no difference which coordinate we choose in order to define differentiability.

A similar process of assigning coordinates locally can also be performed on the quotient $\mathbb{R}/2\pi\mathbb{Z}$. If $I = (a, b)$ is a small open interval (say of length $|I| = b - a < 2\pi$) then the restriction to $I$ of the quotient function $x \mapsto x + 2\pi\mathbb{Z}$ is injective, and so can be used to assign coordinates to its image in $\mathbb{R}/2\pi\mathbb{Z}$. The translations between different coordinate patches are now maps of the form $x \mapsto x + 2\pi k$ for some integer $k$, which are differentiable. Thus, like the unit circle, the quotient $\mathbb{R}/2\pi\mathbb{Z}$ acquires the structure of a *differentiable 1-manifold*. Having done this, we can verify that the function $t + 2\pi\mathbb{Z} \mapsto e^{it}$ is not only a group isomorphism, but is also continuous and in fact differentiable in both directions, that is, it gives an isomorphism of differentiable manifolds.

### 1.1.3 The Theory of the Donut, in a Nutshell

The situation for the donut is similar. Instead of the circle, we consider an algebraic curve defined by a polynomial of degree 3; this is called a *cubic* curve. We also assume that the curve is *nonsingular*, which roughly, means that it is smooth. As is true of the unit circle, we can then show that there is a way to add points on the curve (in what is known as the *chord-and-tangent* method). A little like the rational parameterisation of the unit circle, this method uses the fact that typically, a line will intersect a cubic curve at three points; this gives us a method of taking two points on the curve and obtaining a third. For this method to work, we need to extend the curve to the projective plane (we add "points at infinity").

As with the unit circle, by using projections onto lines, we can locally assign coordinates to points on the curve, and so consider the curve as a differentiable manifold. The new ingredient however is the use of complex numbers rather than real numbers. We allow complex numbers to be the coordinates of points on the curve (we look at pairs of complex numbers satisfying the polynomial equation). This means that "lines" look like $\mathbb{C}$ rather than $\mathbb{R}$, and the coordinates that we give points are complex numbers rather than real numbers. So topologically, we get a two-dimensional surface. Indeed we get a *Riemann surface*, one on which we can perform complex differentiation.

We can understand the topological isomorphism of the quotient $\mathbb{R}/2\pi\mathbb{Z}$ with the circle as taking the closed interval $[0, 2\pi]$ and "gluing" its two end-points; the quotient process identifies all other closed intervals $[2\pi k, 2\pi(k+1)]$ with $[0, 2\pi]$. Intuitively, we obtain a circle by taking a piece of string and connecting its two end-points. A donut shape (which in mathematics is called a *torus*) is obtained by taking a square and gluing opposite sides to each other. This can be more easily envisioned by first identifying two sides to obtain a cylinder, and then the resulting circles to obtain the donut shape; see Fig. 1.2. This process of gluing also allows us to view the torus as the topological product of two circles; see Exercise 8.111.

To obtain a group operation on the torus, we view it as the quotient of the complex plane by a *two-dimensional lattice* $\Gamma$, for example the collection of all points $n + im$ where $n, m \in \mathbb{Z}$ are integers; for a more general picture see Fig. 8.2. Then $\mathbb{C}$ can be viewed as tiled by parallelograms, and in the quotient $\mathbb{C}/\Gamma$, these parallelograms are first identified with each other, and then opposite sides are glued—hence,

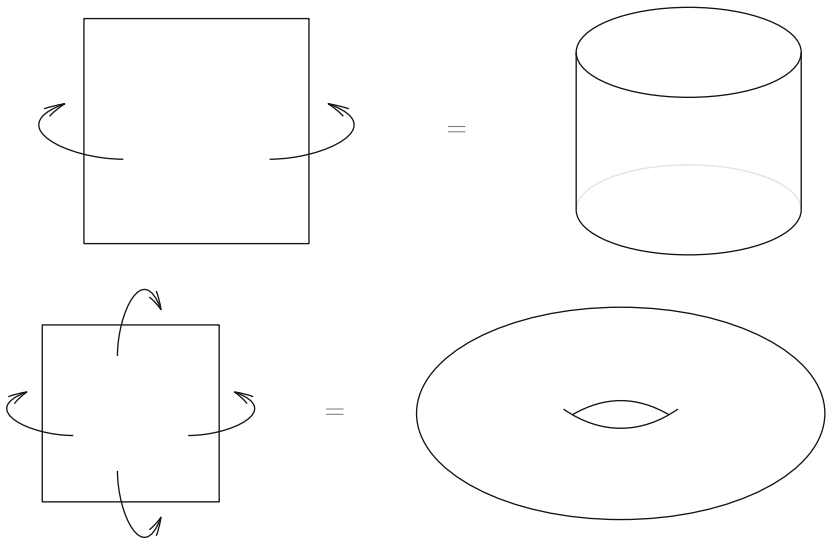


**Fig. 1.2** A cylinder is obtained by gluing two opposite sides of a square; the torus is then obtained by gluing the resulting opposite circles

topologically, $\mathbb{C}/\Gamma$ is a torus. The lattice $\Gamma$ is a subgroup of the additive group of the complex numbers, and so the quotient $\mathbb{C}/\Gamma$ is a group. Also, we can use restrictions of the quotient map $z \mapsto z + \Gamma$ to small open balls, to assign local coordinates to points on the torus. The translations between coordinate patches are shifts by elements of the lattice $\Gamma$, and so the resulting structure is a Riemann surface.

Fixing a two-dimensional lattice $\Gamma$ in the complex plane, Weierstrass used power series (like our definition of the map $t \mapsto e^{it}$) to define a complex differentiable function $\wp$ which has $\Gamma$ as its set of periods (just as $2\pi\mathbb{Z}$ is the set of periods of $t \mapsto e^{it}$). Such a function is called *elliptic*. It induces a well-defined function on the torus $\mathbb{C}/\Gamma$, which is (complex) differentiable in the sense of Riemann surfaces. The function $\wp$ satisfies a certain differential equation which means that we can use it and its derivative to parameterise a complex, nonsingular cubic curve, often denoted by $E_\Gamma$. Because the curve is parameterised by an elliptic function, it is called an *elliptic curve*.

Analogously to the circle, Poincaré and Weil observed that this parameterisation induces an isomorphism between the torus $\mathbb{C}/\Gamma$ and the curve $E_\Gamma$, which is both a group isomorphism and an isomorphism of Riemann surfaces. This is the isomorphism theorem. While the modern formulation was given by Weil, its origins go back to Abel's work on elliptic functions and their addition formulas.

## 1.2 Overview of the Book

### 1.2.1 Part I: Algebraic Curves

The first part of the book lays out the theory of algebraic curves over algebraically closed fields, studies their tangents and intersections, and culminates in the group structure of a nonsingular cubic curve.

#### Affine and Projective Curves

An algebraic curve is the collection of points $(x, y)$ satisfying a polynomial equation $f(x, y) = 0$. For example, the polynomial $f(x, y) = x^2 + y^2 - 1$ defines the unit circle.

Our first task will be to connect divisibility of polynomials and containment between the curves that they define. If $f$ and $g$ are polynomials and $f$ divides $g$ (there is some $h$ such that $g = fh$) then for all $a$ and $b$, if $f(a, b) = 0$ then $g(a, b) = 0$, so the curve $f = 0$ is contained in the curve $g = 0$. The converse of this fact, called Study's Lemma, requires certain assumptions and a tweaking of the definitions. We must, for example, work over an algebraically closed field (such as the complex numbers). Otherwise, we have many polynomials defining the empty curve. We also need to consider curves with repeated components, which leads us to the concept of a *multiset*: a set in which some elements can occur more than once. To properly define algebraic curves as multisets, we break polynomials up into irreducible factors (factors that cannot be further presented as nontrivial products of other polynomials). If $f = f_1 f_2 \cdots f_k$ is an *irreducible factorisation* of

a polynomial $f$, then we will define the curve $f = 0$ to be the *multiset sum* of the curves $f_1 = 0$, $f_2 = 0, \ldots, f_k = 0$. If a point belongs to more than one component $f_i = 0$, then it belongs to the curve $f = 0$ more than once. *Unique factorisation* in polynomial rings (Theorem 2.24) shows that this definition is unambiguous. Multisets, unique factorisation, and other algebraic tools are discussed in Chap. 2; in Chap. 3 we define algebraic curves and prove Study's lemma. A key tool from *elimination theory*, called the *resultant*, is introduced in that chapter as well, and is used in the proof of Study's lemma.

*Projective geometry* is in some ways simpler than usual Euclidean (or *affine*) geometry: for example, in the projective plane, any two distinct lines intersect at exactly one point. To achieve that, extra points "at infinity" are added to the affine plane, where previously parallel lines can intersect. For example, all vertical lines $x = a$ intersect at the "vertical point at infinity", and in general, each new point at infinity corresponds to a direction of lines in the affine plane. In turn, we will see that other curves also have some "missing points" at infinity which should be added to the curve. For example, to get the *projective closure* of the affine parabola $y = x^2$ we add the vertical point at infinity; to the hyperbola $xy = 1$ we add both the vertical and the horizontal points at infinity. We discuss projective geometry in Chap. 4.

### Intersections of Curves

The *degree* of an algebraic curve is the degree of the polynomial which defines the curve. Bézout's Theorem says that two projective curves of degrees $n$ and $m$ respectively, with no common component, intersect in $nm$ many points. This generalises the fact that any two distinct lines in the projective plane intersect at a unique point, as lines are the curves of degree 1. One of the consequences of Bézout's theorem is that a line intersects a cubic curve (a curve of degree 3) at 3 points. This allows us to define the chord-and-tangent group operation on an elliptic curve.

A few caveats should be noted. We need to work in the projective plane: the case of parallel lines is the simplest explanation why. We also need to work over an algebraically closed field; this was already necessary for Study's lemma. The third point to note is that we need to count the intersections correctly. Some points of intersection may coincide, so we need to count them more than once. An example is the curves $y = x^2$ and $y = -x^2$, intersecting at the origin. If we perturb the curves only a little, we see that we get two points of intersection near the origin, and so we say that the curves intersect twice at the origin.

A special case is the intersection of curves with lines. Here we are drawn to consider the *tangents* to a curve: the tangent to a curve at a point will intersect the curve more than once at that point. We need to pay attention to *singular points* on a curve, where a unique tangent is undefined, since all first-order partial derivatives vanish at that point. It turns out that we can identify multiple tangents at singular points (for example, the origin is singular on the "nodal" cubic curve (Fig. 3.3) and the tangents there are $y = x$ and $y = -x$). The number of such tangents is the *order* of a point on a curve. Theorem 5.34 will give us a connection between orders

of points and intersection multiplicities with lines: (a) The order of a point $p$ on a curve $C$ is the least intersection multiplicity with $C$ at $p$ of any line passing through $p$; (b) A line passing through $p$ is a tangent to $C$ at $p$ if and only if the multiplicity of intersection of that line with $C$ at $p$ is strictly greater than the order of $p$ on $C$.

Although this idea is intuitive, it is not easy to properly define intersection multiplicities. In Chap. 5 we consider only the intersection of lines with curves, using parameterisations of lines. Such parameterisations are harder to come by in the general case, so in Chap. 6 we use elimination theory instead. In Chap. 15, armed with analytic tools (in particular, Riemann surfaces for multi-valued functions), we use parameterisations for the general case.

#### Elliptic Curves

An *elliptic curve* is a projective cubic curve that has no singular points. (As mentioned above, this terminology is justified by several steps: *elliptic integrals* measure the arc-length of curves, such as ellipses; *elliptic functions* are the inverses of elliptic integrals; *elliptic curves* are parameterised by elliptic functions. Ellipses are not elliptic curves, as they are conic, not cubic). In Chap. 7 we show that the *chord-and-tangent* rule makes an elliptic curve an abelian group. By Bézout's theorem, every line intersects an elliptic curve $C$ at three points (multiplicities counted). The group operation on the curve is characterised by the requirement that three collinear points on the curve add up to 0. The main challenge we will face is proving associativity of this operation (Theorem 7.19); the technical difficulties arise from tangent lines and repeated intersections.

We will also consider *normal forms* for elliptic curves, for which the identity element is usually the vertical point at infinity; these normal forms will crop up again in the third part of the book, when we consider the isomorphism theorem for complex tori.

### 1.2.2 Part II: Riemann Surfaces

Riemann surfaces are connected surfaces on which we can locally assign coordinates in a way that allows us to perform complex analysis. We define and study them in Chap. 12. In Chaps. 8–11 we build up the required tools, encompassing topology, and real and complex analysis.

#### Three Kinds of Surfaces

A *surface* is a mathematical object that *locally* looks like an open disc in the two-dimensional plane. As discussed above, this means that we can cover it by "patches", each of which is equipped with a bijection between that patch and an open subset of $\mathbb{R}^2$. These bijections are called *charts*. Each chart assigns *local coordinates* to the points in the patch (its domain). Then we can use these coordinates to define what it means for a function on the surface to be continuous, or differentiable.

The catch is that the same point may be covered by different patches, and so may be assigned different coordinates by different charts. So it is conceivable that a function may be differentiable according to one system of coordinates, but not according to the other. To avoid this problem, we need to examine the *transition function* which tells us how to translate from one system of coordinates to the other. Two charts are *compatible* if the transition function is nice, and "nice" depends on the context.

First, in Chap. 8, we consider only topology: we only require the transition functions to be continuous. This allows us to define what it means for a function on the surface to be continuous, but more complicated notions are not yet available: it is possible, for example, that a function on the surface is differentiable according to one set of coordinates but not another. In Chap. 9 we make the extra requirement that transition functions be continuously differentiable. This precludes the problem just described, and so enables us to discuss differentiability of functions on our surfaces. Finally, in Chap. 12, we require the transition functions to be *analytic*, and this allows us to perform complex analysis on our surfaces. We call these *holomorphic surfaces*.

*Riemann surfaces* are holomorphic surfaces which are *connected*: roughly, we cannot neatly break them up into pieces without "tearing". This notion is studied in Chap. 9, along with the important class of *simply* connected spaces: spaces without "holes", such as the plane and the sphere, but not the punctured plane or the torus.

### Analytic Functions

A function is analytic if it is the sum of a power series: $f(x) = \sum a_n x^n$. This is the sum of infinitely many functions, but in some sense, analytic functions are close to polynomials. For example: a polynomial has only finitely many roots. An analytic function may have infinitely many zeros, but they must be *isolated* from each other, so on a bounded closed disc, for example, an analytic function can have only finitely many zeros. Analytic functions are rigid, in the sense that their local behaviour—their values on some small region—determine the entire function. Other functions, even if they are infinitely differentiable, do not behave so rigidly.

Derivatives of complex-valued functions of a complex variable are defined in the same way as the familiar derivative, except that in the ratio $(f(z+h) - f(z))/h$ we use the arithmetic of complex numbers, and the limit as $h \to 0$ is taken over all complex numbers $h$ close to 0. The amazing fact about complex differentiation is that every complex differentiable function is analytic. Thus, for example, the derivative of a complex differentiable function is itself differentiable! This yields a beautiful theory, with results such as a the open mapping theorem, stating that analytic functions map open sets to open sets.

### Real Analysis, Complex Analysis, and Path Integrals

The principal tool used to show that every (continuously) complex-differentiable function is analytic (Theorem 11.67) is Cauchy's Integral Formula, which states that the value of a complex differentiable function at a point $p$ is determined by its values along a path that goes around $p$, via an integration process of the function

along that path. Thus, complex analysis relies heavily on the concept of the *path integral* (also called a *line* integral).

The complex plane $\mathbb{C}$ is naturally identified with the real plane $\mathbb{R}^2$, and so we think of functions $f\colon \mathbb{C} \to \mathbb{C}$ also as *vector fields* $f\colon \mathbb{R}^2 \to \mathbb{R}^2$. Seen in this light, the complex derivative is a special case of the multi-variable real derivative; the precise relationship is given by the Cauchy-Riemann Equations. Similarly, the complex path integral can be expressed as the combination of real path integrals (see Proposition 11.13). We thus review real differential calculus in Chap. 9, and devote Chap. 10 to studying real path integrals and vector fields. Throughout, though, we observe other connections between topology and analysis. For example, the purely topological concept of the *winding number* of a path around a point (how many times it goes around the point), introduced in Chap. 9, can be characterised using either real (Proposition 10.32) or complex (Proposition 11.21) path integrals. In Chap. 11 we bring everything together and present the basics of complex analysis.

**Finally, Riemann Surfaces**

In Chap. 12 we define Riemann surfaces. We first concentrate on the *Riemann sphere*, which is obtained by adding a "pole" to the complex plane. Equivalently, this is the "point at infinity" which is added to the complex affine line; in other words, the Riemann sphere is identified with the projective line over the complex numbers. A complex-differentiable map between Riemann surfaces is called *holomorphic*, and a holomorphic map to the Riemann sphere is called *meromorphic*. Thus, a meromorphic function is like an analytic function, except that we also allow an "infinite value"; so a function such as $z \mapsto 1/z$ can be considered defined at 0 as well.

Studying meromorphic functions also yields results about analytic functions (functions to $\mathbb{C}$); the *calculus of residues*, counting zeros and poles (points mapped to $\infty$) of meromorphic functions, results in the open mapping theorem mentioned above.

In addition, in Chap. 12, we investigate compact Riemann surfaces. We examine the Riemann surfaces for the logarithm and the $n$th roots; these allow us to deal properly with multi-valued functions, and are used in Chap. 15 to parameterise curves near singular points. We also introduce *meromorphic differentials* on Riemann surfaces. These generalise differential forms that are used in Chap. 10 for integration along paths, and are used to invert the isomorphism theorem in Chap. 14.

### 1.2.3 Part III: Curves and Surfaces

The third part of the book ties the first two together. In Chap. 13 we show that nonsingular curves in the complex plane are Riemann surfaces. The main tool used is the analytic implicit function theorem. We also review intersection multiplicities of lines with curves; we see that over the complex numbers, we can recover our original intuition that intersection multiplicity is determined by perturbing the line a little.

In Chap. 14 we prove the isomorphism theorem for complex tori and elliptic curves. We start with elliptic functions, which we define and analyse in general, and then proceed to our main example: Weierstrass's function $\wp$. We verify that $\wp$ and its derivative parameterise an elliptic curve, and show that the induced map is an isomorphism between a complex torus and the curve, which preserves both algebraic and analytic structure. We then prove the *inversion theorem* (also known as the *uniformisation theorem* for elliptic curves), showing the reverse: every elliptic curve is obtained in this way.

As mentioned, in Chap. 15 we return to intersection multiplicities using parameterisations of curves. We pass between formal power series and the analytic functions that they define, except that at singular points, we need to choose $n$th roots as well; this brings into play *fractional* power series, and the Riemann surfaces for roots. Our study allows us to isolate separate *places* of a curve near a singular point, and to show that these have implicit definitions.

Finally, in Chap. 16, we provide a brief history of elliptic functions, and illustrate their historical development by looking once again at the circular functions.

## 1.3 Preliminaries, and Some Notation

We will strive to keep notation standard, and to point out divergence from common practice. We assume basic set theory. For example, we use $A \cap B$ to denote the intersection of two sets $A$ and $B$, $A \cup B$ is their union, $A \setminus B$ is the set-theoretic difference (the elements of $A$ which are not in $B$). If $\mathcal{A}$ is a collection of sets, then $\bigcup \mathcal{A}$ is the union of all the sets in $\mathcal{A}$; similarly we use $\bigcap \mathcal{A}$. The *domain* of a function $f$, denoted $\operatorname{dom} f$, is the collection of all possible inputs for $f$: those $x$ for which $f(x)$ is defined. The *range* (or *image*) of $f$, denoted by $\operatorname{range} f$, is the collection of all the outputs of $f$. A function is *onto* a set $Y$ (surjective) if $Y = \operatorname{range} f$. The *pointwise image* of a set $Y$ under a function $f$, denoted by $f[Y]$, is the collection of all outputs of $f$ on inputs from $Y$: $f[Y] = \{f(x) : x \in Y\}$. We do not assume that $Y \subseteq \operatorname{dom} f$ (see Remark 8.7). The *pointwise pullback* of a set $Y$ under a function $f$, denoted by $f^{-1}[Y]$, is the collection of all pre-images by $f$ of the outputs in $Y$: $f^{-1}[Y] = \{x : f(x) \in Y\}$. We let $f\restriction_X$ denote the restriction of $f$ to a set $X$; so $f[X]$ is the range of $f\restriction_X$. The composition of two functions $f$ and $g$ is denoted by $g \circ f$; we apply $f$ first, that is, $(g \circ f)(x) = g(f(x))$. We do not assume that $\operatorname{range} f \subseteq \operatorname{dom} g$ (again, see Remark 8.7), although in many instances this will be the case.

We assume the reader is familiar with notions such as a function being 1–1 (injective); bijections between sets; and equivalence relations and their equivalence classes. A set is *countable* if it is finite, or there is a bijection between it and the set $\mathbb{N}$ of natural numbers. The set $\mathbb{Z}$ of integers (positive and negative whole numbers, and 0) is countable, as is the set $\mathbb{Q}$ of rational numbers (fractions of integers). On the other hand the set $\mathbb{R}$ of real numbers and $\mathbb{C}$ of complex numbers are uncountable (not

countable). We assume some familiarity with basic operations on complex numbers (addition, multiplication, conjugation, inverses). We assume basic combinatorics, for example, that $\binom{n}{k} = n!/(k!(n-k)!)$, where $\binom{n}{k}$ is the number of $k$-element subsets of an $n$-element set.

# Part I

# Algebraic Curves


# Algebra

# 2

In this chapter we discuss the algebraic tools that we will need. Some of the presentation—for example, of abelian groups, or Gauss's lemma—will be fairly standard, and so at times we give only sketches of proofs, with references to other texts. In contrast, we expect that notions such as rings of formal power series may be new to the reader, so we present them in greater detail. When discussing unique factorisation, we introduce the notion of a multiset, which will be useful throughout the book. We also develop linear algebra and the determinant over integral domains which may fail to be fields.

## 2.1 Polynomials and Power Series

Throughout this book, we attempt to be as concrete as possible. For example, we will not define general vector spaces over a field $F$, rather, we will just deal with subspaces of $F^n$. However, we need to start somewhere. Informally speaking, an *integral domain* is a generalisation of number systems such as the integers, rationals or real numbers: objects that can be added and multiplied. More formally, an integral domain consists of a set $R$ equipped with two binary operations $\cdot_R$ and $+_R$, one unary operation $-_R$, and two distinct designated elements $0_R$ and $1_R$, provided that it satisfies the following conditions:

**Associativity**: for all $a, b$ and $c$ in $R$, $a+_R(b+_R c) = (a+_R b)+_R c$ and $a \cdot_R (b \cdot_R c) = (a \cdot_R b) \cdot_R c$.
**Commutativity**: for all $a$ and $b$ in $R$, $a +_R b = b +_R a$ and $a \cdot_R b = b \cdot_R a$.
**Distributivity**: for all $a, b$ and $c$ in $R$, $a \cdot_R (b +_R c) = (a \cdot_R b) +_R (a \cdot_R c)$.
**Identity elements**: for all $a \in R$, $a +_R 0_R = a$ and $a \cdot_R 1_R = a$.
**Additive inverses**: for all $a \in R$, $a +_R (-_R a) = 0_R$.
**No zero divisors**: for all $a, b \in R$, if $a \cdot_R b = 0_R$ then $a = 0_R$ or $b = 0_R$.



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_2

If we omit the last condition (no zero divisors), then we get the notion of a *commutative ring with unity*. We will not really use more general kinds of rings (non-commutative rings, or rings without a multiplicative identity element) so we will just call them *rings*. The standard examples are number systems such as $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ (equipped with the usual addition, multiplication, negation, and 0 and 1). The latter three are also *fields*: they are integral domains satisfying the extra property of having multiplicative inverses for nonzero elements:

**Multiplicative Inverses**: for all $a \in R$, if $a \neq 0_R$ then there is some $b \in R$ such that $a \cdot_R b = 1_R$.

Another standard example is $\mathbb{Z}/(n)$ (for $n \geqslant 2$), the ring of integers modulo $n$, equipped with addition, negation and multiplication mod $n$. Here is another:

*Exercise 2.1* Let $R = \mathbb{Z}^2 = \mathbb{Z} \times \mathbb{Z}$ be the collection of all ordered pairs of integers. For $(a, b), (c, d) \in R$, let

- $(a, b) +_R (c, d) = (a + c, b + d)$ and $-_R(a, b) = (-a, -b)$; and
- $(a, b) \cdot_R (c, d) = (ac + 2bd, bc + ad)$.

Also let $0_R = (0, 0)$ and $1_R = (1, 0)$. (a) Show that $(R; +_R, -_R, \cdot_R, 0_R, 1_R)$ is a ring. (b) Show that there is some $(a, b) \in R$ such that $(a, b) \cdot_R (a, b) = (2, 0)$. (c) Show that $R$ is an integral domain. (Hint: $\sqrt{2}$ is irrational.) «

There are many ways of manufacturing rings and integral domains from existing ones. The two main ones we will use are forming rings of polynomials and formal power series. The former may be familiar: a (formal) *polynomial* with coefficients from a ring $R$ is an object of the form

$$a_0 + a_1 x + a_2 x^2 + \cdots + a_d x^d$$

where $a_0, a_1, \ldots, a_d$ are elements of $R$. We usually think of them as functions (and we will get to that later this chapter), but right now we just think of them as formal objects, and what interests us right now is that we can add and multiply them: for example, $f = x^2 + 2x + 1$ and $g = x^3 - 3x$ are both polynomials with coefficients in $\mathbb{Z}$; $f + g = x^3 + x^2 - x + 1$ and $fg = x^5 + 2x^4 - 2x^3 - 6x^2 - 3x$. Formal power series are similar, except that we allow "infinite sums".

Suppose that $(R; +_R, \cdot_R, -_R, 0_R, 1_R)$ is a ring. Let $x$ be what we call an *indeterminate* or a *variable*. A *formal power series in $x$ with coefficients in $R$* is an object of the form

$$a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots,$$

(usually abbreviated as $\sum_{k \in \mathbb{N}} a_k x^k$), where $a_0, a_1, a_2, \ldots$ is an *infinite* sequence of elements of $R$; these are the coefficients of the formal power series. Two formal power series $\sum a_k x^k$ and $\sum b_k x^k$ are equal when $a_0 = b_0$, $a_1 = b_1$, $a_2 = b_2$,.... The collection of all formal power series with coefficients in $R$ is denoted by $R[\![x]\!]$.

On this collection we define a ring structure. Let $\alpha = \sum a_k x^k$ and $\beta = \sum b_k x^k$ be two formal power series. We define:

$$\begin{aligned}
\alpha +_{R[\![x]\!]} \beta &= (a_0 +_R b_0) + (a_1 +_R b_1)x + (a_2 +_R b_2)x^2 + \cdots, \\
-_{R[\![x]\!]} \alpha &= (-_R a_0) + (-_R a_1)x + (-_R a_2)x^2 + \cdots, \\
\alpha \cdot_{R[\![x]\!]} \beta &= (a_0 \cdot_R b_0) \; + \\
&\qquad ((a_0 \cdot_R b_1) +_R (a_1 \cdot_R b_0))\; x \; + \\
&\qquad ((a_0 \cdot_R b_2) +_R (a_1 \cdot_R b_1) +_R (a_2 \cdot_R b_0))\; x^2 \; + \cdots, \\
0_{R[\![x]\!]} &= 0_R + 0_R x + 0_R x^2 + \cdots, \quad \text{and} \\
1_{R[\![x]\!]} &= 1_R + 0_R x + 0_R x^2 + 0_R x^3 + \cdots
\end{aligned}$$

In particular, formal power series are multiplied like polynomials, by distributing the product and "collecting terms"; the point is that for each $k$ there are only finitely many pairs $i$ and $j$ such that $x^i x^j = x^k$.

*Exercise 2.2* Show that in $\mathbb{Z}[\![x]\!]$, $(1 - x) \cdot (1 + x + x^2 + x^3 + \cdots) = 1$. «

### 2.1.1 The Category of Rings

By identifying $a \in R$ with the "constant" series $a + 0_R x + 0_R x^2 + \cdots$ we think of $R$ as a subring of $R[\![x]\!]$. We need to define this formally.

A ring $S$ is a *subring* of a ring $R$ if $S \subseteq R$ and the ring structure on $S$ is the one inherited from the ring structure on $R$. Namely, $0_S = 0_R$, $1_S = 1_R$, and for all $a, b \in S$, $a +_S b = a +_R b$, $a \cdot_S b = a \cdot_R b$ and $-_S a = -_R a$. For example, the integers $\mathbb{Z}$ form a subring of the rationals $\mathbb{Q}$, which form a subring of the reals $\mathbb{R}$: $3 + 5 = 8$ whether we think of these numbers as integers, rationals or reals.

A subring of an integral domain is an integral domain, but a subring of a field is not necessarily a field. If $R$ is a ring, $S \subseteq R$, $0_R, 1_R \in S$ and $S$ is closed under the operations of $R$—for all $a, b \in S$, $a +_R b \in S$, $a \cdot_R b \in S$ and $-_R a \in S$—then $S$, equipped with the restrictions to $S$ of the operations of $R$, is a ring, and is therefore a subring of $R$.

*Exercise 2.3* Let $\rho$ be a complex number such that $\rho^2 \in \mathbb{Z}$ (for example, $\rho = i$, $\rho = \sqrt{2}$ or $\rho = i\sqrt{2}$). Let $\mathbb{Z}[\rho] = \{a + \rho b : a, b \in \mathbb{Z}\}$. (a) Show that $\mathbb{Z}[\rho]$ is a subring of $\mathbb{C}$. (b) Show that $\mathbb{Z}[\rho]$ is *minimal* in the following sense: if $S$ is a subring of $\mathbb{C}$ such that $\rho \in S$, then $\mathbb{Z}[\rho] \subseteq S$.[1] «

[1] An element of $\mathbb{Z}[i]$ is called a *Gaussian integer*.

Let $R$ and $S$ be rings. A *ring homomorphism* from $R$ to $S$ is a function $\psi : R \to S$ such that:

- $\psi(0_R) = 0_S$ and $\psi(1_R) = 1_S$; and
- for all $a, b \in R$, $\psi(a +_R b) = \psi(a) +_S \psi(b)$, $\psi(a \cdot_R b) = \psi(a) \cdot_S \psi(b)$, and $\psi(-_R a) = -_S \psi(a)$.

Ring homomorphisms are the *structure-preserving* functions. An example is the map $m \mapsto m \pmod n$, which is a ring homomorphism from $\mathbb{Z}$ to $\mathbb{Z}/(n)$. The composition of ring homomorphisms is a ring homomorphism. The image of a ring homomorphism is a subring of the range.

*Remark 2.4* The definitions we gave of subrings and ring homomorphisms are somewhat non-standard, in that they require that $1_S = 1_R$ (for subrings) and that $\psi(1_R) = 1_S$ (for homomorphisms). These conditions are often dropped, giving examples such as the even integers being a subring of $\mathbb{Z}$ (a ring without unity), or the map $m \mapsto 2m$ being a ring homomorphism from $\mathbb{Z}$ to $\mathbb{Z}$. On the other hand, the conditions $\psi(0_R) = 0_S$ and $\psi(-_R a) = -_S \psi(a)$ follow from the other ones. «

An *isomorphism* is a homomorphism which is also a bijection: one-to-one (injective) and onto (surjective). Two rings $R$ and $S$ are *isomorphic* if there is an isomorphism between them. We write $R \cong S$. Two rings being isomorphic means that they "are the same" ring, except that the elements are labelled differently. The inverse of an isomorphism is an isomorphism, and the composition of isomorphisms is an isomorphism. This implies that being isomorphic is an equivalence relation.

*Exercise 2.5* Let $R$ be the ring from Exercise 2.1, and define $\psi : R \to \mathbb{R}$ by letting $\psi(a, b) = a + b\sqrt{2}$. Show that $\psi$ is an isomorphism from $R$ to the subring $\mathbb{Z}[\sqrt{2}]$ (Exercise 2.3) of $\mathbb{R}$. «

An *embedding* of $R$ into $S$ is an isomorphism from $R$ to a subring of $S$. Equivalently, it is a one-to-one homomorphism from $R$ to $S$. A homomorphism $\psi : R \to S$ is an embedding if and only if for all nonzero $a \in R$, $\psi(a) \neq 0_S$.

**Proposition 2.6** *The map $a \mapsto (a + 0_R x + 0_R x^2 + \cdots)$ is an embedding of $R$ into $R[\![x]\!]$.*

***Proof*** For preservation of multiplication, note that for all $k > 0$, the coefficient of $x^k$ in the product of $a + 0_R x + 0_R x^2 + \cdots$ and $b + 0_R x + 0_R x^2 + \cdots$ is $a \cdot 0 + 0 \cdot 0 + \cdots + 0 \cdot 0 + 0 \cdot b = 0_R$. □

We *identify* $R$ with its image under this embedding, which means that we consider $R$ as a subring of $R[\![x]\!]$. We also identify the variable $x$ with the formal power series $0 + 1 \cdot x + 0 \cdot x^2 + 0 \cdot x^3 + \cdots$. This is justified because according to our definition of multiplying formal power series, $x \cdot x = 0 + 0 \cdot x + 1 \cdot x^2 + 0 \cdot x^3 + 0 \cdot x^4 + \cdots$, and so can be rightfully named $x^2$, as defined above. Similarly

$x^3 = 0 + 0 \cdot x + 0 \cdot x^2 + 1 \cdot x^3 + 0 \cdot x^4 + 0 \cdot x^5 + \cdots$. We call formal power series of the form $ax^k$ *monomials*.

**Proposition 2.7** *If $R$ is an integral domain then so is $R[\![x]\!]$.*

***Proof*** Let $\alpha = \sum a_k x^k$ and $\beta = \sum b_k x^k$ be nonzero elements of $R[\![x]\!]$. This means that there are $m$ and $n$ such that $a_m \neq 0_R$ and $b_n \neq 0_R$. Choose least such $m$ and $n$. The coefficient of $x^{m+n}$ in $\alpha\beta$ is $a_m \cdot_R b_n$, which is nonzero, and so $\alpha \cdot_{R[\![x]\!]} \beta \neq 0$. □

### Simplifying Notation

If no confusion will ensue, we drop the subscript $R$ from the operations and designated elements of $R$, so we write $a+b$ instead of $a+_R b$, $ab$ instead of $a \cdot_R b$ and so on, even if $+_R$ and $\cdot_R$ are not the usual operations of addition and multiplication on numbers.

We let $a - b$ abbreviate $a + (-b)$.

We drop parentheses and let $\cdot_R$ take precedence over $+_R$; so for $a$, $b$ and $c$ in $R$, $a + bc$ means $a +_R (b \cdot_R c)$, not $(a +_R b) \cdot_R c$.

For $a \in R$ we let $a^0 = 1_R$, $a^1 = a$, and in general, for $n \geqslant 1$, $a^n = a \cdot a^{n-1}$ be the result of multiplying (in the ring $R$) $n$ many $a$'s.

### Derived Properties

Properties shared by all rings are derived from the axioms from p. 17. We list some.

**Lemma 2.8** *Let $R$ be a ring, and let $a, b, c \in R$. (a) If $a + b = 0$ then $b = -a$. (b) $-(-a) = a$. (c) If $a + b = a + c$ then $b = c$. (d) $a \cdot 0 = 0$. (e) $(-a)b = a \cdot (-b) = -(ab)$. (f) $a(b - c) = ab - ac$.*

***Proof*** Most are easy and are left as an exercise. For (d), note that $a \cdot 0 = a \cdot (0+0) = a \cdot 0 + a \cdot 0$, and subtract. □

## 2.1.2 Back to Formal Power Series

### Infinite Sums

We have defined formal power series to be expressions of the form $a_0 + a_1 x + a_2 x^2 + \ldots$, and explained how to add, subtract and multiply these objects. This doesn't quite explain the notation: is this really an infinite sum of monomials?

Infinite sums, in general, cannot be defined, since it is not clear how to add infinitely many elements of the ring $R$. Under some circumstances, though, we can.

Suppose, for example, that

$$\begin{aligned}\alpha_0 &= a_{0,0} + a_{0,1}x + a_{0,2}x^2 + a_{0,3}x^3 + a_{0,4}x^4 + \cdots,\\ \alpha_1 &= \quad a_{1,1}x + a_{1,2}x^2 + a_{1,3}x^3 + a_{1,4}x^4 + \cdots,\\ \alpha_2 &= \qquad\quad a_{2,2}x^2 + a_{2,3}x^3 + a_{2,4}x^4 + \cdots,\end{aligned}$$

is an infinite sequence of formal power series, with the first $m$ coefficients of $\alpha_m$ being 0. Then we can define

$$\begin{aligned}\sum \alpha_m = a_{0,0} &+\\ &(a_{0,1} + a_{1,1})\, x +\\ &(a_{0,2} + a_{1,2} + a_{2,2})\, x^2 + \cdots;\end{aligned}$$

in general, we can define $\sum_m \alpha_m$ if for every $k$, for all but finitely many $m$, the coefficient of $x^k$ in $\alpha_m$ is $0_R$. In particular, if $\alpha_m = a_m x^m$, then $\sum \alpha_m = a_0 + a_1 x + a_2 x^2 + \cdots$, so this gives the infinite sum notation some formal sense.

### Several Variables

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a tuple of distinct indeterminate variables. A formal power series in $\boldsymbol{x}$ with coefficients in a ring $R$ is an object of the form

$$\sum_{(k_1,k_2,\ldots,k_n)\in\mathbb{N}^n} a_{k_1,\ldots,k_n} x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n}$$

where $a_{k_1,\ldots,k_n} \in R$. For example, with two variables $x$ and $y$, elements of $R[\![x, y]\!]$ are of the form

$$\begin{aligned}&a_{0,0} + a_{1,0}x + a_{2,0}x^2 + a_{3,0}x^3 + \cdots\\ &+ a_{0,1}y + a_{1,1}xy + a_{2,1}x^2y + a_{3,1}x^3y + \cdots\\ &+ a_{0,2}y^2 + a_{1,2}xy^2 + a_{2,2}x^2y^2 + a_{3,2}x^3y^2 + \cdots\\ &\vdots\end{aligned}$$

For brevity we write, for $\boldsymbol{k} = (k_1, \ldots, k_n) \in \mathbb{N}^n$, $a_{\boldsymbol{k}}\boldsymbol{x}^{\boldsymbol{k}}$ for $a_{\boldsymbol{k}} x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n}$. As expected, two formal power series $\sum a_{\boldsymbol{k}}\boldsymbol{x}^{\boldsymbol{k}}$ and $\sum b_{\boldsymbol{k}}\boldsymbol{x}^{\boldsymbol{k}}$ are equal when $a_{\boldsymbol{k}} = b_{\boldsymbol{k}}$ for all $\boldsymbol{k} \in \mathbb{N}^n$.

Addition and multiplication of formal power series with several variables works as one would expect:

$$\sum a_{\boldsymbol{k}}\boldsymbol{x}^{\boldsymbol{k}} + \sum b_{\boldsymbol{k}}\boldsymbol{x}^{\boldsymbol{k}} = \sum (a_{\boldsymbol{k}} + b_{\boldsymbol{k}})\boldsymbol{x}^{\boldsymbol{k}};$$

negation is similar. Powers are additive in multiplication. Using the notation $(k_1, \ldots, k_n) + (m_1, \ldots, m_n) = (k_1 + m_1, \ldots, k_n + m_n)$, in brief

$$\left(\sum a_{\boldsymbol{k}} \boldsymbol{x}^{\boldsymbol{k}}\right) \cdot \left(\sum b_{\boldsymbol{k}} \boldsymbol{x}^{\boldsymbol{k}}\right) = \sum_{\boldsymbol{k} \in \mathbb{N}^n} \sum_{\boldsymbol{i}+\boldsymbol{j}=\boldsymbol{k}} (a_{\boldsymbol{i}} b_{\boldsymbol{j}}) \boldsymbol{x}^{\boldsymbol{k}},$$

Noting that for each $\boldsymbol{k}$ there are only finitely many pairs of tuples $(\boldsymbol{i}, \boldsymbol{j})$ such that $\boldsymbol{i} + \boldsymbol{j} = \boldsymbol{k}$.

As with one variable, we obtain a ring $R[\![\boldsymbol{x}]\!]$ and an embedding of $R$ into $R[\![\boldsymbol{x}]\!]$ mapping $a \in R$ to the "constant" formal power series $a$ (with zero coefficient for $\boldsymbol{x}^{\boldsymbol{k}}$ for $\boldsymbol{k} \neq \boldsymbol{0}$). We similarly refer to the elements $x_1$, $x_2$, etc. A *monomial* is a power series of the form $a\boldsymbol{x}^{\boldsymbol{i}}$, that is, the product of a constant with finitely many indeterminates.

If $R$ is an integral domain then so is $R[\![\boldsymbol{x}]\!]$. This can also be seen by induction on the number of variables, using the fact that

$$R[\![x, y]\!] \cong R[\![x]\!][\![y]\!]$$

—via the map $\sum a_{k,m} x^k y^m \mapsto \sum_m \left(\sum_k a_{k,m} x^k\right) y^m$—and similarly for more variables.

### 2.1.3 More on Polynomials

We can now identify polynomials as a special kind of power series: those which are the sum of finitely many monomials, that is, those which have only finitely many nonzero coefficients. The collection of polynomials in $\boldsymbol{x}$ with coefficients from $R$ is denoted by $R[\boldsymbol{x}]$. This is a subring of $R[\![\boldsymbol{x}]\!]$, as the sum and product of polynomials is a polynomial. The embedding of $R$ into $R[\![\boldsymbol{x}]\!]$ is actually into $R[\boldsymbol{x}]$ (every constant power series is a polynomial); as above, we identify $a \in R$ with the constant polynomial $a$, so we consider $R$ as a subring of $R[\boldsymbol{x}]$. If $R$ is an integral domain then so is $R[\boldsymbol{x}]$ (recall that a subring of an integral domain is an integral domain). The isomorphism between $R[\![x, y]\!]$ and $R[\![x]\!][\![y]\!]$ restricts to a bijection between $R[x, y]$ and $R[x][y]$, so the two rings are isomorphic by this map.[2]

#### The Degree of a Polynomial

The *degree* of a monomial $ax_1^{k_1} x_2^{k_2} \cdots x_n^{k_n}$ is $k_1 + k_2 + \cdots + k_n$. The degree of a nonzero polynomial $f$ is the maximum of the degrees of the monomials appearing in $f$. It is denoted by $\deg f$. Note that this only makes sense for polynomials; other formal power series will have monomials of unbounded degrees.

[2] Note that $R[\![x]\!][y]$ is a proper subring of $R[y][\![x]\!]$; the latter allows unbounded powers of $y$ as the coefficients of powers of $x$ vary.

The zero polynomial has degree $-\infty$. The nonzero polynomials of degree 0 are the nonzero constant polynomials, i.e., the elements of $R \setminus \{0\}$. A polynomial of degree 1 is called *linear*. There are more such names; *quadratic* for degree 2, *cubic* for degree 3, *quartic* for degree 4, *quintic* for degree 5, and so on.

In one variable, the degree of a nonzero polynomial $f = \sum a_k x^k$ is the greatest number $d$ such that $a_d \neq 0$. The coefficient $a_d$ is called the *leading coefficient* of $f$.

By considering leading coefficients we observe:

**Proposition 2.9** *Suppose that $R$ is an integral domain. Let $f, g \in R[x]$.*

(a) $\deg(f+g) \leqslant \max\{\deg f, \deg g\}$. *If* $\deg f \neq \deg g$ *then* $\deg(f+g) = \max\{\deg f, \deg g\}$.
(b) $\deg(fg) = \deg f + \deg g$. □

In the latter equality, we agree that $-\infty + d = -\infty$ for all $d \in \mathbb{N}$.

For polynomials of several variables, we can measure the degree of a polynomial relative to only some of the variables present; for a polynomial $f$ and a tuple $\boldsymbol{z}$ of indeterminates, we write $\deg_{\boldsymbol{z}} f$ for the degree of $f$ where $\boldsymbol{z}$ are considered variables; other indeterminates may appear in the coefficients. For example, let $f = x^2y^3 + x^4$. Then $\deg_{x,y} f = 5$, $\deg_x f = 4$ and $\deg_y f = 3$.[3] For the analogue of Proposition 2.9 for more than one variable see p. 75.

## Polynomial Substitution

As discussed above, in general rings $R$, it is not clear how to sum an infinite series of elements. Thus, formal power series are just that—formal objects; they do not define functions on $R$. Even when infinite sums are sometimes defined, like in $\mathbb{R}$, some power series define functions on all of $\mathbb{R}$, some on an interval, some only converge at 0. This will be discussed at length in the second part of the book.

Polynomials, however, do define functions: if $f = a_0 + a_1 x + \ldots a_d x^d$, then the function determined by $f$ maps $b \in R$ to $a_0 + a_1 b + \cdots + a_d b^d$, where addition and multiplication is computed, of course, in $R$. We can generalise this to more than one variable. If $R$ is a ring, $\boldsymbol{x} = (x_1, \ldots, x_n)$ is a tuple of indeterminates, and $f = \sum a_{\boldsymbol{k}} \boldsymbol{x}^{\boldsymbol{k}}$ is a polynomial in $R[\boldsymbol{x}]$, then $f$ defines a function from $R^n$ to $R$ (which we also denote by $f$) by mapping $\boldsymbol{b} \in R^n$ to $f(\boldsymbol{b}) = \sum_{\boldsymbol{k} \in \mathbb{N}^n} a_{\boldsymbol{k}} \boldsymbol{b}^{\boldsymbol{k}}$. This is a finite sum since $f$ is a polynomial.

If $f = a$ is a constant (an element of $R$), then $f(\boldsymbol{b}) = a$ for all $\boldsymbol{b} \in R^n$, i.e., $f$ defines a constant function. If $f = x_i$ then $f(b_1, b_2, \ldots, b_n) = b_i$ (this is a projection function). We note that the function defined by $f$ depends not only on $f$ but also on the ordering we imposed on the variables. For example, the polynomial $f = y - x^2$ defines the function $(a, b) \mapsto b - a^2$ if the variables are ordered as $(x, y)$, but defines the function $(a, b) \mapsto a - b^2$ if they are ordered as $(y, x)$; this, even

[3] The point is that $R[x, y] \cong R[x][y] \cong R[y][x]$, but degrees in these rings are computed differently.

though $R[x, y]$ and $R[y, x]$ are isomorphic, and we really have been considering them as the same ring.

*Remark 2.10* Note that like formal power series, we treat polynomials as *formal objects* which we add and multiply; they are not identical to the functions that they define. In fact, if $R$ is finite, then there are only finitely many functions from $R$ to $R$, but $R[x]$ is infinite; many polynomials define the same function from $R$ to $R$. However, in the cases that we will eventually deal with, $R$ will be an infinite integral domain, and in this case, we will show that two distinct polynomials in $R[\boldsymbol{x}]$ must define different functions (see Proposition 2.18). For a concrete counter-example when $R$ is finite, see Exercise 2.71. «

For a fixed $\boldsymbol{b} \in R^n$, the map $f \mapsto f(\boldsymbol{b})$ is a ring homomorphism from $R[\boldsymbol{x}]$ to $R$. If $S$ is a subring of $R$ then $S[\boldsymbol{x}]$ is a subring of $R[\boldsymbol{x}]$ and so the restriction of this map to $S[\boldsymbol{x}]$ is a ring homomorphism from $S[\boldsymbol{x}]$ to $R$. In fact:

**Proposition 2.11** *If $S$ is a subring of $R$ and $\boldsymbol{b} \in R^n$, then $f \mapsto f(\boldsymbol{b})$ is the unique ring homomorphism from $S[\boldsymbol{x}]$ to $R$ which is the identity on $S$ and maps each $x_i$ to $b_i$.* □

*Remark 2.12* We can substitute polynomials into other polynomials. Let $f \in R[\boldsymbol{x}]$; let $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)$ be another sequence of variables (not necessarily disjoint from $\boldsymbol{x}$). We can substitute an $n$-tuple $\boldsymbol{g} = (g_1, g_2, \ldots, g_n)$ of polynomials in $R[\boldsymbol{y}]$ into $f$, obtaining the polynomial $f(\boldsymbol{g}) \in R[\boldsymbol{y}]$. Substitution is transitive: for $\boldsymbol{b} \in R^m$, $f(\boldsymbol{g})(\boldsymbol{b}) = f(\boldsymbol{g}(\boldsymbol{b}))$ (where as expected $\boldsymbol{g}(\boldsymbol{b}) = (g_1(\boldsymbol{b}), \ldots, g_n(\boldsymbol{b}))$). Formally, this can be proved using Proposition 2.11: fixing $\boldsymbol{g}$ and $\boldsymbol{b}$, both functions $f \mapsto (f(\boldsymbol{g}))(\boldsymbol{b})$ and $f \mapsto f(\boldsymbol{g}(\boldsymbol{b}))$ are homomorphisms from $R[\boldsymbol{x}]$ to $R$ extending the identity on $R$ and mapping each $x_i$ to $g_i(\boldsymbol{b})$. «

## 2.2 Unique Factorisation

We turn to the study of factorisation, mainly in polynomial rings.

### 2.2.1 Divisibility in Integral Domains

Let $R$ be an integral domain. For nonzero $a, b \in R$, we say that *$a$ divides $b$* in $R$ (and write $a \mid b$) if $b = ac$ for some $c \in R$. The ring matters: 2 does not divide 5 in the integers $\mathbb{Z}$ but as elements of the rationals $\mathbb{Q}$, 2 does divide 5. Indeed in a field divisibility is not interesting: all nonzero elements divide each other.

The divisibility relation is reflexive and transitive (it is a *pre partial ordering*). This implies that the relation "$a$ divides $b$ and $b$ divides $a$" is an equivalence relation on $R$. This equivalence relation is called *association* and is denoted simply by $a \sim b$

(or $\sim_R$, when more than one ring is considered). In the integers, each integer $n$ is an associate of itself and of $-n$.

Another way to characterise association is by considering *units*. A unit of $R$ is an element which has a multiplicative inverse in $R$. In other words, an element which divides 1. The collection of units of $R$ is denoted by $R^*$. For example $\mathbb{Z}^* = \{-1, 1\}$. The collection of units contains 1 and is closed under taking products. $R$ is a field exactly when $R^* = R \setminus \{0\}$. $R^*$ itself is an *association class*—an equivalence class of the association relation. This follows from the fact that 1 divides every element of $R$.

*Exercise 2.13* By Exercise 2.2, $1 - x$ and $1 + x + x^2 + x^3 + \cdots$ are units of $\mathbb{Z}[\![x]\!]$. Generalise this as follows. Let $R$ be an integral domain. Show that a formal power series $a_0 + a_1x + a_2x^2 + \cdots$ in $R[\![x]\!]$ is a unit if and only if $a_0$ has a multiplicative inverse in $R$. «

The following fact is simple and useful.

*Exercise 2.14* Two elements $a$ and $b$ of $R$ are associates if and only if there is some unit $u \in R^*$ such that $a = bu$. «

*Remark 2.15* The divisibility relation induces a partial ordering (transitive, reflexive, antisymmetric) on the collection of association classes: the class $[a]_\sim$ of $a$ divides the class $[b]_\sim$ of $b$ if $a$ divides $b$; this does not depend on the representatives chosen. «

### Divisibility in $R[x]$

Again we assume that $R$ is an integral domain. If $f, g \in R[x]$ and $f$ divides $g$ (in $R[x]$), then $\deg f \leqslant \deg g$; this follows from Proposition 2.9(2). Hence if $f \sim g$ then $\deg f = \deg g$. It follows that all units of $R[x]$ are constants, and in fact, $(R[x])^* = R^*$; for $a, b \in R$, $a$ divides $b$ in $R$ if and only if $a$ divides $b$ in $R[x]$. Two polynomials $f = a_0 + a_1x + \cdots a_dx^d$ and $g = b_0 + b_1x + \cdots b_ex^e$ are associates if and only if $\deg f = \deg g$ and there is some constant unit $\lambda \in R^*$ such that $b_0 = \lambda a_0, b_1 = \lambda a_1, \ldots$

Again let $f = a_0 + a_1x + \cdots a_dx^d$ and $g = b_0 + b_1x + \cdots b_ex^e$ in $R[x]$ (with $a_d, b_e \neq 0$). Suppose that $d \geqslant e$. If $b_e$ divides $a_d$, say $a_d = b_ec$, then we can subtract $cx^{d-e}g$ from $f$ to obtain another polynomial $f_1$ of degree strictly smaller than $f$. We replace $f$ by $f_1$ (but keep $g$) and do the same, and obtain a sequence $f_2, f_3, \ldots$. Either we get stuck at some point because the leading coefficient of $g$ does not divide the leading coefficient of some $f_i$; or we obtain a polynomial $f_i$ of degree smaller than $e = \deg g$. In the latter case, we note that all the polynomials we subtracted were multiples of $g$; rearranging, we can write $f = qg + r$ with $\deg r < \deg g$. We call $q$ the *quotient* and $r$ the *remainder* obtained by dividing $f$ by $g$. Note that these, if exist, are unique: if $q_0g + r_0 = f = q_1g + r_1$ then subtracting, $g \mid (r_0 - r_1)$ and as $\deg(r_0 - r_1) < \deg g$ it must be that $r_0 = r_1$, whence $q_0 = q_1$ as well. Thus, $g$ divides $f$ if and only if $r = 0$. This process of long division is guaranteed to succeed

if $b_e$, the leading coefficient of $g$, divides all the nonzero elements of $R$, i.e., if it is a unit. There are two special cases:

- When $R$ is a field (so every nonzero $b \in R$ is a unit);
- $b_e = 1$.

In the second case we call $g$ a *monic* polynomial.

Let $f \in R[x]$. An element $a \in R$ is a *root* of $f$ if $f(a) = 0$ (recall that each formal polynomial defines a function $R \to R$ by substitution). For any $a \in R$, dividing $f$ by the monic polynomial $x - a$ we obtain a quotient $q$ and remainder $r$; since $\deg r < \deg(x - a) = 1$ we see that $r \in R$ is a constant. We substitute $a$ into the equation $f = (x - a)q + r$ (formally, this uses Proposition 2.11). Since $r$ is a constant, $r(a) = r$ and we get $r = r(a) = f(a) - (a - a)q(a) = f(a)$. We conclude:

**Theorem 2.16** *An element $a$ of an integral domain $R$ is a root of $f \in R[x]$ if and only if $x - a$ divides $f$ in $R[x]$. The number of roots of any nonzero polynomial $f \in R[x]$ is at most* $\deg f$.

The last part follows by induction on $\deg f$: if $a$ is a root of $f$, then $f = (x-a)g$ for some $g \in R[x]$; if $b$ is another root of $f$, then it must be a root of $g$ (using the fact that $R$ is an integral domain). And $\deg g = \deg f - 1$ (Proposition 2.9).

*Exercise 2.17* Show that if $R$ is an infinite integral domain then distinct polynomials in $R[x]$ define distinct functions from $R$ to $R$. (Contrast with Exercise 2.71.) «

In more than one variable, polynomial equations $f(x_1, \ldots, x_n) = 0$ can have infinitely many solutions, so certainly an analogue of Theorem 2.16 for polynomials in several variables is false. The following is a weaker statement, however it entails Exercise 2.17 in several variables: if $\boldsymbol{x}$ is an $n$-tuple of variables and $R$ is an infinite integral domain, then distinct polynomials in $R[\boldsymbol{x}]$ define distinct functions from $R^n$ to $R$.

**Proposition 2.18** *Let $R$ be an infinite integral domain, and let $\boldsymbol{x}$ be an $n$-tuple of variables. For every nonzero polynomial $f \in R[\boldsymbol{x}]$ there is some tuple $\boldsymbol{b} \in R^n$ such that $f(\boldsymbol{b}) \neq 0$.*

***Proof*** We prove the proposition by induction on $n$. For $n = 1$ it follows immediately from Theorem 2.16, using the fact that $R$ is infinite. Let $n > 1$, and suppose that the proposition holds for $n - 1$. Let $f \in R[\boldsymbol{x}]$ be nonzero. Identifying $R[\boldsymbol{x}]$ with $R[x_1, \ldots, x_{n-1}][x_n]$ we write $f = f_0 + f_1 \cdot x_n + \cdots + f_d \cdot x_n^d$, where $f_0, f_1, \ldots, f_d \in R[x_1, \ldots, x_{n-1}]$ and $d = \deg_{x_n} f$ (so $f_d \neq 0$). By induction, there is some tuple $\boldsymbol{a} \in R^{n-1}$ such that $f_d(\boldsymbol{a}) \neq 0$.

The polynomial $f(\boldsymbol{a}, x_n) = f_0(\boldsymbol{a}) + f_1(\boldsymbol{a}) \cdot x_n + \cdots + f_d(\boldsymbol{a}) \cdot x_n^d$ in $R[x_n]$ is nonzero, as $f_d(\boldsymbol{a}) \neq 0$. By the case $n = 1$, there is some $b \in R$ such that $f(\boldsymbol{a}, b) \neq 0$, which completes the proof. □

An inspection of the proof of Proposition 2.18 shows that as the set of potential inputs, we do not need to take all of $R$. The same proof shows:

**Proposition 2.19** *Let $R$ be an integral domain, and let $S \subseteq R$ be infinite. Let $\boldsymbol{x}$ be an $n$-tuple of variables. For every nonzero polynomial $f \in R[\boldsymbol{x}]$ there is some tuple $\boldsymbol{b} \in S^n$ such that $f(\boldsymbol{b}) \neq 0$.* □

### 2.2.2 Unique Factorisation Domains

A nonzero element $a$ of an integral domain $R$ is *irreducible* if it is not a unit, but it is not the product of two non-units. That is, if $a = bc$, then at least one of $b$ or $c$ are units. In other words, $a$ is irreducible if whenever $b \mid a$ then $b$ is a unit or $b \sim a$. Irreducibility is invariant under association: if $a \sim b$ then $a$ is irreducible if and only if $b$ is.[4] In $\mathbb{Z}$ the irreducible elements are the prime numbers $p$ and their additive inverses $-p$. Non-unit elements which are not irreducible are called *reducible*.

*Exercise 2.20* Let $F$ be a field. Show that every linear polynomial $f \in F[\boldsymbol{x}]$ is irreducible. On the other hand, show that $2x$ is reducible in $\mathbb{Z}[x]$. «

*Exercise 2.21* Let $F$ be a field. The *order* of a formal power series $f = \sum a_n x^n$, denoted by $\operatorname{ord}(f)$, is the least $k$ such that $a_k \neq 0$; in other words, it is the greatest $k$ such that $x^k$ divides $f$. We let $\operatorname{ord}(0) = \infty$. By Exercise 2.13, $f$ is a unit of $F[\![x]\!]$ if and only if $\operatorname{ord}(f) = 0$.

(a) Show that for $f, g \in F[\![x]\!]$ we have $\operatorname{ord}(fg) = \operatorname{ord}(f) + \operatorname{ord}(g)$ and $\operatorname{ord}(f + g) \geqslant \min\{\operatorname{ord}(f), \operatorname{ord}(g)\}$, with equality whenever $\operatorname{ord}(f) \neq \operatorname{ord}(g)$.[5] (b) Show that two formal power series $f, g \in F[\![x]\!]$ are associates if and only if $\operatorname{ord}(f) = \operatorname{ord}(g)$. (c) Show that up to association, $x$ is the unique irreducible element of $F[\![x]\!]$.[6] «

As indicated in the overview chapter, we will be interested in *factoring* polynomials into irreducible components, and will want to ensure that this can be done in essentially only one way. This is analogous to the integers: the fundamental theorem of arithmetic says that every nonzero natural number is the product of prime numbers, and that this prime decomposition is unique. Unique in what sense? $12 = 2 \cdot 2 \cdot 3 = 2 \cdot 3 \cdot 2$, so there is not a unique *sequence* of prime numbers whose

[4] In the language of partial ordering, $a$ is irreducible if $[a]_\sim$ is minimal above $R^*$.

[5] We state that $\infty > n$ for all $n$ and that $\infty + n = \infty + \infty = \infty$ for all $n$.

[6] We say that $F[\![x]\!]$ is a *local* ring.

product is 12; the decomposition is unique up to permutation. In other words, there is a unique *multiset* of primes whose product is 12.

### Multisets

As mentioned in the introduction, a multiset is like a set, except that we allow elements to appear more than once. In contrast with sequences, the order of appearance doesn't matter. There is no agreed notation for multisets. In this book we will use square brackets: $[2, 2, 3]$ is the multiset in which 2 appears twice and 3 appears once. So $[2, 2, 3] = [2, 3, 2]$ but $[2, 2, 3] \neq [2, 3]$ (whereas for sets, $\{2, 2, 3\} = \{2, 3\}$). For a multiset $A$ and an element $a$ of $A$, we let $m_a(A)$, the *multiplicity of $a$ in $A$*, be the number of times $a$ appears in $A$. We will only allow finite multiplicities. $m_a(A) = 0$ means that $a$ does not appear in $A$. Every set is a multiset; in a set, all multiplicities are 1 (for elements) or 0 (for non-elements).

The *underlying set* $\lfloor A \rfloor$ of a multiset $A$ is the collection of elements that appear in $A$, with multiplicities forgotten. Formally, $\lfloor A \rfloor = \{a : m_a(A) > 0\}$.

The *size* of a multiset is the sum of the multiplicities of all of its elements. For example, the size of the multiset $[2, 2, 3]$ is 3. If $A$ and $B$ are multisets, then we say that $A$ is a *subset* of $B$, and write $A \subseteq B$, if for all $a$, $m_a(A) \leqslant m_a(B)$.[7] This definition agrees with the familiar notion when $A$ and $B$ are sets. If $A$ and $B$ are multisets and $A \subseteq B$ then $\lfloor A \rfloor \subseteq \lfloor B \rfloor$. Two multisets $A$ and $B$ are equal if and only if $A \subseteq B$ and $B \subseteq A$. If $A$ is a multiset and $B$ is a set, we say that $A$ is a multiset of elements of $B$ if $\lfloor A \rfloor \subseteq B$.

If $A$ and $B$ are multisets, then the *sum* (or *disjoint union*) of $A$ and $B$, denoted by $A + B$, is the multiset defined by letting $m_a(A + B) = m_a(A) + m_a(B)$ for all $a$. This can be generalised to taking the sum $\sum \mathcal{C}$ of a finite multiset $\mathcal{C}$ of multisets.

If $R$ is a ring and $A$ is a finite multiset of elements of $R$, then we can write $\sum A$ for the sum of all elements of $A$ and $\prod A$ for the product of all elements of $A$. For example in $\mathbb{Z}$, $\sum[2, 2, 3] = 7$ and $\prod[2, 2, 3] = 12$. This is well-defined because addition and multiplication are both associative and commutative. (In non-commutative rings, we cannot take products of multisets, only of sequences.) We let $\sum \varnothing = 0_R$ and $\prod \varnothing = 1_R$.

### Unique Factorisation

In $\mathbb{Z}$, we would like to call a multiset such as $[2, 2, 3]$ an irreducible factorisation of 12. Even though we are now using multisets, this is not quite unique: $[-2, 2, -3]$ is also an irreducible factorisation of 12. Thus, irreducible factorisations are unique, but only up to association.

Therefore, we define, for an integral domain $R$, an *irreducible factorisation* (or *decomposition*) of an element $a \in R$ to be a (finite) multiset $A$ of irreducible elements of $R$ satisfying $\prod A \sim a$ (rather than requiring $\prod A = a$). This makes the definition invariant under association. (It also ensures that units other than 1 have irreducible factorisations, namely the empty set.) Let $A$ and $B$ be finite multisets of

[7] We should really be saying "submultiset" (or perhaps "multisubset"?) but that's a bit awkward.

elements of $R$. We write $A =_\sim B$ if there is a bijection $f : A \to B$ such that for all $a \in A, a \sim f(a)$. In other words, if we can list the elements of $A$ as $[a_1, a_2, \ldots, a_n]$ and of $B$ as $[b_1, b_2, \ldots, b_n]$ such that $a_1 \sim b_1, a_2 \sim b_2, \ldots, a_n \sim b_n$. If $A =_\sim B$ then $\prod A \sim \prod B$. Irreducible decompositions in $\mathbb{Z}$ are unique in that if $A$ and $B$ are two irreducible factorisations of $a \in \mathbb{Z}$, then $A =_\sim B$. Another way to say this is: if $A$ and $B$ are finite multisets of irreducible elements of $\mathbb{Z}$, and $\prod A \sim \prod B$, then $A =_\sim B$.

**Definition 2.22** An integral domain $R$ is a *unique factorisation domain* if:

(i) Every nonzero $a \in R$ has an irreducible factorisation; and
(ii) If $A$ and $B$ are finite multisets of irreducible elements of $R$, and $\prod A \sim \prod B$, then $A =_\sim B$.

Thus the fundamental theorem of arithmetic is the statement that $\mathbb{Z}$ is a unique factorisation domain.

Just like equality of multisets up to association, we can define a subset relation up to association: if $A$ and $B$ are finite multisets of nonzero elements of $R$, then we write $A \subseteq_\sim B$ if there is an injective $f : A \to B$ such that for all $a \in A, a \sim f(a)$. That is, if we can arrange $A = [a_1, a_2, \ldots, a_m]$ and $B = [b_1, b_2, \ldots, b_n]$ (with $m \leqslant n$) so that $a_1 \sim b_1, a_2 \sim b_2, \ldots, a_m \sim b_m$. If $A \subseteq_\sim B$ then $\prod A$ divides $\prod B$.

*Exercise 2.23* If $R$ is a unique factorisation domain, $A$ and $B$ are finite multisets of irreducible elements of $R$, and $\prod A \mid \prod B$, then $A \subseteq_\sim B$. «

### 2.2.3 Unique Factorisation in Polynomial Rings

Our main goal is the following:

**Theorem 2.24** *Let $F$ be a field, and let $\boldsymbol{x}$ be a tuple of variables. The ring $F[\boldsymbol{x}]$ is a unique factorisation domain.*

#### One Variable

We first examine the case of a single variable.

**Proposition 2.25** *If $F$ is a field, then $F[x]$ is a unique factorisation domain.*

Let us sketch the proof, starting with existence. By induction on the degree of a polynomial $f \in F[x]$ we show that $f$ is the product of irreducible polynomials. If $f$ is irreducible then this is immediate. If not, then $f = gh$ where neither $g$ nor $h$ are units. Since all nonzero constants are units, we have $\deg g, \deg h > 0$. Since $\deg f = \deg g + \deg h$ we must have $\deg g, \deg h < \deg f$. By induction,

both $g$ and $h$ have irreducible decompositions; taking the disjoint union, we get an irreducible decomposition for $f$.

For uniqueness, we use the following lemma. In an integral domain $R$, an element $a$ is called *prime* if for all nonzero $b, c \in R$, if $a \mid bc$ then $a \mid b$ or $a \mid c$. A prime element is always irreducible.

**Lemma 2.26** *Let $R$ be an integral domain in which every irreducible element is prime. Then $R$ satisfies part (ii) of Definition 2.22 (the uniqueness part).*

***Proof*** We show that if $A$, $B$ are finite multisets of irreducible elements of $R$ and $\prod A \mid \prod B$, then $A \subseteq_\sim B$. This is done by induction on the size of $A$. Let $B = [b_1, \ldots, b_n]$. If $A$ has a unique element $a$, then $a \mid b_1 \cdots b_n$ implies $a \mid b_i$ for some $i$, as $a$ is prime. Since $b_i$ is irreducible and $a$ is not a unit, we have $a \sim b_i$. If $A$ has more than one element, write $A = A' + A''$ for two nonempty subsets of $A$. Since $\prod A' \mid \prod B$, by induction, there is some $B' \subseteq B$ such that $A' =_\sim B'$. Write $B = B' + B''$. After dividing by $\prod A'$, we see that $\prod A'' \mid \prod B''$; by induction again, we have $A'' \subseteq_\sim B''$. □

So to prove Proposition 2.25, it remains to show that every irreducible $g \in F[x]$ is prime. Let $g$ be irreducible, and suppose that $g \mid fh$ for some $f, h \in F[x]$. Since we can divide with remainder in $F[x]$, we apply the *Euclidean algorithm*. Starting with $f_0 = f$ and $f_1 = g$, divide $f_0$ by $f_1$ and obtain a quotient $q_1$ and remainder $f_2$. Next, divide $f_1$ by $f_2$ and obtain a quotient $q_2$ and remainder $f_3$. Keep going until we get $f_{n+1} = 0$. This must happen since $\deg f_1 > \deg f_2 > \cdots$. Now by reverse induction on $i \leqslant n$, using the equation $f_{i-1} = q_i f_i + f_{i+1}$, we see that $f_n \mid f_i$. Hence $f_n$ divides both $f_0 = f$ and $f_1 = g$. Since $g$ is irreducible, $f_n$ is a unit or $f_n \sim g$. In the latter case $g \mid f$ and we're done.

Suppose that $f_n$ is a unit. By induction on $i = 2, \ldots, n$, we see that $f_i$ is a *linear combination* of $f$ and $g$: there are $\alpha_i, \beta_i \in F[x]$ such that $f_i = \alpha_i f + \beta_i g$. Since $f_n$ is a unit, we can divide by it, and obtain $\alpha, \beta \in F[x]$ such that $\alpha f + \beta g = 1$. Then $h = h(\alpha f + \beta g) = \alpha fh + \beta gh$. Since $g \mid fh$ (and certainly $g \mid \beta gh$), $g$ divides $h$. This concludes the proof of Proposition 2.25.

*Exercise 2.27* Show that if $R$ is a unique factorisation domain, then every irreducible element of $R$ is prime. «

*Remark 2.28* Most textbooks take a wider approach. One usually defines the notion of a *Euclidean domain*, which is an integral domain in which one can divide with remainder (with respect to some *Euclidean norm*, a measure of size that tells us that the remainder is "smaller" than the element we divide by; in $\mathbb{Z}$ we can take the absolute value, in $F[x]$ we take the degree). One then shows that every Euclidean domain is a unique factorisation domain, often by using the intermediate notion of a *principal ideal domain*, integral domains in which every ideal is generated by a single element; every Euclidean domain is an principal ideal domain, and every principal ideal domain is a unique factorisation domain. The arguments are mostly

an abstraction of the argument above. One of the related notions is that of a *greatest common divisor*, showing that it exists and is a linear combination of the elements involved.

Unique factorisation is used well beyond polynomial rings; a standard example is subrings of the algebraic numbers, for example $\mathbb{Z}[i]$, $\mathbb{Z}[\sqrt{2}]$, $\mathbb{Z}[\sqrt{-2}]$ and others. For more, see [Rot00, Sect. 3.5] and [Art91, Chap. 11]. «

### Interlude: Algebraically Closed Fields

A field $F$ is *algebraically closed* if every nonconstant polynomial $f \in F[x]$ has a root in $F$. For example, the fields $\mathbb{Q}$ and $\mathbb{R}$ are not algebraically closed, because the nonconstant polynomial $x^2 + 1$ has no root in $\mathbb{Q}$ or in $\mathbb{R}$.

A field $F$ is algebraically closed if and only if the irreducible polynomials in $F[x]$ are precisely the linear ones. In one direction, if all polynomials of degree $\geqslant 2$ are reducible, then every nonconstant polynomial is the product of linear polynomials; each of these have roots. In the other direction, if $F$ is algebraically closed, $f \in F[x]$ and $\deg f \geqslant 2$, then as $f$ has a root $a$, we know that $x - a$ divides $f$ (Theorem 2.16) and so $f$ is reducible.

By counting degrees (Proposition 2.9), we see that if $F$ is algebraically closed, the size of the irreducible factorisation of a nonzero polynomial $f \in F[x]$ is precisely $\deg f$. Of course some linear factors may appear more than once. It makes sense to think about the collection of roots of a nonzero polynomial $f \in F[x]$ as a multiset: the *multiplicity* of the root $a$ is the multiplicity of the polynomial $x - a$ in the irreducible factorisation of $f$. In other words, the largest number $k$ such that $(x - a)^k$ divides $f$. For example, in $\mathbb{C}[x]$, the polynomial $x^3 - 2ix^2 - x$ is factored as $(x - i)^2 x$, and so the multiset of roots is $[i, i, 0]$. Thus, if we count with multiplicities, if $F$ is algebraically closed, then every nonzero $f \in F[x]$ has *precisely* $\deg f$ many roots. In Chap. 3 we generalise this idea to solutions of polynomial equations with more than one variable.

**Proposition 2.29** *Every algebraically closed field is infinite.*

***Proof*** Let $F$ be a finite field. Let

$$f = 1 + \prod_{a \in F} (x - a).$$

Then $f \in F[x]$ has no root in $F$. □

The following is known as the fundamental theorem of algebra.

**Theorem 2.30** *The field $\mathbb{C}$ of complex numbers is algebraically closed.*

We will give a proof of the fundamental theorem of algebra using analytical methods in Chap. 11, see p. 304.

### Gauss's Lemmas

We return to Theorem 2.24. This is proved by induction on the number of variables, using:

**Proposition 2.31** *If $R$ is a unique factorisation domain then so is $R[x]$.*

We give a sketch of the proof, as the argument appears in many texts; see, for example, [Rot00, Sect. 6.2], [Art91, Sect. 11.3], or [Sti94, Sect. 4.6] for $R = \mathbb{Z}$.

Fix a unique factorisation domain $R$. The first step is to decompose polynomials in $R[x]$ into a constant part and a *primitive* part. A polynomial $f \in R[x]$ is called *primitive* if every $a \in R$ dividing $f$ must be a unit. Note that $a \in R$ divides a polynomial $f$ if and only if it divides all of its coefficients.

**Lemma 2.32** *Every $g \in R[x]$ is the product $af$ where $a \in R$ and $f \in R[x]$ is primitive.*

The idea is to let $a$ be the greatest common divisor of the coefficients of $g$. That this exists follows from the fact that $R$ is a unique factorisation domain: we let $a$ be the product of all elements $c^k$ where $c \in R$ is irreducible and $k$ is greatest such that $c^k$ divides all nonzero coefficients of $g$. Note that $c \in R$ is irreducible in $R$ if and only if it is irreducible in $R[x]$.

We can then quickly dispense with existence: Every $g \in R[x]$ has an irreducible factorisation. We write $g = af$ where $f$ is primitive; since $R$ is a unique factorisation domain, $a$ is the product of irreducible elements of $R$ (and hence of $R[x]$). And since $f$ is primitive, every polynomial dividing $f$ is primitive as well, and if $f = hk$ where $h, k \in R[x]$ are not units, then $k, h \notin R$, i.e., $\deg k, \deg h > 0$, whence $\deg k, \deg h < \deg f$. So just as for $F[x]$, we can prove by induction on the degree of a primitive polynomial $f$ that it is the product of irreducible polynomials.

*Remark 2.33* The decomposition of a polynomial into a constant and a primitive part is unique up to association. Suppose that $a, b \in R$, $f, g \in R[x]$ are primitive, and $af \sim bg$. Then $a$ divides $bg$; so it divides every coefficient of $bg$. Since $g$ is primitive, $b$ is the greatest common divisor of the coefficients of $bg$; and so $a$ must divide $b$. By symmetry $a \sim b$, and so $f \sim g$. «

Uniqueness is more difficult. The argument relies on three lemmas, each named after Gauss. The main one is:

**Lemma 2.34**

(a) *The product of two primitive polynomials is primitive.*
(b) *Every irreducible $p \in R$ is prime in $R[x]$.*

***Sketch of proof*** The same argument gives both parts. Let $f, g \in R[x]$ and suppose that $fg$ is not primitive; since $R$ is a unique factorisation domain, there is some

irreducible $p \in R$ which divides $fg$; we show that $p$ divides $f$ or $p$ divides $g$. The "real proof" is to work in $(R/(p))[x]$, where $R/(p)$ is the quotient ring, which is an integral domain because $p$ is prime in $R$ (Exercise 2.27); so $(R/(p))[x]$ is an integral domain as well. We haven't defined quotient rings so we sketch a direct argument. Let $f = \sum a_i x^i$ and $g = \sum b_i x^i$. Suppose, for a contradiction, that $p$ divides neither $f$ nor $g$. Then there are $k \leqslant \deg f$ and $m \leqslant \deg g$ such that $p \nmid a_k$ and $p \nmid b_m$; choose minimal such. The coefficient $c_{k+m}$ of $x^{k+m}$ in $fg$ is the sum $\sum_{i+j=k+m} a_i b_j$; for any pair $(i, j) \neq (k, m)$ such that $i + j = k + m$, we have $i < k$ or $j < m$, whence $p$ divides $a_i b_j$; but since $p$ is prime in $R$, it does not divide $a_k b_m$, and so does not divide $c_{k+m}$, and so does not divide $fg$. □

The next two lemmas relate divisibility in $R[x]$ to divisibility in $F[x]$, where $F$ is the *field of fractions* of $R$. For any integral domain $R$, we can mimic the creation of the rationals $\mathbb{Q}$ from the integers $\mathbb{Z}$ to obtain a field $F \supseteq R$ which is minimal with respect to containing $R$; every element of $F$ is of the form $a/b$ for some $a, b \in R$. The idea is to start with all pairs $(a, b) \in R^2$ (with $b \neq 0$), and identify two pairs $(a, b)$ and $(c, d)$ if they should represent the same fraction: namely, if in $R$ we have $ad = bc$. This is an equivalence relation on the set of pairs $(a, b)$. We let $F$ be the collection of equivalence classes. On $F$ we define addition, subtraction and multiplication the way that fractions ought to behave: letting, temporarily, $[a, b]$ denote the equivalence class of the pair $(a, b)$, we define $[a, b] \cdot [c, d] = [ac, bd]$ and $[a, b] + [c, d] = [ad + bc, bd]$. We need to check that these operations are well-defined on equivalence classes (do not depend on the choice of pairs $(a, b)$ and $(c, d)$ in the classes); that these operations, together with declaring that $0_F = [0, 1]$ and $1_F = [1, 1]$, make $F$ into a field; and that $a \mapsto [a, 1]$ is an embedding of $R$ into $F$. These are technical but not difficult. Note that the fact that $R$ doesn't have zero divisors is used in the very definition of our operations: for example, letting $[a, b] \cdot [c, d] = [ac, bd]$ assumes that $bd \neq 0$. For more details, see for example [Rot00, Theorem 3.16].

*Example 2.35* The fraction field of $F[\boldsymbol{x}]$ is denoted by $F(\boldsymbol{x})$, the *field of formal rational functions* with coefficients in $F$. A formal rational function $f/g$ defines a partial function $F^n \to F$, which is defined on the points $\boldsymbol{a} \in F^n$ for which $g(\boldsymbol{a}) \neq 0$. «

*Exercise 2.36* A *formal Laurent series* with coefficients in a ring $R$ is an object of the form $a_m x^m + a_{m+1} x^{m+1} + \cdots$ where $m$ is an *integer*: that is, we allow negative exponents, but only finitely many of them. We let $R(\!(x)\!)$ denote the collection of formal Laurent series with coefficients from $R$. (a) Define addition and multiplication of formal Laurent series; show that with these operations, $R(\!(x)\!)$ is an integral domain, and that $R[\![x]\!]$ is a subring of $R(\!(x)\!)$. (b) Show that if $F$ is a field then $F(\!(x)\!)$ is a field, which is isomorphic to the field of fractions of $F[\![x]\!]$. (c) Define the order $\operatorname{ord}(f)$ of a formal Laurent series $f = \sum a_n x^n$ to be the least $k \in \mathbb{Z}$ such that $a_k \neq 0$. Show that Exercise 2.21(a) holds for $f, g \in F(\!(x)\!)$ as well. «

Armed with the field of fractions $F$ of our unique factorisation domain $R$, let $f, g \in R[x]$. If $g \mid f$ in $R[x]$, then certainly $g \mid f$ in $F[x]$, as $R[x] \subseteq F[x]$. The converse is false. For example, if $a \in R \setminus R^*$, then for any $f \in R[x]$, $af \nmid f$ in $R[x]$ but $af \mid f$ in $F[x]$, as $a \in F^*$. The failure is because $af$ is not primitive.

**Lemma 2.37** *Let $f, g \in R[x]$, and suppose that $f$ is primitive. If $f \mid g$ in $F[x]$ then $f \mid g$ in $R[x]$.*

***Sketch of proof*** Write $g = fh$ with $h \in F[x]$. The coefficients of $h$ are fractions of elements of $R$; by clearing denominators, we find $a \in R$ such that $ah \in R[x]$. So $ahf = ag$. Write $ah = b\bar{h}$ with $b \in R$ and $\bar{h} \in R[x]$ primitive. Similarly write $ag = c\bar{g}$ with $\bar{g}$ primitive. So $b(\bar{h}f) = c\bar{g}$. By Lemma 2.34, $\bar{h}f$ is primitive. By Remark 2.33, $\bar{h}f \sim \bar{g}$. □

**Lemma 2.38** *If $f \in R[x]$ is nonconstant and irreducible in $R[x]$, then it is irreducible in $F[x]$.*

***Sketch of proof*** Say $f$ is reducible in $F[x]$. Let $g$ be a proper divisor of $f$ in $F[x]$; since $F$ is a field, $0 < \deg g < \deg f$. As above, there is some primitive $\bar{g} \in R[x]$ such that $g \sim \bar{g}$ in $F[x]$. Then $\bar{g} \mid f$ in $F[x]$ and so in $R[x]$ as well; since $\deg \bar{g} = \deg g < f$, this division is proper, so $f$ is reducible in $R[x]$. □

***Proof of Proposition 2.31*** It remains to prove uniqueness of factorisations; by Lemma 2.26, it suffices to show that every irreducible $f \in R[x]$ is prime in $R[x]$ as well. Part (b) of Lemma 2.34 takes care of the constants. Suppose that $f$ is nonconstant and irreducible in $R[x]$. Then it is primitive. Suppose that $f \mid gh$ in $R[x]$. Since $f$ is irreducible in $F[x]$ (Lemma 2.38) and $F[x]$ is a unique factorisation domain (Proposition 2.25), $f$ is prime in $F[x]$ (Exercise 2.27), so $f$ divides either $g$ or $h$ in $F[x]$. Since $f$ is primitive, it divides either $g$ or $h$ in $R[x]$ as well (Lemma 2.37). □

## 2.3 Groups

There is nothing nonstandard about how we would present groups, and so we only give a brief survey. The material appears in hundreds of texts; see for example [Rot00, Chap. 2], [Sti94, Chap. 7], and [Art91, Chap. 2].

### 2.3.1 The Category of Groups

The initial development of the theory of groups is similar to that of rings. Fundamentally, though, these are different kinds of objects, groups morally being collections of symmetries or operations, whereas rings are generalisations of number systems.

A *group* is a set $G$, equipped with a binary operation $\cdot_G$, a unary operation $a \mapsto (a)_G^{-1}$ and a designated element $1_G$, which satisfy the following properties:

**Associativity**: for all $a, b$ and $c$ in $G$, $a \cdot_G (b \cdot_G c) = (a \cdot_G b) \cdot_G c$.
**Identity element**: for all $a \in G$, $a \cdot_G 1_G = 1_G \cdot_G a = a$.
**Inverses**: for all $a \in G$, $a \cdot_G (a)_G^{-1} = (a)_G^{-1} \cdot_G a = 1_G$.

As is immediately checked from the definitions, if $(R; +_R, \cdot_R, -_R, 0_R, 1_R)$ is a ring, then $(R; +_R, -_R, 0_R)$ is a group (this is the *additive group* of $R$). Hence, for example, $(\mathbb{Z}; +, -, 0)$, $(\mathbb{Q}; +, -, 0)$, $(\mathbb{R}; +, -, 0)$, $(\mathbb{C}; +, -, 0)$ are groups. On the other hand retaining multiplication rather than addition does not result in a group; $0_R$ never has a multiplicative inverse. On the other hand, $(R^*; \cdot_R, 1_R)$ is a group, the *multiplicative group of units*; the point is that the multiplicative inverse of a unit is a unit. So for example $\mathbb{Z}^* = \{1, -1\}$ with multiplication is a group, and so is $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$.

All the groups mentioned so far have the property that their binary operation $\cdot_G$ is commutative as well, a property that is not required in the definition of a group. Groups with this property are called *abelian*. We briefly mention the standard example of a non-abelian group.

*Example 2.39* Recall that a *permutation* of a set $X$ is a bijection (one-to-one and onto function) from $X$ to itself. The collection of all permutations of $X$ is denoted by $S_X$. Equipped with function composition this is a group (in particular, the composition of two permutations is a permutation). The identity element is the identity function $\mathrm{id}_X$. We write $S_n$ for $S_{\{1,2,\ldots,n\}}$. «

*Notation 2.40* We use notational conventions similar to the ones we used for integral domains. We often drop the subscript $G$ and write $a \cdot b$ or even just $ab$ instead of $a \cdot_G b$; we write 1 instead of $1_G$, and $a^{-1}$ instead of $(a)_G^{-1}$. We write $G$ instead of $(G; \cdot_G, (\,)_G^{-1}, 1_G)$. In light of associativity, we write $abc$ instead of $a(bc)$ or $(ab)c$.

We write $a^n$ for $\overbrace{a \cdot a \cdots a}^{n \text{ times}}$ and also let $a^{-n} = (a^{-1})^n$.

We will use *additive* notation for abelian groups: we write $+$ for $\cdot_G$, $-a$ for $a^{-1}$, 0 for $1_G$ and $na$ for $a^n$. «

As with rings (see Lemma 2.8), we use the group axioms to derive properties that are shared by all groups. Here is a short list.

**Lemma 2.41** *Let $G$ be a group and let $a, b \in G$. (a) If $ba = 1$ or $ab = 1$, then $b = a^{-1}$. (b) $(a^{-1})^{-1} = a$. (c) If $ab = ac$ then $b = c$; similarly if $ba = ca$. (d) $(ab)^{-1} = b^{-1}a^{-1}$.* [8] *(e) $a^2 = a$ if and only if $a = 1_G$.* □

[8] For an illustration for why $(ab)^{-1}$ is not $a^{-1}b^{-1}$, let $a$ stand for the operation "wind down the car window" and let $b$ denote the operation "stick your head out".

### Subgroups

A group $H$ is a *subgroup* of a group $G$ if $H \subseteq G$, $1_H = 1_G$ and for all $a, b \in H$, $a \cdot_H b = a \cdot_G b$ and $(a)_H^{-1} = (a)_G^{-1}$. Actually, the other requirements follow just from $a \cdot_H b = a \cdot_G b$. As with rings, closure under the operations is a necessary and sufficient condition for a subset of $G$ to be the domain of a subgroup of $G$: if $G$ is a group, $H \subseteq G$, $1_G \in H$, and for all $a, b \in H$, $a \cdot_G b \in H$ and $(a)_G^{-1} \in H$, then the restriction to $H$ of the operations of $G$ makes $H$ a subgroup of $G$. A subgroup of an abelian group is abelian.

*Example 2.42* The unit circle $S$ is a subgroup of $\mathbb{C}^*$. «

A concept which we encountered in passing in the context of rings (see e.g. Exercise 2.3) is that of a *generated* substructure. Let $G$ be a group, and let $A$ be a subset of $G$. There is a subgroup $H$ of $G$ with the following two properties: (i) $A \subseteq H$; (ii) if $K$ is a subgroup of $G$ and $A \subseteq K$, then $H \subseteq K$. In other words, $H$ is the $\subseteq$-least subgroup of $G$ which contains all the elements of $A$. The uniqueness of such $H$ follows immediately from its definition. To show such a subgroup always exists, we construct it either from above or from below. From above, we take the intersection of all subgroups of $G$ which are supersets of $H$, and show that this intersection is a subgroup. From below, we start with $A$ and "throw in" the necessary elements to construct a subgroup: the generated subgroup is the collection of finite products $a_1 a_2 \cdots a_n$ where each $a_i$ is either in $A$ or the inverse of an element of $A$ (the empty product gives us $1_G$). The subgroup of $G$ generated by $A$ is denoted by $\langle A \rangle_G$.

For an example see Example 2.55 below.

### Group Homomorphisms

A *group homomorphism* from a group $G$ to a group $H$ is a function $\psi : G \to H$ such that $\psi(1_G) = 1_H$ and for all $a, b \in G$, $\psi(a \cdot_G b) = \psi(a) \cdot_H \psi(b)$ and $\psi((a)_G^{-1}) = (\psi(a))_H^{-1}$. When it is clear if we are using rings or groups we just write "homomorphism". As for subgroups, the other conditions follow from $\psi(a \cdot_G b) = \psi(a) \cdot_H \psi(b)$, so this is the definition you usually see.

As with ring homomorphisms, a composition of two group homomorphisms is a group homomorphisms. If $\psi : G \to H$ is a group homomorphism then the image of $\psi$ is a subgroup of $H$ and the preimage $\psi^{-1}[K]$ of a subgroup $K$ of $H$ is a subgroup of $G$.

*Example 2.43* The example from the introduction: let $f : \mathbb{R} \to \mathbb{C}$ by letting $f(t) = e^{it}$. Then $f(0) = 1$ and for all $s, t \in \mathbb{R}$, $f(t + s) = f(t) f(s)$. Thus $f$ is a homomorphism from the *additive* group of the reals $(\mathbb{R}; +, 0)$ to the *multiplicative* group of non-zero complex numbers $(\mathbb{C}^*; \cdot, 1)$. The range of $f$ is the unit circle $S$.«

*Example 2.44* Let $G$ be a group and let $a \in G$. The map $n \mapsto a^n$ is a group homomorphism from the integers $(\mathbb{Z}; +)$ to $G$ (the proof is actually not that short—try it). The range of this homomorphism is the subgroup of $G$ generated by $a$. «

*Example 2.45* If $G$ and $H$ are groups, then $G \times H$ is a group, with pointwise operations: $(a, b) \cdot (c, d) = (a \cdot_G c, b \cdot_G d)$ and similarly for the identity and inverses. This is the *direct product* of $G$ and $H$. «

The *kernel* of a group homomorphism $\psi : G \to H$ is

$$\ker \psi = \{a \in G \,:\, \psi(a) = 1_H\}.$$

The kernel of $\psi$ is a subgroup of $G$. For example, the kernel of the homomorphism of Example 2.43 is the subgroup $(2\pi\mathbb{Z}; +)$, which is a subgroup of $(\mathbb{R}; +)$. A homomorphism is one-to-one if and only if its kernel is trivial (the subgroup $\{1_G\}$). A one-to-one homomorphism is called an *embedding*.

A homomorphism is an *isomorphism* if it is one-to-one and onto. An embedding is an isomorphism between its domain and its range. Groups $G$ and $H$ are isomorphic if there is an isomorphism between them. We write $G \cong H$. Since the composition of isomorphisms is an isomorphism, the inverse of an isomorphism is an isomorphism, and the identity map is an isomorphism, being isomorphic is an equivalence relation on groups.

*Exercise 2.46* Let $\rho$ be an irrational complex number such that $\rho^2 \in \mathbb{Z}$, for example $\rho = i$ or $\rho = \sqrt{2}$. Let $G = (\mathbb{Z}[\rho]; +, -, 0)$ be the additive group of the ring $\mathbb{Z}[\rho]$ (Exercise 2.3). Show that $G$ is isomorphic to the group $\mathbb{Z} \times \mathbb{Z}$.[9] «

### 2.3.2 Quotient Groups

The quotient of a group $G$ by a subgroup $H$ is the group you get once you imagine that all the elements of $H$ are the same as the identity of $G$. An example is the construction of $\mathbb{Z}/(n)$, the integers modulo $n$. In some instances this cannot be done. The general theory involves the notion of a normal subgroup. All difficulties disappear if we assume that $G$ is abelian, and so here we only develop the theory of quotients of abelian groups.

Let $H$ be a subgroup of an abelian group $G$. We say that $a, b \in H$ are *equivalent modulo* $H$ if $a - b \in H$. This is an equivalence relation on $G$. The equivalence class of $a \in G$ is $a + H = \{a + h \,:\, h \in H\}$, called the $H$*-coset* of $a$. Thus, the cosets form a partition of $G$ and they all have the same size (the map $h \mapsto a + h$ is a bijection between $H$ and $a + H$). One consequence is *Lagrange's theorem*, which does not actually require the group $G$ to be abelian:

**Theorem 2.47** *If $G$ is a finite group and $H$ is a subgroup of $G$ then $|H|$ divides $|G|$.*

[9] Note that the ring $\mathbb{Z}[\rho]$ is *not* isomorphic to the ring $\mathbb{Z} \times \mathbb{Z}$.

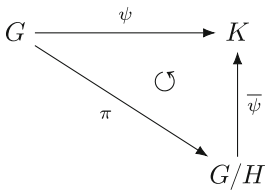



**Fig. 2.1** The diagram *commutes*: $\overline{\psi} \circ \pi = \psi$

Equivalence modulo $H$ is not only an equivalence relation, it is a *congruence* relation: if $a+H = b+H$ then $(-a)+H = (-b)+H$; if in addition, $c+H = d+H$ then $(a + c) + H = (b + d) + H$. Thus we can define addition and negation on the collection of cosets, which is denoted by $G/H$. These operations make $G/H$ a group.

The *quotient map* is the map $\pi\colon G \to G/H$ defined by $\pi(a) = a + H$. The kernel of this map is $H$. There is duality between quotients and surjective (onto) homomorphisms. Suppose that $\psi\colon G \to K$ is a homomorphism from an abelian group $G$ onto a group $K$. Let $H = \ker\psi$. Then $K$ is isomorphic to $G/H$; in fact, the map $\overline{\psi}$ mapping $a + H \mapsto \psi(a)$ is the unique isomorphism $\overline{\psi}\colon G/H \to K$ such that $\overline{\psi} \circ \pi = \psi$. See Fig. 2.1.

*Exercise 2.48* The homomorphism $t \mapsto e^{it}$ from $\mathbb{R}$ onto the unit circle $S$ (Example 2.43) shows that as a subgroup of $(\mathbb{C}^*; \cdot)$, the unit circle is isomorphic to the quotient $\mathbb{R}/2\pi\mathbb{Z}$. «

### 2.3.3 Cyclic Groups

A group $G$ is called *cyclic* if it is generated by a single element: $G = \langle a\rangle_G$ for some $a \in G$. The group $(\mathbb{Z}; +)$ is cyclic: $\mathbb{Z} = \langle 1\rangle_{\mathbb{Z}}$. It is the only infinite cyclic group (up to isomorphism). For each natural number $n \geqslant 1$, there is a unique cyclic group of size $n$, denoted by $C_n$. It can be realised as the quotient group $\mathbb{Z}/n\mathbb{Z}$, which is generated by the coset $1 + n\mathbb{Z}$. All cyclic groups are isomorphic to either $\mathbb{Z}$ or $C_n$ for some $n$; they are all abelian. Every subgroup of $\mathbb{Z}$ is of the form $n\mathbb{Z}$ for some $n$, thus, they are all cyclic.

*Example 2.49* For $n \geqslant 1$ let $\omega_n = e^{2\pi i/n}$. $(\omega_n)^n = 1$ but $(\omega_n)^k \neq 1$ for all positive $k < n$. Hence the subgroup of the unit circle $S$ generated by $\omega_n$ is isomorphic to $C_n$. The elements of this subgroup are precisely the complex numbers $\alpha$ satisfying $\alpha^n = 1$. (That there are no others follows from Theorem 2.16.) They are called the $n$th *roots of unity*. «

The *order* of an element $a$ of a group $G$ is the size of the (cyclic) subgroup $\langle a\rangle_G$ of $G$ generated by $a$. If the order $m = o_G(a)$ of $a$ is finite, then $\langle a\rangle_G = \{a^0, a^1, a^2, \ldots, a^{m-1}\}$; and for all $n \in \mathbb{Z}$, $a^n = 1_G$ if and only if $m$ divides $n$.

If $G$ is finite then the order of an element $a \in G$, which by definition is the size of a subgroup of $G$, must divide $|G|$ by Lagrange's theorem (Theorem 2.47). Thus:

**Proposition 2.50** *If $G$ is finite and $n = |G|$, then for all $a \in G$, $a^n = 1_G$.*

### The Characteristic of a Ring

We end with an application to integral domains. Let $R$ be a ring. If there is some $m$ such that

$$m1_R = \underbrace{1_R + 1_R + \cdots + 1_R}_{m \text{ times}} = 0_R$$

then the least such $m$ is called the *characteristic* of $R$, denoted by $\text{char}(R)$. If there is no such $m$ then we say that the characteristic of $R$ is 0. If the characteristic is positive, then it equals the order of $1_R$ in the additive group of $R$. If $R$ is an integral domain of positive characteristic, then its characteristic must be a prime number: if $\text{char}(R) = kn$ for $k, n \geqslant 2$, then $k1_R$ and $n1_R$ are nonzero in $R$, but their product is $0_R$. In particular, the characteristic of a field is either 0 or a prime number.

## 2.3.4 The Symmetric Group

There is a lot to say about the symmetric groups $S_n$ (Example 2.39). However, we will only really use them to develop the determinant function later in this chapter, so we review briefly. Fixing $n$, the *cycle* $(a_1\, a_2\, a_3 \ldots a_k)$ is the permutation $\sigma \in S_n$ which maps $a_1$ to $a_2$, $a_2$ to $a_3$, $\ldots$, $a_{k-1}$ to $a_k$ and $a_k$ back to $a_1$; all other elements of $\{1, 2, \ldots, n\}$ are mapped to themselves. A *transposition* is a cycle of length 2.

Every permutation $\sigma \in S_n$ is the product (composition) of transpositions. This is because $\sigma \in S_n$ is the product of disjoint cycles (every number is moved by at most one of the cycles), and every cycle is the product of transpositions. In fact:

*Exercise 2.51* Show that every transposition in $S_n$ (and thus every permutation in $S_n$) is the product of transpositions of the form $(k\ k+1)$ for some $k < n$. «

The set of disjoint cycles making up a permutation is unique, but there are many ways to write a permutation as the product of transpositions. However the *parity* of the number of transpositions is fixed: every permutation is the product of an even number of transpositions, or of an odd number of transpositions, but never both.

This uses the *sign homomorphism*. For a permutation $\sigma \in S_n$, let $C(\sigma)$ be the number of pairs $(i, j)$ with $i < j$ but $\sigma(i) > \sigma(j)$. The sign $\text{sgn}(\sigma)$ is $(-1)^{C(\sigma)}$. That is, $\text{sgn}(\sigma) = 1$ if $\sigma$ reverses the ordering of an even number of pairs, and $-1$ otherwise. The sign of a transposition $(k\ k+1)$ is $-1$; by counting, in fact, we can show that the sign of any transposition is $-1$.

The sign is a group homomorphism from $S_n$ to $\mathbb{Z}^* = \{1, -1\}$ (equipped with multiplication). That is, for $\sigma, \tau \in S_n$, $\operatorname{sgn}(\sigma\tau) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\tau)$. One way to see this is the following exercise.

*Exercise 2.52* For $f \in \mathbb{Z}[x_1, \ldots, x_n]$ and $\sigma \in S_n$ let $\sigma f = f\left(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)}\right)$. (a) Show that for $\sigma, \tau \in S_n$, $(\sigma\tau)f = \sigma(\tau f)$. (b) Show that $f \mapsto \sigma f$ is a permutation of $\mathbb{Z}[\boldsymbol{x}]$, and that it is a ring homomorphism (Proposition 2.11, with $S = \mathbb{Z}$ and $R = \mathbb{Z}[\boldsymbol{x}]$). (c) Let $P = \prod_{i<j\leqslant n}(x_j - x_i)$. Show that for every $\sigma \in S_n$, $\sigma P = \operatorname{sgn}(\sigma) \cdot P$. (d) Show that $\sigma \mapsto \operatorname{sgn}\sigma$ is a group homomorphism from $S_n$ to $\mathbb{Z}^*$. «

It follows that for all $\sigma \in S_n$, $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\sigma^{-1})$.

## 2.4 Linear Algebra Over Integral Domains

We quickly review familiar elements of linear algebra: matrix multiplication, linear maps, independence, and so on. The twist is that we will sometimes need to work over an integral domain which is not necessarily a field, so we will need to keep track of where we use invertibility. For example, we can have nonsingular matrices which are not invertible.

We will only need spaces of the form $R^n$ and their subspaces, so we do not give an axiomatic definition of an abstract vector space.

*Remark 2.53* The analogue of vector spaces over integral domains is a *module*. Every finite-dimensional vector space over a field is isomorphic to $F^n$ for some $n$, but if $R$ is not a field then there will be many finitely-generated modules not isomorphic to any $R^n$. For example, the $\mathbb{Z}$-modules are precisely the abelian groups. As we will only use $R^n$ and its subspaces, we will use our own terminology: linear spaces. «

### 2.4.1 Matrices, Linear Spaces, and Linear Maps

Let $R$ be an integral domain. We let $M_{n\times m}(R)$ be the collection of $n \times m$-matrices ($n$ rows, $m$ columns) with entries from $R$. If $n = m$ we write $M_n(R)$. We let $R_n = M_{1\times n}(R)$ be the collection of rows of length $n$, and $R^n = M_{n\times 1}(R)$ be the collection of columns of height $n$. $R$ is identified with $R^1$ and with $R_1$.

We fix some notation: the entry of a matrix $A$ at the $i$th row and $j$th column is usually denoted by $a_{i,j}$. We denote rows by $\underline{u}$ and columns by $\bar{v}$. The $i$th row of a matrix $A$ is denoted by $\underline{a_i}$, the $j$th column by $\bar{a}_j$. A row whose entries are all $0_R$ is denoted by $\underline{0}$, similarly for columns. We let $\bar{e}_i$ and $\underline{e_i}$ denote the unit columns and rows (with $1_R$ at position $i$, and $0_R$ elsewhere).

Operations on matrices: addition, multiplication, multiplication by a scalar—follow standard definitions, but using the operations of $R$. For example for a row $\underline{v}$ and column $\bar{u}$ both of the same length $n$, we let $\underline{v}\bar{u} = \underline{v} \cdot \bar{u} = (v_1 \cdot_R u_1) +_R (v_2 \cdot_R u_2) +_R \cdots +_R (v_n \cdot_R u_n)$. Each $M_{n\times m}(R)$ is an abelian group using matrix addition. Matrix multiplication is associative (this is a short calculation); with addition and multiplication of matrices, $M_n(R)$ is a non-commutative ring. The multiplicative identity is of course the matrix $I_n$ which has $1_R$ along the main diagonal and $0_R$ elsewhere.

*Remark 2.54* In a non-commutative ring, a multiplicative inverse of an element $a$ is an element $b$ satisfying $ab = ba = 1$. Like commutative rings, the collection of units together with multiplication is a group (but of course it may be non-abelian). In the ring of matrices $M_n(R)$, unit elements are called *invertible matrices*. The group of invertible matrices is called the *general linear group*, and is denoted by $\mathrm{GL}_n(R)$. «

*Example 2.55* We name three elements of $\mathrm{GL}_2(\mathbb{C})$ (Remark 2.54): $\mathbf{i} = \left(\begin{smallmatrix} i & 0 \\ 0 & -i \end{smallmatrix}\right)$, $\mathbf{j} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, and $\mathbf{k} = \left(\begin{smallmatrix} 0 & i \\ i & 0 \end{smallmatrix}\right)$. Let $Q = \langle \mathbf{i}, \mathbf{j}, \mathbf{k} \rangle_{\mathrm{GL}_2(\mathbb{C})}$ be the subgroup of $\mathrm{GL}_2(\mathbb{C})$ which is generated by the elements $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$. This group is called the *quaternion group*. By listing the elements of $Q$, show that it contains eight elements. Show that $Q$ is non-abelian. «

## Linear Spaces

Here is our non-standard definition:

**Definition 2.56** A *linear $R$-space* is a subgroup $U$ of some $R^n$ which is closed under scalar multiplication: for all $\lambda \in R$ and $\bar{u} \in U$, $\lambda\bar{u} \in U$. A linear $R$-space $W$ is a *subspace* of a linear $R$-space $U$ if $W \subseteq U$.

For a set $\mathsf{a} \subseteq U$ we write $\langle \mathsf{a} \rangle$ for the *span* of $\mathsf{a}$, or the *linear subspace generated by* $\mathsf{a}$, to be the collection of *linear combinations* of elements of $\mathsf{a}$, i.e., all columns of the form $\sum_{i\leqslant k} \lambda_i \bar{u}_i$ where $\bar{u}_1, \ldots, \bar{u}_k \in \mathsf{a}$ and $\lambda_1, \ldots, \lambda_k \in R$. Like generated subgroups, this is the $\subseteq$-least linear space which contains $\mathsf{a}$ as a subset. We agree that $\bar{0}$ is the result of adding up the empty set of columns, so the trivial subspace $\{\bar{0}\}$ is the span of the empty set.

*Example 2.57* For a nonzero $\bar{a} \in \mathbb{R}^n$, the space $\langle \bar{a} \rangle$ is the line in $\mathbb{R}^n$ passing through $\bar{a}$ and the origin $\bar{0}$. «

**Linear Maps**

Let $U$ and $V$ be linear $R$-spaces. A function $T: U \to V$ is *linear* if it preserves addition and scalar multiplication: $T(\bar{u}+\bar{v}) = T(\bar{u})+T(\bar{v})$ and $T(\lambda\bar{u}) = \lambda T(\bar{u})$ for all $\bar{u}, \bar{v} \in U$ and $\lambda \in R$. Equivalently, $T$ is linear if it preserves linear combinations: $T(\sum \lambda_i \bar{u}_i) = \sum \lambda_i T(\bar{u}_i)$. Preservation of addition means that a linear map is a group homomorphism from the subgroup $U$ of $R^n$ to the subgroup $V$ of $R^m$. The kernel of a linear map $T: U \to V$ (which is the collection of all $\bar{u}$ which $T$ maps to $\bar{0}$) is a linear subspace of $U$.

For a matrix $A \in M_{n\times m}(R)$ we define $T_A: R^m \to R^n$ by letting $T_A(\bar{u}) = A\bar{u}$. The map $T_A$ is linear, and every linear map from $R^m$ to $R^n$ is $T_A$ for some matrix $A$; distinct matrices give rise to distinct linear maps. This is because linear maps are determined by their values on the unit columns $\bar{e}_i$. Namely, $T_A(\bar{e}_i) = \bar{a}_i$ (the $i$th column of $A$). The range of $T_A$ is the subspace of $R^n$ spanned by the columns of $A$: a linear combination $\sum_i \lambda_i \bar{a}_i$ of the columns of $A$ is precisely the image $T_A(\bar{\lambda}) = A\bar{\lambda}$ of the column of scalars $\bar{\lambda} \in R^m$.

*Example 2.58* Let $\theta$ be an angle, and let $A_\theta = \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right) \in M_2(\mathbb{R})$. The map $T_{A_\theta}: \mathbb{R}^2 \to \mathbb{R}^2$ is the counter-clockwise rotation of the plane by $\theta$ radians. «

**Invertible and Nonsingular Matrices**

Matrix multiplication corresponds to composition: $T_{AB} = T_A \circ T_B$; this follows from the associativity of matrix multiplication. $T_{I_n}$ is the identity map on $R^n$. As a result, we see that:

- A matrix $A \in M_n(R)$ is invertible if and only if $T_A$ is 1–1 and onto, that is, if and only if $T_A$ is a linear *automorphism* of $R^n$.

On the other hand, we define:

- A matrix $A \in M_n(R)$ is *nonsingular* if $T_A$ is 1–1.

Thus, every invertible matrix is nonsingular. These notions are not equivalent: for example, when $R = \mathbb{Z}$, the $1 \times 1$ matrix (2) defines the map $n \mapsto 2n$ from $\mathbb{Z}$ to $\mathbb{Z}$, which is 1–1 but not onto. However, if $R$ is a field, then a matrix is nonsingular if and only if it is invertible. This can be seen by analysing the process of Gauss-Jordan elimination; for details, see, for example, [Art91, Sect. 1.2] or [Str76, Sect. 1.6]. This process, in general, only works over fields, because it involves dividing rows by scalars to make certain entries 1.

Gaussian elimination shows that over a field, a matrix is either singular, because it can be transformed by row operations to a matrix with a zero row; or it is the product of *elementary* matrices (the results of applying row operations to the identity matrix), and each of these is invertible. The analysis, by the way, shows that if $AB = I_n$ then $A$ is invertible and $B = A^{-1}$; that is, there is no need to verify that $BA = I_n$ as well.

Note that since each linear map is a group homomorphism, a matrix $A$ is nonsingular if and only if the kernel of $T_A$ is the trivial space $\{\bar{0}\}$.

### 2.4.2 Dimension and Complements

The material here is standard, as we restrict ourselves to working over a field $F$. See [Art91, Sects. 3.3 and 4.2] or [Str76, Sect. 2.3].

Recall that a collection of columns $\mathsf{a} = \{\bar{a}_1, \ldots \bar{a}_m\}$ is *linearly independent* if the only way to obtain $\bar{0}$ as a linear combination $\sum \lambda_i \bar{a}_i$ is by the trivial linear combination $\lambda_1 = \lambda_2 = \cdots = \lambda_m = 0$. As we noted, linear combinations of a collection of the columns $\bar{a}_1, \ldots \bar{a}_m$ are nothing other than elements of the range of $T_A$, where $A = (\bar{a}_1, \ldots, \bar{a}_m)$ is the matrix whose columns are the elements of $\mathsf{a}$. Hence, $\mathsf{a} = \{\bar{a}_1, \ldots \bar{a}_m\}$ is linearly independent if and only if the kernel of $T_A$ is trivial, if and only if $A$ is nonsingular. That is, if and only if every element of the span $\langle \mathsf{a} \rangle \subseteq F^n$ can be *uniquely* written as a linear combination of the columns in $\mathsf{a}$. The size of a linearly independent subset of $F^n$ is at most $n$. A subset of a linearly independent set is linearly independent.

A *basis* of a linear $F$-space $U$ is a linearly independent set $\mathsf{a} \subseteq U$ which spans $U$. All bases of a linear $F$-space have the same size.

**Proposition 2.59** *Let $U$ be a linear $F$-space and let $\mathsf{a} \subset U$ be linearly independent. There is some basis of $U$ which contains $\mathsf{a}$. In particular (since the empty set is linearly independent), every linear $F$-space has a basis.*

We can thus define the *dimension* of $U$ to be the size of any basis of $U$; we write $\dim U$. The dimension of $F^n$ is $n$. If $U$ is a proper subspace of $W$ then $\dim U < \dim W$, since a basis for $U$ is not a basis for $W$ but can be extended to a basis for $W$. Two linear spaces over $F$ of the same dimension are isomorphic. In fact:

**Proposition 2.60** *Let $\{\bar{a}_1, \ldots, \bar{a}_m\}$ be a basis of a linear space $U \subseteq F^n$. Let $\bar{b}_1, \ldots, \bar{b}_m$ be a sequence of columns in some $F^k$ (not necessarily distinct). Then there is a unique linear map $T : U \to F^k$ such that $T(\bar{a}_i) = \bar{b}_i$ for all $i \leqslant m$.*

Note that Proposition 2.59 fails when we don't work over fields; $\{2\} \subset \mathbb{Z}$ is a linearly independent subset of $\mathbb{Z}$, but does not extend to a basis.

Let $U$ and $V$ be subspaces of a linear $F$-space $W$. The subspace generated by $U \cup V$ is $U + V = \{u + v : u \in U \ \& \ v \in V\}$. If $U \cap V = \{\bar{0}\}$ then $\dim(U + V) = \dim U + \dim V$. We conclude:

**Proposition 2.61** *If $U$ and $V$ are subspaces of a linear $F$-space $W$, and $\dim U + \dim V > \dim W$, then $U \cap V \neq \{\bar{0}\}$.*

If $U + V = W$ and $U \cap V = \{\bar{0}\}$ then we write $W = U \oplus V$ and we say that $V$ is a *linear complement* of $U$ in $W$ (and vice-versa). Proposition 2.59 implies that every $U \subseteq W$ has a complement in $W$.

Suppose that $T : U \to V$ is linear, and let $Q \subseteq U$ be a linear complement of the kernel $\ker T$ of $T$. The restriction $T\restriction_Q$ of $T$ to $Q$ is 1–1 (as $Q \cap \ker T = \{\bar{0}\}$). The range $T[Q]$ of this restriction is the range of $T$; so if $T$ is onto $V$ then $T\restriction_Q$ is

onto $V$ as well, in which case $\dim Q = \dim V$. We obtain the *dimension formula* for linear maps:

**Theorem 2.62** *Let $U$ and $V$ be linear $F$-spaces, and let $T: U \to V$ be a linear map which is onto $V$. Then* $\dim U = \dim V + \dim(\ker T)$.

**Corollary 2.63** *Let $U$ and $V$ be a linear $F$-spaces. The following are equivalent for $W \subseteq U$:*

*(1) $W = \ker T$ for some linear $T: U \to V$ which is onto $V$;*
*(2) $W$ is a linear subspace of $U$ and* $\dim W = \dim U - \dim V$.

***Proof*** (1)⇒(2) follows from Theorem 2.62. For (2)⇒(1), by Proposition 2.59, let $Q$ be a linear complement of $W$ in $U$. Since $\dim Q = \dim V$, by Proposition 2.60 we can define a linear map $T$ from $U$ to $V$ which restricts to an isomorphism from $Q$ to $V$ and maps all of $W$ to 0; then $W = \ker T$. □

### 2.4.3 The Determinant

Back to working over an integral domain $R$ which may fail to be a field, the development of the determinant is fairly standard; see for example [Art91, Sects. 1.3 and 1.5] or [Str76, Chap. 4]. However, we cannot apply arguments which rely on Gauss-Jordan elimination. We give a brief sketch.

Define a *volume function* to be a function $V$ defined on $M_n(R)$ satisfying the following: (a) if $A$,$B$ and $C$ have the same columns except for the $i$th one, and $\bar{c}_i = \bar{a}_i + \bar{b}_i$ then $V(C) = V(A) + V(B)$; (b) if $D$ is obtained from $A$ by multiplying a column by a scalar $\lambda \in R$, then $V(D) = \lambda V(A)$; (c) If $A$ has two *adjacent* columns which are equal, then $V(A) = 0$. The idea is that $|V(A)|$ is the volume of the $n$-dimensional "parallelogram" determined by the columns of $A$; the sign depends on orientation. The first two properties together are the *multilinear* property of $V$.

We first show that every volume function is determined by $V(I_n)$. Let $V$ be a volume function. Observe that if $B$ is obtained from $A$ by exchanging two adjacent columns, then $V(B) = -V(A)$. As the collection of transpositions $(k\ k+1)$ (for $k < n$) generates all of $S_n$ (Exercise 2.51), and the sign of a transposition is $-1$, we conclude that for all $\sigma \in S_n$, if we permute the columns by $\sigma$, that is, replace $A = (\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_n)$ by $\sigma A = (\bar{a}_{\sigma(1)}, \bar{a}_{\sigma(2)}, \ldots, \bar{a}_{\sigma(n)})$, then we get $V(\sigma A) = \operatorname{sgn}(\sigma) V(A)$. Similarly, we observe that if $A$ has two identical columns then $V(A) = 0$.[10] We can now obtain the "closed form" of the determinant.

[10] If the characteristic of $R$ is 2, then this cannot be deduced from $V(A) = -V(A)$. For a general proof, we permute $A$ to get a matrix $\sigma A$ with two adjacent identical columns; then $V(A) = \operatorname{sgn}(\sigma) V(\sigma A) = \operatorname{sgn}(\sigma) \cdot 0 = 0$.

Repeatedly using multilinearity,

$$V(A) = \sum_{f:\{1,2,\ldots,n\}\to\{1,2,\ldots,n\}} \prod_{i\leqslant n} a_{f(i),i} \cdot V\left(\bar{e}_{f(1)}, \bar{e}_{f(2)}, \ldots, \bar{e}_{f(n)}\right)$$

(for the first step, note that $\bar{a}_1 = \sum a_{j,1}\bar{e}_j$, and so $V(A) = \sum a_{j,1}V(A_{j,1})$, where the first column of $A_{j,1}$ is $\bar{e}_j$ and the other columns are the same as $A$; now keep going, one column at a time.) If $f:\{1,2,\ldots,n\} \to \{1,2,\ldots,n\}$ is not a permutation then it is not one-to-one. In this case $V$ is applied to a matrix with two equal columns and so evaluates to 0. Hence we can discard the terms in the above sum contributed by functions which are not permutations. When $f = \sigma$ is a permutation then $V$ evaluates to $\operatorname{sgn}(\sigma) \cdot V(I_n)$, yielding

$$V(A) = V(I_n) \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i\leqslant n} a_{\sigma(i),i}.$$

The right hand side only depends on $V(I_n)$.

We can then prove existence: by induction on $n$, we prove the existence of a volume function $A \mapsto \det(A)$ on $M_n(R)$ satisfying $\det(I_n) = 1$. If this has already been done for $n-1$, then for any $i \leqslant n$, the column $i$ expansion:

$$A \mapsto (-1)^{1+i}a_{1,i}\det(A_{1,i}) + (-1)^{2+i}a_{2,i}\det(A_{2,i}) + \cdots + (-1)^{n+i}a_{n,i}\det(A_{n,i}),$$

where $A_{j,i}$ is the $(j,i)$-*minor* of $A$, obtained from $A$ by erasing the $j$th row and $i$th column—is a volume function on $M_n(R)$, mapping $I_n$ to 1. We can conclude that expansion along any column will yield the same result, which we call $\det(A)$. For any volume function $V$, $V(A) = V(I_n)\cdot\det(A)$.

Armed with these results, and with the closed form, we can derive familiar properties of the determinant.

**Proposition 2.64** *For all $A, B \in M_n(R)$, $\det(BA) = \det(B)\det(A)$.*

To see this, let $V(A) = \det(BA)$, and show that this is a volume function; it maps $I_n$ to $\det(B)$. Note that this gives us another geometric interpretation of the determinant: Since the columns of $BA$ are $T_B(\bar{a}_1), \ldots, T_B(\bar{a}_n)$, $\det(B)$ is the ratio by which $T_B$ changes the volumes of objects.

**Proposition 2.65** *For any $A \in M_n(R)$, $\det A = \det A^{\mathsf{t}}$.*

Here $A^{\mathsf{t}}$ is the *transpose*, obtained by reflecting across the main diagonal. To see this we can use the closed form, using the fact that $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\sigma^{-1})$ and that $\sigma \mapsto \sigma^{-1}$ is a permutation of $S_n$. Proposition 2.65 also implies that we can use row expansions to calculate determinants.

**Proposition 2.66** *If $A$ contains a zero row or column, then* $\det(A) = 0$.

For columns, it suffices to apply the second multilinearity condition and multiply the zero column by zero. For rows, take the transpose.

### The Effect of Row Operations

We can now deduce the well-known effects on the determinant by row operations. We have already observed that exchanging two columns multiplies the determinant by $-1$; by taking the transpose, this applies to rows as well. We argue similarly for multiplying rows by a scalar. Finally, for the operation of adding the multiple of one row to the other, we again use the multilinearity of the determinant: If $C$ is obtained from $A$ by adding the multiple of one row to another, then $C = A + B$, where $B$ has the same rows as $A$, except for one, and one row of $B$ is the scalar multiple of another; so $\det(B) = 0$ and $\det(C) = \det(A) + \det(B) = \det(A)$.

### Polynomial Substitution

Using the closed form (or by induction), the determinant is a polynomial in the entries of a matrix: there is a polynomial $f \in R[x_{1,1}, x_{1,2}, \ldots, x_{n,n}]$ such that $\det A = f(a_{1,1}, a_{1,2}, \ldots, a_{n,n})$.

Let $\boldsymbol{y} = (y_1, \ldots, y_m)$ be a tuple of variables, and let $G \in M_n(R[\boldsymbol{y}])$ be a square matrix whose entries $g_{i,j}$ are polynomials in $R[\boldsymbol{y}]$; $\det(G) \in R[\boldsymbol{y}]$. For $\boldsymbol{a} \in R^m$ we let $G(\boldsymbol{a})$ be the matrix in $M_n(R)$ whose entries are $g_{i,j}(\boldsymbol{a})$. The determinant commutes with substitution: $\det(G(\boldsymbol{a})) = (\det G)(\boldsymbol{a})$. To see this, let $f$ be the polynomial which defines the determinant (it is the same polynomial over $R$ and over $R[\boldsymbol{y}]$). Then $\det G = f(g_{1,1}, g_{1,2}, \ldots, g_{n,n})$ and $\det(G(\boldsymbol{a})) = f(g_{1,1}(\boldsymbol{a}), g_{1,2}(\boldsymbol{a}), \ldots, g_{n,n}(\boldsymbol{a}))$. The desired equality follows from the fact that polynomial substitution is translated to composition of the defined functions, see p. 24.

## 2.4.4 Detecting Singularity

Over a field, we know that $A \in M_n(F)$ is singular if and only if it is not invertible if and only if $\det(A) = 0$. This can be deduced by using Gaussian elimination: we first analyse how row operations affect the determinant; in particular, they never change 0 to nonzero or the other way. If $A$ is singular, then by row operations it can be transformed to a matrix with a zero row, and so $\det(A) = 0$. If not, then it is the product of elementary matrices, all of which have nonzero determinant; we then use the multiplicative property.

Over a general integral domain $R$ we cannot argue thus. But Cramer's method does work. Suppose that $A \in M_n(R)$ is invertible ($A \in \mathrm{GL}_n(R)$). Then $1 = \det(I_n) = \det(AA^{-1}) = \det A \cdot \det A^{-1}$. Thus, $\det A$ is a unit of $R$. The multiplicative property of the determinant shows that the restriction of the determinant function to $\mathrm{GL}_n(R)$ is a group homomorphism from $\mathrm{GL}_n(R)$ to $R^*$.

For $A \in M_n(R)$ let $c_{i,j} = (-1)^{i+j} \det A_{i,j}$, and let $\mathrm{Adj}(A) = C^{\mathsf{t}}$ (called the *adjugate of* $A$). The entry at $(i,i)$ of the matrix $A \cdot \mathrm{Adj}(A)$ is $\sum_{j \leqslant n} a_{i,j} c_{i,j}$ which equals $\det A$ by considering row expansion. For $k \neq i$, the entry of $A \cdot \mathrm{Adj}(A)$ at $(i,k)$ is $\sum_{j \leqslant n} a_{i,j} c_{k,j} = \sum_{j \leqslant n} (-1)^{j+k} a_{i,j} \det A_{k,j}$. This is the row expansion of the determinant which is obtained from $A$ by replacing the $k$th row by the $i$th row of $A$. Of course such a matrix has two equal rows, and so the value in this case is 0. That is, $A \cdot \mathrm{Adj}(A) = (\det A) \cdot I_n$. Similarly, by considering column expansions, we see that $\mathrm{Adj}(A) \cdot A = (\det A) I_n$. Thus, if $\det A$ is a unit of $R$, then $\frac{1}{\det A} \mathrm{Adj}(A) = A^{-1}$ and $A$ is invertible. That is:

**Proposition 2.67** *A matrix $A \in M_n(R)$ is invertible if and only if* $\det A \in R^*$.

If $R = F$ is a field then $F^* = F \setminus \{0\}$ and a square matrix is invertible if and only if it is nonsingular, so we derive the fact that a matrix $A \in M_n(F)$ is singular if and only if $\det A = 0$. By considering the field of fractions we can extend this fact to integral domains as well.

**Theorem 2.68** *Let $R$ be an integral domain. A matrix $A \in M_n(R)$ is singular if and only if* $\det A = 0$.

***Proof*** Let $F = \mathrm{Frac}(R)$ be the field of fractions of $R$ (see p. 34). Certainly $M_n(R) \subseteq M_n(F)$. The theorem follows from the fact that $A$ is singular as an element of $M_n(R)$ if and only if it is singular in $M_n(F)$. That is, $T_A \colon R^n \to R^n$ is injective if and only if the extension of this map to $T_A \colon F^n \to F^n$ is injective. In the nontrivial direction, suppose that $A$ is singular in $M_n(F)$; there is a nonzero column $\bar{b} \in F^n$ in the kernel of $T_A$. The entries of $\bar{b}$ are fractions of elements of $R$; multiplying by the denominators we get some nonzero $\alpha \in R$ such that $\alpha \cdot \bar{b} \in R^n$. Since $R$ is an integral domain, $\alpha \bar{b}$ is nonzero, and $T_A(\alpha \bar{b}) = \alpha T_A(\bar{b}) = 0$. □

Using the transpose, we see:

**Corollary 2.69** *A square matrix $A \in M_n(R)$ is singular if and only if there is a nonzero row $\underline{\alpha} \in R_n$ such that $\underline{\alpha} A = \underline{0}$.* □

## 2.5 Further Exercises

**Rings, Integral Domains, Polynomials**

2.70 Let $X$ be a nonempty set. Recall that $\mathcal{P}(X)$, the *power set* of $X$, is the collection of all subsets of $X$. For $A, B \subseteq X$, define $A \cdot B = A \cap B$ and $A + B = (A \cup B) \setminus (A \cap B)$. Show that $\mathcal{P}(X)$, together with the operations $+$ and $\cdot$ and the designated elements $\varnothing$ and $X$, is a ring.

2.71 Let $p$ be a prime number. Use the binomial formula to show that for all $a, b \in \mathbb{Z}/(p)$, $(a+b)^p = a^p + b^p$. Conclude that for all $a \in \mathbb{Z}/(p)$, $a^p = a$. Conclude that the two polynomials $x$ and $x^p$ (which are distinct polynomials in $(\mathbb{Z}/(p))[x]$) define the same function from $\mathbb{Z}/(p)$ to itself.

2.72 For $(a, b, c), (a', b', c') \in \mathbb{Z}^3$, let

- $(a, b, c) + (a', b', c') = (a + a', b + b', c + c')$, and $-(a, b, c) = (-a, -b, -c)$; and
- $(a, b, c) \cdot_* (a', b', c') = (aa' + bc' + cb', ab' + ba' + cc', ac' + bb' + ca')$.

(a) Show that $R = \left(\mathbb{Z}^3; +, -, \cdot_*, (0, 0, 0), (1, 0, 0)\right)$ is a ring. (b) Show that there are three distinct elements $\alpha \in \mathbb{Z}^3$ such that in $R$, $\alpha^3 = 1_R$. (c) Show that $R$ is not an integral domain.

2.73 Show that every finite integral domain is a field.

2.74 (a) Find a multiplicative inverse for $1 + x$ in $\mathbb{Z}[\![x]\!]$.
(b) Similarly for $1 + 2x + 3x^2 + 4x^3 + \cdots$ (Hint: the latter is the formal derivative of $1 + x + x^2 + x^3 + \cdots$.)

2.75 Let $R$ be the collection of polynomials $f \in \mathbb{Q}[x]$ whose constant coefficient is an integer (so polynomials such as $\frac{3}{2}x$ and $1 + \frac{5}{4}x^3 + \frac{17}{8}x^7$, but not $\frac{1}{2} + x$). For shorthand, we write $R = \mathbb{Z} + x\mathbb{Q}[x]$, as its elements are the polynomials of the form $n + xf$ for $f \in \mathbb{Q}[x]$. (a) Show that $R$ is a subring of $\mathbb{Q}[x]$. (b) Show that $R^* = \mathbb{Z}^* = \{1, -1\}$. (c) Show that every irreducible element of $R$ is either an irreducible element of $\mathbb{Z}$, or is of the form $1 + xf$ or $-1 + xf$ for some $f \in \mathbb{Q}[x]$. (d) Conclude that $x$ has no irreducible factorisation in $R$. (For a more extreme example, see Exercise 15.9.)

2.76 Let $R = \mathbb{Z}/(8)$. Find a polynomial $f \in R[x]$ of degree 2 which has more than two roots in $R$. (Hence, the condition that $R$ be an integral domain is necessary for Theorem 2.16.)

2.77 Let $R = \mathbb{Z}[\sqrt{-6}]$ (see Exercise 2.3). For $z = a + b\sqrt{-6} \in R$ let $N(z) = a^2 + 6b^2$. (a) Show that for all $w, z \in R$, $N(zw) = N(z)N(w)$. (b) Show that for all $z \in R^*$, $N(z) = 1$, and so that $R^* = \{1, -1\}$. (c) Show that for all $z \in R$, $N(z) \neq 2, 5$. (d) Show that $2, 5, 2 + \sqrt{-6}$ and $2 - \sqrt{-6}$ are irreducible in $R$, and that no two are associates. (e) Show that in $R$, $[2, 5]$ and $[2 + \sqrt{-6}, 2 - \sqrt{-6}]$ are two irreducible factorisations of 10 which are not associates. Hence $R$ is not a unique factorisation domain. (f) Nonetheless, using $N$, show that every nonzero $a \in R$ has some irreducible factorisation.

2.78 Let $A$ be a set of prime numbers. Let $\mathbb{Q}_{(A)}$ be the collection of all rational numbers of the form $a/b$, where $a, b$ are integers, $b \neq 0$, and for all $p \in A$, $p$ does not divide $b$.

Show that $\mathbb{Q}_{(A)}$ is a subring of $\mathbb{Q}$. Classify the units of $\mathbb{Q}_{(A)}$. Show that $\mathbb{Q}_{(A)}$ is a unique factorisation domain. Show that $\mathbb{Q}_{(\{2,3\})}$ has two association classes of irreducible elements.

**Groups**

2.79 Let $\alpha = e^{2\pi i\theta}$ be an element of the unit circle $S$. Show that the order of $\alpha$ in the group $S$ is finite if and only if $\theta$ is rational.

2.80 Let $R = \mathbb{Z}[\sqrt{2}]$ (see Exercise 2.3); let $G = R^*$. (a) Show that $1 + \sqrt{2}$ is a unit of $\mathbb{Z}[\sqrt{2}]$. (b) Conclude that $\mathbb{Z}[\sqrt{2}]$ contains infinitely many units. (c) Show that $1 + \sqrt{2}$ is the smallest element of $G$ greater than 1. (As in Exercise 2.77, let $N(a + b\sqrt{2}) = a^2 - 2b^2$; show that $N(zw) = N(z)N(w)$ for $z, w \in R$.) (d) Show that $G \cong \mathbb{Z} \times C_2$. (e) Use this to classify all solutions to Pell's equation $x^2 = 2y^2 + 1$.[11]

2.81 Which elements of $(\mathbb{Z} \times \mathbb{Z})/\langle(2, 4)\rangle_{\mathbb{Z}\times\mathbb{Z}}$ have finite order?

2.82 Show that if $\sigma_1, \sigma_2, \ldots, \sigma_k$ are pairwise disjoint cycles in $S_n$, with $\sigma_i$ of length $m_i$, then the order of $\sigma_1\sigma_2 \ldots \sigma_k$ is the least common multiple of $\{m_1, m_2, \ldots, m_k\}$.

Let $\sigma \in S_6$ be the permutation mapping 1 to 3, 2 to 1, 3 to 4, 4 to 5, 5 to 6, and 6 to 2. Calculate $\sigma^{100}$.

2.83 Let $V$ be the subgroup of $S_4$ generated by $\{(1, 2)(3, 4), (1, 3)(2, 4)\}$.[12] Show that $V$ is isomorphic to $C_2 \times C_2$.

**Linear Algebra**

2.84 For a square matrix $A \in M_n(R)$, we let the *trace* of $A$ be $\mathrm{tr}(A) = \sum_{i\leqslant n} a_{i,i}$, the sum of the elements on the main diagonal of $A$. Show that if $A \in M_{n\times m}(R)$ and $B \in M_{m\times n}(R)$, then $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. Use this to show that if $n \neq m$ then $R^n$ and $R^m$ are not linearly isomorphic.[13]

2.85 Let $U$ and $V$ be linear subspaces of $R^n$. Show that if $U \cup V$ is a linear subspace of $R^n$ then either $U \subseteq V$ or $V \subseteq U$.

[11] This can be used to find rational approximations of $\sqrt{2}$.

[12] $V$ is called the *Klein Viergruppe*.

[13] Dedicated to Ken Pledger.

2.86 Let $F$ be a finite field; let $q = |F|$. (a) Show that $F^2$ is the union of $q+1$ many 1-dimensional linear subspaces. (b) Show though that for any linear $F$-space $V$, for all $n \leqslant q$, if $V_1, V_2, \ldots, V_n$ are subspaces of $V$ such that $\bigcup_{j \leqslant n} V_j$ is also a subspace of $V$, then there is some $i \leqslant n$ such that $V_j \subseteq V_i$ for all $j \leqslant n$.

2.87 Let $n$ be an even number. Show that there is a linear map $T\colon F^n \to F^n$ such that for all $\underline{u} \in F^n$, $T(T(\underline{u})) = -\underline{u}$.

Is there a linear map $T\colon \mathbb{R} \to \mathbb{R}$ such that for all $u \in R$, $T(T(u)) = -u$? What about $\mathbb{C}$?

2.88 Suppose that the characteristic of $F$ is not 2. Let $U$ be a linear $F$-space. Show that the following are equivalent for a nonempty subset $H$ of $U$:

(1) there is some $\bar{u} \in U$ and some linear subspace $V$ of $U$ such that $H = \bar{u} + V = \{\bar{u} + \bar{v} : \bar{v} \in V\}$;
(2) there is some linear $F$-space $W$, some $\bar{w} \in W$, and a linear map $T\colon U \to W$ such that $H = T^{-1}\{\bar{w}\}$;
(3) for all $\bar{u}, \bar{v} \in H$ and all $c \in F$, $c\bar{u} + (1-c)\bar{v} \in H$.

A subset of a linear $F$-space $U$ satisfying these conditions is called an *affine subspace* of $U$. The *dimension* of the affine subspace $\bar{u} + V$ is defined to be $\dim V$. A 1-dimensional affine subspace is called a line. A 2-dimensional affine subspace is called a plane. Show that every two elements of $F^2$ lie on a line, and that any three elements of $F^3$ lie on a plane.

2.89 Let $U$ be a linear $F$-space. A linear map $T\colon U \to U$ is called a *projection* if $T \circ T = T$. Give an example of a projection $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ whose range is a 1-dimensional subspace of $\mathbb{R}^2$. Show that if $T$ is a projection, then the range of $T$ is a linear complement of $\ker T$ in $U$.

**Determinants**

2.90 Let $A, B \in M_n(R)$. Show that if $AB = I_n$ then $BA = I_n$.

2.91 Is

$$\det \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 7 & 8 & 11 & 2 & 1 \\ 0 & 3 & 1 & 4 & -1 \\ -2 & 5 & 7 & 7 & 1 \\ 0 & 1 & 0 & 2 & 4 \end{pmatrix} = \frac{3}{4}?$$

2.92 Let $f = a_0 + a_1 x + \cdots + a_d x^d$ be a polynomial in $R[x]$. Show that

$$\det \begin{pmatrix} a_0 & a_1 & a_2 & \dots & a_d \\ -x & 1 & & & \\ & -x & 1 & & \\ & & \ddots & \ddots & \\ & & & -x & 1 \end{pmatrix} = f.$$

2.93 Let $x_0, x_1, x_2, \dots, x_n$ be variables. The *Vandermonde* matrix of dimension $n+1$ is

$$A = \begin{pmatrix} 1 & x_0 & (x_0)^2 & \cdots & (x_0)^n \\ 1 & x_1 & (x_1)^2 & \cdots & (x_1)^n \\ 1 & x_2 & (x_2)^2 & \cdots & (x_2)^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & (x_n)^2 & \cdots & (x_n)^n \end{pmatrix}.$$

(a) Show that $\det(A)$ is a polynomial, all of whose monomials have degree $n(n+1)/2$.
(b) Show that for all $0 \leqslant i \neq j \leqslant n$, $x_j - x_i$ divides $\det(A)$. (Hint: consider $\mathbb{Z}[x_0, \dots, \hat{x_i}, \dots, x_n][x_i]$ and substitute $x_j$ for $x_i$.)
(c) Show that $\det(A) \sim \prod_{0 \leqslant i < j \leqslant n} (x_j - x_i)$.
(d) In fact, by considering the product of the diagonal, show that $\det(A) = \prod_{i<j \leqslant n} (x_j - x_i)$.
(e) Let $R$ be an integral domain, let $a_0, a_1 \dots, a_n \in R$, and let $f = \sum_{i \leqslant n} b_i x^i$ be a polynomial in $R[x]$ of degree $\leqslant n$. Observe that

$$\begin{pmatrix} 1 & a_0 & (a_0)^2 & \cdots & (a_0)^n \\ 1 & a_1 & (a_1)^2 & \cdots & (a_1)^n \\ 1 & a_2 & (a_2)^2 & \cdots & (a_2)^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & a_n & (a_n)^2 & \cdots & (a_n)^n \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} f(a_0) \\ f(a_1) \\ f(a_2) \\ \vdots \\ f(a_n) \end{pmatrix}.$$

Use this to show that if $R$ is an integral domain then any $f \in R[x]$ has at most $\deg f$ many roots in $R$.

2.94 Let $R$ be the collection of matrices in $M_2(\mathbb{Z})$ of the form $\left(\begin{smallmatrix} a & 2b \\ b & a \end{smallmatrix}\right)$. Show that $R$ is a commutative subring of $M_2(\mathbb{Z})$; show that it is isomorphic to the ring $\mathbb{Z}[\sqrt{2}]$.

2.95 Let $R$ be the collection of matrices in $M_2(\mathbb{Z}[i])$ of the form

$$\begin{pmatrix} z & w \\ -\bar{w} & \bar{z} \end{pmatrix};$$

Here $\bar{z}$ denotes usual complex conjugation.

(a) Show that $M$ is a (noncommutative) subring of $M_2(\mathbb{Z}[i])$.

(b) Show that the group of quaternions $Q$ (Example 2.55) is a subset of $R$.

(c) Show that for all $M \in R$ there is a unique quadruple of integers $(a, b, c, d)$ such that $M = aI_2 + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$.

(d) Show that $\det(aI_2 + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}) = a^2 + b^2 + c^2 + d^2$.

(a) Conclude that if $m$ and $n$ are natural numbers, and both $n$ and $m$ are sums of four square integers, then $nm$ is also the sum of four square integers.[14]

[14] Hence, the *four square theorem*—which states that every natural number is the sum of four squares—follows from the fact that every prime number is the sum of four squares.


# Affine Space

# 3

We are interested in curves in the plane which are defined by polynomial equations. Lines, circles, parabolas, ellipses and hyperbolas are such curves, but so are more complicated curves such as the cardioid (Fig. 3.1).

Every algebraic curve in $\mathbb{R}^2$ is defined by an equation $f(x, y) = 0$, where $f$ is a polynomial in $\mathbb{R}[x, y]$. In this chapter we start our systematic study of such curves. Our first main goal is obtaining a correspondence between curves and the polynomials that define them. In an ideal world each curve would be defined by a unique polynomial. This is not so, for three separate reasons. Let $f \in \mathbb{R}[x, y]$ be a polynomial.

(1) The polynomial $(x^2 + 1)f$ defines the same curve as $f$ does.
(2) The polynomial $f^2$ defines the same curve as $f$ does.
(3) If $\lambda \in \mathbb{R}$ is nonzero, then the polynomial $\lambda f$ defines the same curve as $f$ does.

These three obstructions are dealt with as follows. (1) follows from the fact that $\mathbb{R}$ is not algebraically closed. We will thus change the field with which we work. Usually we will use the complex field instead. We will take polynomials $f \in \mathbb{C}[x, y]$ and study the subsets of $\mathbb{C}^2$ that they define by the equation $f(x, y) = 0$. Because of the analogy with the real case we call these sets curves. However $\mathbb{C}^2$ has four real dimensions and the subsets defined by polynomials usually will have two real dimensions, so they are actually surfaces. In this part of the book we do not use the topological or analytic structure of the complex numbers, and so rather than restricting ourselves to the complex field we work with any algebraically closed field.

(2) is not solved by passing to an algebraically closed field. In order to separate between the curve defined by $f$ and the curve defined by $f^2$ we expand our notion of curve. We define our curves to be *multisets* of points (see p. 29). If $f$ defines the curve $C$ then we will let the curve defined by $f^2$ be $C + C$: two copies of $C$. To give a proper definition we will use the fact that $F[x, y]$ is a unique factorisation domain. For an irreducible polynomial $f$ we use the original definition: the curve $f = 0$, which below we denote by $V_{\mathbb{A}^2}(f)$, will be the set of all points $(a, b)$ for



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_3

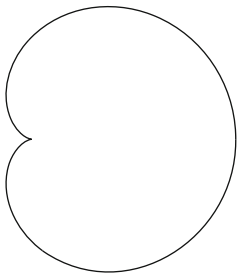


**Fig. 3.1** The cardioid is the trace left by a point on one circle, as that circle rolls along another circle. Its defining equation is $(x^2 + y^2 - 2x)^2 = 4(x^2 + y^2)$

which $f(a, b) = 0$. If $f$ is reducible then we let $[f_1, f_2, \ldots, f_m]$ be an irreducible decomposition of $f$, and define the curve $f = 0$ to be the multiset sum of the curves $f_1 = 0, f_2 = 0, \ldots, f_m = 0$. The uniqueness of irreducible factorisation will imply that this is well-defined.

The irreducible decomposition of $f$ is unique only up to association, which brings us to (3). Here the problem cannot be overcome by changing the field or even the notion of curve; we just have to accept that a curve is not defined by a unique polynomial, but rather by an association class of polynomials. Recall that $F[x, y]^* = F^* = F \setminus \{0\}$, so two polynomials $f$ and $g$ are associates if and only if $f = \lambda g$ for some nonzero scalar $\lambda$.

While our goal is studying curves, we will need to consider solutions to polynomial equations of more than two variables. These will be used, for example, in defining projective curves. Hence we will define the *algebraic hypersurfaces* of $n$-dimensional space, which generalise curves to higher dimensions. Our main goal in this chapter is Study's Lemma, which guarantees the uniqueness (up to scalar multiples) of a polynomial defining a given hypersurface.

Along the way, we introduce an important algebraic tool in *elimination theory*: the *resultant*, which tells us whether two polynomials have a nonconstant common factor. Beyond this chapter, the resultant will play a major role when we consider, in Chap. 6, how curves intersect each other.

## 3.1 Definition of Hypersurfaces

Fix a field $\mathbb{K}$. For a positive natural number $n \geqslant 1$ we let $\mathbb{A}^n(\mathbb{K}) = \mathbb{K}^n$. This is called *n-dimensional affine space over* $\mathbb{K}$. The affine *plane* is the case $n = 2$. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$.

**Definition 3.1** For an irreducible polynomial $f \in \mathbb{K}[\boldsymbol{x}]$, we let

$$V_{\mathbb{A}^n(\mathbb{K})}(f) = \left\{ \boldsymbol{a} \in \mathbb{A}^n(\mathbb{K}) \,:\, f(\boldsymbol{a}) = 0 \right\}.$$

When $\mathbb{K}$ is fixed we write $\mathbb{A}^n$ for $\mathbb{A}^n(\mathbb{K})$, and so also write $V_{\mathbb{A}^n}(f)$. Informally we write "the hypersurface $f = 0$" when we mean $V_{\mathbb{A}^n}(f)$, or even "the hypersurface

$f = g$" when we mean $V_{\mathbb{A}^n}(f - g)$. We also say that the polynomial $f$ *defines* the hypersurface $V_{\mathbb{A}^n}(f)$.

*Example 3.2* In two dimensions we use the variables $x, y$ rather than $x_1, x_2$. $V_{\mathbb{A}^2(\mathbb{R})}(x)$ is the $y$-axis, and $V_{\mathbb{A}^2(\mathbb{R})}(y)$ is the $x$-axis. The polynomial $x^2 + y^2 - 1$ is irreducible in $\mathbb{R}^2$ (as it is not the product of linear polynomials); so Definition 3.1 applies, and $V_{\mathbb{A}^2(\mathbb{R})}(x^2 + y^2 - 1)$ is the unit circle.

In three dimensions we use the variables $x, y, z$. For example, $V_{\mathbb{A}^3(\mathbb{R})}(x)$ is the $yz$-plane. «

Now we want to define the hypersurface defined by any given polynomial in $\mathbb{K}[\boldsymbol{x}]$, not necessarily irreducible. For the following lemma, recall that for $f, g \in \mathbb{K}[\boldsymbol{x}]$, if $f$ is irreducible and $g \sim f$ then $g$ is irreducible. Since $(\mathbb{K}[\boldsymbol{x}])^* = \mathbb{K}^* = \mathbb{K} \setminus \{0\}$ (see p. 26), $f \sim g$ if and only if $g = \lambda f$ for some nonzero constant $\lambda \in \mathbb{K}$. Since $\mathbb{K}$ is an integral domain, in this case, for all $\boldsymbol{a} \in \mathbb{K}^n$, $f(\boldsymbol{a}) = 0$ if and only if $g(\boldsymbol{a}) = 0$. Hence:

**Lemma 3.3** *If $f \in \mathbb{K}[\boldsymbol{x}]$ is irreducible and $g \sim f$ then $V_{\mathbb{A}^n}(f) = V_{\mathbb{A}^n}(g)$.*

If $[f_1, \ldots, f_m]$ and $[g_1, \ldots, g_k]$ are two irreducible factorisations of a polynomial $f$ then $m = k$ and after rearranging, $f_i \sim g_i$ for $i = 1, \ldots, m$ (Theorem 2.24). Thus, Lemma 3.3 implies that the following definition makes sense (does not depend on the choice of irreducible factorisation):

**Definition 3.4** Let $f \in \mathbb{K}[\boldsymbol{x}]$ be nonzero. We let

$$V_{\mathbb{A}^n}(f) = V_{\mathbb{A}^n}(f_1) + V_{\mathbb{A}^n}(f_2) + \cdots + V_{\mathbb{A}^n}(f_m),$$

where $[f_1, f_2, \ldots, f_m]$ is an irreducible factorisation of $f$.

If $f$ is a nonzero constant, then the empty multiset is an irreducible factorisation of $f$, so we get $V_{\mathbb{A}^n}(f) = \emptyset$. If $f$ is the constant 0, then we define $V_{\mathbb{A}^n}(f) = \mathbb{A}^n$. If $f$ is irreducible then $[f]$ is an irreducible factorisation of $f$, so Definition 3.4 agrees with Example 3.2 for irreducible polynomials.

If $f \sim g$ then $f$ and $g$ have the same irreducible factorisations, so Lemma 3.3 holds for all polynomials. Indeed, if $f$ divides $g$, $[f_1, \ldots, f_k]$ is an irreducible factorisation of $f$, and $[g_1, \ldots, g_m]$ is an irreducible factorisation of $g$, then $k \leqslant m$ and after rearranging, $f_i \sim g_i$ for $i \leqslant k$ (in the notation of Sect. 2.2, $[f_1, \ldots, f_k] \subseteq_\sim [g_1, \ldots, g_m]$). It follows that $V_{\mathbb{A}^n}(f) \subseteq V_{\mathbb{A}^n}(g)$, in the sense of multisets. The converse of this fact, known as Study's Lemma, only holds if $\mathbb{K}$ is algebraically closed. It is a special case of Hilbert's Nullstellensatz.

*Example 3.5* $V_{\mathbb{A}^2(\mathbb{R})}(x^2y)$ is the union of the $x$-axis and two copies of the $y$-axis, because $[x, x, y]$ is an irreducible factorisation of $x^2y$. Thus the origin $o = (0, 0)$ appears 3 times on $V_{\mathbb{A}^2(\mathbb{R})}(x^2y)$. «

Since $\mathbb{K}$ is an integral domain, if $[f_1, \ldots, f_m]$ is an irreducible factorisation of $f$ then $f(\boldsymbol{a}) = 0$ if and only if $f_i(\boldsymbol{a}) = 0$ for some $i \leqslant m$. It follows that

$$\lfloor V_{\mathbb{A}^n}(f) \rfloor = \left\{ \boldsymbol{a} \in \mathbb{A}^n \,:\, f(\boldsymbol{a}) = 0 \right\};$$

recall that $\lfloor C \rfloor$, the *underlying set* of a multiset $C$, is the collection of all elements of $C$ with multiplicities forgotten.

A multiset of points of $\mathbb{A}^n$ of the form $V_{\mathbb{A}^n}(f)$ for some nonzero polynomial $f \in \mathbb{K}[\boldsymbol{x}]$ is called an *algebraic hypersurface* of $\mathbb{A}^n$. For $n = 3$, an algebraic hypersurface of $\mathbb{A}^3$ is called an *algebraic surface* of $\mathbb{A}^3$. For $n = 2$, an algebraic hypersurface of $\mathbb{A}^2$ is called an *algebraic curve* of $\mathbb{A}^2$.

*Exercise 3.6* Show that a multiset of elements of $\mathbb{A}^1$ is an algebraic hypersurface of $\mathbb{A}^1$ if and only if it is finite. «

In this terminology, Proposition 2.18 says that if $\mathbb{K}$ is infinite, then no algebraic hypersurface contains all of $\mathbb{A}^n$; Proposition 2.19 says that no algebraic hypersurface contains $S^n$ where $S \subseteq \mathbb{K}$ is infinite.

## 3.2 The Resultant

The resultant is used to test whether two polynomials have a common factor. In this section, fix a unique factorisation domain $R$, and let $x$ be a variable (not already in $R$). Below, by a *nonconstant* polynomial we mean an element of $R[x] \setminus R$ (even if $R$ itself is a ring of polynomials).

### 3.2.1 The Sylvester Matrix

Let $f \in R[x]$ be nonconstant, and let $d \geqslant \deg f$. Write $f = a_0 + a_1x + \cdots + a_dx^d$ (if $d > \deg f$ then $a_d = 0$). Fix $e \geqslant 1$. Given a polynomial $k = c_0 + c_1x + \cdots + c_{e-1}x^{e-1}$ in $R[x]$ of degree less than $e$, we write its coefficients in a row $(c_0, c_1, \ldots, c_{e-1})$ of length $e$. The product $hf$ has degree less than $d + e$, and so we write its coefficients in a row of length $d + e$. The map taking the row $(c_0, \ldots, c_{e-1})$ to the row of coefficients of $hf$ is linear, and is given by multiplying (on the right)

by the $e \times (d+e)$-matrix

$$M^{d,e}(f) = \begin{pmatrix} a_0 & a_1 & \cdots & a_d & & & \\ & a_0 & a_1 & \cdots & a_d & & \\ & & a_0 & a_1 & \cdots & a_d & \\ & & & \ddots & & & \ddots \\ & & & & a_0 & a_1 & \cdots & a_d \end{pmatrix}.$$

*Exercise 3.7* Show that the transpose $M^{d,e}(f)^{\mathsf{t}}$ is nonsingular. «

Now given two polynomials $f = a_0 + a_1 x + \cdots + a_d x^d$ and $g = b_0 + b_1 x + \cdots b_e x^e$ of degrees at most $d$ and $e$, we consider the map taking two polynomials $k, h \in R[x]$ with $\deg k < e$ and $\deg h < d$ and producing the coefficients of the linear combination $kf + hg$. In terms of rows of coefficients, we write the coefficients of $k$ and $h$ sequentially in a row of length $d+e$, and we obtain a row of coefficients of length $d+e$ as well. It is linear, and is given by stacking the two matrices $M^{d,e}(f)$ and $M^{e,d}(g)$ on top of one another:

**Definition 3.8** Let $f, g \in R[x]$ be nonconstant; let $d \geqslant \deg f$, $e \geqslant \deg g$. The $d, e$-*Sylvester matrix* of $f$ and $g$ is

$$M^{d,e}(f,g) = \begin{pmatrix} M^{d,e}(f) \\ M^{e,d}(g) \end{pmatrix} = \begin{pmatrix} a_0 & a_1 & \cdots & a_d & & & & \\ & a_0 & a_1 & \cdots & a_d & & & \\ & & a_0 & a_1 & \cdots & a_d & & \\ & & & \ddots & & & \ddots & \\ & & & & a_0 & a_1 & \cdots & a_d \\ b_0 & b_1 & & \cdots & & b_e & & \\ & \ddots & & & & & \ddots & \\ & & b_0 & b_1 & & \cdots & & b_e \end{pmatrix}$$

where $f = \sum a_i x^i$ and $g = \sum b_i x^i$. If $d = \deg f$ and $e = \deg g$ then we write $M(f,g)$ for $M^{d,e}(f,g)$.

Note that $M^{d,e}(f,g)$ is square (of size $(d+e) \times (d+e)$), and that the main diagonal of $M^{d,e}(f,g)$ contains $e$ many $a_0$'s and $d$ many $b_e$'s.

**Lemma 3.9** *Let $f$ and $g$ be nonconstant polynomials in $R[x]$, and let $d \geqslant \deg f$ and $e \geqslant \deg g$. The matrix $M^{d,e}(f,g)$ is singular if and only if there are nonzero polynomials $h$ and $k$ in $R[x]$ such that $\deg h < d$, $\deg k < e$ and $kf = hg$.*

***Proof*** $M = M^{d,e}(f,g)$ is nonsingular if and only if $\underline{\gamma}M = \underline{0}$ implies $\underline{\gamma} = \underline{0}$ (Corollary 2.69). If $\underline{\alpha}$ and $\underline{\beta}$ are the rows of coefficients of polynomials $k$ and $h$ as above, then $(\underline{\alpha}, -\underline{\beta})M$ is the row of coefficients of $kf - hg$, so $(\underline{\alpha}, -\underline{\beta}) \cdot M = \underline{0}$ if and only if $kf = hg$.

This shows that $M$ is singular if and only if there are polynomials $h$ and $k$ (with $\deg h < d$, $\deg k < e$), *at least one of which is nonzero*, such that $kf = hg$. However $R[x]$ is an integral domain, and $f$ and $g$ are nonzero, so one of $k$ or $h$ being nonzero implies that the other is nonzero as well. □

**Lemma 3.10** *Let $f$ and $g$ be nonconstant polynomials in $R[x]$. The polynomials $f$ and $g$ have a nonconstant common factor if and only if there are nonzero polynomials $h$ and $k$ in $R[x]$ such that* $\deg h < \deg f$, $\deg k < \deg g$ *and* $kf = hg$.

***Proof*** If $p \in R[x]$ is a nonconstant common factor of $f$ and $g$ then $h = f/p$ and $k = g/p$ are as required. The other direction uses the fact that $R[x]$ is a unique factorisation domain (Proposition 2.31). Suppose that $\deg h < \deg f$ and $f$ divides $hg$. The irreducible nonconstant factors of $f$ appear in $hg$, and they cannot all appear in $h$ as $\deg h < \deg f$; and so one of them is also a factor of $g$. □

### 3.2.2 The Resultant, Common Roots, and More Variables

The Sylvester matrix is square, and so has a determinant.

**Definition 3.11** Let $f, g \in R[x]$ be nonconstant and let $d \geqslant \deg f$, $e \geqslant \deg g$. The *$d, e$-resultant* of $f$ and $g$, denoted by $\operatorname{res}^{d,e}(f,g)$, is the determinant of the Sylvester matrix $M^{d,e}(f,g)$.

If $d = \deg f$ and $e = \deg g$ then we write $\operatorname{res}(f,g)$ for $\operatorname{res}^{d,e}(f,g)$; $\operatorname{res}(f,g)$ is known as the *resultant* of $f$ and $g$.

Note that since the entries of $M^{d,e}(f,g)$ are elements of $R$, so is $\operatorname{res}^{d,e}(f,g)$. The fact that the Sylvester matrix is singular if and only if its determinant is zero (Theorem 2.68), together with Lemmas 3.9 and 3.10 shows:

**Theorem 3.12** *Nonconstant polynomials $f$ and $g$ in $R[x]$ have a nonconstant common factor $p \in R[x]$ if and only if* $\operatorname{res}(f,g) = 0$. □

An important special case is when $f$ and $g$ are polynomials (in one variable) over an algebraically closed field $\mathbb{K}$. In this case the irreducible polynomials are

the linear ones, and two polynomials $f$ and $g \in \mathbb{K}[x]$ have a common nonconstant factor if and only if they have a common root.

### Adding More Variables

If $R$ is itself a polynomial ring $S[\boldsymbol{y}]$ for some unique factorisation domain $S$, then we may need to mention the fact that when calculating $M^{d,e}(f,g)$ and $\operatorname{res}^{d,e}(f,g)$ we are doing so *with respect to the variable* $x$. This we do by writing $M_x^{d,e}(f,g)$ and $\operatorname{res}_x^{d,e}(f,g)$. We drop the subscript when the variable is clear from the context.

*Example 3.13* Let $f = x - y^2$ and $g = x^3y$ be elements of $\mathbb{Z}[x,y]$. Then $M_y(f,g) = M_y^{2,1}(f,g) = \begin{pmatrix} x & 0 & -1 \\ 0 & x^3 & 0 \\ 0 & 0 & x^3 \end{pmatrix}$, so $\operatorname{res}_y(f,g) = x^7$ (which is indeed an element of $\mathbb{Z}[x]$), and $M_x(f,g) = M_x^{1,3}(f,g) = \begin{pmatrix} -y^2 & 1 & 0 & 0 \\ 0 & -y^2 & 1 & 0 \\ 0 & 0 & -y^2 & 1 \\ 0 & 0 & 0 & y \end{pmatrix}$, and so $\operatorname{res}_x(f,g) = -y^7$.

To illustrate what happens if we pick dimensions greater than the degrees, check that $\operatorname{res}_y^{2,2}(f,g) = -x^7$ and that $\operatorname{res}_x^{2,3}(f,g) = -y^8$.

We note though that if $d > \deg f$ and $e > \deg g$ then $\operatorname{res}^{d,e}(f,g) = 0$, as the last column of the Sylvester matrix is zero. «

*Exercise 3.14* Let $f, g \in R[x]$, and let $d > \deg f$ and $e > \deg g$. The fact that $\operatorname{res}^{d,e}(f,g) = 0$ implies that there are polynomials $k$ and $h$ in $R[x]$ such that $\deg h < d$, $\deg k < e$ and $kf = hg$. Find such polynomials. «

Suppose that $f, g \in R[y,x]$. Fix $d \geqslant \deg_x f$ and $e \geqslant \deg_x g$. For any $a \in R$, $f(a,x)$ and $g(a,x)$ are polynomials in $R[x]$, with $\deg f(a,x) \leqslant d$ and $\deg g(a,x) \leqslant e$. The entries of $M_x^{d,e}(f,g)$ are polynomials in $R[y]$. The matrix $M^{d,e}(f(a,x), g(a,y))$ is obtained from the matrix $M_x^{d,e}(f,g)$ by substituting $a$ into each of the entries of $M^{d,e}(f)$, and so, after taking the determinant, $\operatorname{res}^{d,e}(f(a,x), g(a,x))$ is the result of substituting $a$ into the polynomial $\operatorname{res}_x^{d,e}(f,g)$ (see the discussion on p. 47).

However, it is possible that $d = \deg_x f$, $e = \deg_x g$, but $d > \deg f(a,x)$ or $e > \deg g(a,x)$, or both. For example, take $R = \mathbb{Z}$, $f = 3yx^2 + (y^3+3)x + (y^5+7y)$, and $a = 0$; then $\deg_x f = 2$ but $\deg f(a,x) = 1$. In this case we say that the leading coefficient of $f$ *vanishes* when we substitute $a$ for $y$ in $f$. In this case, we will have $M_x(f,g) = M_x^{d,e}(f,g)$, but $M(f(a,x), g(a,x)) \neq M^{d,e}(f(a,x), g(a,x))$, becase the last two matrices have different sizes; and it will be impossible to derive any connection between the resultants $\operatorname{res}_x(f,g)$ and $\operatorname{res}(f(a,x), g(a,x))$.

We can replace $y$ by an $m$-tuple of variables $\boldsymbol{y}$. The following lemma says that when the leading coefficients of $f$ and $g$ do not vanish, the resultant "commutes" with substitution:

**Lemma 3.15** *Let $f, g \in R[\boldsymbol{y}, x]$ be nonconstant in $x$ and let $\boldsymbol{a} \in R^m$. If* $\deg f(\boldsymbol{a}, x) = \deg_x f$ *and* $\deg g(\boldsymbol{a}, x) = \deg_x g$ *then*

$$\operatorname{res}(f, g)(\boldsymbol{a}) = \operatorname{res}_x(f(\boldsymbol{a}, x), g(\boldsymbol{a}, x)). \qquad \square$$

**Proposition 3.16** *Let $f, g \in R[\boldsymbol{y}, x]$ be nonconstant in $x$. Let $\boldsymbol{a} \in R^m$. The following are equivalent:*

*(1) Either $f(\boldsymbol{a}, x)$ and $g(\boldsymbol{a}, x)$ have a nonconstant common factor in $R[x]$, or* both $\deg_x f(\boldsymbol{a}, x) < \deg_x f$ *and* $\deg_x g(\boldsymbol{a}, x) < \deg_x g$.
*(2)* $(\operatorname{res}_x(f, g))(\boldsymbol{a}) = 0$.

***Proof*** Let $r = \operatorname{res}_x(f, g)$. Let $d = \deg_x f$ and $e = \deg_x g$. There are three cases. If $\deg_x f(\boldsymbol{a}, x) = d$ and $\deg_x g(\boldsymbol{a}, x) = e$ then $\operatorname{res}(f(\boldsymbol{a}, x), g(\boldsymbol{a}, x)) = r(\boldsymbol{a})$; in this case the equivalence is by Theorem 3.12. If $\deg_x f(\boldsymbol{a}, x) < d$ and $\deg_x g(\boldsymbol{a}, x) < e$ then as observed above, $r(\boldsymbol{a}) = \operatorname{res}_x^{d,e}(f(\boldsymbol{a}, x), g(\boldsymbol{a}, x)) = 0$ (the last column of $M_x^{d,e}(f, g)(\boldsymbol{a})$ is zero).

Finally, suppose, without loss of generality, that $\deg_x g(\boldsymbol{a}, x) = e$ but that $\deg_x f(\boldsymbol{a}, x) < d$. While $\operatorname{res}(f(\boldsymbol{a}, x), g(\boldsymbol{a}, x))$ may not be the same as $r(\boldsymbol{a})$, still $r(\boldsymbol{a}) = \operatorname{res}^{d,e}(f(\boldsymbol{a}, x), g(\boldsymbol{a}, x))$. Lemma 3.9 says that $r(\boldsymbol{a}) = 0$ if and only if there are polynomials $k, h \in R[x]$ such that $\deg h < d$, $\deg k < e$ and $k(x) f(\boldsymbol{a}, x) = h(x) g(\boldsymbol{a}, x)$. An examination of the proof of Lemma 3.10 shows in fact that $g(\boldsymbol{a}, x)$ dividing $k(x) f(\boldsymbol{a}, x)$ with $\deg k < e = \deg g(\boldsymbol{a}, x)$ is equivalent to $f(\boldsymbol{a}, x)$ and $g(\boldsymbol{a}, x)$ having a nonconstant common factor in $R[x]$. $\square$

### 3.2.3 The Resultant is a Linear Combination

The last fact we need (for now) about the resultant $\operatorname{res}(f, g)$ is the fact that it is a linear combination of $f$ and $g$ over $R[x]$.

**Proposition 3.17** *Let $f, g \in R[x]$ be nonconstant. There are polynomials $h, k \in R[x]$ such that* $\operatorname{res}(f, g) = kf + hg$.

In fact, we get such polynomials with $\deg k < \deg g$ and $\deg h < \deg f$.

***Proof*** Let $d = \deg f$ and $e = \deg g$. Let $\bar{u}_0, \bar{u}_1, \ldots, \bar{u}_{d+e-1}$ be the columns of the Sylvester matrix $M(f, g) = M^{d,e}(f, g)$ (so the entries of these columns are in $R$).

Let $\bar{w}$ be the column

$$\bar{w} = \bar{u}_0 + x\bar{u}_1 + \cdots + x^{d+e-1}\bar{u}_{d+e-1} = \begin{pmatrix} f \\ xf \\ x^2 f \\ \vdots \\ x^{e-1} f \\ g \\ xg \\ \vdots \\ x^{d-1} g \end{pmatrix}.$$

Let $A = (\bar{w}, \bar{u}_1, \bar{u}_2, \ldots, \bar{u}_{d+e-1})$ be the matrix obtained from $M(f, g)$ by replacing the first column by $\bar{w}$. By the multi-linearity of the determinant function,

$$\det(A) = \det(\bar{w}, \bar{u}_1, \bar{u}_2, \ldots, \bar{u}_{d+e-1}) = \sum_{j=0}^{d+e-1} x^j \det(\bar{u}_j, \bar{u}_1, \bar{u}_2, \ldots, \bar{u}_{d+e-1}).$$

For $j > 0$, the matrix $(\bar{u}_j, \bar{u}_1, \bar{u}_2, \ldots, \bar{u}_{d+e-1})$ contains two repeated columns, and so its determinant is zero; so $\det A = \det(\bar{u}_0, \ldots, \bar{u}_{d+e-1}) = \operatorname{res}(f, g)$. Now, develop $\det A$ along its first column $\bar{w}$:

$$\operatorname{res}(f, g) = \det A = \sum_{j=1}^{d+e} (-1)^{j+1} w_j \det(A_{j,1}) =$$

$$\sum_{i=0}^{e-1} (-1)^i \det(A_{i+1,1}) x^i f + \sum_{i=0}^{d-1} (-1)^{i+e} \det(A_{i+e+1,1}) x^i g$$

so we let $k = \sum_{i=0}^{e-1} (-1)^i \det(A_{i+1,1}) x^i$ and $h = \sum_{i=0}^{d-1} (-1)^{i+e} \det(A_{i+e+1,1}) x^i$.

The bound on the degree of $k$ and $h$ follows from the observation that every entry of $A_{j,1}$ (for $j \leqslant d + e$) is in $R$, so $\det A_{j,1}$ is also in $R$. □

## 3.3 Study's Lemma

**Study's Lemma** *Suppose that* $\mathbb{K}$ *is algebraically closed. Let* $f, g \in \mathbb{K}[x_1, x_2, \ldots, x_n]$ *be nonzero polynomials. Then*

$$f \text{ divides } g \iff V_{\mathbb{A}^n(\mathbb{K})}(f) \subseteq V_{\mathbb{A}^n(\mathbb{K})}(g).$$

Study's lemma implies that if $\mathbb{K}$ is algebraically closed, then $f \sim g$ if and only if $V_{\mathbb{A}^n}(f) = V_{\mathbb{A}^n}(g)$. In particular, Study's lemma implies that if $f$ is a nonconstant polynomial, then $V_{\mathbb{A}^n(\mathbb{K})}(f) \neq \emptyset$. This is a good exercise to begin with:

*Exercise 3.18* Show directly that if $\mathbb{K}$ is algebraically closed, then for any nonconstant polynomial $f \in \mathbb{K}[x_1, \ldots, x_n]$, the hypersurface $V_{\mathbb{A}^n}(f)$ is nonempty. «

For the rest of this section, we assume that the field $\mathbb{K}$ is algebraically closed. Before we prove Study's lemma, we present a couple of consequences.

To begin with, let $C$ be a a hypersurface of $\mathbb{A}^n$. If $f$ and $g$ are two polynomials defining $C$, then $f \sim g$. This implies that $f$ is irreducible if and only if $g$ is irreducible. We can therefore define:

**Definition 3.19** An algebraic hypersurface is *irreducible* if the polynomials defining it are irreducible.

Similarly, $f \sim g$ implies $\deg f = \deg g$, and so when $\mathbb{K}$ is algebraically closed, hypersurfaces have a well-defined notion of degree:

**Definition 3.20** The *degree* of a hypersurface $C$ of $\mathbb{A}^n$, denoted by $\deg C$, is the degree of the polynomials defining $C$.

A *conic curve* is a curve defined by a polynomial of degree 2 (also known as a *quadratic* curve), a *cubic curve* is a curve defined by a polynomial of degree 3, a *quartic* by a polynomial of degree 4 and so on.

Uniqueness of the polynomial defining a hypersurface, together with unique factorisation in $\mathbb{K}[\boldsymbol{x}]$, allows us to identify the *irreducible components* of a hypersurface in $\mathbb{A}^n$. If $C$ is a hypersurface in $\mathbb{A}^n$, then the irreducible components of $C$ are the irreducible hypersurfaces $D$ such that $D \subseteq C$. Study's lemma tells us that the irreducible components of $V_{\mathbb{A}^n}(f)$ are the hypersurfaces $V_{\mathbb{A}^n}(g)$ where $g$ is an irreducible factor of $f$.

**Proposition 3.21** *Let* $f$ *and* $g$ *be nonzero polynomials in* $\mathbb{K}[\boldsymbol{x}]$. *Then* $\lfloor V_{\mathbb{A}^n}(f) \rfloor \subseteq \lfloor V_{\mathbb{A}^n}(g) \rfloor$ *if and only if every irreducible factor of* $f$ *also divides* $g$.

***Proof*** Suppose that every irreducible factor of $f$ divides $g$. Let $\boldsymbol{a} \in \mathbb{K}^n$. If $f(\boldsymbol{a}) = 0$ then $h(\boldsymbol{a}) = 0$ for some irreducible $h \mid f$. Since $h$ divides $g$ as well, we have $g(\boldsymbol{a}) = 0$. Hence $\lfloor V_{\mathbb{A}^n}(f) \rfloor \subseteq \lfloor V_{\mathbb{A}^n}(g) \rfloor$.

In the other direction, suppose that $\lfloor V_{\mathbb{A}^n}(f) \rfloor \subseteq \lfloor V_{\mathbb{A}^n}(g) \rfloor$. Let $h$ be an irreducible factor of $f$. Then $V_{\mathbb{A}^n}(h)$ is a set, and so is a subset of $\lfloor V_{\mathbb{A}^n}(f) \rfloor$, and so is a subset of $\lfloor V_{\mathbb{A}^n}(g) \rfloor$, which in turn is a subset of the multiset $V_{\mathbb{A}^n}(g)$. That is, $V_{\mathbb{A}^n}(h) \subseteq V_{\mathbb{A}^n}(g)$. By Study's lemma, $h$ divides $g$. □

The following result allows us to find the irreducible components of a hypersurface using its underlying set.

**Proposition 3.22** *Let $C$ be a hypersurface of $\mathbb{A}^n$, and let $A_1, \ldots, A_k$ be irreducible hypersurfaces of $\mathbb{A}^n$ such that $\lfloor C \rfloor = \bigcup_{i \leqslant k} A_i$. Then the irreducible components of $C$ are $A_1, \ldots, A_k$.*

***Proof*** Choose $g$ deifning $C$; for $i \leqslant k$, choose $f_i$ defining $A_i$, and let $f = f_1 \cdot f_2 \cdots f_k$. Then $\lfloor C \rfloor = \left\lfloor \sum_{i \leqslant k} A_i \right\rfloor = \lfloor V_{\mathbb{A}^n}(f) \rfloor$. Proposition 3.21 says that $f$ and $g$ have the same irreducible factors, namely $f_1, \ldots, f_k$. □

### 3.3.1 Proof of Study's Lemma

We observed earlier that the left-to-right direction of Study's lemma holds for any field $\mathbb{K}$, not necessarily algebraically closed. We let $f, g \in \mathbb{K}[\boldsymbol{x}]$ be nonzero polynomials such that $V_{\mathbb{A}^n}(f) \subseteq V_{\mathbb{A}^n}(g)$, and show that $f$ divides $g$.

First, we argue that we may assume that $f$ is irreducible. Suppose that Study's lemma is known to hold when the first polynomial is irreducible. Then we can prove the full Study's lemma by induction on the number of irreducible factors of $f$. Let $f$ be any polynomial, and suppose that $V_{\mathbb{A}^n}(f) \subseteq V_{\mathbb{A}^n}(g)$. Let $p$ be an irreducible factor of $f$. Then $V_{\mathbb{A}^n}(p) \subseteq V_{\mathbb{A}^n}(g)$, and so by assumption, $p$ divides $g$. Then we can "peel off" $p$ from both sides: $V_{\mathbb{A}^n}(f) = V_{\mathbb{A}^n}(p) + V_{\mathbb{A}^n}(f/p)$, and $V_{\mathbb{A}^n}(g) = V_{\mathbb{A}^n}(p) + V_{\mathbb{A}^n}(g/p)$; we conclude that $V_{\mathbb{A}^n}(f/p) \subseteq V_{\mathbb{A}^n}(g/p)$. Since $f/p$ has fewer irreducible factors than $f$, by induction, $f/p$ divides $g/p$, and so overall we get $f \mid g$.

Assuming now that $f$ is irreducible, $V_{\mathbb{A}^n}(f)$ is a set (all the elements of $V_{\mathbb{A}^n}(f)$ have multiplicity 1), so the assumption is that for all $\boldsymbol{a} \in \mathbb{K}^n$, if $f(\boldsymbol{a}) = 0$ then $g(\boldsymbol{a}) = 0$.

If $f$ is constant then it is a unit so it certainly divides $g$. Otherwise, $\deg_{x_i} f > 0$ for some $i \leqslant n$. For notational simplicity, we assume $i = n$. We use the following lemma:

**Lemma 3.23** *Let $f, g \in \mathbb{K}[x_1, \ldots, x_n]$. If $\deg_{x_n} f > 0$, $\deg_{x_n} g = 0$ and $V_{\mathbb{A}^n}(f) \subseteq V_{\mathbb{A}^n}(g)$, then $g = 0$.*

Note that under the hypotheses of the lemma, we have $g \in \mathbb{K}[x_1, \ldots, x_{n-1}]$, but we are still considering $V_{\mathbb{A}^n}(g)$, as $g$ is also an element of $\mathbb{K}[x_1, \ldots, x_n]$; but since $x_n$ does not appear in $g$, we have $V_{\mathbb{A}^n}(g) = V_{\mathbb{A}^{n-1}}(g) \times \mathbb{K}$. Thus, the lemma says that no hypersurface of $\mathbb{A}^n$ can be contained in a *cylinder* $D \times \mathbb{K}$ in $\mathbb{A}^n$ defined by a hypersurface $D$ of $\mathbb{A}^{n-1}$, unless it is a cylinder itself.

***Proof*** We prove the contrapositive. Suppose that $g \neq 0$. Let $d = \deg_{x_n} f$, and write $f = f_0 + f_1 x_n + \cdots + f_d x_n^d$, with $f_0, f_1, \ldots, f_d \in \mathbb{K}[x_1, \ldots, x_{n-1}]$; so $f_d \neq 0$. Since $\mathbb{K}[x_1, \ldots, x_{n-1}]$ is an integral domain, $g f_d \neq 0$. By Proposition 2.18 there is some $\boldsymbol{a} \in \mathbb{K}^{n-1}$ such that $(g f_d)(\boldsymbol{a}) \neq 0$; so $g(\boldsymbol{a}) \neq 0$ and $f_d(\boldsymbol{a}) \neq 0$. The latter shows that $\deg_{x_n} f(\boldsymbol{a}, x_n) = d > 0$; since $\mathbb{K}$ is algebraically closed, there is some $b \in \mathbb{K}$ such that $f(\boldsymbol{a}, b) = 0$, so $(\boldsymbol{a}, b) \in V_{\mathbb{A}^n}(f)$. But $g(\boldsymbol{a}) \neq 0$ means that $(\boldsymbol{a}, b) \notin V_{\mathbb{A}^n}(g)$. Hence $V_{\mathbb{A}^n}(f) \not\subseteq V_{\mathbb{A}^n}(g)$. □

We conclude that $\deg_{x_n} g > 0$. Let $r = \operatorname{res}_{x_n}(f, g)$ (which is defined since $\deg_{x_n} f, \deg_{x_n} g > 0$); it is an element of $\mathbb{K}[x_1, \ldots, x_{n-1}]$. Let $\boldsymbol{a} \in V_{\mathbb{A}^n}(f)$. By assumption, $g(\boldsymbol{a}) = 0$ as well. Since $r$ is a linear combination of $f$ and $g$ (Proposition 3.17), $r(\boldsymbol{a}) = 0$. So $V_{\mathbb{A}^n}(f) \subseteq V_{\mathbb{A}^n}(r)$. As $\deg_{x_n} r = 0$, Lemma 3.23 ensures that $r = 0$. Hence $f$ and $g$ have a common divisor $h$ such that $\deg_{x_n} h > 0$ (Theorem 3.12). Since $f$ is irreducible and $h$ is not a unit, we must have $h \sim f$, and so $f$ divides $g$.

This completes the proof of Study's Lemma.

## 3.4 Affine Lines and Rational Parameterisations

Algebraic plane curves are defined *implicitly*, as the set of points which satisfy an equation. We can also determine curves *parametrically*, by "listing" their points. For example, the unit circle is both given by the equation $x^2 + y^2 = 1$ and as the collection of points $(\sin t, \cos t)$ for $t \in \mathbb{R}$. The simplest example is that of lines.

### 3.4.1 Affine Lines

In Exercise 2.88 we defined the notion of an *affine subspace* of $\mathbb{A}^n$—these are the subsets of $\mathbb{A}^n$ of the form $\boldsymbol{a} + U$, where $U$ is a linear subspace of $\mathbb{K}^n$. The dimension of the subspace $\boldsymbol{a} + U$ is defined to be the dimension of $U$. Similarly, an *affine map* from $\mathbb{A}^n$ to $\mathbb{A}^m$ is a map of the form $\boldsymbol{p} \mapsto T(\boldsymbol{p}) + \boldsymbol{b}$ where $T : \mathbb{A}^n \to \mathbb{A}^m$ is linear and $\boldsymbol{b} \in \mathbb{A}^m$.

*Exercise 3.24* Show that the following are equivalent for $W \subseteq \mathbb{A}^n$: (i) $W$ is the zero set $\{\boldsymbol{p} \in \mathbb{A}^n : f(\boldsymbol{p}) = \boldsymbol{0}\}$ of some affine map $f : \mathbb{A}^n \to \mathbb{A}^m$ which is onto $\mathbb{A}^m$; (ii) $W$ is an affine subspace of $\mathbb{A}^n$ of dimension $n - m$. (See Corollary 2.63). «

Affine maps from $\mathbb{A}^n$ to $\mathbb{A}^1$ are precisely the maps defined by linear polynomials $f \in \mathbb{K}[x_1, \ldots, x_n]$ (polynomials which are constant or have degree 1): the function defined by $f = a_1x_1 + a_2x_2 + \cdots + a_nx_n + b$ is the affine map $T + b$, where $T$ is the linear map given by the row matrix $\begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix}$. The affine map defined by a linear polynomial $f$ is onto $\mathbb{A}^1$ if and only $f$ is nonconstant. Note that all polynomials of degree 1 are irreducible. Thus, Exercise 3.24 says that the affine *hyperplanes* of $\mathbb{A}^n$—the affine subspaces of dimension $n - 1$—are precisely the algebraic hypersurfaces of $\mathbb{A}^n$ of degree 1.

A *line* in $\mathbb{A}^n$ is an affine subspace of dimension 1. Lines in $\mathbb{A}^2$ are the hyperplanes of $\mathbb{A}^2$. We thus get:

**Proposition 3.25** *The lines of $\mathbb{A}^2$ are the algebraic curves of degree 1.*

*Exercise 3.26* Show that if $\boldsymbol{p}$ and $\boldsymbol{q}$ are distinct points in $\mathbb{A}^n$ then there is a unique line which passes through $\boldsymbol{p}$ and $\boldsymbol{q}$. «

**Definition 3.27** A *linear parameterisation* in $\mathbb{A}^n$ is an injective affine map from $\mathbb{A}^1$ to $\mathbb{A}^n$.

The images of injective affine maps from $\mathbb{A}^m$ to $\mathbb{A}^n$ are precisely the $m$-dimensional affine subspaces of $\mathbb{A}^n$, and so lines are the images of linear parameterisations. A parameterisation *of* a line $\ell$ is a parameterisation whose range is $\ell$. A linear parameterisation $\psi : \mathbb{A}^1 \to \mathbb{A}^n$ is defined by a tuple of linear polynomials in one variable: $\psi(a) = (\psi_1(a), \ldots, \psi_n(a))$ where $\psi_i \in \mathbb{K}[t]$ is linear. In particular, linear parameterisations of lines in $\mathbb{A}^2$ are of the form $\psi(a) = (\psi_x(a), \psi_y(a))$ with $\psi_x, \psi_y \in \mathbb{K}[t]$ linear and at least one of $\psi_x$ and $\psi_y$ nonconstant.

*Example 3.28* If $\boldsymbol{a}, \boldsymbol{b}$ are distinct points in $\mathbb{A}^2$ then $\psi(t) = (1 - t)\boldsymbol{a} + t\boldsymbol{b}$ is the unique parameterisation of the line passing through $\boldsymbol{a}$ and $\boldsymbol{b}$ satisfying $\psi(0) = \boldsymbol{a}$ and $\psi(1) = \boldsymbol{b}$. «

### 3.4.2 Rational Parameterisations

To parameterise more complicated curves, we need more complicated functions. We now consider parameterisations by *rational* functions. Recall that the field of fractions of $\mathbb{K}[\boldsymbol{x}]$ (denoted by $\mathbb{K}(\boldsymbol{x})$) is called the field of formal rational functions in $\boldsymbol{x}$ (with coefficients in $\mathbb{K}$); see Example 2.35. A formal rational function $h = f/g \in \mathbb{K}(x_1, \ldots, x_n)$ defines a partial function on $\mathbb{A}^n$, $\boldsymbol{a} \mapsto f(\boldsymbol{a})/g(\boldsymbol{a})$. Note that the values of the function do not depend on the presentation of $h$ as a ratio of polynomials, but the *domain* of the function, namely $\mathbb{A}^n \setminus V_{\mathbb{A}^n}(g)$—does; for that reason we usually ask that $f$ and $g$ do not have a common factor, so that $h$ is given in simplest terms.

A *rational map* from $\mathbb{A}^m$ to $\mathbb{A}^n$ is a map of the form $\boldsymbol{a} \mapsto (h_1(\boldsymbol{a}), \ldots, h_n(\boldsymbol{a}))$, where $h_i \in \mathbb{K}(x_1, \ldots, x_m)$. To parameterise a curve, we take rational maps from $\mathbb{A}^1$. Since such maps will not be defined on every input, we sometimes must omit some "bad points". We thus define:

**Definition 3.29** A *rational parameterisation* of a curve $C$ in $\mathbb{A}^2$ is a rational map $\psi : \mathbb{A}^1 \to \mathbb{A}^2$ such that for some finite $Q \subset \mathbb{A}^1$ and finite $P \subset C$, the restriction $\psi\restriction_{\mathbb{A}^1 \setminus Q}$ is a bijection between $\mathbb{A}^1 \setminus Q$ and $C \setminus P$.

A curve $C$ is *rational* if it has a rational parameterisation.

The standard example is the graph of a function: if $C$ is the curve $y = f(x)$, with $f \in \mathbb{K}[x]$, then $a \mapsto (a, f(a))$ is a rational parameterisation of $C$. Another example is the rational parameterisation of the unit circle given in Chap. 1 (see Fig. 1.1): here $\psi(a) = (2a/(a^2+1), (a^2-1)/(a^2+1))$ is a bijection between $\mathbb{A}^1(\mathbb{R})$ and the unit circle excluding one point. If we replace $\mathbb{R}$ by $\mathbb{C}$, we again get an "almost bijection"; the parameterisation though is undefined at $i$ and $-i$.

Which curves are rational? The answer relies on the topological notion of the *genus* of a curve, which we will not discuss. A rational curve with no repeated components must be irreducible: see Exercise 6.49.

Assuming that a given curve $C$ is rational, how do we find a parameterisation? Sometimes, the technique used for the unit circle can be applied. We find a point $q \in C$ such that every line through $q$ (or at least all but finitely many) intersects $C$ in one other point. We then associate with every $a \in \mathbb{K}$ some line $\ell_a$ passing through $q$; and we let $\psi(a)$ be the other point of intersection of the line $\ell_a$ with $C$. For example, we can choose $q$ to be the origin $o$, and let $\ell_a$ be the line $y = ax$; see Exercises 3.47 and 3.48 below.

The opposite question is how to find an implicit definition of a curve which is given as the range of a parameterisation. Such a definition always exists.

**Proposition 3.30** *If $\psi : \mathbb{A}^1 \to \mathbb{A}^2$ is a rational map, then the range of $\psi$ is contained in an algebraic curve. If $\mathbb{K}$ is algebraically closed, then we can choose the curve to contain at most one point outside the range of $\psi$.*

***Proof*** Write $\psi = (\psi_x, \psi_y)$; let $g_x, h_x, g_y, h_y \in \mathbb{K}[t]$ such that $\psi_x = g_x/h_x$, $\psi_y = g_y/h_y$, and $g_x$ and $h_x$ have no nonconstant common factor, and similarly $g_y$ and $h_y$. Define the polynomials $p(x, y, t) = xh_x(t) - g_x(t)$ and $q(x, y, t) = yh_y(t) - g_y(t)$; even though $y$ does not appear in $p$ and $x$ does not appear in $q$, we think of both as polynomials in $\mathbb{K}[x, y][t]$ so we take $r = \mathrm{res}_t(p, q)$, which is a polynomial in $\mathbb{K}[x, y]$. However, any common factor of $p$ and $q$ must be in $\mathbb{K}[t]$. Viewing $p$ as a polynomial in $\mathbb{K}[t][x]$ we see that such a common factor must divide both $g_x(t)$ and $h_x(t)$; since $g_x$ and $h_x$ were chosen to have no nonconstant common factor, $p$ and $q$ have no nonconstant common factor. By Theorem 3.12, $r \neq 0$, so $r$ defines a curve in $\mathbb{A}^2$.

If $(b, c) = \psi(a)$ then $p(b, c, a) = q(b, c, a) = 0$; by Proposition 3.16, $r(b, c) = 0$. Hence the range of $\psi$ is contained in the curve $V_{\mathbb{A}^2}(r)$. In the other direction, let $(b, c) \in V_{\mathbb{A}^2}(r)$. Proposition 3.16 gives us two possibilities: either $p(b, c, t)$ and $q(b, c, t)$ have a nonconstant common component, or $\deg_t p(b, c, t) < \deg_t p$ and $\deg_t q(b, c, t) < \deg_t q$. Viewed as polynomials in $\mathbb{K}[x, y][t]$, the coefficients of $p$ and $q$ are linear; this means that the second possibility happens for precisely one point $(b, c)$. This point may or may not be in the range of $\psi$. If $\mathbb{K}$ is algebraically closed, then when $p(b, c, t)$ and $q(b, c, t)$ have a common component, they have a common root $a$, i.e., $bh_x(a) = g_x(a)$ and $ch_y(a) = g_y(a)$. Since $g_x$ and $h_x$ don't have a common root, we cannot have $h_x(a) = 0$, and similarly, $h_y(a) \neq 0$; so $(b, c) = \psi(a)$. Hence, if $\mathbb{K}$ is algebraically closed, the range of $\psi$ is the underlying set of the curve $r = 0$, possibly missing one point. □

For examples, see Exercise 3.43. For more on rational parameterisations, see, for example, [Gib98, Chap. 8, Sect. 14.3].

## 3.5 Further Exercises

In the following exercises, unless otherwise stated, let $\mathbb{K}$ be any field. We write $\mathbb{A}^n$ for $\mathbb{A}^n(\mathbb{K})$.

**Lines**

3.31 Let $\ell = V_{\mathbb{A}^2(\mathbb{Z}/(5))}(2x + 3y + 1)$. List the elements of $\ell$. Find elements $\bar{u}$ and $\bar{v}$ of $(\mathbb{Z}/(5))^2$ such that $\ell$ is the 1-dimensional affine subspace $\langle \bar{u} \rangle_{(\mathbb{Z}/(5))^2} + \bar{v}$.

3.32 For each $a \in \mathbb{Z}/(3)$, let $C_a = V_{\mathbb{A}^2(\mathbb{Z}/(3))}\left(x^2 + y^2 + a\right)$. (a) List the elements of $\lfloor C_a \rfloor$. (b) Show that the polynomial $x^2 + y^2 + a$ is irreducible. (c) Show that the plane $\mathbb{A}^2(\mathbb{Z}/(3))$ contains exactly twelve lines.

3.33 Let $\ell \subset \mathbb{A}^2$ be a line, and let $p \in \mathbb{A}^2 \setminus \ell$. Show that there is a unique line $\ell'$ which passes through $p$ such that $\ell \cap \ell'$ is empty (we say that $\ell$ and $\ell'$ are parallel).[1]

**Irreducible Polynomials**

3.34 Suppose that $\mathbb{K}$ is algebraically closed. Let $C$ be a conic or cubic curve in $\mathbb{A}^2(\mathbb{K})$ (recall that this means a curve of degree 2 or 3). Show that $C$ is irreducible if and only if it does not contain a line.

[1] This, of course, is Euclid's *fifth postulate*.

3.35 (a) Let $f \in \mathbb{C}[x]$ be a polynomial of odd degree. Show that $y^2 - f$ is irreducible. (Hint: if not, factor $f$ in $\mathbb{C}[x][y]$ as a product of linear terms, and compare coefficients.) (b) Let $f$ and $g$ be polynomials in $\mathbb{C}[x]$ with no common nonconstant factor. Show that $yf + g$ is irreducible. (c) Let $f \in \mathbb{C}[x]$. Show that $y^3 + f$ is reducible if and only if there is some $g \in \mathbb{C}[x]$ such that $f = g^3$. [Bix06, Examples 1.6, 1.8, and 1.9]

**Resultants, Curves and Study's Lemma**

3.36 Let $R$ be a subring of an integral domain $S$. Let $f, g \in R[x]$. Show that $f$ and $g$ have a nonconstant common factor in $R[x]$ if and only if they have a nonconstant common factor in $S[x]$.

3.37 Let $f = x^2 + 3x + 1$ and $g = x^2 - 4x + 1$ be polynomials in $\mathbb{C}[x]$. Do $f$ and $g$ have a common root in $\mathbb{C}$? Find $\operatorname{res}(f, g)$.

3.38 Let $C = \{(r\cos\theta, r\sin\theta) : r = \sin(3\theta)\}$. Show that $C$ is the underlying set of an algebraic curve of $\mathbb{A}^2(\mathbb{R})$. (Hint: use the familiar identities $r^2 = x^2 + y^2$, $x = r\cos\theta$ and $y = r\sin\theta$, as well as the identity $\sin 3\theta = 3\sin\theta - 4\sin^3\theta$.)

3.39 The purpose of this exercise is to give a simplified argument for the main case of Study's lemma, for $n = 2$. Suppose that $\mathbb{K}$ is algebraically closed, that $f, g \in \mathbb{K}[x, y]$, that $f$ is irreducible, that $\deg_y f, \deg_y g > 0$, and that $V_{\mathbb{A}^2}(f) \subseteq V_{\mathbb{A}^2}(g)$.

Show that for all but finitely many $a \in \mathbb{K}$, $\operatorname{res}_y(f, g)(a) = \operatorname{res}_y(f(a, y), g(a, y))$. Also show that if $\deg_y f(a, y) > 0$, then $f(a, y)$ and $g(a, y)$ have a common root in $\mathbb{K}$. Conclude that $\operatorname{res}_y(f, g)$ has infinitely many roots in $\mathbb{K}$. Hence $\operatorname{res}_y(f, g) = 0$; conclude that $f$ divides $g$.

3.40 Let $f$ and $g$ be nonzero polynomials in $\mathbb{K}[x, y]$ which have no common factor. Proposition 3.17 shows that there are polynomials $h, k \in \mathbb{K}[x, y]$ such that $hf + kg \in \mathbb{K}[x] \setminus \{0\}$. We give a proof of this fact which does not use Proposition 3.17.

Let $F = \mathbb{K}(x)$ be the field of formal rational functions (Example 2.35). Show that $f$ and $g$ have no nonconstant common factor in $F[y]$. Conclude that there are elements $\alpha, \beta \in F[y]$ such that $\alpha f + \beta g = 1$ (consult the proof of Proposition 2.25). Use this to show that there are elements $h, k \in \mathbb{K}[x, y]$ such that $hf + kg \in \mathbb{K}[x] \setminus \{0\}$.

3.41 Suppose that $\mathbb{K}$ is algebraically closed. Let $C$ be an algebraic hypersurface of $\mathbb{A}^n$ which has no repeated components. Show that $C$ is irreducible if and only if there are no distinct hypersurfaces $C_1, C_2$ of $\mathbb{A}^n$ such that $\lfloor C \rfloor = \lfloor C_1 \rfloor \cup \lfloor C_2 \rfloor$.

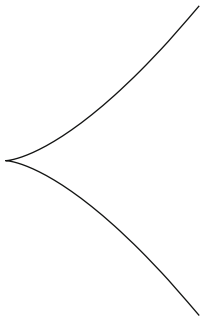


**Fig. 3.2** The cuspidal cubic $y^2 = x^3$

3.42 Suppose that $\mathbb{K}$ is algebraically closed. Show that if $n \geqslant 2$, then every nonempty hypersurface of $\mathbb{A}^n$ contains infinitely many points. In particular, every curve in $\mathbb{A}^2(\mathbb{C})$ contains infinitely many points.

**Parameterisations**

3.43 For the following three rational parameterisations from $\mathbb{R}$ to $\mathbb{R}^2$ find the curve which is parameterised and decide whether the parameterisation is a bijection between $\mathbb{R}$ and the curve. (i) $(t^2 + t^3, t^3 + t^4)$; (ii) $(1 + t, t + t^3)$; (iii) $(t^2/(1 + t^2), t^3/(1 + t^2))$. [Gib98, Example 14.3.1]

3.44 Find a rational parameterisation of the circle $x^2 + y^2 = 5$ in $\mathbb{A}^2(\mathbb{Q})$. (Note that over $\mathbb{Q}$, this circle does not intersect the axes; you need to find a rational point on this circle.)

3.45 Let $f$ and $g$ be polynomials in $\mathbb{K}[t]$, which define a polynomial parameterisation $a \mapsto (f(a), g(a))$ of a curve in $\mathbb{A}^2$. Show that the degree of that curve is at most $\max\{\deg f, \deg g\}$.

**Important Curves**

3.46 For $n \geqslant 1$, let $C_n = V_{\mathbb{A}^2(\mathbb{Q})}(x^n + y^n - 1)$. Show that $C_n \nsubseteq \{(0, \pm 1), (\pm 1, 0)\}$ if and only if there are nonzero integers $a$, $b$ and $c$ such that $a^n + b^n = c^n$. [2]

3.47 Show that the curve $y^2 = x^3$ (Fig. 3.2), called the *cuspidal cubic* curve, is irreducible. Find a rational parameterisation of this curve. (Hint: try the family of lines through the origin.)

[2] Of course, Fermat's last theorem, proved by Wiles and Taylor, states that if $n \geqslant 3$, then there is no such triple of integers.

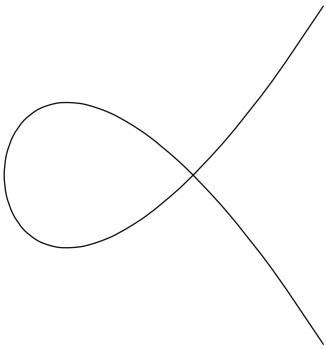


**Fig. 3.3** The nodal cubic $y^2 = x^3 + x^2$

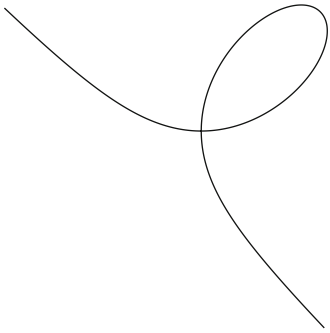


**Fig. 3.4** The folium of Descartes $y^3 + x^3 = 3xy$

3.48 Show that the curve $y^2 = x^2 + x^3$ (Fig. 3.3), called the *nodal cubic* curve, is irreducible. Find a rational parameterisation of this curve.

3.49 Find a rational parameterisation of the *Folium of Descartes*, the curve $x^3 + y^3 = 3xy$ in $\mathbb{A}^2(\mathbb{R})$ (Fig. 3.4).


# Projective Space

# 4

A need to extend the affine plane comes from a consideration of intersections of curves. The simplest case is that of lines. We want any two distinct lines to intersect at a single point. In the affine plane, we have parallel lines. So we add "points at infinity": one for each direction of lines in the affine plane. For example, all vertical lines $x = a$ intersect at the "vertical point at infinity". But note that there is only one vertical point at infinity, not two, as distinct lines should not intersect at more than one point. Adding the vertical point at infinity causes the vertical lines to "wrap around" from $-\infty$ to $\infty$. We can thus envision the projective plane as the result of adding to the affine plane a circle around the plane, at "infinite distance", except that opposite points on that circle are identified.

Topologically, the plane $\mathbb{R}^2$ is identical to the open unit disc. So the real projective plane can be envisioned as the result of taking the closed unit disc and gluing opposite points on the unit circle. This is the same as taking the unit sphere $S^2$ in 3-dimensional space, and gluing all opposite points: first glue together the open northern and southern hemispheres, to obtain the open unit disc; then glue opposite points on the equator as we did with the unit circle. Opposite points on the unit sphere are the points of intersection with the sphere of lines passing through the origin. And so, yet another construction of the projective plane is by taking the collection of all lines in 3-dimensional space passing through the origin, and squeezing each such line to a single point.

This definition gives rise to *homogeneous coordinates* for points on the projective plane: we give a point in the projective plane all the coordinates of nonzero points on the line which was squeezed to obtain that point. So a point in 2-dimensional projective plane has *three* coordinates, but they are not unique: multiplying the coordinates by the same scalar results in the same point. The projective point $(2:2:1)$ is *the same point* as the projective point $(4:4:2)$. The old affine plane is now identified as a subset of the projective plane by the identification $(x, y) \mapsto (1:x:y)$; see Fig. 4.2. The line at infinity consists of the points $(0:x:y)$.



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_4

Homogeneous coordinates allow us to define projective curves (and in general, projective hypersurfaces in higher dimensional projective spaces). However, we cannot take zero sets of just any polynomial. For example, if we're going to use a polynomial $f$ to define a zero-set in the projective plane, it had better be the case that $f(2, 2, 1) = 0$ if and only if $f(4, 4, 2) = 0$. Otherwise there would be ambiguity as to whether the point $(2:2:1) = (4:4:2)$ lies on the curve $f = 0$ or not. The polynomials for which this is never a problem are the *homogeneous* ones: polynomials in which every monomial has the same degree.

The process of *homogenisation* of a polynomial gives the *projective closure* of an affine curve. Consider, for example, the parabola $y = x^2$. In homogeneous coordinates, a point $(a, a^2) = (1:a:a^2)$ on the parabola is the same as $(e:ea:ea^2)$ for any nonzero $e$. So the projective points $(w:x:y)$ with $w \neq 0$ which lie on the parabola are precisely those which satisfy $wy = x^2$. Algebraically, what we did is multiply each monomial of $y - x^2$ by the appropriate power of $w$ so that we get a homogeneous equation. Having done that, we can also set $w = 0$ to get the points at infinity. In the case of the parabola, this forces $x = 0$. Since there is no projective point $(0:0:0)$ (the point $(0, 0, 0)$ does not determine a line through the origin), and all the points $(0:0:b)$ for $b \neq 0$ are actually the same point $(0:0:1)$, we get a single point at infinity which lies on the projective closure of the parabola. That point, by the way, is the vertical point at infinity: the homogenisation of an equation $x = a$ of a vertical line is $x = aw$; now set $w = 0$ to see that $(0:0:1)$ is the point at infinity on the line $x = aw$.

In this chapter we start with the investigation of homogeneous polynomials; we define projective space, and then show how to identify affine space as a subset of projective space. After verifying that Study's lemma holds for projective curves, we also discuss *changes of coordinates*, which are used to simplify equations of curves. These will be useful, for example, in classifying conic curves. We prove a Four Point Lemma, which guarantees the existence of many changes of coordinates. We also consider products of projective spaces, which will be used in the subsequent two chapters. And finally, we introduce the idea of *duality* between points and lines in the projective space. The duality principle allows us to deduce theorems from their duals; an example is Desargues' Theorem.

## 4.1 Homogeneous Polynomials

In this section let $R$ be an integral domain and let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be an $n$-tuple of variables. A polynomial $f \in R[\boldsymbol{x}]$ is *homogeneous* if every monomial which appears in $f$ has the same degree (namely $\deg f$).

Like degree, the notion of homogeneity depends on the choice of variables. For example, the polynomial $3x^2yz + x^3 + y^2xz^4$ is not $x, y, z$-homogeneous but is $x, y$-homogeneous (when we think of it as a polynomial with coefficients from $\mathbb{Z}[z]$).

**Factoring Homogeneous Polynomials**

By grouping together all monomials of a given degree, every polynomial $f \in R[\boldsymbol{x}]$ can be written uniquely as the sum of homogeneous polynomials

$$f = f_{(0)} + f_{(1)} + \cdots + f_{(d)}$$

with $f_{(i)}$ homogeneous of degree $i$. A linear combination of homogeneous polynomials of the same degree $d$ is also homogeneous of degree $d$, or is 0. In particular, if $f \sim g$ then $f$ is homogeneous if and only if $g$ is. If $f$ is homogeneous of degree $d$ and $g$ is homogeneous of degree $e$, then $fg$ is homogeneous of degree $d + e$. In fact, if $f = \sum f_{(i)}$ and $g = \sum g_{(i)}$ (with $f_{(i)}$, $g_{(i)}$ homogeneous of degree $i$) then $(fg)_{(k)} = \sum_{i+j=k} f_{(i)} g_{(j)}$. This shows that the degree formula for products (Proposition 2.9) extends to polynomials in more than one variable: $\deg fg = \deg f + \deg g$. We conclude that if $R$ is a field, $f$ and $g$ are nonconstant, and $f$ is a proper divisor of $g$, then $\deg f < \deg g$.

The following proposition will be used to define projective hypersurfaces as multisets.

**Proposition 4.1** *A factor of a homogeneous polynomial is homogeneous.*

***Proof*** Let $f, g \in R[\boldsymbol{x}]$, and suppose that $f$ is not homogeneous. Write $f = f_{(b)} + f_{(b+1)} + \cdots + f_{(d)}$ where $f_{(i)}$ is homogeneous of degree $i$, and $f_{(b)}$, $f_{(d)}$ are nonzero (and $b < d$). Similarly let $g = g_{(c)} + \cdots + g_{(e)}$ where $c \leqslant e$. Then $(fg)_{(b+c)} = f_{(b)} g_{(c)}$ and $(fg)_{(d+e)} = f_{(d)} g_{(e)}$, which are both nonzero, and $b + c < d + e$, so $fg$ is not homogeneous. □

## A Characterisation of Homogeneity

**Proposition 4.2** *Suppose that $R$ is infinite. The following are equivalent for a polynomial $f \in R[\boldsymbol{x}]$ and a natural number $d$:*

(1) *$f$ is the zero polynomial, or is homogeneous of degree $d$.*
(2) *$f(tx_1, \ldots, tx_n) = t^d f$ (as elements of the ring $R[\boldsymbol{x}, t]$).*
(3) *For all $\boldsymbol{a} \in R^n$ and all $\lambda \in R$, $f(\lambda \boldsymbol{a}) = \lambda^d f(\boldsymbol{a})$.*

***Proof*** The equivalence of (2) and (3) follows from Proposition 2.18: the polynomial $f(tx_1, \ldots, tx_n) - t^d f$ is the zero polynomial if and only if it evaluates to 0 on all tuples $(\boldsymbol{a}, \lambda) \in R^{n+1}$.

For any monic monomial $\boldsymbol{x}^{\boldsymbol{m}} = x_1^{m_1} \cdots x_n^{m_n}$ of degree $d = m_1 + \cdots + m_n$, $(t\boldsymbol{x})^{\boldsymbol{m}} = t^d \boldsymbol{x}^{\boldsymbol{m}}$, and the property extends to taking linear combinations; this shows that (1) implies (2).

Let $f \in R[\boldsymbol{x}]$; as above write $f = f_{(0)} + f_{(1)} + \cdots + f_{(e)}$ with each $f_{(i)}$ homogeneous of degree $i$. We just observed that for all $i$, $f_{(i)}(t\boldsymbol{x}) = t^i f_{(i)}$; so $f(t\boldsymbol{x}) = \sum_{i \leqslant e} t^i f_{(i)}$. If (2) holds then $t^d f = \sum_{i \leqslant e} t^i f_{(i)}$. This holds in $R[\boldsymbol{x}, t] =$

$R[\boldsymbol{x}][t]$, and so the coefficients of each $t^i$ on both sides must be equal, which means that $f = f_{(d)}$ and $f_{(i)} = 0$ for $i \neq d$; so $f$ is homogeneous of degree $d$ if $f_{(d)} \neq 0$, otherwise $f = 0$.

Note that for the equivalence of (1) and (2), we didn't need $R$ to be infinite. □

## 4.2 Projective Space

Let $\mathbb{K}$ be a field, and let $n \geqslant 1$.

**Definition 4.3** *Projective space* $\mathbb{P}^n(\mathbb{K})$ is the collection of all 1-dimensional linear subspaces of $\mathbb{K}^{n+1}$.

Note! the *elements* of $\mathbb{P}^n(\mathbb{K})$ (which we call *points*) are *subsets* of $\mathbb{K}^{n+1}$. The projective *plane* is the case $n = 2$.

Every nonzero tuple $\boldsymbol{a} \in \mathbb{K}^{n+1}$ is an element of exactly one 1-dimensional linear subspace of $\mathbb{K}^{n+1}$, namely the linear subspace $\langle \boldsymbol{a} \rangle$ it generates; two nonzero tuples $\boldsymbol{a}$ and $\boldsymbol{b}$ generate the same 1-dimensional subspace if and only if they are nonzero scalar multiples of each other.

**Definition 4.4** The *projection map* $\pi_n \colon \mathbb{K}^{n+1} \setminus \{\mathbf{0}\} \to \mathbb{P}^n(\mathbb{K})$ is defined by letting $\pi_n(\boldsymbol{a}) = \langle \boldsymbol{a} \rangle$.

The map $\pi_n$ is onto $\mathbb{P}^n(\mathbb{K})$; if $|\mathbb{K}| > 2$ then $\pi_n$ is not one-to-one.

*Notation 4.5* If $\boldsymbol{a} = (a_0, a_1, \ldots, a_n)$ is a nonzero element of $\mathbb{K}^{n+1}$, then we also denote $\pi_n(\boldsymbol{a})$ by $(a_0 \colon a_1 \colon \cdots \colon a_n)$. We call the tuple $\boldsymbol{a}$ a *presentation* of the point $\pi_n(\boldsymbol{a})$. Summing up, projective space $\mathbb{P}^n(\mathbb{K})$ is the collection of points $(a_0 \colon a_1 \colon \cdots \colon a_n)$, where $a_0, a_1, \ldots, a_n \in \mathbb{K}$ are not all zero, and the point $(a_0 \colon a_1 \colon \cdots \colon a_n)$ *equals* the point $(b_0 \colon b_1 \colon \cdots \colon b_n)$ if and only if there is some nonzero scalar $\lambda \in \mathbb{K}$ such that $b_0 = \lambda a_0, b_1 = \lambda a_1, \ldots, b_n = \lambda a_n$. «

We again fix $\mathbb{K}$, and write $\mathbb{P}^n$ for $\mathbb{P}^n(\mathbb{K})$. As in the affine case, irreducible algebraic hypersurfaces of $\mathbb{P}^n$ will be defined to be the zero sets of polynomials. Because of the ambiguity of presentation of points in projective space, polynomials do not define functions on $\mathbb{P}^n$, and some do not have invariant zero sets: there are polynomials $f \in \mathbb{K}[\boldsymbol{x}]$ and nonzero $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{K}^{n+1}$ such that $\pi_n(\boldsymbol{a}) = \pi_n(\boldsymbol{b})$, but such that $f(\boldsymbol{a}) = 0$ and $f(\boldsymbol{b}) \neq 0$. By Proposition 4.2, this does not happen if $f$ is homogeneous.

**Definition 4.6** Let $f \in \mathbb{K}[x_0, \ldots, x_n]$ be an irreducible homogeneous polynomial. The projective hypersurface defined by $f$ is

$$V_{\mathbb{P}^n}(f) = \left\{(a_0 : a_1 : \cdots : a_n) \in \mathbb{P}^n \ : \ f(a_0, a_1, \ldots, a_n) = 0\right\}.$$

We write $V_{\mathbb{P}^n(\mathbb{K})}(f)$ when we need to clarify which field we are working with. Note that $V_{\mathbb{P}^n}(f)$ is the image $\pi_n[V_{\mathbb{A}^{n+1}}(f)]$.

To extend to reducible homogeneous polynomials, we again use irreducible factorisations. For this we use the fact that factors of homogeneous polynomials are homogeneous (Proposition 4.1); this ensures that if $f$ is homogeneous, then $V_{\mathbb{P}^n}(h)$ is defined for each irreducible factor $h$ of $f$. Thus, if $f \in \mathbb{K}[x_0, x_1, \ldots, x_n]$ and $[f_1, \ldots, f_m]$ is an irreducible factorisation of $f$, then we define

$$V_{\mathbb{P}^n}(f) = V_{\mathbb{P}^n}(f_1) + V_{\mathbb{P}^n}(f_2) + \cdots + V_{\mathbb{P}^n}(f_m).$$

This does not depend on the choice of irreducible factorisation, because as in the affine case, if $g \sim h$ are irreducible and homogeneous then $V_{\mathbb{P}^n}(g) = V_{\mathbb{P}^n}(h)$. If $f$ is a nonzero constant then $V_{\mathbb{P}^n}(f) = \varnothing$. We let $V_{\mathbb{P}^n}(0) = \mathbb{P}^n$.

A multiset of points of $\mathbb{P}^n$ is an algebraic hypersurface of $\mathbb{P}^n$ if it is $V_{\mathbb{P}^n}(f)$ for some nonzero homogeneous polynomial $f$. For any homogeneous polynomial $f \in \mathbb{K}[x_0, \ldots, x_n]$, the underlying set of the hypersurface defined by $f$ is

$$\lfloor V_{\mathbb{P}^n}(f) \rfloor = \left\{(a_0 : a_1 : \cdots : a_n) \in \mathbb{P}^n \ : \ f(a_0, a_1, \ldots, a_n) = 0\right\}.$$

If $\mathbb{K}$ is infinite then no hypersurface is the entire projective space:

**Proposition 4.7** *If $\mathbb{K}$ is infinite then for any nonzero homogeneous $f \in \mathbb{K}[x_0, \ldots, x_n]$, $\lfloor V_{\mathbb{P}^n}(f) \rfloor \neq \mathbb{P}^n$.*

***Proof*** Since $\mathbb{K}$ is infinite, so is $\mathbb{K} \setminus \{0\}$. Proposition 2.19 says that there is some $\boldsymbol{a} \in (\mathbb{K} \setminus \{0\})^{n+1}$ with $f(\boldsymbol{a}) \neq 0$. Then $\boldsymbol{a} \neq \mathbf{0}$, so $\pi_n(\boldsymbol{a})$ is defined; and $\pi_n(\boldsymbol{a}) \notin V_{\mathbb{P}^n}(f)$. □

As in the affine case, we see that if $g$ is homogeneous and $f$ divides $g$ then $V_{\mathbb{P}^n}(f) \subseteq V_{\mathbb{P}^n}(g)$. Study's Lemma holds in the projective context as well. We show this in two steps.

**Lemma 4.8** *Suppose that $\mathbb{K}$ is algebraically closed. Let $f \in \mathbb{K}[x_0, \ldots, x_n]$ be nonzero and homogeneous. Then $V_{\mathbb{P}^n}(f) \neq \varnothing$ if and only if $f$ is nonconstant.*

***Proof*** If $f$ is a nonzero constant then $V_{\mathbb{P}^n}(f) = \varnothing$. Suppose that $\deg f > 0$. There is some $i \leqslant n$ such that $\deg_{x_i} f > 0$; for notational convenience we assume $i = 0$. Let $d = \deg_{x_0} f$. Write $f = \sum_{k \leqslant d} f_k x_0^k$, with $f_k \in \mathbb{K}[x_1, \ldots, x_n]$ and $f_d \neq 0$. Since $f_d \neq 0$, as in the proof of Proposition 4.7, there is some nonzero tuple $\boldsymbol{a} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ such that $f_d(\boldsymbol{a}) \neq 0$ (an algebraically closed field is

infinite, Proposition 2.29). So the polynomial $f(x_0, \boldsymbol{a})$ has degree $d > 0$. Since $\mathbb{K}$ is algebraically closed, there is some $b \in \mathbb{K}$ such that $f(b, \boldsymbol{a}) = 0$. The tuple $(b, \boldsymbol{a})$ is not the zero tuple, so $\pi_n(b, \boldsymbol{a}) \in V_{\mathbb{P}^n}(f)$. □

**Projective Study's Lemma** *Suppose that $\mathbb{K}$ is an algebraically closed field. Let $f, g \in \mathbb{K}[x_0, x_1, \ldots, x_n]$ be nonzero homogeneous polynomials. Then*

$$f \text{ divides } g \iff V_{\mathbb{P}^n}(f) \subseteq V_{\mathbb{P}^n}(g).$$

***Proof*** We first argue as we did for the first part of the proof of Study's lemma to see that it is sufficient to show that if $V_{\mathbb{P}^n}(f) \subseteq V_{\mathbb{P}^n}(g)$ and $f$ is irreducible, then $f$ divides $g$. The conclusion $f \mid g$ will be obtained from Study's lemma, once we show that $V_{\mathbb{A}^{n+1}}(f) \subseteq V_{\mathbb{A}^{n+1}}(g)$. As $f$ is irreducible, $V_{\mathbb{A}^{n+1}}(f) = \left\lfloor V_{\mathbb{A}^{n+1}}(f) \right\rfloor$, and so what we need to show is that for all $\boldsymbol{a} \in \mathbb{A}^{n+1}$, if $f(\boldsymbol{a}) = 0$, then $g(\boldsymbol{a}) = 0$. Since $f$ and $g$ are homogeneous, the fact that $V_{\mathbb{P}^n}(f) \subseteq V_{\mathbb{P}^n}(g)$ ensures that for all *nonzero* $\boldsymbol{a} \in \mathbb{A}^{n+1}$, if $f(\boldsymbol{a}) = 0$ then $g(\boldsymbol{a}) = 0$. The inclusion is proved once we show that $g(\mathbf{0}) = 0$. The point is that if $h$ is nonzero and homogeneous then $h(\mathbf{0}) = 0$ if and only if $\deg h > 0$. Since we assume that $f$ is irreducible, it is nonconstant, and so $V_{\mathbb{P}^n}(f)$ is nonempty (here we use Lemma 4.8); whence $V_{\mathbb{P}^n}(g)$ is nonempty, from which we conclude that $g$ is nonconstant. □

The fact that Study's lemma holds in the projective context allows us to draw the same conclusions as we did in the affine case. We call a projective hypersurface $C$ *irreducible* if the polynomials defining $C$ are irreducible, and we let $\deg C$ be the degree of the polynomials defining $C$.

We can again let, for a hypersurface $C$ in $\mathbb{P}^n$, the *irreducible components* of $C$ be the irreducible hypersurfaces $D$ contained in $C$; the irreducible components of $V_{\mathbb{P}^n}(f)$ are the hypersurfaces $V_{\mathbb{P}^n}(g)$ where $g$ is an irreducible factor of $f$. The proofs of Propositions 3.21 and 3.22 carry over to the projective case with no change, using the projective version of Study's lemma.

*Notation 4.9* When $n = 2$ we rename the variables $(x_0, x_1, x_2)$ as $(w, x, y)$. When $n = 1$ we rename the variables $(x_0, x_1)$ as $(w, x)$. «

## 4.3 Projective Lines and Maps

The simplest irreducible hypersurfaces are hyperplanes. These are a special kind of projective subspaces.

**Definition 4.10** A *projective subspace* of $\mathbb{P}^n$ is a subset of $\mathbb{P}^n$ of the form $\pi_n[W]$, where $W$ is a linear subspace of $\mathbb{K}^{n+1}$.

When it is clear that we are dealing with projective subspaces of $\mathbb{P}^n$, we refer to them simply as *subspaces*. If $W \neq U$ are distinct linear subspaces of $\mathbb{K}^{n+1}$ then

$\pi_n[W] \neq \pi_n[U]$, and so we can unambiguously define the *dimension* of a projective subspace $\pi_n[W]$ to be $\dim W - 1$.

The 0-dimensional subspaces of $\mathbb{P}^n$ are the points in $\mathbb{P}^n$. A 1-dimensional subspace of $\mathbb{P}^n$ is called a *line*, and a 2-dimensional subspace is called a *plane*. Because the only $(n+1)$-dimensional subspace of $\mathbb{K}^{n+1}$ is $\mathbb{K}^{n+1}$ itself, the only $n$-dimensional subspace of $\mathbb{P}^n$ is $\mathbb{P}^n$ itself. An $(n-1)$-dimensional subspace of $\mathbb{P}^n$ is called a *hyperplane* of $\mathbb{P}^n$. The hyperplanes of $\mathbb{P}^2$ are the lines of $\mathbb{P}^2$.

Homogeneous linear polynomials in $\mathbb{K}[x_0, \ldots, x_n]$ are irreducible, so all algebraic hypersurfaces of $\mathbb{P}^n$ of degree 1 are irreducible. Homogeneous linear polynomials do not have constant terms, and so are of the form $a_0x_0 + a_1x_1 + \cdots + a_nx_n$, with not all $a_i$ zero. The functions from $\mathbb{K}^{n+1}$ to $\mathbb{K}$ defined by homogeneous linear polynomials are precisely the linear maps from $\mathbb{K}^{n+1}$ onto $\mathbb{K}$ (compare with Sect. 3.4). By Corollary 2.63, the zero sets $V_{\mathbb{A}^{n+1}}(f)$ of linear homogeneous polynomials $f$ are precisely the $n$-dimensional linear subspaces of $\mathbb{A}^{n+1}$. Taking the image under $\pi_n$, we get:

**Proposition 4.11** *The algebraic hypersurfaces of $\mathbb{P}^n$ of degree 1 are precisely the hyperplanes of $\mathbb{P}^n$.*

Hence, the lines in $\mathbb{P}^2$ are the subsets of $\mathbb{P}^2$ defined by homogeneous linear polynomials in $\mathbb{K}[w, x, y]$, that is, by polynomials of the form $ew + ax + by$ with $(e, a, b) \in \mathbb{K}^3 \setminus \{\mathbf{0}\}$.

**Theorem 4.12** *Every two distinct points in $\mathbb{P}^n$ lie on a unique line.*

This can be extended to subspaces of a higher dimension, see Exercise 4.60.

***Proof*** Let $p, q \in \mathbb{P}^n$ be distinct points, which recall, technically, are both 1-dimensional subspaces of $\mathbb{K}^{n+1}$. If $W$ is a linear subspace of $\mathbb{K}^{n+1}$ then $p \in \pi_n[W]$ if and only if $p \subseteq W$ (as subspaces of $\mathbb{K}^{n+1}$).

There is a unique 2-dimensional linear subspace of $\mathbb{K}^{n+1}$ which contains both 1-dimensional subspaces $p$ and $q$, namely the subspace $W$ generated by their union (or by any two nonzero elements one from $p$ and one from $q$). Then $\pi_n[W]$ is the unique line passing through both $p$ and $q$. □

We denote the unique line that passes through points $p$ and $q$ in $\mathbb{P}^n$ by $\overline{pq}$.

*Exercise 4.13* Let $p = (p_w : p_x : p_y)$ and $q = (q_w : q_x : q_y)$ be distinct points in $\mathbb{P}^2$. Show that the equation of the line $\overline{pq}$ is given by

$$\det \begin{pmatrix} w & x & y \\ p_w & p_x & p_y \\ q_w & q_x & q_y \end{pmatrix} = 0.$$

«

The following theorem shows that the projective plane $\mathbb{P}^2$ satisfies at least one of the properties which motivated its definition.

**Theorem 4.14** *Any two distinct lines in $\mathbb{P}^2$ intersect in a unique point.*

***Proof*** Let $\ell_1$ and $\ell_2$ be distinct lines in $\mathbb{P}^2$. Let $W_1$ and $W_2$ be the 2-dimensional subspaces of $\mathbb{K}^3$ such that $\ell_1 = \pi_2[W_1]$ and $\ell_2 = \pi_2[W_2]$. Let $p = W_1 \cap W_2$. $p$ is a linear subspace of $\mathbb{K}^3$; since it is the intersection of two distinct 2-dimensional subspaces of $\mathbb{K}^3$, its dimension is at most 1; however two 2-dimensional subspaces of $\mathbb{K}^3$ cannot have trivial intersection as $2 + 2 > 3$ (Proposition 2.61). Hence $\dim p = 1$; $p$ is the unique 1-dimensional subspace of $\mathbb{K}^3$ which is a subset of both $W_1$ and $W_2$, and so $p$ is the unique point in $\ell_1 \cap \ell_2$. □

### 4.3.1 Projective Maps

Let $T : \mathbb{K}^{n+1} \to \mathbb{K}^{m+1}$ be an *injective* linear transformation. Then $T$ maps $k$-dimensional subspaces of $\mathbb{K}^{n+1}$ to $k$-dimensional subspaces of $\mathbb{K}^{m+1}$. In particular, it maps lines through $\mathbf{0}$ to lines through $\mathbf{0}$, and so induces an injective map $\bar{T} : \mathbb{P}^n \to \mathbb{P}^m$; see Fig. 4.1.

A *projective map* is a map of the form $\bar{T}$ for some injective linear $T$. The image of a projective map from $\mathbb{P}^n$ to $\mathbb{P}^m$ is an $n$-dimensional projective subspace of $\mathbb{P}^m$. We usually use the notation $\boldsymbol{\psi}$ to denote a linear map which induces a projective map $\psi$, and call $\boldsymbol{\psi}$ a *linear presentation* of $\psi$.

*Exercise 4.15* A projective map $\psi$ will have many linear presentations, but like points, they are all nonzero constant multiples of each other. Let $\boldsymbol{\psi}$ and $\boldsymbol{\varphi}$ be linear presentations of projective maps $\psi, \varphi : \mathbb{P}^n \to \mathbb{P}^m$. Show that $\psi = \varphi$ if and only if $\boldsymbol{\varphi} = \lambda\boldsymbol{\psi}$ for some nonzero $\lambda \in \mathbb{K}$.

(Hint: suppose that $\psi = \varphi$. For each nonzero $\boldsymbol{p} \in \mathbb{K}^{n+1}$ there is a unique nonzero $\lambda(\boldsymbol{p}) \in \mathbb{K}$ such that $\boldsymbol{\varphi}(\boldsymbol{p}) = \lambda(\boldsymbol{p})\boldsymbol{\psi}(\boldsymbol{p})$. We need to show that $\lambda(\boldsymbol{p})$ is constant. To show that $\lambda(\boldsymbol{p}) = \lambda(\boldsymbol{q})$ for nonzero $\boldsymbol{p}$ and $\boldsymbol{q}$, consider two cases, depending on

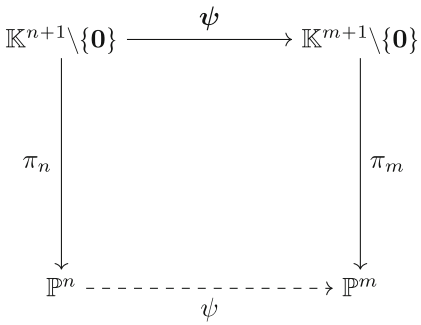



**Fig. 4.1** A projective map

whether $\boldsymbol{p}$ and $\boldsymbol{q}$ are linearly independent or not. If yes, consider $\lambda(\boldsymbol{p})$, $\lambda(\boldsymbol{q})$, and $\lambda(\boldsymbol{p}+\boldsymbol{q})$.) «

Suppose that $\boldsymbol{\psi}$ is a linear presentation of $\psi$. We write $\boldsymbol{\psi} = (\boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \ldots \boldsymbol{\psi}_m)$; each $\boldsymbol{\psi}_i$ is a linear map from $\mathbb{K}^{n+1}$ to $\mathbb{K}$, and so is defined by a linear homogeneous polynomial $\psi_i \in \mathbb{K}[t_0, t_1, \ldots, t_n]$; it is nonzero since $\boldsymbol{\psi}$ is injective. Informally, we write $\psi = (\psi_0 : \psi_1 : \cdots : \psi_m)$.

**Definition 4.16** A *projective linear parameterisation* is a projective map $\psi : \mathbb{P}^1 \to \mathbb{P}^m$.

The images of projective linear parameterisations are projective lines; if $\ell$ is the range of a projective linear parameterisation $\psi$ then we say that $\psi$ is a parameterisation *of* $\ell$. Compare with Definition 3.27. We usually use the variables $s, t$ for the polynomials defining a linear presentation a parameterisation. A triple $\boldsymbol{\psi} = (\psi_w, \psi_x, \psi_y)$ of linear homogeneous polynomials in $\mathbb{K}[s, t]$ is a linear presentation of a projective linear parameterisation if and only if the polynomials are not all scalar multiples of each other.

*Example 4.17* Let $\boldsymbol{p}$ and $\boldsymbol{q}$ be parameterisations of distinct points $p$ and $q$ in $\mathbb{P}^n$. Then $\boldsymbol{\psi}_{\boldsymbol{p},\boldsymbol{q}}(s, t) = (s\boldsymbol{p} + t\boldsymbol{q})$ is a linear presentation of a parameterisation $\psi_{\boldsymbol{p},\boldsymbol{q}}$ of the line $\overline{p\,q}$. Every projective linear presentation of $\overline{p\,q}$ is of this form.

Note that $\psi_{\boldsymbol{p},\boldsymbol{q}}$ is not determined by $p$ and $q$, as we can choose presentations of $p$ and $q$ which differ by different constant multiples. That is, unlike the affine case (Example 3.28), there are many linear parameterisations of $\overline{p\,q}$ which map $(1:0)$ to $p$ and $(0:1)$ to $q$. «

## 4.4 Embedding Affine Space into Projective Space

In this section we see how we can view projective space as an extension of affine space. Let

$$H_\infty = V_{\mathbb{P}^n}(x_0)$$

and let

$$U = \mathbb{P}^n \setminus H_\infty = \left\{(a_0 : a_1 : \ldots : a_n) \in \mathbb{P}^n \;:\; a_0 \neq 0\right\}.$$

Define a map $\rho : \mathbb{A}^n \to \mathbb{P}^n$ by mapping

$$\rho(a_1, \ldots, a_n) = (1 : a_1 : a_2 : \cdots : a_n).$$

The map $\rho$ is a bijection between $\mathbb{A}^n$ and $U$: if $a_0 \neq 0$ then

$$(a_0 : a_1 : a_2 : \cdots : a_n) = \left(1 : \frac{a_1}{a_0} : \frac{a_2}{a_0} : \cdots : \frac{a_n}{a_0}\right)$$

so $\rho$ is onto $U$. If $\rho(\boldsymbol{a}) = \rho(\boldsymbol{b})$ then $(1, \boldsymbol{b}) = \lambda(1, \boldsymbol{a})$ for some nonzero $\lambda \in \mathbb{K}$; but then $\lambda = 1$ so $\boldsymbol{b} = \boldsymbol{a}$.

The map $\rho$ preserves algebraic hypersurfaces. To see this, we need the notions of *homogenisation* and *dehomogenisation* of polynomials.

**Definition 4.18** Let $R$ be a unique factorisation domain; let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a sequence of variables, and let $x$ be another variable.

For $f \in R[\boldsymbol{y}, x]$ we let $f^{\flat x}$ be the result of substituting 1 for $x$ in $f$:

$$f^{\flat x} = f(\boldsymbol{y}, 1).$$

For $f \in R[\boldsymbol{y}]$, write $f$ as the sum of homogeneous polynomials $f_{(d)} + f_{(d-1)} + \cdots + f_{(1)} + f_{(0)}$ where $d = \deg f$. We let

$$f^{\sharp x} = f_{(d)} + x f_{(d-1)} + x^2 f_{(d-2)} + \cdots + x^d f_{(0)}.$$

*Exercise 4.19* Let $f \in R[\boldsymbol{y}]$ and let $d = \deg f$. Show that in the field of formal rational functions $R(\boldsymbol{y}, x)$ (Example 2.35),

$$f^{\sharp x} = x^d f\left(\frac{y_1}{x}, \cdots, \frac{y_n}{x}\right).$$

«

**Proposition 4.20**

(a) *The map $g \mapsto g^{\flat x}$ is a ring homomorphism from $R[\boldsymbol{y}, x]$ to $R[\boldsymbol{y}]$. The map $f \mapsto f^{\sharp x}$ from $R[\boldsymbol{y}]$ to $R[\boldsymbol{y}, x]$ is not a homomorphism, but preserves multiplication: $(fg)^{\sharp x} = f^{\sharp x} g^{\sharp x}$.*

(b) *For all $f \in R[\boldsymbol{y}]$, $f^{\sharp x}$ is $x, \boldsymbol{y}$-homogeneous of degree $\deg f$.*

(c) *For all $f \in R[\boldsymbol{y}]$, $(f^{\sharp x})^{\flat x} = f$. If $g \in R[\boldsymbol{y}, x]$ is homogeneous and not divisible by $x$, then $(g^{\flat x})^{\sharp x} = g$. The range of the map $f \mapsto f^{\sharp x}$ is the collection of homogeneous polynomials in $R[\boldsymbol{y}, x]$ which are not divisible by $x$.*

(d) *For all $f \in R[\boldsymbol{y}]$, the nonconstant factors of $f^{\sharp x}$ are precisely the polynomials $h^{\sharp x}$ where $h$ is a nonconstant factor of $f$. A polynomial $f \in R[\boldsymbol{y}]$ is irreducible if and only if $f^{\sharp x}$ is irreducible.*

***Proof*** Most of these are straightforward and we omit many details. For substitution being a ring homomorphism, see Proposition 2.11. For a quick proof that $f \mapsto f^{\sharp x}$

is multiplicative, use Exercise 4.19; for $f, g \in R[\boldsymbol{y}]$,

$$(fg)^{\sharp x} = x^{\deg(fg)}(fg)\left(\tfrac{y_1}{x}, \ldots, \tfrac{y_n}{x}\right) = \\ x^{\deg f} f\left(\tfrac{y_1}{x}, \ldots, \tfrac{y_n}{x}\right) x^{\deg g} g\left(\tfrac{y_1}{x}, \ldots, \tfrac{y_n}{x}\right) = f^{\sharp x} \cdot g^{\sharp x}.$$

The main point really is (b); the polynomial $f^{\sharp x}$ is defined to be the sum of homogeneous polynomials, all of degree $\deg f$. For (c), let $g \in R[\boldsymbol{y}, x]$ be homogeneous of degree $d$; write $g = g_d + g_{d-1}x + \cdots + g_1 x^{d-1} + g_0 x^d$, where $g_i \in R[\boldsymbol{y}]$. Since $g$ is homogeneous, each $g_i$ is homogeneous of degree $i$. $g^{\flat x} = \sum_i g_i$, so $(g^{\flat x})_{(i)} = g_i$. If $x$ does not divide $g$ then $g_d \neq 0$, and this shows that $\deg g^{\flat x} = d$, from which we conclude that $(g^{\flat x})^{\sharp x} = g$.

We prove (d). Let $f \in R[\boldsymbol{y}]$, and let $h$ be a divisor of $f^{\sharp x}$. By Proposition 4.1, $h$ is homogeneous; since $x$ does not divide $f^{\sharp x}$, it does not divide $h$ either. By (c), $h = (h^{\flat x})^{\sharp x}$. Since $g \mapsto g^{\flat x}$ is a ring homomorphism, $h^{\flat x}$ divides $(f^{\sharp x})^{\flat x} = f$. On the other hand, if $g$ divides $f$ then (a) ensures that $g^{\sharp x}$ divides $f^{\sharp x}$. Thus, the divisors of $f^{\sharp x}$ are precisely the polynomials of the form $g^{\sharp x}$ for divisors $g$ of $f$.

For a constant $a \in R$ we have $a^{\sharp x} = a$. Since the units of $R[\boldsymbol{y}]$ and of $R[\boldsymbol{y}, x]$ coincide with the units of $R$, we see that $g \mapsto g^{\sharp x}$ maps units to units and nonunits to nonunits. It follows that $g \in R[\boldsymbol{y}]$ is irreducible if and only if $g^{\sharp x}$ is irreducible. The map $g \mapsto g^{\sharp x}$ preserves degrees, so the nonconstant factors of $f$ are mapped to the nonconstant factors of $f^{\sharp x}$. □

*Remark 4.21* Even though $f \mapsto f^{\sharp x}$ is not a homomorphism it will be useful to see how it treats addition. Let $f, g \in R[\boldsymbol{y}]$; let $d = \deg f$ and $e = \deg g$. If $d > e$ then $(f+g)^{\sharp x} = f^{\sharp x} + x^{d-e} g^{\sharp x}$. If $d = e$ then there may be cancellations; possibly $c = \deg(f+g) < d$. In this case $x^{d-c}(f+g)^{\sharp x} = f^{\sharp x} + g^{\sharp x}$. «

We return to working over a field $\mathbb{K}$. Fixing variables $x_0, x_1, \ldots, x_n$, we write $f^\flat$ for $f^{\flat x_0}$ and $f^\sharp$ for $f^{\sharp x_0}$. When $\mathbb{K}$ is algebraically closed, we extend this notation to algebraic hypersurfaces: For an algebraic hypersurface $C$ in $\mathbb{P}^n$, we let $C^\flat = V_{\mathbb{A}^n}(g^\flat)$, where $g$ defines $C$; Projective Study's Lemma implies that this does not depend on the choice of $g$. Similarly, for an algebraic hypersurface $D$ in $\mathbb{A}^n$, we let $D^\sharp = V_{\mathbb{P}^n}(f^\sharp)$, where $f$ defines $D$.[1] In what follows, we use the notation $C^\flat$ and $D^\sharp$ even when $\mathbb{K}$ is not algebraically closed. This is for clarity of presentation; when $\mathbb{K}$ is not algebraically closed, we should replace hypersurfaces by polynomials.

Let $C$ be an irreducible algebraic hypersurface in $\mathbb{P}^n$, other than $H_\infty$. Then $x_0$ does not divide $g$ defining $C$. By Proposition 4.20, $C^\flat$ is irreducible. Then by definition of $g^\flat$ and $\rho$,

$$\rho[C^\flat] = C \cap U.$$

[1] Algebraic geometry does not use well temperament: $C^\sharp \neq D^\flat$.

Now let $C$ be any algebraic hypersurface in $\mathbb{P}^n$. Let $C'$ be the result of removing from $C$ any copies of $H_\infty$ that it may contain. Then $C \restriction_U = C' \restriction_U$. Here recall that by $D \restriction_U$ we mean that we remove from $D$ the points that are not in $U$, but other points retain their multiplicity. Also, $C^\flat = C'^\flat$. (In terms of polynomials, if $g$ defines $C$ and $C$ contains $e$ many copies of $H_\infty$, then $g = x_0^e g'$, where $g'$ defines $C'$, and $x_0$ does not divide $g'$; $g^\flat = g'^\flat$).

Let $D_1, D_2, \ldots, D_m$ be the irreducible components of $C'$. Then by Proposition 4.20, $D_1^\flat, D_2^\flat, \ldots, D_m^\flat$ are the irreducible components of $C^\flat = C'^\flat$. From $\rho[D_i^\flat] = D_i \cap U$ we get

$$\rho[C^\flat] = \rho[D_1^\flat] + \cdots + \rho[D_m^\flat] = (D_1 \cap U) + \cdots + (D_m \cap U) = C \restriction_U$$

(here we extend $\rho$ to multisets of elements of $\mathbb{A}^n$, and agree that multiplicities are preserved).

This allows us to *identify* $\mathbb{A}^n$ with $U$ using the map $\rho$. Having performed this identification, $\rho[C^\flat] = C \restriction_U$ is reformulated as

$$C^\flat = C \restriction_{\mathbb{A}^n} .$$

Thus the restriction to $\mathbb{A}^n$ of an algebraic hypersurface of $\mathbb{P}^n$ is an algebraic hypersurface of $\mathbb{A}^n$. In the other direction, Proposition 4.20 implies that for any algebraic hypersurface $D$ of $\mathbb{A}^n$, $(D^\sharp)^\flat = D$, and so

$$D = D^\sharp \restriction_{\mathbb{A}^n} .$$

We call $D^\sharp$ the *projective closure* (or *projective completion*) of $D$. It is obtained by adding to $D$ points from $H_\infty$. Summing up:

**Proposition 4.22** *The algebraic hypersurfaces of $\mathbb{A}^n$ are precisely the restrictions to $\mathbb{A}^n$ of the algebraic hypersurfaces of $\mathbb{P}^n$.* □

The hyperplace $H_\infty$ is called the *hyperplane at infinity*, so we say that $D^\sharp$ is obtained from $D$ by adding points at infinity.

*Example 4.23* As discussed in the introduction to this chapter, the projective closure of the parabola $y = x^2$ is the projective curve $wy = x^2$, and the extra point at infinity is $(0:0:1)$. «

*Exercise 4.24* Suppose that $\mathbb{K}$ is algebraically closed. Let $D$ be an algebraic hypersurface in $\mathbb{A}^n$. Show that $D^\sharp$ is the unique algebraic hypersurface $C$ in $\mathbb{P}^n$ such that $D = C \restriction_{\mathbb{A}^n}$ and $H_\infty$ is not a component of $C$. «

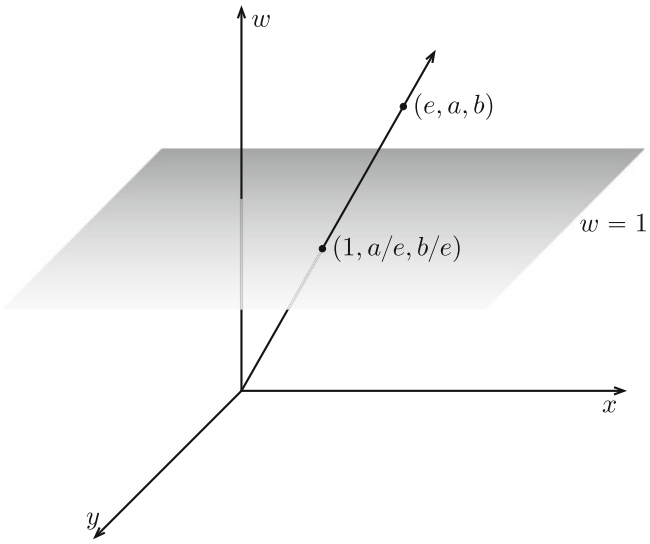



**Fig. 4.2** Affine plane embedded into projective plane. Each line through the origin is a point of $\mathbb{P}^2$. Those lines which are not on the $xy$ plane are points in $\mathbb{A}^2$, which we identify with the plane $w = 1$. The lines through the origin on the $xy$ plane are the points at infinity

*Exercise 4.25* Show that the map $(a_1 : a_2 : \cdots : a_n) \mapsto (0 : a_1 : a_2 : \cdots : a_n)$ is a projective map which is a bijection between $\mathbb{P}^{n-1}$ and the hyperplane $H_\infty$ of $\mathbb{P}^n$. Show that the images in $H_\infty$ of algebraic hypersurfaces of $\mathbb{P}^{n-1}$ are precisely the restrictions to $H_\infty$ of the algebraic hypersurfaces of $\mathbb{P}^n$ which do not contain $H_\infty$. «

### Vertical and Horizontal Projective Lines

Primarily we are interested in the plane: $n = 2$. We can visualise the embedding of the affine plane into the projective plane as in Fig. 4.2. As mentioned above (Notation 4.9), in this case we name $(x_0, x_1, x_2)$ by $(w, x, y)$. A point $(a, b) \in \mathbb{A}^2$ is identified with the point $(1 : a : b) \in \mathbb{P}^2$. Instead of $H_\infty$ we write $\ell_\infty$ for $V_{\mathbb{P}^2}(w)$. This is the *line at infinity*.

Let $\ell$ be a projective line $ew + ax + by = 0$ in $\mathbb{P}^2$. Then $\ell$ does not contain the line at infinity if and only if $\ell$ does not equal the line at infinity, that is, if and only if $a \neq 0$ or $b \neq 0$. In this case, $\ell \cap \mathbb{A}^2$ is the affine line $e + ax + by = 0$ (see Sect. 3.4).

Let $L$ be an affine line $e + ax + by = 0$, and let $\ell = L^\sharp$ be its projective closure, given by the equation $ew + ax + by = 0$. Then $\ell \cap \ell_\infty$ is the singleton containing the point $(0 : b : -a)$. This point corresponds to the *direction* of the line $L$; $L$'s slope is $-a/b$ if $L$ is not vertical ($b \neq 0$). The projective closures of all the affine lines which are parallel to $L$ all intersect at this point.

A projective line is *horizontal* if it is defined by the equation $ew = by$. If $L$ is an affine line then its projective closure $\ell = L^\sharp$ is horizontal if and only if $L$ is. However, we can choose $b = 0$ as well, so the line at infinity $\ell_\infty$ is horizontal as

well. A projective line $\ell$ is horizontal if and only if the *horizontal point at infinity* $(0:1:0)$ lies on $\ell$.

Similarly, a projective line is *vertical* if it is defined by $ew = ax$. The vertical projective lines are the projective closures of the affine vertical lines $x = a$, and the line at infinity (which is thus both horizontal and vertical). A projective line $\ell$ is vertical if and only if the *vertical point at infinity* $(0:0:1)$ lies on $\ell$.

At least in the real case, as mentioned in the introduction to this chapter, the identification between points on the line at infinity and directions of lines in the affine plane $\mathbb{A}^2$ gives us a way to conceptualise the projective plane $\mathbb{P}^2$. The points of $\mathbb{P}^2$ at "infinite distance" from the origin form some kind of circle of infinite radius, except that opposite points are identified.

### Affine Cover and the Riemann Sphere

The choice of the variable $x_0$ makes notation easy but is not essential. We could have chosen any variable $x_i$ and then identify $\mathbb{A}^n$ with $U_i = \mathbb{P}^n \setminus V_{\mathbb{P}^n}(x_i)$ by the map $\boldsymbol{a} \mapsto (a_1 : a_2 : \cdots : a_{i-1} : 1 : a_i : \cdots : a_n)$. Under this choice, $H_i = V_{\mathbb{P}^n}(x_i)$ plays the role of the hyperplane at infinity, and the projective closure of $V_{\mathbb{A}^n}(f)$ under this identification is $V_{\mathbb{P}^n}(f^{\natural x_i})$. So we get $n+1$ ways of embedding $\mathbb{A}^n$ into $\mathbb{P}^n$. Projective space is covered by these $n+1$ copies of $\mathbb{A}^n$, that is, $\mathbb{P}^n = \bigcup_{i \leqslant n} U_i$.

The simplest example is $n = 1$. In this case

$$H_0 = \mathbb{P}^1 \setminus U_0 = V_{\mathbb{P}^1}(x_0) = \{(0:1)\}$$

and

$$H_1 = \mathbb{P}^1 \setminus U_1 = V_{\mathbb{P}^1}(x_1) = \{(1:0)\}.$$

The point $(0:1)$ is the "point at infinity" under the identification of the affine line $\mathbb{A}^1$ with $U_0$ by $\rho_0(a) = (1:a)$, and the point $(1:0)$ is the "point at infinity" under the identification of the affine line $\mathbb{A}^1$ with $U_1$ by $\rho_1(a) = (a:1)$. The standard identification, though, for both $\mathbb{P}^1$ and $\mathbb{P}^2$, will be using $\rho_0$.

In the case $\mathbb{K} = \mathbb{R}$, $\mathbb{P}^1(\mathbb{R})$ is topologically equivalent to the unit circle: it is obtained by adding a single point to both "ends" of the real line $\mathbb{A}^1(\mathbb{R})$. In the case $\mathbb{K} = \mathbb{C}$, $\mathbb{P}^1(\mathbb{C})$ turns out to be topologically equivalent to the unit sphere in $\mathbb{R}^3$, and so is called the *Riemann sphere*. It is obtained by taking the complex affine line $\mathbb{A}^1(\mathbb{C}) = \mathbb{C}$, which is topologically identical to the real plane $\mathbb{R}^2$, and adding a point to the entire "boundary" of the plane. See Exercises 8.55 and 8.56. $\mathbb{P}^1(\mathbb{C})$ is quite different from $\mathbb{P}^2(\mathbb{R})$, which is not embeddable into $\mathbb{R}^3$.

### Algebraic Subsets of the Projective Line

Let $C$ be an algebraic hypersurface of $\mathbb{P}^1$. Then $C$ is the sum of $C\restriction_{\mathbb{A}^1}$ and a number of copies of the point at infinity $(0:1)$. Since $C\restriction_{\mathbb{A}^1}$ is an algebraic hypersurface of $\mathbb{A}^1$, it is finite (Exercise 3.6), and so $C$ is finite as well. When $\mathbb{K}$ is algebraically closed, we get a complete understanding of the hypersurfaces of $\mathbb{P}^1$.

**Proposition 4.26** *Suppose that* $\mathbb{K}$ *is algebraically closed.*

(a) *A homogeneous polynomial in* $\mathbb{K}[w, x]$ *is irreducible if and only if it is linear.*
(b) *An algebraic hypersurface* $C$ *of* $\mathbb{P}^1$ *contains precisely* $\deg C$ *many points (counted with multiplicities).*

***Proof*** Every linear polynomial is irreducible. Let $f \in \mathbb{K}[w, x]$ be homogeneous and irreducible; either $f \sim w$, or $w$ does not divide $f$. In the latter case $f = (f^\flat)^\sharp$ so $f^\flat \in \mathbb{K}[x]$ is irreducible, and is linear; and $\deg f = \deg f^\flat$.

For (2), let $f \in \mathbb{K}[w, x]$ be homogeneous; by (1) (and Proposition 4.1), $f$ is the product of $\deg f$ many linear polynomials. For each homogeneous linear polynomial $ew - ax$, the projective hypersurface $ew = ax$ contains precisely one point $(a : e)$. □

If $f \in \mathbb{K}[w, x]$ is homogeneous, then we call a point in $V_{\mathbb{P}^1}(f)$ a *root* of $f$.

*Exercise 4.27* Let $C$ be a curve in $\mathbb{P}^2$ which does not contain the line at infinity. Show that $C$ intersects the line at infinity $\ell_\infty$ in at most $\deg C$ many points (see Exercise 4.25). In particular, this intersection is finite. «

## 4.5 Changes of Coordinates

A projective map $\psi : \mathbb{P}^n \to \mathbb{P}^m$ is injective and its range is an $n$-dimensional subspace of $\mathbb{P}^m$; so is onto $\mathbb{P}^m$ if and only if $n = m$, in which case $\psi^{-1}$ is also a projective map.

**Definition 4.28** A *change of coordinates* of $\mathbb{P}^n$ is a projective map from $\mathbb{P}^n$ to itself.

### 4.5.1 Change of Variable

Let $\boldsymbol{\alpha}$ be a linear presentation of a change of coordinates $\alpha$ of $\mathbb{P}^n$; in other words, it is an invertible linear map from $\mathbb{K}^{n+1}$ to itself. Recall that for a linear presentation of a projective map $\boldsymbol{\psi} = (\boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n)$ we let $\psi_i$ be the homogeneous linear polynomial which defines the coordinate map $\boldsymbol{\psi}_i$.

We let $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^{-1}$ so that we can write $(\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_n)$ for the linear polynomials defining the coordinate maps of $\boldsymbol{\alpha}^{-1}$. We now define a map $\boldsymbol{\alpha}^* : \mathbb{K}[\boldsymbol{x}] \to \mathbb{K}[\boldsymbol{x}]$ by letting

$$\boldsymbol{\alpha}^*(f(x_0, x_1, \ldots, x_n)) = f\left(\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_n\right).$$

By Proposition 2.11, this is a ring homomorphism. Let $A$ be the matrix such that $\boldsymbol{\alpha} = T_A$. Since $\hat{\alpha}_i$ is the polynomial which defines multiplication by the $i$th row of $A^{-1}$, $\boldsymbol{\alpha}^*$ is the unique ring homomorphism from $\mathbb{K}[\boldsymbol{x}]$ to itself such that

$$\begin{pmatrix} \boldsymbol{\alpha}^*(x_0) \\ \boldsymbol{\alpha}^*(x_1) \\ \vdots \\ \boldsymbol{\alpha}^*(x_n) \end{pmatrix} = A^{-1} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

*Example 4.29* Let $\alpha\colon \mathbb{P}^2 \to \mathbb{P}^2$ be the change of coordinates $\alpha(e:a:b) = (b:a-b:e)$, which maps the origin $(1:0:0)$ to the vertical point at infinity, fixes the horizontal point at infinity, and maps the vertical point at infinity to $(1:-1:0)$ (which under our identification of $\mathbb{A}^2$ with a subset of $\mathbb{P}^2$ by $(a,b) = (1:a:b)$ is the point $(-1,0)$ on the $x$-axis).

A linear presentation of $\alpha$ is $\boldsymbol{\alpha}(e,a,b) = (b, a-b, e)$; its inverse is the map $\boldsymbol{\alpha}^{-1}(e,a,b) = (b, a+e, e)$, and so $\boldsymbol{\alpha}^*(f(w,x,y)) = f(y, w+x, w)$. «

Let $\boldsymbol{\alpha} = T_A$ and $\boldsymbol{\beta} = T_B$ be invertible linear maps from $\mathbb{K}^{n+1} \to \mathbb{K}^{n+1}$. Write $\hat{A} = A^{-1}$ so the entries of $A^{-1}$ are $\hat{a}_{i,j}$ for $i, j = 0, 1 \ldots, n$; so $\hat{\alpha}_i = \sum_j \hat{a}_{i,j} x_j$. Since $\boldsymbol{\beta}^*$ is a ring homomorphism, for each $i \leqslant n$,

$$\boldsymbol{\beta}^*(\boldsymbol{\alpha}^*(x_i)) = \sum_j \hat{a}_{i,j} \boldsymbol{\beta}^*(x_j),$$

so

$$\begin{pmatrix} \boldsymbol{\beta}^*(\boldsymbol{\alpha}^*(x_0)) \\ \boldsymbol{\beta}^*(\boldsymbol{\alpha}^*(x_1)) \\ \vdots \\ \boldsymbol{\beta}^*(\boldsymbol{\alpha}^*(x_n)) \end{pmatrix} = A^{-1} \begin{pmatrix} \boldsymbol{\beta}^*(x_0) \\ \boldsymbol{\beta}^*(x_1) \\ \vdots \\ \boldsymbol{\beta}^*(x_n) \end{pmatrix} = A^{-1} B^{-1} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Since $T_{A^{-1}B^{-1}} = (\boldsymbol{\beta} \circ \boldsymbol{\alpha})^{-1}$, $\boldsymbol{\beta}^* \circ \boldsymbol{\alpha}^*$ and $(\boldsymbol{\beta} \circ \boldsymbol{\alpha})^*$ agree on $x_0, x_1, \ldots, x_n$; by Proposition 2.11, $\boldsymbol{\beta}^* \circ \boldsymbol{\alpha}^* = (\boldsymbol{\beta} \circ \boldsymbol{\alpha})^*$.

In particular, if $\boldsymbol{\beta} = \boldsymbol{\alpha}^{-1}$ then $(\boldsymbol{\alpha}^*)^{-1} = (\boldsymbol{\alpha}^{-1})^*$, which implies that $\boldsymbol{\alpha}^*$ is a ring *automorphism* of $\mathbb{K}[\boldsymbol{x}]$, namely, an invertible homomorphism.

**Proposition 4.30** *Let $\boldsymbol{\alpha}$ be a linear presentation of a change of coordinates $\alpha$ of $\mathbb{P}^n$; let $f \in \mathbb{K}[\boldsymbol{x}]$.*

(a) $\deg \boldsymbol{\alpha}^*(f) = \deg f$.
(b) *$f$ is homogeneous if and only if $\boldsymbol{\alpha}^*(f)$ is homogeneous.*
(c) *The function on $\mathbb{K}^{n+1}$ defined by $\boldsymbol{\alpha}^*(f)$ is $f \circ \boldsymbol{\alpha}^{-1}$.*
(d) *If $f$ is homogeneous then $\alpha\,[V_{\mathbb{P}^n}(f)] = V_{\mathbb{P}^n}(\boldsymbol{\alpha}^*(f))$.*

Thus, if $C$ is a hypersurface of $\mathbb{P}^n$ and $\alpha$ is a change of coordinates of $\mathbb{P}^n$ then $\alpha[C]$ is a hypersurface of $\mathbb{P}^n$, of the same degree as $C$.

***Proof*** Since $\boldsymbol{\alpha}^*$ is a ring homomorphism, we have

$$\boldsymbol{\alpha}^*(x_0^{m_0}x_1^{m_1}\cdots x_n^{m_n}) = \boldsymbol{\alpha}^*(x_0)^{m_0}\cdot\boldsymbol{\alpha}^*(x_1)^{m_1}\cdots\boldsymbol{\alpha}^*(x_n)^{m_n};$$

each $\boldsymbol{\alpha}^*(x_i)$ is homogeneous and linear, so $\boldsymbol{\alpha}^*(\boldsymbol{x}^{\boldsymbol{m}})$ is homogeneous of degree $\sum m_i = \deg(\boldsymbol{x}^{\boldsymbol{m}})$. Extending linearly we obtain (a) and (b) (for (b) also consider the inverse of $\boldsymbol{\alpha}^*$).

By definition, the tuple of polynomials $(\boldsymbol{\alpha}^*(x_0),\ldots,\boldsymbol{\alpha}^*(x_n))$ defines the function $\boldsymbol{\alpha}^{-1}$ on $\mathbb{K}^{n+1}$; (c) follows from the fact that polynomial substitution defines composition (Remark 2.12).

If $f \in \mathbb{K}[\boldsymbol{x}]$ is homogeneous and a nonzero $\boldsymbol{p} \in \mathbb{K}^{n+1}$ is a presentation of a point $p \in \mathbb{P}^n$ then $p \in V_{\mathbb{P}^n}(\boldsymbol{\alpha}^*(f))$ if and only if $(f\circ\boldsymbol{\alpha}^{-1})(\boldsymbol{p}) = 0$ if and only if $\boldsymbol{\alpha}^{-1}(\boldsymbol{p}) \in V_{\mathbb{A}^{n+1}}(f)$ if and only if $\alpha^{-1}(p) \in V_{\mathbb{P}^n}(f)$. This shows that (d) holds for irreducible homogeneous polynomials in $\mathbb{K}[\boldsymbol{x}]$. This is extended to all homogeneous polynomials by taking sums of the irreducible components, noting that if $[f_1,\ldots,f_m]$ is an irreducible decomposition of $f$ then $[\boldsymbol{\alpha}^*(f_1),\ldots,\boldsymbol{\alpha}^*(f_m)]$ is an irreducible decomposition of $\boldsymbol{\alpha}^*(f)$, using the fact that $\boldsymbol{\alpha}^*$ is a ring automorphism of $\mathbb{K}[\boldsymbol{x}]$. □

We call $\boldsymbol{\alpha}^*$ a *change of variable* for $\mathbb{P}^n$.

*Example 4.31* Let $\alpha$ be the change of coordinates from Example 4.29. It maps the $x$-axis $y = 0$ to the line at infinity $w = 0$ (it sends the point $(e:a:0)$ to $(0:a:e)$, and indeed $\boldsymbol{\alpha}^*(y) = w$. «

*Exercise 4.32* Let $\boldsymbol{\alpha}(e,a,b) = (e+b,a,e-b)$. (a) Compute $\boldsymbol{\alpha}^*$. (b) Show that $\boldsymbol{\alpha}^*$ maps $x^2+y^2-w^2$ to $x^2-wy$. (c) Conclude that $\boldsymbol{\alpha}$ maps $S^\sharp$ (the projective closure of the unit circle) to the projective closure $P^\sharp$ of the parabola $P$ given by $y = x^2$.[2] «

*Remark 4.33* The origin of the terminology "change of coordinates" and "change of variable" is an alternative understanding of the nature of projective space. An abstract vector space $V$ over $\mathbb{K}$ does not have a canonical basis. If $\dim V = n+1$ then every linear isomorphism between $V$ and $\mathbb{K}^{n+1}$ gives a way to assign coordinates to the elements of $V$. We can define $\mathbb{P}(V)$ to be the set of 1-dimensional subspaces of $V$. Any choice of (affine) coordinates for $V$ then gives a choice of projective coordinates for $\mathbb{P}(V)$. We can define algebraic hypersurfaces of $\mathbb{P}(V)$ by

[2] Note that if $\mathbb{K} = \mathbb{R}$, then $S^\sharp = S$, but if $\mathbb{K} = \mathbb{C}$ then $S^\sharp$ contains two points at infinity. What this shows is that $S^\sharp$ and $P^\sharp$ are geometrically the same, even though $S$ and $P$ are not. Taking the projective closure added the "missing points" that were required to get the isomorphism.

using any system of coordinates. Proposition 4.30 shows that a different choice of coordinates doesn't change the hypersurfaces; it changes the equations that define them. The change of coordinates $\alpha$ tells us how to translate from one coordinate system to another. The change of variable $\boldsymbol{\alpha}^*$ tells us how the equations defining a hypersurface change with the coordinates. The variables $x_0, \ldots, x_n$ are replaced by "new variables" $\hat{x}_0 = \boldsymbol{\alpha}^*(x_0), \ldots, \hat{x}_n = \boldsymbol{\alpha}^*(x_n)$, and the equation defining the hypersurface changes from $f(\boldsymbol{x}) = 0$ to $f(\hat{\boldsymbol{x}}) = 0$.

While we don't use abstract spaces like $\mathbb{P}(V)$ in this book, this is a useful point of view. We will use changes of coordinates to simplify calculations. Suppose that we want to show that a hypersurface $C$ of $\mathbb{P}^n$ has some property, call it $P$. We will usually show that the property $P$ is *invariant* under changes of coordinates: the hypersurface $C$ has the property $P$ if and only if $\alpha[C]$ has the property $P$ (we also say that $P$ is a "geometric" property). We find a change of coordinates $\alpha$ such that the equation defining $\alpha[C]$ is relatively simple. This will help us verify that $\alpha[C]$ has the property $P$, from which we conclude that $C$ has the property as well.

In the alternative view the hypersurface $C$ didn't move: we just found a way to give new coordinates that would simplify the equation defining $C$. This view will be implicit in our proofs: instead of mentioning $\alpha$ and working with $\alpha[C]$, we will usually say "after changing coordinates, the equation for $C$ is…" For this to work, of course, we need to: (i) show that the property $P$ is invariant under changes of coordinates; and (ii) show that there is some change of coordinates making the equation nice; this is the role of the Four Point Lemma. «

### 4.5.2 Four Point Lemma

We say that points in a set $X \subseteq \mathbb{P}^2$ are *collinear* if $X$ is contained in a line of $\mathbb{P}^2$. Theorem 4.12 says that any two points in $\mathbb{P}^2$ are collinear, but it is easy to find three points in $\mathbb{P}^2$ which are not collinear.

**Definition 4.34** Points in a set $X \subseteq \mathbb{P}^2$ are said to *lie in general position* if no three distinct points in $X$ are collinear.

**Four Point Lemma** *If $p_1, p_2, p_3, p_4$ and $q_1, q_2, q_3, q_4$ are two quadruples of points in $\mathbb{P}^2$, both of which lie in general position, then there is a change of coordinates of $\mathbb{P}^2$ which maps each $p_i$ to $q_i$.*

The condition is necessary since any change of coordinates maps lines to lines, and so collinear points to collinear points.

***Proof*** Consider first a simpler task. Suppose that we only had three non-collinear points $p_1$, $p_2$ and $p_3$ which we wanted to map to three non-collinear points $q_1$, $q_2$ and $q_3$. For $i = 1, 2, 3$ let $\boldsymbol{p_i}$ be a presentation of $p_i$ and $\boldsymbol{q_i}$ be a presentation of $q_i$.

The fact that $p_1$, $p_2$ and $p_3$ are not collinear means that $\dim \langle \boldsymbol{p_1}, \boldsymbol{p_2}, \boldsymbol{p_3} \rangle = 3$. For otherwise, $W = \langle \boldsymbol{p_1}, \boldsymbol{p_2}, \boldsymbol{p_3} \rangle$ has dimension $\leqslant 2$, which would imply that $p_1$, $p_2$ and $p_3$ all lie on the line $\pi_2[W]$ (if $\dim W = 1$ then $p_1$, $p_2$ and $p_3$ are the same point). Thus $\{\boldsymbol{p_1}, \boldsymbol{p_2}, \boldsymbol{p_3}\}$ is a basis of $\mathbb{K}^3$. Similarly, $\{\boldsymbol{q_1}, \boldsymbol{q_2}, \boldsymbol{q_3}\}$ is also a basis of $\mathbb{K}^3$. We can then define the unique linear map $\boldsymbol{\alpha} \colon \mathbb{K}^3 \to \mathbb{K}^3$ such that $\boldsymbol{\alpha}(\boldsymbol{p_i}) = \boldsymbol{q_i}$ (Proposition 2.60), and this map is invertible since $\{\boldsymbol{q_1}, \boldsymbol{q_2}, \boldsymbol{q_3}\}$ is a basis of $\mathbb{K}^3$. Hence the induced change of coordinates $\alpha$ maps $p_i$ to $q_i$ for $i = 1, 2, 3$.

We can manage a fourth point. Let $p_i, q_i$ $(i = 1, 2, 3, 4)$ be as in the hypothesis, and again choose presentations $\boldsymbol{p_i}$ and $\boldsymbol{q_i}$. Because $\{\boldsymbol{p_1}, \boldsymbol{p_2}, \boldsymbol{p_3}\}$ is a basis of $\mathbb{K}^3$, $\boldsymbol{p_4}$ is a linear combination of the elements of this basis: there are scalars $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{K}$ such that $\boldsymbol{p_4} = \lambda_1 \boldsymbol{p_1} + \lambda_2 \boldsymbol{p_2} + \lambda_3 \boldsymbol{p_3}$. Similarly, there are scalars $\mu_1, \mu_2$ and $\mu_3 \in \mathbb{K}$ such that $\boldsymbol{q_4} = \mu_1 \boldsymbol{q_1} + \mu_2 \boldsymbol{q_2} + \mu_3 \boldsymbol{q_3}$.

The main point is that none of $\lambda_1$, $\lambda_2$ or $\lambda_3$ can be 0. Suppose, for example, that $\lambda_3 = 0$. Then $\boldsymbol{p_4} \in \langle \boldsymbol{p_1}, \boldsymbol{p_2} \rangle$, and this means that $p_4$ lies on the line $\overline{p_1\, p_2}$, contrary to hypothesis. Similarly, all of $\mu_1$, $\mu_2$ and $\mu_3$ are nonzero.

For $i = 1, 2, 3$, let $\boldsymbol{p_i}' = \lambda_i \boldsymbol{p_i}$ and let $\boldsymbol{q_i}' = \mu_i \boldsymbol{q_i}$. Then $\boldsymbol{p_i}'$ is another presentation of $p_i$, and similarly for $q_i$. Also, $\{\boldsymbol{p_1}', \boldsymbol{p_2}', \boldsymbol{p_3}'\}$ and $\{\boldsymbol{q_1}', \boldsymbol{q_2}', \boldsymbol{q_3}'\}$ are bases of $\mathbb{K}^3$. We now take a linear map $\boldsymbol{\alpha} \colon \mathbb{K}^3 \to \mathbb{K}^3$ such that $\boldsymbol{\alpha}(\boldsymbol{p_i}') = \boldsymbol{q_i}'$ for $i = 1, 2, 3$. As above, $\boldsymbol{\alpha}$ is invertible, and so induces a change of coordinates $\alpha$ of $\mathbb{P}^2$; $\alpha$ maps $p_1$ to $q_1$, $p_2$ to $q_2$ and $p_3$ to $q_3$. For $i = 4$, $\boldsymbol{p_4} = \boldsymbol{p_1}' + \boldsymbol{p_2}' + \boldsymbol{p_3}'$ and $\boldsymbol{q_4} = \boldsymbol{q_1}' + \boldsymbol{q_2}' + \boldsymbol{q_3}'$, so $\boldsymbol{\alpha}(\boldsymbol{p_4}) = \boldsymbol{\alpha}(\boldsymbol{p_1}' + \boldsymbol{p_2}' + \boldsymbol{p_3}') = \boldsymbol{q_1}' + \boldsymbol{q_2}' + \boldsymbol{q_3}' = \boldsymbol{q_4}$, so $\alpha(p_4) = q_4$. □

In the proof of the four point lemma we showed the simpler

**Three Point Lemma** *If $p_1$, $p_2$ and $p_3$ are not collinear and neither are $q_1$, $q_2$ and $q_3$, then there is a change of coordinates of $\mathbb{P}^2$ which maps each $p_i$ to $q_i$.* □

For fewer points the argument is even simpler. If $p_1$ and $p_2$ is *any* pair of distinct points in $\mathbb{P}^2$ and so are $q_1$ and $q_2$, then there is a change of coordinates which maps $p_1$ to $q_1$ and $p_2$ to $q_2$. If we don't want to run through the argument again, we can simply pick $p_3 \notin \overline{p_1\, p_2}$ and $q_3 \notin \overline{q_1\, q_2}$ and use the three point lemma.

*Remark 4.35* Let $\ell$ and $\ell'$ be two lines in $\mathbb{P}^2$. There is a change of coordinates $\alpha$ of $\mathbb{P}^2$ such that $\alpha[\ell] = \ell'$. For let $p_1$ and $p_2$ be two distinct points on $\ell$, and let $q_1$ and $q_2$ be two distinct points on $\ell'$. As discussed above, there is a change of coordinates $\alpha$ of $\mathbb{P}^2$ which maps $p_1$ to $q_1$ and $p_2$ to $q_2$. Projective maps map lines to lines (for changes of coordinates this also follows from Proposition 4.30). Since $q_1$ and $q_2$ are elements of $\alpha[\ell]$, this line must equal $\overline{q_1\, q_2} = \ell'$, the unique line which passes through $q_1$ and $q_2$ (Theorem 4.12). «

*Exercise 4.36* Let $C$ be a curve in $\mathbb{P}^2$ and let $\ell$ be a line. Use Exercise 4.27 and Remark 4.35 to show that $C$ and $\ell$ intersect in at most $\deg C$ many points. «

*Exercise 4.37* Generalise the three and four point lemmas to $\mathbb{P}^n$ for $n \geqslant 3$. «

## 4.6 Spaces of Curves

For $d \geqslant 1$, we let $\mathbb{G}_d$ be the collection of all curves of $\mathbb{P}^2$ of degree $d$. That is,

$$\mathbb{G}_d = \left\{ V_{\mathbb{P}^2}(f) \,:\, f \in \mathbb{K}[w, x, y] \text{ is homogeneous and } \deg f = d \right\}.$$

The space of curves has the structure of a projective space by using the coefficients of a polynomial $f$ to be homogeneous coordinates of the hypersurface $V_{\mathbb{P}^2}(f)$. We fix a list $m_0, m_1, \ldots m_k$ of the monic monomials (monomials with coefficient 1) in $\mathbb{K}[w, x, y]$ of degree $d$:

$$(m_0, \ldots, m_k) = (w^d, w^{d-1}x, w^{d-2}x^2, \ldots, wx^{d-1}, x^d, w^{d-2}xy, \ldots, y^d);$$

the particular order in which we place these monomials is not important.

*Exercise 4.38* Show that $k + 1 = 1 + 2 + 3 + \cdots + (d + 1)$ and so $k = d(d + 3)/2$. «

If $a = (a_0 : a_1 : \ldots : a_k) \in \mathbb{P}^k$ then we let

$$\iota_d(a) = V_{\mathbb{P}^2}\left(a_0 m_0 + a_1 m_1 + \cdots + a_k m_k\right).$$

The map is well-defined because $V_{\mathbb{P}^2}\left(a_0 m_0 + a_1 m_1 + \cdots + a_k m_k\right)$ equals $V_{\mathbb{P}^2}\left(\lambda a_0 m_0 + \lambda a_1 m_1 + \cdots + \lambda a_k m_k\right)$ when $\lambda \neq 0$. If $\mathbb{K}$ is algebraically closed then Projective Study's Lemma shows that $\iota_d$ is a bijection between $\mathbb{P}^k$ and $\mathbb{G}_d$. In this case we use the map $\iota_d$ to give a geometric structure to $\mathbb{G}_d$. For example, a *change of coordinates* of $\mathbb{G}_d$ is a map of the form $\iota_d \circ \alpha \circ (\iota_d)^{-1}$, where $\alpha$ is a change of coordinates of $\mathbb{P}^k$; an *m-dimensional subspace* of $\mathbb{G}_d$ is a subset of $\mathbb{G}_d$ of the form $\iota_d[U]$ where $U$ is an $m$-dimensional subspace of $\mathbb{P}^k$ and so on. In particular we will use 1-dimensional subspaces:

**Definition 4.39** Let $d \geqslant 1$. A *linear family of curves* of degree $d$ is the image under $\iota_d$ of a line in $\mathbb{P}^k$.

Let $C = \iota_d(p)$ and $D = \iota_d(q)$ be distinct curves of degree $d$. Assuming that $\iota_d$ is a bijection between $\mathbb{P}^k$ and $\mathbb{G}_d$, the linear family $\iota_d[\overline{p\,q}]$ is denoted by $\overline{C\,D}$. If $f$ defining $C$ and $g$ defining $D$ are "presentations" of $C$ and $D$ then the map $(e : a) \mapsto V_{\mathbb{P}^2}(ef + ag)$ is a linear parameterisation of the linear family $\overline{C\,D}$ (see Example 4.17).

### 4.6.1 The Dual Plane

The most prominent space of curves is $\mathbb{G}_1$, the space of lines in $\mathbb{P}^2$. In this case we have $k = 2$, and even if $\mathbb{K}$ is not algebraically closed, $\iota_1$ is a bijection between $\mathbb{P}^2$ and $\mathbb{G}_1$. The space $\mathbb{G}_1$ is commonly denoted by $\check{\mathbb{P}}^2$, and is called the *dual projective plane*. The duality is in that the bijection $\iota_1 \colon \mathbb{P}^2 \to \check{\mathbb{P}}^2$ maps points to lines, and implicitly, lines to points, as we now show.

We write $\iota$ instead of $\iota_1$. For definiteness, let us assume that for the definition of $\iota$ we use the list of monomials $w, x, y$; so $\iota(e:a:b) = V_{\mathbb{P}^2}(ew + ax + by)$.

**Proposition 4.40** *The following are equivalent for a subset $X$ of $\check{\mathbb{P}}^2$:*

(1) *$X$ is a linear family of lines.*
(2) *$X$ is the collection of all lines in $\mathbb{P}^2$ which pass through a fixed point $p$.*

*More precisely, for every $q \in \mathbb{P}^2$, $\iota[\iota(q)]$ is the collection of all lines in $\mathbb{P}^2$ which pass through $q$.*

A linear family of lines is sometimes also called a *pencil of lines*.

***Proof*** The main point is that for $p, q \in \mathbb{P}^2$, $p \in \iota(q)$ if and only if $q \in \iota(p)$. For if $\boldsymbol{p} = (p_w, p_x, p_y) \in \mathbb{K}^3 \setminus \{\mathbf{0}\}$ is a presentation of $p$ and $\boldsymbol{q} = (q_w, q_x, q_y)$ is a presentation of $q$, then $p \in \iota(q)$ if and only if $p \in V_{\mathbb{P}^2}(q_w w + q_x x + q_y y)$ if and only if $q_w p_w + q_x p_x + q_y p_y = 0$; the last condition is symmetric in $\boldsymbol{p}$ and $\boldsymbol{q}$.

Let $\ell$ be a line in $\mathbb{P}^2$; let $q = \iota^{-1}(\ell)$. Then

$$\iota[\ell] = \{\iota(p) \,:\, p \in \iota(q)\} = \{\iota(p) \,:\, q \in \iota(p)\}\,,$$

that is, $\iota[\ell]$ is the collection of all lines which pass through $q$. □

*Example 4.41* A line is horizontal if and only if it passes through the horizontal point at infinity $(0:1:0)$; the family of horizontal lines is the image under $\iota$ of the $x$-axis. «

*Exercise 4.42* Show that the changes of coordinates of $\check{\mathbb{P}}^2$ are the maps of the form $\ell \mapsto \alpha[\ell]$, where $\alpha$ is a change of coordinates of $\mathbb{P}^2$. «

#### Principle of Duality

Proposition 4.40 implies the *principle of duality*. Suppose that $P$ is a statement about lines and points in $\mathbb{P}^2$. The *dual of $P$* is obtained from $P$ by exchanging the roles of points and lines. The principle of duality says that if $P$ is true then so is its dual. The simplest example is: the dual of the statement "every two points lie on a unique line" (Theorem 4.12) is "every two lines intersect at a unique point" (Theorem 4.14).

*Exercise 4.43* Here is an application of the principle of duality. Suppose that $\mathbb{K}$ is infinite. Use Propositions 4.7 and 4.40 to show that if $X \subset \mathbb{P}^2$ is finite then there is

some line which doesn't pass through any point in $X$. Interpret this as a statement about linear subspaces of $\mathbb{K}^3$ and prove this statement directly. «

### 4.6.2 Desargues' Theorem

A *triangle* in $\mathbb{P}^2$ is formalised as an ordered triple of points which are not collinear (they are the vertices of the triangle). Let $abc$ and $a'b'c'$ be two triangles which do not share a vertex. The triangles are *perspective through the point* $v$ if $v$ lies on the lines $\overline{a\,a'}$, $\overline{b\,b'}$ and $\overline{c\,c'}$. The point $v$ is called the *centre* of the perspectivity; in this case we say that they are *centrally perspective* (or *perspective through a point*). The triangles are *perspective through a line* $\ell$ if the intersections of the corresponding sides: $\overline{a\,b} \cap \overline{a'\,b'}$, $\overline{a\,c} \cap \overline{a'\,c'}$, $\overline{b\,c} \cap \overline{b'\,c'}$ all lie on $\ell$. The line $\ell$ is called the *axis* of the perspectivity and so we say that the triangles are *axially perspective* (or *perspective through a line*). See Fig. 4.3.

Note that if any of the lines above coincide then the triangles are trivially both centrally and axially perspective, so we only consider pairs of triangles for which the lines are distinct.

**Proposition 4.44** *If two triangles are perspective through a point then they are perspective through a line.*

***Proof*** Say that the triangles $abc$ and $a'b'c'$ are perspective through the point $v$. No three of the points $\{a, b, c, v\}$ are collinear. By the Four Point Lemma, after a change

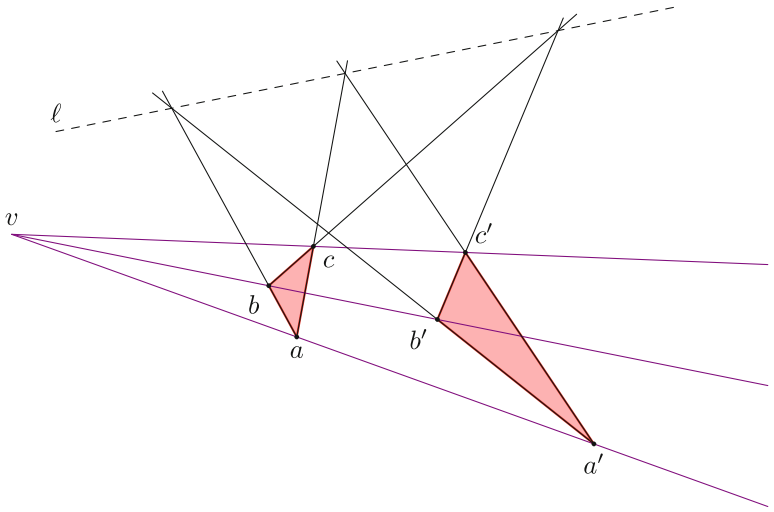



**Fig. 4.3** Desargues' theorem

of coordinates we assume that $a = o = (1:0:0)$ is the origin, $b = (0:1:0)$ is the horizontal point at infinity, $c = (0:0:1)$ is the vertical point at infinity, and that $v = (1:1:1)$. Thus the line $\overline{a\,b}$ is the $x$-axis (defined by $y = 0$); the line $\overline{a\,c}$ is the $y$-axis (defined by $x = 0$); and the line $\overline{b\,c}$ is the line at infinity ($w = 0$).

The line $\overline{a\,v}$ is defined by $y = x$; since $a' \in \overline{a\,v}$ and is distinct from $a$ and from $v$ we can write $a' = (\alpha:1:1)$ for some $\alpha \neq 0, 1$. The line $\overline{b\,v}$ is given by $y = w$, and so we can write $b' = (1:\beta:1)$; and similarly $c' = (1:1:\gamma)$.

The line $\overline{a'\,b'}$ is parameterised as $(\alpha s + t : s + \beta t : s + t)$ for $(s:t) \in \mathbb{P}^2$ (see Example 4.17), and it intersects the line $y = 0$ when $s = -t$, which gives the point $(\alpha - 1 : 1 - \beta : 0)$; similarly, the line $\overline{a\,c}$ and $\overline{a'\,c'}$ intersect at $(1-\alpha:0:\gamma-1)$; and $\overline{b\,c}$ and $\overline{b'\,c'}$ intersect at $(0:\beta-1:1-\gamma)$. The vectors $(\alpha-1, 1-\beta, 0)$, $(1-\alpha, 0, \gamma-1)$ and $(0, \beta - 1, 1 - \gamma)$ are linearly dependent (their sum is the zero vector) and so span a subspace of $\mathbb{K}^3$ of dimension 2; this implies that the three intersection points are collinear. □

The dual of Proposition 4.44 is its converse. Thus, the converse follows from Propositions 4.40 and 4.44. Together we get:

**Desargues' Theorem** *Two triangles in $\mathbb{P}^2$ are perspective through a point if and only if they are perspective through a line.*

*Exercise 4.45* Verify the assertion made before the statement of Desargues' theorem. «

*Remark 4.46* *Pappus' theorem* is the following: let $\ell$ and $\ell'$ be two distinct lines; let $a, b, c \in \ell$ be distinct and $a', b', c' \in \ell'$ be distinct, with none of the points being the point of intersection of $\ell$ and $\ell'$. Then the three points of intersection $\overline{a\,b'} \cap \overline{a'\,b}$, $\overline{a\,c'} \cap \overline{a'\,c}$ and $\overline{b\,c'} \cap \overline{b'\,c}$ are collinear. See Fig. 7.5.

A proof in $\mathbb{P}^2$ can be given in a similar way to the proof we just gave for Desargues' theorem. For a different proof see Exercise 7.55. «

## 4.7 Products of Projective Spaces

Intuitively speaking, we see that $\mathbb{A}^{n+k}$ can be thought of as the product $\mathbb{A}^n \times \mathbb{A}^k$. In particular, if $C$ is a hypersurface of $\mathbb{A}^{n+k}$ and $\boldsymbol{a} \in \mathbb{A}^n$, then the section $C\restriction_{\boldsymbol{a}} = [\boldsymbol{b} \in \mathbb{A}^k : (\boldsymbol{a}, \boldsymbol{b}) \in C]$ is a hypersurface of $\mathbb{A}^k$. This resembles the product of topological spaces or of measure spaces. It is less clear what kind of "space" would be the product of two projective spaces. $\mathbb{P}^n \times \mathbb{P}^k$ is not $\mathbb{P}^{n+k}$, or a subset of $\mathbb{P}^{n+k+1}$: the map

$$(a_0 : \cdots : a_n), (b_0 : \cdots : b_k) \mapsto (a_0 : \cdots : a_n : b_0 : \cdots : b_k)$$

is not well-defined. For example, in $\mathbb{P}^2 \times \mathbb{P}^1$, the pair $((1:2:4),(1:0))$ equals the pair $((2:4:8),(1:0))$, but in $\mathbb{P}^4$, the point $(1:2:4:1:0)$ does not equal the point $(2:4:8:1:0)$.

Like $\mathbb{P}^n$, we first conceive of $\mathbb{P}^n \times \mathbb{P}^k$ as the image of a subset of affine space. Combining the maps $\pi_n \colon \mathbb{K}^{n+1} \setminus \{\mathbf{0}\} \to \mathbb{P}^n$ and $\pi_k \colon \mathbb{K}^{k+1} \setminus \{\mathbf{0}\} \to \mathbb{P}^k$, we obtain the map

$$\pi_{n,k} = \pi_n \times \pi_k \colon \left(\mathbb{K}^{n+1} \setminus \{\mathbf{0}\}\right) \times \left(\mathbb{K}^{k+1} \setminus \{\mathbf{0}\}\right) \to \mathbb{P}^n \times \mathbb{P}^k,$$

by letting $\pi_{n,k}(\boldsymbol{a}, \boldsymbol{b}) = (\pi_n(\boldsymbol{a}), \pi_k(\boldsymbol{b}))$.

The polynomials in $\mathbb{K}[x_0, \ldots, x_n, y_0, \ldots, y_k]$ which can define hypersurfaces of $\mathbb{P}^n \times \mathbb{P}^k$ are the $\boldsymbol{x}, \boldsymbol{y}$-*bihomogeneous* ones.

**Definition 4.47** Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be disjoint tuples of variables, and let $R$ be a unique factorisation domain. A polynomial $f \in R[\boldsymbol{x}, \boldsymbol{y}]$ is $\boldsymbol{x}, \boldsymbol{y}$-*bihomogeneous* if it is both $\boldsymbol{x}$-homogeneous and $\boldsymbol{y}$-homogeneous.

Every monomial in $R[\boldsymbol{x}, \boldsymbol{y}]$ is of the form $a\boldsymbol{x}^{\boldsymbol{m}}\boldsymbol{y}^{\boldsymbol{m}'}$. The $\boldsymbol{x}$-degree of this monomial is $m_1 + m_2 + \cdots + m_n$, and the $\boldsymbol{y}$-degree is $m'_1 + m'_2 + \cdots + m'_k$ (here $\boldsymbol{x}$ is an $n$-tuple of variables and $\boldsymbol{y}$ is a $k$-tuple of variables). If $f \in R[\boldsymbol{x}, \boldsymbol{y}]$ is $\boldsymbol{x}, \boldsymbol{y}$-bihomogeneous then not only is it $\boldsymbol{x}, \boldsymbol{y}$-homogeneous (of degree $\deg_{\boldsymbol{x}} f + \deg_{\boldsymbol{y}} f$), but in fact both $\deg_{\boldsymbol{x}} m$ and $\deg_{\boldsymbol{y}} m$ are constant for all monomials $m$ which appear in $f$.

We let $\deg_{\boldsymbol{x},\boldsymbol{y}} f = (\deg_{\boldsymbol{x}} f, \deg_{\boldsymbol{y}} f)$; this is the *bidegree* of $f$. If $f$ is $\boldsymbol{x}, \boldsymbol{y}$-bihomogeneous, then for all $\boldsymbol{a} \in R^n$, $f(\boldsymbol{a}, \boldsymbol{y})$ is $\boldsymbol{y}$-homogeneous (of degree $\deg_{\boldsymbol{y}} f$, or is the zero polynomial), and for all $\boldsymbol{b} \in R^k$, $f(\boldsymbol{x}, \boldsymbol{b})$ is $\boldsymbol{x}$-homogeneous (of degree $\deg_{\boldsymbol{x}} f$, or is the zero polynomial).

Returning to $\mathbb{P}^n \times \mathbb{P}^k$, we let $\boldsymbol{x} = (x_0, \ldots, x_n)$ and $\boldsymbol{y} = (y_0, \ldots, y_k)$. If $f \in \mathbb{K}[\boldsymbol{x}, \boldsymbol{y}]$ is $\boldsymbol{x}, \boldsymbol{y}$-bihomogeneous of bidegree $(d, e)$ then Proposition 4.2 applied separately to $\boldsymbol{x}$ and $\boldsymbol{y}$ shows that for all $\boldsymbol{a} \in \mathbb{K}^{n+1}$, $\boldsymbol{b} \in \mathbb{K}^{k+1}$, and $\lambda, \mu \in \mathbb{K}$, $f(\lambda\boldsymbol{a}, \mu\boldsymbol{b}) = \lambda^d \mu^e f(\boldsymbol{a}, \boldsymbol{b})$. It follows that if $\boldsymbol{a}, \boldsymbol{a}' \in \mathbb{K}^{n+1} \setminus \{\mathbf{0}\}$ and $\boldsymbol{b}, \boldsymbol{b}' \in \mathbb{K}^{k+1} \setminus \{\mathbf{0}\}$, and $\pi_{n,k}(\boldsymbol{a}, \boldsymbol{b}) = \pi_{n,k}(\boldsymbol{a}', \boldsymbol{b}')$, then $f(\boldsymbol{a}, \boldsymbol{b}) = 0$ if and only if $f(\boldsymbol{a}', \boldsymbol{b}') = 0$. We can thus let, for irreducible bihomogeneous $f \in \mathbb{K}[\boldsymbol{x}, \boldsymbol{y}]$,

$$V_{\mathbb{P}^n \times \mathbb{P}^k}(f) = \left\{ \pi_{n,k}(\boldsymbol{a}, \boldsymbol{b}) \,:\, \boldsymbol{a} \in \mathbb{K}^{n+1} \setminus \{\mathbf{0}\} \;\&\; \boldsymbol{b} \in \mathbb{K}^{k+1} \setminus \{\mathbf{0}\} \;\&\; f(\boldsymbol{a}, \boldsymbol{b}) = 0 \right\}.$$

If $f \in \mathbb{K}[\boldsymbol{x}, \boldsymbol{y}]$ is bihomogeneous, then every divisor of $f$ is also bihomogeneous; this follows immediately from Proposition 4.1, applying it once for $\boldsymbol{x}$ and once for $\boldsymbol{y}$. Hence, an irreducible factorisation of a bihomogeneous polynomial consists of bihomogeneous polynomials, and so we can define, for any bihomogeneous polynomial $f \in \mathbb{K}[\boldsymbol{x}, \boldsymbol{y}]$,

$$V_{\mathbb{P}^n \times \mathbb{P}^k}(f) = V_{\mathbb{P}^n \times \mathbb{P}^k}(f_1) + \cdots + V_{\mathbb{P}^n \times \mathbb{P}^k}(f_m),$$

where $[f_1, \ldots, f_m]$ is any irreducible factorisation of $f$. As usual, this does not depend on the choice of factorisation.

Further analysis of the hypersurfaces of $\mathbb{P}^n \times \mathbb{P}^k$ follows similarly to Sect. 4.2. For bihomogeneous $f \in \mathbb{K}[\boldsymbol{x}, \boldsymbol{y}]$ and $\boldsymbol{a} \in \mathbb{K}^{n+1} \setminus \{\boldsymbol{0}\}$, $\boldsymbol{b} \in \mathbb{K}^{k+1} \setminus \{\boldsymbol{0}\}$, $\pi_{n,k}(\boldsymbol{a}, \boldsymbol{b}) \in V_{\mathbb{P}^n \times \mathbb{P}^k}(f)$ if and only if $f(\boldsymbol{a}, \boldsymbol{b}) = 0$. We get an analogue of Proposition 4.7 with a similar proof:

**Proposition 4.48** *Let $f \in \mathbb{K}[\boldsymbol{x}, \boldsymbol{y}]$ be bihomogeneous and nonzero. Then $\left| V_{\mathbb{P}^n \times \mathbb{P}^k}(f) \right| \neq \mathbb{P}^n \times \mathbb{P}^k$.*

***Proof*** Let $S = \mathbb{K} \setminus \{0\}$; it is infinite, so (Proposition 2.19) there is some $(\boldsymbol{a}, \boldsymbol{b}) \in S^{n+k}$ such that $f(\boldsymbol{a}, \boldsymbol{b}) \neq 0$. Then $\boldsymbol{a} \in \mathbb{K}^{n+1} \setminus \{\boldsymbol{0}\}$ and $\boldsymbol{b} \in \mathbb{K}^{k+1} \setminus \{\boldsymbol{0}\}$, and $(\pi_n(\boldsymbol{a}), \pi_k(\boldsymbol{b})) \notin V_{\mathbb{P}^n \times \mathbb{P}^k}(f)$. □

We now show that Study's Lemma applies for $\mathbb{P}^n \times \mathbb{P}^k$ as well. We need an analogue of Lemma 4.8.

**Lemma 4.49** *Suppose that $\mathbb{K}$ is algebraically closed. Let $f \in \mathbb{K}[\boldsymbol{x}, \boldsymbol{y}]$ be bihomogeneous and nonzero. Then $\deg_{\boldsymbol{y}} f = 0$ if and only if there is some hypersurface $C$ of $\mathbb{P}^n$ such that $C \neq \mathbb{P}^n$ and $V_{\mathbb{P}^n \times \mathbb{P}^k}(f) \subseteq C \times \mathbb{P}^k$.*

***Proof*** In one direction, if $\deg_{\boldsymbol{y}} f = 0$, that is, if $f \in \mathbb{K}[\boldsymbol{x}]$, then $V_{\mathbb{P}^n \times \mathbb{P}^k}(f) = V_{\mathbb{P}^n}(f) \times \mathbb{P}^k$, and because $f$ is nonzero, $C \neq \mathbb{P}^n$ (Proposition 4.7).

In the other direction, suppose that $f \notin \mathbb{K}[\boldsymbol{x}]$, that is, that $\deg_{\boldsymbol{y}} f > 0$. Let $h \in \mathbb{K}[\boldsymbol{x}]$ be $\boldsymbol{x}$-homogeneous and nonzero (but possibly a constant). Write $f = \sum_{\boldsymbol{m} \in \mathbb{N}^k} f_{\boldsymbol{m}} \boldsymbol{y}^{\boldsymbol{m}}$, with $f_{\boldsymbol{m}} \in \mathbb{K}[\boldsymbol{x}]$. Since $f \neq 0$, find some $\boldsymbol{m} \in \mathbb{N}^k$ such that $f_{\boldsymbol{m}} \neq 0$. Since $f_{\boldsymbol{m}} h \neq 0$, Proposition 2.19 gives us a nonzero $\boldsymbol{a} \in \mathbb{A}^{n+1}$ such that $(f_{\boldsymbol{m}} h)(\boldsymbol{a}) \neq 0$. Hence $p = \pi_n(\boldsymbol{a})$ is not in $V_{\mathbb{P}^n}(h)$, and $f_{\boldsymbol{m}}(\boldsymbol{a}) \neq 0$. The latter implies that $f(\boldsymbol{a}, \boldsymbol{y})$ is $\boldsymbol{y}$-homogeneous of degree $\deg_{\boldsymbol{y}} f$, which is greater than zero. Then Lemma 4.8 tells us that there is some nonzero $\boldsymbol{b} \in \mathbb{A}^{k+1}$ such that $f(\boldsymbol{a}, \boldsymbol{b}) = 0$. Let $q = \pi_k(\boldsymbol{b})$. Then $(p, q) \in V_{\mathbb{P}^n \times \mathbb{P}^k}(f)$, and so $V_{\mathbb{P}^n \times \mathbb{P}^k}(f) \nsubseteq V_{\mathbb{P}^n}(h) \times \mathbb{P}^k$. □

This is the corresponding version of Study's lemma:

**Proposition 4.50** *Let $f, g \in \mathbb{K}[\boldsymbol{x}, \boldsymbol{y}]$ be nonzero and bihomogeneous. If $f$ divides $g$ then $V_{\mathbb{P}^n \times \mathbb{P}^k}(f) \subseteq V_{\mathbb{P}^n \times \mathbb{P}^k}(g)$. If, in addition, $\mathbb{K}$ is algebraically closed, and $V_{\mathbb{P}^n \times \mathbb{P}^k}(f) \subseteq V_{\mathbb{P}^n \times \mathbb{P}^k}(g)$, then $f$ divides $g$.*

***Proof*** One direction is immediate. Suppose then that $\mathbb{K}$ is algebraically closed, and that $V_{\mathbb{P}^n \times \mathbb{P}^k}(f) \subseteq V_{\mathbb{P}^n \times \mathbb{P}^k}(g)$. As in the affine and projective case, we may assume that $f$ is irreducible.

We first dispose of degenerate cases. Suppose that $g \in \mathbb{K}[\boldsymbol{x}]$, that is, that $\deg_{\boldsymbol{y}} g = 0$. Then $V_{\mathbb{P}^n \times \mathbb{P}^k}(g) = V_{\mathbb{P}^n}(g) \times \mathbb{P}^k$. Lemma 4.49 then tells us that $f \in \mathbb{K}[\boldsymbol{x}]$ as well, so $V_{\mathbb{P}^n \times \mathbb{P}^k}(f) = V_{\mathbb{P}^n}(f) \times \mathbb{P}^k$. We conclude that $V_{\mathbb{P}^n}(f) \subseteq V_{\mathbb{P}^n}(g)$,

and then conclude that $f$ divides $g$ by appealing to the projective version of Study's lemma.

Of course, if $\deg_{\boldsymbol{x}} g = 0$ then the argument is identical. So we assume that $\deg_{\boldsymbol{x}} g > 0$ and $\deg_{\boldsymbol{y}} g > 0$. This implies that for all $\boldsymbol{a} \in \mathbb{A}^{n+1}$ and all $\boldsymbol{b} \in \mathbb{A}^{k+1}$, $g(\boldsymbol{a}, \boldsymbol{0}) = 0$ and $g(\boldsymbol{0}, \boldsymbol{b}) = 0$. This, on top of the assumption that $V_{\mathbb{P}^n \times \mathbb{P}^k}(f) \subseteq V_{\mathbb{P}^n \times \mathbb{P}^k}(g)$, tells us that for all $\boldsymbol{a} \in \mathbb{A}^{n+1}$ and $\boldsymbol{b} \in \mathbb{A}^{k+1}$, if $f(\boldsymbol{a}, \boldsymbol{b}) = 0$ then $g(\boldsymbol{a}, \boldsymbol{b}) = 0$. Since $f$ is irreducible, this tells us that $V_{\mathbb{A}^{(n+1)+(k+1)}}(f) \subseteq V_{\mathbb{A}^{(n+1)+(k+1)}}(g)$. Then $f \mid g$ is a consequence of the original Study's lemma. □

As in the affine case and the projective case, if $\mathbb{K}$ is algebraically closed then all polynomials defining a hypersurface $C$ of $\mathbb{P}^n \times \mathbb{P}^k$ are associates (scalar multiples of each other), and so we can define the notion of an irreducible hypersurface and the *bidegree* of a hypersurface in $\mathbb{P}^n \times \mathbb{P}^k$. Identical definitions describe the irreducible components of a hypersurface in $\mathbb{P}^n \times \mathbb{P}^k$, and the same argument gives the analogue of Proposition 3.22:

**Proposition 4.51** *Suppose that $\mathbb{K}$ is algebraically closed. Let $C$ be a hypersurface in $\mathbb{P}^n \times \mathbb{P}^k$, and let $A_1, \ldots, A_m$ be irreducible hypersurfaces in $\mathbb{P}^n \times \mathbb{P}^k$ such that $\lfloor C \rfloor = \bigcup_{i \leqslant m} A_i$. Then the irreducible components of $C$ are $A_1, \ldots, A_m$.* □

Changes of coordinates of $\mathbb{P}^n$ and of $\mathbb{P}^k$ can be joined to changes of coordinates of $\mathbb{P}^n \times \mathbb{P}^k$. Let $\boldsymbol{\alpha}$ be a linear presentation of a change of coordinates $\alpha$ of $\mathbb{P}^n$, and let $\boldsymbol{\beta}$ be a linear presentation of a change of coordinates $\beta$ of $\mathbb{P}^k$. We call $\alpha \times \beta$, defined by $(p, q) \mapsto (\alpha(p), \beta(q))$, a change of coordinates of $\mathbb{P}^n \times \mathbb{P}^k$; it is a permutation of the point of $\mathbb{P}^n \times \mathbb{P}^k$. The associated change of variable $(\boldsymbol{\alpha} \times \boldsymbol{\beta})^*$ (a ring automorphism of $\mathbb{K}[\boldsymbol{x}, \boldsymbol{y}]$) has the same definition as above (induced by the invertible linear map $\boldsymbol{\alpha} \times \boldsymbol{\beta}$ from $\mathbb{A}^{(n+1)+(k+1)}$ to itself). Namely, for all $f \in \mathbb{K}[\boldsymbol{x}, \boldsymbol{y}]$,

$$(\boldsymbol{\alpha} \times \boldsymbol{\beta})^*(f) = f\big(\boldsymbol{\alpha}^*(x_0), \ldots, \boldsymbol{\alpha}^*(x_n), \boldsymbol{\beta}^*(y_0), \ldots, \boldsymbol{\beta}^*(y_k)\big).$$

The arguments involved in proving Proposition 4.30 yield:

**Proposition 4.52** *Let $f \in \mathbb{K}[\boldsymbol{x}, \boldsymbol{y}]$ be $\boldsymbol{x}, \boldsymbol{y}$-bihomogeneous. Then $(\boldsymbol{\alpha} \times \boldsymbol{\beta})^*(f)$ is also $\boldsymbol{x}, \boldsymbol{y}$-bihomogeneous, of bidegree $\deg_{\boldsymbol{x},\boldsymbol{y}} f$, and*

$$(\alpha \times \beta)\left[V_{\mathbb{P}^n \times \mathbb{P}^k}(f)\right] = V_{\mathbb{P}^n \times \mathbb{P}^k}((\boldsymbol{\alpha} \times \boldsymbol{\beta})^*(f)).$$

*Remark 4.53* Even though the naive attempt above to realise $\mathbb{P}^n \times \mathbb{P}^k$ as a subset of $\mathbb{P}^{n+k+1}$ (mapping the pair $(\pi_n(\boldsymbol{a}), \pi_k(\boldsymbol{b}))$ to $\pi_{n+k}(\boldsymbol{a}, \boldsymbol{b})$) does not work, there is a way to realise $\mathbb{P}^n \times \mathbb{P}^k$ as the subset of a projective space, of higher dimension. The *Segre embedding* is the mapping

$$(a_0 : a_1 : \cdots : a_n), (b_0 : b_1 : \cdot : b_k) \mapsto (a_0 b_0 : a_0 b_1 : \cdots : a_i b_j : \cdots : a_n b_k)$$

(we take all pairs $a_i b_j$). The image is usually not an algebraic hypersurface but it is an algebraic variety, defined as the common set of zeros of a number of polynomials, in this case $z_{i,j}z_{k,l} = z_{i,l}z_{k,j}$. «

## 4.8 Further Exercises

**Homogeneous Polynomials and First Definitions**

4.54 Suppose that $\mathbb{K}$ is finite; let $q = |\mathbb{K}|$. Show that $|\mathbb{P}^n(\mathbb{K})| = 1+q+q^2+\cdots+q^n$.

4.55 Suppose that polynomials $f$ and $g$ in $\mathbb{K}[\boldsymbol{x}]$ are homogeneous and nonzero, with $\deg f = \deg g + 1$. Suppose that $f$ and $g$ have no common factor. Show that $f + g$ is irreducible.

4.56 Find irreducible factorisations for the following polynomials in $\mathbb{C}[x, y]$: (i) $x^2 + xy + y^2$; (ii) $x^3 + y^3$; (iii) $x^3 + x^2y + xy^2 + y^3$; (iv) $x^4 + x^2y^2 + y^4$.

4.57 Suppose that $\mathbb{K}$ is algebraically closed. Let $f \in \mathbb{K}[x_1, \ldots, x_n]$ and suppose that for all $\lambda \in \mathbb{K}^*$ and all $\boldsymbol{a} \in \mathbb{K}^n$, $f(\boldsymbol{a}) = 0$ if and only if $f(\lambda\boldsymbol{a}) = 0$. Show that $f$ is homogeneous.

**Lines and Subspaces**

4.58 Find the point of intersection of the following lines in $\mathbb{P}^2(\mathbb{R})$: (i) $x + 2y = 6w$ and $3x + 4y = 15w$; (ii) $2x + 3y = 6w$ and $x = y + 3w$; (iii) $3x + y = 2w$ and $6x + 2y + 5w = 0$.

4.59 For the following pairs of points in $\mathbb{P}^2(\mathbb{R})$, find an equation for the line passing through them: (i) $(3:4:-1)$ and $(1:2:5)$; (ii) $(0:4:5)$ and $(0:1:-3)$; (iii) $(2:3:5)$ and $(0:4:1)$.

4.60 Let $n \geqslant 1$. (a) Show that the intersection of $n$ many hyperplanes in $\mathbb{P}^n$ is nonempty. Give an example though of two lines in $\mathbb{P}^3(\mathbb{R})$ whose intersection is empty. (b) Let $k \leqslant n$. Show that if $p_0, \ldots, p_k$ are $(k+1)$-many points in $\mathbb{P}^n$, which do not all lie on any $(k-1)$-dimensional subspace of $\mathbb{P}^n$, then there is a unique $k$-dimensional subspace of $\mathbb{P}^n$ containing all the points $p_0, \ldots, p_k$.

**Cubic Curves**

4.61 (a) Let $C$ be the nodal cubic curve of Exercise 3.48. Show that the map $(e:a) \mapsto (e^3:ea^2 - e^3:a^3 - e^2a)$ is well-defined map from $\mathbb{P}^1$ to the projective closure $C^\sharp$ of $C$, and that this map extends a rational parameterisation of $C$. (b) Find a similar parameterisation for the projective closure of the cuspidal cubic curve of Exercise 3.47, and for the folium of Descartes (Exercise 3.49). (c) What happens when we work over $\mathbb{C}$ instead of $\mathbb{R}$?

4.62 Let $\mathbb{K} = \mathbb{C}$. Let $Q_0 = V_{\mathbb{P}^3}(wy - x^2)$, $Q_1 = V_{\mathbb{P}^3}(wz - xy)$, and $Q_2 = V_{\mathbb{P}^3}(xz - y^2)$. Let $D = Q_0 \cap Q_1 \cap Q_2$. (a) Show that the map $(e:a) \mapsto (e^3:e^2a:ea^2:a^3)$ from $\mathbb{P}^1$ to $\mathbb{P}^3$ is well-defined, and is a bijection between $\mathbb{P}^1$ and $D$. (b) Show that for distinct $i, j \in \{0, 1, 2\}$, $Q_i \cap Q_j$ is the union of $D$ with a line in $\mathbb{P}^3$. (c) Show that no four distinct points on $D$ lie on a hyperplane of $\mathbb{P}^3$. (Hint: consider the Vandermonde determinant, Exercise 2.93. The curve $D$ is called a *twisted cubic*.)

4.63 Recall that a 2-dimensional subspace in $\mathbb{P}^n$ is called a *(projective) plane*. (a) Show that a line in $\mathbb{P}^3$ and a plane in $\mathbb{P}^3$ not containing that line intersect at a single point. (b) Let $p = (0:0:1:0)$, and let $H = V_{\mathbb{P}^3}(y)$. For $q \in \mathbb{P}^3 \setminus \{p\}$, let $\psi$ be the point of intersection of the line $\overline{pq}$ with $H$.[3] Let $D$ be the twisted cubic curve given in Exercise 4.62. Let $\mu$ be a projective map which is a bijection between $\mathbb{P}^2$ and $H$, and let $E = \mu^{-1}\psi[D]$. Show that $E$ is an algebraic curve of $\mathbb{P}^2$.

**Changes of Coordinates**

4.64 Let $\boldsymbol{\alpha}^*$ be the change of variable of $\mathbb{C}[w, x, y]$ defined by $\boldsymbol{\alpha}^*(f(w, x, y)) = f(x - 3y + w, 2x, 4x - y)$. (a) Compute the change of coordinates induced by $\boldsymbol{\alpha}$. (b) Find $\alpha[V_{\mathbb{P}^2}(y - 3x - 2w)]$ and $\alpha[V_{\mathbb{P}^2}(x^2 - y^2 - w^2)]$.

4.65 For the four points $p_0, p_1, p_2, p_3 \in \mathbb{P}^2(\mathbb{R})$ below, find a change of coordinates of $\mathbb{P}^2(\mathbb{R})$ which maps $p_0$ to $(1:0:0)$, $p_1$ to $(0:1:0)$, $p_2$ to $(0:0:1)$, and $p_3$ to $(1:1:1)$: (i) $(0:2:1)$, $(1:2:-1)$, $(0:1:0)$, $(1:3:2)$; (ii) $(1:1:0)$, $(1:2:0)$, $(0:1:1)$, $(0:1:-1)$; (iii) $(3:0:5)$, $(0:1:2)$, $(1:0:1)$, $(3:1:4)$.

4.66 Find a change of coordinates of $\mathbb{P}^2(\mathbb{R})$ which maps the $x$-axis to the line $y + x = 0$, the $y$-axis to the (projective closure of the) line $y = -1$, the line at infinity to the $x$-axis, and the point $(1, 1, 1)$ to the vertical point at infinity.

4.67 (a) Show that if $p_1, p_2$ and $p_3$ are distinct points in $\mathbb{P}^1$, and so are $q_1, q_2$ and $q_3$, then there is a unique change of coordinates of $\mathbb{P}^1$ which maps $p_i$ to $q_i$. (b) Let $p_1, p_2, p_3, p_4$ and $q_1, q_2, q_3, q_4$ be two quadruples of points in $\mathbb{P}^2$, both of

[3] The map $\psi$ is called the *projection* from the point $p$ onto $H$.

which lie in general position. Show that there is a *unique* change of coordinates of $\mathbb{P}^2$ which maps each $p_i$ to $q_i$. (c) Generalise this to $n \geqslant 3$.

4.68 (a) Let $p_1$, $p_2$ and $p_3$ be distinct *collinear* points in $\mathbb{P}^2$ and let $q_1, q_2$ and $q_3$ also be distinct collinear points in $\mathbb{P}^2$. Show that there is a change of coordinates of $\mathbb{P}^2$ mapping each $p_i$ to $q_i$. (b) Let $p_1$, $p_2$ and $p_3$ be three distinct points in $\mathbb{P}^2$ which lie on a line $\ell$. Let $\alpha$ be a change of coordinates of $\mathbb{P}^2$ which fixes each of $p_1$, $p_2$ and $p_3$. Show that $\alpha(p) = p$ for all $p \in \ell$.

4.69 Let $C$ be the curve $w^2y+wy^2 = wx^2+2wxy+x^2y+2xy^2$ in $\mathbb{P}^2(\mathbb{R})$. (a) Let $\alpha$ be the change of coordinates of $\mathbb{P}^2$ which maps $(1:0:0)$ to itself, $(0:1:0)$ to itself, $(-1:0:1)$ to $(0:0:1)$, and $(0:1:1)$ to $(1:1:1)$. Find the polynomial defining $\alpha[C]$. (b) What are the irreducible components of $C$? [Wal50, Example 1.3], [Kun05, Chap. 2, Example 2]

4.70 For $A \in \mathrm{GL}_{n+1}(\mathbb{K})$ let $\alpha_A$ be the change of coordinates of $\mathbb{P}^n$ induced by $\boldsymbol{\alpha} = T_A$. Show that $A \mapsto \alpha_A$ is a group homomorphism from $\mathrm{GL}_{n+1}(\mathbb{K})$ to the group of permutations $S(\mathbb{P}^n)$; its image is the subgroup consisting of the changes of coordinates of $\mathbb{P}^n$. What is the kernel of this map?[4]

4.71 (a) Show that the affine subspaces of $\mathbb{A}^n$ are the restrictions to $\mathbb{A}^n$ of the projective subspaces of $\mathbb{P}^n$ (see Sect. 3.4). (b) An *affine change of coordinates* is a bijective affine map $\beta\colon \mathbb{A}^n \to \mathbb{A}^n$. Show that a map from $\mathbb{A}^n$ to itself is an affine change of coordinates if and only if it is the restriction to $\mathbb{A}^n$ of a change of coordinates of $\mathbb{P}^n$ which maps the hyperplane at infinity $H_\infty$ to itself.

4.72 Show that there is no affine change of coordinates of $\mathbb{A}^2(\mathbb{R})$ that maps the parabola $y = x^2$ to the unit circle.

**Spaces of Curves**

4.73 Show the dual of the four point lemma: if $\ell_1$, $\ell_2$, $\ell_3$ and $\ell_4$ are distinct lines in $\mathbb{P}^2$, no three of which intersect at a single point $p$, and $\ell'_1$, $\ell'_2$, $\ell'_3$ and $\ell'_4$ are lines with the same property, then there is a change of coordinates $\alpha$ of $\mathbb{P}^2$ such that for each $i = 1, 2, 3, 4$, $\alpha[\ell_i] = \ell'_i$.

4.74 Suppose that $\mathbb{K}$ is algebraically closed. Show that for any $p \in \mathbb{P}^2$, the collection of curves of degree $d$ which pass through $p$ is a hyperplane of $\mathbb{G}_d$.

[4] The quotient of $\mathrm{GL}_{n+1}(\mathbb{K})$ by this kernel, which is isomorphic to the group of changes of coordinates of $\mathbb{P}^n$, is called the *projective general linear group* $\mathrm{PGL}_n(\mathbb{K})$. Projective general linear groups play a role in the classification of finite simple groups.

Conclude that every five points in $\mathbb{P}^2$ lie on a conic curve (a curve of degree 2), and that every nine points in $\mathbb{P}^2$ lie on a cubic curve.

**Möbius Maps and the Cross-Ratio**

4.75 A *fractional linear map* is the restriction to $\mathbb{A}^1(\mathbb{K})$ of a change of coordinates of $\mathbb{P}^1(\mathbb{K})$.[5] Show that the fractional linear maps are the functions of the form

$$z \mapsto \frac{az+b}{cz+d},$$

where $ad \neq bc$.[6]

4.76 Let $p_2$, $p_3$ and $p_4$ be distinct points in $\mathbb{P}^1$. By Exercise 4.67, there is a unique change of coordinates $\alpha$ of $\mathbb{P}^1$ which maps $p_2$ to $(1:1)$, $p_3$ to $(1:0)$ and $p_4$ to $(0:1)$. For any point $p_1 \in \mathbb{P}^1$, we let the *cross-ratio* of $p_1$, $p_2$, $p_3$ and $p_4$, often denoted by $(p_1, p_2; p_3, p_4)$ be the point $\alpha(p_1)$.

(a) Show that for any $p_1 \in \mathbb{P}^1$ and any change of coordinates $\beta$ of $\mathbb{P}^1$, $(\beta(p_1), \beta(p_2); \beta(p_3), \beta(p_4)) = (p_1, p_2; p_3, p_4)$. (b) Show that if $p_1$, $p_2$, $p_3$ and $p_4$ are four distinct points in $\mathbb{P}^1$, and $q_1$, $q_2$, $q_3$ and $q_4$ are also four distinct points in $\mathbb{P}^1$, then there is a change of coordinates $\alpha$ of $\mathbb{P}^1$ mapping each $p_i$ to $q_i$, if and only if $(p_1, p_2; p_3, p_4) = (q_1, q_2; q_3, q_4)$. (c) Extend part (b) to more points: if $p_1, \ldots, p_n$ and $q_1, \ldots, q_n$ are tuples of distinct points in $\mathbb{P}^1$ (with $n \geqslant 4$), then there is a change of coordinates $\alpha$ of $\mathbb{P}^1$ mapping each $p_i$ to $q_i$ if and only if for all $k = 4, \ldots, n$, $(p_1, p_2; p_3, p_k) = (q_1, q_2; q_3, q_k)$. (d) Show that an injective function $\alpha\colon \mathbb{P}^1 \to \mathbb{P}^1$ is a change of coordinates of $\mathbb{P}^1$ if and only if for any four distinct points $p_1, \ldots, p_4 \in \mathbb{P}^1$ we have $(p_1, p_2; p_3, p_4) = (\alpha(p_1), \alpha(p_2); \alpha(p_3), \alpha(p_4))$. (e) We identify distinct numbers $a_1$, $a_2$, $a_3$ and $a_4$ in $\mathbb{K}$ with points in $\mathbb{P}^1$ in the usual way and so speak of $(a_1, a_2; a_3, a_4)$ as a number in $\mathbb{K}$. (So technically, $(a_1, a_2; a_3, a_4) = \rho^{-1}(\rho(a_1), \rho(a_2); \rho(a_3), \rho(a_4))$ where recall $\rho(a) = (1:a)$. Note that indeed $(\rho(a_1), \rho(a_2); \rho(a_3), \rho(a_4)) \in \mathbb{A}^2$.) Show that

$$(a_1, a_2; a_3, a_4) = \frac{a_3 - a_1}{a_3 - a_2} \cdot \frac{a_4 - a_2}{a_4 - a_1}.$$

(This explains the name "cross-ratio".) (f) Conclude that fractional linear maps (Exercise 4.75) preserve the cross-ratio of quadruples of complex numbers.[7]

[5] Note that if $\alpha$ is a change of coordinates of $\mathbb{P}^1$ and $p_\infty = (0:1)$ is the standard point at infinity, then the fractional linear map $\alpha|_{\mathbb{A}^1}$ is not defined everywhere; it is defined on $\mathbb{A}^1 \setminus \{\alpha^{-1}(p_\infty)\}$.

[6] When $\mathbb{K} = \mathbb{C}$, fractional linear maps are also called *Möbius transformations*.

[7] It can be shown that Möbius transformations map circles and lines to circles and lines, and that four points in $\mathbb{A}^1(\mathbb{C})$ lie on a circle or a line if and only if their cross ratio is real.

4.77 Let $a_1, a_2, a_3, a_4$ be distinct numbers in $\mathbb{K}$. Let $r = (a_1, a_2; a_3, a_4)$ be the cross-ratio of $a_1$, $a_2$, $a_3$ and $a_4$ (Exercise 4.76).

Show that for all $\sigma \in S_4$,

$$\big(a_{\sigma(1)}, a_{\sigma(2)}; a_{\sigma(3)}, a_{\sigma(4)}\big) \in \{r, 1/r, 1-r, 1/(1-r), (r-1)/r, r/(r-1)\},$$

and that $(a_{\sigma(1)}, a_{\sigma(2)}; a_{\sigma(3)}, a_{\sigma(4)}) = r$ if and only if $\sigma$ is an element of the Klein Viergruppe (Exercise 2.83).

**Conic Curves**

4.78 Show that there is a change of coordinates of $\mathbb{P}^2(\mathbb{R})$ which maps the projective closure of the parabola $y = x^2$ to the projective closure of the hyperbola $xy = 1$.

4.79 Let $f \in \mathbb{C}[w, x, y]$ be a homogeneous quadratic polynomial (polynomial of degree 2). Let $C = V_{\mathbb{P}^2}(f)$. (a) Recall that a square matrix $A \in M_n(\mathbb{C})$ is *symmetric* if $A$ equals its transpose $A^{\mathsf{t}}$. Show that there is a symmetric matrix $A \in M_3(\mathbb{C})$ such that

$$f = (w, x, y)A \begin{pmatrix} w \\ x \\ y \end{pmatrix}.$$

(b) Show that if $C$ is irreducible then there is a change of coordinates $\mathbb{P}^2$ which maps $C$ to the curve $w^2 + x^2 + y^2 = 0$. (Use the fact that every symmetric matrix is diagonalisable; this was not covered in Chap. 2.) Thus, in $\mathbb{P}^2(\mathbb{C})$ there is only one irreducible conic curve up to a change of coordinates.[8] (b) What happens in $\mathbb{P}^2(\mathbb{R})$?

4.80 Let $\alpha$ be a change of coordinates of $\mathbb{P}^2$, mapping a point $p$ to a point $q$. We let

$$C = \bigcup \{\ell \cap \alpha[\ell] \,:\, p \in \ell\}.$$

(a) Show that $\ell \mapsto \alpha[\ell]$ is a bijection between the linear family of lines passing through $p$ and the linear family of lines passing through $q$. (b) Show that $p, q \in C$. (c) Suppose that $p = (0:0:1)$ and $q = (0:1:0)$. Show that there is a fractional linear transformation $g$ such that for $a \in \mathbb{K}$ and $\ell$ the line $x = a$, $\ell \cap \alpha[\ell]$ is the point $(a, g(a))$. (d) Conclude that for any $\alpha$, $p$ and $q$ as above, $C$ is contained in a conic curve.[9]

[8] For an alternative proof see Exercise 5.60.

[9] This is *Jakob Steiner's theorem*.

4.81 We use Steiner's theorem from Exercise 4.80 to give an alternative proof of the fact that five points in lie on a conic curve (Exercise 4.74). Let $p, q, a, b$ and $c$ be points in $\mathbb{P}^2$. (a) Show that if the points do not lie in general position, then they lie on a reducible conic. (b) Assume that the points lie in general position. Let $\alpha$ be a change of coordinates which maps $p$ to $q$, $a$ to $a$, $b$ to $b$, and $c$ to $c$. Show that the conic curve derived by Steiner's theorem for $\alpha$, $p$ and $q$ passes through $p, q, a, b$ and $c$.

4.82 Let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}, \mathbb{Q}\}$. Let $C$ be the projective closure of the unit circle $x^2 + y^2 = 1$ in $\mathbb{A}^2(\mathbb{K})$. (a) Show that the map $a \mapsto (a^2 + 1 : 2a : a^2 - 1)$ extends to a bijection between $\mathbb{P}^1(\mathbb{K})$ to $C$ (see Fig. 1.1). (b) Show that the complex solutions for the Pythagorean equation $w^2 = x^2 + y^2$ are the triples of the form $\left(b^2 + c^2, 2bc, b^2 - c^2\right)$ for $b, c \in \mathbb{C}$. (c) Show that the *real* solutions for the Pythagorean equation are the triples of the form $\left(\pm(b^2 + c^2), 2bc, b^2 - c^2\right)$ for $b, c \in \mathbb{R}$. (d) Show that the *integer* solutions for the Pythagorean equation are the triples of the form

$$\left((b^2 + c^2)d, 2bcd, (b^2 - c^2)d\right)$$

where $b, c, d \in \mathbb{Z}$, $b$ and $c$ are relatively prime, and not both odd; and the triples of the form

$$\left((b^2 + c^2)d/2, bcd, (b^2 - c^2)d/2\right),$$

where $b, c, d \in \mathbb{Z}$, $b$ and $c$ are relatively prime and both are odd.

**Desargues' Theorem**

4.83 The statement of Desargues' theorem makes sense in 3-dimensional space. Show that it holds in $\mathbb{P}^3$. (Separate to two cases, depending on whether the two triangles are coplanar or not.)


# Tangents

# 5

In this chapter we investigate how curves intersect with lines; recall that one of the uses of this is showing that in general, a cubic curve intersects a line in three points, and so any two points on the curve determine a third—this gives us a method of generating new solutions to an equation, given some solutions. The first section of this chapter is less formal, motivating our definitions by looking at the affine case first. Later, we make formal definitions of higher-order tangents, and multiplicity of intersection of a line with a curve. We then show that these notions are *geometric*, meaning that they are invariant under changes of coordinates. Finally, we show how each of these notions can be defined using the other (Theorem 5.34 and Proposition 5.36).

## 5.1 Introduction: Affine Tangents and Intersections with Lines

Suppose that a pair of differentiable functions $\psi_x, \psi_y \colon \mathbb{R} \to \mathbb{R}$ are used to parameterise a curve which is also given implicitly by an equation $f = 0$. Let $t_0 \in \mathbb{R}$, and let $p = (\psi_x(t_0), \psi_y(t_0))$ be the point on the curve which corresponds to "time" $t_0$. If $(\psi_x'(t_0), \psi_y'(t_0)) \neq (0, 0)$ then the tangent to the curve at the point $p$ is the line passing through $p$ whose direction is the vector $(\psi_x'(t_0), \psi_y'(t_0))$; this of course generalises the case that the curve is the graph of $\psi_y$ (so $\psi_x(t) = t$ and $f = y - \psi_y(x)$). Implicitly differentiating the equation $f(\psi_x, \psi_y) = 0$ using the chain rule, and plugging in $t = t_0$ we get

$$\frac{\partial f}{\partial x}(p) \cdot \psi_x'(t_0) + \frac{\partial f}{\partial y}(p) \cdot \psi_y'(t_0) = 0. \tag{5.1}$$



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_5

If $p = (a, b)$ then the equation of the tangent is $\psi_y'(t_0)(x - a) = \psi_x'(t_0)(y - b)$; using Eq. (5.1) we see that a scalar multiple of that equation is the equation

$$\frac{\partial f}{\partial x}(p) \cdot (x - a) + \frac{\partial f}{\partial y}(p) \cdot (y - b) = 0 \tag{5.2}$$

(using the fact that at least one of $\psi_x'(t_0)$ and $\psi_y'(t_0)$ is nonzero). Thus, even if a curve $f = 0$ is only given implicitly and without a parameterisation, we define the tangent to the curve at a point $p = (a, b)$ to be given by Eq. (5.2). Of course, this can be done only if at least one of $\frac{\partial f}{\partial x}(p)$ or $\frac{\partial f}{\partial y}(p)$ is nonzero. If both are zero then the point $p$ is called *singular* on the curve. Examples of singular points are cusps (such as in Fig. 3.2) or self-crossings (as in Fig. 3.3), but sometimes are harder to detect graphically, for example the origin on the curve $y^4 = x^3$.

Now suppose that we are given an algebraic curve $V_{\mathbb{A}^2(\mathbb{K})}(f)$ over a field $\mathbb{K}$ which is not necessarily $\mathbb{R}$ (for example $\mathbb{Q}$, but possibly even finite fields such as $\mathbb{Z}/(p)$). In these cases there is no apparent sense to the expression $\frac{\partial f}{\partial x}$ because we can't take limits. We use *formal differentiation* by simply *defining* the derivative of $\sum_k a_k x^k$ to be $\sum_k k a_k x^{k-1}$. We use the notation $D^x f$ rather than $\frac{\partial f}{\partial x}$. Partial derivatives are computed by treating other variables as constant. We give the details in Sect. 5.2 below. Using formal differentiation, we mimic Eq. (5.2) and define the tangent to a curve $f = 0$ to be the line defined by the equation

$$D^x f(p) \cdot (x - a) + D^y f(p) \cdot (y - b) = 0. \tag{5.3}$$

For simplicity now suppose that $p = o = (0, 0)$ is the origin; the following can be generalised to all points in $\mathbb{A}^2$ by translation. Write $f = f_{(m)} + f_{(m+1)} + \cdots + f_{(d)}$ where $d = \deg f$, $f_{(k)}$ is homogeneous of degree $k$, and $m$ is the least such that $f_{(m)}$ is nonzero. This number $m$ is called the *order* of the origin on the curve, denoted by $o_o(f)$. The origin $o$ lies on the curve $f = 0$ if and only if $f_{(0)} = 0$, that is, if and only if its order is greater than zero. Using Taylor's formula we see that $f_{(k)} = \sum_{i+j=k} \frac{1}{i!j!}(D^{x^i y^j} f)(o) \cdot x^i y^j$. So $m$ is the least such that some $m$th-order partial derivative of $f$ is nonzero. In particular, $o$ is a singular point on the curve $f = 0$ if and only if $m > 1$. If $o$ is nonsingular then $f_{(1)}$ is the equation of the tangent to $f$ at $o$. Suppose that $o$ is singular but also assume that $\mathbb{K}$ is algebraically closed. Since $f_{(m)}$ is homogeneous (in two variables), it is the product of $m$ linear polynomials (Proposition 4.26). These can be considered as the *higher order tangents* to the curve at $o$. See for example Fig. 5.1.

### 5.1.1 Intersection Multiplicities

If a line intersects a curve at some point, then sometimes a small perturbation of the line will result in more points of intersection. The two main examples are when the line is tangent to a curve (Fig. 5.2), and when the point is singular (Fig. 5.3); Fig. 5.4

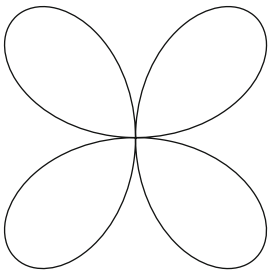


**Fig. 5.1** The *quadrifolium* $(x^2 + y^2)^3 = 4x^2y^2$. The lowest order term is $4x^2y^2$; the curve has four tangents at the origin, two copies of the $x$-axis and two copies of the $y$-axis


**Fig. 5.2** The line $\ell$ intersects the parabola $y = x^2$ at the origin. Perturbing the line a bit gives two points of intersection

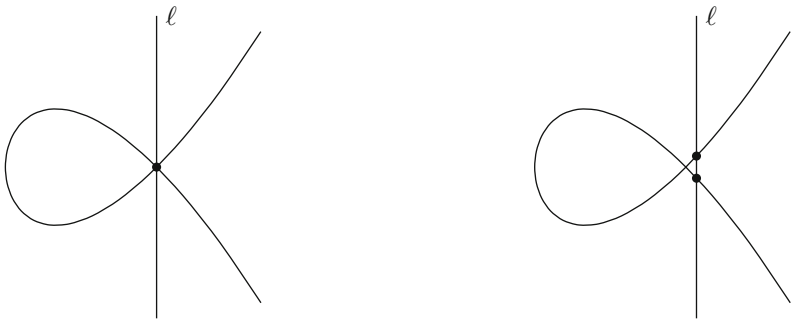



**Fig. 5.3** The line $\ell$ intersects the nodal cubic curve $y^2 = x^3 + x^2$ at the singular point $o$. Perturbing the line a bit gives two points of intersection

illustrates both cases at once. Roughly, if when moving the line a bit, we get $k$ points of intersection, we say that the *multiplicity of intersection* is $k$, or less formally, that the line intersects the curve $k$ *times* at that point.

Let $f \in \mathbb{K}[x, y]$ define a curve and let $\ell$ be a line. Let $\psi = (\psi_x, \psi_y)$ be a linear parameterisation of $\ell$ (Definition 3.27). The roots of the polynomial $f_\psi = f(\psi_x, \psi_y) \in \mathbb{K}[t]$ give the parameters which correspond to the points of intersection of the line $\ell$ with the curve $f = 0$. In the cases $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$ we can make small perturbations to the line $\ell$ by making small changes to the coefficients of the linear polynomials $\psi_x$ and $\psi_y$. This results in small changes to the coefficients of the intersection polynomial $f_\psi$, which in turn results in small

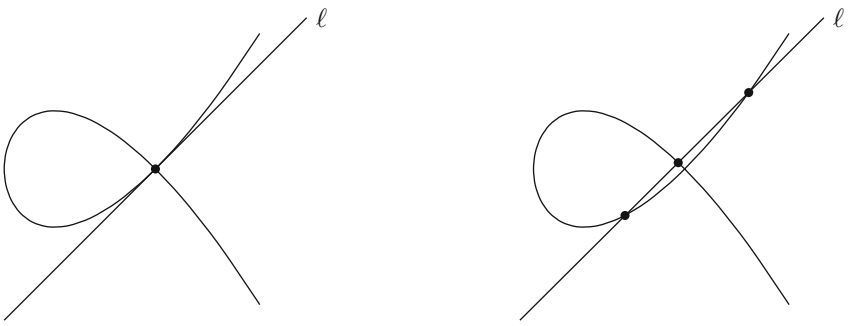



**Fig. 5.4** The line $\ell$ intersects the same nodal cubic curve $y^2 = x^3 + x^2$ at the singular point $o$. Perturbing the line a bit gives three points of intersection

changes to the roots of these polynomials (this is actually not that easy to show; see Proposition 12.29). If near a point $\lambda$, the intersection polynomials close to $f_\psi$ have $k$ distinct roots, all converging to $\lambda$ as the corresponding lines converge to $\ell$, then $\lambda$ will be a $k$-fold root of $f_\psi$: $(t-\lambda)^k$ will divide $f_\psi$ but $(t-\lambda)^{k+1}$ will not. We will then define the intersection multiplicity $i_p(f, \ell)$ of $\ell$ and the curve $f = 0$ at a point $p = \psi(\lambda)$ to be the number of times $(t-\lambda)$ divides $f_\psi$. As with tangents, this definition does not rely on continuity considerations and can be made over any field. It needs to be shown though that the intersection number $i_p(f, \ell)$ does not depend on the parameterisation $\psi$ we chose for $\ell$. Another case to be noted is when $\ell$ is a component of the curve $f = 0$; this happens if and only if $f_\psi$ is the zero polynomial. In that case we define $i_p(f, \ell) = \infty$.

*Exercise 5.1* Show that $\deg f_\psi \leqslant \deg f$. Conclude that any line, which is not a component of the curve $f = 0$, intersects that curve in at most $\deg f$ many points. «

Again for simplicity of computation, we consider the origin point $o = (0, 0)$. An affine line $\ell$ passing through the origin is given by an equation $ay = bx$, and a linear parameterisation of that line is $\psi(t) = (at, bt)$; so the intersection polynomial is $f_\psi = f(at, bt)$. Again write $f = f_{(m)} + f_{(m+1)} + \cdots + f_{(d)}$ where $m$ is the order of the origin on the curve $f = 0$. Then $f_\psi = f_{(m)}(at, bt) + \cdots + f_{(d)}(at, bt)$. Each $f_{(k)}(at, bt)$ is homogeneous of degree $k$. This shows that the multiplicity $i_o(f, \ell)$ is at least the order $o_o(f)$ of the origin on the curve. Further, the multiplicity $i_o(f, \ell)$ is strictly greater than the order $m = o_o(f)$ if and only if $(f_{(m)})_\psi = f_{(m)}(at, bt) = 0$, i.e., if and only if $\ell$ is a component of the curve $V_{\mathbb{A}^2}(f_{(m)})$. But that curve is the sum of the higher-order tangents to the curve at $o$, so $i_o(f, \ell) > m$ if and only if $\ell$ is one of these tangents. Put together, this gives us a characterisation of order and tangency in terms of multiplicity of intersection, which in fact works for all points $p$. Namely: the order of $p$ on the curve $f = 0$ is the smallest intersection

number $i_p(f, \ell)$, as $\ell$ varies over all the lines that pass through $p$. A line $\ell$ passing through $p$ is a tangent to the curve at $p$ if and only if $i_p(f, \ell) > o_p(f)$.

### 5.1.2 Homogeneous Coordinates

Since our main aim is to work in the projective plane, we need to translate the notions of tangents, order, and intersection multiplicity to homogeneous coordinates. For this we use an important identity named after Euler:

**Euler's Relation** *Let $f \in R[x_1, \ldots, x_n]$ be homogeneous of degree $d$. Then*

$$x_1 D^{x_1} f + x_2 D^{x_2} f + \cdots + x_n D^{x_n} f = d \cdot f.$$

We prove Euler's relation on p. 111 below.

Let $p = (a, b)$ be a point in the affine plane, again identified with the subset of $\mathbb{P}^2$ by choosing the line at infinity to be $w = 0$. So $\boldsymbol{p} = (1, a, b)$ is a presentation of $p$. Suppose that $p$ lies on the projective curve $f = 0$. A calculation shows that $D^x f(\boldsymbol{p}) = D^x f^\flat(a, b)$ and the same holds for $y$. Since $p$ lies on the curve, $f(\boldsymbol{p}) = 0$, so Euler's relation says that $D^w f(\boldsymbol{p}) = -a \cdot D^x f(\boldsymbol{p}) - b \cdot D^y f(\boldsymbol{p})$. This shows that the projective closure of the tangent to the curve at $p$ is neatly given by the equation

$$w \cdot D^w f(\boldsymbol{p}) + x \cdot D^x f(\boldsymbol{p}) + y \cdot D^y f(\boldsymbol{p}) = 0, \tag{5.4}$$

and that $p$ is singular if and only if all three partial derivatives vanish at $\boldsymbol{p}$. Similar equations hold for higher tangents if $p$ is singular on the curve.[1]

It is not difficult to see that the equation for the tangent to a curve $C$ at a nonsingular point $p$ does not depend on the presentation $\boldsymbol{p}$ of $p$ and also does not depend on the choice of polynomial $f$ defining $C$ (both vary only up to a nonzero scalar multiple). In this sense it is a geometric rather than merely an algebraic construct. However, for the notions of singularity and tangency to be genuinely geometric we also require that they remain invariant under changes of coordinates of the projective plane (see Sect. 4.5). This is motivated by considering changes of coordinates as relabelling of points rather than permutations of points.

*Exercise 5.2* Let $C$ be a curve in $\mathbb{P}^2(\mathbb{R})$, and let $p \in C$. Let $\alpha$ be a change of coordinates of $\mathbb{P}^2(\mathbb{R})$. Show that $p$ is singular on $C$ if and only if $\alpha(p)$ is singular on $\alpha[C]$. Suppose that $p$ is nonsingular on $C$; let $\ell$ be the tangent to $C$ at $p$. Show that $\alpha[\ell]$ is the tangent to $\alpha[C]$ at $\alpha(p)$. (Hint: use the chain rule and

[1] Note how the equation of the line only depends on the partial derivatives of $f$ at $\boldsymbol{p}$; it does not involve the coordinates of $p$. In the affine case, calculations are much simpler at the origin. Projective space does not have such a "preferred point".

Proposition 4.30(c). Below we will extend this argument to other fields by using a formal version of the chain rule.) «

*Exercise 5.3* Show that the tangent to a line $\ell$ at any point $p \in \ell$ is $\ell$ itself. «

## 5.2 Formal Partial Derivatives

We begin our formal treatment of tangency and intersection with lines by reviewing facts about formal partial derivatives. In this section let $R$ be an integral domain. For a polynomial $f = a_0 + a_1 x + \cdots + a_d x^d$ in $R[x]$ we let $D^x f = a_1 + 2a_2 x + 3a_3 x^2 + \cdots + d \cdot a_d x^{d-1}$. For now, we do not assume that $\mathbb{Z}$ is a subring of $R$, in other words, the characteristic of $R$ could be positive (see p. 40). In the expression $k \cdot a_k$, the first $k$ stands for $k1_R = 1_R + 1_R + \cdots + 1_R$ ($k$ times), so $k \cdot a_k$ is an element of $R$ and so $D^x f \in R[x]$. Positive characteristic can lead to unexpected behaviour. For example over $R = \mathbb{Z}/(p)$, $D^x x^p = 0$ even though the polynomial $x^p$ is not constant.

If $f$ is a polynomial in several variables $\boldsymbol{x} = x_1, \ldots, x_n$ with coefficients in $R$, then a partial derivative $D^{x_i}$ is defined by considering $f$ as a polynomial in $R[x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n][x_i]$.

We note that for any monic monomial $m \in R[\boldsymbol{x}]$ and any $i \leqslant n$, $D^{x_i} m$ is a monomial of total degree $\deg m - 1$, or possibly 0 if $\mathrm{char}(R) > 0$. Thus for a homogeneous polynomial $f \in R[\boldsymbol{x}]$, $D^{x_i} f$ is also homogeneous; usually of degree $\deg f - 1$, except that possibly $D^{x_i} f = 0$ if $\mathrm{char}(R) > 0$.

### 5.2.1 Properties of Derivatives

Familiar properties of the derivative hold in the formal setting. For example, for $f, g \in R[x]$, $D^x(f+g) = D^x f + D^x g$ and $D^x(fg) = f \cdot D^x g + g \cdot D^x f$. Since these are formal derivatives we cannot use manipulations of limits to show these identities. Rather, we need to calculate coefficients. If $f = \sum a_k x^k$ and $g = \sum b_k x^k$ then the coefficient of $x^{k-1}$ in both $D^x(f+g)$ and $D^x f + D^x g$ is $k(a_k + b_k)$, and so these polynomials are equal. For the product, the coefficient of $x^{k-1}$ in $D^x(fg)$ is $k \sum_{i+j=k} a_i b_j$; in $f \cdot D^x g$ is $\sum_{i+j=k} j \cdot a_i b_j$; and in $g \cdot D^x f$ is $\sum_{i+j=k} i \cdot a_i b_j$.

Iterating taking partial derivatives, we write $D^{x_i x_j} f$ for $D^{x_i}(D^{x_j} f)$, $D^{x^2} f$ for $D^x(D^x f)$, and so on. The order of differentiation does not matter: for $f \in R[x, y]$, $D^{xy} f = D^{yx} f$. Again this is done by comparing coefficients; if $f = \sum a_{m,k} x^m y^k$ then the coefficient of $x^{m-1} y^{k-1}$ in both sides is $(mk) a_{m,k}$.

Taking partial derivatives commutes with substitution: if $f \in R[\boldsymbol{x}, y]$ and $\boldsymbol{a} \in R^n$, then $(D^y f)(\boldsymbol{a}, y) = D^y(f(\boldsymbol{a}, y))$. It is important, of course, that we do not substitute a value for the variable with respect to which we take the derivative.

### The Chain Rule

Again let $\boldsymbol{x} = (x_1, \ldots, x_n)$. For $f \in R[\boldsymbol{x}]$ we let $Df$ be the row $(D^{x_1} f, D^{x_2} f, \ldots, D^{x_n} f)$. If $\boldsymbol{f} = (f_1, \ldots, f_m)$ is an $m$-tuple of polynomials from $R[\boldsymbol{x}]$ (which we really think of as a column), then we let $D\boldsymbol{f}$ be the $m \times n$-matrix $\begin{pmatrix} Df_1 \\ Df_2 \\ \vdots \\ Df_m \end{pmatrix}$. The chain rule then says that for any $g \in R[y_1, \ldots, y_m]$, $D(g(\boldsymbol{f})) = (Dg)(\boldsymbol{f}) \cdot D\boldsymbol{f}$. Unravelling, this means that for all $i \leqslant n$,

$$D^{x_i}(g(\boldsymbol{f})) = D^{x_i} f_1 \cdot (D^{y_1} g)(\boldsymbol{f}) + D^{x_i} f_2 \cdot (D^{y_2} g)(\boldsymbol{f}) + \cdots + D^{x_i} f_m \cdot (D^{y_m} g)(\boldsymbol{f}).$$

We can prove the chain rule essentially by "induction on the complexity of $g$". We prove it holds for any constant $g \in R$ (for all $\boldsymbol{f}$); we prove it holds for the variables $g = y_1, g = y_2, \ldots, g = y_m$; and then, assuming that the chain rule holds for two polynomials $g, h \in R[\boldsymbol{y}]$, we show it holds for both $g + h$ and for $gh$. Since every polynomial in $R[\boldsymbol{y}]$ can be built up from the variables in $\boldsymbol{y}$ and the constants by taking sums and products, this will show that the chain rule holds for all $g \in R[\boldsymbol{y}]$.

If $g \in R$ is a constant then the chain rule reduces to the equation $0 = 0 + 0 + \cdots + 0$. If $g = y_k$ then $D^{y_j} g = 0$ for $j \neq k$, and $D^{y_k} g = 1$, so $\sum_{j \leqslant k} D^{x_i} f_j \cdot (D^{y_j} g)(\boldsymbol{f}) = D^{x_i} f_k$, while $g(\boldsymbol{f}) = f_k$, giving the desired equality.

Now suppose that the chain rule is known to hold for two polynomials $g$ and $h$ in $R[\boldsymbol{y}]$. Then we use the sum rule and the product rule to show it holds for $g + h$ and for $gh$. For example, for the product,

$$\begin{aligned} D^{x_i}((gh)(\boldsymbol{f})) &= g(\boldsymbol{f}) D^{x_i}(h(\boldsymbol{f})) + h(\boldsymbol{f}) D^{x_i}(g(\boldsymbol{f})) = \\ &g(\boldsymbol{f}) \sum_{j \leqslant m} D^{x_i} f_j \cdot (D^{y_j}(h))(\boldsymbol{f}) + h(\boldsymbol{f}) \sum_{j \leqslant m} D^{x_i} f_j \cdot (D^{y_j}(g))(\boldsymbol{f}) = \\ &\sum_{j \leqslant m} D^{x_i} f_j \cdot \big((g D^{y_j} h)(\boldsymbol{f}) + (h D^{y_j} g)(\boldsymbol{f})\big) = \sum_{j \leqslant m} D^{x_i} f_j \cdot D^{y_j}(gh)(\boldsymbol{f}); \end{aligned}$$

the calculation for $g + h$ is easier.

*Remark 5.4* The definitions and analysis so far can be done not only for polynomials but also for formal power series (Sect. 2.1). The chain rule works if $\boldsymbol{f}$ is a tuple of formal power series, but we have proved it only if $g$ is a polynomial. Indeed the substitution $g(\boldsymbol{f})$ may be undefined if $g$ is a formal power series. We take this up in Chap. 15. «

### Euler's Relation

We can now deduce Euler's Relation: for homogeneous $f \in R[\boldsymbol{x}]$ of degree $d$, $d \cdot f = \sum_{i \leqslant n} x_i D^{x_i} f$. Let $t$ be a new variable. By Proposition 4.2, $f(t\boldsymbol{x}) = t^d f$.

The chain rule implies that

$$D^t(f(t\boldsymbol{x})) = \sum_{j\leqslant n} D^t(tx_j)\cdot(D^{x_j}f)(t\boldsymbol{x}) = \sum_{j\leqslant n} x_j(D^{x_j}f)(t\boldsymbol{x}).$$

Since $t$ does not appear in $f$, it is considered a constant when taking a formal derivative with respect to $t$; so $D^t(t^d f) = d\cdot t^{d-1}f$. Overall we get $\sum_{i\leqslant n} x_i(D^{x_i}f)(t\boldsymbol{x}) = dt^{d-1}f$; substitute $t=1$ to get Euler's relation.

**Taylor Expansions**

Now we work over a field $\mathbb{K}$. The standard proof of Taylor's formula holds for formal differentiation as well. Let $f = \sum a_k x^k$ be a polynomial in $\mathbb{K}[x]$. By induction on $k$ we see that

$$D^{x^k}f = \frac{k!}{0!}a_k + \frac{(k+1)!}{1!}a_{k+1}x + \frac{(k+2)!}{2!}a_{k+2}x^2 + \cdots;$$

Substituting 0 in $D^{x^k}f$ we get $k!a_k = (D^{x^k}f)(0)$. In other words,

$$f = \sum_{k=0}^{\deg f} \frac{(D^{x^k}f)(0)}{k!}x^k.$$

We notice though that the last step (showing that $a_k = (D^{x^k}f)(0)/k!$) involved dividing by $k!$, which seems innocuous, except that possibly $k! = 0_{\mathbb{K}}$ if $\mathrm{char}(\mathbb{K}) > 0$. Since $\mathbb{K}$ is an integral domain, if $k < \mathrm{char}(\mathbb{K})$ then $k! \neq 0_{\mathbb{K}}$, and so Taylor's formula holds provided that either $\mathrm{char}(\mathbb{K}) = 0$ or $\mathrm{char}(\mathbb{K}) > \deg f$.

*Exercise 5.5* Let $f \in \mathbb{K}[x]$ and let $a \in \mathbb{K}$. Show that if $\mathrm{char}(\mathbb{K}) = 0$ or $\mathrm{char}(\mathbb{K}) > \deg f$ then

$$f = \sum_{k=0}^{\deg f} \frac{D^{x^k}f(a)}{k!}(x-a)^k. \qquad \text{«}$$

*Exercise 5.6* Let $f \in \mathbb{K}[x, y]$. Under the same assumption, show that

$$f = \sum_{i,j} \frac{D^{x^i y^j}f(0,0)}{i!j!}x^i y^j$$

(the sum taken for pairs $(i, j)$ with $i + j \leqslant \deg f$). Generalise to more variables. «

### 5.2.2 The Discriminant

We give an application. Let $f \in R[x]$, and suppose that $\deg f \geqslant 2$. If $\mathrm{char}(R) = 0$ or $\mathrm{char}(R) > \deg f$ then $\deg D^x f = \deg f - 1 \geqslant 1$, in which case $\mathrm{res}_x(f, D^x f)$ is defined. We define the *discriminant* $\mathrm{disc}_x(f)$ to be this resultant. For now we assume that these conditions hold.

**Proposition 5.7** *Suppose that $R$ is a unique factorisation domain. Then $f$ has a nonconstant repeated factor if and only if* $\mathrm{disc}_x(f) = 0$.

***Proof*** $\mathrm{disc}_x(f) = 0$ if and only if $f$ and $D^x f$ have a nonconstant common factor (Theorem 3.12). Thus it suffices to show that $f$ and $D^x f$ have a nonconstant common factor if and only if $f$ has a repeated nonconstant factor. Since $R$ is a unique factorisation domain we can look for irreducible nonconstant factors on both sides. Indeed, we see that a nonconstant irreducible polynomial $g \in R[x]$ is common factor of $f$ and $D^x f$ if and only if $g^2$ divides $f$. Suppose that $g$ is an irreducible nonconstant factor of $f$; let $h = f/g$. Since $D^x f = g D^x h + h D^x g$ and $g$ divides $g D^x h$, it divides $D^x f$ if and only if it divides $h D^x g$. Since $\deg D^x g < \deg g$, $g$ cannot divide $D^x g$, so $g$ divides $D^x f$ if and only if $g$ divides $h$. And $g$ divides $h$ if and only if $g^2$ divides $f$. □

*Example 5.8* Let $\mathbb{K}$ be an algebraically closed field whose characteristic is not 2. Let $f = ax^2 + bx + c \in \mathbb{K}[x]$. Then $D^x f = 2ax + b$, and so

$$\mathrm{disc}_x(f) = \begin{vmatrix} c & b & a \\ b & 2a & 0 \\ 0 & b & 2a \end{vmatrix} = 4a^2c - b(2ab - ab) = a(4ac - b^2).$$

Since $f$ is the product of two linear polynomials, corresponding to the two roots of $f$, we see that $f$ has a repeated root if and only if $b^2 = 4ac$, as we know from the quadratic formula. «

*Exercise 5.9* Let $a, b \in \mathbb{K}$, and let $f = x^3 + ax + b$. Show that $\mathrm{disc}_x(f) = 4a^3 + 27b^2$. «

## 5.3 Higher Order Tangents

Having given an informal treatment of tangents and singularity in Sect. 5.1, we directly define higher-order tangents and the order of a point on a curve. For the rest of the chapter, when dealing with a curve of degree $d$, we assume that $\mathrm{char}(\mathbb{K}) = 0$ or $\mathrm{char}(\mathbb{K}) > d$. To keep things simple, we assume that $\mathbb{K}$ is algebraically closed.

We saw that to obtain the higher-order tangents, we had to keep taking all possible partial derivatives, until we obtain a nonzero equation. We do the same in homogeneous coordinates, generalising the equation of the tangent in projective coordinates.

Let $C$ be a curve in $\mathbb{P}^2$, and let $p \in \mathbb{P}^2$ be a point. Let $f$ define $C$, and let $\boldsymbol{p}$ be a presentation of $p$. We let

$$\begin{aligned}
\partial^0_{\boldsymbol{p}} f &= f(\boldsymbol{p}),\\
\partial^1_{\boldsymbol{p}} f &= w \cdot D^w f(\boldsymbol{p}) + x \cdot D^x f(\boldsymbol{p}) + y \cdot D^y f(\boldsymbol{p}),\\
\partial^2_{\boldsymbol{p}} f &= ww \cdot D^{ww} f(\boldsymbol{p}) + wx \cdot D^{wx} f(\boldsymbol{p}) + wy \cdot D^{wy} f(\boldsymbol{p}) +\\
&\quad xw \cdot D^{xw} f(\boldsymbol{p}) + xx \cdot D^{xx} f(\boldsymbol{p}) + xy \cdot D^{xy} f(\boldsymbol{p}) +\\
&\quad yw \cdot D^{yw} f(\boldsymbol{p}) + yx \cdot D^{yx} f(\boldsymbol{p}) + yy \cdot D^{yy} f(\boldsymbol{p}),
\end{aligned}$$

and in general, for $k \in \mathbb{N}$, we let

$$\partial^k_{\boldsymbol{p}} f = \sum_{\bar{v}} v_1 v_2 \cdots v_k \cdot D^{v_1 v_2 \cdots v_k} f(\boldsymbol{p});$$

the sum is taken over all *ordered* choices $\bar{v} = (v_1, \ldots, v_k)$ where each $v_i$ is one of $w$, $x$ or $y$.[2] By its definition, $\partial^k_{\boldsymbol{p}} f$ is either 0 or a homogeneous polynomial of degree $k$. If $f' = \lambda f$ and $\boldsymbol{p'} = \mu \boldsymbol{p}$ where $\lambda, \mu \in \mathbb{K}^*$, then $\partial^k_{\boldsymbol{p'}} f' = \lambda \mu^{d-k} \partial^k_{\boldsymbol{p}} f$. This shows that (since $\mathbb{K}$ is algebraically closed) we can unambiguously define:

**Definition 5.10** $\ell^k_p C = V_{\mathbb{P}^2}\left(\partial^k_{\boldsymbol{p}} f\right)$.

This does not depend on the choice of $f$ defining $C$ and presentation $\boldsymbol{p}$ of $p$. The curve $\ell^k_p C$, defined by the equation $\partial^k_{\boldsymbol{p}} f = 0$, is the $k$th-order tangent to $C$ at $p$.

*Remark 5.11* The main usage of $\mathbb{K}$ being algebraically closed in this section is for showing that the higher-order tangent is the sum of lines (Corollary 5.22). Other than that, we use algebraic closure when passing between a curve and the association class of the polynomials defining it. Many of the results in this section hold even if $\mathbb{K}$ is not algebraically closed, provided that we distinguish between non-associate polynomials, even if they define the same curve. In the next section we also use algebraic closure for counting the number of intersections of a line with a curve (Theorem 5.27). «

[2] Formally, over all functions from $\{1, 2, \ldots, k\} \to \{w, x, y\}$.

Let $w^m x^i y^j$ be a monic monomial, and let $k = m + i + j$. The number of choices $\bar{v}$ as above yielding $v_1 v_2 \cdots v_k = w^m x^i y^j$ is

$$\binom{k}{m} \cdot \binom{k-m}{i} = \frac{k!}{m!(k-m)!} \cdot \frac{(k-m)!}{i!(k-m-i))!} = \frac{k!}{m!i!j!}$$

(we first choose $m$ many locations for $w$, and from the rest we choose $i$ many locations for $x$). Since the order of differentiation does not matter we see that

$$\partial_{\boldsymbol{p}}^k f = k! \cdot \sum_{m+i+j=k} \frac{D^{w^m x^i y^j} f(\boldsymbol{p})}{m!i!j!} w^m x^i y^j \tag{5.5}$$

(the sum is taken over all triples $(m, i, j)$ with $m + i + j = k$.)

**Lemma 5.12** *Let $d = \deg C$. For all $p \in \mathbb{P}^2$, $\ell_p^d C = C$.*

***Proof*** Let $h \in \mathbb{K}[w, x, y]$ be a monomial of degree $d$: $h = a w^m x^i y^j$. If $m' + i' + j' = d$ but $(m, i, j) \neq (m', i', j')$ then $D^{w^{m'} x^{i'} y^{j'}} h = 0$, as either $m' > m$ or $i' > i$ or $j' > j$. On the other hand $D^{w^m x^i y^j} h = m!i!j!a$. This shows that (for any nonzero $\boldsymbol{p} \in \mathbb{K}^3$) $\partial_{\boldsymbol{p}}^d h = d! \cdot h$. This equality is preserved under addition, and so for any homogeneous polynomial $f \in \mathbb{K}[w, x, y]$ of degree $d$ we have $\partial_{\boldsymbol{p}}^d f = d! \cdot f$. The lemma follows because we assumed that $d! \neq 0_{\mathbb{K}}$. □

In particular, we conclude that for all $p \in \mathbb{P}^2$, $\ell_p^d C \neq \mathbb{P}^2$ ($\partial_{\boldsymbol{p}}^d f$ is nonzero). This makes the following definition meaningful:

**Definition 5.13** Let $C$ be a curve in $\mathbb{P}^2$, and let $p \in \mathbb{P}^2$. The *order* of $p$ on $C$, denoted by $o_p(C)$, is the least natural number $k \leqslant \deg C$ such that $\ell_p^k C \neq \mathbb{P}^2$.

In other words, $o_p(C)$ is the least $k$ such that some $k$th-order partial derivative of $f$ defining $C$ is nonzero at $\boldsymbol{p}$. We have $\ell_p^0 C = \mathbb{P}^2$ if $p \in C$, and otherwise $\ell_p^0 C = \varnothing$. This shows that $o_p(C) > 0$ if and only if $p \in C$. If $o_p(C) \geqslant 2$ then we say that $p$ is *singular* on $C$. If $o_p(C) = 2$ we call $p$ a *double point* of $C$, if $o_p(C) = 3$ then we call it a *triple point*, etc. A *singular curve* is a curve which has a singular point. A curve is *nonsingular* if it is not singular, i.e., if no point $p \in C$ is singular on $C$.

**Definition 5.14** We let $\ell_p C = \ell_p^k C$ for $k = o_p(C)$.

So if $p \in C$ is nonsingular on $C$ then $\ell_p C$, defined by Eq. (5.4), is the tangent to $C$ at $p$. We will show that $\ell_p C$ is the sum of $o_p(C)$ many lines, each of which we call a tangent to $C$ at $p$ (as mentioned above, this does use the assumption that $\mathbb{K}$ is algebraically closed).

*Exercise 5.15* Let $C$ be the projective closure of the nodal cubic $y^2 = x^3 + x^2$ (Exercise 3.48). (a) Show that the origin $o = (1:0:0)$ is the unique singular point on $C$. (b) Show that $\ell_o C$ is the sum of the two lines $y = x$ and $y = -x$. (c) Verify by calculation that $\ell_p^3 C = C$ for all $p \in \mathbb{P}^2$. «

### The Affine Higher Order Tangent

Let $f \in \mathbb{K}[x, y]$ be nonzero. Let $C = V_{\mathbb{P}^2}(f^\sharp)$ be the projective closure of the affine curve $f = 0$. We consider higher tangents at the origin $o = (1:0:0) = (0, 0)$. Let $\boldsymbol{o} = (1, 0, 0)$. We elaborate upon the calculations we made in Sect. 5.1.

The values of the partial derivatives of $f^\sharp$ at $\boldsymbol{o}$ can be copied over from those of $f$: for all $i$, $j$ and $m$ with $m + i + j \leqslant d$,

$$D^{w^m x^i y^j} f^\sharp(\boldsymbol{o}) = \frac{(d - (i + j))!}{(d - (m + i + j))!} D^{x^i y^j} f(0, 0). \tag{5.6}$$

This is proved by induction on $m$. For $m = 0$, $D^{x^i y^j} f^\sharp(\boldsymbol{o}) = D^{x^i y^j} f(0, 0)$ follows from definitions: powers of $w$ in $f^\sharp$ are treated as constant and then evaluate to 1. For $m > 1$, we use Euler's Relation: since $D^{w^m x^i y^j} f^\sharp$ is homogeneous, of degree $d - (m + i + j)$,

$$w D^{w^{m+1} x^i y^j} f^\sharp + x D^{w^m x^{i+1} y^j} f^\sharp + y D^{w^m x^i y^{j+1}} f^\sharp = (d - (m + i + j)) \cdot D^{w^m x^i y^j} f^\sharp;$$

substituting $\boldsymbol{o}$ gives $D^{w^{m+1} x^i y^j} f^\sharp(\boldsymbol{o}) = (d - (m + i + j)) D^{w^m x^i y^j} f^\sharp(\boldsymbol{o})$.

Write $f = f_{(0)} + \cdots f_{(d)}$ with $f_{(k)}$ homogeneous of degree $k$. By Exercise 5.6,

$$f_{(k)} = \sum_{i+j=k} \frac{D^{x^i y^j} f(0, 0)}{i! j!} x^i y^j.$$

Let $k$ be the least such that $f_{(k)} \neq 0$. Then whenever $i + j < k$, $D^{x^i y^j} f(0, 0) = 0$. Then Eqs. (5.5) and (5.6) imply:

(a) For all $k' < k$, $\partial_{\boldsymbol{o}}^{k'}(f^\sharp) = 0$;
(b) $\partial_{\boldsymbol{o}}^{k}(f^\sharp) = k! f_{(k)}$.

So $o_o(C) = k$, and $\ell_o C = V_{\mathbb{P}^2}(f_{(k)})$. Now since $f_{(k)} \in \mathbb{K}[x, y]$ is homogeneous of degree $k$, and $\mathbb{K}$ is algebracially closed, $f_{(k)}$ is the product of $k$-many linear homogeneous polynomials in $\mathbb{K}[x, y]$, i.e., polynomials of the form $ay - bx$ where $(a, b) \neq (0, 0)$; these define lines which pass through the origin. We summarise:

**Proposition 5.16** *Let $f \in \mathbb{K}[x, y]$ be nonzero, and let $C = V_{\mathbb{P}^2}(f^\sharp)$.*

(a) *The order $o_o(C)$ of the origin on $C$ is the least $k$ such that $f_{(k)} \neq 0$. This is the least $k$ such that some $k$th-order partial derivative $D^{x^i y^j} f$ does not vanish at $(0, 0)$.*

(b) $\ell_o C = V_{\mathbb{P}^2}\big(f_{(o_{o(C)})}\big)$, *and is the sum of* $o_o(C)$*-many lines which pass through the origin.* □

*Exercise 5.17* Let $f \in \mathbb{K}[x, y]$; let $C = V_{\mathbb{P}^2}(f^\sharp)$; let $p = (a, b) \in \mathbb{A}^2$. Show that if $o_p(C) = 1$ then the restriction of $\ell_p C$ to $\mathbb{A}^2$ is given by Eq. (5.3). «

### 5.3.1 The Moduli Space of Tangents

We work toward showing that the order of a point on a curve and the higher order tangent are geometric concepts: they are invariant under changes of coordinates. We consider the parameterised collection of all tangents. For a curve $C$ in $\mathbb{P}^2$ and $k \leqslant \deg C$ we let

$$\ell^k C = \left\{(p, q) \in \mathbb{P}^2 \times \mathbb{P}^2 \,:\, q \in \ell_p^k C\right\}.$$

This is the underlying set of a hypersurface of $\mathbb{P}^2 \times \mathbb{P}^2$ (see Sect. 4.7). Fix the variables $\boldsymbol{u} = (u_0, u_1, u_2)$ and $\boldsymbol{v} = (v_0, v_1, v_2)$ which we will use to define hypersurfaces of $\mathbb{P}^2 \times \mathbb{P}^2$. Then

$$\partial^k(f(\boldsymbol{u})) = \sum_{\sigma\,:\,\{1,2,\ldots,k\}\to\{0,1,2\}} v_{\sigma(1)} v_{\sigma(2)} \cdots v_{\sigma(k)} \cdot D^{u_{\sigma(1)} u_{\sigma(2)} \cdots u_{\sigma(k)}}(f(\boldsymbol{u})) \tag{5.7}$$

is $\boldsymbol{u}, \boldsymbol{v}$-bihomogeneous (of bidegree $(d-k, k)$) and defines a hypersurface of $\mathbb{P}^2 \times \mathbb{P}^2$ whose underlying set is $\ell^k C$; here of course $f$ defines $C$ and $f(\boldsymbol{u})$ is the substitution which replaces $w$ by $u_0$, $x$ by $u_1$ and $y$ by $u_2$. For a presentation $\boldsymbol{p}$ of a point $p \in \mathbb{P}^2$,

$$\partial_{\boldsymbol{p}}^k f = \Big(\partial^k(f(\boldsymbol{u}))\Big)(\boldsymbol{p}, w, x, y).$$

In general, for $g \in \mathbb{K}[\boldsymbol{u}, \boldsymbol{v}]$ we let

$$\partial g = v_0 D^{u_0} g + v_1 D^{u_1} g + v_2 D^{u_2} g.$$

We can define $\partial^k g$ in a similar way by replacing $f(\boldsymbol{u})$ by $g$ in Eq. (5.7); the point is that this can be done by iterating $\partial$ for $k$ times, i.e. $\partial^{k+1} g = \partial(\partial^k g)$. So we can reason about $\partial^k$ inductively in a way which we cannot with $\partial_{\boldsymbol{p}}^k$. Note that if $g$ is $\boldsymbol{u}, \boldsymbol{v}$-bihomogeneous, say of bidegree $(d, e)$, then $\partial^k g$ is also bihomogeneous, of bidegree $(d-k, e+k)$.

As a first application we show the following, which says that the $k$th-order tangent at $p$ does indeed pass through $p$.

**Proposition 5.18** *Let $C$ be a curve in $\mathbb{P}^2$, and let $p \in C$. Then for all $k \leqslant \deg C$, $p \in \ell^k_p C$.*

We argue indirectly.

**Lemma 5.19** *Let $g \in \mathbb{K}[\boldsymbol{u}, \boldsymbol{v}]$ be bihomogeneous and let $p \in \mathbb{P}^2$. If $(p, p) \in V_{\mathbb{P}^2\times\mathbb{P}^2}(g)$ then $(p, p) \in V_{\mathbb{P}^2\times\mathbb{P}^2}(\partial g)$.*

***Proof*** Let $\boldsymbol{p} = (e, a, b)$ be a presentation of $p$; so we are assuming that $g(\boldsymbol{p}, \boldsymbol{p}) = 0$, and need to show that $(\partial g)(\boldsymbol{p}, \boldsymbol{p}) = 0$. Let $d = \deg_{\boldsymbol{u}} g$. Euler's relation applied with respect to $\boldsymbol{u}$ gives

$$u_0 D^{u_0} g + u_1 D^{u_1} g + u_2 D^{u_2} g = d \cdot g.$$

Substituting $\boldsymbol{p}$ for both $\boldsymbol{u}$ and $\boldsymbol{v}$, we get

$$e \cdot (D^{u_0} g)(\boldsymbol{p}, \boldsymbol{p}) + a \cdot (D^{u_1} g)(\boldsymbol{p}, \boldsymbol{p}) + b \cdot (D^{u_2} g)(\boldsymbol{p}, \boldsymbol{p}) = d \cdot g(\boldsymbol{p}, \boldsymbol{p}) = 0.$$

The expression on the left equals $(\partial g)(\boldsymbol{p}, \boldsymbol{p})$. □

***Proof of Proposition 5.18*** Let $f$ define $C$. We treat $f(\boldsymbol{u})$ as a polynomial in $\mathbb{K}[\boldsymbol{u}, \boldsymbol{v}]$ (in which the variables $\boldsymbol{v}$ do not appear). Thus $V_{\mathbb{P}^2\times\mathbb{P}^2}(f(\boldsymbol{u})) = C \times \mathbb{P}^2$. Since $p \in C$, $(p, p) \in V_{\mathbb{P}^2\times\mathbb{P}^2}(f(\boldsymbol{u}))$. By induction on $k$, Lemma 5.19 shows that $(p, p) \in V_{\mathbb{P}^2\times\mathbb{P}^2}(\partial^k(f(\boldsymbol{u})))$, so $(p, p) \in \ell^k C$. □

### 5.3.2 Invariance of the Higher Order Tangent

Recall (see Sect. 4.7) that if $\alpha$ is a change of coordinates of $\mathbb{P}^2$ then $\alpha \times \alpha$ is a change of coordinates of $\mathbb{P}^2 \times \mathbb{P}^2$; if $\boldsymbol{\alpha}^*$ is an associated change of variable then $g \mapsto g\,(\boldsymbol{\alpha}^*(u_0), \boldsymbol{\alpha}^*(u_1), \boldsymbol{\alpha}^*(u_2), \boldsymbol{\alpha}^*(v_0), \boldsymbol{\alpha}^*(v_1), \boldsymbol{\alpha}^*(v_2))$ is an associated change of variable $(\boldsymbol{\alpha} \times \boldsymbol{\alpha})^*$ of $\mathbb{K}[\boldsymbol{u}, \boldsymbol{v}]$, where of course $\begin{pmatrix} \boldsymbol{\alpha}^*(u_0) \\ \boldsymbol{\alpha}^*(u_1) \\ \boldsymbol{\alpha}^*(u_2) \end{pmatrix} = A^{-1} \cdot \begin{pmatrix} u_0 \\ u_1 \\ u_2 \end{pmatrix}$ and $\begin{pmatrix} \boldsymbol{\alpha}^*(v_0) \\ \boldsymbol{\alpha}^*(v_1) \\ \boldsymbol{\alpha}^*(v_2) \end{pmatrix} = A^{-1} \cdot \begin{pmatrix} v_0 \\ v_1 \\ v_2 \end{pmatrix}$; $A$ is the matrix such that $\boldsymbol{\alpha} = T_A$.

Fix such $\boldsymbol{\alpha}$. For brevity let $\hat{g} = (\boldsymbol{\alpha} \times \boldsymbol{\alpha})^*(g)$ for all $g \in \mathbb{K}[\boldsymbol{u}, \boldsymbol{v}]$.

**Lemma 5.20** *For all $g \in \mathbb{K}[\boldsymbol{u}, \boldsymbol{v}]$, $\hat{\partial g} = \partial\left(\hat{g}\right)$.*

***Proof*** Let $B = A^{-1}$, so $B = (b_{i,j})_{i,j\in\{0,1,2\}}$. Let $i, j \in \{0, 1, 2\}$. Since $\hat{u_j} = \sum_i b_{j,i} u_i$ we have $D^{u_i} \hat{u_j} = b_{j,i}$. Since $\hat{v_j}$ does not mention $u_i$, $D^{u_i} \hat{v_j} = 0$. The

chain rule then implies that

$$\begin{pmatrix} D^{u_0}\hat{g} \\ D^{u_1}\hat{g} \\ D^{u_2}\hat{g} \end{pmatrix} = B^{\mathsf{t}} \cdot \begin{pmatrix} D^{\hat{u}_0} g \\ D^{\hat{u}_1} g \\ D^{\hat{u}_2} g \end{pmatrix}$$

and so

$$\partial(\hat{g}) = (v_0, v_1, v_2) \cdot \begin{pmatrix} D^{u_0}\hat{g} \\ D^{u_1}\hat{g} \\ D^{u_2}\hat{g} \end{pmatrix} = (v_0, v_1, v_2) \cdot B^{\mathsf{t}} \cdot \begin{pmatrix} D^{\hat{u}_0} g \\ D^{\hat{u}_1} g \\ D^{\hat{u}_2} g \end{pmatrix}.$$

Of course (as a $1 \times 1$-matrix) $\partial(\hat{g})$ is equal to its transpose and so

$$\partial(\hat{g}) = (D^{\hat{u}_0} g, D^{\hat{u}_1} g, D^{\hat{u}_2} g) \cdot B \cdot \begin{pmatrix} v_0 \\ v_1 \\ v_2 \end{pmatrix} = (D^{\hat{u}_0} g, D^{\hat{u}_1} g, D^{\hat{u}_2} g) \cdot \begin{pmatrix} \hat{v}_0 \\ \hat{v}_1 \\ \hat{v}_2 \end{pmatrix} = \hat{\partial g},$$

using the fact that $(\boldsymbol{\alpha} \times \boldsymbol{\alpha})^*$ is a ring homomorphism. □

By induction on $k$ we see that

$$\hat{\partial^k} g = \partial^k \hat{g}.$$

We use this to show that the higher-order tangents are invariant under changes of coordinates.

**Proposition 5.21** *Let $C$ be a curve in $\mathbb{P}^2$, and let $p \in \mathbb{P}^2$. Let $\alpha$ be a change of coordinates of $\mathbb{P}^2$. Then for all $k \leqslant \deg C$,*

$$\ell^k_{\alpha(p)} \alpha[C] = \alpha \left[ \ell^k_p C \right],$$

*and so $o_{\alpha(p)}(\alpha[C]) = o_p(C)$.*

***Proof*** Let $f$ define $C$, and let $\boldsymbol{p}$ be a presentation of $p$. Let $\boldsymbol{\alpha}$ be linear presentation of $\alpha$. Let $\boldsymbol{q} = \boldsymbol{\alpha}(\boldsymbol{p})$, so $\boldsymbol{q}$ is a presentation of $q = \alpha(p)$. We show that $\boldsymbol{\alpha}^* \left( \partial^k_{\boldsymbol{p}} f \right) = \partial^k_{\boldsymbol{q}} (\boldsymbol{\alpha}^*(f))$ (recall Proposition 4.30).

Again let $\hat{g} = (\boldsymbol{\alpha} \times \boldsymbol{\alpha})^*(g)$. By definition, $\boldsymbol{\alpha}^*(w)(\boldsymbol{u}) = \hat{u_0}$, $\boldsymbol{\alpha}^*(w)(\boldsymbol{v}) = \hat{v_0}$, $\boldsymbol{\alpha}^*(x)(\boldsymbol{u}) = \hat{u_1}$, and so on. From this we conclude:

(a) $\hat{f}(\boldsymbol{u}) = (\boldsymbol{\alpha}^*(f))(\boldsymbol{u})$: We get the same result if we first replace the variables of $f$ by $\boldsymbol{u}$ and then apply $(\boldsymbol{\alpha} \times \boldsymbol{\alpha})^*$, or if we first apply $\boldsymbol{\alpha}^*$ to $f$ and then substitute $\boldsymbol{u}$ for $(w, x, y)$.

(b) For all $g \in \mathbb{K}[\boldsymbol{u}, \boldsymbol{v}]$, $\boldsymbol{\alpha}^*(g(\boldsymbol{p}, w, x, y)) = \hat{g}(\boldsymbol{q}, w, x, y)$; for this recall that $\boldsymbol{\alpha}^*$ is defined using $\boldsymbol{\alpha}^{-1}$ rather than $\boldsymbol{\alpha}$.

Applying (b) to $g = \partial^k(f(\boldsymbol{u}))$,

$$\boldsymbol{\alpha}^*\left(\partial^k_{\boldsymbol{p}} f\right) = \boldsymbol{\alpha}^*\left(\partial^k(f(\boldsymbol{u}))(\boldsymbol{p}, w, x, y)\right) = \partial^k(\hat{f}(\boldsymbol{u}))(\boldsymbol{q}, w, x, y).$$

By Lemma 5.20, $\partial^k(\hat{f}(\boldsymbol{u})) = \partial^k(f(\hat{\boldsymbol{u}}))$ and so by (a),

$$\partial^k(\hat{f}(\boldsymbol{u}))(\boldsymbol{q}, w, x, y) = \partial^k\left((\boldsymbol{\alpha}^*(f))(\boldsymbol{u})\right)(\boldsymbol{q}, w, x, y) = \partial^k_{\boldsymbol{q}}(\boldsymbol{\alpha}^*(f))$$

as required. □

**Corollary 5.22** *Let $C$ be a curve in $\mathbb{P}^2$ and let $p \in \mathbb{P}^2$. Then $\ell_p C$ is the sum of $o_p(C)$ many lines, each of which passes through $p$.*

***Proof*** After a change of coordinates we may assume that the line at infinity $\ell_\infty$ is not a component of $C$, and that $p = o$ is the origin. Then $C$ is the projective closure of its restriction to $\mathbb{A}^2$, and the result follows from Proposition 5.16. □

These lines are called the *higher order tangents* to $C$ at $p$. Some tangents can repeat. A singular point $p$ is called *ordinary* if $\ell_p C$ is the sum of $o_p(C)$ *distinct* lines.

**Corollary 5.23** *Let $C$ and $D$ be curves in $\mathbb{P}^2$ and let $p \in \mathbb{P}^2$. Then $o_p(C + D) = o_p(C) + o_p(D)$ and $\ell_p(C + D) = \ell_p C + \ell_p D$.*

***Proof*** Again change coordinates so that $\ell_\infty$ is a component of neither $C$ nor $D$ and $p$ is the origin. Let $f$ define $C$ and $g$ define $D$; let $k = o_p(C)$ and $m = o_p(D)$. Then $f = f_{(k)} +$ higher order terms and $g = g_{(m)} +$ higher order terms, and so $fg = f_{(k)}g_{(m)} +$ higher order terms. Again use Proposition 5.16. □

As a result, for any curve $C$ and point $p \in \mathbb{P}^2$, $o_p(C) \geqslant m_p(C)$, the multiplicity of $p$ on $C$ (the number of irreducible components of $C$ that $p$ lies on). We will use this to show that a nonsingular curve is irreducible (Theorem 6.5).

## 5.4 The Intersection of a Line with a Curve

As indicated in Sect. 5.1 we develop the notion of intersection multiplicity with a line and relate it to the higher-order tangents. We first work in the projective plane.

### 5.4.1 Definition of Intersection Multiplicity

Let $C$ be a curve in $\mathbb{P}^2$, and let $f$ define $C$. Let $\psi$ be a projective linear parameterisation of a projective line $\ell$ (Definition 4.16); let $\boldsymbol{\psi}$ be a linear presentation of $\psi$. Recall that we let $\psi_w, \psi_x, \psi_y$ be the linear homogeneous polynomials in $\mathbb{K}[s,t]$ which define the components of $\boldsymbol{\psi}$.

Define

$$f_{\boldsymbol{\psi}} = f(\psi_w, \psi_x, \psi_y)$$

which is a polynomial in $\mathbb{K}[s,t]$. This polynomial is homogeneous, of the same degree as $f$, or it is the zero polynomial. For $r \in \mathbb{P}^1$, $f_{\boldsymbol{\psi}}(r) = 0$ if and only if the point $\psi(r)$ lies on $C$. In other words, the roots of $f_{\boldsymbol{\psi}}$ in $\mathbb{P}^1$ correspond by the map $\psi$ to the points of intersection of $C$ and $\ell$. This shows that $f_{\boldsymbol{\psi}} = 0$ if and only if $\ell \subseteq C$.

Provided $\ell$ is not a component of $C$, we want to define the multiplicity of intersection of $C$ and $\ell$ at a point $\alpha(r)$ to be the multiplicity of $r$ as a root of $f_{\boldsymbol{\psi}}$ (its multiplicity in the multiset $V_{\mathbb{P}^1}(f_{\boldsymbol{\psi}})$). This does not depend on the choice of $f$ defining $C$; by Exercise 4.15, it also does not depend on the choice of linear presentation $\boldsymbol{\psi}$ of the parameterisation $\psi$. However, we also need to ensure that this does not depend on the choice of parameterisation $\psi$ of $\ell$.

**Lemma 5.24** *Let $\psi$ and $\varphi$ be two linear parameterisations of a line $\ell$ in $\mathbb{P}^2$; let $\boldsymbol{\psi}$ and $\boldsymbol{\varphi}$ be linear presentations of $\psi$ and $\varphi$. Let $f \in \mathbb{K}[w,x,y]$ be homogeneous. Then for all $p \in \ell$, the multiplicity of $\psi^{-1}(p)$ in $V_{\mathbb{P}^1}(f_{\boldsymbol{\psi}})$ is the same as the multiplicity of $\varphi^{-1}(p)$ in $V_{\mathbb{P}^1}(f_{\boldsymbol{\varphi}})$.*

***Proof*** Let $W = \operatorname{range}\boldsymbol{\psi} = \operatorname{range}\boldsymbol{\varphi}$ be the 2-dimensional subspace of $\mathbb{K}^3$ such that $\pi_2[W] = \ell$. Then $\boldsymbol{\theta} = \boldsymbol{\varphi}^{-1} \circ \boldsymbol{\psi}$ is a linear isomorphism from $\mathbb{K}^2$ to itself; it is a linear presentation of the change of coordinates $\theta = \varphi^{-1} \circ \psi$ of $\mathbb{P}^1$. To prove the lemma, we show that $\varphi^{-1} \circ \psi$ maps $V_{\mathbb{P}^1}(f_{\boldsymbol{\psi}})$ to $V_{\mathbb{P}^1}(f_{\boldsymbol{\varphi}})$; equivalently, by Proposition 4.30(d), that $\boldsymbol{\theta}^*(f_{\boldsymbol{\psi}}) = f_{\boldsymbol{\varphi}}$. This can be checked "manually", but we can also consider functions: $f_{\boldsymbol{\varphi}}$ defines the function $f \circ \boldsymbol{\varphi} \colon \mathbb{K}^2 \to \mathbb{K}$, while by Proposition 4.30(c), $\boldsymbol{\theta}^*(f_{\boldsymbol{\psi}})$ defines the same map $f \circ \boldsymbol{\psi} \circ \boldsymbol{\theta}^{-1} = f \circ \boldsymbol{\psi} \circ \boldsymbol{\psi}^{-1} \circ \boldsymbol{\varphi}$. Since $\mathbb{K}$ is infinite, this shows the equality of the polynomials (Proposition 2.18). □

We can therefore define:

**Definition 5.25** Let $C$ be a curve in $\mathbb{P}^2$, let $\ell$ be a line which is not a component of $C$, and let $p \in \ell$. We let $i_p(C,\ell)$ be the multiplicity of $\psi^{-1}(p)$ in $V_{\mathbb{P}^1}(f_{\boldsymbol{\psi}})$, where $f$ defines $C$ and $\psi$ is a linear parameterisation of $\ell$.

We extend this definition to lines $\ell$ which are components of $C$ by letting $i_p(C, \ell) = \infty$ for all $p \in \ell$. In any case, we see that for all $p \in \ell$, $p \in \ell \cap C$ if and only if $i_p(C, \ell) > 0$.

*Example 5.26* Let $f \in \mathbb{K}[w, x, y]$ be homogeneous. The map $\psi(s:t) = (0:s:t)$ is a linear parameterisation of the line at infinity $w = 0$, so the multiplicity of intersection of the curve $f = 0$ with the line at infinity $\ell_\infty$ at a point $(0:a:b)$ is the multiplicity of the root $(a:b)$ of the polynomial $f(0, x, y)$. «

**Bézout for a Line**

We get a special case of Bézout's theorem. It says that if "counted properly", i.e., with intersection multiplicities, then every line $\ell$ intersects a curve $C$ in exactly $\deg C$ many points (or is a component of $C$). Here it is essential that we: (i) work in the projective plane; and (ii) assume that $\mathbb{K}$ is algebraically closed.

**Theorem 5.27** *Let $C$ be a curve in $\mathbb{P}^2$, and suppose that a line $\ell$ is not a component of $C$. Then*

$$\sum_{p \in \ell \cap C} i_p(C, \ell) = \deg C.$$

***Proof*** Let $\boldsymbol{\psi}$ be a presentation of a linear parameterisation of $\ell$; pick $f$ defining $C$. The polynomial $f_{\boldsymbol{\psi}}$ has degree $\deg C$ and so defines a hypersurface of $\mathbb{P}^1$ which is a multiset containing $\deg C$ many points (Proposition 4.26). □

*Exercise 5.28* Let $C$ and $D$ be curves in $\mathbb{P}^2$, let $\ell$ be a line, and let $p \in \ell$. Show that $i_p(C + D, \ell) = i_p(C, \ell) + i_p(D, \ell)$. «

### 5.4.2 Invariance of Multiplicity of Intersection with a Line

We show that multiplicity of intersection with a line is invariant under changes of coordinates, and is thus a geometric notion.

**Proposition 5.29** *Let $C$ be a curve in $\mathbb{P}^2$, $\ell$ be a line, and $p \in \ell$. Let $\alpha$ be a change of coordinates of $\mathbb{P}^2$. Then $i_{\alpha(p)}(\alpha[C], \alpha[\ell]) = i_p(C, \ell)$.*

We give two proofs of this proposition.

***First proof of Proposition 5.29*** Pick $f$ defining $C$ and a presentation $\boldsymbol{\psi}$ of a linear parameterisation $\psi$ of $\ell$. Let $\boldsymbol{\alpha}$ be a linear presentation of $\alpha$. Then $\boldsymbol{\alpha} \circ \boldsymbol{\psi}$ is a presentation of the linear parameterisation $\alpha \circ \psi$ of $\alpha[\ell]$. To prove the lemma, it suffices to verify that $(\boldsymbol{\alpha}^*(f))_{\boldsymbol{\alpha} \circ \boldsymbol{\psi}} = f_{\boldsymbol{\psi}}$; as in the proof of Lemma 5.24, this follows from the fact that both polynomials define the function $f \circ \boldsymbol{\psi}$. □

We give another proof of Proposition 5.29, in the spirit of the proof of Proposition 5.21. This proof will be used in the next chapter.

To set things up we again fix a linear automorphism $\boldsymbol{\alpha}$ of $\mathbb{K}^3$. We now introduce eight new variables: $\boldsymbol{u} = (u_0, u_1, u_2)$, $\boldsymbol{v} = (v_0, v_1, v_2)$, and $(s, t)$. We extend the theory of products of projective spaces to the product of three spaces: $\mathbb{P}^2 \times \mathbb{P}^2 \times \mathbb{P}^1$. Trihomogeneous polynomials (in $\boldsymbol{u}, \boldsymbol{v}, (s, t)$) define hypersurfaces of this product. The map $\boldsymbol{\alpha} \times \boldsymbol{\alpha} \times \mathrm{id}_{\mathbb{K}^2}$ is a linear automorphism of $\mathbb{K}^3 \times \mathbb{K}^3 \times \mathbb{K}^2$ which induces a change of coordinates $\alpha \times \alpha \times \mathrm{id}_{\mathbb{P}^1}$ of $\mathbb{P}^2 \times \mathbb{P}^2 \times \mathbb{P}^1$ and a change of variable $(\boldsymbol{\alpha} \times \boldsymbol{\alpha} \times \mathrm{id}_{\mathbb{K}^2})^*$ of $\mathbb{K}[\boldsymbol{u}, \boldsymbol{v}, s, t]$. We abbreviate $(\boldsymbol{\alpha} \times \boldsymbol{\alpha} \times \mathrm{id}_{\mathbb{K}^2})^*(g)$ by $\hat{g}$.

Let $f \in \mathbb{K}[w, x, y]$. In the same way that $\partial^k$ uniformises the $k$th-order tangents, the general intersection polynomial

$$f_{\boldsymbol{u},\boldsymbol{v}} = f(s\boldsymbol{u} + t\boldsymbol{v})$$

uniformises the intersection polynomials $f_{\boldsymbol{\psi}}$ (see Example 4.17).

**Lemma 5.30** $(\boldsymbol{\alpha}^*(f))_{\boldsymbol{u},\boldsymbol{v}} = \hat{f_{\boldsymbol{u},\boldsymbol{v}}}$.

***Proof*** As in the proof of Proposition 5.21 (where we used $\boldsymbol{\alpha}^*(x)(\boldsymbol{u}) = \hat{u_1}$ etc.), $\boldsymbol{\alpha}^*(x)(s\boldsymbol{u} + t\boldsymbol{v}) = s\hat{u_1} + t\hat{v_1}$ etc., and so

$$(\boldsymbol{\alpha}^*(f))_{\boldsymbol{u},\boldsymbol{v}} = f\left(\boldsymbol{\alpha}^*(w), \boldsymbol{\alpha}^*(x), \boldsymbol{\alpha}^*(y)\right)(s\boldsymbol{u} + t\boldsymbol{v}) = f\left(s\hat{\boldsymbol{u}} + t\hat{\boldsymbol{v}}\right) =$$

$$f_{\boldsymbol{u},\boldsymbol{v}}(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}, s, t) = \hat{f_{\boldsymbol{u},\boldsymbol{v}}}. \quad \square$$

***Second proof of Proposition 5.29*** Using the notation of the first proof, we again need to show that $(\boldsymbol{\alpha}^*(f))_{\boldsymbol{\alpha}\circ\boldsymbol{\psi}} = f_{\boldsymbol{\psi}}$. Let $\boldsymbol{p} = \boldsymbol{\psi}(1, 0)$ and let $\boldsymbol{q} = \boldsymbol{\psi}(0, 1)$; so $\boldsymbol{\psi}(s, t) = s\boldsymbol{p} + t\boldsymbol{q}$. Now $f_{\boldsymbol{\psi}} = f_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{p}, \boldsymbol{q}, s, t)$, and

$$(\boldsymbol{\alpha}^*(f))_{\boldsymbol{\alpha}\circ\boldsymbol{\psi}} = (\boldsymbol{\alpha}^*(f))_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{\alpha}(\boldsymbol{p}), \boldsymbol{\alpha}(\boldsymbol{q}), s, t) = \hat{f_{\boldsymbol{u},\boldsymbol{v}}}(\boldsymbol{\alpha}(\boldsymbol{p}), \boldsymbol{\alpha}(\boldsymbol{q}), s, t).$$

Since the tuple of polynomials $(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ defines the map $\boldsymbol{\alpha}^{-1} \times \boldsymbol{\alpha}^{-1}$ on $\mathbb{K}^6$, we see that for all $(\boldsymbol{a}, \boldsymbol{b}) \in \mathbb{K}^6$ and $g \in \mathbb{K}[\boldsymbol{u}, \boldsymbol{v}, s, t]$,

$$\hat{g}(\boldsymbol{a}, \boldsymbol{b}, s, t) = g\left(\boldsymbol{\alpha}^{-1}(\boldsymbol{a}), \boldsymbol{\alpha}^{-1}(\boldsymbol{b}), s, t\right).$$

We apply this to $g = f_{\boldsymbol{u},\boldsymbol{v}}$ and $\boldsymbol{a} = \boldsymbol{\alpha}(\boldsymbol{p})$, $\boldsymbol{b} = \boldsymbol{\alpha}(\boldsymbol{q})$. $\square$

### Affine Calculations

Let $\psi$ be an *affine* linear parameterisation of an affine line $\ell$ in $\mathbb{A}^2$ (Definition 3.27). Then $\psi$ extends to a projective linear parameterisation $\psi^\sharp$ of the projective closure $\ell^\sharp$ of $\ell$: if $\psi = (\psi_x, \psi_y) = (a_x t + b_x, a_y t + b_y)$ then we let $\psi^\sharp(s:t) =$

$(s:\psi_x^{\sharp s}:\psi_y^{\sharp s}) = (s:a_xt+b_xs:a_yt+b_ys)$; so $\psi^\sharp(1:t)$ is the point $\psi(t)$ (under the usual identification of $\mathbb{A}^2$ in $\mathbb{P}^2$), while $\psi^\sharp(0:1)$ is the point of intersection of $\ell^\sharp$ with the line at infinity.

*Remark 5.31* Note that not every projective linear parameterisation $\psi = (\psi_w:\psi_x:\psi_y)$ is obtained as this "projective closure" of an affine linear parameterisation: if $\psi_w(s,t)$ is not a multiple of $s$ (that is, if $\psi(0:1)$ is not on the line at infinity), then the restriction of $\psi$ to $\mathbb{A}^1$ is not linear but the rational map $(\psi_x^{\flat s}/\psi_w^{\flat s}, \psi_y^{\flat s}/\psi_w^{\flat s})$. «

Let $f \in \mathbb{K}[x,y]$; let $C = V_{\mathbb{A}^2}(f)$; let $C^\sharp = V_{\mathbb{P}^2}(f^\sharp)$ be the projective closure of $C$. For $p \in \mathbb{A}^2$ we write $i_p(C,\ell)$ for $i_p(C^\sharp,\ell^\sharp)$. Now

$$\left((f^\sharp)_{\boldsymbol{\psi}^\sharp}\right)^{\flat s} = f^\sharp(1,\psi_x,\psi_y) = f(\psi_x,\psi_y);$$

so $V_{\mathbb{P}^1}((f^\sharp)_{\boldsymbol{\psi}^\sharp}) \cap \mathbb{A}^1 = V_{\mathbb{A}^1}(f_\psi)$, where $f_\psi = f(\psi_x,\psi_y)$. We conclude:

**Lemma 5.32** *Let $\ell$ be a line in $\mathbb{A}^2$, let $C = V_{\mathbb{A}^2}(f)$ be a curve, and let $\psi$ be an affine parameterisation of $\ell$. Then $i_p(C,\ell)$ is the multiplicity of $\psi^{-1}(p)$ as a root of $f_\psi$.*

This allows us to simplify calculations: for a projective curve $D$, a projective line $L$, and a point $p \in \mathbb{A}^2$, we can find $i_p(D,L)$ by using the affine equation of $D\restriction_{\mathbb{A}^2}$ and an affine parameterisation of $L\restriction_{\mathbb{A}^2}$.

*Example 5.33* If $\ell$ is the line $ay = bx$ which passes through the origin, we can use the affine parameterisation $\psi(t) = (at,bt)$; then $i_o(C,\ell)$ is the multiplicity of 0 as a root of $f(at,bt)$. «

### 5.4.3 Tangents and Intersections with Lines

We can now carry out the analysis of tangency and intersection numbers that was mentioned in Sect. 5.1.

**Theorem 5.34** *Let $C$ be a curve in $\mathbb{P}^2$ and let $p \in \mathbb{P}^2$. A line $\ell$ passing through $p$ is a tangent to $C$ at $p$ if and only if $i_p(C,\ell) > o_p(C)$. If not, then $i_p(C,\ell) = o_p(C)$.*

***Proof*** The concepts of order, $k$th-order tangents and multiplicity of intersection are invariant under changes of coordinates (Propositions 5.21 and 5.29). So we may change coordinates so that $p = o$ is the origin, and the line at infinity is not a component of $C$. Pick $f \in \mathbb{K}[x,y]$ which defines the restriction $C\restriction_{\mathbb{A}^2}$. Then by

Example 5.33, $i_p(C, \ell)$ is the multiplicity of 0 as a root of $f(at, bt)$, where $\ell$ is the line $ay = bx$. Write $f = \sum_k f_{(k)}$, with $f_{(k)}$ homogeneous of degree $k$. Then $f(at, bt) = \sum_{k \leqslant d} f_{(k)}(at, bt)$, and $f_{(k)}(at, bt)$ is homogeneous of degree $k$. By Proposition 5.16, $o_p(C)$ is the least $k$ such that $f_{(k)} \neq 0$, so $i_p(C, \ell) \geqslant o_p(C)$. This of course includes the case $i_o(C, \ell) = \infty$, i.e., when $\ell$ is a component of $C$.

Let $m = o_p(C)$. By Proposition 5.16, $V_{\mathbb{P}^2}(f_{(m)})$ is the sum of the $m$-many tangents to $C$ at $p$, and $\ell$ is one of these tangents if and only if $(f_{(m)})_\psi = f_{(m)}(at, bt) = 0$. So $\ell$ is a tangent to $C$ at $p$ if and only if $i_p(C, \ell) > m$. □

### Defining Multiplicities Using Tangents

Theorem 5.34 shows how to define order and tangents given intersection multiplicity with lines. We go the other direction. Let $f \in \mathbb{K}[w, x, y]$ be homogeneous, defining a curve $C$; let $\boldsymbol{p} = (p_0, p_1, p_2)$ be a presentation of a point $p \in \mathbb{P}^2$. Let $\ell$ be a line passing through $p$; let $\boldsymbol{q} = (q_0, q_1, q_2)$ be a presentation of a point $q \in \ell$ other than $p$. Use the presentation $\boldsymbol{\psi}(s, t) = s\boldsymbol{p} + t\boldsymbol{q}$ of a parameterisation of $\ell$. By definition, $i_p(C, \ell)$ is the multiplicity of the root $(1:0)$ of $f_{\boldsymbol{\psi}}$. This is the same as the multiplicity of 0 as a root of the dehomogenisation $(f_{\boldsymbol{\psi}})^{\flat s} = f(\boldsymbol{p} + t\boldsymbol{q})$. In turn, this is the least $k$ such that $\left(D^{t^k}\left(f(\boldsymbol{p} + t\boldsymbol{q})\right)\right)(0) \neq 0$.

**Lemma 5.35** *For all $k \leqslant d$,*

$$\left(D^{t^k}\left(f(\boldsymbol{p} + t\boldsymbol{q})\right)\right)(0) = (\partial^k_{\boldsymbol{p}} f)(\boldsymbol{q}).$$

***Proof*** Use variables $\boldsymbol{u}, \boldsymbol{v}, t$ as above. For brevity let $\boldsymbol{h} = \boldsymbol{p} + t\boldsymbol{v}$. For any $g \in \mathbb{K}[\boldsymbol{u}, \boldsymbol{v}]$, by the chain rule, $D^t(g(\boldsymbol{h}, \boldsymbol{v})) = (\partial g)(\boldsymbol{h}, \boldsymbol{v})$, and so by induction, $D^{t^k}(g(\boldsymbol{h}, \boldsymbol{v})) = (\partial^k g)(\boldsymbol{h}, \boldsymbol{v})$. Note that $f(\boldsymbol{h}) = (f(\boldsymbol{u}))(\boldsymbol{h}, \boldsymbol{v})$, and so overall $D^{t^k}(f(\boldsymbol{h})) = \left(\partial^k (f(\boldsymbol{u}))\right)(\boldsymbol{h}, \boldsymbol{v})$. The lemma is obtained by substituting $\boldsymbol{v} = \boldsymbol{q}$ and $t = 0$. □

As a result, we see that $i_p(C, \ell)$ is the least $k$ such that $q \notin \ell^k_p C$. But $q \in \ell$ was arbitrary. On the other hand, if $p \in C$ then $p \in \ell^k_p C$ for all $k$ (Proposition 5.18). This shows:

**Proposition 5.36** *Let $C$ be a curve in $\mathbb{P}^2$, let $p \in C$ and let $\ell$ be a line passing through $p$. Let $k = i_p(C, \ell)$. Then $\ell \subseteq \ell^{k'}_p C$ for all $k' < k$, but $\ell \cap \ell^k_p C = \{p\}$.* □

If $\ell$ is a component of $C$ this means that $\ell \subseteq \ell^k_p C$ for all $k \leqslant \deg C$. Note that Proposition 5.36 gives us another proof of Theorem 5.34.

*Example 5.37*

(i) Consider the parabola $y = x^2$. It is nonsingular; the tangent to the parabola at the origin is the $x$-axis. The intersection multiplicity of this tangent with the parabola at the origin is greater than 1 (the order of the origin on the parabola),

and is at most 2 since the degree of the parabola is 2. This can be verified by direction computation.

(ii) The origin is an ordinary double point of the cubic curve $y^2 = x^3 + x^2$ (Exercises 3.48 and 5.15). Ordinary double points are sometimes called *nodes*, and so this is called the *nodal* cubic. The tangents must intersect the curve thrice at the origin.

(iii) The origin is a double point on the cubic $y^2 = x^3$ (Exercise 3.47), but the two tangents coincide. A non-ordinary double point is sometimes called a *cusp*, whence this is the *cuspidal cubic*. «

### 5.4.4 Simple Intersections Are the Norm

We show that most lines in a linear family of lines intersect a curve simply.

**Proposition 5.38** *Let $C$ be a curve in $\mathbb{P}^2$ which has no repeated component. Let $\mathcal{L}$ be the linear family of lines which pass through some point $q$ (see Proposition 4.40). For all but finitely many lines $\ell \in \mathcal{L}$, for all $p \neq q$ on $\ell$, $i_p(C, \ell) \leqslant 1$.*

***Proof*** By changing coordinates, we may assume that $q = (0:0:1)$ is the vertical point at infinity; so $\mathcal{L}$ is the family of vertical lines. The restrictions to $\mathbb{A}^2$ of the lines in $\mathcal{L}$ (other than the line at infinity) are the affine vertical lines $x = a$ for $a \in \mathbb{K}$, and the points on such a line $x = a$ are of course the points $(a, b)$ for $b \in \mathbb{K}$.

Let $f$ be the dehomogenisation of a polynomial which defines $C$. For each $a \in \mathbb{K}$, $t \mapsto (a, t)$ is a linear parameterisation of the affine line $x = a$, so by Lemma 5.32, for all $p = (a, b) \in \mathbb{A}^2$, $i_p(C, x = a)$ is the multiplicity of $b$ as a root of $f(a, t)$. (If $f(a, t)$ is the zero polynomial then the projective vertical line $x = aw$ is a component of $C$.)

We view $f$ as a polynomial in $\mathbb{K}[x][y]$. Suppose that $\deg_y f \leqslant 1$. Then for all $a \in \mathbb{K}$, $\deg f(a, t) \leqslant 1$; for all but finitely many $a \in \mathbb{K}$, $f(a, t)$ is not the zero polynomial. For such $a$, for all $b$, the multiplicity of $b$ as a root of $f(a, t)$ is at most 1.

Suppose then that $\deg_y f \geqslant 2$. Since $C$ has no repeated components, $f$ has no repeated factor. In particular, it has no repeated factor in $\mathbb{K}[x, y] \setminus \mathbb{K}[x]$. Hence, the discriminant $g = \operatorname{disc}_y(f)$ is nonzero (Proposition 5.7). For all but finitely many $a \in \mathbb{K}$, $\deg_y f(a, y) = \deg_y f$ and $\deg_y D^y f(a, y) = \deg_y D^y f$. (Why? write $f = \sum_{i=0}^{d} f_i y^i$, with $f_d \neq 0$; the polynomial $f_d$ has finitely many roots.) Since $(D^y f)(a, y) = D^y(f(a, y))$, by Lemma 3.15, for such $a$, $g(a) = \operatorname{disc}_y(f(a, y))$. Since $g \neq 0$, it has finitely many roots. Hence, by Proposition 5.7 again, for all but finitely many $a \in \mathbb{K}$, $f(a, y)$ has no repeated roots, i.e., the multiplicity of any root $b$ of $f(a, t)$ is at most 1. □

**Corollary 5.39** *Let $C$ be a curve in $\mathbb{P}^2$ which has no repeated component; let $\mathcal{L}$ be the linear family of lines which pass through some point $q$. Let $m = o_q(C)$. Then for all but finitely many lines $\ell \in \mathcal{L}$, $\ell$ intersects $C$ at $\deg C - m$ many distinct points other than $q$.*

***Proof*** By Theorem 5.34, for all but finitely many $\ell \in \mathcal{L}$, $i_q(C, \ell) = o_q(C)$. The corollary then follows from Proposition 5.38 and Theorem 5.27. □

The following shows that the degree of a curve can be recovered by examining intersections with lines.

**Corollary 5.40** *Let $C$ be a curve in $\mathbb{P}^2$ which has no repeated components; let $\mathcal{L}$ be the linear family of lines which pass through some point $q$. Suppose that $q \notin C$, or that $q \in C$ and is nonsingular on $C$. Then all but finitely many lines $\ell \in \mathcal{L}$ intersect $C$ at $\deg C$-many distinct points.*

***Proof*** If $q \notin C$ then $o_q(C) = 0$, and the corollary follows directly from Corollary 5.39. Otherwise, $o_q(C) = 1$; then by Corollary 5.39, for all but finitely many $\ell \in \mathcal{L}$, $\ell$ intersects $C$ in $\deg C - 1$ many distinct points other than $q$; but of course, every $\ell \in \mathcal{L}$ intersects $C$ at $q$ as well. □

**Corollary 5.41** *A curve with no repeated components has only finitely many singular points.*

***Proof*** Let $C$ be a curve with no repeated components. Let $q$ be any point not on $C$. Let $\mathcal{L}$ be the family of lines which pass through $q$. Let $\ell_1, \ldots, \ell_m$ be the lines in $\mathcal{L}$ which intersect $C$ at some point $p$ in multiplicity greater than 1. By Theorem 5.34, if $p \in C$ is singular on $C$, then the line $\overline{pq}$ is one of the lines $\ell_1, \ldots, \ell_m$. Each such line $\ell_i$ intersects $C$ in at most finitely many points. □

For bounds on the number of singular points, see Exercises 6.8 and 6.58.

## 5.5 Further Exercises

When calculating we assume $\mathbb{K} = \mathbb{C}$.

**Differentiating Polynomials**

5.42 Suppose that $F$ is an algebraically closed field, and suppose that $\mathrm{char}(F) = 0$ or $\mathrm{char}(F) > n$. Show that $F$ contains $n$ distinct $n$th *roots of unity*, namely elements $a \in F$ such that $a^n = 1_F$.

5.43 Let $f \in \mathbb{K}[x_1, \ldots, x_n]$ and suppose that $\sum x_i D^{x_i} f = d \cdot f$. Show that $f$ is homogeneous of degree $d$ (or is 0).

**Theory**

5.44

(a) Let $f \in \mathbb{K}[w, x, y]$ be homogeneous, and let $k < \deg f$. Let $p \in \mathbb{P}^2$. Suppose that each $k$th-order partial derivative of $f$ vanishes at $p$, i.e. that $p \in V_{\mathbb{P}^2}(D^{w^i x^j y^m} f)$ whenever $i + j + m = k$. Show that $o_p\left(V_{\mathbb{P}^2}(f)\right) > k$. (That is: we're not assuming that for $k' < k$, all partial derivatives of order $k'$ vanish at $p$, but that follows.)

(b) Let $d \geqslant 1$. Assume that $\mathbb{K}$ is algebraically closed. Let $p \in \mathbb{P}^2$ and let $k \leqslant d$. Show that the collection of curves $C$ of degree $d$ such that $o_p(C) \geqslant k$ is a projective subspace of $\mathbb{G}_d$ of co-dimension $\binom{k+1}{2}$ (see Sect. 4.6 and Exercise 4.74; by a change of coordinates, you may assume that $p$ is the origin).

5.45 Let $p = (a, b)$ be a singular point on an affine curve $f = 0$. Show that $p$ is an ordinary double point of the curve if and only if $((D^{xy} f)(a, b))^2 \neq (D^{x^2} f)(a, b) \cdot (D^{y^2} f)(a, b)$.

5.46 Show directly (without using Corollary 5.23) that if $C$ is a curve and $p$ lies on more than one component of $C$ (i.e. $m_p(C) > 1$) then $p$ is singular on $C$.

5.47 Let $C$ be a curve in $\mathbb{P}^2$, let $p \in C$, and suppose that $o_p(C) = \deg(C)$. Show that $C$ is the sum of $\deg(C)$ many lines (not necessarily distinct), all passing through $p$.

5.48 Let $C$ be a curve in $\mathbb{P}^2$, and let $\mathcal{L}$ be a linear family of lines. Show that only finitely many lines $\ell \in \mathcal{L}$ are tangent to $C$ at any point.

**Examples**

5.49 Is the graph of the sine function $\{(a, \sin a) : a \in \mathbb{R}\}$ an algebraic curve of $\mathbb{A}^2(\mathbb{R})$?

5.50 Let $f \in \mathbb{K}[x]$ be nonconstant. Show that the projective closure $D$ of $y = f$ intersects the line at infinity only at the vertical point at infinity. Show that the order of that point on $D$ is $\deg f - 1$, and that the $\ell_\infty$ is the only tangent to $D$ at that point.

5.51 Let $C$ be the projective closure of the curve $y^3 = x^3 + 3xy + 1$. Find the points of intersection of $C$ with the line at infinity. Without calculating derivatives, explain why these points are nonsingular on $C$. (Recall that $\mathbb{K} = \mathbb{C}$, not $\mathbb{R}$.)

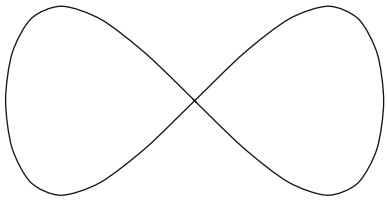


**Fig. 5.5** The eight curve $y^2 = x^2 - x^4$

5.52 For the following affine curves, find: the order of the curve at the origin; the tangents to the curve at the origin; and how many times each tangent intersects the curve at the origin: (i) $y = x^3 + 2x$; (ii) $y^2 = x^3$; (iii) $(x^2 + y^2)^2 = xy^2$.

5.53 Give an example of an irreducible curve $C$ in $\mathbb{P}^2(\mathbb{C})$ and a point $p$ such that: (i) $p$ is nonsingular on $C$ and the tangent to $C$ at $p$ intersects $C$ at $p$ five times ($i_p(C, \ell_p C) = 5$). (ii) $p$ is a double point on $C$ and $C$ has a single tangent at $p$ (so $\ell_p C$ consists of two copies of a line) which intersects $C$ at $p$ five times. (iii) $p$ is a double point on $C$ and $C$ has two tangents at $p$, one of which intersects $C$ at $p$ four times and the other seven times.

5.54 Let $C$ be the curve $y^2 = x^2 - x^4$ (Fig. 5.5). Calculate the intersection multiplicities of $C$ at the origin with every line which passes through the origin; use this to find the order of the origin on $C$, and the tangents to $C$ at the origin.

5.55 Find the singular points on the projective closures of the following curves and find the tangents at each singular point:

(i) $y^3 + x^3 + 3xy^2 + 3x^2y + 2xy = y^2 + x^2$;
(ii) $x^4 + y^4 = x^2y^2$;
(iii) $y^2 = (5 - x^2)(4x^4 - 20x^2 + 25)$.

[Kir92, Example 2.2],[Gib98, Example 7.3.3],[Ful69, Example 3.2]

5.56 An *asymptote* of an affine curve $C$ is an affine line, whose projective closure is tangent to the projective closure of $C$, at a point which lies on the line at infinity. (a) Find an asymptote of the folium of Descartes $x^3 + y^3 = 3xy$ (Exercise 3.49 and Fig. 3.4). (b) Find asymptotes of the hyperbola $xy = 1$. (c) Show that the parabola $y - x^2$ and that the nodal cubic $y^2 = x^3 + x^2$ do not have asymptotes.

**Conic Curves**

5.57 Show that a conic curve is singular if and only if it is reducible.

5.58 Let $f = 2a_{w,x}wx + 2a_{w,y}wy + 2a_{x,y}xy + a_{w,w}w^2 + a_{x,x}x^2 + a_{y,y}y^2$ be a homogeneous quadratic polynomial; let $C = V_{\mathbb{P}^2}(f)$. Show that $C$ is singular if and only if the matrix

$$\begin{pmatrix} a_{w,w} & a_{w,x} & a_{w,y} \\ a_{w,x} & a_{x,x} & a_{x,y} \\ a_{w,y} & a_{x,y} & a_{y,y} \end{pmatrix}$$

is singular. Let $p = (e:a:b)$ be a nonsingular point of $C$; show that $\ell_p C = V_{\mathbb{P}^2}(eD^w f + aD^x f + bD^y f)$.

5.59 Let $C$ be an irreducible conic curve in $\mathbb{P}^2$. (a) Show that no line is a tangent to $C$ at more than one point. (b) Show that no three distinct points on $C$ are collinear.

5.60 In this exercise we show that up to changes of coordinates there is only one irreducible conic curve in $\mathbb{P}^2$. (Compare with Exercise 4.79.) Let $C$ be an irreducible conic curve in $\mathbb{P}^2(\mathbb{C})$. (By Exercise 5.57 $C$ is nonsingular; by Exercise 5.59, distinct points on $C$ have distinct tangents.) (a) Show that we can change coordinates so that the vertical point at infinity $(0:0:1)$ lies on $C$ and the tangent to $C$ at that point is the line at infinity; and further the origin lies on $C$ and the tangent at that point is the $x$-axis. (b) Show that after such a change of coordinates there is some $a \in \mathbb{C}^*$ such that $C = V_{\mathbb{P}^2}(x^2 + awy)$. (c) Show that after another change of coordinates, $C$ is the projective parabola $V_{\mathbb{P}^2}(x^2 - wy)$. (d) What happens in $\mathbb{P}^2(\mathbb{R})$? In $\mathbb{P}^2(\mathbb{Q})$?

5.61 Let $C$ be an irreducible conic curve in $\mathbb{P}^2$, and let $p \in \mathbb{P}^2$. Show that if $p \notin C$ then there are exactly two lines which pass through $p$ and are tangent to $C$ at some point of $C$. (Hint: change coordinates to $w^2 + x^2 + y^2 = 0$ and use Exercise 5.58.)

**Singular Cubic Curves**

5.62 Let $C$ be a cubic curve in $\mathbb{P}^2$. (a) Show that if $C$ is irreducible then $C$ has at most one singular point (Consider the line passing through two singular points). (b) How many singular points does $C$ have if it is the sum of a line and an irreducible conic? How many if it is the sum of three lines?

5.63 In this exercise we show that up to changes of coordinates there are two singular irreducible cubic curves, the nodal and the cuspidal. Let $C$ be an irreducible singular cubic curve in $\mathbb{P}^2$. By Exercise 5.62, $C$ has a unique singular point $p$. (a) Show that $p$ is a double point. (b) Show that if $p$ is ordinary then $C$ can be mapped by a change of coordinates to the projective closure of $x^3 + y^3 = 3xy$, (the

folium of Descartes, Fig. 3.4). Hence all projective nodal cubic curves (including the projective closure of $y^2 = x^3 + x^2$) are the same up to changes of coordinates. (Hint: first determine the singular point and its tangents, so that the cubic is given by an equation $ax^3 + by^3 + cx^2y + dy^2x + ewxy$. Note that $a, b, e$ are all nonzero since $C$ is irreducible. Then ensure that $a = b = 1$ and $e = -1$. Finally change the variable $w$ to $w + cx + dy$.) (c) Show that if $p$ is not ordinary then $C$ can be mapped to the projective closure of $y^2 = x^3$ by a change of coordinates. Hence all projective cuspidal cubic curves are the same up to changes of coordinates. (d) Explain why a cuspidal cubic curve cannot be mapped by a change of coordinates to a nodal cubic curve.

5.64 Let $C$ be the projective curve $x^3 + y^3 + w^3 = 3xyw$. Find the singular points of $C$; find the irreducible components of $C$.

# Bézout's Theorem

# 6

Bézout's theorem states that two curves with no common component intersect each other in the maximal number of points, generalising the case of the intersection of a curve and a line (Theorem 5.27). The main challenge is to give a definition of the multiplicity of intersection between two curves.

As with lines, the intuition is that two curves $C$ and $D$ intersect $k$ times at a point $p$ if when we move the curves just a little bit, we will get $k$ many points of intersection. When we work over the complex field $\mathbb{C}$, the topological structure does allow us to perturb curves; Proposition 13.47 and Exercise 13.76 are perhaps the closest we get to the original idea, at least for lines. Over general fields, the notion of proximity does not have any meaning, and so we need another approach.

The definition for the intersection with lines (Definition 5.25) uses parameterisations of lines. However, most curves do not admit rational parameterisations. One possible approach is to use more general parameterisations, using analytic, rather than rational functions. Using these more complicated parameterisations, we can mimic the definition for lines. Some work goes into showing the existence of parameterisations. We take this up in Chap. 15.

In this chapter we will use the resultant. Very roughly, the idea is as follows. Let $C = V_{\mathbb{A}^2}(f)$ and $D = V_{\mathbb{A}^2}(g)$ be curves, and let $p \in C \cap D$. By a change of coordinates, we arrange that $p$ is the unique point of intersection of $C$ and $D$ on some vertical line $x = a$. The polynomials $f(a, y)$ and $g(a, y)$ have a common factor, namely $y - b$ where $p = (a, b)$; so the resultant $\operatorname{res}_y(f, g)$ (with respect to $y$) evaluates to 0 at $x = a$. The multiplicity of intersection is then the multiplicity of zero of this resultant at $x = a$, i.e., the number of times $(x - a)$ divides the resultant.

This is a relatively simple definition, but there are some drawbacks. The main technical drawback is that it is difficult to prove that this definition gives a result which is invariant under changes of coordinates. We will need to work quite hard to show this. In fact, what we will do, is give a more elaborate definition (Definition 6.18) which makes both Bézout's theorem and geometric invariance straightforward, and then show that it is equivalent to the original definition



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_6

(Proposition 6.25); and furthermore, both definitions agree with the definition that we gave in the previous chapter when one of the curves is a line (Proposition 6.27).

The definitions using the resultant are perhaps not very intuitive. Another reason to believe that the definition we give is a reasonable one, is to show that it is *categorical* (Theorem 6.39). Namely: we show that our definition satisfies a list of properties, which are to be expected of any reasonable definition of intersection multiplicity; and then show that there can be at most one assignment of numbers to intersections that satisfies all these properties.

Throughout, we assume that our base field $\mathbb{K}$ is algebraically closed, and that the characteristic of $\mathbb{K}$ is either 0 or greater than the degrees of the polynomials we are dealing with, so that formal derivatives of nonconstant polynomials do not vanish.

## 6.1 A First Look at the Intersection of Curves

Let $f, g \in \mathbb{K}[w, x, y]$ be homogeneous. Let $(e:a) \in \mathbb{P}^1$. By Theorem 3.12, $\operatorname{res}_y(f(e,a,y), g(e,a,y)) = 0$ if and only if $f(e,a,y)$ and $g(e,a,y)$ have a common root, if and only if the projective curves $f = 0$ and $g = 0$ intersect at a point which lies on the vertical line $ex = aw$, other than the vertical point at infinity $(0:0:1)$. This is because the line contains the vertical point at infinity and the points $(e:a:b)$ for $b \in \mathbb{K}$.

If $\deg_y f = \deg f(e,a,y)$ and $\deg_y g = \deg g(e,a,y)$ then

$$\operatorname{res}_y(f(e,a,y), g(e,a,y)) = \big(\operatorname{res}_y(f,g)\big)(e,a)$$

(Lemma 3.15). This certainly happens if the monomial $y^{\deg f}$ appears in $f$ and in $g$. We make use of the following:

*Remark 6.1* Let $f \in \mathbb{K}[w, x, y]$ be homogeneous of degree $d$. If the vertical point at infinity does not lie on the curve $f = 0$ then the monomial $y^d$ appears in $f$: all other monomials contain either $w$ or $x$ and so evaluate to 0 at that point. Hence $\deg_y f = d$, and further, for all $(e,a) \in \mathbb{K}^2$, $\deg f(e,a,y) = d$. «

Let $r = \operatorname{res}_y(f,g)$. If the projective curves defined by $f$ and $g$ have no common component then $r$ is nonzero. We conclude that if neither curve contains the vertical point at infinity then the two curves intersect on the line $ex = aw$ if and only if $r(e,a) = 0$ (of course the curves do not intersect at the vertical point at infinity).

We will shortly see that $r$ is homogeneous. Note that the map taking $(e:a)$ to the line $ex = aw$ is a bijection between $\mathbb{P}^2$ and the family of vertical lines (it is the map

$(e:a) \mapsto \iota(-a:e:0)$, see Sect. 4.6). Thus, the vertical lines on which the curves intersect correspond to the roots of the polynomial $r$, i.e. to the points in $V_{\mathbb{P}^1}(r)$. The multiplicity of a root of $r$ will correspond to the multiplicity of intersection of the curves on the corresponding vertical line.

### 6.1.1 The Resultant of Homogeneous Polynomials Is Homogeneous

Let $R$ be a unique factorisation domain, and let $(\boldsymbol{x}, y)$ be a tuple of variables.

**Proposition 6.2** *Let $f, g \in R[\boldsymbol{x}, y]$ be nonconstant and $\boldsymbol{x}, y$-homogeneous of degrees $d$ and $e$ respectively. Then $\operatorname{res}_y^{d,e}(f, g)$ is either 0 or $\boldsymbol{x}$-homogeneous of degree $de$.*

Here recall the definition of $\operatorname{res}_y^{d,e}(f, g)$ (Definition 3.11), which is defined for positive $d \geqslant \deg_y f$ and $e \geqslant \deg_y g$, not necessarily only $d = \deg_y f$ and $e = \deg_y g$. In this case we have $d = \deg_{\boldsymbol{x},y} f$ and not $\deg_y f$, and similarly $e = \deg_{\boldsymbol{x},y} g$. In fact we may even have $\deg_y f = 0$ or $\deg_y g = 0$, but $d, e > 0$ so this resultant is defined. The conclusion is that $\deg_{\boldsymbol{x}} \operatorname{res}_y^{d,e}(f, g) = de$ (or the resultant is 0).

***Proof*** Contrary to previous indexation, write $f = f_d + f_{d-1}y + \cdots f_0 y^d$ and $g = g_e + g_{e-1}y + \cdots g_0 y^e$, with $f_i, g_j \in R[\boldsymbol{x}]$. The point is that because $f$ is $\boldsymbol{x}, y$-homogeneous of degree $d$, for each $k \leqslant d$, $f_k$ is $\boldsymbol{x}$-homogeneous of degree $k$, and similarly for $g_l$.

Let $r = \operatorname{res}_y^{d,e}(f, g)$; so $r \in R[\boldsymbol{x}]$ is the determinant of the Sylvester matrix

$$M = M_y^{d,e}(f, g) = \begin{pmatrix} f_d & f_{d-1} & \cdots & f_0 & & & \\ & f_d & f_{d-1} & \cdots & f_0 & & \\ & & f_d & f_{d-1} & \cdots & f_0 & \\ & & & \ddots & & & \ddots \\ & & & & f_d & f_{d-1} & \cdots & f_0 \\ g_e & g_{e-1} & & \cdots & & g_0 & & \\ & \ddots & & & & & \ddots & \\ & & g_e & g_{e-1} & & \cdots & & g_0 \end{pmatrix}.$$

Let $t$ be a new variable. When we substitute $t\boldsymbol{x} = (tx_1, \ldots, tx_n)$ for $\boldsymbol{x}$ in every polynomial in the Sylvester matrix $M$, by Proposition 4.2 we get

$$M(t\boldsymbol{x}) = \begin{pmatrix} t^d f_d & t^{d-1} f_{d-1} & \cdots & & f_0 & & & \\ & t^d f_d & t^{d-1} f_{d-1} & \cdots & & f_0 & & \\ & & t^d f_d & t^{d-1} f_{d-1} & \cdots & & f_0 & \\ & & & \ddots & & & \ddots & \\ & & & & t^d f_d & t^{d-1} f_{d-1} & \cdots & f_0 \\ t^e g_e & t^{e-1} g_{e-1} & & \cdots & & g_0 & & \\ & \ddots & & & & & \ddots & \\ & & t^e g_e & t^{e-1} g_{e-1} & & \cdots & & g_0 \end{pmatrix},$$

and the point is that $r(t\boldsymbol{x}) = \det(M(t\boldsymbol{x}))$.

Let $N$ be the matrix

$$\begin{pmatrix} t^{d+e-1} f_d & t^{d+e-2} f_{d-1} & \cdots & & t^{e-1} f_0 & & & \\ & t^{d+e-2} f_d & t^{d+e-3} f_{d-1} & \cdots & & t^{e-2} f_0 & & \\ & & t^{d+e-3} f_d & t^{d+e-4} f_{d-1} & \cdots & & t^{e-3} f_0 & \\ & & & \ddots & & & \ddots & \\ & & & & t^d f_d & t^{d-1} f_{d-1} & \cdots & f_0 \\ t^{d+e-1} g_e & t^{d+e-2} g_{e-1} & & \cdots & & t^{d-1} g_0 & & \\ & \ddots & & & & & \ddots & \\ & & t^e g_e & t^{e-1} g_{e-1} & & \cdots & & g_0 \end{pmatrix}.$$

That is, for $k = 1, \ldots, d+e$, the $k$th column of $N$ is obtained from the $k$th column of $M$ by multiplying the column by $t^{d+e-k}$. Multiplying a column by a constant causes the determinant to be multiplied by the same constant. Hence

$$\det N = t^b \cdot \det M = t^b \cdot r,$$

where

$$b = (d+e-1) + (d+e-2) + \cdots + 0 = \binom{d+e}{2}.$$

On the other hand, we observe that for $k = 1, \ldots, e$, the $k$th row of $N$ is obtained from the $k$th row of $M(t\boldsymbol{x})$ by multiplying the row by $t^{e-k}$, and that for $k = 1, \ldots, d$, the $(e+k)$th row of $N$ is obtained from the $(e+k)$th row of $M(t\boldsymbol{x})$ by multiplying by $t^{d-k}$. This shows that

$$\det N = t^a \cdot \det(M(t\boldsymbol{x})) = t^a \cdot r(t\boldsymbol{x}),$$

where

$$a = (e-1) + (e-2) + \cdots + 0 + (d-1) + (d-2) + \cdots + 0 = \binom{d}{2} + \binom{e}{2}.$$

Putting it all together, we get $t^b r = t^a r(t\boldsymbol{x})$, so $r(t\boldsymbol{x}) = t^{b-a} \cdot r$. We calculate:

$$b-a = \binom{d+e}{2} - \binom{d}{2} - \binom{e}{2} = \frac{(d+e)(d+e-1)}{2} - \frac{d(d-1)}{2} - \frac{e(e-1)}{2} = de.$$

The proposition now follows from Proposition 4.2. □

*Exercise 6.3* Let $f, g \in R[\boldsymbol{x}, y]$ be $\boldsymbol{x}, y$-homogeneous and nonconstant, and let $d \geqslant \deg_y f$, $e \geqslant \deg_y f$ be positive. Show that $\operatorname{res}_y^{d,e}(f, g)$ is $\boldsymbol{x}$-homogeneous, although its degree is not necessarily $de$. Note that in contrast with Proposition 6.2 here we are not assuming that $\deg_{\boldsymbol{x},y} f = d$ and that $\deg_{\boldsymbol{x},y} g = e$, only that they are homogeneous of *some* nonzero degrees; $\deg_{\boldsymbol{x},y} f$ may be greater than $d$ or smaller than $d$, and similarly for $g$ and $e$. «

### 6.1.2 A Weak Version of Bézout's Theorem

We return to our investigation of the intersection of curves $f = 0$ and $g = 0$ in $\mathbb{P}^2$ with no common component, using the resultant $r = \operatorname{res}_y(f, g)$. We assume that the vertical point at infinity lies on neither curve. So $\deg f = \deg_y f$ and $\deg g = \deg_y g$ and so Proposition 6.2 says that $r$ is $w, x$-homogeneous of degree $\deg f \cdot \deg g$ (it is nonzero since the curves have no common component, so certainly have no common component which mentions $y$). We observed that $(e:a)$ is a root of $r$ if and only if the vertical line $ex = aw$ contains a point of the intersection of the two curves. We get a string of corollaries.

**Theorem 6.4** *Any two curves in $\mathbb{P}^2$ intersect.*

***Proof*** Let the curves be defined by polynomials $f$ and $g$. We may assume that they have no common component (or we're done). The sum of the curves (defined by $fg$) is a curve and so avoids some point $p$ (Proposition 4.7). After a change of coordinates $p$ is the vertical point at infinity.

Let $r = \mathrm{res}_y(f, g)$. Since $r$ is nonconstant it has a root $(e:a)$; so the curves intersect on the line $ex = aw$. □

Here is an application:

**Theorem 6.5** *Every nonsingular curve in* $\mathbb{P}^2$ *is irreducible.*

***Proof*** Suppose that a curve $C$ is reducible; let $D$ and $E$ be two irreducible components of $C$ (possibly equal). Take $p \in D \cap E$ (using Theorem 6.4). Then $p$ is a singular point of $C$ (see Corollary 5.23). □

Two curves without a common component intersect in only finitely many points.

**Proposition 6.6** *Let $C$ and $D$ be curves in $\mathbb{P}^2$ with no common component. Then $C \cap D$ contains at most* $\deg C \cdot \deg D$ *many distinct points.*

(Note that if both $C$ and $D$ contain repeated components, then $C \cap D$, which is the multiset intersection, can contain some points several times.)

***Proof*** Suppose, for a contradiction, that there are more than $\deg C \cdot \deg D$ many distinct points in the intersection of $C$ and $D$. Let $M$ be a set of $\deg C \cdot \deg D + 1$ many points in $C \cap D$. We claim that after a change of coordinates we can assume that the vertical point at infinity lies on neither $C$ nor $D$, and that no two points in $M$ lie on the same vertical line. It is enough to take a point $p$ which does not lie on $C$, on $D$ or on any line which passes through two points in $M$, and move that point to the vertical point at infinity. There is such a point because the curve consisting of the sum of $C$, $D$ and the finitely many lines which pass through two points of $M$ is not all of $\mathbb{P}^2$.

Having changed coordinates, choose $f$ defining $C$ and $g$ defining $D$, and let $r = \mathrm{res}_y(f, g)$. Then each point of $M$ corresponds to a distinct root of $r$ (note that every point other than the vertical point at infinity lies on a unique vertical line). This contradicts $\deg r = \deg f \cdot \deg g$. □

*Remark 6.7* Let $C$ and $D$ be curves in $\mathbb{P}^2$ with no common component. Then Proposition 6.6, together with its proof, show that we can change coordinates so that the vertical point at infinity lies on neither $C$ nor $D$, and each vertical line meets $C \cap D$ in at most one point. «

*Exercise 6.8* Let $C$ be a curve with no repeated components; let $d = \deg C$. Improving on Corollary 5.41, show that $C$ has at most $d(d-1)$ many singular points.[1] «

### A Naïve Definition of Intersection Multiplicity

Again let $f = 0$ and $g = 0$ be two curves with no common component, assume that the vertical point at infinity lies on neither, and let $r = \operatorname{res}_y(f, g)$. We will define intersection multiplicities so that in this situation, the multiplicity of a root $(e:a)$ of $r$ is the sum of the multiplicities of intersection of the curves on the points on the vertical line $ex = aw$. In particular, if each vertical line contains at most one point of intersection, then the multiplicity of intersection at a point equals the multiplicity of the corresponding root of $r$. The fact that the degree of $r$ is precisely $\deg f \cdot \deg g$ gives Bézout's theorem: the number of intersections, with multiplicities counted, is the maximal number possible given by Proposition 6.6.

We could take this as a *definition*: the multiplicity of intersection would be defined to be the multiplicity of the corresponding root of $r$. Remark 6.7 shows that we can always put curves in a position which allows such a definition. However there are many ways in which we can change coordinates to arrive at an acceptable position; it is not easy to show that the result does not depend on the choice of coordinate change.

We take a detour: we give another, more complicated definition of multiplicity of intersection. The advantage of this definition is that it can be made in every situation (no special conditions on the curves), and we can prove it is invariant under coordinate changes. Once this is done, we can show that when curves are positioned correctly, the new definition and the intended one agree.

## 6.2 The Homogeneous Resultant

We devise an analogue of the resultant that tells when homogeneous polynomials in two variables have a nonconstant common factor.

Let $R$ be a unique factorisation domain, and let $f$ and $g$ be nonconstant $x, y$-homogeneous polynomials in $R[x, y]$; so $d = \deg_{x,y} f$ and $e = \deg_{x,y} g$ are both positive. We write

$$f = a_0x^d + a_1x^{d-1}y + \cdots + a_{d-1}xy^{d-1} + a_dy^d$$

and

$$g = b_0x^e + b_1x^{e-1}y + \cdots + b_{e-1}xy^{e-1} + b_ey^e,$$

[1] For a better bound see Exercise 6.58.

and let

$$M_{x,y}(f,g) = \begin{pmatrix} a_0 & a_1 & \cdots & a_d & & & \\ & a_0 & a_1 & \cdots & a_d & & \\ & & a_0 & a_1 & \cdots & a_d & \\ & & & \ddots & & & \ddots \\ & & & & a_0 & a_1 & \cdots & a_d \\ b_0 & b_1 & & \cdots & & b_e & \\ & \ddots & & & & & \ddots \\ & & b_0 & b_1 & & \cdots & & b_e \end{pmatrix}$$

and $\operatorname{res}_{x,y}(f,g) = \det M_{x,y}(f,g)$. This is called the *homogeneous resultant* of $f$ and $g$. Immediately we get:

**Lemma 6.9** $\operatorname{res}_{x,y}(f,g) = \operatorname{res}_y^{d,e}(f^{\flat x}, g^{\flat x})$. □

It is perfectly possible that both $\deg_y f$ and $\deg_x f$ are smaller than $d$ (we allow $a_0 = a_d = 0$; for example let $f = xy$). For the homogeneous resultant we use $\deg_{x,y} f$. In some ways this results in a simpler theory. For example, if $\boldsymbol{z} = (z_1, \ldots, z_m)$ and $f \in R[\boldsymbol{z}, x, y]$ is $x, y$-homogeneous of degree $d$ then for all $\boldsymbol{a} \in R^m$, $f(\boldsymbol{a}, x, y)$ is either 0 or $x, y$-homogeneous of degree $d$. Thus:

**Lemma 6.10** *If $f, g \in R[\boldsymbol{z}, x, y]$ are $x, y$-homogeneous, $\boldsymbol{a} \in R^m$ and $f(\boldsymbol{a}, x, y)$ and $g(\boldsymbol{a}, x, y)$ are nonzero then* $\operatorname{res}_{x,y}(f(\boldsymbol{a}, x, y), g(\boldsymbol{a}, x, y)) = \big(\operatorname{res}_{x,y}(f,g)\big)(\boldsymbol{a})$. □

In contrast with substitutions into the usual resultant (Lemma 3.15), we do not need to worry about leading coefficients vanishing and the degree dropping.

**A Symmetry Between the Variables**

Starting with the homogeneous Sylvester matrix $M_{x,y}(f,g)$, exchange the first column with the last, the second with the second last, etc. We obtain the mirror

image

$$\begin{pmatrix} & & & & a_d & a_{d-1} & \cdots & a_0 \\ & & & \iddots & & & \iddots & \\ & & a_d & a_{d-1} & \cdots & a_0 & & \\ & a_d & a_{d-1} & \cdots & a_0 & & & \\ a_d & a_{d-1} & \cdots & a_0 & & & & \\ & & b_e & b_{e-1} & & \cdots & & b_0 \\ & \iddots & & & & & \iddots & \\ b_e & b_{e-1} & & \cdots & & b_0 & & \end{pmatrix}.$$

Now a permutation of the rows gives us

$$M_{y,x}(f,g) = \begin{pmatrix} a_d & a_{d-1} & \cdots & a_0 & & & & \\ & a_d & a_{d-1} & \cdots & a_0 & & & \\ & & a_d & a_{d-1} & \cdots & a_0 & & \\ & & & \ddots & & & \ddots & \\ & & & & a_d & a_{d-1} & \cdots & a_0 \\ b_e & b_{e-1} & & \cdots & & b_0 & & \\ & \ddots & & & & & \ddots & \\ & & b_e & b_{e-1} & & \cdots & & b_0 \end{pmatrix}$$

Thus $\operatorname{res}_{x,y}(f,g) = \pm \operatorname{res}_{y,x}(f,g)$. Applying Lemma 6.9 we get:

**Lemma 6.11** $\operatorname{res}_{x,y}(f,g) = \pm \operatorname{res}_x^{d,e}(f^{\flat y}, g^{\flat y})$. □

### 6.2.1 Main Property of the Homogeneous Resultant

The analogue of Theorem 3.12 is the following.

**Theorem 6.12** *Let $f$ and $g$ be nonconstant $x, y$-homogeneous polynomials in $R[x, y]$. Then $f$ and $g$ have a nonconstant common factor if and only if* $\operatorname{res}_{x,y}(f,g) = 0$.

***Proof*** Let $d = \deg_{x,y} f$ and $e = \deg_{x,y} g$; as above, write $f = \sum_i a_i x^{d-i} y^i$ and $g = \sum_i b_i x^{e-i} y^i$.

There are four cases, depending on the values of $a_d$ and of $b_e$.

**1:** $a_d \neq 0$ **and** $b_e \neq 0$ In this case, we see that $\deg_y f^{\flat x} = d$ and $\deg_y g^{\flat x} = e$ and so $\operatorname{res}_{x,y}(f, g) = \operatorname{res}_y(f^{\flat x}, g^{\flat x})$. Thus $\operatorname{res}_{x,y}(f, g) = 0$ if and only if $f^{\flat x}$ and $g^{\flat x}$ have a nonconstant common factor in $R[y]$. However, the assumption that $a_d \neq 0$ and $b_e \neq 0$ shows that $x$ divides neither $f$ nor $g$, and so $f = (f^{\flat x})^{\sharp x}$ and similarly for $g$ (see Proposition 4.20). The nonconstant common factors of $f$ and $g$ are the polynomials $h^{\sharp x}$, where $h$ is a nonconstant common factor of $f^{\flat x}$ and $g^{\flat x}$ (again see Proposition 4.20). In particular, $f$ and $g$ have a nonconstant common factor if and only if $f^{\flat x}$ and $g^{\flat x}$ have a nonconstant common factor.

**2:** $a_d = 0$ **and** $b_e = 0$ In this case, we see that $\operatorname{res}_{x,y}(f, g) = 0$, because the last column of $M_{x,y}(f, g)$ is a column of zeros. And we see that the nonconstant polynomial $x$ is a common factor of $f$ and $g$.

**3:** $a_d = 0$ **and** $b_e \neq 0$ Let $k$ be the number of times $x$ divides $f$. So

$$f = x^k \cdot \left(a_0 x^{d-k} + a_1 x^{d-k-1} y + \cdots + a_{d-k} y^{d-k}\right),$$

and $a_{d-k} \neq 0$. Let $h = f/x^k$, so $x$ does not divide $h$. The degree of $h$ is $d-k$. Since

$$M_{x,y}(f, g) = \begin{pmatrix}
a_0 & a_1 & \cdots & a_{d-k} & 0 & \cdots & 0 & & \\
 & a_0 & a_1 & \cdots & a_{d-k} & 0 & \cdots & 0 & \\
 & & \ddots & & & \ddots & \ddots & & \ddots \\
 & & & a_0 & a_1 & \cdots & a_{d-k} & 0 \cdots & 0 \\
b_0 & b_1 & b_2 & \cdots & b_e & & & & \\
 & b_0 & b_1 & b_2 & \cdots & b_e & & & \\
 & & b_0 & b_1 & b_2 & \cdots & b_e & & \\
 & & & \ddots & & & & \ddots & \\
 & & & & b_0 & b_1 & b_2 & \cdots & b_e
\end{pmatrix}.$$

Repeatedly developing the determinant along the last columns shows that

$$\operatorname{res}_{x,y}(f, g) = \pm (b_e)^k \cdot \operatorname{res}_{x,y}(h, g).$$

Now with $h$ and $g$ we are back in case (1)—because $x$ divides neither $h$ nor $g$—so we conclude that $\operatorname{res}_{x,y}(f, g) = 0$ if and only if $h$ and $g$ have a nonconstant common factor. The fact that $f = x^k \cdot h$ and that $x$ does not divide $g$ shows that $h$ and $g$ have a nonconstant common factor if and only if $f$ and $g$ have a nonconstant common factor (this of course uses unique factorisation in $R[x, y]$).

**4:** $a_d \neq 0$ **and** $b_e = 0$ This reduces to case (3), because $M_{x,y}(g, f)$ is obtained from $M_{x,y}(f, g)$ by a sequence of row-exchanges, and so $\text{res}_{x,y}(g, f) = \pm \text{res}_{x,y}(f, g)$. □

We obtain a homogeneous analogue of Proposition 3.16.

**Proposition 6.13** *Let $f, g \in R[z, x, y]$ be $x, y$-homogeneous and let $\boldsymbol{a} \in R^n$. The following are equivalent:*

- *Either $f(\boldsymbol{a}, x, y) = 0$ or $g(\boldsymbol{a}, x, y) = 0$; or $f(\boldsymbol{a}, x, y)$ and $g(\boldsymbol{a}, x, y)$ have a nonconstant common factor in $R[x, y]$;*
- $(\text{res}_{x,y}(f, g))(\boldsymbol{a}) = 0$.

***Proof*** If $f(\boldsymbol{a}, x, y) = 0$ then $(M_{x,y}(f, g))(\boldsymbol{a})$ contains a zero row and so $(\text{res}_{x,y}(f, g))(\boldsymbol{a}) = 0$. If $f(\boldsymbol{a}, x, y)$ and $g(\boldsymbol{a}, x, y)$ are both nonzero then the equivalence follows from Lemma 6.10 and Theorem 6.12. □

## 6.3 Multiplicity of Intersection and Bézout's Theorem

In this section we give a definition of intersection multiplicities. It generalises the idea of examining the intersection of curves $C$ and $D$ by examining the lines which intersect $C \cap D$. Indeed we examine all parameterisations of lines.

### 6.3.1 Coding Lines in $\mathbb{P}^2 \times \mathbb{P}^2$

A pair of distinct points $(q, r) \in \mathbb{P}^2 \times \mathbb{P}^2$ determines the line $\overline{qr}$. If $\mathcal{L}$ is a collection of lines then the corresponding subset of $\mathbb{P}^2 \times \mathbb{P}^2$ is $\{(q, r) : q \neq r \ \& \ \overline{qr} \in \mathcal{L}\}$.

The pairs in the *diagonal*

$$\Delta = \left\{(p, p) : p \in \mathbb{P}^2\right\} \subset \mathbb{P}^2 \times \mathbb{P}^2$$

do not determine lines. We have to add them though to get a hypersurface. For a point $p \in \mathbb{P}^2$, the linear family of lines which pass through $p$ is coded by

$$L_p = \Delta \cup \{(q, r) : q \neq r \ \& \ p \in \overline{qr}\}.$$

Since two points are always collinear, another description of $L_p$ is: the collection of pairs $(q, r)$ such that $\{p, q, r\}$ are collinear. If $p \neq s$ then it is easy to find a line which passes through $p$ and not through $s$, and so $L_p \neq L_s$.

**Lemma 6.14** *$L_p$ is an irreducible hypersurface of $\mathbb{P}^2 \times \mathbb{P}^2$ of bidegree* (1, 1).

***Proof*** Fix a presentation $\boldsymbol{p}$ of $p$. For presentations $\boldsymbol{q}$ and $\boldsymbol{r}$ of points $q$ and $r$, $\{p, q, r\}$ are collinear if and only if the dimension of the subspace $\langle \boldsymbol{p}, \boldsymbol{q}, \boldsymbol{r} \rangle$ is at most 2 if and only if the matrix $\begin{pmatrix} \boldsymbol{p} \\ \boldsymbol{q} \\ \boldsymbol{r} \end{pmatrix}$ is singular if and only if the determinant of that matrix is 0. Fix tuples of variables $\boldsymbol{u} = (u_0, u_1, u_2)$ and $\boldsymbol{v} = (v_0, v_1, v_2)$ used for defining hypersurfaces in $\mathbb{P}^2 \times \mathbb{P}^2$ (see Sect. 4.7). Let

$$f = \det \begin{pmatrix} \boldsymbol{p} \\ \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix}$$

Examining the closed form for the determinant (see page 46) we see that $f$ is the sum of products $\pm p_k u_i v_j$ and so $f$ is a $\boldsymbol{u}, \boldsymbol{v}$-bihomogeneous polynomial in $\mathbb{K}[\boldsymbol{u}, \boldsymbol{v}]$ of bidegree (1,1); and we have verified that $L_p = \left\lfloor V_{\mathbb{P}^2 \times \mathbb{P}^2}(f) \right\rfloor$. It remains to show that $f$ is irreducible.

By considering possible degrees, though, we see that if $f$ were reducible then it would be the product of two bihomogeneous polynomials of bidegrees (1, 0) and (0, 1). But then $L_p$ would be the union of $\ell \times \mathbb{P}^2$ and $\mathbb{P}^2 \times \ell'$ for lines $\ell$ and $\ell'$. However $L_p$ does not contain a subset of the form $\ell \times \mathbb{P}^2$: let $\ell$ be a line. Pick $q \in \ell$ distinct from $p$ and $r \notin \overline{p\,q}$. Then $(q, r) \in (\ell \times \mathbb{P}^2) \setminus L_p$. □

### 6.3.2 The Resultant of the General Intersection Polynomials

Let $f \in \mathbb{K}[w, x, y]$ be homogeneous of degree $d$. Recall the general intersection polynomial (Sect. 5.4, page 123)

$$f_{\boldsymbol{u},\boldsymbol{v}} = f(s\boldsymbol{u} + t\boldsymbol{v}) \in \mathbb{K}[\boldsymbol{u}, \boldsymbol{v}, s, t].$$

It is the sum of monomials of the form

$$s^k \cdot \boldsymbol{u}^k \cdot t^{d-k} \cdot \boldsymbol{v}^{d-k},$$

for some $k \leqslant d$, where of course by $\boldsymbol{u}^k$ we actually mean $\boldsymbol{u}^{\boldsymbol{k}}$ where $k_0 + k_1 + k_2 = k$, and similarly for $\boldsymbol{v}^{d-k}$. This shows that $f_{\boldsymbol{u},\boldsymbol{v}}$ is: $s, t$-homogeneous; $s, \boldsymbol{v}$-homogeneous; $t, \boldsymbol{u}$-homogeneous; and $\boldsymbol{u}, \boldsymbol{v}$-homogeneous, all of degree $d$.

For the rest of the section, fix $f, g \in \mathbb{K}[w, x, y]$ nonconstant and homogeneous of degrees $d$ and $e$. Since $f_{\boldsymbol{u},\boldsymbol{v}}$ and $g_{\boldsymbol{u},\boldsymbol{v}}$ are both $s, t$-homogeneous we can take their homogeneous resultant

$$R_{f,g} = \mathrm{res}_{s,t}\left(f_{\boldsymbol{u},\boldsymbol{v}}, g_{\boldsymbol{u},\boldsymbol{v}}\right).$$

The aim of this subsection is to show that $R_{f,g}$, which is an element of $\mathbb{K}[\boldsymbol{u}, \boldsymbol{v}]$, is $\boldsymbol{u}, \boldsymbol{v}$-bihomogeneous of bidegree $(de, de)$; and that the irreducible components of $V_{\mathbb{P}^2\times\mathbb{P}^2}\left(R_{f,g}\right)$ are various $L_p$ for $p \in V_{\mathbb{P}^2}(f) \cap V_{\mathbb{P}^2}(g)$. We will then define the multiplicity of intersection at a point $p$ to be the number of copies of $L_p$ in this hypersurface. Our presentation follows [vdW45],[SK59, Ch.II].

### Bihomogeneity of $R_{f,g}$

**Lemma 6.15** *Either $R_{f,g} = 0$ or $R_{f,g}$ is $\boldsymbol{u}, \boldsymbol{v}$-bihomogeneous of bidegree $(de, de)$.*

***Proof*** Suppose that $R_{f,g} \neq 0$. We see that ${f_{\boldsymbol{u},\boldsymbol{v}}}^{\flat s}$ is $t, \boldsymbol{u}$-homogeneous of degree $d$, and similarly ${g_{\boldsymbol{u},\boldsymbol{v}}}^{\flat s}$ is $t, \boldsymbol{u}$-homogeneous of degree $e$. Because $R_{f,g} = \mathrm{res}_t^{d,e}\left({f_{\boldsymbol{u},\boldsymbol{v}}}^{\flat s}, {g_{\boldsymbol{u},\boldsymbol{v}}}^{\flat s}\right)$ (Lemma 6.9), $R_{f,g}$ is $\boldsymbol{u}$-homogeneous of degree $de$ (Proposition 6.2).

The intersection polynomials are $s, \boldsymbol{v}$-homogeneous; a similar argument using dehomogenisation with respect to $t$ and using Lemma 6.11, shows that $R_{f,g}$ is also $\boldsymbol{v}$-homogeneous of degree $de$. □

### The Structure of the Hypersurface Defined by $R_{f,g}$

Let $\boldsymbol{q}$ and $\boldsymbol{r}$ be presentations of points $q$ and $r$ in $\mathbb{P}^2$. Let $f_{\boldsymbol{q},\boldsymbol{r}} = f_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{q}, \boldsymbol{r}, s, t) = f(s\boldsymbol{q} + t\boldsymbol{r})$; when $q \neq r$, this is the intersection polynomial $f_{\boldsymbol{\psi}}$, where $\boldsymbol{\psi} = \boldsymbol{\psi}_{\boldsymbol{q},\boldsymbol{r}}(s, t) = s\boldsymbol{q} + t\boldsymbol{r}$ is the presentation of the parameterisation $\psi_{\boldsymbol{p},\boldsymbol{q}}$ of $\overline{pq}$ (Example 4.17). In that case, $f_{\boldsymbol{q},\boldsymbol{r}} = 0$ if and only if the line $\overline{qr}$ is a component of $V_{\mathbb{P}^2}(f)$. In any case ($q = r$ or not), if both $f_{\boldsymbol{q},\boldsymbol{r}}$ and $g_{\boldsymbol{q},\boldsymbol{r}}$ are nonzero, then as they are homogeneous polynomials in two variables, they have a nonconstant common component if and only if they have a common root in $\mathbb{P}^1$.

**Lemma 6.16**

$$\left\lfloor V_{\mathbb{P}^2\times\mathbb{P}^2}(R_{f,g})\right\rfloor = \bigcup\left\{L_p \,:\, p \in V_{\mathbb{P}^2}(f) \cap V_{\mathbb{P}^2}(g)\right\}.$$

***Proof*** We show: (1) $R_{f,g}(\boldsymbol{q}, \boldsymbol{q}) = 0$ for all $\boldsymbol{q}$; and (2) if $q \neq r$ then $R_{f,g}(\boldsymbol{q}, \boldsymbol{r}) = 0$ if and only if the line $\overline{qr}$ intersects $V_{\mathbb{P}^2}(f) \cap V_{\mathbb{P}^2}(g)$. We use Proposition 6.13.

For (1), we note that $s + t$ is a nonconstant factor of both $f_{\boldsymbol{q},\boldsymbol{q}}$ and $g_{\boldsymbol{q},\boldsymbol{q}}$. For (2) there are two cases. If (say) $f_{\boldsymbol{q},\boldsymbol{r}} = 0$ then $\overline{qr}$ is a component of the curve $f = 0$; the curve $g = 0$ intersects that line and so the line intersects $V_{\mathbb{P}^2}(f) \cap V_{\mathbb{P}^2}(g)$; and we know that $R_{f,g}(\boldsymbol{q}, \boldsymbol{r}) = 0$. In the other case suppose that both $f_{\boldsymbol{q},\boldsymbol{r}}$ and $g_{\boldsymbol{q},\boldsymbol{r}}$ are nonzero. We know that $p \in \overline{qr}$ is on both curves $f = 0$ and $g = 0$ if and only if $\psi_{\boldsymbol{q},\boldsymbol{r}}^{-1}(p)$ is a root of both $f_{\boldsymbol{q},\boldsymbol{r}}$ and $g_{\boldsymbol{q},\boldsymbol{r}}$. Thus the existence of a common root, which is equivalent to $R_{f,g}(\boldsymbol{q}, \boldsymbol{r}) = 0$, is also equivalent to the condition $\overline{qr} \cap V_{\mathbb{P}^2}(f) \cap V_{\mathbb{P}^2}(g) \neq 0$.

It follows that if $p \in V_{\mathbb{P}^2}(f) \cap V_{\mathbb{P}^2}(g)$ then $L_p \subseteq V_{\mathbb{P}^2\times\mathbb{P}^2}(R_{f,g})$. On the other hand let $(q, r) \in V_{\mathbb{P}^2\times\mathbb{P}^2}(R_{f,g})$; we show that $(q, r) \in L_p$ for some $p \in V_{\mathbb{P}^2}(f) \cap V_{\mathbb{P}^2}(g)$. If $q \neq r$ then of course $(q, r) \in L_p$ for any $p \in \overline{qr} \cap V_{\mathbb{P}^2}(f) \cap V_{\mathbb{P}^2}(g)$. If

$q = r$ then $(q, r) \in L_p$ for all $p$; the union on the right hand side is nonempty since $V_{\mathbb{P}^2}(f) \cap V_{\mathbb{P}^2}(g)$ is nonempty (Theorem 6.4).[2] □

If $f$ and $g$ have a common factor then so do $f_{\boldsymbol{u},\boldsymbol{v}}$ and $g_{\boldsymbol{u},\boldsymbol{v}}$, in which case $R_{f,g} = 0$. If $f$ and $g$ do not have a common factor then $V_{\mathbb{P}^2}(f) \cap V_{\mathbb{P}^2}(g)$ is finite (Proposition 6.6). For any finite set of points in $\mathbb{P}^2$ we can find a line which does not pass through any of these points (see Exercise 4.43); it follows that $V_{\mathbb{P}^2\times\mathbb{P}^2}(R_{f,g})$ does not contain every pair in $\mathbb{P}^2 \times \mathbb{P}^2$, whence $R_{f,g} \neq 0$. In this case we see that the irreducible components of $V_{\mathbb{P}^2\times\mathbb{P}^2}(R_{f,g})$ are precisely the hypersurfaces $L_p$ for $p \in V_{\mathbb{P}^2}(f) \cap V_{\mathbb{P}^2}(g)$; this follows from Proposition 4.51. To summarise:

**Proposition 6.17** *$R_{f,g} = 0$ if and only if $f$ and $g$ have a common component. If $R_{f,g} \neq 0$ then the irreducible components of $V_{\mathbb{P}^2\times\mathbb{P}^2}(R_{f,g})$ are $L_p$ for $p \in V_{\mathbb{P}^2}(f) \cap V_{\mathbb{P}^2}(g)$.* □

### 6.3.3 Intersection Multiplicity and Bézout's Theorem

**Definition 6.18** Let $C$ and $D$ be curves in $\mathbb{P}^2$ with no common component. Let $f$ define $C$ and $g$ define $D$. For $p \in \mathbb{P}^2$ let $i_p(C, D)$ be the number of times the hypersurface $L_p$ appears in $V_{\mathbb{P}^2\times\mathbb{P}^2}(R_{f,g})$.

For the definition to be completely rigorous we need to show that it does not depend on the choice of polynomials $f$ and $g$ defining $C$ and $D$. But suppose that $f' = \lambda f$ and $g' = \mu g$ for nonzero $\lambda, \mu \in \mathbb{K}$. Then $f'_{\boldsymbol{u},\boldsymbol{v}} = \lambda f_{\boldsymbol{u},\boldsymbol{v}}$ and $g'_{\boldsymbol{u},\boldsymbol{v}} = \mu g_{\boldsymbol{u},\boldsymbol{v}}$ and this shows that $M_{s,t}(f'_{\boldsymbol{u},\boldsymbol{v}}, g'_{\boldsymbol{u},\boldsymbol{v}})$ is obtained from $M_{s,t}(f_{\boldsymbol{u},\boldsymbol{v}}, g_{\boldsymbol{u},\boldsymbol{v}})$ by multiplying the first $e$ rows by $\lambda$ and the last $d$ rows by $\mu$. So $R_{f',g'} = \lambda^e \mu^d R_{f,g}$ (as usual $d = \deg f$ and $e = \deg g$), whence $V_{\mathbb{P}^2\times\mathbb{P}^2}(R_{f',g'}) = V_{\mathbb{P}^2\times\mathbb{P}^2}(R_{f,g})$.

Proposition 6.17 implies:

**Proposition 6.19** *Let $C$ and $D$ be curves in $\mathbb{P}^2$ with no common component, and let $p \in \mathbb{P}^2$. Then $i_p(C, D) > 0$ if and only if $p \in C \cap D$.* □

The fact that the bidegree of each $L_p$ is (1,1) and that the bidegree of $R_{f,g}$ is $(de, de)$ shows that $R_{f,g}$ has precisely $de$ many irreducible components. Since $L_p \neq L_q$ if $p \neq q$ no component is counted twice. This gives:

[2] Alternatively, since the diagonal $\Delta$ is a proper subset of an irreducible hypersurface it is not a hypersurface of $\mathbb{P}^2 \times \mathbb{P}^2$; hence $V_{\mathbb{P}^2\times\mathbb{P}^2}(R_{f,g})$ must be a proper superset of the diagonal. This gives another proof that the curves $f = 0$ and $g = 0$ intersect.

**Bézout's Theorem** *If $C$ and $D$ are curves in $\mathbb{P}^2$ with no common component then*

$$\sum_{p \in \mathbb{P}^2} i_p(C, D) = \deg C \cdot \deg D. \qquad \square$$

We again emphasise that the theorem holds only when $\mathbb{K}$ is algebraically closed and when the curves are projective rather than affine. Bézout's theorem shows that the bound given by Proposition 6.6 is always achieved if intersections are "counted properly".

### The Intersection Multiset

**Definition 6.20** Let $C$ and $D$ be curves in $\mathbb{P}^2$ with no common component. We let $C \cdot D$ be the multiset of points defined by $m_p(C \cdot D) = i_p(C, D)$. That is, $C \cdot D$ is the multiset with underlying set $C \cap D$, in which the multiplicity of each $p \in C \cap D$ is $i_p(C, D)$.

So Bézout's theorem states that $|C \cdot D| = \deg C \cdot \deg D$.

## 6.3.4 Geometric Invariance

We show that multiplicity of intersection is invariant under changes of coordinates. Let $C$ and $D$ be curves in $\mathbb{P}^2$ which have no common component. If $\alpha$ is a change of coordinates of $\mathbb{P}^2$, then $\alpha[C]$ and $\alpha[D]$ are curves which have no common component. We prove:

**Proposition 6.21** *Let $C$ and $D$ be curves in $\mathbb{P}^2$ which have no common component. Let $\alpha$ be a change of coordinates of $\mathbb{P}^2$. Then for all $p$,*

$$i_p(C, D) = i_{\alpha(p)}(\alpha[C], \alpha[D]) .$$

Fix a change of coordinates $\alpha$ of $\mathbb{P}^2$, and let $\boldsymbol{\alpha}$ be a linear presentation of $\alpha$. As in Sect. 5.4 we use the extension of $\alpha$ to changes of coordinates of both $\mathbb{P}^2 \times \mathbb{P}^2$ and $\mathbb{P}^2 \times \mathbb{P}^2 \times \mathbb{P}^1$. The definition of intersection multiplicity shows that Proposition 6.21 follows from the following two lemmas:

**Lemma 6.22** *For all $p \in \mathbb{P}^2$, $(\alpha \times \alpha)[L_p] = L_{\alpha(p)}$.*

**Lemma 6.23** *Let $f, g \in \mathbb{K}[w, x, y]$ be nonconstant and homogeneous. Then $(\boldsymbol{\alpha} \times \boldsymbol{\alpha})^*(R_{f,g}) = R_{\boldsymbol{\alpha}^*(f), \boldsymbol{\alpha}^*(g)}$.*

***Proof of Lemma 6.22*** Since $\alpha$ maps lines to lines, $\{q, r, p\}$ are collinear if and only if $\{\alpha(q), \alpha(r), \alpha(p)\}$ are collinear. $\square$

For the proof of Lemma 6.23 we use the notation from Sect. 5.4: for $g \in \mathbb{K}[\boldsymbol{u}, \boldsymbol{v}, s, t]$ let $\hat{g} = (\boldsymbol{\alpha} \times \boldsymbol{\alpha} \times \mathrm{id}_{\mathbb{K}^2})^*(g)$. Note that if $g \in \mathbb{K}[\boldsymbol{u}, \boldsymbol{v}]$ then $\hat{g} = (\boldsymbol{\alpha} \times \boldsymbol{\alpha})^*(g)$.

***Proof of Lemma 6.23*** Lemma 5.30 says that $(\boldsymbol{\alpha}^*(f))_{\boldsymbol{u},\boldsymbol{v}} = \hat{f_{\boldsymbol{u},\boldsymbol{v}}}$ for all $f \in \mathbb{K}[w, x, y]$. So it remains to show that for $s, t$-homogeneous $h, k \in \mathbb{K}[\boldsymbol{u}, \boldsymbol{v}, s, t]$,

$$\widehat{\mathrm{res}_{s,t}(h, k)} = \mathrm{res}_{s,t}(\hat{h}, \hat{k}) \tag{6.1}$$

Since $g \mapsto \hat{g}$ is a ring homomorphism and $\hat{s} = s$, $\hat{t} = t$, if $h = \sum h_i s^{d-i} t^i$ then $\hat{h} = \sum \hat{h}_i s^{d-i} t^i$. Thus $M_{s,t}(\hat{h}, \hat{k})$ is obtained from $M_{s,t}(h, k)$ by applying the map $g \mapsto \hat{g}$ to every entry. The determinant is a polynomial in the entries, so the fact that $g \mapsto \hat{g}$ is a ring homomorphism gives Eq. (6.1). □

## 6.4 Coincidence with Earlier Definitions

We show that the definition we gave for multiplicity of intersection agrees with the attempted definition from Sect. 6.1. We also show that it agrees with the definition we already gave for the intersection of a curve and a line (Definition 5.25).

### 6.4.1 Using the Family of Vertical Lines

Let $R$ be a unique factorisation domain. Let $f = f(y, z) \in R[y, z]$. By $f(z, y)$ we of course mean the polynomial obtained by exchanging $y$ and $z$: if $f(y, z) = y^2 z$ then $f(z, y) = z^2 y$.

*Remark 6.24* Let $f, g \in R[y, z] \setminus R[y]$, i.e. $\deg_z f, \deg_z g > 0$, so $\mathrm{res}_z(f, g)$ is defined. Exchanging variables, $\mathrm{res}_y(f(z, y), g(z, y))$ is defined. As expected, renaming variables does not really matter:

$$\mathrm{res}_y\,(f(z, y), g(z, y)) = (\mathrm{res}_z(f(y, z), g(y, z)))\,(z).$$

For the entries of $M_z(f(y, z), g(y, z))$ are simply the entries of $M_y(f(z, y), g(z, y))$ but with $y$ instead of $z$. «

**Proposition 6.25** *Let $C$ and $D$ be curves in $\mathbb{P}^2$ with no common component. Suppose that the vertical point at infinity $(0:0:1)$ lies on neither $C$ nor $D$. Let $f$ define $C$ and $g$ define $D$. Then for all $(e:a) \in \mathbb{P}^1$, the multiplicity of the root $(e:a)$ of $\mathrm{res}_y(f, g)$ is the sum of $i_p(C, D)$ for the points $p$ on the line $ex = aw$.*

In particular, if there is a unique point $p \in C \cap D$ on that line, then $i_p(C, D)$ equals the multiplicity $m_{(e:a)}\left(V_{\mathbb{P}^1}(\mathrm{res}_y(f, g))\right)$.

***Proof*** Let

$$r = R_{f,g}(w, x, y, 0, 0, 1);$$

that is, in $R_{f,g}(\boldsymbol{u}, \boldsymbol{v})$ we substitute a presentation of the vertical point at infinity for $\boldsymbol{v}$, and substitute the variables $(w, x, y)$ for $\boldsymbol{u}$. For a presentation $\boldsymbol{p} = (e, a, b)$ of a point $p = (e:a:b)$ let $h_{\boldsymbol{p}} = \det\begin{pmatrix}\boldsymbol{p}\\ \boldsymbol{u}\\ \boldsymbol{v}\end{pmatrix}$ be a polynomial which defines the hypersurface $L_p$ (see the proof of Lemma 6.14). When $p \neq (0:0:1)$ (i.e. when $(e, a) \neq (0, 0)$),

$$h_{\boldsymbol{p}}(w, x, y, 0, 0, 1) = \det\begin{pmatrix} e & a & b \\ w & x & y \\ 0 & 0 & 1 \end{pmatrix} = ex - aw$$

defines the point $(e:a)$ in $\mathbb{P}^1$. Since $R_{f,g}$ is the product of $h_{\boldsymbol{p}}^{i_p(C,D)}$ for $p \in C \cap D$, $r$ is the product of the resulting $ex - aw$ and we conclude that the multiplicity of $(e:a)$ in $V_{\mathbb{P}^1}(r)$ is the sum of $i_p(C, D)$ for $p \in C \cap D$ of the form $(e:a:b)$ for some $b \in \mathbb{K}$; since the vertical point at infinity is not in $C \cap D$, these are precisely the points in $C \cap D$ on the line $ex = aw$. The proposition is proved once we show:

$$r = \mathrm{res}_y(f, g). \tag{6.2}$$

Let $d = \deg f$ and $e = \deg g$. The polynomial

$$f_{\boldsymbol{u},\boldsymbol{v}}(w, x, y, 0, 0, 1) = f(sw, sx, sy + t)$$

is nonzero (set $s = 0, t = 1$ to get $f(0, 0, 1)$ which is nonzero, as the vertical point at infinity does not lie on $C$). The same holds for $g$ and so

$$r = \mathrm{res}_{s,t}\left(f(sw, sx, sy + t), g(sw, sx, sy + t)\right)$$

(Lemma 6.10 for $\boldsymbol{a} = (w, x, y)$, applying the lemma over the domain $R = \mathbb{K}[w, x, y]$). Dehomogenising with respect to $s$, by Lemma 6.9,

$$r = \mathrm{res}_t^{d,e}\left(f(w, x, y + t), g(w, x, y + t)\right).$$

The assumption on $C$ and $D$ implies that the monomial $y^d$ appears in $f$ and $y^e$ appears in $g$ (Remark 6.1) and so in fact $r = \mathrm{res}_t\left(f(w, x, y + t), g(w, x, y + t)\right)$. Exchanging the variables $y$ and $t$ in $f(w, x, y + t)$ does not change the polynomial and so (Remark 6.24) $r = \mathrm{res}_y\left(f(w, x, t + y), g(w, x, t + y)\right)(y)$ where on the

right $y$ is substituted for $t$. As observed above, $r \in \mathbb{K}[w, x]$ (it is the product of polynomials $ex - aw$), that is, neither $y$ nor $t$ appear in $r$; this implies that

$$r = \mathrm{res}_y\left(f(w, x, t + y), g(w, x, t + y)\right).$$

Now substitute $t = 0$ in both sides. On the left we get $r$. Using the fact that $y^d$ appears in $f$ and $y^e$ in $g$, on the right we get $\mathrm{res}_y(f, g)$ (Lemma 3.15), which establishes Eq. (6.2). □

*Remark 6.26* For calculations is it neater to pass to affine coordinates. Suppose that $C$ and $D$ have no common component and that neither contains the vertical point at infinity. Then $\mathrm{res}_y(f, g)^\flat = \mathrm{res}_y(f^\flat, g^\flat)$ (Lemma 3.15 and Remark 6.1). Thus, if $C \cap D$ intersects the affine line $x = a$ at exactly one point $p$ then $i_p(C, D)$ is the multiplicity of the root $a$ of the polynomial $\mathrm{res}_y\left(f^\flat, g^\flat\right)$. «

### 6.4.2 Intersecting Lines

**Proposition 6.27** *Let $C$ be a curve in $\mathbb{P}^2$, let $\ell$ be a line and let $p \in \ell$. Definitions 5.25 and 6.18 for $i_p(C, \ell)$ agree.*

***Proof*** If $p \notin C \cap \ell$ then we know that both definitions give 0. So we assume that $p \in C \cap \ell$. Both definitions are invariant under coordinate changes (Propositions 5.29 and 6.21) so we change coordinates so that: (1) $p$ is the origin; (2) $\ell$ is the $x$-axis; and (3) the vertical point at infinity doesn't lie on $C$ (pick points $q \in \ell \setminus \{p\}$ and $r \notin \ell \cup C$ and move $p$ to the origin, $q$ to the horizontal point at infinity and $r$ to the vertical point at infinity; we use the Three Point Lemma, since $\{p, q, r\}$ are not collinear).

Only $p$ lies on both the $y$-axis and on $\ell \cap C$. So by Proposition 6.25 and Remark 6.26, $i_p(C, \ell)$ according to Definition 6.18 is the multiplicity of the root 0 of the polynomial $\mathrm{res}_y(f, y)$, where $f$ is the dehomogenisation of a polynomial defining $C$. On the other hand the multiplicity according to Definition 5.25 is the multiplicity of the root 0 of the polynomial $f(t, 0)$ (Example 5.33).

So we calculate. Write $f = a_0(x) + a_1(x)y + \cdots + a_d(x)y^d$. Then $f(t, 0) = a_0(t)$ and

$$\mathrm{res}_y(f, y) = \begin{vmatrix} a_0 & a_1 & a_2 & \cdots & a_d \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \end{vmatrix} = a_0(x)$$

which gives the desired equality. □

## 6.5 Categoricity of Multiplicity of Intersection

In practice, most calculations of multiplicities of intersection don't use either definition given in this chapter. Rather, we identify properties of multiplicity of intersection which help reduce complicated calculations to simpler ones. We then show that these properties actually determine the multiplicity of intersection function.

### 6.5.1 Symmetry

**Proposition 6.28** *Let $C$ and $D$ be curves with no common component. For all $p \in \mathbb{P}^2$, $i_p(C, D) = i_p(D, C)$.*

***Proof*** As observed above, $R_{f,g} = \pm R_{g,f}$ (by permuting the rows) and so the proposition follows directly from Definition 6.18. □

### 6.5.2 Products

We work toward showing:

**Proposition 6.29** *Let $C$, $D$ and $E$ be curves in $\mathbb{P}^2$; suppose that no component of $C$ is a component of $D$ or of $E$. Then $C \cdot (D + E) = C \cdot D + C \cdot E$, i.e., for all $p \in \mathbb{P}^2$,*

$$i_p(C, D + E) = i_p(C, D) + i_p(C, E).$$

As a result we see that multiplicity of intersection is determined by its values for irreducible curves.

Let $R$ be a unique factorisation domain and let $\boldsymbol{y} = (y_1, \ldots, y_d)$, $\boldsymbol{z} = (z_1, \ldots, z_e)$. Let

$$f(x, \boldsymbol{y}) = (x - y_1)(x - y_2) \cdots (x - y_d)$$

and

$$g(x, \boldsymbol{z}) = (x - z_1)(x - z_2) \cdots (x - z_e).$$

Note that both $f$ and $g$ are $x, \boldsymbol{y}, \boldsymbol{z}$-homogeneous, and $\deg_x f = d$, $\deg_x g = e$.

**Lemma 6.30**

$$\operatorname{res}_x(g, f) = \prod_{i \leqslant d, j \leqslant e} (y_i - z_j).$$

***Proof*** Let $r = \operatorname{res}_x(g, f)$. By Proposition 6.2, $r$ is $\boldsymbol{y}, \boldsymbol{z}$-homogeneous, of degree $de$.

Let $D = R[y_2, y_3, \ldots, y_d, \boldsymbol{z}]$, so $R[x, \boldsymbol{y}, \boldsymbol{z}] = D[x, y_1]$. We think of both $f$ and $g$ as elements of $D[x, y_1]$, so we write $f(x, y_1)$ and $g(x, y_1)$ even though $y_1$ doesn't appear in $g$.

Let $j \leqslant e$. We substitute $z_j$ for $y_1$ in $f$ and $g$:

$$f(x, z_j) = (x - z_j)(x - y_2)(x - y_3) \ldots (x - y_d)$$

and $g(x, z_j) = g$. Both are polynomials in $D[x]$.

Certainly $\deg_x f(x, z_j) = d$ so by Lemma 3.15

$$r(z_j) = \operatorname{res}_x(g(x, z_j), f(x, z_j)) = \operatorname{res}_x(g, f(x, z_j))$$

(note that $r(z_j) \in D$). Now $x - z_j$ is nonconstant as an element of $D[x]$, and is a common factor of $f(x, z_j)$ and $g$, so $r(z_j) = 0$ (Theorem 3.12). So $y_1 - z_j$ divides $r$ in $D[y_1]$ (Theorem 2.16).

We picked $y_1$ for notational simplicity; the argument holds for any $y_i$, and so for all $i \leqslant d$ and $j \leqslant e$ we have $(y_i - z_j) \mid r$. Since the polynomials $y_i - z_j$ are irreducible and pairwise non-associate, and the number of them is $de = \deg r$, we see that

$$r = \alpha \prod_{i \leq d, j \leq e} (y_i - z_j),$$

for some $\alpha \in R$. We show that $\alpha = 1$.

$f(x, \mathbf{0}) = x^d$, so $\deg_x f(x, \mathbf{0}) = d = \deg_x f$. Substituting $\mathbf{0}$ for $\boldsymbol{y}$ in $g$ yields $g$, so as above $r(\mathbf{0}, \boldsymbol{z}) = \operatorname{res}_x(g, f(x, \mathbf{0}))$. But this determinant we can calculate. Write

$g = a_0 + a_1 x + \cdots + a_e x^e$. We have $a_0 = (-1)^e z_1 z_2 \cdots z_e$, and

$$\operatorname{res}_x(g, f(x, \mathbf{0})) = \begin{vmatrix} a_0 & a_1 & \cdots & a_e & & & \\ & a_0 & a_1 & \cdots & a_e & & \\ & & a_0 & a_1 & \cdots & a_e & \\ & & & \ddots & & & \ddots \\ & & & & a_0 & a_1 & \cdots & a_e \\ 0 & 0 & & \cdots & & 1 & & \\ & \ddots & & & & & \ddots & \\ & & 0 & 0 & & \cdots & & 1 \end{vmatrix} = a_0^d.$$

So

$$r(\mathbf{0}, \mathbf{z}) = (-1)^{de} z_1^d z_2^d \cdots z_e^d.$$

On the other hand

$$\prod_{i \leqslant d, j \leqslant e} (0 - z_j) = (-1)^{de} z_1^d z_2^d \cdots z_e^d,$$

so $\alpha = 1$. □

By substituting elements of $R$ into the polynomials above (Lemma 3.15), we get:

**Corollary 6.31** *Let $R$ be an integral domain; let $b_1, \ldots, b_d$, $c_1, \ldots, c_e \in R$. Let $f = \prod_{i \leqslant d}(x - b_i)$ and $g = \prod_{j \leqslant e}(x - c_j)$ (in $R[x]$). Then*

$$\operatorname{res}_x(g, f) = \prod_{i \leqslant d, j \leqslant e} (b_i - c_j).$$ □

Let $f, g \in R[x]$ have degrees $d$ and $e$, and let $\alpha, \beta \in R$ be nonzero. The Sylvester matrix $M_x(\alpha f, \beta g)$ is the result of multiplying the first $e$ rows of $M_x(f, g)$ by $\alpha$, and the last $d$ rows by $\beta$, and so

$$\operatorname{res}_x(\alpha f, \beta g) = \alpha^e \beta^d \operatorname{res}_x(f, g). \tag{6.3}$$

This gives:

**Lemma 6.32** *Let $R$ be an integral domain; let $f, g, h \in R[x]$, and assume that all three are products of linear polynomials in $R[x]$. Then*

$$\operatorname{res}_x(f, gh) = \operatorname{res}_x(f, g) \cdot \operatorname{res}_x(f, h).$$

***Proof*** Write $f = \alpha \prod_{i \leqslant d}(x - a_i)$; $g = \beta \prod_{j \leqslant e}(x - b_j)$; $h = \gamma \prod_{l \leqslant m}(x - c_l)$. Then by Corollary 6.31 and Eq. (6.3),

$$\operatorname{res}_x(f, g) = \alpha^e \beta^d \prod_{i \leqslant d, j \leqslant e} (b_j - a_i);$$

$$\operatorname{res}_x(f, h) = \alpha^m \gamma^d \prod_{i \leqslant d, l \leqslant m} (c_l - a_i); \text{ and}$$

$$\operatorname{res}_x(f, gh) = \alpha^{e+m} (\beta\gamma)^d \prod_{i \leqslant d, j \leqslant e, l \leqslant m} (b_j - a_i)(c_l - a_i),$$

which gives the required equality. □

Recall that we are assuming that $\mathbb{K}$ is algebraically closed; so Lemma 6.32 implies:

**Corollary 6.33** *For all nonconstant $f, g, h \in \mathbb{K}[x]$,*

$$\operatorname{res}_x(f, gh) = \operatorname{res}_x(f, g) \cdot \operatorname{res}_x(f, h).$$

Now we want to extend Corollary 6.33 by replacing $\mathbb{K}$ by other integral domains. For the "proper" way to do it, see Exercise 6.45; this uses tools that we have not developed. We will require one particular such domain, namely polynomial rings $\mathbb{K}[\boldsymbol{y}]$:

**Proposition 6.34** *Let $\boldsymbol{y} = y_1, \ldots, y_n$. For all $f, g, h \in \mathbb{K}[\boldsymbol{y}, x] \setminus \mathbb{K}[\boldsymbol{y}]$,*

$$\operatorname{res}_x(f, gh) = \operatorname{res}_x(f, g) \cdot \operatorname{res}_x(f, h).$$

***Proof*** Let $d = \deg_x f$, $e = \deg_x g$ and $m = \deg_x h$. Write $f = \sum f_i x^i$, $g = \sum g_i x^i$, $h = \sum h_i x^i$ with $f_i, g_i, h_i \in \mathbb{K}[\boldsymbol{y}]$. Let $\boldsymbol{a} \in \mathbb{K}^n$, and suppose that $f_d g_e h_m(\boldsymbol{a}) \neq 0$; so $\deg_x f(\boldsymbol{a}, x) = d$, $\deg_x g(\boldsymbol{a}, x) = e$ and $\deg_x h(\boldsymbol{a}, x) = m$. In that case, $\operatorname{res}_x(f(\boldsymbol{a}, x), g(\boldsymbol{a}, x)) = \operatorname{res}_x(f, g)(\boldsymbol{a})$, and similarly for the pairs $(f, h)$ and $(f, gh)$. By Corollary 6.33,

$$\operatorname{res}_x(f(\boldsymbol{a}, x), g(\boldsymbol{a}, x)h(\boldsymbol{a}, x)) = \operatorname{res}_x(f(\boldsymbol{a}, x), g(\boldsymbol{a}, x)) \cdot \operatorname{res}_x(f(\boldsymbol{a}, x), h(\boldsymbol{a}, x)),$$

and so

$$\operatorname{res}_x(f, gh)(\boldsymbol{a}) = \operatorname{res}_x(f, g)(\boldsymbol{a}) \cdot \operatorname{res}_x(f, h)(\boldsymbol{a}).$$

This holds for all $\boldsymbol{a} \in \mathbb{K}^n \setminus V_{\mathbb{A}^n}(f_d g_e h_m)$. Let $k = \operatorname{res}_x(f, gh) - \operatorname{res}_x(f, g) \operatorname{res}_x(f, h)$. So $k f_d g_e h_m(\boldsymbol{a}) = 0$ for all $\boldsymbol{a} \in \mathbb{K}^n$. Hence $k f_d g_e h_m = 0$ (Proposition 2.18). Since $f_d g_e h_m \neq 0$ we get $k = 0$, which completes the proof. □

***Proof of Proposition 6.29*** We have $(gh)_{\boldsymbol{u},\boldsymbol{v}} = g_{\boldsymbol{u},\boldsymbol{v}} h_{\boldsymbol{u},\boldsymbol{v}}$. By Proposition 6.34 and Lemma 6.9 we get

$$R_{f,gh} = R_{f,g} R_{f,h},$$

so the proposition follows from Definition 6.18. □

### 6.5.3 Infinite Multiplicities

The additivity of multiplicity of intersection allows us to extend the notion to cases of curves with common components. Let $C$ and $D$ be curves. If $p$ lies on a common component of $C$ and $D$ then we let $i_p(C, D) = \infty$. Otherwise, let $K$ be the sum of the common components of $C$ and $D$; then $C \setminus K$ and $D \setminus K$ are curves with no common component and we let $i_p(C, D) = i_p(C \setminus K, D \setminus K)$.

We can then define $C \cdot D$ as above using the extended notion of multiplicities. The multiset $C \cdot D$ is the sum of $(C \setminus K) \cdot (D \setminus K)$ and infinitely many copies of $K$.[3] It is still the case that $\lfloor C \cdot D \rfloor = \lfloor C \cap D \rfloor$.

*Exercise 6.35* Show that Proposition 6.29 holds under this extended definition of multiplicities: for any curves $C$, $D$ and $E$, $C \cdot (D + E) = C \cdot D + C \cdot E$. «

### 6.5.4 Shifts

**Proposition 6.36** *Let $R$ be a unique factorisation domain. Let $f, g, h \in R[x]$, and let $d \geqslant \deg f$, $e \geqslant \deg g, \deg f + \deg h$. Then*

$$\operatorname{res}_x^{d,e}(f, g) = \operatorname{res}_x^{d,e}(f, fh + g).$$

***Proof*** The first $e$ rows of $M^{d,e}(f, g)$ equal the first $e$ rows of $M^{d,e}(f, fh + g)$, namely they consist of the matrix $M^{d,e}(f)$ (see Sect. 3.2). The last $d$ rows of $M^{d,e}(f, fh+g)$ are the matrix $M^{e,d}(fh+g)$ which is the sum $M^{e,d}(fh)+M^{e,d}(g)$. We show that the rows of $M^{e,d}(fh)$ are linear combinations of the rows of $M^{d,e}(f)$. This will show that $M^{d,e}(f, fh+g)$ is obtained from $M^{d,e}(f, g)$ by row operations that do not change the determinant, namely the addition of scalar multiples of some rows to other rows.

[3] This is one of the few occasions in which we allow elements to appear infinitely many times in a multiset.

Take for example $\underline{u}_0$, the first row of $M^{e,d}(fh)$. It is the row of coefficients of $fh$ of length $d+e$, and so equals the product $\underline{b}_0 \cdot M^{d,e}(f)$, where $\underline{b}_0$ is the row of coefficients of $h$ of length $e$, and so is a linear combination of the rows of $M^{d,e}(f)$. The second row of $M^{e,d}(fh)$ is the row of coefficients of $xfh$ of length $d+e$ and equals $\underline{b}_1 \cdot M^{d,e}(f)$ where $\underline{b}_1$ is the row of coefficients of $xh$ of length $e$, and so on. The fact that $e \geqslant \deg f + \deg h$ shows that for all $j < d$, $\deg(x^j h) < e$ and so the row $\underline{b}_j$ of coefficients of $x^j h$ of length $e$ does not miss any coefficients, and multiplied by $M^{d,e}(f)$ gives the $(j+1)$st row of $M^{e,d}(fh)$. □

Applying Lemma 6.9 gives:

**Lemma 6.37** *If $f, g, h \in R[x, y]$ are nonconstant, $x, y$-homogeneous,* $\deg g = \deg h + \deg f$, *and $fh + g \neq 0$, then*

$$\operatorname{res}_{x,y}(f, g) = \operatorname{res}_{x,y}(f, fh + g).$$

In the following lemma and later we write $i_p(f, g)$ for $i_p\left(V_{\mathbb{P}^2}(f), V_{\mathbb{P}^2}(g)\right)$.

**Proposition 6.38** *Let $f, g, h \in \mathbb{K}[w, x, y]$ be homogeneous, with* $\deg g = \deg f + \deg h$. *Suppose that $fh + g \neq 0$. Then for all $p \in \mathbb{P}^2$,*

$$i_p(f, g) = i_p(f, fh + g).$$

Note that if $fh + g \neq 0$ then it is homogeneous of degree $\deg g$. Also note that if $f$ and $g$ have no common factor then $fh + g \neq 0$.

***Proof*** For any point $p$, $p \in V_{\mathbb{P}^2}(f) \cap V_{\mathbb{P}^2}(g)$ if and only if $p \in V_{\mathbb{P}^2}(f) \cap V_{\mathbb{P}^2}(fh + g)$: if $f(\boldsymbol{p}) = 0$ then $g(\boldsymbol{p}) = (fh + g)(\boldsymbol{p})$. Further, the common factors of $f$ and $g$ are the common factors of $f$ and $fh+g$: if $k \mid f$ then $k \mid g$ if and only if $k \mid (fh+g)$. Hence for all $p$, $i_p(f, g) = \infty$ if and only if $i_p(f, fh+g) = \infty$. Otherwise, $p$ does not belong to any common components. Dividing by the greatest common divisor $k$ of $f$ and $g$ we now compare $i_p(f/k, g/k)$ and $i_p(f/k, (f/k) \cdot h + (g/k))$. Thus after renaming, we assume that $f$ and $g$ have no common factors.

Now we observe that $(fh+g)_{\boldsymbol{u},\boldsymbol{v}} = f_{\boldsymbol{u},\boldsymbol{v}} h_{\boldsymbol{u},\boldsymbol{v}} + g_{\boldsymbol{u},\boldsymbol{v}}$; so by Lemma 6.37, $R_{f,g} = R_{f,fh+g}$, whence the result follows from Definition 6.18. □

### 6.5.5 Categoricity of Multiplicity of Intersection

**Theorem 6.39** *Multiplicity of intersection is the unique function (defined on pairs of curves and points) satisfying:*

1. *For any change of coordinates $\alpha$ of $\mathbb{P}^2$, $i_{\alpha(p)}(\alpha[C], \alpha[D]) = i_p(C, D)$.*
2. *$i_p(C, D) = i_p(D, C)$.*

3. $i_p(C, D) > 0$ *if and only if* $p \in C \cap D$; $i_p(C, D) = \infty$ *if and only if* $p$ *lies on a common component of* $C$ *and* $D$.
4. *If* $\ell$ *and* $\ell'$ *are lines and* $p$ *is their point of intersection then* $i_p(\ell, \ell') = 1$.
5. $i_p(C, D + E) = i_p(C, D) + i_p(C, E)$.
6. *If* $\deg g = \deg f + \deg h$ *and* $fh + g \neq 0$ *then* $i_p(f, g) = i_p(f, fh + g)$.

***Proof*** First we mention that the multiplicity of intersection $i_p(C, D)$ which we defined satisfies all the properties listed. They have all been proved earlier: see Propositions 6.21, 6.28, 6.19, 5.27 and 6.29 and Theorem 6.38. We need to observe that the properties hold when we extend to pairs of curves with common components (and infinite multiplicities); for example see Exercise 6.35.

Now we prove uniqueness. Let $j_p(C, D)$ be a function which satisfies all the properties. Again we write $j_p(f, g)$ for $j_p(V_{\mathbb{P}^2}(f), V_{\mathbb{P}^2}(g))$. For each triple $(p, f, g)$ we show that $j_p(f, g) = i_p(f, g)$. Properties (3) and (5) show that we may assume that $f$ and $g$ have no common factor. The rest is done in three steps.

**I.** For any homogeneous $f$ which is not divisible by $x$, for any $p$, $j_p(x, f) = i_p(x, f)$.

This we prove by induction on $d = \deg_x f$. Suppose (I) has been shown for all $f$ such that $\deg_x f < d$.

If $d = 0$ then $f \in \mathbb{K}[w, y]$ and so is a product of linear polynomials. For each linear polynomial $g$ not associate to $x$ we have $i_p(g, x) = j_p(g, x)$ (properties (3) and (4)). So in this case the equality follows from (5).

Suppose that $d > 0$. Let $h \in \mathbb{K}[w, y]$ be the coefficient of $x^d$ in $f$. Note that $h$ is homogeneous. Then $\deg_x(f - x^d h) < d$. $x$ does not divide $f - x^d h$ and so by induction, $j_p(x, f - x^d h) = i_p(x, f - x^d h)$. However by (6) we have $j_p(x, f) = j_p(x, f - x^d h)$ and $i_p(x, f) = i_p(x, f - x^d h)$.

**II.** For any line $\ell$ and any curve $D$ which does not contain $\ell$, for any $p \in \mathbb{P}^2$, $j_p(\ell, D) = i_p(\ell, D)$.

Let $\alpha$ be a change of coordinates such that $\alpha[\ell] = V_{\mathbb{P}^2}(x)$. Then $V_{\mathbb{P}^2}(x)$ is not a component of $\alpha[C]$; by (1) and (I),

$$j_p(\ell, D) = j_{\alpha(p)}(\alpha[\ell], \alpha[D]) = i_{\alpha(p)}(\alpha[\ell], \alpha[D]) = i_p(\ell, D).$$

**III.** For any homogeneous $f$ and $g$ with no common factor, for any $p \in \mathbb{P}^2$, $j_p(f, g) = i_p(f, g)$.

This is proved by induction on $\deg_y f + \deg_y g$. If $\deg_y f = 0$ then $f$ is the product of linear polynomials and then (II) and (5) yield the result; of course the case $\deg_y g = 0$ is the same.

Suppose, then, that $d = \deg_y f > 0$ and $e = \deg_y g > 0$. By using (2) and (5) we may assume that both $f$ and $g$ are irreducible. We may assume that $e \geqslant d$; otherwise we switch $f$ and $g$, using property (2). Also, in light of (II), we may assume that neither $f$ nor $g$ are linear.

Write $f = \sum_{i \leqslant d} f_i y^i$ and $g = \sum_{i \leqslant e} g_i y^i$ with $f_i, g_j \in \mathbb{K}[w, x]$. Every $f_i$ and $g_j$ is homogeneous and so a product of linear polynomials. Let

$$h = f_d g - y^{e-d} g_e f.$$

Then $h$ is homogeneous, and is designed so that $\deg_y h < e$. Also note that $f$ does not divide $h$ (it does not divide $g$, and does not divide $f_d$). So $h \neq 0$. By induction, since $f$ and $h$ have no common factor, $j_p(f, h) = i_p(f, h)$. By (6), since $h \neq 0$, $j_p(f, h) = j_p(f, f_d g)$, and the same holds for $i_p$. By (5), $j_p(f, f_d g) = j_p(f, f_d) + j_p(f, g)$, and the same holds for $i_p$. Since $f_d$ is a product of linear polynomials, by (II) and (5), $j_p(f, f_d) = i_p(f, f_d)$. Now $j_p(f, g) = i_p(f, g)$ follows by subtraction. □

## 6.6 Affine Calculations

To actually calculate intersection multiplicities we usually use the properties of multiplicity of intersection listed in Theorem 6.39 rather than apply the definitions. That this is always possible is indicated by the proof of Theorem 6.39. Many calculations, though, are simpler than the method described in the proof of that proposition. In most cases, it is simpler to use affine coordinates.

For affine curves $C$ and $D$ in the plane $\mathbb{A}^2$ and a point $(a, b) \in \mathbb{A}^2$ we let $i_{(a,b)}(C, D)$ equal $i_{(1:a:b)}(C^\sharp, D^\sharp)$ where $C^\sharp$ and $D^\sharp$ are the projective closures of $C$ and $D$. In terms of polynomials, for $f, g \in \mathbb{K}[x, y]$ we let $i_{(a,b)}(f, g) = i_{(1:a:b)}(f^\sharp, g^\sharp)$. Note that common factors of $f$ and $g$ correspond to common factors of $f^\sharp$ and $g^\sharp$ (Proposition 4.20).

The properties from Theorem 6.39 translate to the affine version as follows.

**Proposition 6.40** *For nonzero $f, g, h \in \mathbb{K}[x, y]$ and $p \in \mathbb{A}^2$,*

1. $i_p(f, g) = i_p(g, f)$.
2. $i_p(f, g) > 0$ *if and only if* $f(p) = g(p) = 0$; $i_p(f, g) = \infty$ *if and only if* $h(p) = 0$ *for some common factor $h$ of $f$ and $g$.*
3. *If $f$ and $g$ are linear then* $i_p(f, g) \leqslant 1$.
4. $i_p(f, gh) = i_p(f, g) + i_p(f, h)$.
5. *If* $fh + g \neq 0$ *then* $i_p(f, g) = i_p(f, fh + g)$.

***Proof*** These are mostly quite immediate, using Proposition 4.20. For (4) use the fact that $(gh)^\sharp = g^\sharp h^\sharp$. (5) needs an explanation. Let $d = \deg(fh)$ and $e = \deg(g)$. Suppose that $d > e$. Then (Remark 4.21) $(fh + g)^\sharp = (fh)^\sharp + w^{d-e} g^\sharp$ and is certainly nonzero. Since $p \notin \ell_\infty$, $i_p(f^\sharp, g^\sharp) = i_p(f^\sharp, w^{d-e} g^\sharp)$ which by Proposition 6.38 is $i_p(f^\sharp, f^\sharp h^\sharp + w^{d-e} g^\sharp)$ which equals $i_p(f, fh + g)$. If $d < e$ then $(fh + g)^\sharp = w^{e-d}(fh)^\sharp + g^\sharp$ and then we simply use the equality $i_p(f^\sharp, g^\sharp) = i_p(f^\sharp, f^\sharp \cdot w^{e-d} h^\sharp + g^\sharp)$. If $d = e$ then we could get a cancellation;

$c = \deg(fh + g)$ may be smaller than $d$; then $w^{d-c}(fh+g)^\sharp = (fh)^\sharp + g^\sharp$ (Remark 4.21 again). In this case $i_p(f^\sharp, (fh+g)^\sharp) = i_p(f^\sharp, w^{d-c}(fh+g)^\sharp) = i_p(f^\sharp, f^\sharp h^\sharp + g^\sharp) = i_p(f^\sharp, g^\sharp)$ giving the desired equality. □

*Example 6.41 ([Bix06, p.9])* Let $\mathbb{K} = \mathbb{C}$. Let $f = y - x^2$ and $g = y^3 + 2xy + x^6$; we calculate $i_o(f, g)$. We first eliminate $y$ from the second polynomial. Using long division with respect to $y$, we see that upon dividing $g$ by $f$ we get the remainder $2x^3 + 2x^6$, that is, $g = fh + (2x^3 + 2x^6)$ for some $h \in \mathbb{K}[x, y]$. By property (5) of Proposition 6.40, $i_o(f, g) = i_o(f, 2x^3 + 2x^6)$. For the second step we pull out an $x^3$ factor and use property (4) to see that $i_o(f, g) = i_o(f, x^3) + i_o(f, 2 + 2x^3)$. Noting that $o \notin V_{\mathbb{A}^2}(2 + 2x^3)$, property (2) says that $i_o(f, 2 + 2x^3) = 0$. Now using property (4) twice we see that $i_o(f, x^3) = 3i_o(f, x)$. Finally, by property (5) again (and (1)), $i_o(f, x) = i_o(x, y - x^2) = i_o(x, y)$ which is 1 by property (3).[4] Hence $i_o(f, g) = 3$. «

## 6.7 Multiplicities, Orders and Tangents

We prove two properties of intersection multiplicity which reflect the intuition that a tangent is a good approximation of a curve.

**Theorem 6.42** *Let $C$ and $D$ be curves and let $p \in C \cap D$. Then*

$$i_p(C, D) \geqslant o_p(C) \cdot o_p(D).$$

*Equality holds if and only if no line is a tangent to both $C$ and $D$ at $p$.*

***Proof*** If $p$ lies on a common component of $C$ and $D$ then strict inequality certainly holds, and $C$ and $D$ share a tangent at $p$. So we assume that $C$ and $D$ have no common component. After a change of coordinates we suppose that: (1) $p$ is the origin; (2) the vertical point at infinity doesn't lie on $C$ or on $D$; (3) no point on the $y$-axis other than the origin lies on both $C$ and $D$; (4) the $y$-axis is not a tangent to $C$ or to $D$ at $p$; and (5) the $x$-axis is a tangent to $C$ at $p$. [To do this, choose a point $q$ such that $\overline{p\,q}$ intersects $C \cap D$ only at $p$ and is not a tangent to $C$ or to $D$ at $p$, and move $q$ to the vertical point at infinity. Also choose a point $r \notin \overline{p\,q}$ on a tangent to $C$ at $p$, and move $r$ using the Three Point Lemma.]

Let $d = \deg C$ and $e = \deg D$. Let $f$ defining $C\restriction_{\mathbb{A}^2}$ be the dehomogenisation of a polynomial defining $C$; Similarly choose $g$ for $D$. By Remark 6.26, $i_p(C, D)$ is the multiplicity of the root 0 of $\mathrm{res}_y(f, g)$.

[4] Alternatively, the origin is nonsingular on the parabola $y = x^2$ and the $y$-axis is not the tangent to the parabola at the origin, and hence intersects it once there.

Write $f = \sum f_i y^i$ and $g = \sum g_i y^i$. Let $r = o_p(C)$ and $s = o_p(D)$. Since the lowest order terms in $f$ have degree $r$ (Proposition 5.16), $x^r$ divides $f_0$, $x^{r-1}$ divides $f_1$, and so on; and the same holds for $g$. We now manipulate the Sylvester matrix $M_y(f, g)$ in a similar way to that of the proof of Proposition 6.2. We multiply the first row by $x^s$, the second by $x^{s-1}, \ldots,$ all the way to the $s$th row. The rest of the first $e$ rows of $M_y(f, g)$ (the rows with the $f$-coefficients) are left unchanged; note that $s \leqslant e$. Then we multiply the $(e + 1)$st row (the first row with the $g$-coefficients) by $x^r$, the next by $x^{r-1}$, and so on; again note that there are $d$ rows with $g$-coefficients and $r \leqslant d$. An examination of the resulting matrix shows that every entry in the first column is divisible by $x^{r+s}$, every entry of the next is divisible by $x^{r+s-1}$, etc, all they way to the $(r+s)$th column. We thus divide the first column by $x^{r+s}$, the second by $x^{r+s-1}, \ldots$. Call the resulting matrix $N$. The calculation performed in the proof of Proposition 6.2 shows that $\det(N) = \operatorname{res}_y(f, g)/x^{rs}$, and so $x^{rs}$ divides $\operatorname{res}_y(f, g)$, so $i_p(C, D) \geqslant rs$.

We prove the rest of the theorem for the special case $s = r = 1$, i.e., when $p$ is nonsingular on both $C$ and $D$. For a proof of the general theorem, using resultants, see [BK86, Prop.6.1.3]. For a proof using analytic parameterisations, see Chap. 15.

Since we chose the tangent to $C$ at $p$ to be the $x$-axis, we need to show that $i_p(f, g) > 1$ if and only if the $x$-axis is the tangent to $D$ at $p$. Since $p \in C$, $f$ has no constant term; by Proposition 5.16, it has no $x$-term, so by grouping the monomials in which $y$ does not appear, we can write $f = y\alpha(x, y) + x^2\beta(x)$ for polynomials $\alpha$ and $\beta$, where $\alpha(0, 0) \neq 0$. Similarly, since $g(p) = 0$, we write $g = y\gamma(x, y) + x\delta(x)$; the $x$-axis is the tangent to $D$ at $p$ if and only if $\delta(0) = 0$.

Since $\alpha(p) \neq 0$, we have $i_p(f, g) = i_p(f, \alpha g)$. Since $\alpha g - \gamma f = x(\alpha\delta - x\gamma\beta)$, by Proposition 6.40,

$$i_p(f, \alpha g) = i_p(f, \alpha g - \gamma f) = i_p(f, x) + i_p(f, \alpha\delta - x\gamma\beta).$$

Since the $y$-axis is not the tangent to $C$ at $p$, and $p$ is nonsingular on $C$, we have $i_p(f, x) = 1$; so it remains to show that $\delta(0) = 0$ if and only if $i_p(f, \alpha\delta - x\gamma\beta) > 0$. Since $f(p) = 0$, the latter holds if and only if $p \in V_{\mathbb{A}^2}(\alpha\delta - x\gamma\beta)$; setting $x = 0$ and recalling that $\alpha(p) \neq 0$ gives the desired equivalence. □

The following proposition will be used in the next chapter.

**Proposition 6.43** *Let $C$ and $D$ be curves in $\mathbb{P}^2$ with no common component. Suppose that $p \in C$ is nonsingular on $C$. Let $\ell$ be a line which is a component of neither $C$ nor $D$. Suppose that $i_p(D, \ell) < i_p(C, \ell)$. Then $i_p(C, D) = i_p(D, \ell)$.*

Roughly, the proposition says that if $\ell$ is "closer" to $C$ than to $D$ then as far as $D$ is concerned, $\ell$ is a good approximation for $C$ with respect to counting intersections.

***Proof*** If $p \notin D$ then $i_p(D, \ell) = i_p(C, D) = 0$, so we assume that $p \in D$. This implies that $i_p(D, \ell) \geqslant 1$ so $i_p(C, \ell) > 1$ and so, as $p$ is nonsingular on $C$, $\ell = \ell_p C$

is the tangent to $C$ at $p$ (Theorem 5.34). We change coordinates so that: (1) $p$ is the origin; (2) $\ell$ is the $x$-axis.

Let $f$ and $g$ be polynomials in $\mathbb{K}[x, y]$ which define the restrictions of $C$ and $D$ to $\mathbb{A}^2$. Neither have constant terms, so as in the proof of Theorem 6.42, we group the monomials in which $y$ does not appear so we write $f = y\alpha(x, y) + x^r\beta(x)$ and $g = y\gamma(x, y) + x^s\delta(x)$ for polynomials $\alpha$, $\beta$, $\gamma$ and $\delta$; here $r$ and $s$ are chosen to be the highest power of $x$ dividing all monomials with no $y$-terms, so $\beta(0), \delta(0) \neq 0$. Further, $t \mapsto (t, 0)$ is a linear parameterisation of the $x$-axis, and $f(t, 0) = t^r\beta(t)$ and similarly for $g$; by Lemma 5.32, $r = i_p(C, \ell)$ and $s = i_p(D, \ell)$. So we are assuming that $r > s$. Also, the assumption that $\ell$ is the tangent to $C$ at $p$ implies $\alpha(p) \neq 0$.

We need to show that $i_p(f, g) = s$. We manipulate as in the previous proof:

$$i_p(f, g) = i_p(f, \alpha g) = i_p(f, \alpha g - \gamma f) = i_p(f, x^s\alpha\delta - x^r\gamma\beta) = \\ s \cdot i_p(f, x) + i_p(f, \alpha\delta - x^{r-s}\gamma\beta).$$

Since $x = 0$ is not the tangent to $C$ at $p$ and $p$ is nonsingular on $C$, we have $i_p(f, x) = 1$; Since $r > s$, we have $(\alpha\delta - x^{r-s}\gamma\delta)(p) = \alpha(p)\delta(0) \neq 0$ so $i_p(f, \alpha\delta - x^{r-s}\gamma\beta) = 0$. □

For a proof using parameterisations, see Chap. 15.

*Exercise 6.44* Let $C$ and $D$ be curves in $\mathbb{P}^2$ with no common component. Suppose that $p \in C$ is nonsingular on $C$. Let $\ell$ be a line which is a component of neither $C$ nor $D$. Show that if $i_p(C, D) \geqslant i_p(C, \ell)$ then $i_p(D, \ell) \geqslant i_p(C, \ell)$.

Informally, this says that if at $p$, $D$ approaches $C$ at least as closely as $\ell$ approaches $C$, then $\ell$ approaches $D$ at least as closely as it approaches $C$. «

## 6.8 Further Exercises

When calculating we assume $\mathbb{K} = \mathbb{C}$.

6.45 Every integral domain is a subring of an algebraically closed field (one first takes the field of fractions, then the *algebraic closure* of that field). Use this to show the following generalisation of Corollary 6.33 and Proposition 6.34: for any integral domain $R$, for any nonconstant $f, g, h \in R[x]$, $\operatorname{res}_x(f, gh) = \operatorname{res}_x(f, g) \cdot \operatorname{res}_x(f, h)$.

6.46 Let $f = (x - a_1)(x - a_2) \cdots (x - a_d) \in \mathbb{K}[x]$. Show that the discriminant $\operatorname{disc}(f)$ (see Proposition 5.7) equals $\pm \prod_{1 \leqslant i < j \leqslant d} (a_j - a_i)^2$.

6.47 Let $f \in \mathbb{K}[x, y]$ and $g \in \mathbb{K}[x]$. Suppose that the curve $y = g$ is not a component of the curve $f = 0$. Show that for all $\lambda \in \mathbb{K}$, the intersection multiplicity $i_p(f, y - g)$ between the curves $f = 0$ and $y = g$ at the point $p = (\lambda, g(\lambda))$ is the multiplicity of $\lambda$ as a root of the polynomial $f(t, g(t))$. (Hint: divide $f$ by $y - g$ with respect to the variable $y$ (i.e., over the ring $\mathbb{K}[x]$) to get $f = (y - g)h + f(x, g)$, and use Proposition 6.40(5). This exercise generalises Definition 5.25: the map $\lambda \mapsto (\lambda, g(\lambda))$ is a parameterisation of the curve $y = g$, and $f(t, g(t))$ is the associated "intersection polynomial". We could similarly use rational parameterisations (Definition 3.29) to calculate intersection multiplicities; this follows from the results of Chap. 15.)

6.48 In this exercise we sketch an alternative proof of Proposition 6.43, using the resultant. Dehomogenising and changing coordinates, let $p$ be the origin, let $\ell$ be the $x$-axis, and let $f, g \in \mathbb{K}[x, y]$; we assume that $p$ is nonsingular on $f = 0$ and that $i_p(f, \ell) > i_p(g, \ell)$; we assume that $p \in V_{\mathbb{A}^2}(g)$ so $\ell$ is the tangent to $V_{\mathbb{A}^2}(f)$ at $p$. We assume that the curves have been arranged so that $i_p(f, g)$ is the multiplicity of 0 as a root of the resultant $\mathrm{res}_y(f, g)$. Let $s = i_p(g, \ell)$; we need to show that $i_p(f, g) = s$.

(a) Show that $x^s$ divides the first column of the Sylvester matrix $M_y(f, g)$. (b) Let $N$ be the matrix obtained by dividing the first column of $M_y(f, g)$ by $x^s$, and let $N(0)$ be the result of substituting 0 for $x$ in every entry of $N$. Write $f = \sum f_i y^i$ and $g = \sum_j g_j y^j$. Let $b = (g_0/x^s)(0)$. Show that $\det(N(0)) = \pm b \cdot \mathrm{res}_y((f - f_0)/y, (g - g_0)/y)$ (c) Show that $p \notin V_{\mathbb{A}^2}((f - f_0)/y)$. (d) Conclude that $\det(N(0)) \neq 0$. (e) Show that $i_p(f, g) = s$.

6.49 The purpose of this exercise is to show that rational curves (Definition 3.29) are irreducible.

(a) Show that if $A$ and $B$ are irreducible affine curves and $A \neq B$, then $A \setminus B$ contains infinitely many points. (See Exercise 3.42). (b) Let $h = f/g \in \mathbb{K}(t)$ (the field of formal rational functions). Suppose that there are infinitely many $a \in \mathbb{K}$ such that $h(a)$ is defined (i.e. $g(a) \neq 0$) and $h(a) = 0$. Show that $h = 0$. (c) Suppose that $\psi = (\psi_x, \psi_y)$ is a rational parameterisation of an affine curve $f = 0$, which has no repeated components. Show that there is an irreducible factor $h$ of $f$ such that $h(\psi_x(a), \psi_y(a)) = 0$ for infinitely many $a \in \mathbb{K}$. Conclude that $h(\psi_x, \psi_y) = 0$ in the field of rational functions $\mathbb{K}(t)$. (d) Show that all but finitely many points of $f = 0$ lie on $h = 0$; conclude that $f \sim h$ and so $f$ is irreducible.

**Affine Calculations**

6.50 For the following pairs $(f, g)$ of polynomials in $\mathbb{C}[x, y]$, find the affine multiplicity of intersection $i_o(f, g)$ at the origin:

(i) $f = y^3 + 2x^5, g = xy^2 + y - 3x^3$.
(ii) $f = y - x^3, g = y^4 + 6x^3y + x^8$.
(iii) $f = y^3 - x^2, g = y^2 - x^3$.
(iv) $f = y^3 - x^2, g = xy^2 - 4y^2 - x^3$.

(v) $f = y^3 - x^2 - 1, g = y^2 - 2xy^4 + x^5$.

For more see [Bix06, Ex.1.1]

6.51 Find the following, by translating the point to the origin:

(i) $i_{(0,2)}(x^3 + y^2 - 4, x^3y + y^2 - 4)$.
(ii) $i_{(1,2)}(xy - 2, xy^2 - 4)$.
(iii) $i_{(-1,2)}(x^2 + xy + y - 1, xy^2 + 4)$.

For more see [Bix06, Ex.3.4]

**Projective Calculations**

6.52 Find:

(i) $i_{(0:1:2)}(2x^2 + w^2 - xy, y^2 + yw - 4x^2)$.
(ii) $i_{(0:3:-1)}(3y^2 + xy + 2w^2, xy^2 + 3y^3 - w^3)$.
(iii) $i_{(0:1:-1)}(y^2 + xy - w^2, y^2 + 2w^2 - x^2)$.

For more see [Bix06, Ex.3.5]

6.53 For the following $f, g \in \mathbb{C}[w, x, y]$, find the points of intersection of $V_{\mathbb{P}^2}(f)$ and $V_{\mathbb{P}^2}(g)$, and their multiplicities. (Sometimes the resultant computation is quicker.)

(i) $f = y^5 - x(y^2 - xw)^2, g = y^4 + y^3w - x^2w^2$.
(ii) $f = (x^2 + y^2)^2 + 3x^2yw - y^3w, g = (x^2 + y^2)^3 - 4x^2y^2w^2$.
(iii) $f = x^4 + y^4 - y^2w^2, g = x^4 + y^4 - 2y^3w - 2x^2yw - xy^2w + y^2w^2$.
(iv) $f = y^2 - xw, g = y^3 - xw^2 + x^3$.
(v) $f = (x^2 + y^2)w + x^3 + y^3, g = x^3 + y^3 - 2wxy$.

For more see [Gib98, Sec.14.4,14.5], [Wal50, III.3.3], or [Ful69, Ex.5.3]

6.54 For the following $f, g \in \mathbb{C}[x, y]$, find the points of intersection of the projective closures of the affine curves $f = 0$ and $g = 0$, their multiplicities, and the tangents at these points to both curves.

(i) $f = y - x^3, g = y - x^5$;
(ii) $f = y^2 - x^3, g = y^3 - x^4$;
(iii) $f = (x^2 + y^2)^2 - 2y(x^2 + y^2) - x^2, g = x^2 + y^2 - y$.

**Bézout's Theorem**

6.55 Let $C$ and $D$ be two curves which intersect at precisely $\deg C \cdot \deg D$ distinct points. Show that $C$ and $D$ have no common component; show that for all $p \in C \cap D$, $p$ is nonsingular on both $C$ and $D$ and that $\ell_p C \neq \ell_p D$.

6.56 Show that for each partition of 4, namely, [4], [3, 1], [2, 2], [2, 1, 1] or [1, 1, 1, 1], there are irreducible conics $C$ and $D$ in $\mathbb{P}^2(\mathbb{C})$ which realise that partition as their intersection pattern: for example, for [3, 1], this means that $C$ and $D$ intersect at two points $p$ and $q$, with $i_p(C, D) = 3$ and $i_q(C, D) = 1$.

**Applications to Singular Points**

6.57 Let $C$ be a curve and let $d = \deg C$. Suppose that some line $\ell$ contains $> d/2$ many singular points of $C$. Show that $\ell$ is a component of $C$.

6.58 Let $C$ be a curve with no repeated components; let $d = \deg C$. Improving on Exercise 6.8, show that

$$\sum_{p \in C} o_p(C)(o_p(C) - 1) \leqslant d(d-1).$$

Conclude that the number of singular points on $D$ is bounded by $d(d-1)/2$.

6.59 (a) Let $C$ be an irreducible curve of degree 4 (a *quartic*) in $\mathbb{P}^2$. Show that $C$ has at most three singular points. (Hint: suppose that there are four singular points. Let $p$ be another point on $C$. Let $D$ be a conic curve which passes through $p$ and the four singular points (Exercise 4.74). Obtain a contradiction.) (b) Show that the *hypocycloid* quartic $3(x^2+y^2)^2+8x(3y^2-x^2)+6(x^2+y^2) = 1$ has three singular points.[5]

6.60 (a) Let $C$ be a cubic curve with exactly one singular point which is a node (an ordinary double point). Show that $C$ is irreducible. (b) Let $C$ be a quartic curve with exactly one singular point which is a an ordinary triple point. Show that $C$ is irreducible. (c) Let $C$ be a quartic curve with three non-collinear singular points. Assume they are all double points and that they are either all ordinary or all not ordinary. Show that $C$ is irreducible. (d) Conclude that the polynomial $x^2y^2 + x^2 + y^2$, and that the hypocycloid from the previous exercise, are both irreducible [Gib98, Ex.14.5.7].

[5] For a generalisation to higher degrees, see for example [Fis01, Sec.3.8].

6.61 Let $C$ be the projective closure of $x^2y^3 + y^2 + x^2 = 0$. Show that $C$ has three singular points: a triple point, a cusp, and a node. Conclude that $C$ doesn't have a component which is a line [Gib98, Ex.12.5.12].

**The Nine Associated Points**

For the following exercises, we work with the space of cubics $\mathbb{G}_3$ (see Sect. 4.6), which by Exercise 4.38 is isomorphic to $\mathbb{P}^9$ via the map $\iota_3$. Recall the notion of a linear family of curves (Definition 4.39), and more generally, of subspaces of projective space (Definition 4.10). Since any two distinct points in $\mathbb{P}^9$ determine a line, any two distinct cubic curves $C$, $C'$ in $\mathbb{G}_3$ determine a linear family of cubics, which we denote by $\overline{C\,C'}$.

Our aim in the following exercises is to show that if two cubics $C$ and $C'$ intersect in nine distinct points, then any other cubic which passes through eight of these must also pass through the ninth.

6.62 Show that if $C$ and $C'$ are two distinct cubic curves, then for all $p \in C \cap C'$, for every $D \in \overline{C\,C'}$ we have $p \in D$.

For the following exercises, fix two cubics $C$ and $C'$, and suppose that $C \cdot C' = C \cap C' = \{p_1, p_2, \ldots, p_9\}$ consists of nine *distinct* points.

6.63 (a) Show that no line contains four of the points $p_1, \ldots, p_9$. (Hint: $C$ and $C'$ do not have a common component.) (b) Show that no conic curve contains seven of these points. (c) Show that any five of the points $p_1, \ldots, p_9$ lie on a *unique* conic curve. (By Exercise 4.74 there is such a conic.)

Suppose that $D$ is a cubic which does not in the family $\overline{C\,C'}$. By Exercise 4.60, there is a unique plane $P$ of $\mathbb{G}_3$ which contains $C$, $C'$ and $D$. Note that for any distinct $A, B \in P$, the linear family $\overline{A\,B}$ is contained in $P$.

6.64 Show that for any two points $q, r \in \mathbb{P}^2$ and a plane $P$ of $\mathbb{G}_3$, there is some cubic $E \in P$ which passes through both $q$ and $r$.

In the next exercise, we show that if $D$ is a cubic which passes through $p_1, \ldots, p_8$, then $D \in \overline{C\,C'}$.

6.65 Suppose, for a contradiction, that there is a cubic $D$ that is not in $\overline{C\,C'}$ but that all of $p_1, \ldots, p_8$ lie on $D$. Let $P$ be the plane determined by $C, C'$ and $D$. Observe that for all $E \in P$, $p_1, \ldots p_8 \in E$. (a) Suppose that $p_1$, $p_2$ and $p_3$ lie on a line $\ell$. Let $Q$ be a conic curve which passes through $p_4, \ldots, p_8$. Choose a point $q \neq p_1, p_2, p_3$ on $\ell$ and another point $r \notin \ell \cup Q$. By Exercise 6.64, let $E \in P$ be a cubic which passes through $q$ and $r$. Show that $E = \ell + Q$. (b) Conclude

that no three points from $p_1, \ldots, p_8$ are collinear. (c) Show that no six points from $p_1, \ldots, p_8$ lie on a conic curve. (d) Let $\ell = \overline{p_1\, p_2}$, and let $Q$ be a conic passing through $p_3, \ldots, p_7$. Choose $q, r \in \ell$ other than $p_1$ and $p_2$, and let $E \in P$ be a cubic passing through $q$ and $r$. Show that $E = \ell + Q$. Obtain a contradiction.

6.66 Conclude that if two cubics $C$ and $C'$ intersect in nine distinct points, then any other cubic which passes through eight of these must also pass through the ninth.

6.67 Suppose that two cubic curves intersect in nine distinct points. (a) Suppose that six of these points lie on a conic curve. Show that the other three are collinear. (b) Suppose that three of these points are collinear. Show that the other six lie on a conic curve.


# The Elliptic Group

# 7

Let $C$ be a nonsingular cubic curve. Bézout's theorem says that every line intersects $C$ at three points (counting multiplicities). If $p, q \in C$ then the line $\overline{pq}$ intersects the curve at a third point $r$. We let $r = p * q$. In this way we can take two solutions to the cubic equation and manufacture a third. The main aim of this chapter is to show how to modify this binary operation to define an abelian group structure on $C$. This has a variety of applications, including famously to cryptography.

A line $\ell$ intersects the curve $C$ in one of three patterns:

- $C \cap \ell$ contains three distinct points. So $i_p(C, \ell) = 1$ for each $p \in C \cap \ell$. In particular $\ell$ is not a tangent to $C$ at any point (see Theorem 5.34).
- $C \cap \ell$ contains two points $p$ and $q$, with $i_q(C, \ell) = 1$ and $i_p(C, \ell) = 2$. Thus $\ell = \ell_p C$ (and $p * p = q, q * p = p * q = p$).
- $C \cap \ell$ contains one point $p$, so $i_p(C, \ell) = 3$. Again $\ell = \ell_p C$ is the tangent to $C$ at $p$, and $p * p = p$. Such a point is called a *flex* of $C$.

Our first step is an analysis of the flexes of $C$. In particular, we show that they exist. To do that, we use Bézout's theorem, together with an analysis of the *Hessian matrix* of second order partial derivatives and the curve defined by its determinant (Definition 7.5). We will then choose a flex and declare it to be the identity element $0_C$ of the group; the group operation will then be characterised by declaring that three collinear points add up to $0_C$.

In the last part of the chapter, we consider *normal forms* for elliptic curves. Of particular importance are Legendre's normal form (Proposition 7.26), and Weierstrass's normal form (Lemma 7.27): every elliptic curve is equivalent (by a change of coordinates) to one given by an equation $y^2 = 4x^3 - \alpha x - \beta$ where $\alpha^3 \neq 27\beta^2$.

In this chapter we assume that $\mathbb{K}$ is algebraically closed, and that char($\mathbb{K}$) is either 0, or greater than the degree of any curve that we are considering.



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_7

## 7.1 Flexes

**Definition 7.1** Let $C$ be a curve in $\mathbb{P}^2$ and let $p \in C$ be nonsingular. The point $p$ is a *flex* of $C$ if $i_p(C, \ell_p C) \geqslant 3$.

Another terminology is a *point of inflection*. See for example Fig. 7.1. For the terminology consider the following:

*Exercise 7.2* Let $C$ be an affine curve defined by the equation $y = f(x)$ for some $f \in \mathbb{K}[x]$ of degree $> 1$. Show that a point $(a, b) \in C$ is a flex of $C$ if and only if $(D^{x^2} f)(a) = 0$. (See Exercise 13.64 for a generalisation.) «

Since $i_p(C, \ell) \leqslant \deg C$ for any curve $C$, a conic curve has no flexes. We will see that nonsingular cubic curves do have flexes.

### 7.1.1 Flexes and the Second Order Tangent

Recall that $\ell_p^k C$ is defined for all $k \leqslant \deg C$, even if $k > o_p(C)$ (see Sect. 5.3).

**Lemma 7.3** *Let $C$ be a curve of degree $d \geqslant 2$, and let $p \in C$. Then $\ell_p^1(\ell_p^2 C) = \ell_p^1 C$. In particular, $p$ is nonsingular on $C$ if and only if $p$ is nonsingular on $\ell_p^2 C$.*

***Proof*** Let $f$ define $C$ and let $\boldsymbol{p}$ be a presentation of $p$. We show that $\partial_{\boldsymbol{p}}^1(\partial_{\boldsymbol{p}}^2 f)$ is a constant multiple of $\partial_{\boldsymbol{p}}^1 f$. Consider for example the variable $x$: a calculation shows that

$$D^x\left(\partial_{\boldsymbol{p}}^2 f\right) = 2wD^{xw} f(\boldsymbol{p}) + 2xD^{xx} f(\boldsymbol{p}) + 2yD^{xy} f(\boldsymbol{p});$$

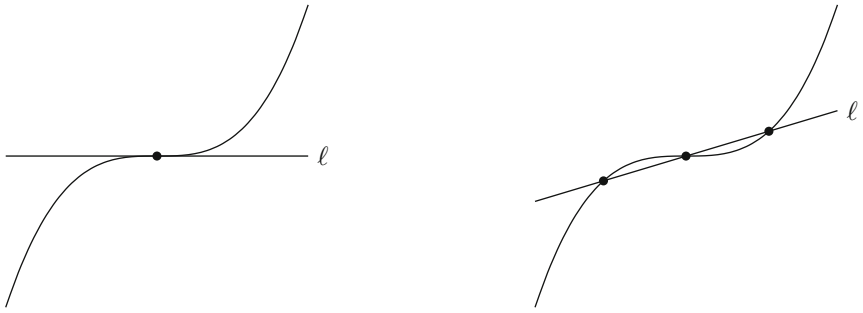



**Fig. 7.1** The line $\ell$ intersects the cubic curve $y = x^3$ at the origin. Perturbing the line a bit gives three points of intersection, witnessing that the origin is a flex of the curve

Since $D^x f$ is homogeneous of degree $d-1$, by Euler's Relation we get

$$D^x\left(\partial_{\boldsymbol{p}}^2 f\right)(\boldsymbol{p}) = 2(d-1)D^x f(\boldsymbol{p}).$$

The same holds for $y$ and $w$. □

**Proposition 7.4** *Let $C$ be a curve of degree at least 2. The following are equivalent for $p \in C$:*

(1) *$p$ is a singular point of $C$, or is a flex of $p$.*
(2) *$\ell_p^2 C = \mathbb{P}^2$ or $\ell_p^2 C$ is reducible.*

***Proof*** First suppose that $p$ is singular on $C$. If $o_p(C) = 2$ then $\ell_p^2 C$ is the sum of two lines, and so is reducible (Corollary 5.22). If $o_p(C) \geqslant 3$ then $\ell_p^2 C = \mathbb{P}^2$ (Definition 5.13). We thus assume that $p$ is nonsingular and show that it is a flex if and only if (2) holds.

By Proposition 5.36, $p$ is a flex if and only if $\ell_p C \subset \ell_p^2 C$. If $p$ is a flex and $\ell_p^2 C \neq \mathbb{P}^2$ then by Study's lemma, $\ell_p C$ is a component of $\ell_p^2 C$ and so $\ell_p^2 C$ is reducible.

In the other direction, certainly if $\ell_p^2 C = \mathbb{P}^2$ then $\ell_p C \subset \ell_p^2 C$ and so $p$ is a flex. Suppose that $\ell_p^2 C \neq \mathbb{P}^2$ and is reducible. Since $\ell_p^2 C$ is a conic curve, it is the sum of two lines $\ell_0$ and $\ell_1$. By Proposition 5.18, $p \in \ell_p^2 C$; without loss of generality, $p \in \ell_0$. By Lemma 7.3, $p$ is nonsingular on $\ell_p^2 C$, and so $p \notin \ell_1$ (Corollary 5.23). Since the tangent to a line $\ell$ at any point of $\ell$ is the line $\ell$ itself, we see that $\ell_0 = \ell_p(\ell_p^2(C))$ (again Corollary 5.23); by Lemma 7.3 again, $\ell_0 = \ell_p C$ and so $\ell_p C$ is a component of $\ell_p^2 C$, from which we conclude that $p$ is a flex of $C$. □

### 7.1.2 The Hessian

We will show the existence of flexes on nonsingular cubic curves using Bézout's theorem, by observing that the flexes of $C$ are the points of intersection of $C$ with another curve, the Hessian curve of $C$.

**Definition 7.5** Let $C$ be a curve in $\mathbb{P}^2$, and let $f$ define $C$. We let

$$H_f = \begin{pmatrix} D^{ww} f & D^{wx} f & D^{wy} f \\ D^{xw} f & D^{xx} f & D^{xy} f \\ D^{yw} f & D^{yx} f & D^{yy} f \end{pmatrix}.$$

We let $\mathcal{H}_C = V_{\mathbb{P}^2}(\det H_f)$.

Let $d = \deg C$. We first note that every entry of $H_f$ is a homogeneous polynomial of degree $d-2$. By its definition, the determinant of $H_f$ is the sum of products of three entries from $H_f$, and so $\det H_f$ is a homogeneous polynomial of degree $3(d-2)$ (or is the zero polynomial); hence $V_{\mathbb{P}^2}(\det H_f)$ is defined. Furthermore, for all $\lambda \in \mathbb{K}^*$, $D^{uv}(\lambda f) = \lambda D^{uv} f$ for all $u, v \in \{w, x, y\}$, and so $H_{\lambda f} = \lambda H_f$, whence $\det H_{\lambda f} = \lambda^3 \det H_f$, so $\det H_{\lambda f} \sim \det H_f$. Thus $\mathcal{H}_C$ does not depend on the choice of $f$ defining $C$. The matrix $H_f$ is called the *Hessian matrix* of $f$ and the curve $\mathcal{H}_C$ is called the *Hessian curve* of $C$.

The main property of the Hessian is the following.

**Proposition 7.6** *Let $C$ be a curve of degree at least 2. Then $C \cap \mathcal{H}_C$ consists of the points $p \in C$ for which $\ell_p^2 C$ is reducible or $\ell_p^2 C = \mathbb{P}^2$.*

With Proposition 7.4 this allows us to conclude:

**Corollary 7.7** *Let $C$ be a curve of degree at least 2. Then $C \cap \mathcal{H}_C$ consists of the singular points of $C$ and the flexes of $C$.* □

To prove Proposition 7.6 we will change coordinates to make calculations simpler. In order to do this we need to show that the Hessian is invariant under changes of coordinates.

**Proposition 7.8** *Let $C$ be a curve in $\mathbb{P}^2$ and let $\alpha$ be a change of coordinates of $\mathbb{P}^2$. Then $\mathcal{H}_{\alpha[C]} = \alpha[\mathcal{H}_C]$.*

***Proof*** Let $f$ define $C$. Let $A \in \mathrm{GL}_3(\mathbb{K})$ be a matrix such that $\boldsymbol{\alpha} = T_A$ is a presentation of $\alpha$. As in Chap. 5 we write $\hat{g}$ for $\boldsymbol{\alpha}^*(g)$. We need to show that $\det \hat{H}_f \sim \det H_{\hat{f}}$.

Since $\begin{pmatrix} \hat{w} \\ \hat{x} \\ \hat{y} \end{pmatrix} = A^{-1} \cdot \begin{pmatrix} w \\ x \\ y \end{pmatrix}$, the chain rule shows that for all $g \in \mathbb{K}[w, x, y]$, $\begin{pmatrix} D^w \hat{g} \\ D^x \hat{g} \\ D^y \hat{g} \end{pmatrix} = (A^{-1})^{\mathsf{t}} \cdot \begin{pmatrix} D^{\hat{w}} g \\ D^{\hat{x}} g \\ D^{\hat{y}} g \end{pmatrix}$ (see the proof of Lemma 5.20), whence $\begin{pmatrix} D^{\hat{w}} g \\ D^{\hat{x}} g \\ D^{\hat{y}} g \end{pmatrix} = A^{\mathsf{t}} \cdot \begin{pmatrix} D^w \hat{g} \\ D^x \hat{g} \\ D^y \hat{g} \end{pmatrix}$. Let $\hat{H}_f$ be the result of applying $\boldsymbol{\alpha}^*$ to each entry of $H_f$. Then

$$\hat{H}_f = A^{\mathsf{t}} \cdot \begin{pmatrix} D^w(D^{\hat{w}} f) & D^w(D^{\hat{x}} f) & D^w(D^{\hat{y}} f) \\ D^x(D^{\hat{w}} f) & D^x(D^{\hat{x}} f) & D^x(D^{\hat{y}} f) \\ D^y(D^{\hat{w}} f) & D^y(D^{\hat{x}} f) & D^y(D^{\hat{y}} f) \end{pmatrix}.$$

Since taking a partial derivative is linear, for any variable $u \in \{w, x, y\}$

$$\begin{pmatrix} D^u D^{\hat{w}} f \\ D^u D^{\hat{x}} f \\ D^u D^{\hat{y}} f \end{pmatrix} = A^{\mathsf{t}} \cdot \begin{pmatrix} D^u D^w \hat{f} \\ D^u D^x \hat{f} \\ D^u D^y \hat{f} \end{pmatrix}.$$

Taking the transpose of the matrix above we see that $\hat{H}_f = A^{\mathsf{t}} \cdot \left(A^{\mathsf{t}} \cdot H_{\hat{f}}\right)^{\mathsf{t}}$. Since $H_f$ is symmetric this equals $A^{\mathsf{t}} \cdot H_{\hat{f}} \cdot A$. Since $\boldsymbol{\alpha}^*$ is a ring automorphism, $\det \hat{H}_f = \det \widehat{H_f}$ and so $\det \widehat{H_f} = (\det A)^2 \cdot \det H_{\hat{f}}$, and $\det A \neq 0$ as $A$ is invertible. □

***Proof of Proposition 7.6*** Let $p \in C$. Fix $f$ defining $C$ and a presentation $\boldsymbol{p}$ of $p$. Note that $p \in \mathcal{H}_C$ if and only if $\det(H_f(\boldsymbol{p})) = 0$ if and only if $H_f(\boldsymbol{p})$ is a singular matrix (recall that $\det(H_f(\boldsymbol{p})) = (\det H_f)(\boldsymbol{p})$).

In one direction, first suppose that $\ell_p^2 C = \mathbb{P}^2$, i.e. $\partial_{\boldsymbol{p}}^2 f = 0$. This implies that $H_f(\boldsymbol{p})$ is the zero matrix, which is singular. Next suppose that $\ell_p^2 C$ is reducible. It is the sum of two lines $\ell$ and $\ell'$. If $\ell \neq \ell'$ then after a change of coordinates we assume that $\ell$ is the $x$-axis and $\ell'$ is the $y$-axis. So $\partial_{\boldsymbol{p}}^2 f \sim xy$. This means that $H_f(\boldsymbol{p})$ has a zero row (in fact only two nonzero entries), so is singular. If $\ell = \ell'$ then after a change of coordinates we assume that $\ell$ is the $x$ axis, so $\partial_{\boldsymbol{p}}^2 f \sim y^2$. In this case $H_f(\boldsymbol{p})$ has only one nonzero entry, and so certainly is singular.

In the other direction suppose that $H_f(\boldsymbol{p})$ is singular. We claim that we can change coordinates so that the last column of $H_f(\boldsymbol{p})$ is the zero column; this would imply that $\partial_{\boldsymbol{p}}^2 f \in \mathbb{K}[w, x]$ (and is homogeneous of course) and so is either reducible or zero. To see that such a change of coordinates is possible, let $B$ be an invertible $3 \times 3$-matrix, let $\boldsymbol{\alpha} = T_{B^{-1}}$ and let $\alpha$ be the induced change of coordinates. The proof of Proposition 7.8 shows that $H_{\boldsymbol{\alpha}^*(f)} = B^{\mathsf{t}} \cdot \boldsymbol{\alpha}^*[H_f] \cdot B$ (we use the fact that $(B^{-1})^{\mathsf{t}} = (B^{\mathsf{t}})^{-1}$). If $\boldsymbol{q} = \boldsymbol{\alpha}(\boldsymbol{p})$ then $\boldsymbol{\alpha}^*[H_f](\boldsymbol{q}) = H_f(\boldsymbol{p})$ (recall that $\boldsymbol{\alpha}^*(g)$ defines $g \circ \boldsymbol{\alpha}^{-1}$) and so $H_{\boldsymbol{\alpha}^*(f)}(\boldsymbol{q}) = B^{\mathsf{t}} \cdot H_f(\boldsymbol{p}) \cdot B$. So it remains to find an invertible matrix $B$ such that the last column of $B^{\mathsf{t}} \cdot H_f(\boldsymbol{p}) \cdot B$ is zero. But this is easy. Since $H_f(\boldsymbol{p})$ is singular there is some nonzero column $\bar{c}$ such that $H_f(\boldsymbol{p}) \cdot \bar{c} = \bar{0}$. There is an invertible matrix $B$ such that $B \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \bar{c}$, and then $\left(B^{\mathsf{t}} \cdot H_f(\boldsymbol{p}) \cdot B\right) \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \bar{0}$, which implies the required property. □

## 7.2 The Group Operation on a Nonsingular Cubic Curve

**Definition 7.9** Let $C$ be a nonsingular cubic curve. For $p \in C$ let $\overline{p\,p}$ be the tangent $\ell_p C$. Let $p, q \in C$ (possibly $p = q$). As discussed above, $C \cdot \overline{p\,q} = [p, q, r]$ for some $r$; we let $r = p * q$.

Since $\overline{p\,q} = \overline{q\,p}$, $p*q = q*p$. If $\ell$ is any line and $C\cdot\ell = [p, q, r]$ then $r = p*q$, as $\ell = \overline{p\,q}$; this is immediate if $p \neq q$; if $p = q$ then the assumption implies that $i_p(C, \ell) > 1$ and so $\ell = \ell_p C$. If $p$ is a flex of $C$ then $C \cdot \ell_p C = [p, p, p]$ and so $p * p = p$. Altogether, we see that if $r = p * q$ then $p * r = q$ and $q * r = p$.

**Proposition 7.10** *Every nonsingular cubic curve has a flex.*

***Proof*** Let $C$ be a nonsingular cubic curve. We observed that $\deg \mathcal{H}_C = 3\cdot(3-2) = 3$, so is nonempty; hence $C \cap \mathcal{H}_C$ is nonempty; now use Corollary 7.7. □

In fact, by Bézout's theorem, $C \cdot \mathcal{H}_C$ contains nine points; we will show that they are all distinct, so every nonsingular cubic curve has nine flexes; see Exercises 7.35 and 7.36.

*Terminology* In the literature, the term *elliptic curve* is defined as either a nonsingular cubic curve, or as a nonsingular cubic curve, together with a choice of a flex. We will mostly mean the latter.[1] «

**Definition 7.11** Let $(C, 0_C)$ be an elliptic curve. For $p, q \in C$ we let $p +_C q = (p * q) * 0_C$.

For brevity we usually write $p + q$. Since $p * q = q * p$ we have $p + q = q + p$. As discussed above, for any $p \in C$, $(p * 0_C) * 0_C = p$ and $(p * 0_C) * p = 0_C$. The first means that $p + 0_C = p$. The second, together with the fact that $0_C * 0_C = 0_C$, implies that $((p * 0_C) * p) * 0_C = 0_C$ and so $p + (p * (0_C)) = 0_C$. We thus let $-p = p * 0_C$. So $p + (-p) = 0_C$, and $-(-p) = (p * 0_C) * 0_C = p$.

*Remark 7.12* Let $p, q$ and $r$ be three collinear points on $C$. Then $(p + q) + r = 0_C$, since $r = p * q = -(p + q)$. «

To show that $(C; +, 0_C)$ is an abelian group, it remains to show that the binary operation $+$ is associative. This takes a little bit of work.

### 7.2.1 The Complement Curve

Suppose that $C$, $G$ and $A$ are curves, and $A \cdot G \subset C \cdot G$. Can we find a curve $B$ such that $C \cdot G$ is the multiset sum of $A \cdot G$ and $B \cdot G$? The answer is immediate if $A$ is a component of $C$ (using Proposition 6.29); but interestingly, we can get a positive answer in other cases too.

[1] These definitions are purely algebraic; often in the literature a topological definition is preferred, using the genus of a curve over the complex numbers.

We use linear families of curves. Recall that for $d \geqslant 1$, the space $\mathbb{G}_d$ of all curves of degree $d$ in $\mathbb{P}^2$ is bijective with $\mathbb{P}^k$ (for $k = d(d+3)/2$) via the map $\iota_d$ (see Sect. 4.6). A linear family of curves is the image of a line under $\iota_d$ (Definition 4.39). If $C = V_{\mathbb{P}^2}(f)$ and $D = V_{\mathbb{P}^2}(g)$ are curves of degree $d$ then $\overline{C\,D}$ is the family of curves $V_{\mathbb{P}^2}(ef + ag)$ for $(e:a) \in \mathbb{P}^1$. If $\alpha$ is a change of coordinates of $\mathbb{P}^2$ then the map $C \mapsto \alpha[C]$ is a change of coordinates of $\mathbb{G}_d$ (see Exercise 4.42 for the case $d = 1$). In particular if $E \in \overline{C\,D}$ then $\alpha[E] \in \overline{\alpha[C]\,\alpha[D]}$, which can be verified directly: $\boldsymbol{\alpha}^*(ef + ag) = e\boldsymbol{\alpha}^*(f) + a\boldsymbol{\alpha}^*(g)$.

The following strengthens Exercise 6.62:

**Lemma 7.13** *Let $C$ and $C'$ be curves of the same degree which have no common component. Let $D, D' \in \overline{C\,C'}$ be distinct. Then $D$ and $D'$ have no common component and $C \cdot C' = D \cdot D'$.*

***Proof*** Let $f$ define $C$ and $g$ define $C'$. If $E \in \overline{C\,C'}$ is distinct from $C$ then $E = V_{\mathbb{P}^2}(\lambda f + g)$ for some $\lambda \in \mathbb{K}$. Then $E$ and $C$ have no common component and $C \cdot E = C \cdot C'$: this follows from Proposition 6.38.

Let $D, D' \in \overline{C\,C'}$ be distinct. If one of $D$ or $D'$ is either $C$ or $C'$ then the observation above suffices. Otherwise, applying this observation twice, we get $C \cdot C' = C \cdot D = D' \cdot D$, using the fact that $\overline{C\,C'} = \overline{C\,D}$. □

**Proposition 7.14** *Let $C$ and $G$ be curves of the same degree with no common component; let $A$ be an irreducible curve of smaller degree, which is not a component of $G$. Suppose that:*

*(*) $A \cdot G \subseteq A \cdot D$ for every $D \in \overline{C\,G}$ distinct from $C$.*

*Then $A \cdot G \subset C \cdot G$, and there is a curve $B$ such that*

$$C \cdot G = (A \cdot G) + (B \cdot G).$$

Note that by Bézout's theorem, since $\deg D = \deg G$, if $A$ is not a component of $D$ then the condition $A \cdot G \subseteq A \cdot D$ is in fact equivalent to $A \cdot G = A \cdot D$.

***Proof*** If $A$ is a component of $C$ then $B = C \setminus A$ would do (Proposition 6.29), so we assume otherwise. By Lemma 7.13, it suffices to show that there is some $D \in \overline{C\,G}$ of which $A$ is a component, as we could then take $B = D \setminus A$ and use $D \cdot G = C \cdot G$ (note that $D \neq G$ by assumption that $A$ is not a component of $G$).

We find $D \in \overline{C\,G}$ of which $A$ is a component. By our assumptions that $A$ is a component of neither $C$ nor $G$, since $A$ is irreducible, it follows that $A$ is not a component of $C+G$. By Study's lemma, find $p \in A$, $p \notin C \cup G$. Now there is some $D \in \overline{C\,G}$ which passes through $p$: this follows from Exercise 4.74; for a direct proof in our case, take $f$ defining $C$ and $g$ defining $G$ and let $D = V_{\mathbb{P}^2}(g(\boldsymbol{p})f - f(\boldsymbol{p})g)$ for some presentation $\boldsymbol{p}$ of $p$. Note that $D \neq C, G$ as $p \notin C, G$.

Since $D \neq C$, by the assumption (*), $A \cdot G \subseteq A \cdot D$. But $p \in A \cdot D$ while $p \notin A \cdot G$, so this containment is proper. However $\deg G = \deg D$, so by Bézout's

theorem, it must be the case that $A$ and $D$ have a common component. Since $A$ is irreducible, that component is $A$ itself. □

Suppose that $C$ and $G$ are curves of the same degree with no common component, $A$ is an irreducible curve of smaller degree which is not a component of $G$, and $A \cdot G \subset C \cdot G$. Then for any $D \in \overline{C\,G}$, every $p \in C \cap G$ also lies in $D$ (see Exercise 6.62), implying $\lfloor A \cap G \rfloor \subseteq \lfloor A \cap D \rfloor$. So the condition $A \cdot G \subseteq A \cdot D$ is really about multiplicities of intersection, and it holds if $A \cap G$ contains $\deg A \cdot \deg G$ many distinct points. This shows:

**Corollary 7.15** *Suppose that $C$ and $G$ are curves of the same degree with no common component. Let $A$ be an irreducible curve of smaller degree which is not a component of $G$. Suppose that $A \cap G$ consists of $\deg A \cdot \deg G$ many distinct points, and suppose that $A \cap G \subseteq C \cap G$. Then there is a curve $B$ such that $C \cdot G = (A \cdot G) + (B \cdot G)$.*

Here is an application.

*Example 7.16 (Pascal's "Mystic" Hexagon)* Let $p_1, p_2 \ldots, p_6$ be six distinct points in $\mathbb{P}^2$ which all lie on an irreducible conic $A$. Let $\ell_1 = \overline{p_1\,p_2}$, $\ell_2 = \overline{p_2\,p_3}$, $\ldots$, $\ell_5 = \overline{p_5\,p_6}$, $\ell_6 = \overline{p_6\,p_1}$. No three of the points $p_1, \ldots, p_6$ are collinear, since a line cannot intersect an irreducible conic in more than two points. Hence the lines $\ell_1, \ell_2, \ldots, \ell_6$ are all distinct.

The line segments $\ell_i$ between $p_i$ and $p_{i+1}$ form a hexagon together. Say that $\ell_1$ is *opposite* $\ell_4$, $\ell_2$ is opposite $\ell_5$, and $\ell_3$ is opposite $\ell_6$. Let $q_1, q_2$ and $q_3$ be the points of intersection of the pairs of opposite points: $q_1$ is the intersection of $\ell_1$ with $\ell_4$, etc. Then $q_1, q_2$ and $q_3$ are collinear. See Fig. 7.2.

To see this, let $G = \ell_1 + \ell_3 + \ell_5$ and let $C = \ell_2 + \ell_4 + \ell_6$. Then $G \cdot C$ consists of the nine points of intersection $\ell_i \cap \ell_j$ for $i$ even and $j$ odd. Six of these are the points $p_i$, the set of which is $G \cap A = G \cdot A$. The other three points are the points $q_1$, $q_2$ and $q_3$. The conditions of Corollary 7.15 hold, and so there is a line $\ell$ such that $G \cdot C = (G \cdot A) + (G \cdot \ell)$, so $q_1, q_2$ and $q_3$ lie on the line $\ell$. «

*Remark 7.17* Observe that this conclusion for Pascal's hexagon follows from Exercise 6.67, once we show that $p_1, \ldots, p_6, q_1, q_2, q_3$ are all distinct. «

### 7.2.2 Associativity of the Group Operation

**Lemma 7.18** *Let $G$ be a nonsingular cubic curve and let $L$ be the sum of three lines (not necessarily distinct). Let $\ell$ be a line such that $(G \cdot \ell) \subset (G \cdot L)$. Then there is a conic curve $B$ such that*

$$G \cdot L = (G \cdot \ell) + (G \cdot B).$$

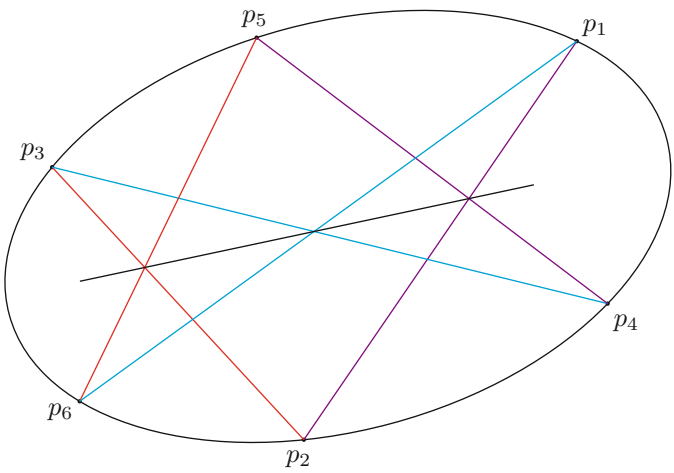



**Fig. 7.2** Pascal's mystic hexagon. "Opposite" sides have the same colours

***Proof*** If $\ell$ is a component of $L$ then we can simply take $B = L \setminus \ell$. We let $L = \ell_1 + \ell_2 + \ell_3$ and assume that $\ell \neq \ell_i$ for $i = 1, 2, 3$. Note that $L$ and $G$ have no common component as $G$ is irreducible (Theorem 6.5).

Proposition 7.14 says that we will be done once we show that for any $D \in \overline{G\,L}$ distinct from $L$, $\ell \cdot G \subseteq \ell \cdot D$. Let $D$ be such a curve. Let $p \in \ell \cap G$; we show that $i_p(G, \ell) \leqslant i_p(D, \ell)$. There are three cases, depending on the value of $i_p(G, \ell)$.

*Case 1:* $i_p(G, \ell) = 1$ In this case we only need to show that $p \in D \cap \ell$, i.e., that $p \in D$. By assumption, $p \in G \cdot L$ so $p \in L$. As we observed before Corollary 7.15, $p \in D$ follows from $p \in G \cap L$ and $D \in \overline{G\,L}$.

*Case 2:* $i_p(G, \ell) = 2$ In this case $\ell$ is the tangent to $G$ at $p$ but $p$ is not a flex of $G$. We need to show that $i_p(D, \ell) \geqslant 2$.

By assumption, $G \cdot \ell \subseteq G \cdot L$, so $i_p(G, L) \geqslant 2$. We have $i_p(G, L) = \sum_{i \leqslant 3} i_p(G, \ell_i)$. Since $\ell_i \neq \ell$ for $i = 1, 2, 3$, no $\ell_i$ is the tangent to $G$ at $p$ and so $i_p(G, \ell_i) \leqslant 1$ and of course equals 1 if and only if $p \in \ell_i$. Hence $p$ belongs to at least two of the lines $\ell_1$, $\ell_2$ and $\ell_3$, and so $p$ is singular on $L$. Since $D \neq L$, for some $\alpha \in \mathbb{K}$, $D = V_{\mathbb{P}^2}(f + \alpha h)$, where $f$ defines $G$ and $h$ defines $L$. Let $\boldsymbol{p}$ be a presentation of $p$. For $u \in \{w, x, y\}$, $D^u(f + \alpha h) = D^u f + \alpha D^u h$; since $D^u h(\boldsymbol{p}) = 0$, $D^u(f + \alpha h)(\boldsymbol{p}) = D^u f(\boldsymbol{p})$; since $p$ is nonsingular on $G$, it is nonsingular on $D$, and further $\ell_p D = \ell_p G = \ell$. Hence $i_p(D, \ell) \geqslant 2$ as required.

*Case 3:* $i_p(G, \ell) = 3$ In this case $\ell$ is the tangent to $G$ at $p$ and $p$ is a flex of $G$. We need to show that $i_p(D, \ell) \geqslant 3$. Again, by assumption, $i_p(G, L) = 3$, which by the argument of the previous case means that $p$ is the point of intersection of all three lines $\ell_1$, $\ell_2$ and $\ell_3$. The same argument shows that $p$ is nonsingular on $D$ and

$\ell_p D = \ell_p G = \ell$. However, $o_p(L) = 3$ so all second derivatives of $h$ are 0 on $\boldsymbol{p}$ as well; linearity again shows that $\ell_p^2 D = \ell_p^2 G$. Recall that $p$ is a flex of $D$ if and only if $\ell_p D \subset \ell_p^2 D$ (and the same holds for $G$, see Proposition 5.36), and so $p$ being a flex of $G$ implies that it is a flex of $D$: $i_p(D, \ell) \geqslant 3$ as required.

□

We are ready to prove the associativity of the group operation. The reason we worked so hard is that the proof that $(p+q)+r = p+(q+r)$ works for any possible constellation of $p$, $q$, $r$ (all distinct, two of them equal, all of them equal, all are different from the identity element $0_C$, some are equal to $0_C$,...); see Exercise 7.32.

**Theorem 7.19** *Let $(C, 0_C)$ be an elliptic curve. The operation $+$ defined above is associative, and so $(C; +, 0_C)$ is an abelian group.*

***Proof*** Let $p, q, r \in C$; we need to show that $p + (q + r) = (p + q) + r$. Let $v = p + q = -(p * q)$ and $u = q + r = -(q * r)$. We need to show that $p + u = -(p * u)$ equals $v + r = -(v * r)$; of course this is the same as showing $p * u = v * r$.

Let

$$L = \overline{p\,q} + \overline{v\,r} + \overline{0_C\,u}.$$

Then

$$C \cdot L = [p, q, -v, v, r, v * r, u, -u, 0_C].$$

Also,

$$C \cdot \overline{q\,r} = [q, r, -u];$$

so $C \cdot \overline{q\,r} \subset C \cdot L$. By Lemma 7.18 there is a conic curve $B$ such that

$$C \cdot B = [p, -v, v, v * r, u, 0_C].$$

Let $\ell = \overline{(-v)\,v}$. We claim that $B \cdot \ell \supseteq [v, -v, 0_C]$; we know that $v, -v, 0_C \in B \cap \ell$, but the question is about multiplicities, if some of these points coincide. Since $C \cdot \ell = [v, -v, 0_C]$ and $[v, -v, 0_C] \subset C \cdot B$, we have $i_s(C, B) \geqslant i_s(C, \ell)$ for all $s \in \{v, -v, 0_C\}$. As $C$ is nonsingular, Exercise 6.44 now states that for all $s \in \{v, -v, 0_C\}$, $i_s(B, \ell) \geqslant i_s(C, \ell)$ which is what is required to show that $B \cdot \ell \supseteq [v, -v, 0_C]$.

By Bézout's theorem, $\ell$ must be a component of $B$, and so $B$ is the sum of $\ell$ and another line $\ell'$; and $C \cdot \ell' = [p, v * r, u]$, from which we conclude that $v * r = p * u$ as required (See Figs. 7.3 and 7.4). □

*Exercise 7.20* Show that for all $p, q, r, s \in C$, $((p * q) * s) * r = p * ((q * r) * s)$. «

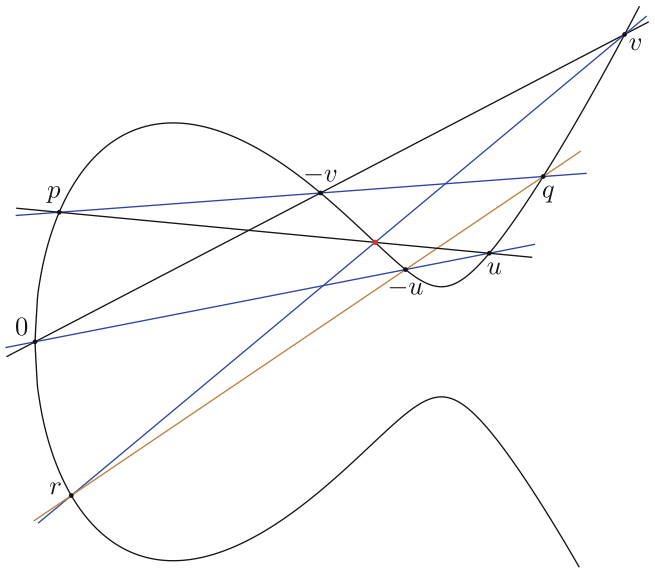



**Fig. 7.3** The operation $+$ is associative. $L$ is the sum of the blue lines. The red point is $p*u = r*v$

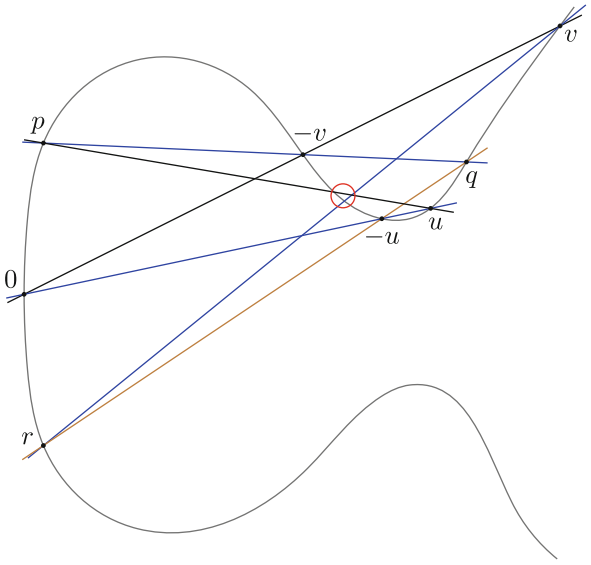



**Fig. 7.4** This curve is not quite cubic; in this case $p * u \neq r * v$

## 7.3 Normal Forms for Nonsingular Cubics

Exercise 5.60 shows that every irreducible conic in $\mathbb{P}^2(\mathbb{C})$ can be mapped by a change of coordinates to the projective closure of the parabola $y = x^2$. In a strong way, this is a *normal form* for irreducible conics. Another such normal form is $x^2 + y^2 + w^2 = 0$ or $x^2 + y^2 = w^2$ (Exercise 4.79). Every irreducible conic is nonsingular, so it is possible to have just one irreducible conic (up to changes of coordinates). This cannot be replicated for cubic curves; in Exercise 5.63 it is shown that there are two non-equivalent singular, irreducible cubic curves (the nodal curve and the cuspidal curve).

In this section we present a couple of normal forms for nonsingular cubic curves. Unlike conics, there are infinitely many pairwise inequivalent nonsingular cubic curves.

We will show that any nonsingular cubic can be transformed by a change of coordinates to the projective closure of the curve

$$y^2 = f(x)$$

where $f \in \mathbb{K}[x]$ has degree 3. Since $\mathbb{K}$ is algebraically closed, $f$ is the product of three linear polynomials.

**Proposition 7.21** *Let $C$ be the projective closure of the curve $y^2 = f$, where $f \in \mathbb{K}[x]$ has degree 3. Then $C$ is nonsingular if and only if the three roots of $f$ in $\mathbb{K}$ are distinct.*

***Proof*** Suppose that $f = \delta(x-\alpha)(x-\beta)(x-\gamma)$, so

$$C = V_{\mathbb{P}^2}\left(y^2 w - \delta(x-\alpha w)(x-\beta w)(x-\gamma w)\right).$$

Suppose that $\alpha$, $\beta$ and $\gamma$ are not all distinct. Without loss of generality $\alpha = \beta$. Then a calculation shows that $(1:\alpha:0)$ is a singular point of $C$.

In the other direction, let $p \in C$ be singular on $C$. Let $g$ be the defining polynomial of $C$ given above. First we argue that $p$ cannot lie on the line at infinity: if $p = (0:a:b)$ then $g(0,a,b) = 0$ implies $a = 0$, But then $b \neq 0$ implies $D^w g(0,0,b) \neq 0$. Hence, we take $p = (1:a:b)$.

Now, $D^y g(1,a,b) = 0$ implies $b = 0$; and then $g(1,a,0) = 0$ implies $a \in \{\alpha, \beta, \gamma\}$. Say $a = \alpha$. Then $D^x g(1,a,0) = 0$ implies $(a-\beta)(a-\gamma) = 0$; whence $\alpha = \beta$ or $\alpha = \gamma$. □

Recall that the Hessian matrix $H_f$ is the $3 \times 3$-matrix of second partial derivatives of $f$.

**Lemma 7.22** *Let $f \in \mathbb{K}[w, x, y]$ be homogeneous of degree $d > 1$. Then*

$$y^2 \det H_f = (d-1)^2 \det \begin{pmatrix} D^{ww} f & D^{wx} f & D^w f \\ D^{xw} f & D^{xx} f & D^x f \\ D^w f & D^x f & \frac{d}{d-1} f \end{pmatrix}.$$

***Proof*** We begin by multiplying the first row of $H_f$ by $w$, the second by $x$, and the third by $y$. Then, we add the first and the second row to the third row, which doesn't change the determinant, and apply Euler's Relation to the polynomials $D^u f$, which are homogeneous of degree $d-1$. We get

$$wxy \det H_f = \begin{vmatrix} wD^{ww} f & wD^{wx} f & wD^{wy} f \\ xD^{xw} f & xD^{xx} f & xD^{xy} f \\ (d-1)D^w f & (d-1)D^x f & (d-1)D^y f \end{vmatrix},$$

so factoring out the scalar $d-1$ and dividing the first row by $w$ and the second by $x$ we get

$$y \det H_f = (d-1) \begin{vmatrix} D^{ww} f & D^{wx} f & D^{wy} f \\ D^{xw} f & D^{xx} f & D^{xy} f \\ D^w f & D^x f & D^y f \end{vmatrix}.$$

We now repeat the trick, this time multiplying the first *column* by $w$, the second by $x$, and the third by $y$; Then we add the first and second columns to the third column and apply Euler's relation to $D^w f$, $D^x f$ and $f$ to get

$$wxy^2 \det H_f = (d-1) \begin{vmatrix} wD^{ww} f & xD^{wx} f & (d-1)D^w f \\ wD^{xw} f & xD^{xx} f & (d-1)D^x f \\ wD^w f & xD^x f & d \cdot f \end{vmatrix}.$$

The final result is obtained by factoring out $d-1$ from the second column, dividing the first column by $w$ and the second by $x$. □

**Proposition 7.23** *Any nonsingular cubic curve can be mapped by a change of coordinates to one which is the projective closure of an affine curve $y^2 = f$, where $f \in \mathbb{K}[x]$ has degree 3.*

*Remark 7.24* Suppose that $C$ is the projective closure of $y^2 = f(x)$ where $f$ is cubic. Then the vertical point at infinity $(0:0:1)$ is the only intersection of $C$ with the line at infinity. Hence, if $C$ is nonsingular, then the vertical point at infinity is a flex of $C$ and the tangent at that point is the line at infinity $w = 0$. «

***Proof of Proposition 7.23*** Let $C$ be a nonsingular cubic curve. By Proposition 7.10, $C$ has a flex. After a change of coordinates, we assume that the vertical point at infinity $(0:0:1)$ is a flex of $C$, and that the tangent to $C$ at that point is the line at infinity $w = 0$. Let $\boldsymbol{p} = (0, 0, 1)$. Picking $g$ defining $C$, we summarise:

- $g(\boldsymbol{p}) = 0$; since $g(\boldsymbol{p})$ is the coefficient of $y^3$ in $g$, we conclude that the monomial $y^3$ does not appear in $g$.
- $D^x g(\boldsymbol{p}) = 0$; looking at the derivative, we see that $D^x g(\boldsymbol{p})$ is the coefficient of $xy^2$ in $g$, and so this monomial does not appear in $g$.
- $D^w g(\boldsymbol{p}) \neq 0$; so the monomial $wy^2$ does appear in $g$.

We also know that $\det H_g(\boldsymbol{p}) = 0$ as $p$ is a flex of $C$. Using Lemma 7.22 we see that

$$0 = 1^2 \det H_g(\boldsymbol{p}) = 4 \det \begin{pmatrix} D^{ww} g(\boldsymbol{p}) & D^{wx} g(\boldsymbol{p}) & D^w g(\boldsymbol{p}) \\ D^{xw} g(\boldsymbol{p}) & D^{xx} g(\boldsymbol{p}) & 0 \\ D^w g(\boldsymbol{p}) & 0 & 0 \end{pmatrix} =$$

$$= -4 \left(D^w g(\boldsymbol{p})\right)^2 D^{xx} g(\boldsymbol{p}),$$

and so (recalling that we assume that $\mathrm{char}(\mathbb{K}) \neq 2$):

- $D^{xx} g(\boldsymbol{p}) = 0$. We conclude that the monomial $x^2 y$ does not appear in $g$.

Overall, we conclude that every monomial which appears in $g$ and which mentions $y$, also mentions $w$. In other words,

$$g = wy(\alpha w + \beta x + \gamma y) + h,$$

where $\alpha, \beta, \gamma \in \mathbb{K}$ with $\gamma \neq 0$, and $h \in \mathbb{K}[w, x]$ homogeneous of degree 3.

Now we change coordinates again. First, by rescaling, we may assume that $\gamma = 1$: use the change of variable $(w, x, y) \mapsto (\gamma w, x, y/\gamma)$. Now, we are searching for a change of variable $\varphi\colon \mathbb{K}[w, x, y] \to \mathbb{K}[w, x, y]$ which would map $g$ to a polynomial of the form $y^2 w + \bar{h}$ for some $\bar{h} \in \mathbb{K}[w, x]$. For simplicity we require that $\varphi(w) = w$ and $\varphi(x) = x$. It is easier to find $\psi = \varphi^{-1}$: so we are looking for a change of variable $\psi$ which maps $w$ to $w$, $x$ to $x$ and $y^2$ to $y(\alpha w + \beta x + y) + \hat{h}$ for some $\hat{h} \in \mathbb{K}[w, x]$. Writing $\psi(y) = aw + bx + cy$, we expand $(aw + bx + cy)^2$ to find that $c = 1$, $a = \alpha/2$ and $b = \beta/2$ are as required. We note that $\psi$ is indeed a change of variable since it is induced by the invertible matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \alpha/2 & \beta/2 & 1 \end{pmatrix}$. Hence, after the change of coordinates which induces $\varphi = \psi^{-1}$, $C$ is defined by the polynomial $wy^2 - \bar{h}$ for some $\bar{h} \in \mathbb{K}[w, x]$. We note that $w$ does not divide $\bar{h}$, as $C$ is irreducible. So $f = \bar{h}^b$ gives the required polynomial. □

*Remark 7.25* In the proof above, we chose any flex $p$ of $C$ and mapped it to the vertical point at infinity. Then we scaled, and applied a change of variable $(w, x, y) \mapsto (w, x, y - \alpha w/2 - \beta x/2)$. The associated change of coordinates fixes

the vertical point at infinity. We conclude that for any nonsingular cubic curve $C$, and every flex $p$ on $C$, there is a change of coordinates which maps $p$ to the vertical point at infinity and the curve $C$ to the projective closure of $y^2 = f(x)$ for some cubic $f$. «

The following is known as *Legendre's normal form*.

**Proposition 7.26** *Any nonsingular cubic curve can be mapped by a change of coordinates to one which is the projective closure of an affine curve*

$$y^2 = x(x-1)(x-\lambda)$$

*where* $\lambda \in \mathbb{K}$, $\lambda \neq 0, 1$.

***Proof*** Let $C$ be a nonsingular cubic. By Proposition 7.23 we can change coordinates so that $C$ is given by the equation $y^2 w = h$ where $h \in \mathbb{K}[w, x]$ is a homogeneous polynomial of degree 3 and $w$ does not divide $h$. So $h = \delta(x-\alpha w)(x-\beta w)(x-\gamma w)$ for some $\alpha, \beta, \gamma, \delta \in \mathbb{K}$ (with $\delta \neq 0$). By Proposition 7.21, $\alpha$, $\beta$ and $\gamma$ are distinct.

We first want to change coordinates so that the roots become 0, 1 and some other root $\lambda$. For this we use a change of variable $\psi$ which maps $y$ to $y$, $w$ to $w$ and $x$ to $ax - cw$ for some $a, c \in \mathbb{K}$. Assuming $a \neq 0$, $\psi$ maps $h$ to

$$\delta(ax-(c+\alpha)w)(ax-(c+\beta)w)(ax-(c+\gamma)w) =$$
$$a^3\delta\left(x-\frac{c+\alpha}{a}w\right)\left(x-\frac{c+\beta}{a}w\right)\left(x-\frac{c+\gamma}{a}w\right)$$

so we get our desired mapping by choosing $c = -\alpha$ and $a = c + \beta = \beta - \alpha$, which is indeed nonzero since $\beta \neq \alpha$. Again this is a legal change of variable since it is induced by the invertible matrix $\begin{pmatrix} 1 & 0 & 0 \\ \alpha & \beta-\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}$. We know that $\lambda = (\gamma-\alpha)/(\beta-\alpha)$ is distinct from 0 and 1 for otherwise $C$ would be singular. Finally, we get rid of the (nonzero) constant $a^3\delta$. Since $\mathbb{K}$ is algebraically closed, $a^3\delta$ has a square root $\sqrt{a^3\delta}$ in $\mathbb{K}$. Replace $y$ by $\sqrt{a^3\delta}\cdot y$; the associated change of coordinates maps the defining equation of $C$ to the desired form. □

Another normal form is one which is associated with Weierstrass. It is given by the affine equation $y^2 = 4x^3 - \alpha x - \beta$ for some $\alpha, \beta \in \mathbb{K}$.

**Lemma 7.27** *Let $C$ be the projective closure of the curve $y^2 = 4x^3 - \alpha x - \beta$. Then $C$ is nonsingular if and only if $\alpha^3 \neq 27\beta^2$.*

***Proof*** By Proposition 7.21, $C$ is nonsingular if and only if the polynomial $f = 4x^3 - \alpha x - \beta$ has three distinct roots in $\mathbb{K}$. Of course $f$ has a repeated root if and only if some irreducible factor of $f$ appears twice in $f$'s irreducible factorisation. By

Proposition 5.7, this happens if and only if the discriminant $\mathrm{disc}(f) = \mathrm{res}_x(f, D^x f)$ is zero. A calculation (see Exercise 5.9) shows that $\mathrm{disc}(f)$ is a constant multiple of $\alpha^3 - 27\beta^2$. □

**Proposition 7.28** *Any nonsingular cubic curve can be mapped by a change of coordinates to one which is the projective closure of an affine curve*

$$y^2 = 4x^3 - \alpha x - \beta$$

*where* $\alpha, \beta \in \mathbb{K}$ *and* $\alpha^3 \neq 27\beta^2$.

***Proof*** By Proposition 7.23, and by scaling $x$, we can assume the equation for $C$ is $y^2w = x^3 + \gamma x^2 w + \varepsilon x w^2 + \delta w^3$ for some $\gamma, \varepsilon, \delta \in \mathbb{K}$. A change of variable $\psi$ which maps $w$ to $w$, $y$ to $y$ and $x$ to $x - aw$ for some $a \in \mathbb{K}$ maps the equation for $C$ to $y^2w = \bar{h}$ where $\bar{h} = (x - aw)^3 + \gamma(x - aw)^2 w + \varepsilon(x - aw)w^2 + \delta w^3$. A calculation shows that the coefficient of $x^2 w$ in $\bar{h}$ is $\gamma - 3a$ so we can easily choose $a$ so that this coefficient is 0. Rescaling $y$ if necessary gives us the desired equation. □

### 7.3.1 Explicit Calculations of the Group Operation

Let the nonsingular cubic curve $C$ be the projective closure of $y^2 = f(x)$. By Remark 7.24 we can choose the vertical point at infinity as the identity element of the group $(C; +, 0_C)$.

**Lemma 7.29** *For every* $p = (a, b) \in C \cap \mathbb{A}^2$, $-p = (a, -b)$.

Here $-p$ is the inverse of $p$ in the group $(C; +, 0_C)$.

***Proof*** Let $\ell = \overline{p\, 0_C}$. It is the vertical line $x = aw$, whose restriction to $\mathbb{A}^2$ is the line $x = a$. $-p = p * o$ is the third point of intersection of this line with $C$. Since $b^2 = f(a)$, we also have $(-b)^2 = f(a)$, that is, $(a, -b) \in C$. If $f(a) \neq 0$ then $\ell \cap C = \{0_C, p, (a, -b)\}$ consists of three distinct points. If $f(a) = 0$ then $-b = b = 0$, and then $\ell \cap C = \{0_C, p\}$; but $\ell$ is not the tangent to $C$ at $0_C$ (the tangent is the line at infinity), so $\ell \cdot C = [0_C, p, p]$. □

*Exercise 7.30* Suppose that $C$ is the projective closure of $y^2 = f(x)$, where $f(x) = a_3x^3 + a_2x^2 + a_1x + a_0$.

(a) Suppose that $\ell$ is a line $y = mx + d$. Let $p_i = (x_i, y_i)$, for $y = 1, 2, 3$, be the three points of intersection of $\ell$ with $C$ (not necessarily distinct). Show that

$$a_3(x_1 + x_2 + x_3) = m^2 - a_2.$$

(Hint: the $x_i$ are the roots of $f - (mx + d)^2$; compare coefficients with $a_3(x - x_1)(x - x_2)(x - x_3)$.)

(b) Show that the tangent to $p = (x, y) \in C$ is vertical if and only if $y = 0$. Calculate rational functions $u$ and $v$ in $\mathbb{K}(x, y)$ such that for all $p = (x, y) \in C$ with $y \neq 0$, $p + p = (u(x, y), v(x, y))$ (Again, addition of points is in the group $(C, 0_C)$).

(c) Calculate rational functions $u$ and $v$ in $\mathbb{K}(x, y, z, w)$ such that for all $p = (x, y)$ and $q = (z, w)$ in $C$ with $p \neq -q, q$, $p + q = (u(x, y, z, w), v(x, y, z, w))$.

(d) How do these calculations simplify when $C$ is in Weierstrass normal form? «

For examples and calculations see Exercise 7.43.

*Exercise 7.31* Suppose that the coefficients $a_0, a_1, a_2, a_3$ in the exercise above are all rational. Let $C(\mathbb{Q}) = C \cap \mathbb{P}^2(\mathbb{Q})$ be the collection *rational points* of $C$: all points $(e : a : b) \in C$ where $e, a$ and $b$ are rational numbers. Show that $C(\mathbb{Q})$ is a subgroup of $C$.[2] «

## 7.4 Further Exercises

7.32 Using the notation of the proof of Theorem 7.19, suppose that the points $0_C$, $p, q, r, u, -u, v, -v, v * r$ are all distinct. Let $L = \overline{p\,q} + \overline{0_C\,u} + \overline{r\,v}$ and $M = \overline{0_C\,v} + \overline{q\,r} + \overline{p\,u}$. Find $L \cap C$ and $M \cap C$; use Exercise 6.66 to show that $p * u = v * r$.

**Flexes**

7.33 Let $f = x^2y^2 + y^2w^2 + w^2x^2$. Find the singular points of $V_{\mathbb{P}^2(\mathbb{C})}(f)$ (See Exercise 6.60.) Show that $H_f = 24 \cdot (9w^2x^2y^2 - (w^2 + x^2 + y^2)f)$. Conclude that $V_{\mathbb{P}^2(\mathbb{C})}(f)$ has no flexes.

7.34 Show that a cuspidal cubic has one flex, and that a nodal cubic has three collinear flexes. (See Exercise 5.63.)

7.35 Let $C = V_{\mathbb{P}^2}(f)$ be a nonsingular cubic curve. Suppose that the origin is a flex of $C$, and that the tangent to $C$ at the origin is the $x$-axis. (a) Show that there is a polynomial $g \in \mathbb{K}[w, x, y]$ such that $g(1, 0, 0) \neq 0$, $D^w g(1, 0, 0) \neq 0$, and $f \sim x^3 + yg$. (b) Show that there are polynomials $h \in \mathbb{K}[w, x]$ and $k \in \mathbb{K}[w, x, y]$ such that $H_f \sim yk + xh$ and $h(1, 0) \neq 0$. (c) Conclude that the origin is nonsingular on $\mathcal{H}_C$ and that the tangent to $\mathcal{H}_C$ at the origin is not the $x$-axis. (d) Conclude that $i_o(C, \mathcal{H}_C) = 1$. (e) Conclude that every nonsingular cubic curve has nine distinct flexes.[3]

[2] A deep theorem of Mordell's says that the group $C(\mathbb{Q})$ is finitely generated.

7.36 (a) Suppose that $C$ is the projective closure of $y^2 = f(x)$ for some cubic $f$ and is nonsingular. Show that the vertical point at infinity is nonsingular on $\mathcal{H}_C$ and that the tangent to $\mathcal{H}_C$ at that point is not the line at infinity. (b) Use Remark 7.25 to give another proof of the fact that a nonsingular cubic curve has nine distinct flexes.

**Generalisations**

7.37 Let $C$ be a nonsingular cubic curve. Show that Definition 7.11 yields a group operation $+_C$ (with identity element $0_C$) even if $0_C$ is not a flex of $C$.[4]

Let $0_C, 0'_C \in C$; let $a = 0_C * 0'_C$. Show that the map $p \mapsto a * p$ is an isomorphism of the groups $(C, 0_C)$ and $(C, 0'_C)$. (Hint: apply Exercise 7.20, and recall that $*$ is commutative. For more details see, for example, [Gib98, Lem.17.4].)

7.38 Let $C$ be a nonsingular cubic curve, and let $\alpha$ be a change of coordinates of $\mathbb{P}^2$. Let $0_C \in C$. Show that $\alpha$ yields an isomorphism between the group $(C, 0_C)$ and the group $(\alpha[C], \alpha(0_C))$.

7.39 Suppose that $C$ is an irreducible singular cubic. Show that we can define $p * q$ for nonsingular $p, q \in C$, and that the result is also nonsingular on $C$. Conclude that the definition made in Exercise 7.37 gives a group operation on the collection of nonsingular points of $C$.

Let $D$ be the affine curve $y = x^3$. Show that the projective closure of $D$ is a cuspidal cubic, and $D$ is the collection of nonsingular points of that closure. Show that $t \mapsto (t, t^3)$ is a parameterisation of $D$ and is a group isomorphism between the additive group $(\mathbb{K}, +)$ and the group on $D$ obtained by choosing $0_C$ to be the origin. Conclude that the group defined on the nonsingular points of any cuspidal cubic is isomorphic to $(\mathbb{K}, +)$.[5]

**The Order of Group Elements**

Let $(C, 0_C)$ be an elliptic curve (with $0_C$ a flex of $C$). Recall that the *order* $o_C(p)$ in the group $C$ of an element $p \in C$ is the size of the cyclic subgroup of $C$ generated by $p$ (see page 39). This is not to be confused with $o_p(C)$, which is always 1, as $C$ is nonsingular.

7.40 Let $p \in C$, $p \neq 0_C$. (a) Show that $o_C(p) = 2$ if and only if the tangent to $C$ at $p$ passes through $0_C$. (b) Show that $o_C(p) = 3$ if and only if $p$ is a flex of $C$.

[3] For a generalisation see [BK86, Thm.7.3.1], which also gives an alternative proof of Corollary 7.7.

[4] If $0_C$ is not a flex then it is not necessarily the case though that $-p = p * 0_C$, and Remark 7.12 may fail.

[5] With a little more work, show that the group on the nonsingular points of a nodal cubic curve is isomorphic to the multiplicative group $(\mathbb{K}^*, \cdot)$.

7.41 Suppose that $C$ is the projective closure of $y^2 = f(x)$ for cubic $f$ and is nonsingular, and let $0_C = (0:0:1)$. (a) Find the points $p \in C$ which have order 2 in the group $C$; show that they are collinear. (b) Conclude that the collection of points of order 1 or 2 in $(C, 0_C)$ form a subgroup of $(C, 0_C)$, isomorphic to $C_2 \times C_2$ (recall that $C_2$ is the cyclic group of size 2). (c) Show that the previous conclusion holds for any elliptic curve $(D, 0_D)$. (Hint: use Exercise 7.37.)

7.42 Let $(C, 0_C)$ be an elliptic curve. Show that a point $p \in C$ has order 6 in the group $(C, 0_C)$ if and only if $p$ is not a flex, but $\ell_p C$ passes through a flex of $C$ other than $0_C$.

7.43 For each of the following affine equations and points $p$ in $\mathbb{A}^2(\mathbb{C})$, show that the (projective closure of the) equation defines a nonsingular cubic curve $C$; that $0_C = (0:0:1)$ is a flex of $C$, and find the order of $p$ in the group $(C, 0_C)$:

(i) $y^2 = x^3 + 4x$, $p = (2, 4)$.
(ii) $y^2 = x^3 + 1$, $p = (2, 3)$
(iii) $y^2 + 7xy = x^3 + 16x$, $p = (2, 2)$.

For more such calculations see [Bix06, Ex.9.1] and [Gib98, Sec.17.4].

7.44 Let $(C, 0_C)$ be an elliptic curve, and let $B$ be a conic curve. Suppose that $C$ and $B$ intersect at six distinct points $p_1, p_2, \ldots, p_6$. (a) Show that $p_1 * p_2$, $p_3 * p_4$, and $p_5 * p_6$ are collinear. (b) Show that in the group $(C, 0_C)$, $p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = 0_C$.[6]

**The Nine Point Configuration**

A *nine point configuration* is a collection of nine points such that any line passing through two points from the collection also passes through a third (but no others).

7.45 Show that the points in the affine plane $\mathbb{A}^2(\mathbb{Z}/(3))$ over the field of three elements form a nine point configuration.

7.46 Let $C$ be a nonsingular cubic curve. (a) Let $p$ and $q$ be distinct flexes of $C$. Show that $p * q$ is also a flex of $C$. (b) Show that the nine flexes of $C$ (Exercises 7.35 and 7.36) form a nine point configuration. (c) Show that the collection of flexes of $C$ forms a subgroup of $(C, 0_C)$ (for any choice of flex $0_C$ of $C$), which is isomorphic to $C_3 \times C_3$.

7.47 Let $C$ be a nonsingular cubic curve. Let $p$, $r$, $s$ and $t$ be distinct points on $C$. Suppose that $r$, $s$ and $t$ are collinear. Further, suppose that $p$ lies on the tangents to $C$ at $r$, at $s$ and at $t$. (a) Show that $p$ is a flex of $C$. (b) Show that if $q \neq p$ is a point on $C$ and $p$ lies on the tangent to $C$ at $q$, then $q$ is one of $r$, $s$ and $t$.

[6] In fact, the result holds even when the points are not all distinct. See for example [Bix06, Ex.10.6].

7.48 (a) Find the unique nine-point configuration in $\mathbb{A}^2$ which contains the points $(1,1)$, $(1,-1)$, $(-1,1)$ and $(-1,-1)$. (Hint: it must also contain the origin.) (b) Let $A$ and $B$ be nine-point configurations in $\mathbb{P}^2$. Show that there is a change of coordinates $\alpha$ of $\mathbb{P}^2$ which maps $A$ to $B$. (c) Show that in $\mathbb{P}^2(\mathbb{R})$ there is no nine-point configuration.

7.49 Let $C$ be a nonsingular cubic curve and let $p$ and $q$ be flexes of $C$. Show that there is a change of coordinates of $\mathbb{P}^2$ which maps $p$ to $q$ and $C$ to itself. (Hint: use Remark 7.25 and Exercise 7.46, and note that $(x, y) \mapsto (-x, y)$ maps $y^2 = f(x)$ to itself.)

7.50 The *Hesse normal form* of a cubic curve is an equation $w^3+x^3+y^3 = 3\lambda wxy$ for some $\lambda \in \mathbb{K}$. (a) Show that the cubic defined by the Hesse normal form with parameter $\lambda$ is singular if and only if $\lambda^3 = -1$. (b) Find the flexes of a nonsingular cubic given in Hesse normal form. (c) Show that any cubic curve which passes through all the flexes of a cubic in Hesse normal form is already in Hesse normal form. (d) Show that every nonsingular cubic curve can be transformed to have Hesse normal form. (Use Exercises 7.46 and 7.48.)

**Pascal's Hexagon**

7.51 Following the notation of Example 7.16, let $p_1, p_2, \ldots, p_6$ be six distinct points in $\mathbb{P}^2$, no three of which are collinear; and let $\ell_1 = \overline{p_1\, p_2}$, $\ell_2 = \overline{p_1\, p_2}$, $\ldots$, $\ell_6 = \overline{p_6\, p_1}$; and for $j = 1, 2, 3$ let $q_j$ be the point of intersection of $\ell_j$ and $\ell_{j+3}$. Prove that $p_1, p_2, \ldots, p_6$ lie on a conic *if and only if* $q_1, q_2$ and $q_3$ are collinear.

7.52 Let $C$ and $D$ be curves of the same degree $d$. Let $p \in C \cap D$, and suppose that $p$ is singular on both $C$ and $D$. Show that $p$ is singular on every curve $E \in \overline{C\, D}$. (Note Exercise 5.44.)

7.53 Let $A$ be an irreducible conic curve, and let $p_1, p_2, p_3$ be distinct points on $A$. For $i = 1, 2, 3$, let $\ell_i$ be the tangent to $A$ at $p_i$, and let $k_i$ be the line connecting the other two points, e.g. $k_1 = \overline{p_2\, p_3}$. (a) Show that the lines $\ell_1, \ell_2, \ell_3, k_1, k_2, k_3$ are all distinct. (b) Let $q_i$ be the point of intersection of $\ell_i$ and $k_i$. Show that the points $q_1$, $q_2$ and $q_3$ are collinear. (One way to do it is to use Proposition 7.14, Exercise 7.52, and Theorem 6.42.)

7.54 Let $A$ be an irreducible conic curve, and let $r, p_1, p_2, p_3, p_4$ be five distinct points on $A$. Let $\ell^*$ be the tangent to $A$ at $r$; let $\ell_1 = \overline{p_1\, p_2}$, $\ell_2 = \overline{p_2\, p_3}$, $\ell_3 = \overline{p_3\, p_4}$, $k_1 = \overline{p_4\, r}$ and $k_2 = \overline{r\, p_4}$. Let $q_1 = \ell_1 \cap k_1$, $q_2 = \ell_2 \cap \ell^*$, and $q_3 = \ell_3 \cap k_2$. (a) Show that $q_1, q_2$ and $q_3$ are collinear. (b) Let $C$ be an irreducible conic curve in $\mathbb{A}^2(\mathbb{R})$, and let $p \in C$. Show how to draw the tangent to $C$ at $p$ if all you have is a straight-edge ruler [Kir92, Ex.3.7].

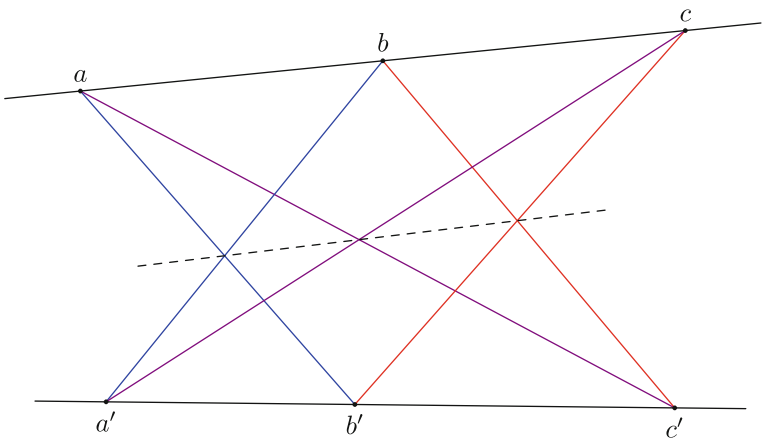



**Fig. 7.5** Pappus' theorem

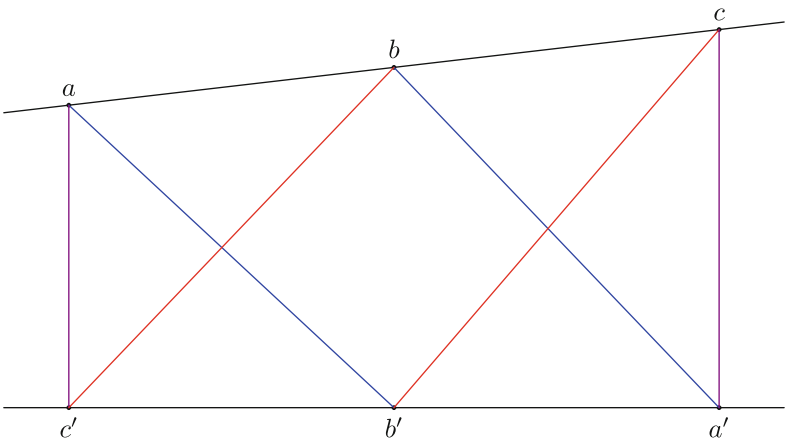



**Fig. 7.6** Pappus' little theorem

7.55 If in Pascal's hexagon we replace the irreducible conic $A$ by a pair of lines, and require that the points $p_1, \ldots, p_6$ avoid the point of intersection of these two lines, then we obtain *Pappus' hexagon theorem* (Remark 4.46); see Fig. 7.5.

(a) Show that Pappus' theorem can be proved by the same technique as Pascal's, using a modification of Proposition 7.14. (The main use of irreducibility of $A$ was at the end of the proof. Now we can only conclude that one of the lines comprising $A$ is a component of $D$. Conclude though that the other line must be a component of the conic $D \setminus A$.)

(b) Suppose that the six points $a, \ldots, c'$ lie in $\mathbb{A}^2$, and that $\overline{a\,b'}$ is parallel to $\overline{a'\,b}$ and that $\overline{a\,c'}$ is parallel to $\overline{a'\,c}$. Conclude that $\overline{b\,c'}$ and $\overline{b'\,c}$ are also parallel (see Fig. 7.6).

# Part II

# Riemann Surfaces


# Quasi-Euclidean Spaces

# 8

This chapter is about topology. The general definition of a topological space is quite abstract. We are mostly concerned with surfaces, which are 2-dimensional manifolds. In general, an $n$-manifold is an object which locally looks like $\mathbb{R}^n$: a space on which we can locally assign coordinates from Euclidean space. To study manifolds, we will also need to consider some of their subsets. And so we give the somewhat non-standard definition of a *quasi-Euclidean space* (Definition 8.48): a subset of a manifold, equipped with its topology (the collection of open sets).

Since we do not define topological spaces axiomatically, our definition of the topology of manifolds is again somewhat nonstandard. We first review the topology of $\mathbb{R}^n$. We then define a chart on a set $M$ to be a bijection between a subset of $M$ and some open set in $\mathbb{R}^n$. Two charts $\varphi$ and $\psi$ are said to be compatible if the associated transition function $\varphi \circ \psi^{-1}$ is a bijection between open subsets of $\mathbb{R}^n$ which is continuous in both directions. An atlas for $M$ is a collection of pairwise compatible charts whose domains cover the set $M$. Given an atlas for $M$, we can define open neighbourhoods in $M$ in terms of pull-backs of neighbourhoods under charts. This approach resembles the standard definition of differentiability in manifolds, which we will discuss in the next chapter.

Having given the definition, we review several topological notions: first and second countability, the Hausdorff property, and most importantly, compactness. We characterise compactness in terms of sequential compactness (convergence of subsequences). Our treatment is fairly standard. We also discuss the completeness of the real numbers.

Finally, we introduce quotients of $\mathbb{R}^n$ by discrete subgroups, one of which is the torus.



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_8

## 8.1 Topology of $\mathbb{R}^n$

We quickly review the topology of $\mathbb{R}^n$, mostly in a sequence of exercises.

Let $n \geqslant 1$. For a point $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ let $|\boldsymbol{a}| = \sqrt{a_1^2 + \cdots + a_n^2}$. For two points $\boldsymbol{a} = (a_1, \ldots, a_n)$ and $\boldsymbol{b} = (b_1, \ldots, b_n) \in \mathbb{R}^n$, the *Euclidean distance* between $\boldsymbol{a}$ and $\boldsymbol{b}$ is

$$d(\boldsymbol{a}, \boldsymbol{b}) = |\boldsymbol{b} - \boldsymbol{a}| = \sqrt{(b_1 - a_1)^2 + \cdots + (b_n - a_n)^2}.$$

A key fact is that Euclidean distance satisfies the *triangle inequality*: for all $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^n$, $d(\boldsymbol{a}, \boldsymbol{c}) \leqslant d(\boldsymbol{a}, \boldsymbol{b}) + d(\boldsymbol{b}, \boldsymbol{c})$ (see Exercise 8.118). Also, $d(\boldsymbol{a}, \boldsymbol{b}) = d(\boldsymbol{b}, \boldsymbol{a})$ and $d(\boldsymbol{a}, \boldsymbol{b}) = 0$ if and only if $\boldsymbol{a} = \boldsymbol{b}$. In modern terminology we say that $d$ is a *metric* on $\mathbb{R}^n$.

Let $\boldsymbol{a} \in \mathbb{R}^n$ and let $r > 0$ be a real number. The *open ball* with centre $\boldsymbol{a}$ and radius $r$ is

$$B(\boldsymbol{a}, r) = \{\boldsymbol{b} \in \mathbb{R}^n \,:\, d(\boldsymbol{a}, \boldsymbol{b}) < r\}.$$

A subset of $\mathbb{R}^n$ is *open* if it is the union of (possibly infinitely many) open balls. A subset $X \subseteq \mathbb{R}^n$ is a *neighbourhood* of a point $\boldsymbol{a} \in \mathbb{R}^n$ if there is some $r > 0$ such that $B(\boldsymbol{a}, r) \subseteq X$.

*Exercise 8.1* (a) Show that every open ball $B$ is a neighbourhood of every point in $B$. (b) Show that if $X \subseteq Y \subseteq \mathbb{R}^n$ and $X$ is a neighbourhood of point $a \in \mathbb{R}^n$, $Y$. (c) Show that a set $X \subseteq \mathbb{R}^n$ is a neighbourhood of a point $\boldsymbol{a} \in X$ if and only if there is an open set $U \subseteq X$ such that $\boldsymbol{a} \in U$. (d) Show that a subset $U \subseteq \mathbb{R}^n$ is open if and only if it is a neighbourhood of every point $\boldsymbol{a} \in U$. «

*Exercise 8.2* (a) Show that the intersection of finitely many neighbourhoods of a point $\boldsymbol{a}$ is a neighbourhood of $\boldsymbol{a}$. (b) Show that the empty set and $\mathbb{R}^n$ are open subsets of $\mathbb{R}^n$. (c) Show that the union of any number (possibly infinite) of open subsets of $\mathbb{R}^n$ is open. (d) Show that the intersection of finitely many open subsets of $\mathbb{R}^n$ is open. (e) Give an example of a countable family of open subsets of $\mathbb{R}$ whose intersection is not open. «

*Exercise 8.3* An open ball $B(\boldsymbol{a}, r)$ is *rational* if the coordinates of its centre are rational and its radius is rational: $\boldsymbol{a} \in \mathbb{Q}^n$ and $r \in \mathbb{Q}$. Show that every open subset of $\mathbb{R}^n$ is the union of rational open balls. Show that there are only countably many rational open balls (whereas there are uncountably many open balls). «

*Exercise 8.4* Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be distinct points in $\mathbb{R}^n$. Show that there are neighbourhoods $X$ of $\boldsymbol{a}$ and $Y$ of $\boldsymbol{b}$ which are disjoint.[1] «

Let $n, m \geqslant 1$ and let $U \subseteq \mathbb{R}^n$ be open. A function $f : U \to \mathbb{R}^m$ is *continuous at a point* $\boldsymbol{a} \in U$ if for every neighbourhood $Y \subseteq \mathbb{R}^m$ of $f(\boldsymbol{a})$, $f^{-1}[Y]$ is a neighbourhood of $\boldsymbol{a}$.

*Exercise 8.5* Let $U \subseteq \mathbb{R}^n$ be open. Show that a function $f : U \to \mathbb{R}^m$ is continuous at a point $\boldsymbol{a} \in U$ if and only if for every $\varepsilon > 0$ there is some $\delta > 0$ such that for all $\boldsymbol{x} \in \mathbb{R}^n$, if $d(\boldsymbol{a}, \boldsymbol{x}) < \delta$ then $\boldsymbol{x} \in U$ and $d(f(\boldsymbol{a}), f(\boldsymbol{x})) < \varepsilon$. «

A function $f : U \to \mathbb{R}^m$ is called *continuous* if it is continuous at every point in $U$.

*Exercise 8.6* Let $U \subseteq \mathbb{R}^n$ be open. Show that a function $f : U \to \mathbb{R}^m$ is continuous if and only if for every open subset $V \subseteq \mathbb{R}^m$, $f^{-1}[V]$ is open. «

*Remark 8.7* Recall that we allow partial compositions of functions. If $f$ and $g$ are functions, then $(g \circ f)(x)$ is defined if $f(x)$ is defined (i.e. $x \in \operatorname{dom} f$) and $g(f(x))$ is defined ($f(x) \in \operatorname{dom} g$). In other words, we define $g \circ f$ even if the image of $f$ is not a subset of the domain of $g$, but this means that the domain of $g \circ f$ may be smaller than the domain of $f$.

Similarly, we allow partial pointwise images of sets by functions: if $f$ is a function and $Y$ is a set then $f[Y] = \{f(x) : x \in Y\}$ is defined even if $Y$ is not a subset of the domain of $f$; of course $f[Y] = f[Y \cap \operatorname{dom} f]$. «

*Exercise 8.8* Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^m$ be open, and let $f : U \to \mathbb{R}^m$ and $g : V \to \mathbb{R}^k$ be continuous. Show that the domain of $g \circ f$ is open, and that $g \circ f$ is continuous. «

*Exercise 8.9* Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}^m$. We write $f = (f_1, \ldots, f_m)$ where $f_i : U \to \mathbb{R}$. Show that $f$ is continuous if and only if each $f_i$ is continuous. «

*Exercise 8.10* Let $m, n \geqslant 1$. We identify $\mathbb{R}^{n+m}$ with $\mathbb{R}^n \times \mathbb{R}^m$. (a) Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^m$. Show that $U \times V \subseteq \mathbb{R}^{n+m}$ is open if and only if both $U$ and $V$ are open. (b) Assuming both are open, let $\bar{n}, \bar{m} \geqslant 1$, and let $f : U \to \mathbb{R}^{\bar{n}}$ and $g : V \to \mathbb{R}^{\bar{m}}$. Show that the function $(f \times g) : (U \times V) \to \mathbb{R}^{\bar{n}+\bar{m}}$, defined by $(f \times g)(\boldsymbol{x}, \boldsymbol{y}) = (f(\boldsymbol{x}), g(\boldsymbol{y}))$, is continuous if and only if both $f$ and $g$ are continuous. «

*Exercise 8.11* (a) Show that addition and multiplication are continuous (as functions from $\mathbb{R}^2$ to $\mathbb{R}$). (b) Show that the function $x \mapsto 1/x$ is continuous on $\mathbb{R} \setminus \{0\}$.

[1] While this is an easy exercise, this property will become important when we discuss manifolds.

(c) Let $U \subseteq \mathbb{R}$ be open; let $f, g : U \to \mathbb{R}$ be continuous. Show that $f + g$ and $fg$ are continuous. (Make use of Exercise 8.9.) (d) Show that $f/g$ (defined on those points $\boldsymbol{x} \in U$ for which $g(\boldsymbol{x}) \neq 0$) has open domain and is continuous. «

*Exercise 8.12* Continuity is sometimes defined in terms of limits; this can be reversed. Let $U \subseteq \mathbb{R}^n$ be open, let $\boldsymbol{a} \in U$, and let $f : U \setminus \{\boldsymbol{a}\} \to \mathbb{R}^m$ be a function. We say that $\lim_{\boldsymbol{x} \to \boldsymbol{a}} f(\boldsymbol{x}) = \boldsymbol{b}$ if the function $\hat{f}$ extending $f$ by defining $\hat{f}(\boldsymbol{a}) = \boldsymbol{b}$ is continuous at $\boldsymbol{a}$. Show that $\lim_{\boldsymbol{x} \to \boldsymbol{a}} f(\boldsymbol{x}) = \boldsymbol{b}$ if and only if for every $\varepsilon > 0$ there is some $\delta > 0$ such that for all $\boldsymbol{x} \in U \setminus \{\boldsymbol{a}\}$, if $d(\boldsymbol{a}, \boldsymbol{x}) < \delta$ then $d(f(\boldsymbol{x}), \boldsymbol{b}) < \varepsilon$. «

*Exercise 8.13* Let $A \in M_{m \times n}(\mathbb{R})$; let $M = \max_{i,j} |a_{i,j}|$ be a bound on the absolute values of the entries of $A$. (a) Show that for all $\boldsymbol{u} \in \mathbb{R}^n$, if for all $i$, $|u_i| \leqslant \varepsilon$, then $|A\boldsymbol{u}| \leqslant m\sqrt{n}M\varepsilon$. (b) Conclude that every linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ is continuous. «

*Exercise 8.14* Let $n \geqslant 1$. Show that the absolute value function $\boldsymbol{x} \mapsto |\boldsymbol{x}|$ from $\mathbb{R}^n$ to $\mathbb{R}$ is continuous. «

## 8.2 Manifolds

Let $M$ be a set, and fix $n \geqslant 1$.

**Definition 8.15** A *chart* for $M$ is a bijection between a subset of $M$ and an open subset of $\mathbb{R}^n$.

Intuitively we think of a chart $\psi : Y \to U$ (where $Y \subseteq M$ and $U \subseteq \mathbb{R}^n$ is open) as an assignment of coordinates to the points in $Y$. A point $y \in Y$ is assigned the coordinates of the point $\psi(y)$.

Suppose that $\psi : Y \to U$ and $\varphi : Z \to V$ are two charts for $M$. Every point $y$ of the intersection $Y \cap Z$ is given two coordinates: one by $\psi$ and one by $\varphi$. The function $\varphi \circ \psi^{-1}$, which is a bijection from $\psi[Y \cap Z]$ (the image of the set $Y \cap Z$ under the function $\psi$) to $\varphi[Y \cap Z]$, gives the translation between one set of coordinates and the other. See Fig. 8.1. This function is called the *transition map* between the two charts. (Note that we are using the partial composition, see Remark 8.7.)

**Definition 8.16** Two charts $\psi$ and $\varphi$ for $M$ are *compatible* if the domain and range of the transition map $\varphi \circ \psi^{-1}$ are open subsets of $\mathbb{R}^n$, and both the transition map and its inverse are continuous.

As discussed in the overview chapter, this notion of compatibility is a topological one; we will later encounter more stringent conditions for compatibility of charts.

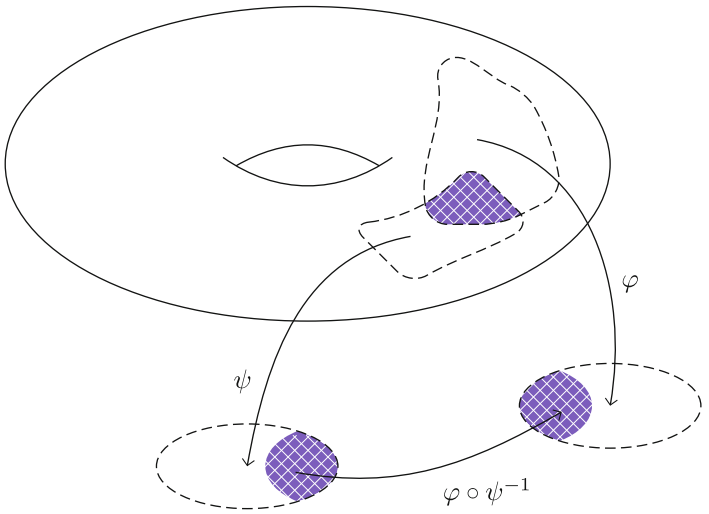



**Fig. 8.1** The transition map between the two charts $\psi$ and $\varphi$

**Definition 8.17** An *atlas* for a set $M$ is a collection $\mathcal{A}$ of charts for $M$ which are pairwise compatible, such that every point of $M$ lies in the domain of at least one chart in $\mathcal{A}$.

### Examples

If $U \subseteq \mathbb{R}^n$ is open then the identity map $\mathrm{id}_U$ on $U$ is a chart for $U$, and the atlas $\{\mathrm{id}_U\}$ containing only this identity map is an atlas for $U$. For all $n \geqslant 1$, $\mathbb{A}^n(\mathbb{R})$ (as a set) is $\mathbb{R}^n$ and so the atlas containing only $\mathrm{id}_{\mathbb{A}^n(\mathbb{R})}$ is an atlas for $\mathbb{A}^n(\mathbb{R})$. The complex numbers $\mathbb{C}$ can be identified with $\mathbb{R}^2$. More precisely, define a bijection $\psi : \mathbb{C} \to \mathbb{R}^2$ by letting $\psi(a + ib) = (a, b)$. Then $\psi$ is a chart for $\mathbb{C}$ and the atlas containing only $\psi$ is an atlas for $\mathbb{C}$. Similarly, $\mathbb{A}^n(\mathbb{C})$ is identified with $\mathbb{R}^{2n}$. And similarly, the collection of matrices $M_{n\times m}(\mathbb{R})$ has a global chart by identifying it with $\mathbb{R}^{nm}$.

*Example 8.18* We define an atlas for the unit circle $S \subseteq \mathbb{R}^2$. Consider the following four charts. The chart $\psi_+$ is defined on $\{(a, b) \in S : a > 0\}$ by letting $\psi_+(a, b) = b$. The chart $\psi_-$ is defined on $\{(a, b) \in S : a < 0\}$ by letting $\psi_-(a, b) = b$. The chart $\varphi_+$ is defined on $\{(a, b) \in S : b > 0\}$ by letting $\varphi_+(a, b) = a$. Similarly define $\varphi_-$.

Every point in the unit circle lies in the domain of at least one of these four charts. Most of them lie in two; but four points in $S$ lie in only one. The range of each of these charts is the interval $(-1, 1)$ which is an open subset of $\mathbb{R}$. The charts are 1–1. To check compatibility consider for example $\psi_+$ and $\varphi_+$. Let $Z = \mathrm{dom}\,\psi_+ \cap \mathrm{dom}\,\varphi_+$. Then $\psi_+[Z] = \varphi_+[Z] = (0, 1)$ and the transition map $\varphi_+ \circ \psi_+^{-1}$ is the map sending $c \in (0, 1)$ to $\sqrt{1 - c^2}$, which is continuous. (See Exercises 8.73 and 8.120 below.) «

*Example 8.19* Let $n \geqslant 1$. Recall the *affine cover* of $\mathbb{P}^n(\mathbb{R})$: the maps $\rho_0, \rho_1, \ldots, \rho_n$ defined by $\rho_i(a_1, a_2, \ldots, a_n) = (a_1 : a_2 : \cdots : a_{i-1} : 1 : a_i : \cdots : a_n)$. They are bijections between $\mathbb{A}^n(\mathbb{R}) = \mathbb{R}^n$ and the subset $U_i = \mathbb{P}^n(\mathbb{R}) \setminus V_{\mathbb{P}^n(\mathbb{R})}(x_i)$ (the collection of points $(a_0 : \cdots : a_n)$ with $a_i \neq 0$). The collection $\{\rho_0^{-1}, \rho_1^{-1}, \ldots, \rho_n^{-1}\}$ is an atlas for $\mathbb{P}^n(\mathbb{R})$.

To check compatibility, for ease of notation we consider $\rho_0$ and $\rho_n$. Let $Z = U_0 \cap U_n$. Then $\rho_0^{-1}[Z] = \mathbb{R}^n \setminus V_{\mathbb{A}^n(\mathbb{R})}(x_n)$ (the collection of $(a_1, \ldots, a_n)$ with $a_n \neq 0$) and $\rho_n^{-1}[Z] = \mathbb{R}^n \setminus V_{\mathbb{A}^n(\mathbb{R})}(x_0)$, both of which are open in $\mathbb{R}^n$. (See Exercise 8.73). For $\boldsymbol{a} = (a_1, \ldots, a_n) \in \rho_0^{-1}[Z]$ we have

$$\rho_0(\boldsymbol{a}) = (1 : a_1 : a_2 : \cdots : a_{n-1} : a_n) = \left(\frac{1}{a_n} : \frac{a_1}{a_n} : \frac{a_2}{a_n} : \cdots : \frac{a_{n-1}}{a_n} : 1\right)$$

and so

$$(\rho_n^{-1} \circ \rho_0)(a_1, \ldots, a_n) = \left(\frac{1}{a_n}, \frac{a_1}{a_n}, \ldots, \frac{a_{n-1}}{a_n}\right)$$

which is continuous. «

*Example 8.20* Let $n \geqslant 1$. The affine cover of $\mathbb{P}^n(\mathbb{C})$ gives an atlas for $\mathbb{P}^n(\mathbb{C})$. The analysis is the same as in Example 8.19, except that the transition maps (for example $(a_1, \ldots, a_n) \mapsto (1/a_n, a_1/a_n, \ldots, a_{n-1}/a_n)$) are defined on tuples of complex numbers (and their domains are identified with subsets of $\mathbb{R}^{2n}$). Showing these maps are continuous boils down to showing that: (i) The map $z \mapsto 1/z$ defined on $\mathbb{C} \setminus \{0\}$ is continuous when viewed as a map from $\mathbb{R}^2 \setminus \{(0,0)\}$ to $\mathbb{R}^2$; this is the map $(a, b) \mapsto (a/(a^2 + b^2), -b/(a^2 + b^2))$. (ii) The map $(z, w) \mapsto zw$ defined on $\mathbb{C}^2$ is continuous when viewed as a map from $\mathbb{R}^4 \to \mathbb{R}^2$; this is the map $(a, b, c, d) \mapsto (ac - bd, ad + bc)$. «

*Example 8.21* We extend the example given in the introduction (see Fig. 1.1) by one dimension. Let $S^2$ be the unit sphere in $\mathbb{R}^3$ (the collection of $\boldsymbol{p} \in \mathbb{R}^3$ such that $|\boldsymbol{p}| = 1$). Define two projections on the sphere. Namely let $\boldsymbol{p}_+ = (0, 0, 1)$ be the north pole of the sphere and let $\boldsymbol{p}_- = (0, 0, -1)$ be the south pole. For $\boldsymbol{q} \in S^2 \setminus \{\boldsymbol{p}_+\}$ let $\sigma_+(\boldsymbol{q})$ be the point of intersection of the line $\overline{\boldsymbol{q}\,\boldsymbol{p}_+}$ with the plane $z = 0$; similarly define $\sigma_-$.

Identifying the plane $z = 0$ in $\mathbb{R}^3$ with the complex plane (via $(x, y, 0) \mapsto x + iy$), show that for $\boldsymbol{q} \in S^2 \setminus \{\boldsymbol{p}_+, \boldsymbol{p}_-\}$, $\sigma_+(\boldsymbol{q}) \cdot \overline{\sigma_-(\boldsymbol{q})} = 1$, where $\overline{z}$ denotes complex conjugation.[2] Conclude that $\{\sigma_+, \overline{\sigma_-}\}$ is an atlas for $S^2$ (the transition map is $z \mapsto 1/z$). «

[2] Recall that the conjugate of $a + ib$ is $a - ib$.

*Remark 8.22* In general, when specifying a collection of elements, many texts also assume that these elements are also *indexed* by some *index set*: $X = \{x_i : i \in I\}$. Set-theoretically, this means that on top of specifying the set $X$ (what are its elements), we also associate with it a function from some set $I$ of "indices" onto $X$, often injective ($x_i \neq x_j$ if $i \neq j$). If the index set $I$ is ordered in some way (for example if $I = \mathbb{N}$ is the set of natural numbers), then this indexing induces an ordering on the elements of the set $X$.

While this may be extremely familiar, it is often unnecessary; we will often just specify the elements of a set. For example, when specifying the atlas $\{\varphi_+, \varphi_-, \psi_+, \psi_-\}$ for the unit circle in Example 8.18, we didn't need to number the four charts in some way. However, sometimes this notation is useful. For example, if an atlas is given as an indexed collection of charts $\mathcal{A} = \{\psi_i : i \in I\}$, then we can use the shorthand $\psi_{i,j}$ for the transition function $\psi_j \circ \psi_i^{-1}$ from $\psi_i$-coordinates to $\psi_j$-coordinates. We will use this later in the book. «

### 8.2.1 Topology of Pre-manifolds

The local coordinates allow us to put a topological structure on a set with an atlas.

**Lemma 8.23** *Let $\mathcal{A}$ be an atlas on a set $M$. The following are equivalent for a point $y \in M$ and a subset $Y \subseteq M$:*

(1) *There is a chart $\psi \in \mathcal{A}$ such that $y \in \operatorname{dom} \psi$ and $\psi[Y]$ is a neighbourhood of $\psi(y)$. (We do not assume that $Y$ is a subset of the domain of $\psi$, see Remark 8.7.)*
(2) *For every chart $\psi \in \mathcal{A}$ such that $y \in \operatorname{dom} \psi$, $\psi[Y]$ is a neighbourhood of $\psi(y)$.*

***Proof*** (2)⇒(1) follows from the requirement on $\mathcal{A}$ that every $y \in M$ is in the domain of at least one chart in $\mathcal{A}$.

Suppose that (1) holds; let $\psi$ witness this. Let $\varphi$ be any chart in $\mathcal{A}$ with $y \in \operatorname{dom} \varphi$; let $Z = \operatorname{dom} \psi \cap \operatorname{dom} \varphi$. Since $\psi[Z]$ is open and $y \in Z$, it is a neighbourhood of $\psi(y)$. It follows that $W = \psi[Y] \cap \psi[Z]$ is a neighbourhood of $\psi(y)$ (Exercise 8.2). Since the transition map $\psi \circ \varphi^{-1}$ is continuous at $\varphi(y)$, the inverse image $(\psi \circ \varphi^{-1})^{-1}[W] = \varphi[\psi^{-1}[W]]$ of $W$ by this transition map is a neighbourhood of $\varphi(y)$; and it is a subset of $\varphi[Y]$, which is therefore also a neighbourhood of $\varphi(y)$. □

**Definition 8.24** When the conditions of Lemma 8.23 hold, we say that $Y$ is an *$\mathcal{A}$-neighbourhood* of $y$.

As with algebraic objects, the atlas $\mathcal{A}$ is often understood but not mentioned, so we just say "neighbourhood".

**Lemma 8.25** *Let $\mathcal{A}$ be an atlas on a set $M$. The following are equivalent for $Y \subseteq M$:*

(1) *$Y$ is the union of sets of the form $\psi^{-1}[U]$, where $\psi \in \mathcal{A}$ and $U \subseteq \operatorname{range}\psi$ is open.*
(2) *$Y$ is an $\mathcal{A}$-neighbourhood of every point $y \in Y$.*

***Proof*** (2)⇒(1): For each $y \in Y$ choose some $\psi_y \in \mathcal{A}$ such that $y \in \operatorname{dom}\psi_y$. Since $\psi_y[Y]$ is a neighbourhood of $\psi_y(y)$, there is some open $U_y \subseteq \psi_y[Y]$ such that $\psi_y(y) \in U_y$. Then $Y = \bigcup_{y\in Y} \psi_y^{-1}[U_y]$.

(1)⇒(2): Let $y \in Y$. There is some $\psi \in \mathcal{A}$ and open $U \subseteq \operatorname{range}\psi$ such that $\psi^{-1}[U] \subseteq Y$ and $y \in \psi^{-1}[U]$. This means that $y \in \operatorname{dom}\psi$, and $U \subseteq \psi[Y]$ and $\psi(y) \in U$, so $\psi[Y]$ is a neighbourhood of $\psi(y)$. It follows that $Y$ is an $\mathcal{A}$-neighbourhood of $y$. □

**Definition 8.26** A subset $Y$ of $M$ satisfying the conditions of Lemma 8.25 is called an *$\mathcal{A}$-open* subset of $M$.

Again $\mathcal{A}$ is usually understood. The domain $\operatorname{dom}\psi$ of any chart $\psi$ is $\mathcal{A}$-open.

*Exercise 8.27* Let $\mathcal{A}$ be an atlas on a set $M$. Show that a subset $X \subseteq M$ is an $\mathcal{A}$-neighbourhood of a point $x \in X$ if and only if there is an $\mathcal{A}$-open set $V \subseteq X$ such that $x \in V$. «

*Exercise 8.28* Let $\mathcal{A}$ be an atlas on a set $M$. Show that: the empty set and $M$ are $\mathcal{A}$-open subsets of $M$; and that the collection of $\mathcal{A}$-open subsets of $M$ is closed under taking arbitrary unions and finite intersections. «

### 8.2.2 Subspaces

Let $\mathcal{A}$ be an atlas on a set $M$, and let $X \subseteq M$ be any subset. We say that a set $Z \subseteq X$ is *$\mathcal{A}$-open in $X$* if $Z = U \cap X$ for some $\mathcal{A}$-open subset $U$ of $M$. We also say that $Z \subseteq X$ is an $\mathcal{A}$-neighbourhood of $x$ in $X$ if there is an $\mathcal{A}$-neighbourhood $Y \subseteq M$ of $x$ such that $Z = Y \cap X$.

*Exercise 8.29* (a) Show that $Z \subseteq X$ is an $\mathcal{A}$-neighbourhood of $x$ in $X$ if and only if there is some $V \subseteq Z$ such that $x \in V$ and $V$ is $\mathcal{A}$-open in $X$. (b) Show that a set $Y \subseteq X$ is $\mathcal{A}$-open in $X$ if and only if it is an $\mathcal{A}$-neighbourhood in $X$ of every point $y \in Y$. (c) Show that the collection of sets which are $\mathcal{A}$-open in $X$ contains the empty set, $X$, and is closed under taking arbitrary unions and finite intersections. «

*Exercise 8.30* Show that if $U \subseteq M$ is $\mathcal{A}$-open, then a subset of $U$ is $\mathcal{A}$-open in $U$ if and only if it is $\mathcal{A}$-open. «

The situation becomes potentially confusing when $X$ itself has an atlas $\mathcal{B}$. We then could have two conflicting notions of open subsets of $X$: those determined by the atlas $\mathcal{B}$, and those inherited from $M$ (and the atlas $\mathcal{A}$) by taking intersections with $X$. *In all such cases we will ensure that these two notions coincide.*

**Definition 8.31** Let $\mathcal{A}$ be an atlas on a set $M$ and let $\mathcal{B}$ be an atlas on a set $X \subseteq M$. We say that $(X, \mathcal{B})$ is a *topological subspace* of $(M, \mathcal{A})$ if a subset of $X$ is $\mathcal{B}$-open if and only if it is $\mathcal{A}$-open in $X$.

*Example 8.32* Let $\mathcal{B} = \{\psi_+, \psi_-, \varphi_+, \varphi_-\}$ be the atlas for the unit circle $S$ defined in Example 8.18. Then $(S, \mathcal{B})$ is a topological subspace of $\mathbb{R}^2$ (equipped with the trivial atlas $\{\mathrm{id}_{\mathbb{R}^2}\}$). For example, the pull-backs $\psi_+^{-1}[U]$ by $\psi_+$ of open sets $U \subseteq (-1, 1)$ are the intersections $S \cap U \times (0, \infty)$, which are open in $S$ in the subspace topology. In the other direction, If $V$ is an open subset of $\mathbb{R}^2$ and $\boldsymbol{x} = (a, b) \in V \cap S$ then $B(\boldsymbol{x}, r) \subseteq V$ for some $r > 0$. Say, for example, $a > 0$; then $V \cap S$ contains $\psi_+^{-1}[(a - r, a + r)]$. Thus, $V \cap S$ is a $\mathcal{B}$-neighbourhood of $\boldsymbol{x}$. «

*Exercise 8.33* Similarly, show that the unit sphere $S^2$ in $\mathbb{R}^3$ equipped with the atlas $\{\sigma_+, \overline{\sigma_-}\}$ (Example 8.21) is a topological subspace of $\mathbb{R}^3$. «

### 8.2.3 The Hausdorff Property

**Definition 8.34** Let $\mathcal{A}$ be an atlas on a set $M$. We say that $(M, \mathcal{A})$ is *Hausdorff* if for any distinct $y, z \in M$ there are disjoint $Y, Z \subseteq M$ such that $Y$ is an $\mathcal{A}$-neighbourhood of $y$ and $Z$ is an $\mathcal{A}$-neighbourhood of $z$.[3]

Exercise 8.4 says that each $\mathbb{R}^n$ (with the trivial atlas) is Hausdorff. In fact, every example of an atlas discussed so far satisfies the Hausdorff property. (For a different example see Exercise 8.124).

*Exercise 8.35* Let $\mathcal{A}$ be an atlas on a set $M$, and suppose that $(M, \mathcal{A})$ is Hausdorff. Show that for all $X \subseteq M$, for all $y, z \in X$, there are $Y, Z \subseteq X$ such that $Y$ is an $\mathcal{A}$-neighbourhood of $y$ in $X$ and $Z$ is an $\mathcal{A}$-neighbourhood of $z$ in $X$. «

*Example 8.36* The unit circle with the atlas $\mathcal{B} = \{\psi_+, \psi_-, \varphi_+, \varphi_-\}$ is Hausdorff. This follows from Exercise 8.35, the fact that $\mathbb{R}^2$ is Hausdorff (Exercise 8.4) and $(S, \mathcal{B})$ is a topological subspace of $\mathbb{R}^2$ (Example 8.32). (Of course, that $(S, \mathcal{B})$ is

[3] We say that $Y$ and $Z$ *separate* $y$ and $z$.

Hausdorff can also be verified directly.) Similarly, the unit sphere $S^2$ is Hausdorff (Exercise 8.33). «

*Exercise 8.37* Let $\mathcal{A}$ be an atlas on a set $M$. Let $y, z \in M$ and suppose that there is a chart $\psi \in \mathcal{A}$ such that $y, z \in \text{dom } \psi$. Show that there are $\mathcal{A}$-neighbourhoods $Y$ of $y$ and $Z$ of $z$ which are disjoint. «

*Example 8.38* Let $n \geqslant 2$. Then $\mathbb{P}^n(\mathbb{R})$ and $\mathbb{P}^n(\mathbb{C})$ (with the affine cover for an atlas) are Hausdorff. For let $p, q \in \mathbb{P}^n$. If there is some $i \leqslant n$ such that $p, q \in U_i$ then neighbourhoods separating $p$ and $q$ are given by Exercise 8.37. Otherwise, for all $i \leqslant n$ either $p_i = 0$ or $q_i = 0$ (where $p = (p_0 : \cdots : p_n)$ and $q = (q_0 : \cdots : q_n)$). For neatness consider the simplest example $n = 1$, so $p = (1:0)$ and $q = (0:1)$. We then let $V_p = \{(1:a) : |a| < 1\}$ and $V_q = \{(a:1) : |a| < 1\}$. Certainly $p \in V_p$ and $q \in V_q$. $V_p = \rho_0[B(0,1)]$ and $V_q = \rho_1[B(0,1)]$ and so are open in $\mathbb{P}^n$ (in the real case $B(0,1) = (-1,1)$, in the complex case $B(0,1) = \{z \in \mathbb{C} : |z| < 1\}$ is the interior of the unit circle). And $V_p \cap V_q = \varnothing$, because if $(1:a) = (b:1)$ and $|a| < 1$ then $|b| > 1$. (Verify the Hausdorff condition when $n \geqslant 2$.) «

### 8.2.4 Topological Countability

**Lemma 8.39** *Let $\mathcal{A}$ be an atlas on a set $M$ and let $X \subseteq M$. Let $x \in X$. There is a collection $\{U_n : n \in \mathbb{N}\}$ of $X$-open neighbourhoods of $x$ such that every $X$-neighbourhood of $x$ is a superset of one of the sets $U_n$, in fact of all but finitely many.*

Note that "all but finitely many" can be obtained from the weaker condition by replacing $U_n$ by $\bigcap_{m \leqslant n} U_m$.

***Proof*** Fix a chart $\psi$ for $M$ such that $x \in \text{dom } \psi$. For every $n > 0$ let $U_n = X \cap \psi^{-1}[B(\psi(x), 1/n)]$. That every neighbourhood of $x$ contains one (and so all but finitely many) of the $U_n$ follows from Lemma 8.23. □

**Definition 8.40** Let $\mathcal{A}$ be an atlas on $M$ and let $X \subseteq M$. A *basis for the topology on $X$* is a collection $\mathcal{U}$ of sets, each $\mathcal{A}$-open in $X$, such that for any $O \subseteq X$, $\mathcal{A}$-open in $X$, $O = \bigcup \{U \in \mathcal{U} : U \subseteq O\}$. (That is: $O$ is the union of all of its subsets which are in $\mathcal{U}$.)

*Exercise 8.41* Let $\mathcal{A}$ be an atlas on $M$ and let $X \subseteq M$. Show that if $\mathcal{U}$ is a basis for the topology on $M$ then $\{U \cap X : U \in \mathcal{U}\}$ is a basis for the topology on $X$. «

The property described in Lemma 8.39 is usually called in the literature the "first countable" property. The "second countable" property is the existence of a

countable basis for the topology. Exercise 8.3 says that there is a countable basis for the topology on $\mathbb{R}^n$.

**Proposition 8.42** *Let $\mathcal{A}$ be an atlas on $M$. If $\mathcal{A}$ is countable then there is a countable basis for the topology on $M$.*

***Proof*** For each chart $\psi \in \mathcal{A}$ consider the collection of sets $\psi^{-1}[B(\boldsymbol{q}, r)]$ where $B(\boldsymbol{q}, r)$ is a rational ball (see Exercise 8.3) contained in the range of $\psi$. This is a countable collection. Combining these collections for all charts in $\mathcal{A}$ gives a countable basis for $M$: every $\mathcal{A}$-open subset of $M$ is the union of pull-backs of balls by charts (Lemma 8.25), and each open ball is the union of rational open balls (Exercise 8.3 and Lemma 8.25). (We use the fact that the union of countably many countable sets is countable; there is a bijection between $\mathbb{N}^2$ and $\mathbb{N}$.) □

*Example 8.43* In particular, if $\mathcal{A}$ is finite then $M$ has a countable basis for its topology. So for example $\mathbb{P}^n(\mathbb{R})$ and $\mathbb{P}^n(\mathbb{C})$ (Examples 8.19 and 8.20) have a countable basis for their topology. «

*Exercise 8.44* Let $\mathcal{A}$ be an atlas on $M$ and let $X \subseteq M$. A set $D \subseteq X$ is *dense in $X$* if it intersects every nonempty subset of $X$ which is $\mathcal{A}$-open in $X$. Show that there is a countable basis for the topology on $X$ if and only if there is a countable subset of $X$ which is dense in $X$. «

For an example of an atlas which does not give a countable basis for the topology see Exercise 8.125.

### 8.2.5 Manifolds

**Definition 8.45** Let $n \geqslant 1$. An *$n$-manifold* is a pair $(M, \mathcal{A})$ where $\mathcal{A}$ is an atlas on $M$ (with charts mapping to $\mathbb{R}^n$) such that: (i) $(M, \mathcal{A})$ is Hausdorff; and (ii) there is a countable basis for the topology on $M$.

*Example 8.46* All the examples we gave so far of atlases give manifolds; in particular, projective spaces $\mathbb{P}^n(\mathbb{R})$ and $\mathbb{P}^n(\mathbb{C})$ are manifolds (Examples 8.38 and 8.43), as is the unit circle. «

*Example 8.47* Suppose that $M = (M, \mathcal{A})$ and $N = (N, \mathcal{B})$ are manifolds ($m$ and $n$-dimensional respectively). (a) Show that for each $\psi \in \mathcal{A}$ and $\varphi \in \mathcal{B}$, the map $\psi \times \varphi$ (to $\mathbb{R}^{m+n}$) is a chart. (See Exercise 8.10.) (b) Show that $\mathcal{A} \times \mathcal{B} = \{\psi \times \varphi : \psi \in \mathcal{A} \ \& \ \varphi \in \mathcal{B}\}$ is an atlas on the Cartesian product $M \times N$. (c) Show that $(M \times N, \mathcal{A} \times \mathcal{B})$ is a manifold. (d) If $X \subseteq M$ and $Y \subseteq N$ are subsets then $X \times Y$ is a subset of $M \times N$. Let $U \subseteq X$ and $V \subseteq Y$. Show that $U \times V$ is $(\mathcal{A} \times \mathcal{B})$-open in $X \times Y$ if and only if $U$ is $\mathcal{A}$-open in $X$ and $V$ is $\mathcal{B}$-open in $Y$. (e) Show that a set

$O \subseteq X \times Y$ is open in $X \times Y$ if and only if it is the union of (possibly infinitely many) sets of the form $U \times V$, where $U$ is $\mathcal{A}$-open in $X$ and $V$ is $\mathcal{B}$-open in $Y$. «

### 8.2.6 Spaces and Continuity

Exercise 8.29 says that if $X$ is a subset of a manifold $M$ then the subsets of $X$ which are open in $X$ satisfy the abstract definition of a topological space. As discussed above, we do not need the full generality of this definition, so we give the following definition:

**Definition 8.48** A *quasi-Euclidean space* is a subset $X$ of a manifold $M$, equipped with the collection of sets which are open in $X$.

*Remark 8.49* By Definition 8.45 and Exercise 8.41, every quasi-Euclidean space has a countable basis for its topology. By Exercise 8.35, every quasi-Euclidean space is Hausdorff. «

The morphisms between spaces are the continuous maps.

**Definition 8.50** Let $X$ and $Y$ be quasi-Euclidean spaces. A function $f\colon X \to Y$ is *continuous* if for every set $U \subseteq Y$ which is open in $Y$, $f^{-1}[U]$ is open in $X$.

As above we can also generalise the pointwise definition: A function $f\colon X \to Y$ is continuous at a point $a \in X$ if for every $U \subseteq Y$ which is a neighbourhood of $f(a)$ in $Y$, $f^{-1}[U]$ is a neighbourhood of $a$ in $X$. A function $f\colon X \to Y$ is continuous if and only if it is continuous at every point $a \in X$.

*Exercise 8.51* Show that the projection map $\pi_n\colon \mathbb{A}^{n+1}(\mathbb{R}) \to \mathbb{P}^n(\mathbb{R})$ (Definition 4.4) is continuous. Show that the same holds when $\mathbb{R}$ is replaced by $\mathbb{C}$. «

*Exercise 8.52* Let $X$, $Y$ and $Z$ be quasi-Euclidean spaces, and let $f\colon X \to Y$ and $g\colon X \to Z$. (a) Show that the pair map $(f, g)\colon X \to Y \times Z$ is continuous if and only if both $f$ and $g$ are continuous. (b) Let $h\colon Y \times Z \to X$ be continuous. Show that for all $y^* \in Y$ and $z^* \in Z$, the maps $z \mapsto h(y^*, z)$ (from $Z$ to $X$) and $y \mapsto h(y, z^*)$ (from $Y$ to $X$) are continuous. (c) Show that the converse does not hold. (d) Show that if $W$ is another quasi-Euclidean space, $f\colon X \to Y$ and $g\colon W \to Z$ then the map $(f \times g)\colon X \times W \to Y \times Z$ is continuous if and only if both $f$ and $g$ are continuous. «

It is not the case that the inverse of a bijective continuous function is always continuous.

*Example 8.53* Let $X = [0, 2\pi)$ be the subset of $\mathbb{R}$ and let $S$ be the unit circle. The function $f(t) = (\cos t, \sin t)$ is a bijection between $X$ and $S$, and is continuous.

However its inverse is not continuous: the set $U = [0, 1)$ is an open subset of $X$, but $f[U]$ contains $(1, 0)$ and is not a neighbourhood of $(1, 0)$ in $S$. «

For this reason the notion of sameness in this category requires bicontinuity.

**Definition 8.54** A *homeomorphism* is a bijection $f\colon X \to Y$ between two quasi-Euclidean spaces such that both $f$ and $f^{-1}$ are continuous.

We say that $X$ and $Y$ are *homeomorphic* if there is a homeomorphism between them.

*Exercise 8.55* Show that the unit circle is homeomorphic to $\mathbb{P}^1(\mathbb{R})$. (Consider the embedding of $\mathbb{A}^1$ into $\mathbb{P}^1$. But see also Fig. 1.1 of the introduction.) «

*Exercise 8.56* Define a function $f\colon S^2 \to \mathbb{P}^1(\mathbb{C})$ by letting $f(\boldsymbol{p}_+) = (1:0)$, $f(\boldsymbol{p}_-) = (0:1)$, and for $\boldsymbol{q} \in S^2 \setminus \{\boldsymbol{p}_+, \boldsymbol{p}_-\}$ let $f(\boldsymbol{q}) = \rho_0(\sigma_+(\boldsymbol{q})) = \rho_1(\overline{\sigma_-(\boldsymbol{q})})$ (see Example 8.21; verify the equality). Show that $f$ is a homeomorphism from $S^2$ to $\mathbb{P}^1(\mathbb{C})$. «

*Exercise 8.57* Show that the composition of continuous functions between quasi-Euclidean spaces is continuous. Show that being homeomorphic is an equivalence relation on quasi-Euclidean spaces. «

*Exercise 8.58* Let $M$ be a manifold and let $\psi$ be a chart for $M$. Show that $\psi$ is a homeomorphism between $\operatorname{dom}\psi$ and $\operatorname{range}\psi$. «

*Exercise 8.59* Let $\mathbb{K}$ be either $\mathbb{R}$ or $\mathbb{C}$. Let $X$ be a quasi-Euclidean space, and let $f_0, f_1, \ldots, f_n\colon X \to \mathbb{K}$ be continuous. Suppose also that for no $x \in X$ do we have $f_0(x) = f_1(x) = \cdots = f_n(x) = 0$. Show that the map $x \mapsto (f_0(x) : f_1(x) : \cdots : f_n(x))$ is a continuous map from $X$ to $\mathbb{P}^n(\mathbb{K})$. «

*Exercise 8.60* Show that changes of coordinates of $\mathbb{P}^n(\mathbb{R})$ (and of $\mathbb{P}^n(\mathbb{C})$) are homeomorphisms. «

The following is essentially the standard definition of a manifold: a space already equipped with a topology is an $n$-manifold if it is locally homeomorphic to $\mathbb{R}^n$.

**Proposition 8.61** *Let $X$ be a quasi-Euclidean space. Let $n \geqslant 1$, and suppose that $\mathcal{A}$ is a family of functions satisfying:* (i) *Every $\psi \in \mathcal{A}$ is a homeomorphism between an open subset of $X$ and an open subset of $\mathbb{R}^n$; and* (ii) *Every point $x \in X$ is in the domain of some $\psi \in \mathcal{A}$. Then $\mathcal{A}$ is an atlas on $X$, $(X, \mathcal{A})$ is an $n$-manifold, and a subset of $X$ is $(X, \mathcal{A})$-open if and only if it is open in $X$. (In other words, the $(X, \mathcal{A})$-topology is identical to the original topology on $X$.)*

***Proof*** To see that two charts $\psi, \varphi$ are compatible, observe that $W = \text{dom}\,\psi \cap \text{dom}\,\varphi$ is the intersection of two open subsets of $X$, and so is open; since $\psi$ is a homeomorphism, $\psi[W]$, the domain of the transition function $\varphi \circ \psi^{-1}$, is open in $\mathbb{R}^n$; the transition function itself is the composition of two continuous functions, and so is continuous.

By Remark 8.49, $X$ is Hausdorff and has a countable basis for its topology; so it remains to check that the $(X, \mathcal{A})$-topology is identical to the $X$-topology. We take a point $x \in X$ and a set $Y \subseteq X$ with $x \in Y$, and show that $Y$ is a neighbourhood of $x$ in $X$ if and only if it is an $\mathcal{A}$-neighbourhood of $x$. Let $\psi : U \to V$ be a chart in $\mathcal{A}$ with $x \in U$. By Lemma 8.23, $Y$ is an $\mathcal{A}$-neighbourhood of $x$ if and only if $\psi[Y] = \psi[Y \cap U]$ is a neighbourhood of $\psi(x)$ in $\mathbb{R}^n$, equivalently in $V$ (as $V$ is open in $\mathbb{R}^n$); since $\psi$ is a homeomorphism, this holds if and only if $Y \cap U$ is a neighbourhood of $x$ in $U$. Since $U$ is open in $X$, $Y$ is a neighbourhood of $x$ in $X$ if and only if $U \cap Y$ is a neighbourhood of $x$ in $U$ (Exercise 8.30). □

*Example 8.62* The atlases defined on the unit circle and the unit sphere (Examples 8.18 and 8.21) satisfy the conditions of Proposition 8.61, yielding the fact that these are indeed manifolds which are topological subspaces of $\mathbb{R}^2$ and $\mathbb{R}^3$ respectively (Example 8.32 and Exercise 8.33). «

## 8.3 Compactness

Let $X$ be a quasi-Euclidean space. An *open cover* of $X$ is a family of open subsets of $X$ whose union is $X$, that is, every point in $X$ belongs to at least one open set in the collection. For example, if $\mathcal{A}$ is an atlas on a set $M$ then $\{\text{dom}\,\psi : \psi \in \mathcal{A}\}$ is an open cover of $M$. A *subcover* of an open cover $\mathcal{O}$ is a subcollection of $\mathcal{O}$ which is also a cover.

**Definition 8.63** A quasi-Euclidean space $X$ is *compact* if every open cover of $X$ has a finite subcover.

An easy example of a compact space is a finite one. A compact space can be thought of as "nearly finite".

**Proposition 8.64** *Let $X$ and $Y$ be quasi-Euclidean spaces and let $f : X \to Y$ be continuous and onto $Y$. If $X$ is compact then so is $Y$.*[4]

***Proof*** If $\mathcal{O}$ is an open cover of $Y$ then $\left\{f^{-1}[O] : O \in \mathcal{O}\right\}$ is an open cover of $X$; a finite subcover of the latter gives a finite subcover of $\mathcal{O}$. □

[4] We say that the continuous image of a compact space is compact.

### 8.3.1 Closed Sets

Compactness can be rephrased in terms of families of closed sets. A subset $Y$ of a quasi-Euclidean space $X$ is *closed* (in $X$) if its complement $X \setminus Y$ is an open subset of $X$. Both $X$ and $\varnothing$ are closed in $X$; the union of finitely many closed sets is closed, and the intersection of any number of closed sets is closed.

*Exercise 8.65* Show that every finite subset of a quasi-Euclidean space is closed. «

*Exercise 8.66* Let $X$ be a quasi-Euclidean space. For each $B \subseteq X$ we let $\overline{B}$, the *closure of* $B$, be the intersection of all closed sets $F \subseteq X$ such that $B \subseteq F$. This is well-defined since $X$ is closed. The closure $\overline{B}$ is the smallest closed set containing $B$ ($\overline{B}$ is closed, $B \subseteq \overline{B}$, and if $F$ is closed and $B \subseteq F$ then $\overline{B} \subseteq F$; compare with the idea of generated subgroups on page 37). Show that the other *Kuratowski closure axioms* hold: (i) $\overline{\varnothing} = \varnothing$; (ii) $\overline{A \cup B} = \overline{A} \cup \overline{B}$ for all $A, B \subseteq X$; and (iii) $\overline{\overline{B}} = \overline{B}$ for all $B \subseteq X$. «

*Exercise 8.67* Let $X$ be a quasi-Euclidean space. Show that a subset $D \subseteq X$ is dense in $X$ (see Exercise 8.44) if and only if its closure is $X$. «

We say that a collection $\mathcal{F}$ of closed subsets of a space $X$ has the *finite intersection property* if the intersection of any finitely many sets from $\mathcal{F}$ is nonempty.

*Exercise 8.68* Show that a quasi-Euclidean space is compact if and only if for any collection $\mathcal{F}$ of closed subsets of $X$ which has the finite intersection property, the intersection $\bigcap \mathcal{F}$ of all the sets in $\mathcal{F}$ is nonempty. «

Below we are concerned with the compactness of subsets of a quasi-Euclidean space. Compactness is equivalent to "relative" compactness. Suppose that $Y \subseteq X$. We call a collection $\mathcal{O}$ of open subsets of $X$ a cover of $Y$ if $Y \subseteq \bigcup \mathcal{O}$. This is equivalent to $\{O \cap Y : O \in \mathcal{O}\}$ being a cover of $Y$ consisting of $Y$-open sets. Then $Y$ is compact if and only if every cover of $Y$ consisting of $X$-open sets has a finite sub-cover.

*Exercise 8.69* Show that if $X$ is a compact space and $A \subseteq X$ is closed, then $A$ is compact. «

*Exercise 8.70* Show that the union of finitely many compact subsets of a space $X$ is compact. «

**Proposition 8.71** *Let $X$ be a quasi-Euclidean space and suppose that $A \subseteq X$ is compact. Then $A$ is a closed subset of $X$.*

***Proof*** We show that $X \setminus A$ is a neighbourhood of every point it contains. Let $x \in X \setminus A$. By the Hausdorff property of $X$ (Exercise 8.35), for every $y \in A$ find disjoint $X$-open sets $U_y$ containing $x$ and $V_y$ containing $y$. The collection $\{V_y : y \in A\}$ is an open cover of $A$. Since $A$ is compact we can find a finite subset $Z$ of $A$ such that $A \subseteq \bigcup_{y \in Z} V_y$. Then $\bigcap_{y \in Z} U_y$ is an $X$-open set containing $x$ and disjoint from $A$. □

**Corollary 8.72** *Let $X$ and $Y$ be quasi-Euclidean spaces; let $f : X \to Y$ be a continuous bijection. If $X$ is compact then $f$ is a homeomorphism.*

***Proof*** We need to show that $f^{-1}$ is continuous. By taking complements, it suffices to show that if $A \subseteq X$ is closed in $X$ then $f[A]$ is closed in $Y$. Let $A \subseteq X$ be closed in $X$. By Exercise 8.69, $A$ is compact. By Proposition 8.64, $f[A]$ is compact. By Proposition 8.71, $f[A]$ is closed in $Y$. □

*Exercise 8.73* (a) Show that every polynomial $f \in \mathbb{R}[x_1, \ldots, x_n]$ defines a continuous function from $\mathbb{R}^n$ to $\mathbb{R}$. (See Exercise 8.10.) (b) Conclude that every algebraic hypersurface of $\mathbb{A}^n(\mathbb{R})$ is closed. (c) Show that every algebraic hypersurface of $\mathbb{P}^n(\mathbb{R})$ is closed. (d) Show that the same holds when $\mathbb{R}$ is replaced by $\mathbb{C}$. «

### 8.3.2 Sequences and Limits

Let $X$ be a quasi-Euclidean space; let $\langle x_n \rangle$ be an infinite sequence of points in $X$; let $x \in X$. We say that the sequence $\langle x_n \rangle$ *converges* to $x$ (and write $\lim_{n \to \infty} x_n = x$) if every neighbourhood of $x$ contains $x_n$ for all but finitely many $n$. The Hausdorff property shows that a sequence can converge to at most one point.

*Exercise 8.74* Let $m \geqslant 1$. (a) Show that a sequence $\langle \boldsymbol{x}_n \rangle$ of points in $\mathbb{R}^m$ converges to a point $\boldsymbol{x} \in \mathbb{R}^m$ if and only if for every $\varepsilon > 0$ there is some $N$ such that for all $n \geqslant N$, $d(\boldsymbol{x}_n, \boldsymbol{x}) < \varepsilon$. (b) Show that if $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ and $\boldsymbol{x}_n = (x_{1,n}, \ldots, x_{m,n})$ then $\langle \boldsymbol{x}_n \rangle$ converges to $\boldsymbol{x}$ if and only if for all $i \leqslant m$, the sequence $\langle x_{i,n} \rangle$ converges to $x_i$. «

Suppose that $X$ is a subspace of a quasi-Euclidean space $Y$. Let $\langle x_n \rangle$ be a sequence of points from $X$, and let $x \in X$. On the face of it, the statement "$\langle x_n \rangle$ converges to $x$" can be interpreted differently in $X$ and in $Y$ by using either $X$-neighbourhoods of $x$ or $Y$-neighbourhoods of $x$.

*Exercise 8.75* Show that in the situation described, $\langle x_n \rangle$ converges to $x$ in the sense of $X$ if and only if $\langle x_n \rangle$ converges to $x$ in the sense of $Y$. «

It is quite possible though that $X$ is a subspace of $Y$ and that $\langle x_n \rangle$ is a sequence of points from $X$ which converges to a point in $Y \setminus X$. For example let $X = \mathbb{Q}$, $Y = \mathbb{R}$ and let $\langle x_n \rangle$ be a sequence of rational approximations of some irrational

number. The following proposition shows that this happens exactly when $X$ is not closed in $Y$.

**Proposition 8.76** *Let $X$ be a quasi-Euclidean space. A subset $A$ of $X$ is closed in $X$ if and only if for any sequence $\langle a_n \rangle$ of points from $A$, if $\langle a_n \rangle$ converges to a point $a \in X$ then $a \in A$.*

***Proof*** Let $A \subseteq X$. Suppose that $A$ is closed, and let $\langle a_n \rangle$ be a sequence of points from $A$ which converges to a point $a \in X$. The set $X \setminus A$ contains none of the points $a_n$ and so cannot be a neighbourhood of $a$. Since $X \setminus A$ is open, it is a neighbourhood of all of its points, so $a \notin X \setminus A$, which means that $a \in A$.

In the other direction suppose that $A$ is not closed in $X$. So there is a point $a \in X \setminus A$ of which $X \setminus A$ is not a neighbourhood; if $U \subseteq X$ is $X$-open and $a \in U$ then $U$ and $A$ are not disjoint. By Lemma 8.39 fix a family $\{U_n\}$ of open neighbourhoods of $a$ with the property that every neighbourhood of $a$ contains all but finitely many of the $U_n$. For each $n$ we choose some point $a_n \in A \cap U_n$. The property of the sets $U_n$ ensure that the sequence $\langle a_n \rangle$ converges to $a$. □

Continuity can be characterised using convergence of sequences.

*Exercise 8.77* Let $X$ and $Y$ be quasi-Euclidean spaces. Show that a function $f : X \to Y$ is continuous at a point $a \in X$ if and only if for any sequence $\langle a_n \rangle$ of points from $X$ which converges to $a$, the sequence $\langle f(a_n) \rangle$ converges to $f(a)$. (Hint: again consider Lemma 8.39.) «

*Exercise 8.78* Let $\langle x_n \rangle$ and $\langle y_n \rangle$ be two convergent sequences of real numbers. (a) show that $\lim_n (x_n + y_n) = \lim x_n + \lim y_n$ and $\lim_n x_n y_n = \lim x_n \cdot \lim y_n$. (Use Exercise 8.77 and the continuity of addition and multiplication, Exercise 8.11.) (b) Show that if for all $n$, $x_n \leqslant y_n$ then $\lim x_n \leqslant \lim y_n$. «

A *subsequence* of a sequence $\langle x_n \rangle$ of points is a sequence obtained by removing some elements of the original sequence. Formally, it is a sequence $\langle x_{f(n)} \rangle$ where $f : \mathbb{N} \to \mathbb{N}$ is strictly increasing.

**Proposition 8.79** *A quasi-Euclidean space $X$ is compact if and only if every sequence of points from $X$ has a subsequence which converges to some point in $X$.*

***Proof*** First suppose that $\langle x_n \rangle$ is a sequence of points from $X$ which has no converging subsequence. We note that the set of points $Z = \{x_n : n \in \mathbb{N}\}$ is infinite. In fact no point can appear infinitely often as some $x_n$; otherwise $\langle x_n \rangle$ has a constant subsequence, and that subsequence converges to that constant value.

For each $y \in X$ there is some open $V_y \subset X$ containing $y$ which contains only finitely many of the points $x_n$. To see this fix a family $\{U_m\}$ of open neighbourhoods of $y$ given by Lemma 8.39. Suppose that each $U_m$ contains infinitely many points $x_n$.

Inductively we define a subsequence $\langle x_{f(n)}\rangle$ of $\langle x_n\rangle$ which converges to $y$: given $x_{f(n-1)}$ we choose $f(n) > f(n-1)$ such that $x_{f(n)} \in U_n$.

Having chosen the sets $V_y$ as described we see that the collection $\{V_y : y \in X\}$ is an open cover of $X$ which has no finite subcover: otherwise the set $Z$ is contained in the union of finitely many sets, each of which contains only finitely many elements of $Z$; this would imply that $Z$ is finite.

In the other direction suppose that every sequence of elements of $X$ has a convergent subsequence. To show that $X$ is compact let $\mathcal{O}$ be an open cover of $X$. We show that $\mathcal{O}$ has a finite subcover in two steps: first we show that $\mathcal{O}$ has a countable subcover $\mathcal{W}$; then we find a finite subcover of $\mathcal{W}$.

Let $\mathcal{U}$ be a countable basis for the topology on $X$ (see Remark 8.49). Let $\mathcal{V}$ be the collection of sets $U \in \mathcal{U}$ such that $U \subseteq O$ for some $O \in \mathcal{O}$. Then $\mathcal{V}$ is countable (it is a subcollection of $\mathcal{U}$). Every set in $\mathcal{O}$ is the union of sets from $\mathcal{U}$, and all of these are in $\mathcal{V}$; so $\bigcup \mathcal{O} = \bigcup \mathcal{V}$, which shows that $\mathcal{V}$ is an open cover of $X$. Each element of $\mathcal{V}$ is a subset of some element of $\mathcal{O}$; choosing one for each element of $\mathcal{V}$ gives us a countable subcover $\mathcal{W}$ of $\mathcal{O}$.[5]

Let $\{W_1, W_2, \dots\}$ be an enumeration of the sets in $\mathcal{W}$, and let $X_n = \bigcup_{m \leqslant n} W_m$. We show that $X_n = X$ for some $n$. If not then for each $n$ we pick some $x_n \in X \setminus X_n$. By assumption, the sequence $\langle x_n\rangle$ has a converging subsequence $\langle x_{f(n)}\rangle$; let $x = \lim_{n\to\infty} x_{f(n)}$. Since $\bigcup \mathcal{W} = \bigcup_n X_n$ is all of $X$, there is some $n$ such that $x \in X_n$. But then $X_n$ (as an open neighbourhood of $x$) contains $x_{f(m)}$ for some $f(m) > n$, which contradicts $X_n \subseteq X_{f(m)}$ and $x_{f(m)} \notin X_{f(m)}$. This is a contradiction, so for some $n$, $\{W_k : k \leqslant n\}$ is an finite subcover of $\mathcal{W}$ and so of $\mathcal{O}$. □

*Exercise 8.80* Let $X$ and $Y$ be compact quasi-Euclidean spaces. Show that $X \times Y$ (see Example 8.47) is compact. «

### 8.3.3 Interlude: Completeness

In the following section we use a basic fact about Euclidean space $\mathbb{R}^n$, namely that it is *complete*. There are many equivalent formulations. We mention two.

A sequence of points $\langle \boldsymbol{x}_m\rangle$ from $\mathbb{R}^n$ is called a *Cauchy sequence* if for all $\varepsilon > 0$ there is some $N$ such that for all $k, m \geqslant N$, $d(\boldsymbol{x}_k, \boldsymbol{x}_m) \leqslant \varepsilon$.

**Proposition 8.81** *If $\langle \boldsymbol{x}_m\rangle$ is convergent then it is a Cauchy sequence.*

***Proof*** Let $\boldsymbol{y} = \lim_m \boldsymbol{x}_m$. Let $\varepsilon > 0$. There is some $N$ such that for all $m \geqslant N$, $d(\boldsymbol{x}_m, \boldsymbol{y}) < \varepsilon$. By the triangle inequality, for all $m, k \geqslant N$,

$$d(\boldsymbol{x}_k, \boldsymbol{x}_m) \leqslant d(\boldsymbol{x}_k, \boldsymbol{y}) + d(\boldsymbol{x}_m, \boldsymbol{y}) < 2\varepsilon.$$ □

[5] Note that this argument holds in any quasi-Euclidean space, not necessarily compact.

Restricted to some subspaces there are Cauchy sequences which do not have limits. For example, any sequence of rational numbers which converges to an irrational number is a Cauchy sequence, but does not have a limit in the subspace $\mathbb{Q}$.

Let $A \subseteq \mathbb{R}$ be nonempty. An *upper bound* for $A$ is a real number $b$ such that $b \geqslant a$ for all $a \in A$. Similarly we define lower bounds. A set is *bounded from above* if it has an upper bound; similarly we define "bounded from below". The set $A$ is *bounded* if it is bounded both from below and from above. And in general, we say that a subset of $\mathbb{R}^n$ is bounded if it is contained in some open ball $B(\boldsymbol{a}, r)$.

A *least upper bound* for $A$ is an upper bound $b$ such that $b \leqslant c$ for every upper bound $c$ of $A$; similarly we define a greatest lower bound. Directly from the definition we see that a least upper bound, if it exists, is unique, and the same holds for the greatest lower bound. We write $\sup A$ for the least upper bound and $\inf A$ for the greatest lower bound. If $A$ is not bounded from above we write $\sup A = \infty$.

*Exercise 8.82* Let $A \subseteq \mathbb{R}$ be bounded from above. (a) Show that $b = \sup A$ if and only if $b$ is an upper bound of $A$, and for all $\varepsilon > 0$ there is some $a \in A$ such that $a > b - \varepsilon$. (b) Give a similar characterisation for $\inf A$. (c) Show that if $A \subset \mathbb{R}$ is closed and nonempty, and $b = \sup A$, then $b \in A$. «

The *completeness* of $\mathbb{R}$ says that:

- Every Cauchy sequence of real numbers has a limit.
- Every subset of $\mathbb{R}$ which is bounded from above has a least upper bound.

*Exercise 8.83* Show that these two statements imply each other. «

*Exercise 8.84* (a) Show that every Cauchy sequence of elements of $\mathbb{R}^n$ has a limit. (b) Show that every subset of $\mathbb{R}$ which is bounded from below has a greatest lower bound. «

*Exercise 8.85* Show that an open subset of $\mathbb{R}$ is the union of a (countable) set of pairwise disjoint open intervals.[6] «

Finally, completeness guarantees the *Archimedean property* of $\mathbb{R}$:

**Proposition 8.86** *For every $c \in \mathbb{R}$ there is some (unique) $n \in \mathbb{Z}$ such that $n \leqslant c < n + 1$.*

***Proof*** By replacing $c$ by $-c$ we may assume that $c \geqslant 0$. It suffices to show that $c < m$ for some natural number $m$; we can then consider the least such $m$. If not,

[6] Here, as open intervals we also accept rays $(a, \infty)$ and $(-\infty, a)$, and $\mathbb{R}$ itself. This property is special to $\mathbb{R}$: it is not true that every open subset of $\mathbb{R}^2$ is the union of pairwise disjoint open balls, or open rectangles.

then $\mathbb{N}$ is a bounded subset of $\mathbb{R}$. But $\mathbb{N}$ cannot have a least upper bound; if $d$ is an upper bound for $\mathbb{N}$, then so is $d - 1$. □

### 8.3.4 Compactness in Euclidean Space

The *diameter* of a subset $A$ of $\mathbb{R}^n$ is

$$\sup \{d(\boldsymbol{x}, \boldsymbol{y}) \,:\, \boldsymbol{x}, \boldsymbol{y} \in A\}.$$

The diameter of an open ball $B(\boldsymbol{a}, r)$ in $\mathbb{R}^n$ is $2r$. The diameter of a set is finite if and only if it is bounded.

*Exercise 8.87* Let $A \subseteq \mathbb{R}^n$ be bounded, and let $\varepsilon > 0$. Show that there is a finite open cover of $A$ consisting of open balls whose diameter is at most $\varepsilon$. «

A sequence $\langle \boldsymbol{a}_k \rangle$ of points is called bounded if the set $\{\boldsymbol{a}_k \,:\, k \in \mathbb{N}\}$ is bounded.

*Exercise 8.88* Let $\langle r_n \rangle$ be a non-decreasing sequence of real numbers (if $n < m$ then $r_n \leqslant r_m$). (a) Show that $\langle r_n \rangle$ converges if and only if it is bounded, in which case $\lim r_n = \sup\{r_n \,:\, n \in \mathbb{N}\}$. (b) Show the result holds for non-increasing sequences, with sup replaced by inf. «

**Proposition 8.89** *Every bounded sequence in $\mathbb{R}^n$ has a converging subsequence.*

***Proof*** Let $\langle \boldsymbol{a}_k \rangle$ be a bounded sequence; let $r$ be the diameter of the set $Z = \{\boldsymbol{a}_k \,:\, k \in \mathbb{N}\}$. We may assume that $Z$ is infinite; otherwise the sequence $\langle \boldsymbol{a}_k \rangle$ has a constant subsequence.

Let $Z_0 = Z$. By recursion we define a decreasing sequence of sets $Z_k$ (this means that $Z_{k+1} \subseteq Z_k$) so that each $Z_k$ is infinite and the diameter of $Z_k$ is at most $r2^{-k}$. Given $Z_k$ we appeal to Exercise 8.87 and find a finite collection $\mathcal{B}$ of balls, each of diameter at most $r2^{-(k+1)}$, whose union contains $Z_k$. Since $\mathcal{B}$ is finite there is at least one $B \in \mathcal{B}$ such that $Z_k \cap B$ is infinite; we choose such $B$ and let $Z_{k+1} = Z_k \cap B$.

Having defined the sequence of sets $\langle Z_k \rangle$ we inductively choose a subsequence $\langle \boldsymbol{a}_{f(k)} \rangle$ of $\langle \boldsymbol{a}_k \rangle$; given $f(k-1)$ we find $f(k) > f(k-1)$ such that $\boldsymbol{a}_{f(k)} \in Z_k$. The sequence $\langle \boldsymbol{a}_{f(k)} \rangle$ is a Cauchy sequence and so has a limit (Exercise 8.84). □

*Exercise 8.90* Show that every compact subset of $\mathbb{R}^n$ is bounded. «

The following is known as the *Heine-Borel theorem*:

**Theorem 8.91** *A subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded.*

***Proof*** Proposition 8.71 and Exercise 8.90 give one direction. In the other, suppose that $A \subset \mathbb{R}^n$ is closed and bounded. Let $\langle \boldsymbol{a}_k \rangle$ be a sequence of points from $A$. The sequence is bounded, and so has a subsequence which converges to some point $\boldsymbol{a} \in \mathbb{R}^n$ (Proposition 8.89). Since $A$ is closed, Proposition 8.76 shows that $\boldsymbol{a} \in A$. □

Here are a couple of corollaries (both use Proposition 8.64.)

*Exercise 8.92* (a) Show that the unit sphere $S^{n-1} = \{\boldsymbol{a} \in \mathbb{R}^n : |\boldsymbol{a}| = 1\}$ in $\mathbb{R}^n$ is closed and bounded. (You can use Exercise 8.14.) (b) Conclude that for all $n \geqslant 1$, $\mathbb{P}^n(\mathbb{R})$ and $\mathbb{P}^n(\mathbb{C})$ are compact (see Exercise 8.51).[7] «

*Exercise 8.93* Let $X$ be a compact quasi-Euclidean space and let $f : X \to \mathbb{R}$ be continuous. Show that $f$ obtains both a maximum and a minimum. (That is, there are $x, y \in X$ such that for all $z \in X$, $f(x) \leqslant f(z) \leqslant f(y)$. You can use Exercise 8.82(c).) «

### Uniform Continuity

Let $A \subseteq \mathbb{R}^n$ and let $f : A \to \mathbb{R}^m$. We say that $f$ is *uniformly continuous* if for every $\varepsilon > 0$ there is some $\delta > 0$ such that for all $\boldsymbol{a}, \boldsymbol{b} \in A$, if $d(\boldsymbol{a}, \boldsymbol{b}) < \delta$ then $d(f(\boldsymbol{a}), f(\boldsymbol{b})) < \varepsilon$. If $f$ is uniformly continuous then it is continuous;[8] uniform continuity requires $\delta$ to be dependent on $\varepsilon$ alone but not on the point at which we are measuring continuity.

**Proposition 8.94** *Suppose that $K \subset \mathbb{R}^n$ is compact. Then any continuous function $f : K \to \mathbb{R}^m$ is uniformly continuous.*

***Proof*** Fix $\varepsilon > 0$. Let $\mathcal{U}$ be the collection of open balls $B(\boldsymbol{a}, r)$ such that for all $\boldsymbol{b} \in K \cap B(\boldsymbol{a}, 2r)$ we have $d(f(\boldsymbol{b}), f(\boldsymbol{a})) < \varepsilon$. The continuity of $f$ implies that $\mathcal{U}$ is an open cover of $K$. Let $B(\boldsymbol{a}_1, r_1), B(\boldsymbol{a}_2, r_2), \ldots, B(\boldsymbol{a}_k, r_k)$ be a finite sub-cover. Let $\delta = \min_{i \leqslant k} r_i$. Let $\boldsymbol{b}, \boldsymbol{c} \in K$ and suppose that $d(\boldsymbol{b}, \boldsymbol{c}) < \delta$. Find some $i \leqslant k$ such that $\boldsymbol{b} \in B(\boldsymbol{a}_i, r_i)$. Then $\boldsymbol{c} \in B(\boldsymbol{a}_i, 2r_i)$ and so $d(f(\boldsymbol{b}), f(\boldsymbol{c})) \leqslant d(f(\boldsymbol{b}), f(\boldsymbol{a}_i)) + d(f(\boldsymbol{c}), f(\boldsymbol{a}_i)) < 2\varepsilon$. □

### Distances from Sets

We defined the distance $d(\boldsymbol{x}, \boldsymbol{y}) = |\boldsymbol{y} - \boldsymbol{x}|$ between two points in $\mathbb{R}^n$; we will need to extend this to two other notions: the distance between a point and a set, and the distance between two sets.

For a nonempty set $C \subseteq \mathbb{R}^n$ and a point $\boldsymbol{x} \in \mathbb{R}^n$ we let

$$d(\boldsymbol{x}, C) = \inf \{d(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{y} \in C\}.$$

[7] In contrast, $\mathbb{A}^n(\mathbb{R})$ and $\mathbb{A}^n(\mathbb{C})$ are not compact.

[8] The converse fails, see Exercise 8.132.

Note that if $\boldsymbol{x} \in C$ then $d(\boldsymbol{x}, C) = 0$. The converse may fail in general. (Consider for example points in the closure of an open ball.)

**Proposition 8.95** *Let $C \subseteq \mathbb{R}^n$ be nonempty.*

(a) *The function $d(\boldsymbol{x}, C) \colon \mathbb{R}^n \to \mathbb{R}$ is continuous.*
(b) *If $C$ is closed, then for all $\boldsymbol{x} \in \mathbb{R}^n$, $d(\boldsymbol{x}, C) = 0$ if and only if $\boldsymbol{x} \in C$.*

***Proof*** For (a), let $\varepsilon > 0$, let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and suppose that $d(\boldsymbol{x}, \boldsymbol{y}) \leqslant \varepsilon$. Find $\boldsymbol{z} \in C$ such that $d(\boldsymbol{x}, \boldsymbol{z}) < d(\boldsymbol{x}, C) + \varepsilon$ (Exercise 8.82). Then $d(\boldsymbol{y}, C) \leqslant d(\boldsymbol{y}, \boldsymbol{z}) \leqslant d(\boldsymbol{x}, C) + 2\varepsilon$. By symmetry, $|d(\boldsymbol{y}, C) - d(\boldsymbol{x}, C)| \leqslant 2\varepsilon$.

For (b), suppose that $C$ is closed and that $d(\boldsymbol{x}, C) = 0$. For every $k$ we can find some $\boldsymbol{z}_k \in C$ such that $d(\boldsymbol{x}, \boldsymbol{z}_k) < 1/k$. Then $\lim \boldsymbol{z}_k = \boldsymbol{x}$. Proposition 8.76 implies that $\boldsymbol{x} \in C$. □

For subsets $A$ and $B$ of $\mathbb{R}^n$ define $d(A, B)$ to be

$$\inf \{d(\boldsymbol{x}, \boldsymbol{y}) \,:\, \boldsymbol{x} \in A \;\&\; \boldsymbol{y} \in B\}.$$

*Exercise 8.96* Show that $d(A, B) = \inf \{d(\boldsymbol{a}, B) \,:\, \boldsymbol{a} \in A\}$. «

**Proposition 8.97** *Suppose that $B$ is closed, $A$ is compact and that $A$ and $B$ are disjoint. Then $d(A, B) > 0$.*

***Proof*** By Proposition 8.95, $d(\boldsymbol{a}, B) > 0$ for all $\boldsymbol{a} \in A$. Since the distance from $B$ is a continuous function on $A$ and $A$ is compact, this function attains a minimum (Exercise 8.93). □

*Exercise 8.98* Give an example of two closed subsets $F$ and $G$ of $\mathbb{R}^n$ which are disjoint but such that $d(F, G) = 0$. «

## 8.4 Quotients by Discrete Subgroups

We discuss quotients of $\mathbb{R}^n$ by discrete subgroups. The torus is one kind of these.

**Discrete Sets**

**Definition 8.99** Let $X$ be a quasi-Euclidean space. A point $x \in X$ is *isolated* if the singleton $\{x\}$ is open in $X$. The space $X$ is called *discrete* if every point in $X$ is isolated.

For example, the set of integers (considered as a subspace of $\mathbb{R}$) is discrete. The definition of being a subspace (Definition 8.31) implies that if $X \subseteq M$ is a subspace of a manifold then $x \in X$ is isolated (in $X$) if and only if there is some $M$-neighbourhood $U$ of $x$ such that $U \cap X = \{x\}$.

*Exercise 8.100* Show that every finite quasi-Euclidean space is discrete. «

**Proposition 8.101** *If a quasi-Euclidean space is discrete and compact then it is finite.*

***Proof*** If $X$ is discrete then $\{\{x\} : x \in X\}$ is an open cover of $X$. A finite sub-cover shows that $X$ is finite. □

*Exercise 8.102* Let $X$ be a quasi-Euclidean space. Show that a set $Z \subset X$ is discrete and closed if and only if there is an open cover of $X$ consisting of sets $U$, each of which intersects $Z$ in at most one point. «

**Discrete Subgroups of $\mathbb{R}^n$**

Let $n \geqslant 1$. We also think of $\mathbb{R}^n$ as an abelian group (equipped with addition of points).

**Lemma 8.103** *A subgroup $G$ of $\mathbb{R}^n$ is discrete if and only if there is some $\delta > 0$ such that for all distinct $\boldsymbol{a}, \boldsymbol{b} \in G$, $d(\boldsymbol{a}, \boldsymbol{b}) \geqslant \delta$.*

***Proof*** If $G$ is not discrete then let $\boldsymbol{a} \in G$ be non-isolated; this means that for all $\delta > 0$ we can find a group element $\boldsymbol{b} \in G \cap B(\boldsymbol{a}, \delta)$ distinct from $\boldsymbol{a}$.

Suppose that $G$ is discrete. Since $\mathbf{0} = 0_{\mathbb{R}^n}$ is in $G$ and is isolated, there is some $\delta > 0$ such that $G \cap B(\mathbf{0}, \delta) = \{\mathbf{0}\}$. Let $\boldsymbol{a}, \boldsymbol{b} \in G$ and suppose that $|\boldsymbol{b} - \boldsymbol{a}| < \delta$. Since $\boldsymbol{b} - \boldsymbol{a} \in G$, we must have $\boldsymbol{b} = \boldsymbol{a}$. □

*Example 8.104* Unlike linear subspaces, there are many kinds of subgroups of $\mathbb{R}^n$; most are not discrete (for example, $\mathbb{Q}$ as a subgroup of $\mathbb{R}$). The canonical example of a discrete subgroup of $\mathbb{R}^n$ is $\mathbb{Z}^n$. We can choose $\delta = 1$. «

Lemma 8.103 and Exercise 8.102 (or Proposition 8.76) imply:

**Corollary 8.105** *Every discrete subgroup of $\mathbb{R}^n$ is closed.*

Indeed there is a characterisation of discrete subgroups of $\mathbb{R}^n$.

**Proposition 8.106** *A subgroup of $\mathbb{R}^n$ is discrete if and only if it is generated by a linearly independent subset of $\mathbb{R}^n$.*

***Proof*** In one direction, let $\mathsf{a} \subset \mathbb{R}^n$ be linearly independent. Let $k = |\mathsf{a}|$; let $G$ be the subgroup of $\mathbb{R}^n$ generated by $\mathsf{a}$, and let $U$ be the linear subspace of $\mathbb{R}^n$ spanned by $\mathsf{a}$. Let $T : \mathbb{R}^k \to U$ be a linear isomorphism obtained by mapping the standard basis of $\mathbb{R}^k$ to $\mathsf{a}$. Then $T[\mathbb{Z}^k] = G$. By Exercise 8.13, $T$ is a homeomorphism, and so restricts to a homeomorphism between $\mathbb{Z}^k$ and $G$. Since $\mathbb{Z}^k$ is discrete, so is $G$.

For the other direction, let $G \subset \mathbb{R}^n$ be a discrete subgroup. Let $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\} \subset G$ be maximal linearly independent; then the linear span of $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\}$ equals the linear span of $G$, call it $U$. Let $H$ be the subgroup generated by $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$. So $H$ is a subgroup of $G$.

Let $P = \left\{\sum_{i \leqslant k} \lambda_i \boldsymbol{u}_i \,:\, \lambda_i \in [0,1]\right\}$. This is the image of the unit cube in $\mathbb{R}^k$ under a linear map (which is continuous), and so $P$ is compact. Further, for every $\boldsymbol{a} \in G$ there is some $\boldsymbol{x} \in P \cap G$ such that $\boldsymbol{a} - \boldsymbol{x} \in H$. Namely, if $\boldsymbol{a} = \sum \mu_i \boldsymbol{u}_i$ (as $\boldsymbol{a} \in U$) then we choose integers $n_i$ such that $n_i \leqslant \mu_i < n_i + 1$ (Proposition 8.86) and let $\boldsymbol{x} = \sum(\mu_i - n_i)\boldsymbol{u}_i$.

Since $G$ is discrete and closed and $P$ is compact, $P \cap G$ is discrete and compact, and so is finite (Proposition 8.101). This shows that $G/H$ is a finite group. Let $q = |G/H|$ be the size of $G/H$. By Proposition 2.50, for all $\boldsymbol{a} \in G$, $q\boldsymbol{a} \in H$. In other words, $G$ is a subgroup of the group generated by $\{(1/q)\boldsymbol{u}_1, \ldots, (1/q)\boldsymbol{u}_k\}$.

For every $\boldsymbol{g} \in G$ write $\boldsymbol{g} = \sum_{i \leqslant k} \alpha_i(\boldsymbol{g})\boldsymbol{u}_i$; so $\alpha_i(\boldsymbol{g})$ is an integer multiple of $1/q$. Let $U_0 = \{\boldsymbol{0}\}$ and for $i = 1, \ldots, k$, let $U_i$ be the linear span of $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_i\}$ (so $U_k = U$). For each such $i$ we choose some $\boldsymbol{w}_i \in G \cap U_i$ such that $\alpha_i(\boldsymbol{w}_i) > 0$, and is smallest among $\alpha_i(\boldsymbol{w})$ for all $\boldsymbol{w} \in G \cap U_i$ for which $\alpha_i(\boldsymbol{w}) > 0$. Since $\boldsymbol{w}_i \in U_i \setminus U_{i-1}$, we see that $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k\}$ is linearly independent and so is a basis for $U$.

It remains to show that $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k\}$ generate $G$. By induction on $i$ we show that $G \cap U_i$ is the subgroup generated by $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_i$. Suppose that this has been shown for $i - 1$. Let $\boldsymbol{g} \in G \cap U_i$. Since $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_i\}$ linearly span $U_i$, we write $\boldsymbol{g} = \sum_{j \leqslant i} \lambda_j \boldsymbol{w}_j$ where $\lambda_1, \ldots, \lambda_i \in \mathbb{R}$. We want to show that each $\lambda_j$ is an integer.

Let $m$ be the integer such that $m \leqslant \lambda_i < m + 1$. Let $\boldsymbol{b} = \boldsymbol{g} - m\boldsymbol{w}_i$. Then $\boldsymbol{b} \in G \cap U_i$. We have both $\boldsymbol{b} - (\lambda_i - m)\boldsymbol{w}_i \in U_{i-1}$ and $\boldsymbol{w}_i - \alpha_i(\boldsymbol{w}_i)\boldsymbol{u}_i \in U_{i-1}$, so $\boldsymbol{b} - (\lambda_i - m)\alpha_i(\boldsymbol{w}_i)\boldsymbol{u}_i \in U_{i-1}$; so $\alpha_i(\boldsymbol{b}) = (\lambda_i - m)\alpha_i(\boldsymbol{w}_i)$. This is non-negative but smaller than $\alpha_i(\boldsymbol{w}_i)$. The minimality of $\alpha_i(\boldsymbol{w}_i)$ means that $\alpha_i(\boldsymbol{b}) = 0$, i.e., that $\lambda_i = m$. So $\boldsymbol{b}$ is in $G \cap U_{i-1}$; by induction it is in the subgroup generated by $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{i-1}$. So $\lambda_j$ for $j < i$ are all integers as well. □

**Quotients by Discrete Subgroups**

Let $G$ be a discrete subgroup of $\mathbb{R}^n$; let $\delta > 0$ be given by Lemma 8.103. If the diameter of a set $U \subset \mathbb{R}^n$ is smaller than $\delta$, then $U$ intersects every coset of $G$ in at most one point. For if $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^n$ are in the same coset of $G$ and $|\boldsymbol{q} - \boldsymbol{p}| < \delta$ then as $\boldsymbol{q} - \boldsymbol{p} \in G$, $\boldsymbol{p} = \boldsymbol{q}$. In other words, $\pi_G\restriction_U$ is one-to-one, where $\pi_G \colon \mathbb{R}^n \to \mathbb{R}^n/G$ is the quotient map (see page 38).

We define an atlas on the quotient group $\mathbb{R}^n/G$ as follows. We let $\mathcal{A} = \mathcal{A}(\mathbb{R}^n, G)$ be the collection of all maps $\psi_U = (\pi_G\restriction_U)^{-1}$, where $U \subset \mathbb{R}^n$ is an open ball and the diameter of $U$ is at most $\delta/2$.

**Proposition 8.107** *Let $G$ be a discrete subgroup of $\mathbb{R}^n$. Then $\mathcal{A} = \mathcal{A}(\mathbb{R}^n, G)$ is an atlas on $\mathbb{R}^n/G$, and $(\mathbb{R}^n/G, \mathcal{A})$ is a manifold. The quotient map $\pi_G : \mathbb{R}^n \to \mathbb{R}^n/G$ is continuous.*

***Proof*** For brevity write $\pi$ for $\pi_G$. Let $U$ and $V$ be open subsets of $\mathbb{R}^n$ of diameter at most $\delta/2$. Let $\boldsymbol{u}, \boldsymbol{u}' \in U$, $\boldsymbol{v}, \boldsymbol{v}' \in V$ and suppose that $\pi(\boldsymbol{u}) = \pi(\boldsymbol{v})$ and $\pi(\boldsymbol{u}') = \pi(\boldsymbol{v}')$. Then $\boldsymbol{v} - \boldsymbol{u}$ and $\boldsymbol{v}' - \boldsymbol{u}'$ are in $G$, and since $|\boldsymbol{u} - \boldsymbol{u}'| < \delta/2$ and $|\boldsymbol{v} - \boldsymbol{v}'| < \delta/2$, $|(\boldsymbol{v} - \boldsymbol{u}) - (\boldsymbol{v}' - \boldsymbol{u}')| < \delta$, so $\boldsymbol{v} - \boldsymbol{u} = \boldsymbol{v}' - \boldsymbol{u}'$. In other words, there is some $\boldsymbol{a} \in G$ such that for all $\boldsymbol{u} \in U$ and $\boldsymbol{v} \in V$, if $\pi(\boldsymbol{u}) = \pi(\boldsymbol{v})$ then $\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{a}$. (If $\pi[U]$ and $\pi[V]$ are disjoint then any $\boldsymbol{a} \in G$ would do, vacuously). So the transition map $\psi_V \circ \psi_U^{-1} = \psi_V \circ (\pi \restriction_U)$ is the restriction of the map $\boldsymbol{x} \mapsto \boldsymbol{x} + \boldsymbol{a}$ to the open set $U \cap (V - \boldsymbol{a}) = \{\boldsymbol{u} \in U : \boldsymbol{u} + \boldsymbol{a} \in V\}$, and its range is the open set $V \cap (U + \boldsymbol{a})$. Thus any two charts $\psi_U$ and $\psi_V$ are compatible. So $\mathcal{A}$ is an atlas on $\mathbb{R}^n/G$.

To show that $(\mathbb{R}^n/G, \mathcal{A})$ is a manifold we need to show it has a countable basis for its topology and that it satisfies the Hausdorff property. For the first we take rational balls. That is, we let $\mathcal{U}$ consists of the sets $\pi[B]$ where $B$ is a rational open ball (Exercise 8.3) of diameter at most $\delta/2$; this is a countable basis. For the Hausdorff property, let $A$ and $B$ be two elements of $\mathbb{R}^n/G$ (two cosets of $G$), and suppose that these cosets are distinct. Then there is some $r > 0$ such that $d(\boldsymbol{a}, \boldsymbol{b}) \geqslant r$ for all $\boldsymbol{a} \in A$ and $\boldsymbol{b} \in B$. Let $U$ be an open ball of radius smaller than $r/2$ around a point $\boldsymbol{a} \in A$, and let $V$ be an open ball of radius smaller than $r/2$ around a point $\boldsymbol{b} \in B$. Then $\pi[U]$ and $\pi[V]$ are disjoint open neighbourhoods of $A$ and $B$ in $\mathbb{R}^n/G$.

Finally, to see that $\pi$ is continuous, let $\langle \boldsymbol{a}_n \rangle$ be a sequence of points in $\mathbb{R}^n$ converging to $\boldsymbol{a}$; a tail of this sequence is contained in an open set $U$ of diameter at most $\delta/2$; since $\pi \restriction_U$ is a homeomorphism it follows that $\langle \pi(\boldsymbol{a}_n) \rangle$ converges to $\pi(\boldsymbol{a})$. □

*Exercise 8.108* Proposition 8.107 shows that $\mathbb{R}/2\pi\mathbb{Z}$ is a manifold. Show that the map $t \mapsto e^{it}$ induces a homeomorphism between $\mathbb{R}/2\pi\mathbb{Z}$ and the unit circle $S$. (See Chap. 1 and Exercise 2.48. You may assume that $t \mapsto e^{it}$ is continuous.) «

When discussing quotient groups in Chap. 2, we have not mentioned a correspondence between subgroups of the quotient $G/H$ and intermediate subgroups $H \subseteq K \subseteq G$ (as we will not need it). Another way of stating this correspondence is that a group homomorphism $\psi : G \to K$ induces a group homomorphism $\bar{\psi} : G/H \to K$ if and only if $H \subseteq \ker \psi$ if and only if $\psi$ is *H-invariant*: if $a + H = b + H$ then $\psi(a) = \psi(b)$. In other words, if $\psi$ is constant on each coset of $H$. In other words, if we can "filter" $\psi$ through the quotient map $\pi_H : G \to G/H$: we have $\psi = \bar{\psi} \circ \pi_H$ for a group homomorphism $\bar{\psi}$ on $G/H$.

The following is the topological version of this fact.

**Proposition 8.109** *Let $G$ be a discrete subgroup of $\mathbb{R}^n$, and let $X$ be a quasi-Euclidean space. A function $f : \mathbb{R}^n/G \to X$ is continuous if and only if the function $f \circ \pi_G : \mathbb{R}^n \to X$ is continuous.*

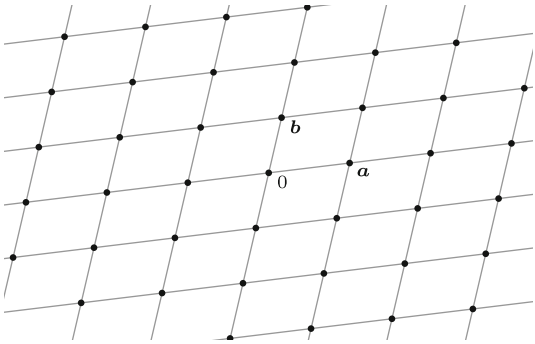



**Fig. 8.2** The lattice $\Gamma$ generated by points $\boldsymbol{a}$ and $\boldsymbol{b}$

***Proof*** If $f$ is continuous then $f \circ \pi_G$ is the composition of continuous functions and so is continuous. Suppose that $f \circ \pi_G$ is continuous. Let $\psi = (\pi_G \restriction_U)^{-1}$ be a chart for $\mathbb{R}^n/G$. Then $(f \circ \pi_G) \circ \psi$ is the composition of continuous functions, so is continuous; and it is the restriction of $f$ to $\operatorname{dom} \psi = \pi_G[U]$, so $f$ is continuous on $\operatorname{dom} \psi$. Since the domains of charts cover $\mathbb{R}^n/G$, $f$ is continuous. □

**The Torus**

A special case is a torus. Let $\Gamma$ be a 2-dimensional discrete subgroup of $\mathbb{R}^2$: it is generated by a pair of linearly independent points (see Fig. 8.2). We call such a subgroup a *lattice* in $\mathbb{R}^2$.

We let

$$T_\Gamma = \mathbb{R}^2/\Gamma$$

be the quotient (quotient group and quotient space); so $\pi_\Gamma : \mathbb{R}^2 \to T_\Gamma$ is the quotient map.

As in the proof of Proposition 8.106, if $\Gamma = \langle \boldsymbol{a}, \boldsymbol{b} \rangle$ is the lattice generated by $\boldsymbol{a}$ and $\boldsymbol{b}$, then the closed parallelogram determined by $\boldsymbol{a}$ and $\boldsymbol{b}$ is $P = P_{\boldsymbol{a},\boldsymbol{b}} = \{s\boldsymbol{a}+t\boldsymbol{b} : s, t \in [0, 1]\}$. We know that $P$ is compact, and that $\pi_\Gamma \restriction_P$ is onto $T_\Gamma$. Thus:

**Proposition 8.110** *The torus $T_\Gamma$ is compact.*

Topologically, we can view the quotient as gluing opposite sides of $P$ (see Fig. 1.2). In this process, the end-points of each side are identified, creating a circle, and the torus is the product of the resulting two circles:

*Exercise 8.111* Extending Exercise 8.108, show that a torus $T_\Gamma$ is homeomorphic to the product $S \times S$. «

In our usual picture of the torus as a donut, one of these circles is a cross-section of the torus; the other becomes a circle "running along" the circumference of the

torus. Some stretching is necessary to embed the torus into 3-dimensional space. The product $S \times S$ is the "pure" torus for which no stretching is necessary; the "inner circumference" and "outer circumference" both have the same length; we need 4 dimensions to realise this.

*Remark 8.112* Below, we will identify $\mathbb{R}^2$ with $\mathbb{C}$ (and note that the addition operations are the same). Note that two complex numbers $\alpha$ and $\beta$ are linearly independent over $\mathbb{R}$ (when considered as points in $\mathbb{R}^2$) if and only if $\beta/\alpha \notin \mathbb{R}$. «

**Topological Groups**

A *topological group* is a group $H$ which is also a quasi-Euclidean space, such that the group operation is a continuous map from $H^2$ to $H$, and the group inverse function is a continuous map from $H$ to itself.

*Exercise 8.113* Show that if $G$ is a discrete subgroup of $\mathbb{R}^n$ then $\mathbb{R}^n/G$ is a topological group. «

Two topological groups are *topologically isomorphic* if there is a group isomorphism $f : G \to H$ which is also a homeomorphism.

*Exercise 8.114* (a) Show that the unit circle $S$, considered as a subgroup of $\mathbb{C}$, is a topological group. (b) Extending Exercise 8.108, show that $S$ is topologically isomorphic to $\mathbb{R}/2\pi\mathbb{Z}$. «

*Exercise 8.115* Show that if $G$ and $H$ are topological groups, then so is $G \times H$. «

*Exercise 8.116* Extending Exercise 8.111, show that a torus $T_\Gamma$ and $S \times S$ are topologically isomorphic. «

*Exercise 8.117* Let $G$ be a topological group. Show that every open subgroup of $G$ is also closed. «

## 8.5 Further Exercises

8.118 For $\boldsymbol{a} = (a_1, \ldots, a_n)$ and $\boldsymbol{b} = (b_1, \ldots, b_n) \in \mathbb{R}^n$ let $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \boldsymbol{a} \cdot \boldsymbol{b}^{\mathsf{t}} = a_1 b_1 + \cdots + a_n b_n$. (a) Prove the *Cauchy-Schwarz inequality*: $|\langle \boldsymbol{a}, \boldsymbol{b} \rangle| \leqslant |\boldsymbol{a}| \cdot |\boldsymbol{b}|$. (Hint: consider the sum of squares $\sum_{i,j \leqslant n} (a_i b_j - a_j b_i)^2$, which is nonnegative.) (b) Conclude that $|\boldsymbol{a} + \boldsymbol{b}| \leqslant |\boldsymbol{a}| + |\boldsymbol{b}|$. (c) Derive the triangle inequality for the Euclidean distance in $\mathbb{R}^n$.

8.119 Again let $\boldsymbol{a} = (a_1, \ldots, a_n)$ and $\boldsymbol{b} = (b_1, \ldots, b_n) \in \mathbb{R}^n$. (a) Use the hint in Exercise 8.118 to show that $|\langle \boldsymbol{a}, \boldsymbol{b} \rangle| = |\boldsymbol{a}| \cdot |\boldsymbol{b}|$ if and only if $\boldsymbol{a}$ and $\boldsymbol{b}$ are linearly dependent over $\mathbb{R}$. (b) Conclude that if $\boldsymbol{a}$ and $\boldsymbol{b}$ are linearly independent then $|\boldsymbol{a} + \boldsymbol{b}| < |\boldsymbol{a}| + |\boldsymbol{b}|$.

8.120 (a) Let $c > 0$ be a real number. Show that if $x \in \mathbb{R}$ and $x^2 < c$, then there is some $y > x$ such that $y^2 < c$. Similarly show that if $x^2 > c$ then there is some $y < x$ such that $y^2 > c$. (b) Use the completeness of $\mathbb{R}$ to show that every positive real number has a square root. (c) Show that the square root function is continuous on $[0, \infty)$.

8.121 Define a function $f : \mathbb{R} \to \mathbb{R}$ by letting $f(x) = 0$ if $x$ is irrational, and $f(x) = 1/m$ if $x$ is rational and $x = n/m$ in lowest terms. Show that $f$ is continuous at the irrational numbers and discontinuous at the rational numbers.[9]

8.122 Show that if $q \in (-1, 1)$ then $\lim_n q^n = 0$. (One way to do this: assume that $q > 0$. Show that $\langle q^n \rangle$ is monotone, and so $a = \lim_n q^n$ exists; show that $qa = a$.)

**Manifolds and Counter-Examples**

8.123 Let $n \geqslant 1$ and let $k \leqslant n$. (a) Show that any $k$-dimensional subspace of $\mathbb{A}^n(\mathbb{R})$ is homeomorphic to $\mathbb{R}^k$. (b) Show that a $k$-dimensional subspace of $\mathbb{P}^n(\mathbb{R})$ (Definition 4.10) is homeomorphic to $\mathbb{P}^k(\mathbb{R})$. (c) Show that the same holds when $\mathbb{R}$ is replaced by $\mathbb{C}$.

8.124 Let $M \subset \mathbb{R}^2$ be the union of $\{(x, 0) : x < 0\}$ with $\{(x, 1) : x \geqslant 0\}$ and $\{(x, -1) : x \geqslant 0\}$. Define an atlas $\mathcal{A}$ on $M$ as follows. It contains two charts: $\psi_+$ is defined on $(x, 0)$ for $x < 0$ and $(x, 1)$ for $x \geqslant 0$; in both cases $\psi_+(x, j) = x$. $\psi_-$ is defined on $(x, 0)$ for $x < 0$ and $(x, -1)$ for $x \geqslant 0$; again we define $\psi_-(x, j) = x$. (a) Show that $\mathcal{A} = \{\psi_+, \psi_-\}$ is an atlas on $M$. (b) Show that $(M, \mathcal{A})$ has a countable basis for its topology. (c) Show that $(M, \mathcal{A})$ is not a topological subspace of $\mathbb{R}^2$. (d) Show that $(M, \mathcal{A})$ is not Hausdorff.

8.125 Define an atlas $\mathcal{A}$ on $\mathbb{R}^2$ as follows. For every $x \in \mathbb{R}$, a chart $\psi_x \in \mathcal{A}$ is defined by letting $\psi_x(x, y) = y$ for all $y \in \mathbb{R}$. (a) Show that $\mathcal{A}$ is an atlas on $\mathbb{R}^2$. (b) Show that $(\mathbb{R}^2, \mathcal{A})$ is Hausdorff. (c) Show that $(\mathbb{R}^2, \mathcal{A})$ does not have a countable basis for its topology.

8.126 Let $\mathcal{A}$ be an atlas on a set $M$. Show that $(M, \mathcal{A})$ is Hausdorff if and only if the diagonal $\Delta = \{(x, x) : x \in M\}$ is an $(\mathcal{A} \times \mathcal{A})$-closed subset of $M \times M$. (See Example 8.47.)

[9] There is no function which is continuous exactly on the rational numbers; see Exercise 13.75.

8.127 (a) Show that every open ball in $\mathbb{R}^n$ is homeomorphic to $\mathbb{R}^n$. (b) Let $M$ be an $n$-dimensional manifold. Show that every point in $M$ has a neighbourhood which is homeomorphic to $\mathbb{R}^n$.

**Compactness**

8.128 Let $E = \{0\} \cup \{1/n : n \in \mathbb{N}\}$. Show directly from the definition (Definition 8.63) that $E$ is compact (as a subset of $\mathbb{R}$).

8.129 Let $A, B$ be disjoint, compact subsets of a quasi-Euclidean space $X$. Show that there are disjoint open subsets $U, V$ of $X$ such that $A \subseteq U$ and $B \subseteq V$.[10]

8.130 (a) Show that every sequence of real numbers has a monotone subsequence (a non-decreasing subsequence, or a non-increasing subsequence). (b) Give an alternative proof of Proposition 8.89 in the case $n = 1$. (c) Show that the general case $n > 1$ follows from the case $n = 1$.

8.131 Suppose that $A \subseteq \mathbb{R}$ is not compact. Show that there is a continuous and unbounded function (a function whose image is unbounded) $f : A \to \mathbb{R}$.

8.132 (a) Show that the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ is continuous but not uniformly continuous. (b) Find an example of a bounded function on $\mathbb{R}$ which is continuous but not uniformly continuous.

8.133 Use Exercise 8.111 to give an alternative proof that the torus is compact.

**Discrete Sets and Groups**

8.134 Give an example of a discrete subset of $\mathbb{R}$ which is not closed.

8.135 For which $\alpha$ and $\beta$ in $\mathbb{R}$ is the subgroup $\langle \alpha, \beta \rangle_{\mathbb{R}}$ discrete?

8.136 Let $G$ be a discrete subgroup of $\mathbb{R}^n$. Show that $\mathbb{R}^n / G$ is compact if and only if $G$ is $n$-dimensional.

8.137 We sketch a variant of the proof of Proposition 8.106 in the case $n = 2$. Let $G$ be a discrete subgroup of $\mathbb{R}^2$. (a) Show that there is some $\boldsymbol{a} \in G$ such that $|\boldsymbol{a}| = \min\{|\boldsymbol{g}| : \boldsymbol{g} \in G \setminus \{\boldsymbol{0}\}\}$. (b) Picking such $\boldsymbol{a}$, show that $G \cap \mathbb{R}\boldsymbol{a} = \mathbb{Z}\boldsymbol{a}$. (c) Suppose that $G \neq \mathbb{Z}\boldsymbol{a}$. Show that there is some $\boldsymbol{b} \in G$ such that $|\boldsymbol{b}| = \min\{|\boldsymbol{g}| : \boldsymbol{g} \in G \setminus \mathbb{R}\boldsymbol{a}\}$. (d) Picking such $\boldsymbol{b}$, suppose that $G \neq \langle \boldsymbol{a}, \boldsymbol{b} \rangle$. Show that there is some $\boldsymbol{w} \in G \setminus \langle \boldsymbol{a}, \boldsymbol{b} \rangle$

[10] In topological language, this says that a compact, Hausdorff space is *normal*.

such that $\boldsymbol{w} = \alpha\boldsymbol{a} + \beta\boldsymbol{b}$ with $0 < \alpha, \beta \leqslant 1/2$. (e) Use Exercise 8.119 to show that $|\boldsymbol{w}| < |\boldsymbol{b}|$, contradicting the choice of $\boldsymbol{b}$.

**Topological Groups**

8.138 Let $H = \mathbb{Z} \times \{0\}$ considered as a subgroup of $\mathbb{R}^2$. (a) Show that $\mathbb{R}^2/H$ is topologically isomorphic to the cylinder $S \times \mathbb{R}$. (b) Show that the cylinder is homeomorphic to $\mathbb{R}^2 \setminus \{\mathbf{0}\}$. (Hint: consider the map $\boldsymbol{p} \mapsto (|\boldsymbol{p}|, \boldsymbol{p}/|\boldsymbol{p}|)$.)

8.139 Let $n \geqslant 2$. Show that the determinant function $\det\colon M_n(\mathbb{R}) \to \mathbb{R}$ is continuous (recall that $M_n(\mathbb{R})$ is identified with $\mathbb{R}^{n^2}$ so is an $n^2$-manifold).

8.140 We consider the general linear group $\mathrm{GL}_n(\mathbb{R})$ as a subspace of $M_n(\mathbb{R})$, and so is a quasi-Euclidean space. (a) Show that $\mathrm{GL}_n(\mathbb{R})$ is an open subset of $M_n(\mathbb{R})$. (b) Show that $\mathrm{GL}_n(\mathbb{R})$ is a topological group. (Hint: for the inverse, consider the adjugate matrix, see page 48.)

8.141 Let $\mathrm{O}_n(\mathbb{R}) \subset \mathrm{GL}_n(\mathbb{R})$ be the collection of orthogonal matrices: those such that $A^{\mathsf{t}}A = I$. Show that $\mathrm{O}_n(\mathbb{R})$ is compact. (Hint: if $A$ is orthogonal then $|A\boldsymbol{x}| = |\boldsymbol{x}|$ for all $\boldsymbol{x}$.)

8.142 Show that $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$ with multiplication of complex numbers is a topological group.

8.143 Let $G$ be a topological group. Show that the closure of a subgroup of $G$ is also a subgroup of $G$.


# Connectedness, Smooth and Simple

# 9

A quasi-Euclidean space is *connected* if it does not consist of two or more pieces which are separated from each other. This resembles the notion of irreducible algebraic curves. In this chapter we investigate connectedness, and the somewhat stronger notion of being *path-connected*, meaning any two points are connected by a continuous path. A third notion is that of a space being *simply connected*, which very roughly, means that it has "no holes". For example, the plane and a disc are simply connected, but the punctured plane $\mathbb{R}^2 \setminus \{\mathbf{0}\}$, the unit circle and the torus are not. Simply connected Riemann surfaces play an important role in the theory of elliptic functions.

The formal definition of simple connectedness has to do with deforming paths, a process formally called path *homotopy*. In some instances it will be very convenient to work with *smooth* paths and homotopies, for example, in the next chapter, in which we consider integration along paths (which is in turn a necessary tool for developing complex analysis). Smoothness, for example, ensures that integrals are well-defined and that paths have finite lengths.

For this purpose we will introduce differentiable manifolds and talk about smooth functions on manifolds. As hinted previously, what we need is for the transition functions to be differentiable. In Sect. 9.3 will review the basics of calculus in more than one dimension, and then we will define differentiable manifolds. The process of "smoothening" paths and homotopies requires smooth *partitions of unity*, which we discuss in Sect. 9.5. This process will allow us, for example, to show that for verifying that a subset of $\mathbb{R}^n$ is simply connected, it suffices to consider only smooth paths.

***Terminology*** *In this book we call a function which is continuously differentiable at every point* ***smooth****. The term is often taken to mean a stronger condition: being differentiable infinitely many times. For many of the results mentioning smoothness, we could take either definition, but we will not verify this.*



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_9

## 9.1 Connectedness, Path and Simple

A quasi-Euclidean space $X$ is called *connected* if we cannot partition it into two nonempty open sets; equivalently if the only subsets of $X$ which are both closed and open are the empty set and $X$ itself.

*Exercise 9.1* Show that the continuous image of a connected space is connected. (In other words, if $X$ is a connected quasi-Euclidean space and $f : X \to Y$ is continuous and onto $Y$, then $Y$ is connected. Compare with Proposition 8.64.) «

*Exercise 9.2* Show that if $X$ is a connected space, $Y$ is a discrete space (Definition 8.99) and $f : X \to Y$ is continuous, then $f$ is constant. «

Recall that a *closed interval* (in the real line) is a subset

$$[a, b] = \{x \in \mathbb{R} : a \leqslant x \leqslant b\}$$

of $\mathbb{R}$, where $a < b$.

**Proposition 9.3** *Every closed interval is connected.*

***Proof*** Let $I = [a, b]$ be a closed interval. Suppose that $I = A \cup B$ with $A$ and $B$ nonempty, disjoint and open. Without loss of generality $a \in A$. Let $c = \inf B$ be the greatest lower bound of $B$ (see page 209). Since $B$ is closed, $c \in B$, so $c > a$. Since $[a, c) \subseteq A$ and $A$ is closed, $c \in A$, a contradiction. □

*Exercise 9.4* Prove the intermediate value theorem: if $I = [a, b]$ is a closed interval and $f : I \to \mathbb{R}$ is continuous then $[f(a), f(b)] \subseteq f[I]$. «

**Definition 9.5** Let $X$ be a quasi-Euclidean space. A *path* in $X$ is a continuous function from a closed interval to $X$.

There is an important physical intuition here. A path $\gamma : I \to X$ traces the movement of a body (say a particle) in the space $X$. The input for $\gamma$ represents time. Say $I = [a, b]$. The movement begins at time $t = a$ (and the particle starts at the point $\gamma(a)$), and the movement concludes at time $t = b$ and at the point $\gamma(b)$. Continuity means that the particle cannot jump instantaneously from one point to another. We say that $\gamma(a)$ and $\gamma(b)$ are the end-points of $\gamma$; alternatively that the path is from $\gamma(a)$ to $\gamma(b)$. An important distinction to make is between the path and its image $\gamma[I]$ (we also write $\gamma[a, b]$): the image tells us the route the particle took, but not how quickly it moved (or in which direction).

**Definition 9.6** A quasi-Euclidean space $X$ is *path-connected* if for any two points $x, y \in X$ there is a path in $X$ from $x$ to $y$.

*Exercise 9.7* Show that the continuous image of a path-connected space is also path-connected. «

*Exercise 9.8* Show that the torus is path-connected. «

**Proposition 9.9** *Every path-connected quasi-Euclidean space is connected.*

***Proof*** Suppose that $X$ is not connected; let $A \subset X$ be both closed and open (but not empty and not all of $X$). Pick a point $a \in A$ and another point $b \in X \setminus A$. There cannot be a path from $a$ to $b$. For suppose that $\gamma : I \to X$ is such a path. Let $Y$ be the image $\gamma[I]$ of the path. The closed interval $I$ is connected, so by Exercise 9.1, $Y$ is connected. But $Y$ is manifestly not connected: $Y \cap A$ is both closed and open in $Y$, and equals neither $Y$ nor the empty set. □

**Concatenation of Paths**

Suppose that $a < b < c$ are real numbers and that $\gamma : [a, b] \to X$ and $\delta : [b, c] \to X$ are paths in $X$ such that $\gamma(b) = \delta(b)$. Then we can naturally define a path $\gamma \hat{} \delta$ from $[a, c] \to X$ by "concatenating" $\gamma$ with $\delta$: travel along $\gamma$ between time $a$ and time $b$, then travel along $\delta$ until time $c$. Formally of course $(\gamma \hat{} \delta)(t)$ is defined as $\gamma(t)$ if $t \in [a, b]$ and as $\delta(t)$ if $t \in [b, c]$. It is not difficult to formally show that $\gamma \hat{} \delta$ is continuous (exercise!).

Actually it is not that important that $b$, the right end-point of the domain of $\gamma$, equals the left end-point of the domain of $\delta$. We can move the domains of paths about, for example by shifting; if $\gamma$ is a path with domain $[a, b]$ and $d$ is a real number then we can define $\widetilde{\gamma}$ with domain $[a+d, b+d]$ by letting $\widetilde{\gamma}(t) = \gamma(t-d)$. We will consider $\widetilde{\gamma}$ as essentially the same as $\gamma$. Thus we can move the domain of $\gamma$ so that it ends at the start of the domain of $\delta$.

In general, the notion of a *re-parameterisation* of a path allows us to not only shift the domain but also, for example, stretch it; see Exercise 9.108.

There are connected spaces which are not path-connected (see Exercise 9.89 for an example.) However we do obtain a restricted converse to Proposition 9.9.

**Proposition 9.10** *A manifold is connected if and only if it is path-connected.*

***Proof*** Let $M$ be a connected manifold; we show that it is path-connected. For $a, b \in M$ let $a \sim b$ if there is a path in $M$ from $a$ to $b$. Concatenation of paths shows that this relation is an equivalence relation on the points of $M$. We need to show there is only one equivalence class.

We show that every equivalence class is open. For let $a \in M$; let $\psi$ be a chart for $M$ with $a \in \operatorname{dom} \psi$; find some small positive $\varepsilon$ such that $B(\psi(a), \varepsilon) \subseteq \operatorname{range} \psi$. For every point $\boldsymbol{c} \in B(\psi(a), \varepsilon)$, the linear path from $\psi(a)$ to $\boldsymbol{c}$ is contained in $B(\psi(a), \varepsilon)$. Composing with $\psi^{-1}$ we see that $\psi^{-1}[B(\psi(a), \varepsilon)]$ is an open neighbourhood of $a$ in $M$ which is a subset of the equivalence class of $a$.

Let $A$ be a $\sim$-equivalence class. Then $M \setminus A$ is the union (possibly empty) of equivalence classes, hence the union of open sets, and so is open. This shows that $A$ is both closed and open in $M$. Since $M$ is connected it follows that $A = M$, as required. □

### 9.1.1 Homotopy; Simple Connectedness

Two paths are "similar" if we can continuously deform one to another. We formalise it as follows. If $H : [0, 1] \times [a, b] \to X$ is a function then for all $s \in [0, 1]$ we define $H_s : [a, b] \to X$ by $H_s(t) = H(s, t)$. If $H$ is continuous then for each $s$, $H_s$ is continuous, i.e. is a path in $X$. We think of $H$ as a continuous deformation from $H_0$ to $H_1$: at time $s$ the path $H_0$ has been deformed into $H_s$.

**Definition 9.11** Let $X$ be a quasi-Euclidean space. A *path homotopy* in $X$ is a continuous function $H : [0, 1] \times [a, b] \to X$ such that for all $s \in [0, 1]$, $H_s(a) = H_0(a)$ and $H_s(b) = H_0(b)$.

That is, not only is $H_0$ deformed into $H_1$, but the deformation fixes the endpoints. We say that $H$ is a path homotopy from $H_0$ to $H_1$. We say that two paths are homotopic if there is a path homotopy between them. Below, we will usually just say "homotopy" rather than "path homotopy"; there are other notions of homotopy, but we will not use them.

*Exercise 9.12* Let $X$ be a quasi-Euclidean space and let $I$ be a closed interval. Show that homotopy is an equivalence relation on the collection of all paths $\gamma : I \to X$ from some point $a$ to some point $b$. «

**Definition 9.13** A path-connected quasi-Euclidean space $X$ is *simply connected* if for every $a, b \in X$, any two paths from $a$ to $b$ (with the same domain) are homotopic.

Simple connectedness is a much stronger property than path-connectedness. Speaking informally, all "reasonable" spaces are path-connected; spaces which are not are exotic, or consist of disjoint pieces. However many reasonable spaces are *not* simply connected, for example the unit circle, or the torus. Simply connected spaces are those which have no "holes", though this intuition may be a bit misleading. For example, as we will shortly see, a sphere is simply connected.

In the case of the circle, it is intuitively clear that the path which travels once round the circle is not homotopic to the constant path; we cannot continuously deform the former to the latter while staying within the circle. We will prove this in the next section.

The following exercise shows that for measuring simple connectedness, we can restrict our attention to loops. A path is called a *loop* (or sometimes, confusingly, "closed") if it starts and ends at the same point ($\gamma(a) = \gamma(b)$). The following characterisation is usually given as a definition of simple connectedness.

*Exercise 9.14* Let $X$ be a path-connected space. Show that $X$ is simply connected if and only if every loop in $X$ is homotopic to a constant path. «

*Example 9.15* A subset $E \subseteq \mathbb{R}^n$ is called *convex* if for all $\boldsymbol{a}, \boldsymbol{b} \in E$, the straight line segment $\{(1-s)\boldsymbol{a} + s\boldsymbol{b} : s \in [0,1]\}$ from $\boldsymbol{a}$ to $\boldsymbol{b}$ is a subset of $E$.

Immediately, a convex subset of $\mathbb{R}^n$ is path-connected. In fact every convex subset of $\mathbb{R}^n$ is simply connected. Two paths $\gamma, \delta : I \to X$ with the same end-points are homotopic by the linear homotopy defined by

$$H_s(t) = (1-s) \cdot \gamma(t) + s \cdot \delta(t).$$

(Show that $H$ is indeed continuous.)

For example, an open ball in $\mathbb{R}^n$, and $\mathbb{R}^n$ itself, are convex, and so simply connected. «

Here is another example. Recall that $S^2$, the unit sphere in $\mathbb{R}^3$, is homeomorphic to the Riemann sphere $\mathbb{P}^1(\mathbb{C})$ (Exercise 8.56).

**Proposition 9.16** *The sphere $\mathbb{P}^1(\mathbb{C})$ is simply connected.*

***Proof*** Recall that $U_0 = \mathbb{P}^1(\mathbb{C}) \setminus \{(0:1)\}$ and $U_1 = \mathbb{P}^1(\mathbb{C}) \setminus \{(1:0)\}$ are both homeomorphic to $\mathbb{R}^2$ (via the maps $\rho_0$ and $\rho_1$), and so are simply connected. Also, $U_0 \cap U_1$ is homeomorphic to the punctured plane $\mathbb{R}^2 \setminus \{\mathbf{0}\}$ (via either map), and so is path-connected.

Let $\gamma : I \to \mathbb{P}^1(\mathbb{C})$ be a loop; we show that it is homotopic to a constant loop. Let $p_0$ be the start- and end-point of $\gamma$. Since changes of coordinates of $\mathbb{P}^1(\mathbb{C})$ are homeomorphisms (Exercise 8.60), we may assume that $p_0 \in U_0 \cap U_1$.

Since $\gamma$ is continuous, each $t \in [a,b]$ has a neighbourhood $V$ in $I$ (which we can take to be an interval) such that $\gamma[V] \subseteq U_0$ or $\gamma[V] \subseteq U_1$. Since $I$ is compact, there is a finite covering of $I$ of such neighbourhoods. Hence, we can break up the domain $I = [a,b]$ of $\gamma$ into finite intervals which $\gamma$ maps into either $U_0$ or $U_1$. That is, we can find $a = t_0 < t_1 < \cdots < t_k = b$ so that for each $i = 1, 2, \ldots, k$, the image $\gamma[t_{i-1}, t_i]$ is contained in $U_0$ or contained in $U_1$. By taking the unions of successive sub-intervals, we may assume that for each $i = 0, 1, \ldots, k$, $\gamma(t_i) \in U_0 \cap U_1$.

Now for each $i$, we can find a path $\eta_i$ (with domain $[t_{i-1}, t_i]$) in $U_0 \cap U_1$ from $\gamma(t_{i-1})$ to $\gamma(t_i)$. And since both $\eta_i$ and $\gamma\restriction_{[t_{i-1},t_i]}$ are contained in either $U_0$ or $U_1$, and either is simply connected, the paths $\eta_i$ and $\gamma\restriction_{[t_{i-1},t_i]}$ are homotopic. Letting $\eta$ be the concatenation $\eta_1 \hat{} \eta_2 \hat{} \cdots \hat{} \eta_k$, and concatenating the homotopies, we see that $\gamma$ and $\eta$ are homotopic in $\mathbb{P}^1(\mathbb{C})$. Since $\eta$ is a path in $U_0 \cap U_1$, it is a path in $U_0$, and so homotopic in $U_0$, and hence in $\mathbb{P}^1(\mathbb{C})$, to a constant loop. □

*Remark 9.17* One could wonder if we worked too hard to prove that the sphere is simply connected. For let $\gamma$ be a loop in $\mathbb{P}^1(\mathbb{C})$. We can then take any point $p \in \mathbb{P}^1(\mathbb{C})$ which is not in the image of $\gamma$, and find a homotopy between $\gamma$ and a constant loop in $\mathbb{P}^1(\mathbb{C})\backslash\{p\}$, which is homeomorphic to $\mathbb{R}^2$ and so simply connected. This is a reasonable plan, except that there are some pathological loops whose image is all of $\mathbb{P}^1(\mathbb{C})$! (see the related Exercise 10.53). Thus, an alternative proof of Proposition 9.16 is to first show that every path in the sphere is homotopic to a path which does not fill all of the sphere. This can be done using Exercise 8.85. «

## 9.2 Lifting Maps

Let $G$ be a discrete subgroup of $\mathbb{R}^n$ and let $X$ be a quasi-Euclidean space.

**Definition 9.18** A *lifting* of a continuous map $f\colon X \to \mathbb{R}^n/G$ is a continuous map $F\colon X \to \mathbb{R}^n$ satisfying $\pi_G \circ F = f$, where $\pi_G\colon \mathbb{R}^n \to \mathbb{R}^n/G$ is the quotient map.

A lifting is a "continuous choice of representatives": since the elements of $\mathbb{R}^n/G$ are cosets of $G$, a lifting of $f$ is a continuous map which for every $x \in X$, chooses a point in the coset $f(x)$.

**Lemma 9.19** *Suppose that $X$ is connected; suppose that $F_0$ and $F_1$ are both liftings of a continuous map $f\colon X \to \mathbb{R}^n/G$. Then there is some fixed $\boldsymbol{g} \in G$ such that $F_1(x) - F_0(x) = \boldsymbol{g}$ for all $x \in X$.*

***Proof*** For every $x \in X$, $F_1(x) - F_0(x) \in G$. The function $F_1 - F_0$ is continuous. Since $X$ is connected and $G$ is discrete, $F_1 - F_0$ must be constant (Exercise 9.2). □

**Proposition 9.20** *Every path in $\mathbb{R}^n/G$ has a lifting.*

***Proof*** Let $\gamma\colon [a,b] \to \mathbb{R}^n/G$ be a path. For every $t \in [a,b]$, $\gamma(t)$ lies in the domain of some chart for $\mathbb{R}^n/G$. Since $[a,b]$ is compact, we can find a finite sequence $a = t_0 < t_1 < \cdots < t_m = b$ such that for each $i = 1,\ldots,m$, $\gamma[t_{i-1},t_i]$ lies in the domain of a chart. We define a lifting $\Gamma(t)$ for $t \in [t_0,t_i]$ by induction on $i$. We start by choosing any chart $\psi_1$ for $\mathbb{R}^n/G$ such that $\gamma[t_0,t_1] \subseteq \operatorname{dom}\psi_1$, and let $\Gamma(t) = \psi_1(\gamma(t))$ for every $t \in [t_0,t_1]$.

Now $\psi_1$ is the inverse of $\pi_G \restriction_{U_1}$ for some small open ball $U_1$ in $\mathbb{R}^n$. Also, $\gamma[t_1,t_2] \subseteq \operatorname{dom}\varphi$ where $\varphi$ is some chart for $\mathbb{R}^n/G$; $\varphi$ is the inverse of $\pi_G \restriction_V$ for some small open ball $V$. Let $\boldsymbol{g} = \varphi(\gamma(t_1)) - \psi_1(\gamma(t_1))$; then $\boldsymbol{g} \in G$. Let $U_2 = V - \boldsymbol{g}$ and let $\psi_2 = (\pi_G \restriction_{U_2})^{-1}$. Then $\psi_2$ is a chart for $\mathbb{R}^n/G$, $\operatorname{dom}\psi_2 = \operatorname{dom}\varphi$ (so $\gamma[t_1,t_2] \subseteq \operatorname{dom}\psi_2$), and $\psi_1(\gamma(t_1)) = \psi_2(\gamma(t_1))$. So we can let $\Gamma(t) = \psi_2(\gamma(t))$ for all $t \in [t_1,t_2]$; we ensured that $\Gamma$ is continuous at $t_1$. This determines $\Gamma(t_2)$; we continue this process for $m$ steps. □

Suppose that $\gamma : [a, b] \to \mathbb{R}^n/G$ is a path. Lemma 9.19 and Proposition 9.20, together with the ability to add a fixed element of $G$, show that for every point $\boldsymbol{p} \in \mathbb{R}^n$ in the coset $\gamma(a)$—any point $\boldsymbol{p}$ satisfying $\pi_G(\boldsymbol{p}) = \gamma(a)$—there is a *unique* lifting $\Gamma$ of $\gamma$ which starts at $\boldsymbol{p}$ (i.e., $\Gamma(a) = \boldsymbol{p}$).

We extend Proposition 9.20 by one dimension.

**Proposition 9.21** *If $C \subseteq \mathbb{R}^2$ is a closed rectangle, then any continuous map $f : C \to \mathbb{R}^n/G$ has a lifting.*

***Proof*** Let $C = [c, d] \times [a, b]$. Let $f^a(s) = f(s, a)$ for $s \in [c, d]$; then $f^a$ is a path in $\mathbb{R}^n/G$, and so we can fix a lifting $\gamma$ of $f^a$. For each $s \in [c, d]$, recall that we let $f_s(t) = f(s, t)$; let $\eta_s$ be the unique lifting of $f_s$ starting at $\gamma(s)$.

The uniqueness of liftings of paths shows that after choosing a starting point for $\gamma$, we *must* define $F(s, t) = \eta_s(t)$. But we need to show that this is continuous; it suffices to show that for each $s$ and $t$, if $s'$ is close to $s$ then $\eta_{s'}(t)$ is close to $\eta_s(t)$. We go about it in a slightly roundabout way.

Fix $s \in [c, d]$. Every $t \in [a, b]$ has a neighbourhood on which $\eta_s = \psi \circ f_s$ for some chart $\psi$; and there is a neighbourhood $A$ of $(s, t)$ in $[c, d] \times [a, b]$ such that $f[A] \subseteq \operatorname{dom} \psi$. By compactness of $[a, b]$, we can obtain a finite sequence $a = t_0 < t_1 < \cdots < t_m = b$, charts $\psi_1, \psi_2, \ldots, \psi_m$ and some $\delta > 0$ such that: (i) $\eta_s = \psi_i \circ f_s$ on $[t_{i-1}, t_i]$; and (ii) for all $u \in [c, d]$ with $|u - s| < \delta$ and all $t \in [t_{i-1}, t_i]$, $f(u, t) \in \operatorname{dom} \psi_i$. Further note that by the definition of the charts, $\psi_i(f_s(t_i)) = \psi_{i+1}(f_s(t_i))$ implies that $\psi_i = \psi_{i+1}$ on some neighbourhood of $f(s, t_i) = f_s(t_i)$; so by shrinking $\delta$, we may assume that $\psi_i(f(s', t_i)) = \psi_{i+1}(f(s, t_i))$ whenever $|s' - s| < \delta$. Further, we may also assume that $\gamma(s') = \psi_1(f(s', a))$ whenever $|s' - s| < \delta$. Thus, the map $h(s', t) = \psi_i(f(s', t))$ for $|s' - s| < \delta$ and $t \in [t_{i-1}, t_i]$ is continuous. And for each $s'$ within $\delta$ of $s$, $t \mapsto h(s', t)$ is a lifting of $f_{s'}$ starting at $\gamma(s')$. Uniqueness of this lifting shows that $h(s', t) = \eta_{s'}(t)$, i.e., $F = h$ for such points; since $h$ is continuous, so is $F$. □

*Remark 9.22* Just like paths, Lemma 9.19 implies that if $C$ is a closed rectangle, $x \in C$ and $f : C \to \mathbb{R}^n/G$ is continuous and $\pi_G(\boldsymbol{p}) = f(x)$, then there is a unique lifting $F$ of $f$ satisfying $F(x) = \boldsymbol{p}$. «

*Remark 9.23* Suppose that $h : [0, 1] \times [a, b] \to \mathbb{R}^n/G$ is a path homotopy: the maps $h^a(s) = h(s, a)$ and $h^b(s) = h(s, b)$ are constant. Then any lifting $H$ of $h$ is a homotopy as well: this is because $H^a$ (the map $s \mapsto H(s, a)$) is a lifting of $h^a$; uniqueness of liftings of paths implies that $H^a$ must be constant, and similarly for $H^b$. «

**Theorem 9.24** *Suppose that $X$ is a simply connected manifold. Then every continuous map $f : X \to \mathbb{R}^n/G$ has a lifting.*

***Proof*** Let $f: X \to \mathbb{R}^n/G$ be continuous. Choose some $x^* \in X$, and then some $\boldsymbol{p} \in \mathbb{R}^n$ such that $\pi_G(\boldsymbol{p}) = f(x^*)$. Let $x \in X$. Suppose that $\gamma$ and $\lambda$ are two paths in $X$ from $x^*$ to $x$. So $f \circ \gamma$ and $f \circ \lambda$ are two paths in $\mathbb{R}^n/G$ from $f(x^*)$ to $f(x)$. Let $\Gamma$ and $\Lambda$ be the unique liftings of $f \circ \gamma$ and $f \circ \lambda$ both starting at $\boldsymbol{p}$. By assumption, there is a homotopy $h$ in $X$ from $\gamma$ to $\lambda$; so $f \circ h$ is a homotopy in $\mathbb{R}^n/G$ from $f \circ \gamma$ to $f \circ \lambda$; by Remark 9.22, let $H$ be the unique lifting of $f \circ h$ such that $H(0, a) = \boldsymbol{p}$ (where $\gamma, \lambda: [a, b] \to X$). By Remark 9.23, $H$ is a homotopy in $\mathbb{R}^n$ from $\Gamma$ to $\Lambda$. Thus, both $\Gamma$ and $\Lambda$ end at the same point. We can therefore define $F(x)$ by choosing any path $\gamma$ in $X$ from $x^*$ to $x$ and letting $F(x)$ be the end-point of the lifting of $f \circ \gamma$ starting at $\boldsymbol{p}$.

Certainly $\pi_G \circ F = f$; it remains to show that $F$ is continuous. First, we observe that for all $x, y \in X$, for any path $\theta: [a, b] \to X$ from $x$ to $y$, if $\Theta$ is the lifting of $f \circ \theta$ which starts at $F(x)$, then the end-point of $\Theta$ is $F(y)$. To see this fix a path $\gamma$ in $X$ from $x^*$ to $x$, and let $\Gamma$ be the lifting of $\gamma$ starting at $\boldsymbol{p}$. Then $\Gamma \hat{} \Theta$ is the lifting of $f \circ (\gamma \hat{} \theta)$ starting at $\boldsymbol{p}$.

Now let $x \in X$ and let $U$ be a small open ball in $\mathbb{R}^n$ around $F(x)$. Let $V$ be a neighbourhood of $x$ in $X$ such that $f[V] \subseteq \pi_G[U]$, i.e., $f[V]$ is contained in the domain of the chart $\psi = (\pi_G \restriction_U)^{-1}$. Since $X$ is a manifold, it is *locally path-connected*: since locally $X$ looks like an open subset of $\mathbb{R}^m$, we can shrink $V$ so that it is homeomorphic to a ball in $\mathbb{R}^m$; so we may assume that $V$ is path-connected.

Let $y \in V$; fix a path $\theta$ in $V$ from $x$ to $y$. Since $f[V] \subseteq \operatorname{dom} \psi$, $\Theta = \psi \circ f \circ \theta$ is the unique lifting of $f \circ \theta$ which starts at $F(x)$; so the end-point $F(y)$ of $\Theta$ is in $U$. So $F[V] \subseteq U$; hence $F$ is continuous at $x$. □

### 9.2.1 The Winding Number

The winding number of a loop in the punctured plane is the number of times it goes around the puncture. To define it, we make use of the notion of a continuous choice of argument of a nonzero complex number.

Using the isomorphism $t + 2\pi\mathbb{Z} \mapsto e^{it}$ from $\mathbb{R}/2\pi\mathbb{Z}$ to the unit circle (Exercise 8.108), we can transfer the notion of a lifting to maps to the unit circle: a *lifting* of a continuous map $f: X \to S$ is a continuous $F: X \to \mathbb{R}$ such that $f = e^{iF}$.

The punctured plane is $\mathbb{R}^2 \setminus \{\mathbf{0}\}$, which we identify with $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$. An *argument* of a nonzero complex number $z \in \mathbb{C}^*$ is any $t \in \mathbb{R}$ satisfying $z = |z|e^{it}$.

**Definition 9.25** Let $X$ be a quasi-Euclidean space and let $f: X \to \mathbb{C}^*$ be continuous. A *continuous choice of argument* for $f$ is a continuous map $F: X \to \mathbb{R}$ such that $f = |f|e^{iF}$.

So a continuous choice of argument for $f\colon X \to \mathbb{C}^*$ is a lifting of the map $f/|f|$ (observe that $f\colon X \to \mathbb{C}^*$ is continuous if and only if both maps $|f|$ and $f/|f|$ are continuous). Hence Proposition 9.20 implies:

**Proposition 9.26** *There is a continuous choice of argument for every path in* $\mathbb{C}^*$.

Let $\gamma\colon [a,b] \to \mathbb{C}^*$ be a loop. If $\theta$ is a continuous choice of argument for $\gamma$ then $\theta(b) - \theta(a) = 2\pi m$ for some $m \in \mathbb{Z}$. Lemma 9.19 implies that this value $m$ does not depend on the choice of $\theta$, so we define:

**Definition 9.27** The *winding number* of a loop $\gamma\colon [a,b] \to \mathbb{C}^*$ is the unique $m$ such that $\theta(b) - \theta(a) = 2\pi m$ for any continuous choice of argument $\theta$ for $\gamma$.

*Example 9.28* If $\gamma$ is a constant loop, then a continuous choice of argument for $\gamma$ is constant; so the winding number of a constant loop is 0. If $\gamma\colon [0, 2\pi] \to S$ is given by $\gamma(t) = e^{it}$, then the identity map on $[0, 2\pi]$ is a continuous choice of argument for $\gamma$; so the winding number of $\gamma$ is 1. «

**Proposition 9.29** *Let* $\gamma_1, \gamma_2\colon [a,b] \to \mathbb{C}^*$ *be two loops with the same end-point. Then* $\gamma_1$ *and* $\gamma_2$ *are homotopic in* $\mathbb{C}^*$ *if and only if they have the same winding number.*

***Proof*** In one direction, suppose that $\gamma_1$ and $\gamma_2$ are homotopic, say via some homotopy $H\colon [0,1] \times [a,b] \to \mathbb{C}^*$. Then $H/|H|$ is a homotopy between $\gamma_1/|\gamma_1|$ and $\gamma_2/|\gamma_2|$. Proposition 9.21 and Remark 9.23 imply that there is a continuous choice of argument $\Theta\colon [0,1] \times [a,b] \to \mathbb{R}$ for $H$ with the maps $s \mapsto \Theta(s,a)$ and $s \mapsto \Theta(s,b)$ constant; this shows that $\gamma_1$ and $\gamma_2$ have the same winding number.

In the other direction, suppose that $\gamma_1$ and $\gamma_2$ have the same winding number $m$; let $\theta_1$ and $\theta_2$ be continuous choices of argument for $\gamma_1$ and $\gamma_2$, which by shifting we may assume both start at the same point $t_0$; then they both end at the same point $t_0 + 2\pi m$. For $i = 1, 2$, the paths $\Gamma_i = (|\gamma_i|, \theta_i)$ (from $[a,b]$ to $\mathbb{R}^+ \times \mathbb{R}$) are continuous with the same start and end-points. Since $\mathbb{R}^+ \times \mathbb{R}$ is simply connected (it is a convex subset of $\mathbb{R}^2$), there is a homotopy $G$ in $\mathbb{R}^+ \times \mathbb{R}$ from $\Gamma_1$ to $\Gamma_2$. Write $G = (G_r, G_t)$. Then the map $G_r e^{iG_t}$ is a homotopy between $\gamma_1$ and $\gamma_2$. □

If $U \subseteq \mathbb{C}^*$ then a continuous choice of argument *on* $U$ is a continuous choice of argument for the identity function on $U$: a map $\theta\colon U \to \mathbb{R}$ with $z = |z|e^{i\theta(z)}$ for all $z \in U$. Theorem 9.24 implies:

**Proposition 9.30** *If* $U \subset \mathbb{C}^*$ *is open and simply connected then there is a continuous choice of argument on* $U$.

*Example 9.31* Fix $t_0 \in \mathbb{R}$. Let $U = \mathbb{C} \setminus \{re^{it_0} : r \geqslant 0\}$ be the result of removing the infinite ray in direction $t_0$ from the plane. For each $z \in U$ we can choose the unique argument $t \in (t_0, t_0 + 2\pi)$. «

**Proposition 9.32** *The following are equivalent for an open and connected $U \subseteq \mathbb{C}^*$:*

(1) *There is a continuous choice of argument on $U$.*
(2) *The winding number of any loop in $U$ is 0.*

***Proof*** First, suppose that $\alpha$ is a continuous choice of argument on $U$. Let $\gamma : [a, b] \to U$ be a loop. Then $\alpha \circ \gamma$ is a continuous choice of argument for $\gamma$, and $\alpha(\gamma(a)) = \alpha(\gamma(b))$.

In the other direction, suppose that the winding number of any loop in $U$ is 0. The argument is similar to that of Theorem 9.24. Fix $z^* \in U$; for any $z \in U$, let $\alpha(z) = \theta(b) - \theta(a)$ where $\theta$ is a continuous choice of argument for some path in $U$ from $z^*$ to $z$; the assumption implies that the value $\alpha(z)$ does not depend on the choice of the path. We observe that for any $z, w \in U$, $\alpha(w) - \alpha(z) = \theta(b) - \theta(a)$, where $\theta : [a, b] \to \mathbb{R}$ is any continuous choice of argument along any path in $U$ from $z$ to $w$.

To see that $\alpha$ is continuous, let $z \in U$; let $V \subseteq U$ be a simply connected open neighbourhood of $z$ (say a small disc). By Proposition 9.30, let $\beta$ be a continuous choice of argument on $V$; by shifting, we may assume that $\beta(z) = \alpha(z)$. Then $\beta = \alpha$ on $V$: for any $w \in V$, for any path $\gamma$ in $V$ from $z$ to $w$, $\beta \circ \gamma$ is a continuous choice of argument for $\gamma$, and so $\alpha(w) - \alpha(z) = \beta(w) - \beta(z)$. □

Since there are loops with winding number $\neq 0$, we conclude:

- There is no continuous choice of argument on all of $\mathbb{C}^*$.

With Proposition 9.30 we conclude:

- The punctured plane is not simply connected.

## 9.3 Differentiability: A Reminder

We quickly recall some basic facts about differentiability of multi-variable functions. For more details see for example [Spi65], [Die69, Ch.8] or [Mun91, Ch.2]. Let $U \subseteq \mathbb{R}^n$ be open, let $\boldsymbol{a} \in U$, and let $f : U \to \mathbb{R}^m$ be a function. Another function $g : U \to \mathbb{R}^m$ is *tangent* to $f$ at $\boldsymbol{a}$ if $\lim_{\boldsymbol{x} \to \boldsymbol{a}} (f(\boldsymbol{x}) - g(\boldsymbol{x}))/|\boldsymbol{x} - \boldsymbol{a}| = 0$ (using the notion of a limit of a function, see Exercise 8.12). This implies that $f(\boldsymbol{a}) = g(\boldsymbol{a})$; so unravelling, this says that for all $\varepsilon > 0$ there is some $\delta > 0$ such that for all $\boldsymbol{x} \in U$, if $|\boldsymbol{x} - \boldsymbol{a}| < \delta$ then $|f(\boldsymbol{x}) - g(\boldsymbol{x})| < \varepsilon |\boldsymbol{x} - \boldsymbol{a}|$. Tangency at a point is an equivalence relation; if $f$ and $g$ are tangent at $\boldsymbol{a}$, and $g$ and $h$ are tangent at $\boldsymbol{a}$, then $f$ and $h$ are tangent at $\boldsymbol{a}$. Among all possible tangents we seek a translate of a linear map: a function $g(\boldsymbol{a} + \boldsymbol{h}) = g(\boldsymbol{a}) + T(\boldsymbol{h})$ where $T : \mathbb{R}^n \to \mathbb{R}^m$ is linear.

*Exercise 9.33* Suppose that $T$ and $S$ are linear maps from $\mathbb{R}^n$ to $\mathbb{R}^m$. Show that if the functions $g(\boldsymbol{a}+\boldsymbol{h}) = \boldsymbol{b} + T(\boldsymbol{h})$ and $f(\boldsymbol{a}+\boldsymbol{h}) = \boldsymbol{b} + S(\boldsymbol{h})$ are tangent at $\boldsymbol{a}$ then $S = T$. «

Thus there is at most one translate of a linear map which is tangent to $f$ at $\boldsymbol{a}$. If there is such a translate $f(\boldsymbol{a}) + T(\boldsymbol{h})$ then we say that $f$ is differentiable at $\boldsymbol{a}$ and write $Df(\boldsymbol{a})$ for the matrix $A$ such that $T = T_A$ is multiplication by $A$. Unravelling again we see that $Df(\boldsymbol{a}) = A$ if and only if for all $\varepsilon > 0$ there is some $\delta > 0$ such that for all $\boldsymbol{h} \in \mathbb{R}^n$, if $|\boldsymbol{h}| < \delta$ then

$$|(f(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a})) - A\boldsymbol{h}| < \varepsilon|\boldsymbol{h}|.$$

*Exercise 9.34* Prove: if $f$ is differentiable at $\boldsymbol{a}$ then it is continuous at $\boldsymbol{a}$. «

If $f$ is differentiable at every point of $U$ then we get a function $Df : U \to M_{m\times n}(\mathbb{R})$. We say that $f$ is *continuously differentiable* at a point $\boldsymbol{a} \in U$ if $f$ is differentiable at all points in a neighbourhood of $\boldsymbol{a}$ and the function $Df$ (defined at least on that neighbourhood) is continuous at $\boldsymbol{a}$. We say that $f$ is *smooth* if it is differentiable on $U$ and the function $Df$ is continuous on $U$.[1]

*Exercise 9.35* (a) Show that a constant function is differentiable at every point and its derivative is the zero matrix. (b) Let $T = T_A : \mathbb{R}^n \to \mathbb{R}^m$ be a linear map. Show that $T$ is differentiable at every point and that $DT = A$ at every point. «

*Exercise 9.36* Show that the function $1/x$ (defined on $\mathbb{R} \setminus \{0\}$) is smooth. «

If $f : U \to \mathbb{R}^m$ then $f = \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix}$, where $f_i : U \to \mathbb{R}$. Recall that for matrix multiplication, $\mathbb{R}^m$ is the space of columns, not rows. However we will sometimes abuse notation and write elements of $\mathbb{R}^m$ as rows.

*Exercise 9.37* Show that a function $f : U \to \mathbb{R}^m$ is differentiable at $\boldsymbol{a}$ if and only if each $f_i$ is differentiable at $\boldsymbol{a}$; in that case

$$Df(\boldsymbol{a}) = \begin{pmatrix} Df_1(\boldsymbol{a}) \\ \vdots \\ Df_m(\boldsymbol{a}) \end{pmatrix}.$$

«

[1] Recall that this is nonstandard terminology, see page 221.

*Notation 9.38* When $m = 1$, $Df$ is a row of length $n$. When $n = 1$, $Df$ is a column of height $m$, so in this case both $f$ and $Df$ map into $\mathbb{R}^m$.

The notation $f'$ and $\dot{f}$ is often used for $Df$. We will use the notation $\dot{f}$ for the case $n = 1$. We will use the notation $f'$ to denote complex differentiation (from Chap. 11 onwards).

When $m = 1$, a common notation for $Df$ is also $\nabla f$, in which case $Df$ is called the *gradient* of $f$. «

**The Operator Norm**

We will use the notion of the *operator norm* of a linear map. Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be linear. By Exercise 8.13, $T$ is continuous; by Exercise 8.14, the map $\boldsymbol{x} \mapsto |T(\boldsymbol{x})|$ is continuous. The closed unit ball $\overline{B}(0, 1)$ in $\mathbb{R}^n$ is compact (Theorem 8.91); by Exercise 8.93, we can define:

**Definition 9.39** For a matrix $A \in M_{m\times n}(\mathbb{R})$ we let

$$\|A\| = \max\{|A\boldsymbol{x}| \,:\, |\boldsymbol{x}| \leqslant 1\}\,.$$

Linearity gives:

**Proposition 9.40** *For all $\boldsymbol{x} \in \mathbb{R}^n$, $|A\boldsymbol{x}| \leqslant \|A\| \cdot |\boldsymbol{x}|$.*

*Exercise 9.41* Show that: (a) If $\|A\| = 0$ then $A$ is the zero matrix. (b) For $A, B \in M_{m\times n}(\mathbb{R})$, $\|A + B\| \leqslant \|A\| + \|B\|$. (c) For $A \in M_{m\times n}(\mathbb{R})$ and $c \in \mathbb{R}$, $\|cA\| = |c| \cdot \|A\|$. (d) For $A \in M_{m\times n}(\mathbb{R})$ and $B \in M_{k\times m}(\mathbb{R})$, $\|BA\| \leqslant \|B\| \cdot \|A\|$. (e) The map $A \mapsto \|A\|$ is continuous. (Hint: use Exercise 8.13.) «

Note that in (d) we can have inequality; for a simple example, consider nonzero matrices $A$ and $B$ such that $BA = 0$.

**The Chain Rule**

**Proposition 9.42** *Let $U \subseteq \mathbb{R}^n$, $V \subseteq \mathbb{R}^m$, $f : U \to V$, $g : V \to \mathbb{R}^k$, and $\boldsymbol{a} \in U$. Suppose that $f$ is differentiable at $\boldsymbol{a}$ and that $g$ is differentiable at $f(\boldsymbol{a})$. Then $g \circ f$ is differentiable at $\boldsymbol{a}$ and $D(g \circ f)(\boldsymbol{a}) = Dg(f(\boldsymbol{a})) \cdot Df(\boldsymbol{a})$.*

This can be intuitively explained: if a translate of the linear map $T$ is tangent to $f$ at $\boldsymbol{a}$, and a translate of $S$ is tangent to $g$ at $f(\boldsymbol{a})$, then the translate of $S \circ T$ is tangent to $g \circ f$ at $\boldsymbol{a}$.

***Sketch of Proof*** Let $\boldsymbol{b} = f(\boldsymbol{a})$, Let $\boldsymbol{h}$ be small, let $\boldsymbol{k} = f(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a})$; let $\boldsymbol{q} = g(f(\boldsymbol{a}+\boldsymbol{h})) - g(f(\boldsymbol{a}))$. Let $A = Df(\boldsymbol{a})$ and let $B = Dg(\boldsymbol{b})$. Then $|\boldsymbol{q} - B\boldsymbol{k}| \leqslant \varepsilon|\boldsymbol{k}|$; and $|\boldsymbol{k} - A\boldsymbol{h}| \leqslant \varepsilon|\boldsymbol{h}|$. By Proposition 9.40, $|A\boldsymbol{h}| \leqslant \|A\| \cdot |\boldsymbol{h}|$, so $|\boldsymbol{k}| \leqslant (\|A\| + \varepsilon)|\boldsymbol{h}|$. Assuming $\varepsilon \leqslant 1$,

$$|\boldsymbol{q} - B\boldsymbol{k}| \leqslant \varepsilon|\boldsymbol{k}| \leqslant \varepsilon(\|A\| + 1)|\boldsymbol{h}|.$$

Again by Proposition 9.40,

$$|B\boldsymbol{k} - BA\boldsymbol{h}| = |B(\boldsymbol{k} - A\boldsymbol{h})| \leqslant \|B\| \cdot |\boldsymbol{k} - A\boldsymbol{h}| \leqslant \|B\| \cdot \varepsilon|\boldsymbol{h}|,$$

so overall $|\boldsymbol{q} - BA\boldsymbol{h}| \leqslant (\|A\| + \|B\| + 1)\varepsilon|\boldsymbol{h}|$. □

*Exercise 9.43* Show that the composition of two smooth functions is smooth. «

*Exercise 9.44* Let $f, g : U \to \mathbb{R}^m$ and let $c \in \mathbb{R}$. Show that $D(f+g) = Df + Dg$ and $D(cf) = c \cdot Df$. (In detail, if $\boldsymbol{a} \in U$ and $f$ and $g$ are differentiable at $\boldsymbol{a}$ then $f + g$ is differentiable at $\boldsymbol{a}$ and $D(f+g)(\boldsymbol{a}) = Df(\boldsymbol{a}) + Dg(\boldsymbol{a})$. You can use the chain rule, together with Exercise 9.37 and Exercise 9.35 applied to function $(\boldsymbol{x}, \boldsymbol{y}) \mapsto \boldsymbol{x} + \boldsymbol{y}$.) «

### 9.3.1 Mean Value Inequalities

The following is related to the mean value theorem. It says that if the speed of a car is bounded by $M$, then the distance it travels is bounded by $M \times$the time it travels. In this section we avoid using the mean value theorem, in favour of arguments that we consider more intuitive. See, for example, [Tuc97] for more details.

**Proposition 9.45** *Let $\gamma : [a, b] \to \mathbb{R}^n$ be continuous, and differentiable on the open interval $(a, b)$. Let $M \geqslant 0$ and suppose that $|\dot{\gamma}(t)| \leqslant M$ for all $t \in (a, b)$. Then $|\gamma(b) - \gamma(a)| \leqslant M \cdot (b - a)$.*

***Proof*** We prove the lemma under the assumption that $|\dot{\gamma}(t)| < M$ for all $t \in (a, b)$. Then, replacing $M$ by $M+\varepsilon$, we could conclude that $|\gamma(b)-\gamma(a)| \leqslant (M+\varepsilon)\cdot(b-a)$ for all $\varepsilon > 0$, giving the desired result.

Also, we may assume that $\gamma$ also has one-sided derivatives at $a$ and $b$, also bounded by $M$—because we can replace $[a, b]$ by $[a + \varepsilon, b - \varepsilon]$ and then use continuity of $\gamma$.

The two assumptions imply that for every $t \in [a, b]$ there is some $\delta > 0$ such that for all $s \in [a, b]$, if $|s - t| < \delta$ then $|\gamma(s) - \gamma(t)| \leqslant M \cdot |s - t|$. It follows that if $s, r \in [a, b]$, $s \leqslant t \leqslant r$ and $|r - s| < \delta$ then $|\gamma(r) - \gamma(s)| \leqslant M \cdot (r - s)$.

Suppose that $|\gamma(b) - \gamma(a)| > M \cdot (b - a)$. Inductively we define a shrinking sequence of closed subintervals $I_k$, starting with $I_0 = [a, b]$. Given $I_k = [a_k, b_k]$

such that $|\gamma(b_k) - \gamma(a_k)| > M \cdot (b_k - a_k)$, let $z$ be the midpoint of $I_k$. Either $|\gamma(b_k) - \gamma(z)| > M \cdot (b_k - z)$ or $|\gamma(z) - \gamma(a_k)| > M \cdot (z - a_k)$; we choose $I_{k+1} = [a_k, z]$ or $I_{k+1} = [z, b_k]$ accordingly.

By completeness of $\mathbb{R}$, $z^* = \lim_k a_k = \lim_k b_k$ exists. Fixing $\delta$ as above for $z^*$, for large enough $k$, $|b_k - a_k| < \delta$ (and $a_k \leqslant z^* \leqslant b_k$) giving a contradiction. □

For $\mathbb{R}$-valued functions we get a lower bound as well:

*Exercise 9.46* Let $f : [a, b] \to \mathbb{R}$ be continuous, and differentiable on $(a, b)$. Let $M \in \mathbb{R}$, and suppose that $\dot{f}(t) > M$ for all $t \in (a, b)$. Show that $f(b) - f(a) > M \cdot (b - a)$. (Use the technique of Proposition 9.45.) «

**Corollary 9.47** *Let $f : [a, b] \to \mathbb{R}$ be continuous and differentiable on $(a, b)$. (a) If $\dot{f} > 0$ on $(a, b)$ then $f$ is (strictly) increasing: if $x < y$ then $f(x) < f(y)$. (b) If $\dot{f} \geqslant 0$ on $(a, b)$ then $f$ is nondecreasing: if $x < y$ then $f(x) \leqslant f(y)$. (c) If $\dot{f} = 0$ on $(a, b)$ then $f$ is constant.*

***Sketch of Proof*** (a) is Exercise 9.46 with $M = 0$. For (b), for every $\alpha > 0$, apply Exercise 9.46 to the function $f(t) + \alpha t$. (c) follows from Proposition 9.45 with $M = 0$ (alternatively, from (b), applied to $f$ and $-f$). □

We can extend Proposition 9.45 to functions of more variables:

**Proposition 9.48** *Suppose that $U \subseteq \mathbb{R}^n$ is open and that $K \subseteq U$ is convex (see Example 9.15); suppose that $f : U \to \mathbb{R}^m$ is differentiable and that $\|Df\| \leqslant M$ on $K$. Then for all $\boldsymbol{a}, \boldsymbol{b} \in K$, $|f(\boldsymbol{b}) - f(\boldsymbol{a})| \leqslant M|\boldsymbol{b} - \boldsymbol{a}|$.*

***Proof*** Let $\gamma : [0, 1] \to U$ be defined by $\gamma(t) = (1 - t)\boldsymbol{a} + t\boldsymbol{b}$. By the chain rule, $|D(f \circ \gamma)| \leqslant M|\boldsymbol{b} - \boldsymbol{a}|$ on $(0, 1)$. Now apply Proposition 9.45. □

The following proposition gives a "uniform modulus of differentiability".

**Proposition 9.49** *Let $U \subseteq \mathbb{R}^n$ be open, $f : U \to \mathbb{R}^m$ be smooth, and $K \subseteq U$ be compact and convex. Then for every $\varepsilon > 0$ there is some $\delta > 0$ such that for all $\boldsymbol{a}, \boldsymbol{b} \in K$, if $|\boldsymbol{b} - \boldsymbol{a}| < \delta$ then*

$$|(f(\boldsymbol{b}) - f(\boldsymbol{a})) - Df(\boldsymbol{a}) \cdot (\boldsymbol{b} - \boldsymbol{a})| \leqslant \varepsilon \cdot |\boldsymbol{b} - \boldsymbol{a}|.$$

***Proof*** This uses the fact that $Df$ is uniformly continuous on $K$ (Proposition 8.94). Given $\varepsilon > 0$, find $\delta > 0$ such that $\|Df(\boldsymbol{b}) - Df(\boldsymbol{a}))\| < \varepsilon$ whenever $\boldsymbol{b}, \boldsymbol{a} \in K$ and $|\boldsymbol{b} - \boldsymbol{a}| < \delta$. Fix $\boldsymbol{a} \in K$. Then $K \cap B(\boldsymbol{a}, \delta)$ is convex (it is the intersection of two convex sets). We apply Proposition 9.48 on that convex set with the function $\boldsymbol{x} \mapsto f(\boldsymbol{x}) - (Df(\boldsymbol{a}))\boldsymbol{x}$. □

*Exercise 9.50* Let $U \subseteq \mathbb{R}$ be open and let $f : U \to \mathbb{R}$ be smooth. Define $G : U^2 \to \mathbb{R}$ by letting

$$G(x, y) = \begin{cases} \frac{f(y)-f(x)}{y-x}, & \text{if } x \neq y; \text{ and} \\ \dot{f}(x), & \text{if } x = y. \end{cases}$$

Show that $G$ is continuous. «

### 9.3.2 Partial Derivatives

Let $f : U \to \mathbb{R}$ where $U \subseteq \mathbb{R}^n$. Let $\boldsymbol{a} \in U$. For $i \leqslant n$ we let $D^i f(\boldsymbol{a})$ be the derivative of the function $t \mapsto f(a_1, \ldots, a_{i-1}, t, a_{i+1}, \ldots, a_n)$, if it exists. This is of course the partial derivative in direction $x_i$. When $n = 2$ we write $D^x f$ for $D^1$ and $D^y$ for $D^2$. (More familiar notation is $\partial f/\partial x_i$, and sometimes $f_{x_i}$.)

*Exercise 9.51* Suppose that $f : U \to \mathbb{R}$ is differentiable at $\boldsymbol{a}$. Show that $Df(\boldsymbol{a}) = (D^1 f(\boldsymbol{a}), D^2 f(\boldsymbol{a}), \ldots, D^n f(\boldsymbol{a}))$. «

**Proposition 9.52** *Let $U \subseteq \mathbb{R}^n$, let $f : U \to \mathbb{R}$ and let $\boldsymbol{a} \in U$. Suppose that for each $i \leqslant n$, $D^i f$ is defined in an open neighbourhood of $\boldsymbol{a}$ and that each $D^i f$ is continuous at $\boldsymbol{a}$. Then $f$ is differentiable at $\boldsymbol{a}$.*

***Sketch of Proof*** Let $\boldsymbol{b}$ be close to $\boldsymbol{a}$. Let $\boldsymbol{c_0} = \boldsymbol{a}$, $\boldsymbol{c_1} = (b_1, a_2, \ldots, a_n)$, $\boldsymbol{c_2} = (b_1, b_2, a_3, \ldots, a_n)$, and so on, until $\boldsymbol{c_n} = \boldsymbol{b}$. Then $f(\boldsymbol{b}) - f(\boldsymbol{a}) = \sum_{i \leqslant n} (f(\boldsymbol{c}_i) - f(\boldsymbol{c}_{i-1}))$. By Proposition 9.49 (applied on some small closed ball around $\boldsymbol{a}$, which is convex and compact),

$$|(f(\boldsymbol{c}_i) - f(\boldsymbol{c}_{i-1})) - (b_i - a_i) \cdot D^i f(\boldsymbol{c}_{i-1})| < \varepsilon \cdot |b_i - a_i| \leqslant \varepsilon \cdot |\boldsymbol{b} - \boldsymbol{a}|.$$

But $|D^i f(\boldsymbol{c}_{i-1}) - D^i f(\boldsymbol{a})| < \varepsilon$, so

$$|(f(\boldsymbol{c}_i) - f(\boldsymbol{c}_{i-1})) - (b_i - a_i) \cdot D^i f(\boldsymbol{a})| < 2\varepsilon \cdot |\boldsymbol{b} - \boldsymbol{a}|. \qquad \square$$

*Exercise 9.53* Let $f, g : U \to \mathbb{R}$. Show that $D(fg) = fDg + gDf$. (Like Exercise 9.44, but use Proposition 9.52 to calculate the derivative of the function $(x, y) \mapsto xy$.) «

*Remark 9.54* Exercises 9.35, 9.44 and 9.53 imply that the function defined by a polynomial $f \in \mathbb{R}[x_1, \ldots, x_n]$ is smooth. In the first part of the book, we used the notation $D^{x_i} f$ for formal differentiation of polynomials over any integral

domain. For polynomials in $\mathbb{R}[x_1, \ldots, x_n]$, the polynomial $D^{x_i} f$ defines the partial derivative $D^i f$ of the function defined by the polynomial $f$. «

**Proposition 9.55** *Suppose that $U \subseteq \mathbb{R}^n$ is open, $f : U \to \mathbb{R}$ is differentiable, $\boldsymbol{c} \in U$ and $f(\boldsymbol{c})$ is a minimum value of $f$. Then $Df(\boldsymbol{c}) = 0$.*

The same of course holds for a maximum value.

***Sketch of Proof*** The case $n = 1$ follows from the definition of the derivative: say $f : (a, b) \to \mathbb{R}$ is differentiable, $c \in (a, b)$, and that $\dot{f}(c) \neq 0$. Without loss of generality, suppose that $q = \dot{f}(c) > 0$. Find some $\delta > 0$ such that $|(f(c + h) - f(c)) - qh| \leqslant q|h|/2$ when $|h| < \delta$; then for $0 < h < \delta$ we have $f(c - h) < f(c) < f(c + h)$.

For the case $n > 1$, for each $i = 1, \ldots, n$, apply the case $n = 1$ to the function $x \mapsto (c_1, \ldots, c_{i-1}, x, c_{i+1}, \ldots, c_n)$ to get $D^i f(\boldsymbol{c}) = 0$. □

### 9.3.3 Inverse Functions

**Inverse Function Theorem** *Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}^n$ be smooth, let $\boldsymbol{a} \in U$ and suppose that $Df(\boldsymbol{a})$ is invertible. Then there is an open neighbourhood $V \subseteq U$ of $\boldsymbol{a}$ such that $f[V]$ is open, the restriction $f \restriction_V$ is a homeomorphism between $V$ and $f[V]$, and its inverse $g = (f \restriction_V)^{-1}$ is differentiable at $f(\boldsymbol{a})$, with $Dg(f(\boldsymbol{a})) = Df(\boldsymbol{a})^{-1}$.*

Note that the value of the derivative of the inverse can be deduced from the chain rule, once we know the inverse is indeed differentiable; but the value is revealed naturally during the proof.

***Sketch of Proof of the Inverse Function Theorem*** First, we claim that there is a neighbourhood $O$ of $\boldsymbol{a}$ on which $f$ is 1–1, indeed, there is some $\alpha > 0$ such that $|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)| \geqslant \alpha|\boldsymbol{x}_1 - \boldsymbol{x}_2|$ for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in O$. To see this, let $A = Df(a)$ and $\alpha = 1/(2\|A^{-1}\|)$. Let $O$ be a small open ball around $\boldsymbol{a}$ (and so convex); by Exercise 9.41(e) and the continuous differentiability of $f$, $\|Df(\boldsymbol{x}) - A\| \leqslant \alpha$ for all $\boldsymbol{x} \in O$. Applying Proposition 9.48 to the function $\boldsymbol{x} \mapsto f(\boldsymbol{x}) - A\boldsymbol{x}$, for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in O$,

$$|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) - A(\boldsymbol{x}_1 - \boldsymbol{x}_2)| \leqslant \alpha|\boldsymbol{x}_1 - \boldsymbol{x}_2|;$$

but by definition of $\alpha$, $|A(\boldsymbol{x}_1 - \boldsymbol{x}_2)| \geqslant 2\alpha|\boldsymbol{x}_1 - \boldsymbol{x}_2|$.

Let $\varepsilon$ be small; let $C = \{\boldsymbol{x} : |\boldsymbol{x} - \boldsymbol{a}| = \varepsilon\}$ be the boundary of $\overline{B}(\boldsymbol{a}, \varepsilon) \subset O$. Since $f$ is injective on $\overline{B}(\boldsymbol{a}, \varepsilon)$, $f(\boldsymbol{a}) \notin f[C]$. Let $r$ be the distance $d(f(\boldsymbol{a}), f[C])$. Since $f[C]$ is compact (Proposition 8.64), it is closed (Proposition 8.71); by Proposition 8.95, $r > 0$. Let $W = B(f(\boldsymbol{a}), r/2)$. Let $\boldsymbol{y}_0 \in W$. Then there is

some $\boldsymbol{x}_0 \in B(\boldsymbol{a}, \varepsilon)$ (note: in the *open* ball) such that $f(\boldsymbol{x}_0) = \boldsymbol{y}_0$. Why? Consider $g(\boldsymbol{x}) = |f(\boldsymbol{x}) - \boldsymbol{y}_0|^2$. Note that $\boldsymbol{y} \mapsto |\boldsymbol{y}|^2$ is the polynomial $\sum y_i^2$; it is smooth, and its derivative at a point $\boldsymbol{y}$ is $2\boldsymbol{y}^\mathsf{t}$ (note that $\boldsymbol{y}$ is a column and the derivative is a row). By the chain rule,

$$Dg(\boldsymbol{x}) = 2(f(\boldsymbol{x}) - \boldsymbol{y}_0)^\mathsf{t} \cdot Df(\boldsymbol{x}).$$

Since $\overline{B}(\boldsymbol{a}, \varepsilon)$ is compact, $g$ has a minimum on that closed ball, at some $\boldsymbol{x}_0$ (Exercise 8.93). Since $|\boldsymbol{y}_0 - f(\boldsymbol{a})| < r/2$, $|\boldsymbol{y}_0 - \boldsymbol{y}| > r/2$ for all $\boldsymbol{y}$ on the boundary $f[C]$, so $g(\boldsymbol{a}) < g(\boldsymbol{x})$ for all $\boldsymbol{x} \in C$; hence $\boldsymbol{x}_0$ is in the open ball $B(\boldsymbol{a}, \varepsilon)$. Thus by Proposition 9.55, $Dg(\boldsymbol{x}_0) = 0$. Since $\varepsilon$ is small and the determinant is continuous, $Df(\boldsymbol{x}_0)$ is invertible; it follows that $f(\boldsymbol{x}_0) = \boldsymbol{y}_0$ as required. So we let $V = f^{-1}[W] \cap B(\boldsymbol{a}, \varepsilon)$.

Observe that $|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)| \geqslant \alpha|\boldsymbol{x}_1 - \boldsymbol{x}_2|$ shows that the inverse $g$ of $f\restriction_V$ is continuous on $W$. We need to show that $g$ is differentiable at $f(\boldsymbol{a})$. Let $\eta > 0$ ($\varepsilon$ is already used). Let $\boldsymbol{k}$ be small; let $\boldsymbol{h} = \boldsymbol{a} - g(f(\boldsymbol{a}) + \boldsymbol{k})$. Since $g$ is continuous, $\boldsymbol{h}$ is small as well, so $|\boldsymbol{k} - A\boldsymbol{h}| < \eta|\boldsymbol{h}|$. We bound $|\boldsymbol{h} - A^{-1}\boldsymbol{k}|$ by a constant multiple of $\eta|\boldsymbol{k}|$. Well, as $|\boldsymbol{k}| \geqslant \alpha|\boldsymbol{h}|$,

$$|\boldsymbol{h} - A^{-1}\boldsymbol{k}| = |A^{-1}A\boldsymbol{h} - A^{-1}\boldsymbol{k}| \leqslant \|A^{-1}\| \cdot |A\boldsymbol{h} - \boldsymbol{k}| < \|A^{-1}\| \cdot \eta|\boldsymbol{h}| < (\|A^{-1}\|/\alpha) \cdot \eta|\boldsymbol{k}|. \quad \square$$

*Exercise 9.56* Show that under the hypotheses of the inverse function theorem, we can ensure that $g = (f\restriction_V)^{-1}$ is smooth on $f[V]$. (Hint: the determinant is continuous, so $Df$ is invertible on an open set; and matrix inversion is continuous.) «

*Exercise 9.57* Let $U \subseteq \mathbb{R}^n$ be open. Show that if $f : U \to \mathbb{R}^n$ is smooth, 1–1, and $Df(\boldsymbol{a})$ is invertible for all $\boldsymbol{a} \in U$, then $f[U]$ is open and $f$ is a homeomorphism from $U$ to $f[U]$, with smooth inverse. «

*Remark 9.58* The hypothesis that $Df(\boldsymbol{a})$ is invertible is not necessary for $f$ to be a homeomorphism: the function $x \mapsto x^3$ is a homeomorphism from $\mathbb{R}$ to $\mathbb{R}$ but has derivative 0 at 0. In Chap. 11 we will see that in the complex context, this cannot happen. «

*Remark 9.59* When $n = 1$, some of the consequences of the inverse function theorem follow from Corollary 9.47. Indeed, suppose that $f : [a, b] \to \mathbb{R}$ is continuous, differentiable on $(a, b)$, and that $\dot{f} > 0$ on $(a, b)$ (we do not need to assume that $\dot{f}$ is continuous on $(a, b)$). Since $f$ is strictly increasing, it is one-to-one; since $[a, b]$ is compact, $f$ is a homeomorphism onto its image $[f(a), f(b)]$ (Corollary 8.72 and Exercise 9.4). So $f^{-1}$ is well-defined and continuous. We can

then show that $f^{-1}$ is differentiable on $(f(a), f(b))$, using a simple version of part of the proof of the inverse function theorem. «

### 9.3.4 Second Derivatives

Suppose that $f : U \to \mathbb{R}$ is smooth. Then the full derivative $Df$ is a function from $U$ to $\mathbb{R}^n$ (where again we confuse rows and columns). If in turn the function $Df$ is smooth then we call $f$ "twice smooth". Applying Proposition 9.52 twice shows that $f$ is twice smooth if and only if for all $i$ and $j$, the second partial derivative $D^{ij} f = D^i(D^j f)$ is defined and continuous on $U$, in which case $DDf$ is the Hessian matrix $(D^{ji} f)_{i,j \leqslant n}$.

**Proposition 9.60** *Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be twice smooth. Then the Hessian matrix $D^{ji} f(\boldsymbol{a})$ is symmetric at every $\boldsymbol{a} \in U$.*

In fact the proof works under the weaker assumption that $f$ is smooth, and twice differentiable at $\boldsymbol{a}$. For an alternative proof see Exercise 10.47.

***Sketch of Proof*** We assume $n = 2$, the general case is identical. So we need to show that $D^{xy} f = D^{yx} f$ at every point $\boldsymbol{a} \in U$. For simplicity of notation assume $\boldsymbol{a} = \mathbf{0} = (0, 0)$.

Let $h \in \mathbb{R}$. The proof relies on the fact that

$$(f(h,h) - f(h,0)) - (f(0,h) - f(0,0)) = (f(h,h) - f(0,h)) - (f(h,0) - f(0,0)).$$

Call this quantity $s(h)$; we show that $D^{yx} f(\mathbf{0}) = \lim_{h \to 0} s(h)/h^2$. By the symmetry above the proof will also show that $D^{xy} f(\mathbf{0})$ equals the same limit.

Let $\varepsilon > 0$ and let $h$ be small; we need to show that

$$|s(h) - h^2 \cdot D^{yx} f(\mathbf{0})| \leqslant \varepsilon |h|^2.$$

Fixing $h$, for $t \in [0, h]$ define $g(t) = f(t, h) - f(t, 0)$; so $s(h) = g(h) - g(0)$. For $\boldsymbol{p} = (p_1, p_2) \in \mathbb{R}^2$ close to $\mathbf{0}$,

$$|(Df(\boldsymbol{p}) - Df(\mathbf{0})) - DDf(\mathbf{0}) \cdot \boldsymbol{p}| \leqslant \varepsilon |\boldsymbol{p}|;$$

taking only the first coordinate we get

$$|\big(D^x f(\boldsymbol{p}) - D^x f(\mathbf{0})\big) - \big(p_1 \cdot D^{xx} f(\mathbf{0}) + p_2 \cdot D^{yx} f(\mathbf{0})\big)| \leqslant \varepsilon |\boldsymbol{p}|.$$

applying this to $\boldsymbol{p} = (t, h)$ and $\boldsymbol{p} = (t, 0)$ we get

$$|\big(D^x f(t,0) - D^x f(\mathbf{0}) - t \cdot D^{xx} f(\mathbf{0})\big)| \leqslant \varepsilon |t| \leqslant \varepsilon |h|$$

and

$$|\big(D^x f(t,h) - D^x f(\mathbf{0}) - t \cdot D^{xx} f(\mathbf{0}) - h \cdot D^{yx} f(\mathbf{0})\big)| \leqslant \varepsilon|(t,h)| \leqslant \sqrt{2}\varepsilon|h|;$$

putting these together and noticing that $\dot{g}(t) = D^x f(t,h) - D^x f(t,0)$ we get

$$|\big(\dot{g}(t) - h \cdot D^{yx} f(\mathbf{0})\big)| \leqslant (1+\sqrt{2})\varepsilon|h|.$$

We will apply this to $t = 0$; but also to observe that for all $t \in [0,h]$,

$$|\dot{g}(t) - \dot{g}(0)| \leqslant 2(1+\sqrt{2})\varepsilon|h|.$$

Applying Proposition 9.45 to the function $g(t) - \dot{g}(0) \cdot t$ we see that

$$|s(h) - h \cdot \dot{g}(0)| \leqslant 2(1+\sqrt{2})\varepsilon|h|^2;$$

Together with $|h \cdot \dot{g}(0) - h^2 \cdot D^{yx} f(\mathbf{0})| \leqslant (1+\sqrt{2})\varepsilon|h|^2$ we get the desired inequality. □

## 9.4 Differentiable Manifolds

When is a function between two manifolds differentiable? In a manifold, small neighbourhoods look like $\mathbb{R}^n$, and so we could apply usual differentiability. However, we need to make sure that the choice of local coordinates does not affect the result. This invites the concept of a differentiable manifold.

**Definition 9.61** A manifold $(M, \mathcal{A})$ is *differentiable* if every transition function is smooth.

Note that the inverse of a transition function is also a transition function, so the definition implies that both a transition function and its inverse are smooth.

*Example 9.62* Every manifold we met in the previous chapter is differentiable. For example, transition maps for the unit circle were $\sqrt{1-x^2}$ (with 0 not in the domain); for projective space, functions such as $(\rho_n^{-1} \circ \rho_0)(a_1, \ldots, a_n) = \left(\frac{1}{a_n}, \frac{a_1}{a_n}, \ldots, \frac{a_{n-1}}{a_n}\right)$, and so on. «

Let $M$ and $N$ be manifolds and let $f : M \to N$ be a continuous function. A *coordinate representation* of $f$ is a function of the form $\varphi \circ f \circ \psi^{-1}$, where $\varphi$ is a chart for $N$ and $\psi$ is a chart for $M$. In other words, we use $\varphi$ and $\psi$ to choose coordinates on a patch of the domain and a patch of the range; and then use these coordinates to describe $f$. This coordinate representation is defined on the

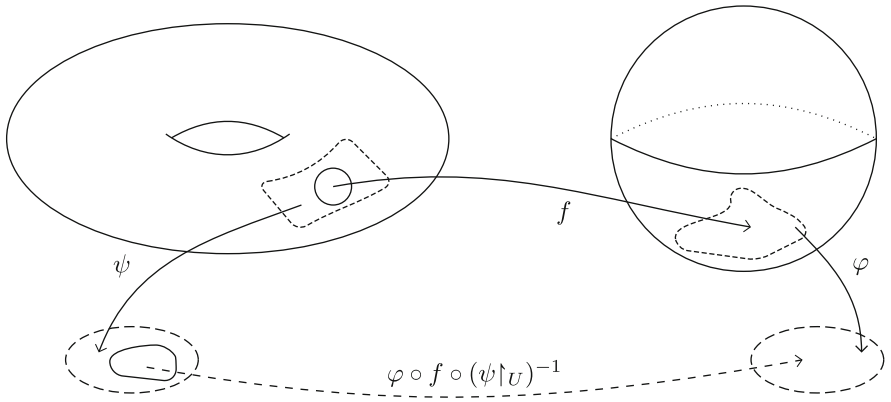



**Fig. 9.1** Giving $f$ coordinates

set $\{\psi(p) : p \in \operatorname{dom}\psi \ \&\ f(p) \in \operatorname{dom}\varphi\} = \psi[\operatorname{dom}\psi \cap f^{-1}[\operatorname{dom}\varphi]]$; this set is open since $f$ is assumed to be continuous. See Fig. 9.1.

**Proposition 9.63** *Let $M$ and $N$ be differentiable manifolds. The following are equivalent for a continuous function $f : M \to N$:*

*(1) Every coordinate representation of $f\restriction_U$ is smooth.*
*(2) For every point $a \in M$ there is a chart $\psi$ for $M$ and a chart $\varphi$ for $N$ such that $a \in \operatorname{dom}\psi$, $f(a) \in \operatorname{dom}\varphi$, and the coordinate representation $\varphi \circ f \circ \psi^{-1}$ is smooth.*

***Proof*** (1)⇒(2): immediate, since $M$ and $N$ are manifolds: for every point $a$ there are charts $\psi$ for $M$ and $\varphi$ for $N$ such that $a \in \operatorname{dom}\psi$ and $f(a) \in \operatorname{dom}\varphi$.

(2)⇒(1): Let $g = \varphi \circ f \circ \psi^{-1}$ be some coordinate representation of $f$. To show that $g$ is smooth it suffices to show that for all $\boldsymbol{c} \in \operatorname{dom} g$ there is an open neighbourhood $V \subseteq \operatorname{dom} g$ of $\boldsymbol{c}$ such that $g$ is smooth on $V$ (formally, the restriction $g\restriction_V$ of $g$ to $V$ is smooth). Fix some point $\boldsymbol{c} \in \operatorname{dom} g$; let $a = \psi^{-1}(\boldsymbol{c})$. By (2), let $\eta$ and $\mu$ be charts for $M$ and $N$ such that $a \in \operatorname{dom}\eta$, $f(a) \in \operatorname{dom}\mu$, and the coordinate representation $h = \mu \circ f \circ \eta$ is smooth. Then on an open neighbourhood of $\boldsymbol{c}$, $g$ equals the composition

$$\left(\varphi \circ \mu^{-1}\right) \circ h \circ \left(\eta \circ \psi^{-1}\right).$$

This is the composition of $h$ with two transition functions, which are assumed to be smooth, and so is smooth (Exercise 9.43). □

We call a function satisfying the conditions of Proposition 9.63 *smooth*. Criterion (2) implies that a function $f : M \to N$ is smooth if and only if there is an open cover $\mathcal{O}$ of $M$ such that for all $O \in \mathcal{O}$, $f\restriction_O : O \to N$ is smooth (recall that an *open*

*cover* of $M$ is a collection $\mathcal{O}$ of open subsets of $M$ such that $M = \bigcup \mathcal{O}$. Also note that an open subset of a manifold is a manifold.)

*Exercise 9.64* Most maps between manifolds, that we have encountered, are smooth. For example, show that the following maps are smooth: (a) the quotient map $\pi_G : \mathbb{R}^n \to \mathbb{R}^n/G$ (where $G$ is a discrete subgroup of $\mathbb{R}^n$). (b) The projection function $\pi_N : \mathbb{A}^{n+1}(\mathbb{R}) \setminus \{\mathbf{0}\} \to \mathbb{P}^n(\mathbb{R})$. (c) The isomorphism between $S \times S$ and $T_\Gamma$ (Exercise 8.111). (d) The homeomorphism between $S$ and $\mathbb{P}^1(\mathbb{R})$ (Exercise 8.55). (e) The homeomorphism between the sphere $S^2$ and the projective complex line $\mathbb{P}^1(\mathbb{C})$ (Exercise 8.56). «

## 9.5 Partitions of Unity

Partitions of unity are used to glue together locally defined functions. This will allow us to smoothen continuous maps.

Let $M$ be a differentiable manifold, and let $f : M \to \mathbb{R}^k$. We let $\operatorname{supp}(f)$, the *support* of $f$, be the *closure* of the set $\{x \in M : f(x) \neq \mathbf{0}\}$ (see Exercise 8.66).

Suppose that $\mathcal{F}$ is a collection of functions from $M$ to $\mathbb{R}^k$; and suppose that for all $x \in M$, $f(x) = \mathbf{0}$ for all but finitely many $f \in \mathcal{F}$. This holds if $x \in \operatorname{supp}(f)$ for only finitely many functions $f \in \mathcal{F}$. Then the sum $\sum_{f \in \mathcal{F}} f(x)$ is well-defined for every $x \in M$, and so gives a function $\sum \mathcal{F}$ from $M$ to $\mathbb{R}^k$.

For differentiability, we need this to happen on neighbourhoods. We say that a collection $\mathcal{G}$ of subsets of $M$ is *locally finite* if there is an open cover $\mathcal{O}$ of $M$ such that every $O \in \mathcal{O}$ intersects only finitely many sets from $\mathcal{G}$. We say that a collection $\mathcal{F}$ of functions from $M$ to $\mathbb{R}^k$ is locally finite if the collection $\{\operatorname{supp}(f) : f \in \mathcal{F}\}$ is locally finite.

**Lemma 9.65** *Suppose that $\mathcal{F}$ is locally finite and that every $f \in \mathcal{F}$ is smooth. Then the sum $\sum \mathcal{F} : M \to \mathbb{R}^k$ is smooth.*

***Proof*** Let $a \in M$. By shrinking, we can find a neighbourhood $U$ of $a$ and a chart $\psi$ for $M$ such that $U \subseteq \operatorname{dom}\psi$, and $U$ intersects $\operatorname{supp}(f)$ for only finitely many $f \in \mathcal{F}$. By Proposition 9.63(1), for each such $f$, the coordinate representation $f \circ (\psi \restriction_U)^{-1}$ is a smooth map from $\psi[U]$ to $\mathbb{R}^k$. It follows that their sum is a smooth map from $\psi[U]$ to $\mathbb{R}^k$. That sum is $\sum \mathcal{F} \circ (\psi \restriction_U)^{-1}$, which is a coordinate representation of $\sum \mathcal{F} \restriction_U$. By Proposition 9.63(2), $\sum \mathcal{F}$ is smooth. □

A *partition of unity* for a manifold $M$ is a locally finite collection $\mathcal{F}$ of smooth functions from $M$ to $[0, 1]$ such that $\sum \mathcal{F}$ is the constant function 1. This notion is useful if we prescribe small neighbourhoods containing the supports of the functions in $\mathcal{F}$. Let $\mathcal{W}$ be an open cover of $M$. We say that a collection $\mathcal{G}$ of subsets of $M$ is *subordinate* to $\mathcal{W}$ if for every $G \in \mathcal{G}$ there is some $W \in \mathcal{W}$ such that $\overline{G} \subseteq W$ (the

closure of $G$ is a subset of $W$). A family $\mathcal{F}$ of functions is subordinate to $\mathcal{W}$ if the collection $\{\text{supp}(f) : f \in \mathcal{F}\}$ is subordinate to $\mathcal{W}$.

**Theorem 9.66** *Let $M$ be a differentiable manifold and let $\mathcal{W}$ be an open cover of $M$. There is a partition of unity for $M$ which is subordinate to $\mathcal{W}$.*

Before we prove Theorem 9.66, we give an example of how partitions of unity may be useful. In general, continuous differentiability is meaningful only for functions defined on open subsets of a manifold. The following lemma allows us to extend the terminology to closed subsets.

**Proposition 9.67** *Let $M$ be a differentiable manifold; let $A \subset M$ be closed, and let $f : A \to \mathbb{R}^k$ be a function. Suppose that every $x \in A$ has an $M$-neighbourhood on which $f$ can be extended to a smooth function. Then $f$ can be extended to a smooth function on $M$.*

***Proof*** We define an open cover $\mathcal{U}$ of $M$ such that for every $U \in \mathcal{U}$ there is a smooth function $f_U : U \to \mathbb{R}^k$ such that $f_U$ agrees with $f$ on $U \cap A$. By assumption, for every $x \in A$ we can find a neighbourhood of $x$ with this property. Since $A$ is closed, we add the set $M \setminus A$ to $\mathcal{U}$ and let $f_{M\setminus A}$ be any smooth function, say a constant one.

Let $\mathcal{F}$ be a partition of unity subordinate to $\mathcal{U}$. For every function $\theta \in \mathcal{F}$ there is some $U_\theta \in \mathcal{U}$ such that $\text{supp}(\theta) \subseteq U_\theta$. For each such $\theta$ we define $g_\theta : M \to \mathbb{R}^k$ by letting, for $x \in M$,

$$g_\theta(x) = \begin{cases} \theta(x) \cdot f_{U_\theta}(x), & \text{if } x \in U_\theta; \text{ and} \\ \mathbf{0}, & \text{otherwise.} \end{cases}$$

Each function $g_\theta$ is smooth. Here we use the fact that the support is closed. On $U_\theta$, $g_\theta$ is the product of a smooth scalar function and a smooth vector function, and so is smooth; on the open set $M \setminus \text{supp}(\theta)$, the function $g_\theta$ is the constant $\mathbf{0}$, and so is smooth.

The collection of functions $\{g_\theta : \theta \in \mathcal{F}\}$ is locally finite because $\text{supp}(g_\theta) \subseteq \text{supp}(\theta)$ (use the fact that $A \subseteq B$ implies $\overline{A} \subseteq \overline{B}$). Thus we let $g = \sum_{\theta\in\mathcal{F}} g_\theta$; by Lemma 9.65, $g$ is smooth. It remains to show that $g$ extends $f$. Suppose that $x \in A$. Then for all $\theta \in \mathcal{F}$, $g_\theta(x) = \theta(x) \cdot f(x)$. Since $\sum \mathcal{F} = 1$,

$$g(x) = \sum_{\theta\in\mathcal{F}} \theta(x) \cdot f(x) = f(x). \qquad \square$$

If the conditions of the proposition hold for a closed set $A \subseteq M$ and a function $f : A \to \mathbb{R}^k$ then we say that $f$ *is smooth on* $A$.

*Exercise 9.68* Let $I = [a, b]$ be a closed interval and let $f\colon [a, b] \to \mathbb{R}^k$ be a function. Show that $f$ is smooth on $I$ if and only if $f \restriction_{(a,b)}$ is smooth, and the one-sided derivatives $\lim_{h\to 0+}(f(a + h) - f(a))/h$ and $\lim_{h\to 0-}(f(b + h) - f(b))/h$ exist and equal the one-sided limits $\lim_{x\to a+} \dot{f}(x)$ and $\lim_{x\to b-} \dot{f}(x)$ respectively. «

*Exercise 9.69* Improve Proposition 9.67 as follows: suppose that in addition to $A$ and $f$ we are given an open set $U \supseteq A$. Show that $f$ can be extended to a smooth function $h\colon M \to \mathbb{R}^k$ such that $\operatorname{supp}(h) \subseteq U$. «

### 9.5.1 Proof of Theorem 9.66

We construct a partition of unity in a number of steps. We first show that every manifold is the union of countably many compact subsets.

**Lemma 9.70** *Let $M$ be a manifold. There is an open cover $\{U_n : n \in \mathbb{N}\}$ of $M$ such that for every $n$, $\overline{U}_n$ is compact and $\overline{U}_n \subseteq U_{n+1}$.*

***Proof*** Let $\mathcal{V}$ be a countable basis for the topology of $M$. By removing some sets, we may assume that for every $V \in \mathcal{V}$ there is some chart $\psi_V$ such that $V \subseteq \operatorname{dom}\psi_V$: for every open set $U$ and every $a \in U$, for any chart $\psi$ such that $a \in \operatorname{dom}\psi$, there is some $V \in \mathcal{V}$ such that $a \in V$ and $V \subseteq U \cap \operatorname{dom}\psi$.

Let $\mathcal{W}$ be the collection of sets $\psi_V^{-1}[B(\boldsymbol{a}, r/2)]$, where $V \in \mathcal{V}$, $\boldsymbol{a}$ and $r$ are rational, and $B(\boldsymbol{a}, r) \subseteq \operatorname{range}\psi_V$. Then $\mathcal{W}$ is a countable open cover of $M$ and for every $W \in \mathcal{W}$, $\overline{W}$ is compact: as $(\psi_V)^{-1}$ is a homeomorphism, it maps the closed ball $\overline{B}(\boldsymbol{a}, r/2)$, which is compact by Theorem 8.91, to $\overline{W}$.

Now define the sequence $U_0, U_1, \ldots$ recursively. Fix an enumeration $W_0, W_1, W_2, \ldots$ of the sets in $\mathcal{W}$. Let $U_0 = W_0$. By induction, suppose that we have defined $U_n$ and that $\overline{U}_n$ is compact. Then there are finitely many elements of $\mathcal{W}$ which cover $\overline{U}_n$; the closure of their union is compact (Exercise 8.66 and 8.70). We let $U_{n+1}$ be their union, together with $W_{n+1}$. □

**Lemma 9.71** *Let $M$ be a manifold. For any open cover $\mathcal{W}$ of $M$ there is a locally finite open cover $\mathcal{O}$ of $M$, subordinate to $\mathcal{W}$, such that the closure of each $O \in \mathcal{O}$ is compact.*[2]

***Proof*** Fix a sequence $U_0, U_1, \ldots$ given by Lemma 9.70; we define an open cover $\mathcal{O}$ by recursion. For notational convenience let $U_{-1} = \emptyset$. At step $n \geqslant 0$ we observe that $\overline{U}_{n+1} \setminus U_n$ is compact (Exercise 8.69). For each $x \in \overline{U}_{n+1} \setminus U_n$ we find some open neighbourhood $O_x$ of $x$ which is (a) disjoint from $\overline{U}_{n-1}$ (which can be done since

[2] In the language of topology, this says that $M$ is *paracompact*.

$\overline{U}_{n-1}$ is closed, Proposition 8.71); and (b) the closure $\overline{O}_x$ is compact and a subset of some set in $\mathcal{W}$. This can be done by choosing $O_x$ to be the inverse image by some chart of a small open ball. We add a finite subcover of $\left\{O_x : x \in \overline{U}_{n+1} \setminus U_n\right\}$ to $\mathcal{O}$. We note that every point in $M$ is in $\overline{U}_{n+1} \setminus U_n$ for some $n$; so $\mathcal{O}$ is indeed an open cover of $M$. By choice of the sets $O_x$, $\mathcal{O}$ is subordinate to $\mathcal{W}$. For each $n$, only finitely many sets in $\mathcal{O}$ intersect $\overline{U}_n$, so the sets $U_n$ witness that $\mathcal{O}$ is locally finite. □

Having secured a locally finite cover we turn to functions. We start by fixing, for all $r > 0$, a smooth function $h_r : \mathbb{R} \to \mathbb{R}$ such that $h_r(x) > 0$ for all $x < r$ and $h_r(x) = 0$ for all $x \geqslant r$; for example $h_r(x) = (x - r)^2$ for all $x \leqslant r$, $h_r(x) = 0$ otherwise.

**Lemma 9.72** *Let $M$ be a differentiable manifold. Suppose that $K \subseteq O \subseteq M$, that $K$ is compact and that $O$ is open. Then there is a smooth function $f : M \to [0, \infty)$ such that $f > 0$ on $K$ and $\mathrm{supp}(f) \subseteq O$.*

***Proof*** For each $x \in K$ we can find a smooth function $f_x : M \to [0, \infty)$ such that $f_x > 0$ on a neighbourhood of $x$ and $\mathrm{supp}(f) \subseteq O$. To see this, let $\psi$ be a chart for $M$ such that $x \in \mathrm{dom}\,\psi$. Find a real number $r > 0$ such that the pullback by $\psi^{-1}$ of the closed ball $\overline{B}(\psi(x), r)$ is contained in $O \cap \mathrm{dom}\,\psi$. Define $f_x$ by letting, for $y \in M$,

$$f_x(y) = \begin{cases} h_r(d(\psi(y), \psi(x))), & \text{if } y \in \mathrm{dom}\,\psi; \text{ and} \\ 0, & \text{otherwise.} \end{cases}$$

where recall that $d(\boldsymbol{a}, \boldsymbol{b})$ denotes the Euclidean distance between the points $\boldsymbol{a}$ and $\boldsymbol{b}$; see Exercise 9.99.

By the compactness of $K$, we can find finitely many points $x_1, x_2, \ldots, x_k \in K$, and neighbourhoods $U_i$ of $x_i$ such that $K \subseteq \bigcup_i U_i$ and $f_{x_i} > 0$ on $U_i$; we let $f = \sum_{i \leqslant k} f_{x_i}$. □

We are now ready to prove Theorem 9.66. We are given a differentiable manifold $M$ and an open cover $\mathcal{W}$ of $M$. We apply Lemma 9.71 *twice*: first we apply it to $\mathcal{W}$, and get a locally finite open cover $\mathcal{O}$ of $M$ subordinate to $\mathcal{W}$; and then apply it to $\mathcal{O}$, and get an open cover $\mathcal{U}$ of $M$ subordinate to $\mathcal{O}$. Take some $U \in \mathcal{U}$; there is some $O \in \mathcal{O}$ such that $\overline{U} \subseteq O$. By Lemma 9.72 let $f_U : M \to [0, \infty)$ be smooth such that $f_U > 0$ on $\overline{U}$ and $\mathrm{supp}(f_U) \subseteq O$. Since $\mathcal{O}$ is locally finite, so is the family $\{f_U : U \in \mathcal{U}\}$; let $g$ be the sum of this family. Because $\mathcal{U}$ is an open cover of $M$, $f_U$ is strictly positive on $U$, and every $f_{U'}$ is non-negative, we see that $g$ is strictly positive, and $f_U \leqslant g$ for all $U$; the desired partition of unity is the collection $\{f_U/g : U \in \mathcal{U}\}$. □

*Remark 9.73* We mentioned above that often the notion of smoothness used is infinite differentiability ($C^\infty$ functions) rather than mere continuous differentiability. One can then require partitions of unity to be $C^\infty$ functions. The only extra difficulty is in getting the functions $h_r$ to be infinitely differentiable; one usually uses variants of the function $e^{-1/x}$. «

## 9.6 Differentiable Connectedness

Paths and the notions of connectedness are more useful when they are differentiable. We will use partitions of unity to smoothen continuous paths and homotopies. For example, we will see that an open subset of $\mathbb{R}^n$ is path-connected if and only if any two points can be connected by a smooth path.

The following lemma says that continuous functions on a differentiable manifold can be closely approximated by smooth ones. Recall the discussion around Proposition 9.67 about continuous differentiability on a closed set.

**Lemma 9.74** *Let $M$ be a differentiable manifold, let $F \subseteq M$ be closed, and let $f\colon F \to \mathbb{R}^k$ be continuous. Let $\varepsilon > 0$. Then there is a smooth function $g\colon M \to \mathbb{R}^k$ such that for all $x \in F$, $d(f(x), g(x)) < \varepsilon$. Further, if $A \subseteq F$ is closed and the restriction $f\restriction_A$ is smooth then we can find such $g$ which agrees with $f$ on $A$.*

***Proof*** Let $\mathcal{W}$ be the collection of open sets $U$ such that for all $a, b \in U \cap F$, $d(f(a), f(b)) < \varepsilon$. The continuity of $f$ implies that $\mathcal{W}$ is an open cover of $M$. Let $\mathcal{F}$ be a partition of unity for $M$ which is subordinate to $\mathcal{W}$. For $\theta \in \mathcal{F}$ fix some $U_\theta \in \mathcal{W}$ such that $\operatorname{supp}(\theta) \subseteq U_\theta$; and pick some $a_\theta \in U_\theta \cap F$, unless $U_\theta$ is disjoint from $F$. For $\theta \in \mathcal{F}$ and $x \in M$ let

$$g_\theta(x) = \begin{cases} \theta(x) \cdot f(a_\theta), & \text{if } U_\theta \cap F \neq \emptyset; \\ \mathbf{0}, & \text{if } U_\theta \cap F = \emptyset. \end{cases}$$

Let $g = \sum_{\theta \in \mathcal{F}} g_\theta$. By Lemma 9.65, $g$ is smooth. Let $x \in F$; let $\mathcal{G}$ be the (finite) collection of $\theta \in \mathcal{F}$ such that $x \in \operatorname{supp}(\theta)$. So $\sum_{\theta \in \mathcal{G}} \theta(x) = 1$. For all $\theta \in \mathcal{G}$, $a_\theta$ is defined, and so

$$|g_\theta(x) - \theta(x) \cdot f(x)| = \theta(x) \cdot |f(a_\theta) - f(x)| < \theta(x) \cdot \varepsilon.$$

Since $f(x) = \sum_{\theta \in \mathcal{G}} \theta(x) \cdot f(x)$,

$$d(g(x), f(x)) \leqslant \sum_{\theta \in \mathcal{G}} |g_\theta(x) - \theta(x) \cdot f(x)| < \sum_{\theta \in \mathcal{G}} \theta(x) \cdot \varepsilon = \varepsilon$$

as required.

Now suppose that $f$ is smooth on $A$. By Proposition 9.67 let $h\colon M \to \mathbb{R}^k$ be a smooth extension of $f\restriction_A$ to $M$. We modify the construction as follows. We let the sets in $\mathcal{W}$ be open sets $U$ satisfying that both $d(f(x), f(y)) < \varepsilon$ for all $x, y \in U \cap F$, and $d(h(x), h(y)) < \varepsilon$ for all $x, y \in U$. Again let $\mathcal{F}$ be a partition of unity subordinate to $\mathcal{W}$; define $U_\theta$ as above, and choose $a_\theta \in U_\theta \cap F$ when possible. Now let, for $\theta \in \mathcal{F}$ and $x \in M$,

$$g_\theta(x) = \begin{cases} \theta(x) \cdot h(x), & \text{if } U_\theta \cap A \neq \emptyset; \\ \theta(x) \cdot f(a_\theta), & \text{if } U_\theta \cap A = \emptyset \text{ but } U_\theta \cap F \neq \emptyset, \\ \mathbf{0}, & \text{if } U_\theta \cap F = \emptyset. \end{cases}$$

Let $g = \sum_{\theta \in \mathcal{F}} g_\theta$. Again let $x \in M$, and let $\mathcal{G}$ be the finite set of $\theta \in \mathcal{F}$ such that $x \in \operatorname{supp}(\theta)$. If $x \in A$ then for all $\theta \in \mathcal{G}$, $U_\theta \cap A \neq \emptyset$ so $g_\theta(x) = \theta(x) \cdot h(x)$ whence $g(x) = h(x) = f(x)$. If $x \in F \setminus A$ then for all $\theta \in \mathcal{G}$, $a_\theta$ is defined, but it is possible that for some $\theta \in \mathcal{G}$, $U_\theta \cap A \neq \emptyset$. But for such $\theta$, $f$ and $h$ are close: fixing $z \in U_\theta \cap A$, we have

$$d(f(x), h(x)) \leqslant d(f(x), f(z)) + d(h(z), h(x)) < 2\varepsilon$$

seeing as $f(z) = h(z)$. So in both cases we have $|g_\theta(x) - \theta(x) \cdot f(x)| \leqslant 2\theta(x)\varepsilon$. □

We now discuss differentiable connectedness. To reiterate, a path $\gamma\colon I \to \mathbb{R}^n$ is smooth if it can be extended to a smooth function on $\mathbb{R}$ (and see Exercise 9.68). We can similarly work with smooth homotopies.

**Proposition 9.75** *Let $U \subseteq \mathbb{R}^n$ be open, and let $\boldsymbol{a}, \boldsymbol{b} \in U$. If there is a path in $U$ from $\boldsymbol{a}$ to $\boldsymbol{b}$ then there is a smooth path in $U$ from $\boldsymbol{a}$ to $\boldsymbol{b}$.*

***Proof*** Let $I = [a, b]$ be a closed interval and let $\gamma\colon I \to U$ be a path. The image $E = \gamma[I]$ of the path is compact (Proposition 8.64); by Proposition 8.97, the distance $d(E, \mathbb{R}^n \setminus U)$ between $E$ and the complement of $U$ is positive. In other words, there is some $\varepsilon > 0$ such that for all $\boldsymbol{a} \in \mathbb{R}^n$, if $d(\boldsymbol{a}, E) < \varepsilon$ then $\boldsymbol{a} \in U$.

Apply Lemma 9.74 with $M = \mathbb{R}$, $F = I$, $f = \gamma$, $A = \{a, b\}$ and $\varepsilon$: we can extend $\gamma\restriction_{\{a\}}$, which is a function defined on a single point, to a smooth function in a neighbourhood of $a$, for example a constant one; and similarly for $\gamma\restriction_{\{b\}}$. Let $\hat{\gamma}$ be the restriction to $[a, b]$ of the function given by the lemma. It is a smooth path from $\gamma(a)$ to $\gamma(b)$. The image of $\hat{\gamma}$ is contained in $U$: for all $t \in [a, b]$, $d(\hat{\gamma}(t), E) \leqslant d(\hat{\gamma}(t), \gamma(t)) < \varepsilon$ and so $\hat{\gamma}(t) \in U$. □

This gives us an extension of Proposition 9.10: if $U \subseteq \mathbb{R}^n$ is open and connected then it is *differentiably path-connected*: every two points are connected by a smooth path. From now on we call an open and connected subset of $\mathbb{R}^n$ a **region**.

*Porism 9.76* The proof of Proposition 9.75 shows that for any path $\gamma : [a, b] \to U$ there is a smooth path $\hat{\gamma} : [a, b] \to U$ with the same end-points such that for all $t \in [a, b]$, $d(\gamma(t), \hat{\gamma}(t)) < \varepsilon$, where $\varepsilon$ is sufficiently small so that $B(\gamma(t), \varepsilon) \subseteq U$ for all $t \in [a, b]$. This implies that $\gamma$ is homotopic in $U$ to $\hat{\gamma}$ by the linear homotopy $\gamma_s(t) = (1 - s)\gamma(t) + s\hat{\gamma}(t)$. So: every path in $U$ is homotopic (in $U$) to a smooth one. «

**Corollary 9.77** *Let $U \subseteq \mathbb{R}^n$ be a region, and let $f : U \to \mathbb{R}^k$ be differentiable such that $Df = \mathbf{0}$ (the zero matrix) on $U$. Then $f$ is constant.*

Note that in contrast with Proposition 9.48, here we do not assume that $U$ is convex.

***Proof*** Let $\boldsymbol{a}, \boldsymbol{b} \in U$. By Proposition 9.75 let $\gamma : [a, b] \to U$ be a smooth path from $\boldsymbol{a}$ to $\boldsymbol{b}$. Let $g = f \circ \gamma$. Then $\dot{g} = \mathbf{0}$ on $[a, b]$ (chain rule). By Proposition 9.45, $g$ is constant, which means that $f(\boldsymbol{b}) = f(\boldsymbol{a})$. □

**Proposition 9.78** *Let $U \subseteq \mathbb{R}^n$ be open, and let $\gamma, \delta : I \to U$ be two smooth paths with the same start and end points. If there is a homotopy in $U$ between $\gamma$ and $\delta$ then there is a smooth homotopy in $U$ between $\gamma$ and $\delta$.*

***Proof*** This is an extension of the argument for Proposition 9.75. Again let $I = [a, b]$ and let $H : [0, 1] \times I \to U$ be a homotopy between $\gamma$ and $\delta$. Again let $E$ be the range of $H$, and let $\varepsilon = d(E, \mathbb{R}^n \setminus U)$, which is positive since $[0, 1] \times I$ is compact.

The proof is complete once we apply Lemma 9.74 with $M = \mathbb{R}^2$, $F = [0, 1] \times I$, $f = H$, $\varepsilon$, and $A$ being the boundary of $F$: the points $(x, y)$ with $x = 0$, or $x = 1$, or $y = a$, or $y = b$. We just need to observe that $H$ is smooth on $A$; we use Proposition 9.67.

Consider, for example, a small open neighbourhood $V$ of the corner $(0, a)$. Map $(x, y)$ in $V$ to $g(x, y) = \widetilde{\gamma}(y)$, where $\widetilde{\gamma}$ is a smooth extension of $\gamma$ to $\mathbb{R}$. Then $g$ agrees with $H$ on $A \cap V$, and is smooth. A similar argument works near any point of $A$. □

**Corollary 9.79** *Let $U$ be a region in $\mathbb{R}^n$ (open and connected). Then $U$ is simply connected if and only if any two smooth paths in $U$ with the same domain and end-points are homotopic in $U$ by a smooth homotopy.*

***Proof*** One direction is Proposition 9.78. In the other direction suppose that any two smooth paths in $U$ (with the same domain and end-points) are homotopic in $U$. Let $\gamma$ and $\delta$ be two paths in $U$ with the same domain $I$ and same end-points $\boldsymbol{a}$ and $\boldsymbol{b}$. By porism 9.76 there are smooth paths $\hat{\gamma}$ and $\hat{\delta}$ in $U$ from $\boldsymbol{a}$ to $\boldsymbol{b}$ (with domain $I$) such that $\gamma$ and $\hat{\gamma}$ are homotopic in $U$, and $\delta$ and $\hat{\delta}$ are homotopic in $U$. The assumption implies that $\hat{\gamma}$ and $\hat{\delta}$ are homotopic in $U$. Combining these three homotopies together (Exercise 9.12) we see that $\gamma$ and $\delta$ are homotopic in $U$. □

### 9.6.1 Piecewise Smooth Paths

Smooth paths are much nicer than general paths (they cannot be space-filling, for example). On the other hand the concatenation of two smooth paths may fail to be smooth; the derivatives at the meeting point may disagree. So it is useful to consider the class of paths which are obtained as concatenations of finitely many smooth paths.

**Definition 9.80** A path $\gamma\colon [a,b] \to M$ is *piecewise smooth* if there are points $a = c_0 < c_1 < \cdots < c_k = b$ such that for all $i < k$, the restriction $\gamma\restriction_{[c_i,c_{i+1}]}$ is smooth.

In other words, if there is a partition of $[a,b]$ into finitely many closed intervals $I_1, I_2, \ldots, I_k$ such that $\gamma$ is smooth on each sub-interval $I_j$.

Just as we like homotopic smooth paths to be homotopic by a smooth homotopy, we want homotopic piecewise smooth paths to be homotopic by a piecewise smooth homotopy. The definition is as expected: a homotopy $H\colon [0,1] \times I \to M$ is *piecewise smooth* if we can partition the rectangle $[0,1] \times I$ into finitely many closed sub-rectangles, on each of which, $H$ is smooth. That is, if there are points $0 = t_0 < t_1 < \cdots < t_k = 1$ and $a = s_0 < s_1 < \cdots < s_m = b$ (where $I = [a,b]$) such that for each $i = 1, \ldots, k$ and $j = 1, \ldots, m$, the restriction of $H$ to $[t_{i-1}, t_i] \times [s_{j-1}, s_j]$ is smooth.

*Remark 9.81* If $H$ is a piecewise smooth homotopy from $\gamma$ to $\delta$ then all the sections of $H$, both horizontal and vertical, are piecewise smooth paths. That is, for all $s$, the path $H_s$ defined by $H_s(t) = H(s,t)$ is piecewise smooth, and for all $t$, the path $H^t$ defined by $H^t(s) = H(s,t)$ is piecewise smooth as well. «

We have an analogue of Proposition 9.78:

**Proposition 9.82** *Let $U \subseteq \mathbb{R}^n$ be open, and let $\gamma, \delta\colon I \to U$ be two piecewise smooth paths with the same start and end points. If there is a homotopy in $U$ between $\gamma$ and $\delta$ then there is a piecewise smooth homotopy in $U$ between $\gamma$ and $\delta$.*

***Proof*** Let $\gamma$ and $\delta$ be piecewise smooth paths in $U$ which are homotopic in $U$. Let $\hat{\gamma}$ and $\hat{\delta}$ be smooth paths in $U$ homotopic to $\gamma$ and $\delta$ respectively (porism 9.76). Then by concatenating homotopies, we see that $\hat{\gamma}$ and $\hat{\delta}$ are homotopic in $U$. By Proposition 9.78, there is a smooth homotopy between them. Recall that the homotopy between $\gamma$ and $\hat{\gamma}$ given by porism 9.76 is linear in the first variable: $H(s,t) = (1-s)\gamma(t) + s\hat{\gamma}(t)$. If $J \subseteq I$ is a closed sub-interval of $I$ on which $\hat{\gamma}$ is smooth, then (as $\gamma$ is also smooth on $J$), $H$ is smooth on $[0,1] \times J$ (consider the two partial derivatives). Hence $H$ is piecewise smooth. Similarly, $\delta$ and $\hat{\delta}$ are homotopic by a piecewise smooth homotopy. We can concatenate these two piecewise smooth

homotopies with the smooth homotopy from $\hat{\gamma}$ and $\hat{\delta}$ in between, and obtain a piecewise smooth homotopy from $\gamma$ to $\delta$. □

Similarly, Corollary 9.79 extends to piecewise smooth paths.

**Proposition 9.83** *Let $U$ be a region in $\mathbb{R}^n$. Then $U$ is simply connected if and only if any two piecewise smooth paths in $U$ with the same domain and end-points are homotopic in $U$ by a piecewise smooth homotopy.*

***Proof*** The second direction follows the proof of Corollary 9.79: the assumption implies that any two smooth paths are homotopic (by a piecewise smooth homotopy), and we use the fact that any path in $U$ is homotopic with a smooth one. The first direction is given by Proposition 9.82. □

## 9.7 Further Exercises

**Topological Connectedness**

9.84 Let $X$ and $Y$ be quasi-Euclidean spaces. (a) Show that if $X$ and $Y$ are connected, then so is $X \times Y$. (b) Show that if $X$ and $Y$ are path-connected, then so is $X \times Y$.

9.85 (a) Show that no two of $(0, 1)$, $(0, 1]$ and $[0, 1]$ are homeomorphic. (Hint: the difficulty is that perhaps there is a homeomorphism which does not preserve order. But consider what happens to connectedness if we remove a point or two.) (b) Show that if $n > 1$ then $\mathbb{R}^n$ and $\mathbb{R}$ are not homeomorphic. (c) Show that $\mathbb{R}^3$ and $\mathbb{R}^2$ are not homeomorphic.

9.86 Let $X$ be a quasi-Euclidean space, and let $Y \subseteq X$ be connected. Show that $\overline{Y}$ (the closure of $Y$ in $X$) is connected.

9.87 Use the Intermediate Value Theorem (Exercise 9.4) to prove that every positive $x \in \mathbb{R}$ has a square root.

9.88 Show that every continuous function $f : [0, 1] \to [0, 1]$ has a *fixed point*: a point $x \in [0, 1]$ such that $f(x) = x$.

9.89 Let $S = \{(x, \sin(1/x)) : x > 0\}$ be the graph of the function $\sin(1/x)$ restricted to $x > 0$. (a) Show that $S$ is path connected. (b) Show that $\overline{S}$ (the closure of $S$ in $\mathbb{R}^2$) is $S \cup \big(\{0\} \times [-1, 1]\big)$. (c) Show that $\overline{S}$ is connected but not path-connected.

9.90 Let $X$ be a quasi-Euclidean space. For $x, y$, write $x \sim y$ if there is a connected set $Y \subseteq X$ such that $x, y \in Y$. (a) Show that $\sim$ is an equivalence relation. (b) Show that for all $x \in X$, the $\sim$-equivalence class of $x$ (called the *connected component of* $x$) is the largest connected set $Y \subseteq X$ such that $x \in Y$. (c) Show that every connected component of $X$ is closed.

9.91 Consider the rational numbers $\mathbb{Q}$ as a quasi-Euclidean space (a subspace of $\mathbb{R}$). Show that the connected components of $\mathbb{Q}$ are the singletons.

9.92 Show that the orthogonal group $\mathrm{O}_n(\mathbb{R})$ (see Exercise 8.141) is disconnected. (Hint: consider the determinant.)

9.93 Let $G$ be a topological group which is a manifold. Show that the connected component (Exercise 9.90) of the identity $1_G$ is a subgroup of $G$.

**Topological Partitions of Unity**

A *topological partition of unity* for a manifold $M$ is defined just like a partition of unity (Sect. 9.5), except that the maps are required to be continuous rather than smooth (which may not make sense if $M$ is not differentiable).

9.94 Show that every manifold has a topological partition of unity, subordinate to any given open cover.

9.95 Let $M$ be a compact $n$-manifold. (a) Show that there are: an open cover $\{U_1, U_2, \ldots, U_k\}$ of $M$; for $i = 1, \ldots, k$, a homeomorphism $f_i$ from $U_i$ to an open subset of $\mathbb{R}^n$; and a continuous function $\psi_i : M \to [0, 1]$ such that $\mathrm{supp}(\psi_i) \subseteq U_i$ and $\sum_{i \leqslant k} \psi_i > 0$. (b) For $i \leqslant k$ define $F_i : M \to \mathbb{R}^n$ by letting $F_i(x) = \psi_i(x) f_i(x)$ if $x \in U_i$, and $f_i(x) = \mathbf{0}$ otherwise. Show that $f_i$ is continuous. (c) Define $F : M \to \mathbb{R}^{k+nk}$ by letting

$$F(x) = (\psi_1(x), \psi_2(x), \ldots, \psi_k(x), F_1(x), F_2(x), \ldots, F_k(x)) .$$

Show that $F$ is a homeomorphism between $M$ and a subset of $\mathbb{R}^{k+nk}$.[3]

9.96 Let $M$ be a manifold, and let $A \subseteq M$ be closed. Show that there is a continuous function $f : M \to \mathbb{R}$ such that $A = f^{-1}\{0\}$ is the zero-set of $f$. (If $M$ is differentiable then $f$ can be chosen smooth. Hint: if $M$ is an open subset of $\mathbb{R}^n$ then we can let $f$ be the distance from $A$. In general, use a partition of unity.)

[3] We say that every compact manifold is *embeddable* in a finite-dimensional Euclidean space $\mathbb{R}^m$. In fact this is true for every manifold, but that is harder to prove.

**Multivariable Differential Calculus**

9.97 Let $A \in M_{m\times n}(\mathbb{R})$, and let $M = \max\{|a_{i,j}| : i \leqslant m, j \leqslant n\}$. Show that $\|A\| \leqslant \sqrt{mn}M$. (Hint: use Exercise 8.118. Note that Exercise 8.13 gives the larger bound $m\sqrt{n}M$.)

9.98 Show that the positive square root function $x \mapsto \sqrt{x}$ is continuous on $[0, \infty)$ and smooth on $(0, \infty)$.

9.99 Let $\boldsymbol{a} \in \mathbb{R}^n$. Show that the function $\boldsymbol{x} \mapsto d(\boldsymbol{x}, \boldsymbol{a})$ is smooth on $\mathbb{R}^n \setminus \{\boldsymbol{a}\}$.

9.100 Define $f : \mathbb{R} \to \mathbb{R}$ by letting $f(x) = x^2 \sin(1/x)$ for $x \neq 0$, $f(0) = 0$. Show that $f$ is differentiable at every point, but that $\dot{f}$ is not continuous.

9.101 Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}$ be a function, and suppose that for all $i \leqslant n$, $D^i f$ exists at every point of $U$ and is bounded on $U$ (but not necessarily continuous). Show that $f$ is continuous.

9.102 Let $d, n \geqslant 1$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable function such that for all $t \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^n$, $f(t\boldsymbol{x}) = t^d f(\boldsymbol{x})$. Show that Euler's relation holds:

$$x_1 D^1 f + x_2 D^2 f + \cdots + x_n D^n f = d \cdot f.$$

(Compare of course with the algebraic Euler's Relation.)

**Differentiable Manifolds**

9.103 Let $Q$ be the boundary of the unit square (that is, $([0, 1] \times \{0, 1\}) \cup (\{0, 1\} \times [0, 1])$). Show that there is an atlas on $Q$ which makes it a differentiable manifold which is a topological subspace of the plane $\mathbb{R}^2$.

A *diffeomorphism* between two differentiable manifolds $M$ and $N$ is a bijection $f : M \to N$ which is smooth and such that $f^{-1}$ is smooth.

9.104 Let $T$ be the result of rotating the circle $\{(0, 2 + \cos t, \sin t) : t \in [0, 2\pi]\}$ (with centre $(0, 2, 0)$ and contained in the $yz$-plane) around the axis $y = 0$. (a) Show that there is an atlas which makes $T$ a smooth manifold, which is a topological subspace of $\mathbb{R}^3$. (b) Show that $T$, the torus $T_\Gamma$ and the torus $S \times S$ are diffeomorphic.

9.105 Let $h(t) = t^3$. (a) Show that $h$ is a chart for $\mathbb{R}$. (b) Show that $h$ is topologically compatible with $\mathrm{id}_{\mathbb{R}}$, but that $(\mathbb{R}, \{h, \mathrm{id}_{\mathbb{R}}\})$ is not a differentiable manifold (that is, $h$ and $\mathrm{id}_{\mathbb{R}}$ are not differentiably compatible). (c) Nonetheless, show that the differentiable manifolds $(\mathbb{R}, h)$ and $(\mathbb{R}, \mathrm{id}_{\mathbb{R}})$ are diffeomorphic.

9.106 Let $M$ be a differentiable manifold. Show that every path in $M$ is homotopic to a piecewise smooth path in $M$.[4]

A *differentiable group* is a group which is a differentiable manifold for which the group operation and inverse are both smooth.[5]

9.107 Show that $(\mathbb{C}^*, \cdot)$ is a differentiable group.

**Paths, Homotopies, the Fundamental Group**

9.108 A *re-parameterisation* of a path $\gamma : I \to X$ is any path of the form $\gamma \circ \varphi$, where $\varphi : J \to I$ is a strictly increasing bijection from a closed interval $J$ to $I$. (a) Show that such a map $\varphi$ is a homeomorphism from $J$ to $I$. (b) Show that the relation "$\theta$ is a re-parameterisation of $\gamma$" is an equivalence relation on paths.[6]

9.109 Suppose that $\gamma$ is a smooth path in $\mathbb{R}^n/G$. Show that any lifting of $\gamma$ to $\mathbb{R}^n$ is smooth.

9.110 A space $X$ is *contractible* if there is a continuous function $F : X \times [0, 1] \to X$ such that $F(-, 0)$ is the identity on $X$ and $F(-, 1)$ is constant. Show that every contractible space is simply connected.

9.111 A set $X \subseteq \mathbb{R}^n$ is *star-like* if there is a point $\boldsymbol{p} \in X$ such that for every $\boldsymbol{q} \in X$, the line segment between $\boldsymbol{p}$ and $\boldsymbol{q}$ is in $X$. (So every convex set is star-like, but not vice-versa.) Show that every star-like subset of $\mathbb{R}^n$ is contractible (and hence simply connected).

9.112 Let $G$ be a discrete subgroup of $\mathbb{R}^n$; let $U \subseteq \mathbb{R}^n/G$ be open. (a) Show that if $U$ is simply connected, then there is a "global chart" on $U$: a continuous $\psi : U \to \mathbb{R}^n$ such that $\pi_G \circ \psi$ is the identity on $U$. (b) Show that if $\psi$ is a global chart on $U$, then it is a homeomorphism with an open subset of $\mathbb{R}^n$.

9.113 Let $U \subset \mathbb{C}^*$ be open and connected. Suppose that $\alpha$ is a continuous choice of argument on $U$. Show that the map $z \mapsto (|z|, \alpha(z))$ is a homeomorphism from $U$ to an open subset of $\mathbb{R}^+ \times \mathbb{R}$. (Hint: $\mathbb{C}^*$ is homeomorphic to $\mathbb{R}^+ \times S$ by $z \mapsto (|z|, z/|z|)$.)

[4] With a little extra work, one can also obtain a smooth path.

[5] If the manifold and the group operations are $C^\infty$ then it is called a *Lie group*.

[6] The physical interpretation is that the particle travelling along $\gamma$ does so at a possibly different speed than one travelling along $\theta$, but both trace the same curve in $X$. The assumption that $\varphi$ is strictly increasing implies that the direction of travel is the same in both cases. Re-parameterisation allows us to treat paths with different domains as essentially the same.

9.114 Let $X$ be a path-connected space, and let $x_0 \in X$. Let $G_{x_0}$ be the collection of all homotopy equivalence classes of loops $\gamma : [0, 1] \to X$ which start and end at $x_0$ (Exercise 9.12). (a) Show that concatenation of loops induces a well-defined binary operation on $G_{x_0}$. (b) Show that this operation makes $G_{x_0}$ a group. (c) Show that for any $x_0, x_1 \in X$, $G_{x_0} \cong G_{x_1}$. The isomorphism type of the groups $G_{x_0}$ is called the *fundamental group* of $X$, denoted by $\pi_1(X)$. (d) Show that $\pi_1(X)$ is trivial if and only if $X$ is simply connected. (e) Show that if $X$ and $Y$ are homeomorphic then $\pi_1(X) \cong \pi_1(Y)$. (In particular, $X$ is simply connected if and only if $Y$ is.)

9.115 (a) Show that the fundamental group $\pi_1(\mathbb{C}^*)$ of the punctured plane is isomorphic to the infinite cyclic group $(\mathbb{Z}; +)$. (b) Show that for any discrete subgroup $G$ of $\mathbb{R}^n$, $\pi_1(\mathbb{R}^n/G) \cong G$ . (In particular, $\pi_1(S) \cong \mathbb{Z}$, so the unit circle is not simply connected.)

9.116 (a) Show that for path-connected spaces $X$ and $Y$, $\pi_1(X \times Y) \cong \pi_1(X) \times \pi_1(Y)$. (b) Use this to show that the fundamental group of the torus is isomorphic to $\mathbb{Z}^2$.


# Path Integrals

# 10

The central piece of machinery that makes complex analysis work is Cauchy's Integral Formula, which says that values of analytic functions are determined by integrating the functions along loops. Such integration is also a key component in the calculus of residues. In this chapter we develop the theory of integration along paths.

In most texts, a path integral such as $\int_\gamma f\,ds$ is defined to be $\int_I f(\gamma(t))|\dot{\gamma}(t)|\,dt$, i.e., is translated back to the usual Riemann integral. Often, not much motivation is given for this definition. We will give a definition using Riemann sums, which does not assume that $\gamma$ is smooth. We will see that the objects that are really being integrated are not functions but *differential forms*.

We then investigate integration of vector fields, in particular, over simply connected domains. One of the applications will give us an alternative characterisation of the winding number for piecewise smooth loops.

## 10.1 Integrating Forms Along Paths

There are two main examples of path integrals. The first, $\int_\gamma f\,ds$ measures the area below the graph of a function $f$ defined on the image of $\gamma$. If we set $f$ to be the constant function 1, we get the length of the curve. The other, $\int_\gamma F \cdot dr$, integrates not a function but a vector field $F$. The physical intuition is that $F$ is a force field, and the integral measures the work done by the force on a particle moving along the path.

There is a framework that includes both of these path integrals, in which we integrate *generalised differential forms*. Informally speaking, the *tangent bundle* of an $n$-manifold $M$ is the result of attaching in a continuous fashion a copy of $\mathbb{R}^n$ to each point of $M$. The vectors in the copy of $\mathbb{R}^n$ attached to a point $p$ are thought of



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_10

as tangents to $M$ at $p$.[1] We do not need this generality; we will only consider *trivial bundles*, of the form $E \times \mathbb{R}^n$, where $E$ is a subset of $\mathbb{R}^n$.

A *vector field* on $M$ is a continuous choice of a tangent vector at every point. Dually, a *differential 1-form* is a continuous choice of a linear operator on each tangent space (a "covector"). In the trivial setting, it is a continuous map from $E \times \mathbb{R}^n$ to $\mathbb{R}$ such that for every point $\boldsymbol{p} \in E$, the restriction of the map to the tangent space $\{\boldsymbol{p}\} \times \mathbb{R}^n$ is linear.

For our development (originally done by Weierstrass, see [Ces58]), we relax the condition that the map on each tangent space be linear; we only require that it respects multiplication by non-negative scalars (so that we include the function taking a vector to its norm). We call these *generalised forms*. When the form is linear (as will be the case when we deal with vector fields, and later in the book with meromorphic forms on Riemann surfaces), we will usually just call them *forms*. There is a wider theory which includes differential $k$-forms for $k > 1$ as well, which can be used to integrate on parameterised manifolds, rather than just paths. See for example [Mun91, Spi65].

**Generalised Forms**

If $E \subseteq \mathbb{R}^n$ then as discussed, the "tangent bundle" is $E \times \mathbb{R}^n$; but we think of pairs $(\boldsymbol{p}, \vec{v}) \in E \times \mathbb{R}^n$ as a pair consisting of a *point* $\boldsymbol{p}$ and a *vector* $\vec{v}$.

**Definition 10.1** Let $E \subseteq \mathbb{R}^n$. A *generalised form* on $E$ is a continuous function $\omega\colon E \times \mathbb{R}^n \to \mathbb{R}$ which respects non-negative scalar multiplication on the second coordinate: for all $\boldsymbol{p} \in E$, scalar $a \geqslant 0$ and $\vec{v} \in \mathbb{R}^n$, $\omega(\boldsymbol{p}, a\vec{v}) = a \cdot \omega(\boldsymbol{p}, \vec{v})$.

If $\omega$ is a generalised form on $E$, $\boldsymbol{p} \in E$ and $\vec{v} \in \mathbb{R}^n$ then we write $\omega_{\boldsymbol{p}}(\vec{v})$ for $\omega(\boldsymbol{p}, \vec{v})$. In other words, the form defines, in a continuous way, maps $\omega_{\boldsymbol{p}}$ from the tangent space at $\boldsymbol{p}$ to $\mathbb{R}$, and these maps are each required to preserve multiplication by non-negative scalars. Note that a form is not required to be differentiable, only continuous (and so we do not assume that $E$ is open).

*Example 10.2* As mentioned above, we will work with two important examples. The second one is that of forms $F \cdot dr$ defined by vector fields on $E$. In this section our guiding example will be the generalised form $ds$, defined by letting

$$ds_{\boldsymbol{p}}(\vec{v}) = |\vec{v}|$$

for all $\boldsymbol{p} \in E$ and $\vec{v} \in \mathbb{R}^n$. (Verify it indeed satisfies Definition 10.1.) «

[1] This makes more sense if we think of the manifold embedded in $\mathbb{R}^m$ for some $m > n$.

*Example 10.3* If $\omega$ is a generalised form on $E$ and $f\colon E \to \mathbb{R}$ is continuous then the generalised form $f\omega$ is defined by letting $(f\omega)_{\boldsymbol{p}} = f(\boldsymbol{p}) \cdot \omega_{\boldsymbol{p}}$. (Verify that this is a generalised form). The main example in this section is $f\,ds$. Also, if $\omega$ and $\eta$ are generalised forms then so is $\omega + \eta$. «

## Definition of the Integral

Our next task is to explain how to integrate a generalised form on $E$ along a path. This is defined using Riemann sums, generalising the familiar definition for one dimension. Let $K = [a, b]$ be a closed interval.

A *tagged partition* of $K$ is a sequence of points $a = x_0 \leqslant r_1 \leqslant x_1 \leqslant r_2 \leqslant x_2 \leqslant \cdots \leqslant x_{k-1} \leqslant r_k \leqslant x_k = b$, such that $x_{i-1} < x_i$ for all $i = 1, \ldots, k$. If $P$ is a tagged partition then a *$P$-interval* is one of the sub-intervals $[x_0, x_1]$, $[x_1, x_2]$, $\ldots$, $[x_{k-1}, x_k]$. So we think of a tagged partition as a partition of $[a, b]$ to finitely many closed intervals, and a choice of one point (a "tag") $r_j$ in each interval $[x_{j-1}, x_j]$. We use the following notation: for a $P$-interval $I = [x_{j-1}, x_j]$, we let $r_I$ be the $P$-tag $r_j$.

For a tagged partition $P$, let $D(P)$ (the *mesh size* of $P$) be the length of the longest $P$-interval.

We define a notion of limit for functions on tagged partitions. Suppose that $\langle b_P \rangle$ is a function which associates with every tagged partition $P$ of $K$ a real number $b_P$. Let $a \in \mathbb{R}$. We say that

$$\lim_{D(P)\to 0} b_P = a$$

if the values $b_P$ approach $a$ as $P$ becomes finer (has smaller mesh size). Formally: if for all $\varepsilon > 0$ there is some $\delta > 0$ such that for every tagged partition $P$ of $K$, if $D(P) < \delta$ then $|b_P - a| < \varepsilon$. (Ensure that the usage of the equality sign is justified: there is at most one number $a$ satisfying this definition.)

We fix: a path $\gamma\colon K \to \mathbb{R}^n$ where $K = [a, b]$ is a closed interval, and a generalised form $\omega$ defined at least on the image $\gamma[K]$.

Suppose that $I = [s, t] \subseteq [a, b]$ is a subinterval. We let $|I| = t - s$ (the length of the interval). We let $\Delta_I\gamma$ be the vector $\gamma(t) - \gamma(s)$. Let $P$ be a tagged partition of $K$. The *partial sum* for $P$ is

$$S_P = S_P(\omega, \gamma) = \sum \left\{ \omega_{\gamma(r_I)}(\Delta_I\gamma) \,:\, I \text{ is a } P\text{-interval} \right\}.$$

That is, for each $P$-interval $I$, we evaluate the form $\omega$ at the tagged point $\gamma(r_I)$ on the vector $\Delta_I\gamma$, and add up the results for all the $P$-intervals. We define

$$\int_\gamma \omega = \lim_{D(P)\to 0} S_P(\omega, \gamma)$$

(if the limit exists; otherwise the integral is undefined). The reader should keep the example $\omega = f\,ds$ in mind; in that case the partial sum is defined by adding up, for each $P$-interval $I$, the product of $f(\gamma(r_I))$ and $|\Delta_I\gamma|$. What we are approximating is the area underneath the graph of $f$ above the image of the path. If $f = 1$ we are approximating the length of the path.

## Properties of the Integral

*Exercise 10.4* Suppose that $\gamma$ is a constant path. Show that for any generalised form $\omega$, $\int_\gamma \omega = 0$. «

*Exercise 10.5* Let $f : \gamma[K] \to \mathbb{R}$ be continuous. Show that

$$\left|\int_\gamma f\,ds\right| \leqslant \int_\gamma |f|\,ds$$

(provided they both exist). «

*Exercise 10.6* Show that the integral is a linear operator: (a) suppose that $\int_\gamma \omega$ exists. Then for all $a \in \mathbb{R}$, $\int_\gamma a\omega$ exists and equals $a\int_\gamma \omega$. (b) suppose that both $\int_\gamma \omega$ and $\int_\gamma \eta$ exist. Then $\int_\gamma(\omega+\eta)$ exists and equals $\int_\gamma \omega + \int_\gamma \eta$. «

**Concatenation of Paths**

**Proposition 10.7** *Let $\gamma$ and $\theta$ be two paths in $\mathbb{R}^n$ which we can concatenate (the end of $\gamma$ is the beginning of $\theta$). Let $\omega$ be a form defined (at least) on the images of $\gamma$ and $\theta$. Suppose that both $\int_\gamma \omega$ and $\int_\theta \omega$ are defined. Then $\int_{\gamma\hat{}\theta}\omega$ is defined and $\int_{\gamma\hat{}\theta}\omega = \int_\gamma\omega + \int_\theta\omega$.*

***Proof*** Suppose that $\operatorname{dom}\gamma = [a,b]$ and $\operatorname{dom}\theta = [b,c]$; let $K = [a,c]$. For brevity let $\zeta = \gamma\hat{}\theta$. We show:

(*) for all $\varepsilon > 0$ there is some $\delta > 0$ such that if $I \subseteq K$ is a sub-interval, $r \in I$ and $|I| < \delta$, then $|\omega_{\zeta(r)}(\Delta_I\zeta)| < \varepsilon$.

To see this, consider the Cartesian product $\zeta[K] \times \overline{B}(\vec{0}, 1)$ of the image of $\zeta$ with the closed unit ball in $\mathbb{R}^n$. It is compact, and so $\omega$ is uniformly continuous on that set (Proposition 8.94). In particular, for every $\varepsilon > 0$ there is some $\eta > 0$ such that for all $\boldsymbol{p} \in \zeta[K]$, for all $\vec{v}, \vec{w} \in \mathbb{R}^n$, if $|\vec{v}|, |\vec{w}| \leqslant 1$ and $|\vec{w} - \vec{v}| < \eta$ then $|\omega_{\boldsymbol{p}}(\vec{v}) - \omega_{\boldsymbol{p}}(\vec{w})| < \varepsilon$. Since $\omega_{\boldsymbol{p}}(\vec{0}) = 0$ for all $\boldsymbol{p}$, this means that for all $\boldsymbol{p} \in \zeta[K]$, for all $\vec{v}$ with $|\vec{v}| < \eta$, we have $|\omega_{\boldsymbol{p}}(\vec{v})| < \varepsilon$.

Next we use the uniform continuity of $\zeta$ to see that there is some $\delta > 0$ such that for all $s, t \in K$, if $|t - s| < \delta$ then $|\zeta(t) - \zeta(s)| < \eta$. That is, if $I \subseteq K$ and $|I| < \delta$ then $|\Delta_I\zeta| < \eta$. Now (*) follows.

Given $\varepsilon > 0$ find $\delta$ given by (*). Let $P$ be a tagged partition of $K$ with $D(P) < \delta$. We want to show that $S_P(\omega, \zeta)$ is close to $\int_\gamma \omega + \int_\theta \omega$. If $b$ is one of the end-points $x_j$ of $P$, that is, if every $P$-interval is a subset of $[a, b]$ or of $[b, c]$, then $S_P(\omega, \zeta)$ is the sum of two partial sum; informally writing, $S_P(\omega, \zeta) = S_{P\restriction_{[a,b]}}(\omega, \gamma) + S_{P\restriction_{[b,c]}}(\omega, \theta)$. By shrinking $\delta$ we may assume that $S_{P\restriction_{[a,b]}}(\omega, \gamma)$ is $\varepsilon$-close to $\int_\gamma \omega$, and similarly for $S_{P\restriction_{[b,c]}}(\omega, \theta)$ and $\int_\theta \omega$. The difficulty is when $b$ is not one of the points of $P$, in which case one $P$-interval encompasses parts of both $[a, b]$ and $[b, c]$. In this case we use (*) above to argue that the contribution of this $P$-interval is negligible.

In detail, let $\widetilde{P}$ be the tagged partition of $K$ obtained from $P$ by breaking the "bad $P$-interval" at $b$ (we add $b$ as one of the points $x_j$; the tags could be anything). By the argument above, we can ensure that $S_{\widetilde{P}} = S_{\widetilde{P}}(\omega, \zeta)$ is within $2\varepsilon$ of $\int_\gamma \omega + \int_\theta \omega$. So it suffices to show that $S_{\widetilde{P}}$ is close to $S_P = S_P(\omega, \zeta)$. The difference is in the contribution of the $P$-interval containing $b$, which in $\widetilde{P}$ is broken into two. But by (*) above, the contribution of each of these three intervals is at most $\varepsilon$, so $|S_{\widetilde{P}} - S_P| < 3\varepsilon$. □

### 10.1.1 The Length of a Path

As mentioned above, to measure the length of the image of a path we add up the lengths of the straight segments determined by the $\gamma$-images of the end-points of sub-intervals. In other words, we define the length of $\gamma$ to be

$$\ell(\gamma) = \int_\gamma ds.$$

*Exercise 10.8* Let $\gamma : K \to \mathbb{R}^n$ be a path. Show that

$$\ell(\gamma) = \sup \{ S_P(ds, \gamma) \ : \ P \text{ is a tagged partition of } K \}.$$

(Hint: show that if $Q$ refines $P$, that is, if each $P$-interval is the union of $Q$-intervals, then $S_P(ds, \gamma) \leqslant S_Q(ds, \gamma)$.) «

**Proposition 10.9** *Let $\gamma : K \to \mathbb{R}^n$ be a path and let $f : \gamma[K] \to \mathbb{R}$ be continuous. If $\int_\gamma f\, ds$ exists then*

$$\left| \int_\gamma f\, ds \right| \leqslant \ell(\gamma) \cdot \max_{\boldsymbol{p} \in \gamma[K]} |f(\boldsymbol{p})|.$$

The maximum is obtained because $\gamma[K]$ is compact; see Exercise 8.93.

***Proof*** Let $M = \max_{\boldsymbol{p} \in \gamma[K]} |f(\boldsymbol{p})|$. By Exercise 10.5 we may assume that $f \geqslant 0$ on $\gamma[K]$. Let $P$ be a tagged partition of $K$; then by Exercise 10.8,

$$0 \leqslant S_P(f\, ds, \gamma) \leqslant M \cdot S_P(ds, \gamma) \leqslant M \cdot \ell(\gamma)$$

So the result is obtained by taking the limit defining $\int_\gamma f\,ds$. □

In fact, if $\ell(\gamma)$ is finite then $\int_\gamma f\,ds$ exists; see Exercise 10.57.

## 10.2 Integrating Along Smooth Paths

Suppose that $\gamma\colon K \to \mathbb{R}^n$ is a smooth path. For all $r \in K$, the derivative vector $\dot\gamma(r) = D\gamma(r)$ is the tangent to $\gamma$ at $\gamma(r)$.

**Lemma 10.10** *Suppose that $\gamma$ is smooth; let $\omega$ be a generalised form on $\gamma[K]$. Then for every $\varepsilon > 0$ there is some $\delta > 0$ such that for any subinterval $I \subseteq K$ such that $|I| < \delta$, for any $r \in I$, and any point $t \in K$ such that $d(t, I) < \delta$,*

$$\left|\omega_{\gamma(r)}(\Delta_I\gamma) - |I| \cdot \omega_{\gamma(t)}(\dot\gamma(t))\right| < |I| \cdot \varepsilon.$$

This means: If $I$ is really small, and $t$ is close to $I$, then the contribution $\omega_{\gamma(r)}(\Delta_I\gamma)$ of $I$ to a partial sum is close to the value of $\omega$ on $\dot\gamma(t)$ at $\gamma(t)$, multiplied by the length of $I$.

Before we prove the lemma we show how we will use it.

**Theorem 10.11** *If $\gamma\colon K \to \mathbb{R}^n$ is smooth then $\int_\gamma \omega$ exists for any generalised form $\omega$ on $\gamma[K]$.*

***Proof*** We first make some simplifications. We claim that it suffices to show:

**A.** For every $\varepsilon > 0$ there is some $\delta > 0$ such that if $P$ and $Q$ are two tagged partitions of $K$ such that $D(P), D(Q) < \delta$, then $|S_P - S_Q| < \varepsilon$.

The reason for this is the completeness of $\mathbb{R}$ (see page 208). For each $n$ choose some tagged partition $P_n$ such that $D(P_n) < 1/n$. (A) implies that the sequence $\langle S_{P_n}\rangle$ is a Cauchy sequence; so it has some limit $a$. (A) again then implies that $a = \lim_{D(P)\to 0} S_P$.

As in Exercise 10.8, we say that $Q$ *refines* $P$ if every $P$-interval is the union of $Q$-intervals. If $Q$ refines $P$ then $D(Q) \leqslant D(P)$. Any two partitions have a common refinement. This implies that it suffices to show:

**B.** For every $\varepsilon > 0$ there is some $\delta > 0$ such that if $P$ and $Q$ are two tagged partitions of $K$ such that $D(P) < \delta$ and $Q$ refines $P$, then $|S_P - S_Q| < \varepsilon$.

To prove (B) we use Lemma 10.10. Given $\varepsilon > 0$ let $\delta$ be given by the lemma. Let $P$ be a tagged partition such that $D(P) < \delta$ and let $Q$ be a refinement of $P$. To avoid confusion, we use the following notation: for a $P$-interval $J$, we let $t_J$ be the $P$-tag for $J$; for a $Q$-interval $I$, let $r_I$ be the $Q$-tag for $I$.

Let $J$ be a $P$-interval; let $\mathcal{I}(J)$ be the collection of $Q$-intervals which are subintervals of $J$. For any $Q$-interval $I \subseteq J$, by Lemma 10.10,

$$\left|\omega_{\gamma(r_I)}(\Delta_I\gamma) - |I| \cdot \omega_{\gamma(t_J)}(\dot\gamma(t_J))\right| < \varepsilon \cdot |I|;$$

adding up all $Q$-intervals $I \subseteq J$ we get

$$\left|\sum_{I\in\mathcal{I}(J)} \omega_{\gamma(r_I)}(\Delta_I\gamma) - |J| \cdot \omega_{\gamma(t_J)}(\dot\gamma(t_J))\right| < \varepsilon \cdot |J|.$$

On the other hand, using the lemma for $J$ we get

$$\left|\omega_{\gamma(t_J)}(\Delta_J\gamma) - |J| \cdot \omega_{\gamma(t_J)}(\dot\gamma(t_J))\right| < \varepsilon \cdot |J|,$$

so overall

$$\left|\omega_{\gamma(t_J)}(\Delta_J\gamma) - \sum_{I\in\mathcal{I}(J)} \omega_{\gamma(r_I)}(\Delta_I\gamma)\right| < 2\varepsilon \cdot |J|.$$

Now adding up for all $P$-intervals $J$ gives us $|S_P - S_Q| < 2\varepsilon|K|$. □

***Proof of Lemma 10.10*** Let $\varepsilon > 0$.

Since the function $s \mapsto |\dot\gamma(s)|$ is continuous on $K$, it is bounded by some $M > 0$ (Exercise 8.93). The Cartesian product $\gamma[K] \times \overline{B}(\vec{0}, M+1)$ is compact. Since $\omega$ is continuous, it is uniformly continuous on that set (Proposition 8.94). So there is some $\eta > 0$ such that for all $s_1, s_2 \in K$ and all $\vec{v}_1, \vec{v}_2 \in \mathbb{R}^n$ such that $|\vec{v}_1|, |\vec{v}_2| \leqslant M+1$, if $|\gamma(s_2) - \gamma(s_1)| < \eta$ and $|\vec{v}_2 - \vec{v}_1| < 2\eta$ then $|\omega_{\gamma(s_2)}(\vec{v}_2) - \omega_{\gamma(s_1)}(\vec{v}_1)| < \varepsilon$. We may assume that $\eta \leqslant 1/2$.

Using the compactness of $K$ and smoothness of $\gamma$, and Proposition 9.49, we see that there is some $\delta > 0$ such that if $I \subseteq K$, $r \in I$ and $|I| < 2\delta$, then: (a) $|\Delta_I\gamma| < \eta$; (b) $|\Delta_I\dot\gamma| < \eta$; and (c) $|\Delta_I\gamma - |I| \cdot \dot\gamma(r)| < |I| \cdot \eta$.

Let $I \subseteq K$ and $r \in I$; suppose that $|I| < \delta$ and $d(t, I) < \delta$. Then $|\dot\gamma(t) - \dot\gamma(r)| < \eta$, and $|\Delta_I\gamma/|I| - \dot\gamma(r)| < \eta$, so $|\Delta_I\gamma/|I| - \dot\gamma(t)| < 2\eta$. Since $|\dot\gamma(t)| \leqslant M$ and $2\eta \leqslant 1$ we have $|\Delta_I\gamma/|I|| \leqslant M+1$. Since $|\gamma(r) - \gamma(t)| < \eta$ we conclude that

$$\left|\omega_{\gamma(r)}(\Delta_I\gamma/|I|) - \omega_{\gamma(t)}(\dot\gamma(t))\right| < \varepsilon.$$

Since $\omega_{\gamma(r)}$ respects multiplication by the non-negative scalar $|I|$, multiplying throughout by $|I|$ yields the desired inequality. □

Recalling Definition 9.80, now Proposition 10.7 implies:

**Corollary 10.12** *If $\gamma : K \to \mathbb{R}^n$ is piecewise smooth then $\int_\gamma \omega$ exists for any generalised form $\omega$ on $\gamma[K]$.*

### 10.2.1 Linear Forms

A generalised form $\omega$ is *linear* if for every $\boldsymbol{p}$, $\omega_{\boldsymbol{p}} \colon \mathbb{R}^n \to \mathbb{R}$ is a linear map. As discussed above, henceforth we will often call linear forms simply *forms*. Among the linear forms, the most basic ones are the forms $dx_i$ (for $i = 1, 2, \ldots, n$), which return the $i$th component of the vector: if $\vec{v} = (v_1, v_2, \ldots, v_n)$ then for all $\boldsymbol{p}$, $(dx_i)_{\boldsymbol{p}}(\vec{v}) = v_i$. The integral $\int_\gamma f\, dx_i$ is similar to $\int f\, ds$, except that only the component of change of $\gamma(t)$ along the $i$th axis is considered, and the contributions can be either positive or negative. When $n = 1$, we write $dx$ for $dx_1$. When $n = 2$, we write $dx$ and $dy$ for $dx_1$ and $dx_2$. For an interval $[a, b]$ and a continuous function $f \colon [a, b] \to \mathbb{R}$ we define

$$\int_a^b f\, dt = \int_\gamma f\, dx$$

where $\gamma \colon [a, b] \to [a, b]$ is the identity path $\gamma(t) = t$. Unravelling the definitions, we observe that this is the familiar notion of the Riemann integral. The path $\gamma$ is smooth, and so the integral exists by Theorem 10.11.

*Exercise 10.13* Let $\gamma \colon [a, b] \to \mathbb{R}^n$ be a path and let $\omega$ be a linear form defined on the image of $\gamma$. Suppose that $\int_\gamma \omega$ exists. Let $-\gamma \colon [-b, -a] \to \mathbb{R}^n$ be the path defined by $-\gamma(t) = \gamma(-t)$. Show that $\int_{-\gamma} \omega = -\int_\gamma \omega$. «

*Remark 10.14* If $a < b$ then we also define $\int_b^a f\, dt = \int_\gamma f\, dx$ where this time $\gamma$ is the path which travels from $b$ to $a$ in constant speed 1. Then Exercise 10.13 implies that $\int_b^a f\, dt = -\int_a^b f\, dt$. «

*Remark 10.15* If $a < b$ then $\int_a^b f\, dt = \int_\gamma f\, ds$ for $\gamma(t) = t$. In fact if $\gamma$ is non-decreasing then $\int_\gamma f\, ds = \int_\gamma f\, dx$; this is because $ds_{\boldsymbol{p}}(r) = r = dx_{\boldsymbol{p}}(r)$ for all $r \geqslant 0$. On the other hand, $\int_b^a f\, dt$ does not equal $\int_\gamma f\, ds$ for $\gamma$ travelling from $b$ to $a$ as above, rather, it equals $-\int_\gamma f\, ds$. «

*Exercise 10.16* Show that if $f \colon [a, b] \to \mathbb{R}$ is continuous and $m \leqslant f(t) \leqslant M$ for all $t \in [a, b]$ then $m(b - a) \leqslant \int_a^b f\, dt \leqslant M(b - a)$. «

### 10.2.2 Relating the General and Familiar Integrals

If $\gamma$ is smooth then integrating along $\gamma$ can be reduced to the usual straight integral, by evaluating the (generalised) form at every point on the tangent to the curve at that point.

**Proposition 10.17** *If $\gamma$ is smooth then for any generalised form $\omega$ on its image,*

$$\int_\gamma \omega = \int_a^b \omega_{\gamma(t)}(\dot\gamma(t))\,dt.$$

(On the right hand side $\omega$ does not play the role of the "differential", rather we are integrating the function $\omega_{\gamma(t)}(\dot\gamma(t)) \colon [a,b] \to \mathbb{R}$.)

***Proof*** Let $\varepsilon > 0$; let $\delta > 0$ be given by Lemma 10.10. Let $P$ be a tagged partition of $K$ such that $D(P) < \delta$; let $I$ be a $P$-interval. So

$$\left|\omega_{\gamma(r_I)}(\Delta_I\gamma) - |I| \cdot \omega_{\gamma(r_I)}(\dot\gamma(r_I))\right| < |I| \cdot \varepsilon.$$

Adding up for all $P$-intervals we get that the difference between $S_P(\omega,\gamma)$ and the partial sum given by $P$ for the integral on the right hand side is bounded by $\varepsilon \cdot |K|$. Taking the limit we obtain equality. □

In particular, if $\gamma \colon [a,b] \to \mathbb{R}^n$ is smooth then Proposition 10.17 says that

$$\ell(\gamma) = \int_a^b |\dot\gamma(t)|\,dt.$$

*Example 10.18* Define $\gamma \colon [0, 2\pi] \to S$ by defining $\gamma(t) = (\cos t, \sin t)$. Then $\dot\gamma(t) = (-\sin t, \cos t)$ so $|\dot\gamma(t)| = 1$. Hence the length of the path (which is the circumference of the unit circle) is $\int_0^{2\pi} dt = 2\pi$. «

Proposition 10.17 also implies that for $j = 1, 2, \ldots, n$, when $\gamma$ is smooth then

$$\int_\gamma f\,dx_j = \int_a^b f(\gamma(t)) \cdot D^j\gamma(t)\,dt.$$

### The Fundamental Theorem of Calculus

This has two parts.

**Theorem 10.19** *If $f \colon [a,b] \to \mathbb{R}$ is continuous then $g(x) = \int_a^x f\,dt$ is smooth and $\dot g = f$.*

(Again note that by *smooth* we mean $C^1$ (continuously differentiable), not $C^\infty$.)

***Sketch of Proof*** For $h > 0$, by Proposition 10.7,

$$g(x+h) - g(x) = \int_x^{x+h} f\,dt;$$

we need to show that

$$\left(\frac{1}{h}\int_x^{x+h} f\,dt\right) - f(x) = \frac{1}{h}\int_x^{x+h} (f(t) - f(x))\,dt \to 0$$

as $h \to 0$. However by Exercise 10.16

$$\frac{1}{h}\left|\int_x^{x+h} (f(t) - f(x))\,dt\right| \leqslant \frac{1}{h}\int_x^{x+h} |f(t) - f(x)|\,dt \leqslant \max_{t\in[x,x+h]} |f(t) - f(x)|,$$

which approaches 0 as $f$ is continuous at $x$. When $h < 0$ we apply the same argument but need to be careful with the signs. □

*Exercise 10.20* Let $\gamma\colon [a,b] \to \mathbb{R}^n$ be smooth and let $f\colon \gamma[a,b] \to \mathbb{R}$ be continuous. Define $g\colon [a,b] \to \mathbb{R}$ by letting $g(t) = \int_{\gamma\restriction_{[a,t]}} f\,ds$. Show that $g$ is smooth and that

$$\dot{g}(t) = f(\gamma(t)) \cdot |\dot{\gamma}(t)|.$$ «

The second part is the following:

**Theorem 10.21** *Let $g\colon [a,b] \to \mathbb{R}$ be smooth. Then*

$$\int_a^b \dot{g}\,dt = g(b) - g(a).$$

***Proof*** We consider $g$ as a smooth path from $[a,b]$ to $\mathbb{R}$. For any sub-interval $[c,d] \subseteq [a,b]$, we have

$$dx_{g(c)}(\Delta_{[c,d]}g) = g(d) - g(c)$$

as $dx_p$ is the identity on $\mathbb{R}$ at every point $p$. This shows that for any tagged partition $P$ of $[a,b]$, $S_P(dx, g)$ equals $g(b) - g(a)$ (it is a telescopic sum); and so $\int_g dx = g(b) - g(a)$.

On the other hand, Proposition 10.17 implies that $\int_g dx = \int_a^b \dot{g}\,dt$. □

Below we give a generalisation (Proposition 10.29).

### Derivative of an Integral Depending on a Parameter

Below we will need to apply Leibniz's integral rule about "differentiating under the integral sign". Here we quickly recall this rule. In the following let $[a,b] \times [c,d]$ be

a closed rectangle in $\mathbb{R}^2$. When we integrate sections, we use the standard notation $\int_a^b f(x, y)\, dx$ (for fixed $y \in [c, d]$), rather than using $dt$.

**Lemma 10.22** *Let $f : [a, b] \times [c, d] \to \mathbb{R}$ be continuous. Then the function $g(y) = \int_a^b f(x, y)\, dx$ is continuous on $[c, d]$.*

***Proof*** Since $[a, b] \times [c, d]$ is compact, $f$ is uniformly continuous (Proposition 8.94). Given $\varepsilon > 0$ find $\delta > 0$ such that for all $\boldsymbol{p}, \boldsymbol{q} \in [a, b] \times [c, d]$, if $|\boldsymbol{q} - \boldsymbol{p}| < \delta$ then $|f(\boldsymbol{q}) - f(\boldsymbol{p})| < \varepsilon$. If $y_1, y_2 \in [c, d]$ and $|y_2 - y_1| < \delta$ then by the linearity of the integral (Exercise 10.6) and by Exercise 10.5,

$$\big|g(y_2) - g(y_1)\big| = \left|\int_a^b f(x, y_2)\, dx - \int_a^b f(x, y_1)\, dx\right| =$$

$$\left|\int_a^b \big(f(x, y_2) - f(x, y_1)\big)\, dx\right| \leqslant \int_a^b \big|f(x, y_2) - f(x, y_1)\big|\, dx \leqslant \varepsilon \cdot (b - a),$$

since $|f(x, y_2) - f(x, y_1)| < \varepsilon$ for all $x \in [a, b]$; we use Proposition 10.9. □

**Proposition 10.23** *Let $f : [a, b] \times [c, d] \to \mathbb{R}$ be smooth. Then the function $g(y) = \int_a^b f(x, y)\, dx$ is smooth on $[c, d]$ and $\dot{g}(y) = \int_a^b D^y f(x, y)\, dx$.*

***Proof*** It suffices to show that $g$ is differentiable on $[c, d]$ and $\dot{g}(y) = \int_a^b D^y f(x, y)\, dx$; smoothness then follows from Lemma 10.22 since $D^y f$ is assumed to be continuous on $[a, b] \times [c, d]$.

Let $\varepsilon > 0$ and let $y \in [c, d]$. Given $h$ small, by the linearity of the integral and Exercise 10.5,

$$\left|g(y + h) - g(y) - h \cdot \int_a^b D^y f(x, y)\, dx\right| =$$

$$\left|\int_a^b \big(f(x, y + h) - f(x, y) - h \cdot D^y f(x, y)\big)\, dx\right| \leqslant$$

$$\int_a^b \big|f(x, y + h) - f(x, y) - h \cdot D^y f(x, y)\big|\, dx \leqslant \varepsilon |h| \cdot (b - a),$$

once we show that for small $h$,

$$\big|f(x, y + h) - f(x, y) - h \cdot D^y f(x, y)\big| \leqslant \varepsilon |h|$$

for all $x \in [a, b]$, which follows from Proposition 9.49. □

## 10.3 Integrating Vector Fields

A *vector field* on a set $E \subseteq \mathbb{R}^n$ is a continuous function $F\colon E \to \mathbb{R}^n$.[2] The form associated with a vector field $F$ is the form $F \cdot dr$ defined by

$$(F \cdot dr)_{\boldsymbol{p}}(\vec{v}) = F(\boldsymbol{p}) \cdot \vec{v}.$$

Here we use physics notation: the *dot product* $\vec{v} \cdot \vec{w}$ of two vectors $\vec{v}, \vec{w} \in \mathbb{R}^n$ is $\sum_{i \leqslant n} v_i \cdot w_i$, where $\vec{v} = (v_1, \ldots, v_n)$ and $\vec{w} = (w_1, \ldots, w_n)$. [3]

*Exercise 10.24* Show that a generalised form $\omega$ on $E$ is linear if and only if it is $F \cdot dr$ for some (unique) vector field $F$ on $E$. «

Proposition 10.17 implies that if $\gamma\colon [a, b] \to E$ is smooth then

$$\int_\gamma F \cdot dr = \int_a^b F(\gamma(t)) \cdot \dot{\gamma}(t)\, dt.$$

In other words, we are integrating the signed length of the projection of $F(\gamma(t))$ onto the tangent to $\gamma$ at $\gamma(t)$; the length is positive if the angle between $F(\gamma(t))$ and $\dot{\gamma}(t)$ is smaller than a right angle, 0 if $F(\gamma(t))$ and $\dot{\gamma}(t)$ are perpendicular, and negative otherwise.[4]

Note that for $i = 1, 2, \ldots, n$, $dx_i = F \cdot dr$ for the constant vector field $F$ mapping every point to the $i$th unit vector $\vec{e}_i$. If $F$ is a vector field then we write $F = (F_1, F_2, \ldots, F_n)$ where $F_i\colon E \to \mathbb{R}$ is the function giving the $i$th component of $F$; then $F \cdot dr = \sum_i F_i\, dx_i$.

**Lemma 10.25** *Let $\gamma\colon K \to \mathbb{R}^n$ be a path and let $F$ be a vector field on $\gamma[K]$. Then*

$$\left| \int_\gamma F \cdot dr \right| \leqslant \ell(\gamma) \cdot \max_{\boldsymbol{p} \in \gamma[K]} |F(\boldsymbol{p})|$$

*(provided the integral exists).*

***Proof*** This is of course similar to the proof of Proposition 10.9. The added ingredient is the Cauchy-Schwarz inequality (see Exercise 8.118). Let $M = \max_{\boldsymbol{p} \in \gamma[K]} |F(\boldsymbol{p})|$. Let $I \subseteq K$ be an interval; let $r \in I$; let $\boldsymbol{p} = \gamma(r)$. Then by

[2] In terms of the tangent bundle, this is a continuous choice of one tangent vector at each point.

[3] Other notation and terminology (the *inner product*) was used in Exercise 8.118; of course $\vec{v} \cdot \vec{w} = \vec{v}\vec{w}^{\mathrm{t}}$, if we're thinking of tangent vectors as rows.

[4] In physical terms, $F$ is a force field; $\int F \cdot dr$ is the work done as a particle is travelling through the field.

said inequality,

$$|F(\boldsymbol{p}) \cdot \Delta_I \gamma| \leqslant |F(\boldsymbol{p})| \cdot |\Delta_I \gamma| \leqslant M \cdot |\Delta_I \gamma|.$$

So for any tagged partition $P$ of $K$, $|S_P(F \cdot dr, \gamma)| \leqslant M \cdot S_P(ds, \gamma)$, which in the limit gives the desired inequality. (If we assume that $\gamma$ is smooth then we can use Proposition 10.17 and Cauchy-Schwarz to reduce the lemma to Proposition 10.9.) □

### 10.3.1 Conservative Vector Fields

**For the rest of this chapter, unless otherwise specified, by "path" we mean "piecewise smooth path"**, and appeal to Corollary 10.12.

**Definition 10.26** Let $U \subseteq \mathbb{R}^n$ be a region (open and connected). A vector field $F$ on $U$ is *conservative* if for any two points $\boldsymbol{a}, \boldsymbol{b} \in U$ and any two paths $\gamma$ and $\delta$ in $U$ from $\boldsymbol{a}$ to $\boldsymbol{b}$, $\int_\gamma F \cdot dr = \int_\delta F \cdot dr$.

That is, the work done by the field does not depend on the path taken from $\boldsymbol{a}$ to $\boldsymbol{b}$.

*Exercise 10.27* Recall that a loop is a path whose start point is also its end-point. Show that a vector field $F$ on $U$ is conservative if and only if for any loop $\gamma$ in $U$, $\int_\gamma F \cdot dr = 0$. «

The term "conservative" again is from physics. The idea is that the force field $F$ has an underlying potential energy and that this energy is conserved. In mathematical terms this says that $F$ is a *gradient field*:

**Proposition 10.28** *Let $U \subseteq \mathbb{R}^n$ be a region. A vector field $F$ on $U$ is conservative if and only if there is a smooth function $f : U \to \mathbb{R}$ such that $F = Df$.*

(Here again $Df$ is the full derivative of $f$, which in this context is often called the gradient of $f$ and is often denoted by $\nabla f$. The function $f$ is called a *potential* of $F$.) To prove Proposition 10.28 we use the following generalisation of the second part of the fundamental theorem of calculus.

**Proposition 10.29** *Let $U \subseteq \mathbb{R}^n$ be a region and let $f : U \to \mathbb{R}$ be smooth. Then for any path $\gamma : [a, b] \to U$,*

$$\int_\gamma Df \cdot dr = f(\gamma(b)) - f(\gamma(a)).$$

***Proof*** By breaking up into a finite sum, we may assume that $\gamma$ is smooth. Let $g = f \circ \gamma$ and let $F = Df$. By the chain rule, $\dot{g} = \dot{\gamma} \cdot (F \circ \gamma) = \sum_{j \leqslant n} D^j \gamma \cdot (F_j \circ \gamma)$. By the second part of the fundamental theorem of calculus (Theorem 10.21), $\int_a^b \dot{g}\, dt = g(b) - g(a) = f(\gamma(b)) - f(\gamma(a))$; and

$$\int_\gamma F \cdot dr = \sum_{j \leqslant n} \int_\gamma F_j \, dx_j = \sum_{j \leqslant n} \int_a^b F_j(\gamma(t)) \cdot D^j \gamma(t) \, dt = \int_a^b \dot{g} \, dt. \qquad \square$$

***Proof of Proposition 10.28*** One direction is covered by Proposition 10.29: if $F = Df$ for some smooth $f : U \to \mathbb{R}$ then for any two points $\boldsymbol{a}, \boldsymbol{b} \in U$, for any path $\gamma$ in $U$ from $\boldsymbol{a}$ to $\boldsymbol{b}$, $\int_\gamma F \cdot dr = f(\boldsymbol{b}) - f(\boldsymbol{a})$, which manifestly does not depend on $\gamma$.

For the other direction, suppose that $F$ is conservative. Fix some point $\boldsymbol{p} \in U$. For any point $\boldsymbol{q} \in U$ let $f(\boldsymbol{q}) = \int_\gamma F \cdot dr$ where $\gamma$ is any (piecewise smooth) path in $U$ from $\boldsymbol{p}$ to $\boldsymbol{q}$ (recall that since $U$ is connected and open, it is path-connected and in fact there is a smooth path in $U$ from $\boldsymbol{p}$ to $\boldsymbol{q}$, see Propositions 9.10 and 9.75).

We need to show that $f$ is differentiable at every point of $U$ and that $Df = F$; we show that $D^j f = F_j$ for all $j \leqslant n$ and appeal to Proposition 9.52, recalling that $F$ is continuous. For simplicity of notation let $j = 1$. Fix $\boldsymbol{q} \in U$; for $t \in \mathbb{R}$ let $\boldsymbol{t} = (t, 0, 0, \ldots, 0)$. Let $h \in \mathbb{R}$. Then $f(\boldsymbol{q} + \boldsymbol{h}) - f(\boldsymbol{q})$ is $\int_\gamma F \cdot dr$ where $\gamma$ is any path from $\boldsymbol{q}$ to $\boldsymbol{q} + \boldsymbol{h}$; for example we take the obvious linear path $\gamma : [0, h] \to \mathbb{R}^n$ given by $\gamma(t) = \boldsymbol{q} + \boldsymbol{t}$; for small enough $h$, the image of this path is contained in $U$. Now $\dot{\gamma} = (1, 0, \ldots, 0)$ at every $t$ and so $\int_\gamma F \cdot dr = \int_0^h F_1(\gamma(t))\, dt$. Let $g(x) = \int_0^x F_1(\boldsymbol{q} + \boldsymbol{t})\, dt$. Then by the first part of the fundamental theorem of calculus (Theorem 10.19), as $F_1$ is continuous,

$$D^1 f(\boldsymbol{q}) = \lim_{h \to 0} \frac{1}{h}(f(\boldsymbol{q} + \boldsymbol{h}) - f(\boldsymbol{q})) = \lim_{h \to 0} \frac{1}{h} \int_0^h F_1(\boldsymbol{q} + \boldsymbol{t})\, dt = \dot{g}(0) = F_1(\boldsymbol{q}).$$

$\square$

*Example 10.30* Define a vector field $G$ on the punctured plane $\mathbb{R}^2 \setminus \{\boldsymbol{0}\}$ by letting $G(\boldsymbol{p}) = \boldsymbol{p}/|\boldsymbol{p}|^2$, that is,

$$G(x, y) = \left( \frac{x}{x^2 + y^2}, \frac{y}{x^2 + y^2} \right).$$

Then $G = Dg$, where $g(x, y) = 1/2 \cdot \ln(x^2 + y^2)$. Hence $G$ is conservative. «

### 10.3.2 The Winding Number Revisited

Now define a vector field $F_{\mathtt{wind}}$ on $\mathbb{R}^2 \setminus \{\mathbf{0}\}$ by letting

$$F_{\mathtt{wind}}(x, y) = \left( \frac{-y}{x^2 + y^2}, \frac{x}{x^2 + y^2} \right).$$

We will shortly observe that $F_{\mathtt{wind}}$ is not conservative (as an hors-d'œuvre, consider $\gamma(t) = (\cos t, \sin t)$ on $[0, 2\pi]$, and check that $F_{\mathtt{wind}}(\gamma(t)) \cdot \dot{\gamma}(t) = 1$ for all $t$, so $\int_\gamma F_{\mathtt{wind}} \cdot dr = 2\pi$.) Locally, however, we can find a potential function for $F_{\mathtt{wind}}$.

Let $\boldsymbol{p}_0 \in \mathbb{R}^2 \setminus \{\mathbf{0}\}$; let $r_0 = |\boldsymbol{p}_0|$ and choose some argument $t_0$ for $\boldsymbol{p}$. Writing in real variables, the map $(r, t) \mapsto re^{it}$ is the map $(r, t) \mapsto (r \cos t, r \sin t)$—the translation from polar to Cartesian coordinates. This is a smooth function on $\mathbb{R}^+ \times \mathbb{R}$ and the determinant of its derivative is $r(\cos^2 t + \sin^2 t) = r \neq 0$. Hence, by the Inverse Function Theorem, there are open neighbourhoods $V$ of $(r_0, t_0)$ and $U$ of $\boldsymbol{p}_0$ between which this map is a homeomorphism, with a smooth inverse. The inverse is the map $\boldsymbol{p} \mapsto (|\boldsymbol{p}|, \alpha(\boldsymbol{p}))$ for some continuous choice of argument $\alpha$ on $U$ (Definition 9.25). Computing the inverse of the derivative of $(r, t) \mapsto (r \cos t, r \sin t)$, we get $D\alpha = F_{\mathtt{wind}}$. With the uniqueness of liftings (Lemma 9.19) we conclude:

**Proposition 10.31** *For any open $U \subseteq \mathbb{R}^2 \setminus \{\mathbf{0}\}$, any continuous choice of argument $\alpha$ on $U$ is smooth, and $D\alpha = F_{\mathtt{wind}}$ on $U$.*

In fact, the vector field $F_{\mathtt{wind}}$ can be used to compute the winding number of loops in the punctured plane (Definition 9.27), at least when they are piecewise smooth.

**Proposition 10.32** *If $\gamma$ is a loop in $\mathbb{R}^2 \setminus \{\mathbf{0}\}$ then*

$$\int_\gamma F_{\mathtt{wind}} \cdot dr = 2\pi m$$

*where $m$ is the winding number of $\gamma$.*

***Proof*** We prove the slightly more general statement: for any (piecewise smooth) path $\gamma : [a, b] \to \mathbb{R}^2 \setminus \{\mathbf{0}\}$ and any continuous choice of argument $\theta$ for $\gamma$,

$$\int_\gamma F_{\mathtt{wind}} \cdot dr = \theta(b) - \theta(a).$$

Let $\theta$ be such a choice of argument. Every point in $\mathbb{R}^2 \setminus \{\mathbf{0}\}$ has a simply connected open neighbourhood in the punctured plane (say a small disc). By compactness of $[a, b]$, there is a partition $\{a = t_0 < t_1 < \cdots < t_k = b\}$ of $[a, b]$ such that

for all $i = 1, \dots, k$, $\gamma[t_{i-1}, t_i] \subset U_i$ where $U_i \subset \mathbb{R}^2 \setminus \{\mathbf{0}\}$ is simply connected. By Proposition 9.30, for each $i$ there is a continuous choice of argument $\alpha_i$ on $U_i$. By shifting, we may assume that $\alpha_i(\gamma(t_i)) = \theta(t_i)$; then by Lemma 9.19, $\theta(t) = \alpha_i(\gamma(t))$ for all $t \in [t_{i-1}, t_i]$. Since $D\alpha_i = F_{\mathtt{wind}}$ on $U_i$, by Proposition 10.29,

$$\int_{\gamma\restriction_{[t_{i-1},t_i]}} F_{\mathtt{wind}} \cdot dr = \alpha_i(\gamma(t_{i+1})) - \alpha_i(\gamma(t_i)) = \theta(t_{i+1}) - \theta(t_i);$$

we now sum up for all $i \leqslant k$. □

## 10.4 Symmetric Vector Fields

We find a criterion for conservativity which is easier to verify. We restrict our discussion to two dimensions, because this is what we will use when we consider complex functions. Fix a region $U \subseteq \mathbb{R}^2$ and let $F : U \to \mathbb{R}^2$ be a *smooth* vector field; we write $F = (F_x, F_y)$. Suppose that $F$ is conservative; so $F = Df$ for some smooth $f : U \to \mathbb{R}$. Since $f$ is twice smooth,

$$D^y F_x = D^{yx} f = D^{xy} f = D^x F_y$$

(Proposition 9.60). We call a smooth vector field $F$ *symmetric* if $D^y F_x = D^x F_y$, that is, if $DF$ is a symmetric matrix at every point. So every smooth conservative vector field is symmetric. In general the converse does not hold; the vector field $F_{\mathtt{wind}}$ used to define the winding number is symmetric.

Our aim now is to prove the following theorem. In more abstract terminology, it says that the first de Rahm cohomology group of a simply connected region in $\mathbb{R}^2$ is trivial.

**Theorem 10.33** *Suppose that $F$ is a symmetric vector field defined on a simply connected region. Then $F$ is conservative.*

*Remark 10.34* Let $U \subseteq \mathbb{R}^2 \setminus \{\mathbf{0}\}$ be a region. By Proposition 10.32, $F_{\mathtt{wind}}$ is conservative on $U$ if and only if every loop in $U$ has winding number 0. Thus, Theorem 10.33 generalises the fact that if $U$ is simply connected then every loop in $U$ has winding number 0 (Propositions 9.30 and 9.32). «

The proof of Theorem 10.33 passes through the concept of local conservation. A vector field $F : U \to \mathbb{R}^n$ is *locally conservative* if every point $\boldsymbol{p} \in U$ has a neighbourhood $V \subseteq U$ such that the restriction $F\restriction_V$ of $F$ to $V$ is conservative. As mentioned, the vector field $F_{\mathtt{wind}}$ above is locally conservative. This is not an accident:

**Proposition 10.35** *Every symmetric vector field is locally conservative.*

The proof of Theorem 10.33 is then completed by the following, which is not restricted to two dimensions:

**Proposition 10.36** *If $U \subseteq \mathbb{R}^n$ is a simply connected region and $F\colon U \to \mathbb{R}^n$ is a locally conservative vector field then $F$ is conservative.*

***Proof of Proposition 10.35*** Let $U \subseteq \mathbb{R}^2$ and let $F\colon U \to \mathbb{R}^2$ be symmetric. Given $\boldsymbol{p} = (p_x, p_y) \in U$ let $V$ be an open rectangle such that $\boldsymbol{p} \in V$ and $V \subseteq U$. We show that $F\restriction_V$ is conservative by showing it is a gradient field (Proposition 10.28). For a point $\boldsymbol{q} \in V$ we let $\gamma_{\boldsymbol{q}}$ be the path from $\boldsymbol{p}$ to $\boldsymbol{q}$ which is travelled in constant speed first parallel to the $x$-axis and then parallel to the $y$-axis: the concatenation of $\gamma\colon [p_x, q_x] \to \mathbb{R}^2$ defined by $\gamma(t) = (t, p_y)$ and $\delta\colon [p_y, q_y] \to \mathbb{R}^2$ defined by $\delta(t) = (q_x, t)$.[5] The path $\gamma_{\boldsymbol{q}}$ is piecewise smooth. We define $f(\boldsymbol{q}) = \int_{\gamma_{\boldsymbol{q}}} F \cdot dr$. This works out to be

$$f(\boldsymbol{q}) = \int_{p_x}^{q_x} F_x(t, p_y)\, dt + \int_{p_y}^{q_y} F_y(q_x, t)\, dt.$$

Fixing $q_x$ and varying $q_y$, the first integral is constant and by the fundamental theorem of calculus (Theorem 10.19)

$$D^y f(\boldsymbol{q}) = D^y \left( \int_{p_y}^{y} F_y(q_x, t)\, dt \right)(q_y) = F_y(\boldsymbol{q}).$$

Similarly,

$$D^x \left( \int_{p_x}^{x} F_x(t, p_y)\, dt \right)(q_x) = F_x(q_x, p_y).$$

By Leibniz's integration rule (Proposition 10.23), as we keep $q_y$ constant and vary $q_x$,

$$D^x \left( \int_{p_y}^{q_y} F_y(x, t)\, dt \right)(q_x) = \int_{p_y}^{q_y} D^x F_y(q_x, t)\, dt = \int_{p_y}^{q_y} D^y F_x(q_x, t)\, dt = F_x(\boldsymbol{q}) - F_x(q_x, p_y),$$

[5] We follow the convention that if $b < a$ then $[a, b]$ denotes $[b, a]$ but the path starts at time $a$ and ends at time $b$ nonetheless.

the last equality following from the second part of the fundamental theorem of calculus (Theorem 10.21); here of course we finally use the symmetry of $F$. So overall $D^x f(\boldsymbol{q}) = F_x(q_x, p_y) + F_x(\boldsymbol{q}) - F_x(q_x, p_y) = F_x(\boldsymbol{q})$. To sum up, $D^x f$ and $D^y f$ exist at every point in the rectangle and equal $F_x$ and $F_y$ respectively. Since $F$ is continuous, Proposition 9.52 says that $f$ is differentiable on the rectangle and that $Df = F$, whence $F\restriction_V$ is conservative. □

Proposition 10.36 follows from the following, recalling that piecewise smooth homotopies characterise simple connectedness (Proposition 9.83).

**Lemma 10.37** *Let $U \subseteq \mathbb{R}^n$ be a region, let $F\colon U \to \mathbb{R}^n$ be a locally conservative vector field, and let $H\colon [0,1] \times [a,b] \to U$ be a piecewise smooth homotopy. Then $\int_{H_0} F \cdot dr = \int_{H_1} F \cdot dr$.*

***Proof*** We will show that every $u \in [0,1]$ has a neighbourhood $O \subseteq [0,1]$ such that the value $\int_{H_s} F \cdot dr$ is constant for $s \in O$. By compactness of $[0,1]$ we can cover $[0,1]$ by finitely many open intervals on which $\int_{H_s} F \cdot dr$ is constant; this implies that $\int_{H_s} F \cdot dr$ is constant on $[0,1]$. Note that each $H_s$ is piecewise smooth as $H$ is piecewise smooth. Hence the integrals are all defined.

Fix $u \in [0,1]$. For all $c \in [a,b]$, $H(u,c)$ has a neighbourhood in $U$ on which $F$ is conservative. Since $[a,b]$ is compact we can find a partition $a = t_0 < t_1 < \cdots < t_k = b$ of $[a,b]$ such that $F$ is conservative on a neighbourhood $U_i$ of $H_u[I_i]$, where $I_i = [t_{i-1}, t_i]$. For every $i$, since $H_u[I_i]$ is compact, there is a positive distance from $H_u[I_i]$ to the complement of $U_i$; since $H$ is uniformly continuous, there is some $\delta_i > 0$ such that $H_s[I_i] \subseteq U_i$ for every $s$ with $|s-u| < \delta_i$. Let $\delta = \min\{\delta_i : i = 1, \ldots, k\}$.

Now take some $s$ such that $|s-u| < \delta$. We show that $\int_{H_u} F \cdot dr = \int_{H_s} F \cdot dr$. This is done by breaking the integrals up and presenting their difference as a telescopic sum of zeros. For $i \leqslant k$ we let $\zeta_i$ be the image under $H$ of the loop which runs at a constant speed along the boundary of the rectangle $[u,s] \times [t_{i-1}, t_i]$ and starts and ends at $(u, t_{i-1})$. That is,

$$\zeta_i = \left(H_u\restriction_{[t_{i-1},t_i]}\right)^\frown \left(H^{t_i}\restriction_{[u,s]}\right)^\frown \left(-H_s\restriction_{[t_{i-1},t_i]}\right)^\frown \left(-H^{t_{i-1}}\restriction_{[u,s]}\right),$$

where we recall that $H^d(t) = H(t,d)$ and for any path $\xi$, $-\xi$ denotes the reverse path travelled from the end-point of $\xi$ to the start point of $\xi$. See Fig. 10.1. The path $\zeta_i$ is piecewise smooth; here we use the piecewise smoothness of $H$. Further, the range of $\zeta_i$ is contained in $U_i$, on which $F$ is conservative; so $\int_{\zeta_i} F \cdot dr = 0$. On

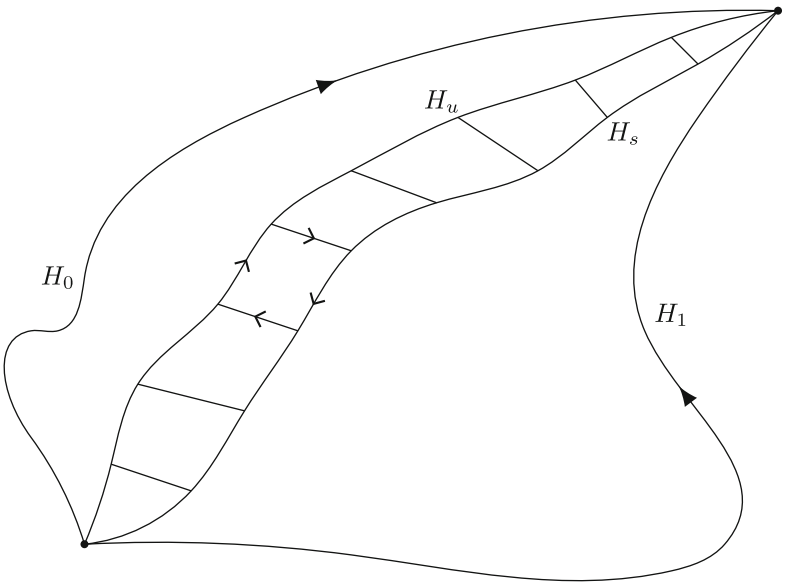



**Fig. 10.1** Locally conservative vector fields and homotopies

the other hand, the additivity of the integral implies that

$$0 = \sum_{i=1}^{k} \int_{\zeta_i} F \cdot dr = \left( \int_{H_u} F \cdot dr - \int_{H_s} F \cdot dr \right) - \left( \int_{H^b\restriction_{[u,s]}} F \cdot dr - \int_{H^a\restriction_{[u,s]}} F \cdot dr \right).$$

However, Since $H^a$ and $H^b$ are constant (the homotopy fixes end-points), $\int_{H^b\restriction_{[u,s]}} F \cdot dr = \int_{H^a\restriction_{[u,s]}} F \cdot dr = 0$. This gives the desired equality. □

*Remark 10.38* Theorem 10.33 gives a somewhat roundabout proof of the fact that the punctured plane is not simply connected, one which avoids lifting homotopies (Proposition 9.21): $F_{\texttt{wind}}$ is symmetric but not conservative on the punctured plane. «

### 10.4.1 Missing a Point

So far the symmetric vector fields we discussed were smooth. It will turn out that we need to relax that condition and prove:

**Proposition 10.39** *Suppose that $U \subseteq \mathbb{R}^2$ is a region, that $F : U \to \mathbb{R}^2$ is a vector field, that $\boldsymbol{p} \in U$ and that $F\restriction_{U\setminus\{\boldsymbol{p}\}}$ is symmetric. Then $F$ is locally conservative.*

Thus, $F$ is required to be defined and continuous at $\boldsymbol{p}$, but not even required to be differentiable at $\boldsymbol{p}$, and the derivative $DF$, or the four associated partial derivatives, are not required to have a limit at $\boldsymbol{p}$.

*Remark 10.40* Note that there is a single step of the proof of Proposition 10.35 that doesn't go through when we do not assume that $F$ is symmetric at $\boldsymbol{p}$: where we used Leibniz's differentiation under the integral sign to show that $D^x\left(\int_{p_y}^{q_y} F_y(x,t)\,dt\right)(q_x) = \int_{p_y}^{q_y} D^x F_y(q_x,t)\,dt$. Indeed, it is not clear what the second integral means, as the function $D^x F_y(q_x, y)$ may be unbounded on the interval $(p_y, q_y)$. More sophisticated machinery from measure theory, namely Lebesgue's dominated convergence theorem, can be used to overcome this problem, but we do not take that route. «

In the rest of the section we give a proof of Proposition 10.39. Fix $U$, $F$ and $\boldsymbol{p}$ as described. Let $C \subseteq U$ be an open rectangle; we show that $F\restriction_C$ is conservative. We assume that $\boldsymbol{p} \in C$ (otherwise we quote Proposition 10.36).

Our first step is the following.

**I.** Let $S \subset C$ be a rectangle one of whose corners is $\boldsymbol{p}$; and let $\gamma$ be the path which follows the boundary of $S$ (say starting and ending at $\boldsymbol{p}$ and travelling at constant speed.) Then $\int_\gamma F \cdot dr = 0$.

For small $h$ let $\gamma_h$ be the path following the boundary of the square whose opposite corners are $\boldsymbol{p}$ and $\boldsymbol{p} + (h, h)$. Now $F\restriction_{C\setminus\{\boldsymbol{p}\}}$ is symmetric and there is a simply connected (in fact, convex) region $V \subset C \setminus \{\boldsymbol{p}\}$ such that $S \setminus \{\boldsymbol{p}\} \subset V$. By Theorem 10.33, $F\restriction_V$ is conservative, and this implies that $\int_\gamma F \cdot dr = \int_{\gamma_h} F \cdot dr$ (see Fig. 10.2). And this holds for all $h > 0$.

However by Lemma 10.25,

$$\left|\int_{\gamma_h} F \cdot dr\right| \leqslant M \cdot 4h$$

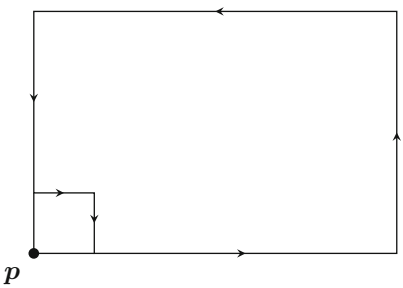



**Fig. 10.2** The integral along the marked path is 0

where $M$ bounds $|F|$ on the closed rectangle $S$; recall that we are still assuming that $F$ is continuous everywhere. Hence this integral is 0.

**II.** Let $S \subset C$ be any rectangle and let $\gamma$ be the loop which travels along the boundary of $S$. Then $\int_\gamma F \cdot dr = 0$.

If $\boldsymbol{p} \notin S$ (not on the boundary or the interior of the rectangle) then this follows from the fact that $F \restriction_V$ is conservative on a simply connected $V$ containing $S$. Otherwise this follows from (I) by breaking $S$ up into two or four rectangles, see Fig. 10.3.

Now for a point $\boldsymbol{q} \in C$ let $f_{\boldsymbol{q}} : C \to \mathbb{R}$ be defined as in the proof of Proposition 10.35, with the starting point being $\boldsymbol{q}$. Namely, for $\boldsymbol{q}$ and $\boldsymbol{r}$ in the rectangle $C$ let $\gamma_{\boldsymbol{q}\to\boldsymbol{r}}$ be the path which travels (in constant speed) first parallel to the $x$-axis from $\boldsymbol{q}$ to $(r_x, q_y)$, and then parallel to the $y$-axis to $\boldsymbol{r}$. Then let $f_{\boldsymbol{q}}(\boldsymbol{r}) = \int_{\gamma_{\boldsymbol{q}\to\boldsymbol{r}}} F \cdot dr$.

**III.** For all $\boldsymbol{q} \in C$, $f_{\boldsymbol{q}} - f_{\boldsymbol{p}}$ is constant on $C$.

Let $\boldsymbol{r} \in C$. We claim that $f_{\boldsymbol{p}}(\boldsymbol{r}) = f_{\boldsymbol{p}}(\boldsymbol{q}) + f_{\boldsymbol{q}}(\boldsymbol{r})$. For consider the "extra" points $\boldsymbol{rp} = (r_x, p_y)$, $\boldsymbol{rq} = (r_x, q_y)$ and $\boldsymbol{qp} = (q_x, p_y)$. Then

$$f_{\boldsymbol{p}}(\boldsymbol{r}) - (f_{\boldsymbol{p}}(\boldsymbol{q}) + f_{\boldsymbol{q}}(\boldsymbol{r})) = \int_\gamma F \cdot dr,$$

where $\gamma$ is the loop around the rectangle whose vertices are $\boldsymbol{qp}$, $\boldsymbol{rp}$, $\boldsymbol{rq}$ and $\boldsymbol{q}$, and so by (II) is 0. See Fig. 10.4

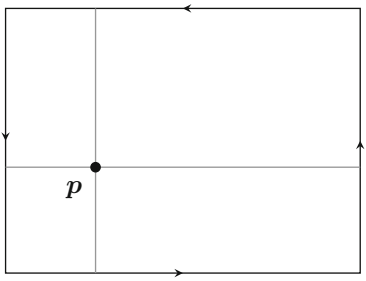



**Fig. 10.3** The integral along the boundary of the big rectangle is the sum of the integrals of along the boundaries of the four smaller ones

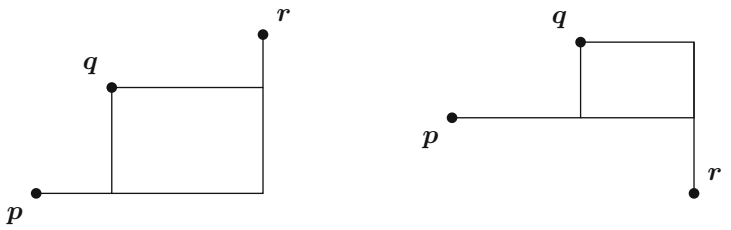



**Fig. 10.4** $f_p(\boldsymbol{r}) = f_p(\boldsymbol{q}) + f_{\boldsymbol{q}}(\boldsymbol{r})$

The penultimate step is:

**IV.** For all $\boldsymbol{q} \in C$, $Df_{\boldsymbol{q}}(\boldsymbol{q}) = F(\boldsymbol{q})$.

This follows from the fundamental theorem of calculus:

$$f_{\boldsymbol{q}}(x, q_y) = \int_{q_x}^{x} F_x(t, q_y)\, dt$$

and the derivative of the latter (with respect to $x$) at $q_x$ is $F_x(\boldsymbol{q})$. The same argument holds for $y$.

Finally, from (III) and (IV), we conclude that $Df_{\boldsymbol{p}} = F$ on $C$; (III) says that $Df_{\boldsymbol{p}} = Df_{\boldsymbol{q}}$ for all $\boldsymbol{q} \in C$. So $F\restriction_C$ is a gradient field, and so conservative. This concludes the proof of Proposition 10.39.

## 10.5 Further Exercises

10.41 Let $g\colon [a, b] \to \mathbb{R}$ be smooth. Show that the length of the graph of $g$ is

$$\int_a^b \sqrt{1 + \dot{g}(t)^2}dt.$$

10.42 Let $f, g\colon (a, b) \to \mathbb{R}$ be continuous, let $c \in (a, b)$ and suppose that $\dot{f} = g$ on $(a, b) \setminus \{c\}$. Show that $\dot{f}(c) = g(c)$.

10.43 Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^m$ be open, and suppose that $F\colon U \to V$ is smooth. Suppose that $\omega$ is a generalised form on $V$. We define the generalised form $F^*\omega$ on $U$, the *pull-back of* $\omega$ *by* $F$, by letting

$$(F^*\omega)_{\boldsymbol{p}}(\vec{v}) = \omega_{F(\boldsymbol{p})}((DF(\boldsymbol{p}))\vec{v}).$$

Let $\gamma$ be a smooth path in $U$. Show that

$$\int_\gamma F^*\omega = \int_{F\circ\gamma} \omega.$$

10.44 Suppose that $\gamma$ is a path in $\mathbb{R}^n$ and that $\omega$ is a form on the image of $\gamma$. Suppose that $\int_\gamma \omega$ exists. Show that for any re-parameterisation $\theta$ of $\gamma$ (see Exercise 9.108), $\int_\theta \omega$ exists and equals $\int_\gamma \omega$. (Hint: the translation from $\gamma$ to $\theta$ is uniformly continuous.)

10.45 Let $\gamma\colon [a, b] \to \mathbb{R}^n$ be a smooth path. Suppose that $\dot{\gamma}(s) \neq \vec{0}$ for all $t \in [a, b]$. Define $\varphi\colon [a, b] \to [0, \ell(\gamma)]$ by letting $\varphi(s) = \ell(\gamma\restriction_{[a,s]})$. (a) Show that $\varphi$ is smooth and $\dot{\varphi}(s) = |\dot{\gamma}(s)|$ for all $s \in [a, b]$. (b) Show that $\varphi$ is strictly increasing.

(c) Let $\theta = \gamma \circ \varphi^{-1}$. Show that $|\dot{\theta}| = 1$ on $[0, \ell(\gamma)]$. (d) Show that for all $t \in [0, \ell(\gamma)]$, $\ell(\theta|_{[0,t]}) = t$.[6]

**Fubini's Theorem**

10.46 Let $f\colon [a,b] \times [c,d] \to \mathbb{R}$ be continuous. (a) For tagged partitions $P$ of $[a,b]$ and $Q$ of $[c,d]$ define $S_{P\times Q}(f)$ to be the sum of products of the form $f(\boldsymbol{p})|A|$ where $A$ is a $P \times Q$-rectangle (the product of a $P$-interval and of a $Q$-interval), $|A|$ is the area of the rectangle and $\boldsymbol{p}$ is the point in $A$ tagged by $P$ and $Q$. Show that the limit of $S_{P\times Q}(f)$ exists as $D(P), D(Q) \to 0$. (b) Show that the limit equals both $\int_a^b \int_c^d f\,dy\,dx$ and $\int_c^d \int_a^b f\,dx\,dy$, and so these two iterated integrals are equal.[7]

10.47 Use Fubini's theorem to prove Proposition 9.60: if $g$ is twice smooth then $D^{xy}g = D^{yx}g$. (Hint: use the fundamental theorem of calculus to show that $\int_a^b \int_c^d D^y D^x g\,dy\,dx$ equals $(g(b,d)-g(b,c))-(g(a,d)-g(a,c))$; use the fact that $(g(b,d)-g(b,c))-(g(a,d)-g(a,c)) = (g(b,d)-g(a,d))-(g(b,c)-g(a,c))$; and the fact that if $h$ is continuous and $h \neq 0$ then $\int h\,dxdy \neq 0$ on some small rectangle.)

10.48 Use Fubini's theorem to prove Leibniz's integral rule (Proposition 10.23). (Hint: $g(t) - g(c) = \int_a^b \int_c^t D^y f dy dx$.)

**Vector Fields**

10.49 (a) Let $F(x,y) = (y,x)$. Show that $F$ is symmetric. (b) Find a potential function for $F$. (c) Do the same for $G(x,y) = (x,y)$.

10.50 (Turetsky) Let $U \subseteq \mathbb{R}^2$ be simply connected. Let $f\colon U \to \mathbb{R}$ be smooth, and let $G\colon U \to \mathbb{R}^2$ be a smooth vector field. Suppose that for all $\boldsymbol{p} \in U$, $Df(\boldsymbol{p})$ and $G(\boldsymbol{p})$ are parallel; and that $G$ is conservative. Show that the vector field $f \cdot G$ (mapping each $\boldsymbol{p} \in U$ to $f(\boldsymbol{p}) \cdot G(\boldsymbol{p})$) is conservative.

[6] The path $\theta$ is called the *path-length re-parameterisation* of $\gamma$.

[7] This is *Fubini's theorem* for continuous functions.

10.51 Suppose that $\gamma$ is a smooth and injective path in $\mathbb{R}^n$, and that $\dot{\gamma} \neq \vec{0}$ at every point. Show that

$$\int_\gamma F \cdot dr = \int_\gamma \left( F \cdot \frac{\dot{\gamma}}{|\dot{\gamma}|} \right) ds.$$

(Here $F \cdot \frac{\dot{\gamma}}{|\dot{\gamma}|}$ is a function from the image of $\gamma$ to $\mathbb{R}$, and is well-defined since $\gamma$ is injective.)

10.52 Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable, and suppose that $f(\mathbf{0}) = 0$. Show that there are functions $g_1, g_2, \ldots, g_n : \mathbb{R}^n \to \mathbb{R}$ such that for all $\boldsymbol{x} \in \mathbb{R}^n$,

$$f(\boldsymbol{x}) = x_1 g_1(\boldsymbol{x}) + x_2 g_2(\boldsymbol{x}) + \cdots + x_n g_n(\boldsymbol{x}).$$

(This uses Exercise 9.102. For a hint see [Spi65, p.34].)

**Plane-Filling and Rectifiable Paths**

10.53 The purpose of this exercise is to construct a "plane-filling curve". For each $n$ we define a division of the unit square $[0, 1]^2$ into $4^n$ many squares $A_1^n, A_2^n, \ldots, A_{4^n}^n$. Each square $A_k^n$ is split into 4 equal sub-squares $A_{4k-3}^{n+1}$, $A_{4k-2}^{n+1}$, $A_{4k-1}^{n+1}$ and $A_{4k}^n$.

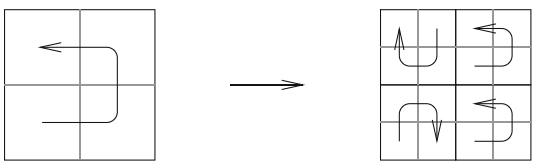
Here are the first few steps, omitting the superscripts $n = 1, 2, 3$:

| | |
|---|---|
| 4 | 3 |
| 1 | 2 |

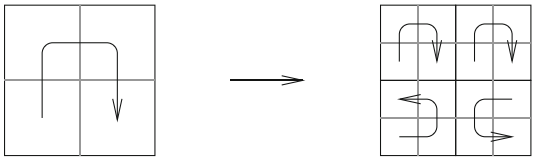
| | | | |
|---|---|---|---|
| 16 | 13 | 12 | 11 |
| 15 | 14 | 9 | 10 |
| 2 | 3 | 8 | 7 |
| 1 | 4 | 5 | 6 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 64 | 63 | 50 | 49 | 48 | 45 | 44 | 43 |
| 61 | 62 | 51 | 52 | 47 | 46 | 41 | 42 |
| 60 | 57 | 56 | 53 | 34 | 35 | 40 | 39 |
| 59 | 58 | 55 | 54 | 33 | 36 | 37 | 38 |
| 6 | 7 | 10 | 11 | 32 | 29 | 28 | 27 |
| 5 | 8 | 9 | 12 | 31 | 30 | 25 | 26 |
| 4 | 3 | 14 | 13 | 18 | 19 | 24 | 23 |
| 1 | 2 | 15 | 16 | 17 | 20 | 21 | 22 |

we follow the pattern



and its reflections and rotations, for example:



(a) Show that for each $k < 4^n$, $A^n_k$ and $A^n_{k+1}$ have a common edge.

For $n \geqslant 1$ and $k = 1, 2, \ldots, 4^n$ let $I^n_k = [(k-1)4^{-n}, k4^{-n}]$.

(b) Show that there is a (unique) function $f : [0, 1] \to [0, 1]^2$ satisfying, for all $n \geqslant 1$ and $k \leqslant 4^n$, $f[I^n_k] \subseteq A^n_k$.
(c) Show that $f$ is continuous.
(d) Show that $f$ is onto $[0, 1]^2$.
(e) Show that $f$ is not injective.
(f) Find distinct $s, t \in [0, 1]$ such that $f(s) = f(t)$.

A set $A \subseteq \mathbb{R}^n$ is *Lebesgue null* if for all $\varepsilon > 0$ there is a sequence $U_1, U_2, \ldots$ of open balls such that $A \subseteq \bigcup_n U_n$ and the sum $\sum_n v(U_n)$ of the volumes of the balls $U_n$ is smaller than $\varepsilon$. Observe that a union of countably many Lebesgue null sets is Lebesgue null.

10.54 Let $n \geqslant 1$; show that the $n$-dimensional unit hypercube $[0, 1]^n$ is not Lebesgue null.

10.55 let $n > 1$ and let $\gamma : I \to \mathbb{R}^n$ be a path, and suppose that $\ell(\gamma) < \infty$. Show that the image of $\gamma$ is Lebesgue null. (Hence, any plane-filling curve has to have infinite length. One way to do this is, given $\varepsilon > 0$, to construct a piecewise linear path $\eta$ such that every point in $\gamma[I]$ is $\varepsilon$-close to $\eta[I]$.)

10.56 (a) Let $M$ be a differentiable manifold, and let $\gamma$ be a piecewise smooth path in $M$. Show that the image of $\gamma$ is not all of $M$.[8] (b) Give an alternative proof that the sphere is simply connected (Proposition 9.16). (Use Exercise 9.106 and see Remark 9.17.)

10.57 Let $\gamma : I \to \mathbb{R}^n$ be a path such that $\ell(\gamma) < \infty$ (but not necessarily smooth or piecewise smooth).

[8] While there is no obviously good notion of the length of a curve in $M$—for that we need to define an appropriate form on $M$—we can use charts to show that $\gamma$ cannot fill any open set in $M$.

(a) Show that for every continuous function $f : \gamma[I] \to \mathbb{R}$, $\int_\gamma f\, ds$ exists.
(b) Show that for every vector field $F : \gamma[I] \to \mathbb{R}^n$, $\int_\gamma F \cdot dr$ exists.[9]

10.58 Define $\gamma : [0, 1] \to \mathbb{R}^2$ by letting $\gamma(0) = \mathbf{0}$ and for $t > 0$, $\gamma(t) = (t, t \sin(1/t))$. Show that $\ell(\gamma) = \infty$.

[9] In fact, for every form $\omega$ on $\gamma[I]$, $\int_\gamma \omega$ exists. This is a bit harder to prove. The hypothesis $\ell(\gamma) < \infty$ is quite strong; it implies that $\gamma$ is differentiable outside a Lebesgue null set, and that $\dot{\gamma}$, while perhaps not continuous, is integrable. such a curve is called *rectifiable*.


# Complex Differentiation

# 11

If a complex function is differentiable on an open set then it is analytic. This remarkable fact makes complex analysis very different from real analysis. In this chapter we finally develop the basics of complex analysis and analytic functions. In the next chapter, this will allow us to define Riemann surfaces: those surfaces on which we can perform complex differentiation.

Our development of the complex derivative and integral, up to Cauchy's Integral Formula, emphasises the link between complex and real differentiation: a function $f\colon \mathbb{C} \to \mathbb{C}$ is complex differentiable if and only if the corresponding function $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ is real differentiable, and the Cauchy-Riemann Equations hold. We later turn to power series and analytic functions, and finally connect the two strands together in Theorem 11.67.

Our work in this chapter will allow us to settle two debts: first, to carry out the development from the introductory chapter of the function $t \mapsto e^{it}$; and second, to give a proof of the fundamental theorem of algebra.

Throughout, we continue to assume that all paths are piecewise smooth, so that path integrals exist.

## 11.1 Complex Derivatives and Integrals

We have identified the complex numbers with the plane $\mathbb{R}^2$, and so we identify functions $f\colon \mathbb{C} \to \mathbb{C}$ with maps $f\colon \mathbb{R}^2 \to \mathbb{R}^2$. For such $f$ we write $f_x$ and $f_y$ for the real and imaginary part of $f$, so $f = f_x + i f_y = \begin{pmatrix} f_x \\ f_y \end{pmatrix}$.



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_11

The key to complex differentiation is the following observation:

**Lemma 11.1** *Let $c \in \mathbb{C}$. Considered as a map $\mathbb{R}^2 \to \mathbb{R}^2$, the function $z \mapsto cz$ is linear, defined by the matrix*

$$M_c = \begin{pmatrix} c_x & -c_y \\ c_y & c_x \end{pmatrix}.$$

*Exercise 11.2* Verify that multiplying by $i$ is the rotation of $\mathbb{R}^2$ by $90°$ and in general, multiplying by $e^{i\theta}$ is the rotation by $\theta$ radians. «

*Exercise 11.3* Show that for $c, d \in \mathbb{C}$ we have $M_{c+d} = M_c + M_d$ and $M_{cd} = M_c M_d$.[1] «

*Exercise 11.4* Let $c, d \in \mathbb{C}$. (a) Show that $M_{\bar{c}} = {M_c}^{\mathsf{t}}$ (where $\bar{c}$ is the complex conjugate $c_x - ic_y$ of $c$). (b) Show that $\det M_c = |c|^2$. (c) Show that $|cd| = |c| \cdot |d|$. (d) Show that if $c \neq 0$ then $M_{c^{-1}} = (M_c)^{-1} = \frac{1}{|c|^2} {M_c}^{\mathsf{t}}$. (e) Show that $\|M_c\| = |c|$. (See Definition 9.39.) «

Lemma 11.1 implies:

**Proposition 11.5** *Let $U \subseteq \mathbb{C}$ be open, let $w \in U$, and let $g : U \to \mathbb{C}$ be a function. Let $c \in \mathbb{C}$. The following are equivalent:*

(1) *The corresponding function $g : U \to \mathbb{R}^2$ is differentiable at $w$, and*

$$Dg(w) = M_c.$$

(2) *For all $\varepsilon > 0$ there is some $\delta > 0$ such that for all $z \in \mathbb{C}$, if $|z - w| < \delta$ then $|(g(z) - g(w)) - (z - w)c| < |z - w|\varepsilon$. In other words,*

$$c = \lim_{z \to w} \frac{g(z) - g(w)}{z - w},$$

*where the ratio is taken as the ratio of two complex numbers.*

When these conditions hold then we say that $g$ is *complex differentiable* at $w$, and write $g'(w) = c$. The $2 \times 2$ matrices $A$ which are $M_c$ for some $c \in \mathbb{C}$ are precisely those satisfying $a_{1,1} = a_{2,2}$ and $a_{1,2} = -a_{2,1}$; so Proposition 11.5 gives the:

[1] Thus, $c \mapsto M_c$ is an embedding of $\mathbb{C}$ into the matrix ring $M_2(\mathbb{R})$.

**Cauchy-Riemann Equations** *A function $g : U \to \mathbb{C}$ is complex differentiable at $w$ if and only if the corresponding function $g : U \to \mathbb{R}^2$ is real differentiable at $w$, and at $w$,*

$$D^x g_x = D^y g_y \qquad \& \qquad D^y g_x = -D^x g_y,$$

*in which case $g' = D^x g_x - iD^y g_x = D^y g_y + iD^x g_y$.*

*Example 11.6* The function $z \mapsto \bar{z}$ is not complex differentiable at any point, as the full derivative of the corresponding vector field is $\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$ at every point. «

**Proposition 11.7** *Let $f$ and $g$ be defined on a neighbourhood of $w$ and differentiable at $w$. Then:*

(a) *$f + g$ is differentiable at $w$ and $(f + g)'(w) = f'(w) + g'(w)$.*
(b) *$fg$ is differentiable at $w$ and $(fg)'(w) = (fg' + f'g)(w)$.*
(c) *If $f$ is constant on a neighbourhood of $w$ then $f'(w) = 0$.*
(d) *The chain rule holds: if $h$ is differentiable at $f(w)$ then $h \circ f$ is differentiable at $w$ and $(h \circ f)'(w) = h'(f(w)) \cdot f'(w)$.*
(e) *If $f(z) = z$ on a neighbourhood of $w$ then $f'(w) = 1$.*
(f) *If $f(z) = 1/z$ on a neighbourhood of $w$ then $f'(w) = -1/w^2$.*
(g) *$f$ is continuous at $w$.*

***Proof*** The proofs for real functions carry over to the complex case. However, we can also use facts about real differentiability to deduce these properties using the characterisation above of complex differentiation. For example, (a) follows from Exercise 9.44 together with the fact that $M_c + M_d = M_{c+d}$ (Exercise 11.3). (c) follows from Exercise 9.35; (d) follows from the chain rule (Proposition 9.42), using the fact that $M_c M_d = M_{cd}$; (e) follows from Exercise 9.35 as well, using the fact that $M_1$ is the identity matrix; (g) follows from Exercise 9.34.

(f) is a calculation; if $f(z) = 1/z = (x/(x^2 + y^2), -y/(x^2 + y^2))$ then $Df = \frac{1}{(x^2+y^2)^2}\left(\begin{smallmatrix} y^2-x^2 & -2xy \\ 2xy & y^2-x^2 \end{smallmatrix}\right)$ which is $M_{-1/z^2}$.

(b) is more elaborate but follows from Exercise 9.53. As a vector field, $fg = \left(\begin{smallmatrix} f_x g_x - f_y g_y \\ f_x g_y + f_y g_x \end{smallmatrix}\right)$. Applying the product rule and recalling that at $w$, $Df = M_{f'}$, gives, at $w$,

$$D((fg)_x) = (f_x Dg_x + g_x Df_x) - (f_y Dg_y + g_y Df_y) = (f_x, -f_y)M_{g'} + (g_x, -g_y)M_{f'};$$

together with a similar calculation for $D((fg)_y)$ we get $D(fg) = M_f M_{g'} + M_g M_{f'} = M_{fg'+f'g}$ at $w$ as required. □

**Proposition 11.8** *Suppose that $f: U \to \mathbb{C}$ is differentiable, that $U$ is connected, and that $f' = 0$ on $U$. Then $f$ is constant.*

***Proof*** Follows from Corollary 9.77, since $M_0$ is the zero matrix. □

Proposition 9.48 and Exercise 11.4(e) imply:

**Proposition 11.9** *Suppose that $f: U \to \mathbb{C}$ is complex differentiable and that $U$ is convex. Suppose that $|f'| \leqslant M$ on $U$. Then for all $a \neq b \in U$, $|f(b) - f(a)| \leqslant M|b - a|$.*

If $U \subseteq \mathbb{C}$ is open, then a function $f: U \to \mathbb{C}$ is *continuously (complex) differentiable* on $U$ if it is complex differentiable at every point of $U$ and the derivative function $f': U \to \mathbb{C}$ is continuous.[2]

*Exercise 11.10* Let $U \subseteq \mathbb{C}$ be open and let $f: U \to \mathbb{C}$ be continuously differentiable. Define $G: U^2 \to \mathbb{C}$ by letting

$$G(z, w) = \begin{cases} \frac{f(w)-f(z)}{w-z}, & \text{if } z \neq w; \text{ and} \\ f'(z), & \text{if } z = w. \end{cases}$$

Show that $G$ is continuous. (Compare with Exercise 9.50.) «

Finally, the real inverse function theorem (for $n = 2$), together with $(M_c)^{-1} = M_{c^{-1}}$ and $\det(M_c) = |c|^2$, which is nonzero if $c \neq 0$, gives a complex inverse function theorem:

**Theorem 11.11** *Let $f: U \to \mathbb{C}$ be continuously differentiable, let $a \in U$, and suppose that $f'(a) \neq 0$. Then there are neighbourhoods $V \subseteq U$ of $a$ and $W$ of $f(a)$ such that $f\restriction_V$ is a homeomorphism between $V$ and $W$, and the inverse $g$ of $f\restriction_V$ is differentiable at $f(a)$, and $g'(f(a)) = 1/f'(a)$.*

*Example 11.12* Let $a \in \mathbb{C}$ be nonzero. There are two complex square roots of $a$; let $b$ be one of them. Of course $b \neq 0$. The function $z \mapsto z^2$ is continuously differentiable and its derivative is nonzero at $b$. So there is a neighbourhood $U$ of $b$ and $V$ of $a$ such that $z \mapsto z^2$ is a homeomorphism from $U$ to $V$; and on $V$ there is a continuously differentiable function which maps $a$ to $b$ and each $w \in V$ to one of the square roots of $w$. Such a choice of a square root, or an $n$th root for any $n \geqslant 2$, cannot be done on a neighbourhood of 0; see Example 12.36.[3] «

[2] We do not use the term *smooth* here, to distinguish between the complex and real derivative.

[3] The construction of a continuously differentiable square root can be done directly by a continuous choice of argument; see Exercise 11.87. We will "resolve" the multi-valued nature of the square root using Riemann surfaces in Chap. 12.

### 11.1.1 Complex Integrals

We introduce two kinds of complex integrals. First, given a continuous function $f\colon [a,b] \to \mathbb{C}$ we can again write $f = f_x + i f_y$ and then we just define

$$\int_a^b f\,dt = \int_a^b f_x\,dt + i\int_a^b f_y\,dt.$$

Next, we consider path integrals of complex functions. A complex (linear) form on $U \subseteq \mathbb{C}$ is a continuous map $\omega\colon U \times \mathbb{C} \to \mathbb{C}$ such that for all $a \in U$, the function $\omega_a(v) = \omega(a,v)$ is linear as a map from $\mathbb{C}$ to $\mathbb{C}$; which of course means that $\omega_a(v) = f(a)\cdot v$ for all $v \in \mathbb{C}$, where $f$ is continuous. We let $dz$ be the form defined by $dz_a(v) = v$; so every complex form is $f\,dz$ for some continuous $f$.

Now for a path $\gamma$ in $U$ we can define $\int_\gamma \omega$ exactly as is done at page 257: Fix some $f\colon U \to \mathbb{C}$ and a path $\gamma\colon K \to U$. For a tagged partition $P$ of $K$ we let

$$S_P(f\,dz,\gamma) = \sum \big\{ f(\gamma(r_I))\cdot \Delta_I\gamma \;:\; I \text{ is a } P\text{-interval}\big\};$$

recall that $r_I$ is the $P$-tag for the interval $I$. Here $\Delta_I\gamma$ (which is a vector in $\mathbb{R}^2$) is treated as a complex number, and the product $f(\gamma(r_I))\cdot\Delta_I\gamma$ is the product of two complex numbers. If the limit exists we let

$$\int_\gamma f\,dz = \lim_{D(P)\to 0} S_P(f\,dz,\gamma).$$

Let $I = [c,d] \subseteq K$ be a subinterval. Let $\Delta_x = (\Delta_I\gamma)_x = \gamma_x(d) - \gamma_x(c)$ where $\gamma = \gamma_x + i\gamma_y$; similarly define $\Delta_y$. Let $r \in I$. Again identifying with functions to $\mathbb{R}^2$,

$$f(\gamma(r))\cdot\Delta_I\gamma = M_{f(\gamma(r))}\cdot\begin{pmatrix}\Delta_x\\ \Delta_y\end{pmatrix} = \begin{pmatrix} f_x(\gamma(r))\Delta_x - f_y(\gamma(r))\Delta_y \\ f_y(\gamma(r))\Delta_x + f_x(\gamma(r))\Delta_y\end{pmatrix},$$

which shows that

$$S_P(f\,dz,\gamma) = S_P(f_x\,dx - f_y\,dy,\gamma) + i\cdot S_P(f_y\,dx + f_x\,dy,\gamma).$$

If $\gamma$ is piecewise smooth (as we are assuming now), real integrals along $\gamma$ exist. Recalling our notation $f_x\,dx - f_y\,dy = (f_x,-f_y)\cdot dr$ and similarly for the imaginary part, we conclude:

**Proposition 11.13** *For any path $\gamma$ in $U$ and continuous $f\colon U \to \mathbb{C}$, the integral $\int_\gamma f\,dz$ exists, and*

$$\int_\gamma f\,dz = \int_\gamma (f_x,-f_y)\cdot dr + i\int_\gamma (f_y,f_x)\cdot dr.$$

*Exercise 11.14* Let $f\colon U \to \mathbb{C}$ be continuous and let $\gamma\colon K \to U$ be path. Show that

$$\left|\int_\gamma f\,dz\right| \leqslant \int_\gamma |f|\,ds$$

(Use the identity $|zw| = |z| \cdot |w|$ and the triangle inequality.) Conclude that

$$\left|\int_\gamma f\,dz\right| \leqslant \ell(\gamma) \cdot \max_{w\in\gamma[K]}\{|f(w)|\}$$ «

*Exercise 11.15* Show that the complex path integral is linear: $\int_\gamma (cf + g)dz = c\int_\gamma f\,dz + \int_\gamma g\,dz$ (where $c \in \mathbb{C}$). «

When $\gamma\colon K \to \mathbb{C}$ is smooth we identify $\dot\gamma(t)$ with the corresponding complex number, which is $\dot\gamma_x + i\dot\gamma_y$. Recall that when $\gamma\colon [a,b] \to \mathbb{C}$ is smooth and $F$ is a vector field then $\int_\gamma F \cdot dr = \int_a^b F(\gamma(t)) \cdot \dot\gamma(t)\,dt$ where here the multiplication is the dot product of vectors. Again letting $f\colon U \to \mathbb{C}$ be continuous, when $\gamma$ is smooth we have

$$\int_\gamma f\,dz = \int_\gamma (f_x, -f_y)\cdot dr + i\int_\gamma (f_y, f_x)\cdot dr =$$

$$\int_a^b \big((f_x(\gamma(t))\cdot\dot\gamma_x(t)) - f_y(\gamma(t))\cdot\dot\gamma_y(t)\big)\,dt +$$

$$i\int_a^b \big((f_y(\gamma(t))\cdot\dot\gamma_x(t)) + f_x(\gamma(t))\cdot\dot\gamma_y(t)\big)\,dt = \int_a^b f(\gamma(t))\cdot\dot\gamma(t)\,dt$$

where in the last integral the multiplication is that of complex numbers, and the integral of the function $f(\gamma(t))\cdot\dot\gamma(t)\colon [a,b] \to \mathbb{C}$ is the first kind of complex integral (integral of a function of a real variable) mentioned above.

## 11.2 Cauchy's Integral Formula

Fix a region (open and connected) $U \subseteq \mathbb{C}$. A *primitive* of a continuous function $f\colon U \to \mathbb{C}$ is a continuously differentiable $g\colon U \to \mathbb{C}$ such that $g' = f$.

**Lemma 11.16** *Let $f : U \to \mathbb{C}$ be continuous. The following are equivalent.*

(1) *Both vector fields $(f_x, -f_y)$ and $(f_y, f_x)$ on $U$ are conservative.*
(2) *For every loop $\gamma$ in $U$, $\int_\gamma f\,dz = 0$. (Recall that we assume that all paths are piecewise smooth, so the integral always exists.)*
(3) *$f$ has a primitive on $U$.*

***Proof*** The equivalence of (1) and (2) follows from Proposition 11.13. The equivalence of (1) and (3) follows from the Cauchy-Riemann Equations. A function $g : U \to \mathbb{C}$ is a primitive of $f$ if and only if $Dg = M_f$ on $U$, in other words, if and only if $Dg_x = (f_x, -f_y)$ and $Dg_y = (f_y, f_x)$—if and only if $g_x$ and $g_y$ show that both vector fields $(f_x, -f_y)$ and $(f_y, f_x)$ are gradient fields, equivalently conservative (Proposition 10.28). □

If $g$ is continuously differentiable, then Proposition 11.13 implies that

$$\int_\gamma g'\,dz = \int_\gamma Dg_x \cdot dr + i \int_\gamma Dg_y \cdot dr.$$

Then by Proposition 10.29 we have following analogue of the second part of the fundamental theorem of calculus:

**Lemma 11.17** *Suppose that $g : U \to \mathbb{C}$ is continuously differentiable. Then for any path $\gamma : [a, b] \to U$,*

$$\int_\gamma g'\,dz = g(\gamma(b)) - g(\gamma(a)).$$

The following lemma says that this is a characterisation of primitives:

**Lemma 11.18** *Let $f, h : U \to \mathbb{C}$ be continuous and suppose that for any path $\gamma : [a, b] \to U$,*

$$\int_\gamma f\,dz = h(\gamma(b)) - h(\gamma(a)).$$

*Then $h$ is a primitive of $f$ on $U$.*

***Proof*** The assumption implies that $\int_\gamma f\,dz = 0$ for loops in $U$, hence there is some $g : U \to \mathbb{C}$ such that $g' = f$. Fix some point $w \in U$. Given $z \in U$, since we assume that $U$ is connected, let $\gamma : [a, b] \to U$ be a path from $w$ to $z$ (which we may take to be smooth by Proposition 9.75). Then $h(z) - h(w) = \int_\gamma f\,dz = g(z) - g(w)$. This implies that $g - h$ is constant on $U$, whence $h' = g' = f$. □

### Continuously Differentiable Functions and Primitives

Let $f : U \to \mathbb{C}$ be continuously differentiable. The Cauchy-Riemann Equations imply that the vector fields $(f_x, -f_y)$ and $(f_y, f_x)$ are symmetric on $U$. Theorem 10.33 (together with Lemma 11.16) then implies:

**Proposition 11.19** *Suppose that $f : U \to \mathbb{C}$ is continuously differentiable and that $U$ is simply connected. Then $f$ has a primitive on $U$.*

In fact, we can omit one point, so we get the following extension of Proposition 11.19:

**Lemma 11.20** *Suppose that $U$ is simply connected, that $p \in U$, that $f : U \to \mathbb{C}$ is continuous and that $f\restriction_{U\setminus\{p\}}$ is continuously differentiable. Then $f$ has a primitive on $U$.*

***Proof*** Again we need to show that the vector fields $(f_x, -f_y)$ and $(f_y, f_x)$ are conservative (on $U$). As before, the Cauchy-Riemann Equations imply that both vector fields are symmetric on $U \setminus \{p\}$. The result then follows from Propositions 10.39 and 10.36. □

### 11.2.1 Winding Numbers in the Complex Plane

Recall that the winding number of a piecewise smooth loop around the origin can be characterised using integration of the vector field $F_{\texttt{wind}}$ (Proposition 10.32). We can also use complex integration:

**Proposition 11.21** *Let $\gamma$ be a loop in $\mathbb{C} \setminus \{0\}$. If the winding number of $\gamma$ is $m$ then*

$$\int_\gamma \frac{dz}{z} = 2\pi i m.$$

***Proof*** Let $G$ be the vector field from Example 10.30. Let $f(z) = z^{-1} = \bar{z}/|z|^2$; so $f_x = x/(x^2 + y^2)$ and $f_y = -y/(x^2 + y^2)$. That is, $(f_x, -f_y) = G$ and $(f_y, f_x) = F_{\texttt{wind}}$. Then by Proposition 11.13,

$$\int_\gamma \frac{dz}{z} = \int_\gamma G \cdot dr + i \int_\gamma F_{\texttt{wind}} \cdot dr;$$

The proposition follows from Proposition 10.32 and the fact that $G$ is conservative. □

The winding number of a loop in $\mathbb{C}\setminus\{0\}$ is its winding number around the origin; by shifting, for every point $p \in \mathbb{C}$ and loop in $\mathbb{C}\setminus\{p\}$ we can define the winding number of $\gamma$ around $p$, which we denote by $\mathrm{wnd}_\gamma(p)$. Proposition 11.21 implies that

$$\mathrm{wnd}_\gamma(p) = \frac{1}{2\pi i}\int_\gamma \frac{dz}{z-p}.$$

Using this notation, Proposition 9.29 says that two loops $\gamma$ and $\eta$ in $\mathbb{C}\setminus\{p\}$ (with the same domain and end-point) are homotopic in $\mathbb{C}\setminus\{p\}$ if and only if $\mathrm{wnd}_\gamma(p) = \mathrm{wnd}_\eta(p)$. And Propositions 9.30 and 9.32 together show:

**Proposition 11.22** *Suppose that $W \subseteq \mathbb{C}$ is simply connected and that $\gamma$ is a loop in $W$. Then $\mathrm{wnd}_\gamma(p) = 0$ for all $p \notin W$.*

**Corollary 11.23** *For any loop $\gamma$, $\mathrm{wnd}_\gamma = 0$ outside a bounded subset of $\mathbb{C}$.*

***Proof*** Since $\gamma[I]$ is compact, for sufficiently large $R$ we have $\gamma[I] \subset B(0, R)$, and $B(0, R)$ is simply connected. □

Another corollary is:

**Proposition 11.24** *For any loop $\gamma\colon I \to \mathbb{C}$, the function $p \mapsto \mathrm{wnd}_\gamma(p)$ is continuous on $\mathbb{C}\setminus\gamma[I]$.*

***Proof*** By shifting, we assume that $0 \notin \gamma[I]$ and show that $p \mapsto \mathrm{wnd}_\gamma(p)$ is continuous at 0. Let $r = d(0, \gamma[I]) = \min\{|z| : z \in \gamma[I]\}$, which is positive (as $\gamma[I]$ is closed). We show that $p \mapsto \mathrm{wnd}_\gamma(p)$ is constant on $B(0, r)$. Suppose that $|p| < r$. Let $\eta(t) = \gamma(t) - p$. By definition, $\mathrm{wnd}_\gamma(p) = \mathrm{wnd}_\eta(0)$. So we show that $\mathrm{wnd}_\eta(0) = \mathrm{wnd}_\gamma(0)$.

To see this, we observe that for all $z \in \gamma[I]$, $|p/z| < 1$ and so $(z-p)/z = 1 - (p/z)$ is in $U = B(1, 1)$. Now $U \subset \mathbb{C}^*$ is simply connected, so there is a continuous choice of argument $\alpha$ on $U$ (Proposition 9.30). Let $\theta$ be a continuous choice of argument for $\gamma$. Then $t \mapsto \theta(t) + \alpha(\eta(t)/\gamma(t))$ is a continuous choice of argument for $\eta$, and this establishes $\mathrm{wnd}_\eta(0) = \mathrm{wnd}_\gamma(0)$. (For an alternative proof see Exercise 11.78.) □

Since $\mathrm{wnd}_\gamma(p)$ is always an integer, this means that $\mathrm{wnd}_\gamma(p)$ is constant on connected open sets.

**Definition 11.25** Let $p \in \mathbb{C}$. A loop $\gamma$ in $\mathbb{C}\setminus\{p\}$ is a *contour* around $p$ if $\mathrm{wnd}_\gamma(p) = 1$.

*Example 11.26* A *parameterised circle* with centre $p$ is a loop $\gamma\colon [0, 2\pi] \to \mathbb{C}$ given by $\gamma(t) = p + re^{it}$ (for some fixed $r > 0$). A parameterised circle with centre $p$ is a contour around $p$. «

More is true:

**Proposition 11.27** *Let $\gamma$ be a parameterised circle. Then $\gamma$ is a contour around every point in its interior, but* $\mathrm{wnd}_\gamma(p) = 0$ *if $p$ lies outside the circle.*

***Proof*** Let $q$ be the circle's centre and $r$ its radius. The interior of the circle $B(q, r)$ and its exterior $\mathbb{C} \setminus \overline{B}(q, r) = \{p : d(p, q) > r\}$ are both connected, and so $\mathrm{wnd}_\gamma$ is constant on both. We know that $\mathrm{wnd}_\gamma(q) = 1$, so $\mathrm{wnd}_\gamma = 1$ on $B(q, r)$; on the other hand, $\{p : d(p, q) > r\}$ is unbounded, and so $\mathrm{wnd}_\gamma = 0$ on that set. □

## The Integral Formula

We are now ready to prove Cauchy's integral formula, which says that the value of a continuously differentiable function at some point is the average of its values on a contour around that point.

**Cauchy's Integral Formula** *Let $U \subseteq \mathbb{C}$ be simply connected, let $p \in U$, let $\gamma$ be a contour in $U$ around $p$, and let $f\colon U \to \mathbb{C}$ be continuously differentiable. Then*

$$f(p) = \frac{1}{2\pi i} \int_\gamma \frac{f(z)}{z - p}\, dz.$$

***Proof*** Define $g\colon U \to \mathbb{C}$ by letting

$$g(z) = \begin{cases} \frac{f(z)-f(p)}{z-p}, & \text{if } z \neq p, \text{ and} \\ f'(p) & \text{if } z = p. \end{cases}$$

Proposition 11.5 implies that $g$ is continuous at $p$; and $g$ is continuously differentiable on $U \setminus \{p\}$. Lemma 11.20 says that $g$ has a primitive on $U$, so by Lemma 11.16, $\int_\gamma g\, dz = 0$. But since $p$ is not in the range of $\gamma$,

$$0 = \int_\gamma g\, dz = \int_\gamma \frac{f(z)}{z - p}\, dz - f(p) \int_\gamma \frac{dz}{z - p}.$$

The integral formula now follows since $\gamma$ is a contour around $p$. □

## 11.3 Uniform Convergence and Power Series

Analytic functions are those which are locally the sum of power series. The theory that we present in this section is pretty standard. It is often part of a real analysis course; all we need to observe is that the same arguments work for complex numbers as well.

### 11.3.1 Absolute Convergence

The topology of $\mathbb{C}$ allows us to not only take the limits of sequences but sometimes add up infinitely many numbers. Let $\langle z_n \rangle$ be an infinite sequence of complex numbers. We say that $\sum_n z_n = w$ if the sequence of partial sums $z_1, z_1 + z_2, z_1 + z_2 + z_3, \ldots$ converges to $w$. If $\sum z_n = w$ for some $w$ then we say that the sum $\sum z_n$ converges.

Analogously to sequences, completeness is utilised via the notion of Cauchy series. A series $\sum z_n$ is a *Cauchy series* if the sequence of partial sums $\left\langle \sum_{j=1}^{n} z_j \right\rangle$ is a Cauchy sequence. This unwraps to the condition: for all $\varepsilon > 0$ there is some $N$ such that for all $m \geqslant n \geqslant N$, $|z_n + z_{n+1} + \cdots + z_m| < \varepsilon$.[4] Completeness of the complex numbers (Exercise 8.84 for $n = 2$) implies that a series converges if and only if it is a Cauchy series.

*Exercise 11.28* (a) Show that if $|z| < 1$ then the geometric series $\sum z^n$ converges. (Hint: consider the division of polynomials $(1 - x^n)/(1 - x)$. Use Exercise 8.122.) (b) Show that $\sum 1/n^2$ converges. (Hint: $1/n^2 \leqslant 1/(n-1) - 1/n$.) «

**Proposition 11.29** *If $\sum z_n$ converges then $\lim_{n\to\infty} z_n = 0$.*

***Proof*** If for some $\varepsilon$, for infinitely many $n$ we have $|z_n| \geqslant \varepsilon$, then $\sum z_n$ is not a Cauchy series, as witnessed by $m = n$; a convergent series is Cauchy (Proposition 8.81). □

*Exercise 11.30* Show that the harmonic series $\sum_n 1/n$ does not converge, giving a counterexample to the converse of Proposition 11.29. (One way to do this is to consider the sum of $1/n$ for $n$ ranging from $2^m + 1$ to $2^{m+1}$.) «

**Definition 11.31** A series $\sum z_n$ *converges absolutely* if the series $\sum |z_n|$ converges.

**Proposition 11.32** *If $\sum z_n$ converges absolutely then it converges, and $\left|\sum z_n\right| \leqslant \sum |z_n|$.*

[4] Informally, we sometimes write $\lim_{n,m\to\infty} \sum_{j=n}^{m} z_j = 0$.

***Proof*** Let $m \geqslant n$. The triangle inequality implies that $|z_n + z_{n+1} + \cdots + z_m| \leqslant |z_n| + |z_{n+1}| + \cdots + |z_m|$. This shows that if $\sum |z_n|$ is a Cauchy series, then so is $\sum z_n$, and so it converges. The inequality follows from applying the triangle inequality to the partial sums, once we observe that $\left|\sum z_n\right| = \lim_{n\to\infty} \left|\sum_{j\leqslant n} z_j\right|$ (the absolute value function is continuous). □

### Rearrangements

Recall (Example 2.39) that $S_{\mathbb{N}}$ is the group of all permutations of the natural numbers. A *rearrangement* of a sequence $\langle z_n \rangle$ is a sequence $\langle z_{\sigma(n)} \rangle$ for some $\sigma \in S_{\mathbb{N}}$. For sequences this is not too interesting:

*Exercise 11.33* Let $\langle x_n \rangle$ be a sequence of points in a quasi-Euclidean space $X$, converging to some point $y \in X$. Show that for all $\sigma \in S_{\mathbb{N}}$, $\lim_n x_{\sigma(n)} = y$. «

A rearrangement of the series $\sum z_n$ is $\sum z_{\sigma(n)}$ for some $\sigma \in S_{\mathbb{N}}$. Unlike sequences, a rearrangement of a convergent series could fail to converge, or converge to a different sum (see Exercise 11.101). This is precluded by absolute convergence.

**Proposition 11.34** *Suppose that a series $\sum z_n$ converges absolutely. Then any rearrangement of $\sum z_n$ converges, to the same sum.*

***Proof*** Let $\sigma \in S_{\mathbb{N}}$; let $\alpha = \sum_n z_n$. To show that $\sum z_{\sigma(n)} = \alpha$, given $\varepsilon > 0$, we first choose $N$ sufficiently large such that $\sum_{n>N} |z_n| < \varepsilon$. Now suppose that $M$ is sufficiently large so that $\{\sigma(1), \sigma(2), \ldots, \sigma(M)\}$ contains all of $\{1, 2, \ldots, N\}$ (in other words, $M \geqslant \max\{\sigma^{-1}(1), \sigma^{-1}(2), \ldots, \sigma^{-1}(N)\}$). Then $\left|\sum_{n\leqslant M} z_{\sigma(n)} - \sum_{n\leqslant N} z_n\right|$ is $|z_{k_1} + z_{k_2} + \cdots + z_{k_l}|$, where $k_1, k_2, \cdots > N$; by the triangle inequality, this is bounded by $|z_{k_1}| + \cdots + |z_{k_l}|$, which in turn is bounded by $\sum_{n>N} |z_n| < \varepsilon$. On the other hand, $\left|\alpha - \sum_{n\leqslant N} z_n\right| < \varepsilon$, so altogether, $\left|\alpha - \sum_{n\leqslant M} z_{\sigma(n)}\right| < 2\varepsilon$. □

## 11.3.2 Uniform Convergence

Let $U \subseteq \mathbb{C}$; let $\langle f_n \rangle$ be a sequence of functions, each $f_n : U \to \mathbb{C}$. For each $z \in U$, $\langle f_n(z) \rangle$ is a sequence of complex numbers, and we can ask if it converges or not. We say that $\lim f_n = f$ *pointwise* if for all $z \in U$, $f(z) = \lim_{n\to\infty} f_n(z)$. The rate of convergence of $\langle f_n(z) \rangle$ as $z$ varies could vary as well. Uniform convergence, a strengthening of pointwise convergence, says that it doesn't.

**Definition 11.35** A sequence $\langle f_n \rangle$ of functions from $U$ to $\mathbb{C}$ converges to $f$ *uniformly* if for all $\varepsilon > 0$ there is some $N$ such that for all $n > N$, for all $z \in U$, $|f(z) - f_n(z)| < \varepsilon$.

In other words, if for all but finitely many $n$, the number $\sup_{z\in U}|f(z) - f_n(z)|$ is finite, and the sequence of these numbers tends to 0.

*Exercise 11.36* Give an example of functions $f, f_n\colon [0,1] \to \mathbb{R}$ such that $\langle f_n\rangle$ converges to $f$ pointwise but not uniformly. «

These notions apply to series by considering the sequence of partial sums. In other words, $\sum f_n = f$ pointwise if for all $z \in U$, $f(z) = \sum_n f_n(z)$; if and only if the sequence $\langle \sum_{m\leqslant n} f_m\rangle$ converges pointwise to $f$. And $f = \sum f_n$ uniformly if $f = \lim_n \sum_{m\leqslant n} f_m$ uniformly.

The completeness of $\mathbb{C}$ implies that the sequence $\langle f_n\rangle$ has a limit if and only if for all $z \in U$, $\langle f_n(z)\rangle$ is a Cauchy sequence. However, we say that $\langle f_n\rangle$ is a Cauchy sequence of functions if this is true uniformly; that is, if for all $\varepsilon > 0$ there is some $N$ such that for all $n, m > N$, for all $z \in U$, $|f_n(z) - f_m(z)| < \varepsilon$.

**Proposition 11.37** *A sequence $\langle f_n\rangle$ of functions is Cauchy if and only if it converges uniformly to some $f\colon U \to \mathbb{C}$.*

***Proof*** Suppose that $f = \lim f_n$ uniformly. The argument proving Proposition 8.81 shows that $\langle f_n\rangle$ is a Cauchy sequence of functions; $\varepsilon$ depends on $N$ but not on $z$. In the other direction, let $\langle f_n\rangle$ be a Cauchy sequence of functions. For all $z \in U$, $\langle f_n(z)\rangle$ is a Cauchy sequence of complex numbers, and so has a limit, which we call $f(z)$. Let $\varepsilon > 0$. If for all $n, m \geqslant N$, for all $z \in U$, $|f_n(z) - f_m(z)| \leqslant \varepsilon$, then, taking the limit as $m \to \infty$, for all $n \geqslant N$, for all $z \in U$, $|f(z) - f_n(z)| \leqslant \varepsilon$. □

The notion applies to series as well; a series $\sum f_n$ is Cauchy if the sequence of partial sums is Cauchy, which amounts to: for all $\varepsilon > 0$ there is some $N$ such that for all $m \geqslant n > N$, for all $z \in U$, $\left|\sum_{j=n}^{m} f_j(z)\right| < \varepsilon$.

*Exercise 11.38* Give an example of a sequence of continuous functions $f_n\colon [0,1] \to \mathbb{R}$ which converges pointwise to a discontinuous function $f\colon [0,1] \to \mathbb{R}$. «

Uniform convergence, on the other hand, preserves continuity.

**Proposition 11.39** *Let $\langle f_n\rangle$ be a sequence of continuous functions which converges uniformly to a function $f$. Then $f$ is continuous.*

***Proof*** Let $z \in U$; let $\varepsilon > 0$. Find some $n$ such that for all $w \in U$, $|f(w) - f_n(w)| \leqslant \varepsilon$. Since $f_n$ is continuous, find some open neighbourhood $W \subseteq U$ of $z$ such that for all $w \in W$, $|f_n(w) - f_n(z)| \leqslant \varepsilon$. Then, using the triangle inequality, for all $w \in W$, $|f(w) - f(z)| \leqslant 3\varepsilon$. □

Since the sum of finitely many continuous functions is continuous (Exercise 8.11), we see that Proposition 11.39 applies to series as well: if $f = \sum f_n$ uniformly, and each $f_n$ is continuous, then so is $f$.

We say that $\sum f_n$ converges absolutely if the series $\sum |f_n|$ converges (that is, if for all $z \in U$, $\sum f_n(z)$ converges absolutely). Putting the adverbs together, we say that $\sum f_n$ converges *absolutely uniformly* if the series $\sum |f_n|$ converges uniformly.

*Exercise 11.40* Let $f_n : (-1, 0) \to \mathbb{R}$ be defined by $f_n(x) = x^n/n$. Show that $\sum f_n$ converges absolutely, and uniformly, but not absolutely uniformly. (The last part will be easier later; you might consider $\ln(1 - x)$.) «

**Proposition 11.41** *If $\sum f_n$ converges absolutely uniformly, then $\sum f_n$ converges both absolutely and uniformly.*

***Proof*** Absolute convergence of $\sum f_n$ is immediate. For uniform convergence, note that $\sum |f_n|$ is a Cauchy series of functions, and so $\sum f_n$ is a Cauchy series of functions. □

The following will help show that power series have a radius of convergence.

**Weierstrass $M$-Test** *Let $\sum M_n$ be a convergent series of positive real numbers. Let $\langle f_n \rangle$ be a sequence of functions from $U$ to $\mathbb{C}$, and suppose that for all $n$, $\sup_{z \in U} |f_n(z)| \leqslant M_n$. Then $\sum f_n$ converges absolutely uniformly.*

***Proof*** $\sum M_n$ is a Cauchy series of real numbers; it follows that $\sum |f_n|$ is a Cauchy series of functions. □

### Convergence on Compact Sets

Let $\langle f_n \rangle$ be a sequence of functions defined on $U$, and let $A \subseteq U$. It is sometimes possible that $\langle f_n \rangle$ converges to $f$ pointwise but not uniformly; but that $\langle f_n \restriction_A \rangle$ converges to $f \restriction_A$ uniformly. We say that $\langle f_n \rangle$ converges uniformly *on* $A$.

*Exercise 11.42* Let $U \subseteq \mathbb{C}$ be open, and let $\langle f_n \rangle$ be a sequence of functions defined on $U$ which converges to some function $f : U \to \mathbb{C}$. Show that the following are equivalent:

(1) $f_n$ converges *locally uniformly* to $f$: there is an open cover $\mathcal{O}$ of $U$ such that $\langle f_n \rangle$ converges uniformly to $f$ on each $O \in \mathcal{O}$;
(2) for every compact subset $D \subset U$, $\langle f_n \rangle$ converges uniformly to $f$ on $D$. «

Proposition 11.39 implies its strengthening:

**Proposition 11.43** *Let $\langle f_n \rangle$ be a sequence of continuous functions which converges locally uniformly to a function $f$. Then $f$ is continuous.*

***Proof*** Let $\mathcal{O}$ be an open cover of $U$ such that $f_n \to f$ uniformly on each $O \in \mathcal{O}$. Then for each $O \in \mathcal{O}$, $f \restriction_O$ is continuous; this implies that $f$ is continuous at each $z \in U$, and hence is continuous. □

### 11.3.3 Power Series

Let $\langle c_n \rangle_{n \geqslant 0}$ be a sequence of complex numbers, and let $a$ be a complex number. The power series with coefficients $\langle c_n \rangle$ and centre $a$ is the series of functions $\sum c_n(z - a)^n$. This series will converge for some $z \in \mathbb{C}$ (for example for $z = a$), but may not converge for all. The domain of the sum is the collection of complex numbers $z$ for which the series converges.

**Lemma 11.44** *Let $\sum c_n(z - a)^n$ be a power series, and suppose that the series converges on some number $b \neq a$. Let $r = |b - a|$. Then the series $\sum c_n(z - a)^n$ converges on $B(a, r)$; in fact, for all $s < r$, this power series converges absolutely uniformly on $B(a, s)$.*

***Proof*** The sum $\sum c_n(b - a)^n$ converges and so the sequence $\langle c_n(b - a)^n \rangle$ tends to 0 (Proposition 11.29) and so is bounded, say by some $M > 0$; this means that $|c_n| r^n \leqslant M$ for all $n$. Let $s < r$. Then for all $z \in B(a, s)$,

$$|c_n(z - a)^n| \leqslant |c_n| s^n = |c_n| r^n \left(\frac{s}{r}\right)^n \leqslant M \left(\frac{s}{r}\right)^n .$$

Since $s/r < 1$, the series $\langle M(s/r)^n \rangle$ is a convergent geometric series (Exercise 11.28). The result then follows from the Weierstrass $M$-Test. □

Let $R = \sup \left\{ r \geqslant 0 : \sum c_n(z - a)^n \text{ converges on } B(a, r) \right\}$, and let $f$ be the sum of the series $\sum c_n(z - a)^n$. Lemma 11.44 says that

$$B(a, R) \subseteq \operatorname{dom} f \subseteq \overline{B}(a, r)$$

and that the series converges absolutely uniformly on $B(a, r)$ for all $r < R$. The number $R$ is called the *radius of convergence* of the series. Of course if $R = \infty$ then $\operatorname{dom} f = \mathbb{C}$; $R = 0$ means that $\operatorname{dom} f = \{a\}$. The restriction of $f$ to $B(a, R)$ is continuous, because $\sum c_n(z - a)^n$ converges locally uniformly on $B(a, R)$ (Proposition 11.43).

On the boundary, the series may converge on some, all, or none of the points; see Exercises 11.102 and 11.103. It may even converge and be discontinuous on the boundary.

The *root test* is used to calculate the radius of convergence. It uses the notion of the *limit superior* of a sequence.

*Exercise 11.45* Let $\langle r_n \rangle$ be a sequence of real numbers, bounded from above. Let $s_n = \sup_{k \geqslant n} r_k$. (a) Show that the sequence $\langle s_n \rangle$ is non-increasing.

Since the sequence $\langle s_n \rangle$ is also bounded from below, we know it has a limit (Exercise 8.88), which we denote by $\limsup_n r_n$. If $\langle r_n \rangle$ is unbounded from above then we write $\limsup_n r_n = \infty$.

(b) Show that $\limsup_n r_n$ is the greatest number $s$ such that for all $\varepsilon > 0$, for all but finitely many $n$, $s \geqslant r_n - \varepsilon$. (c) Show that $\limsup_n r_n$ is the greatest number $s$ such that there is a subsequence of $\langle r_n \rangle$ which converges to $s$. (d) Show that if $\lim a_n = a > 0$ then $\limsup_n (a_n r_n) = a \cdot \limsup_n r_n$. «

**Proposition 11.46** *Let $R$ be the radius of convergence of the power series $\sum c_n(z-a)^n$. Then $1/R = \limsup_n \sqrt[n]{|c_n|}$.*

The root test also holds in the extreme cases $R = 0$ or $R = \infty$, if $1/0$ is understood as $\infty$ and $1/\infty$ as 0.

***Proof*** Let $\alpha = \limsup_n \sqrt[n]{|c_n|}$, and let $b \in \mathbb{C}$. We need to show that if $|b-a| < 1/\alpha$ then $\sum c_n(b-a)^n$ converges, and that if $|b-a| > 1/\alpha$ then $\sum c_n(b-a)^n$ does not converge.

First suppose that $|b-a| < 1/\alpha$, that is, that $\alpha < 1/|b-a|$. By (b) of Exercise 11.45, for all but finitely many $n$, $\sqrt[n]{|c_n|}$ is bounded below $1/|b-a|$; this implies that there is some $q < 1$ such that for all but finitely many $n$, $\sqrt[n]{|c_n|} \leqslant q/|b-a|$. So for all but finitely many $n$, $|c_n(b-a)^n| \leqslant q^n$. It follows that $\sum c_n(b-a)^n$ converges absolutely.

If $1/|b-a| < \alpha$ then there are infinitely many $n$ such that $1/|b-a| < \sqrt[n]{|c_n|}$, which we simplify to $|c_n(b-a)^n| > 1$. Then $\sum c_n(b-a)^n$ does not converge by Proposition 11.29. □

## 11.4 Analytic Functions

**Definition 11.47** Let $U \subseteq \mathbb{C}$ be open. A function $f : U \to \mathbb{C}$ is *analytic* if locally it is the sum of a power series: for every $a \in \operatorname{dom} f$ there is a power series $\sum c_n(z-a)^n$ which converges to $f$ on some open disc containing $a$.

*Exercise 11.48* Let $\sum c_n(z-a)^n$ be a power series with radius of convergence $R > 0$. Show that the sum of the series is analytic on $B(a, R)$. (Hint: use Exercise 11.85

below; if $b \in B(a, R)$, show that $\sum_{n=0}^{\infty} \sum_{k=0}^{n} c_n \binom{n}{k} (b-a)^{n-k}(z-b)^k$ converges absolutely for $z$ sufficiently close to $b$.) «

Analytic functions are rigid compared to merely smooth functions. In some ways they resemble polynomials. Here is an example.

If $f : U \to \mathbb{C}$ then we say that a zero $z$ of $f$ is isolated if it is isolated in the set of zeros of $f$. That is, if there is some neighbourhood of $z$ in which $z$ is the only zero of $f$ (see Definition 8.99).

**Lemma 11.49** *Let $f : U \to \mathbb{C}$ be analytic. If some zero $a$ of $f$ is not isolated then $f = 0$ on a neighbourhood of $a$.*

***Proof*** Let $a$ be a zero of $f$. Let $V \subseteq U$ be an open neighbourhood of $a$ on which $f = \sum c_n (z-a)^n$. Since $f(a) = c_0$ we have $c_0 = 0$. If $c_k = 0$ for all $k$ then $f = 0$ on $V$. Otherwise, let $k$ be the least such that $c_k \neq 0$. Then $f = (z-a)^k \sum_{n \geqslant k} c_n (z-a)^{n-k}$. Define $g(z) = \sum_{n \geqslant k} c_n (z-a)^{n-k}$; it converges on $V$. For $z \neq a$ in $V$, $(z-a)^k \neq 0$; and $g(a) \neq 0$. Since $g$ is continuous (see after Lemma 11.44), there is a neighbourhood of $a$ on which $g \neq 0$. Also $(z-a)^k \neq 0$ if $z \neq a$. It follows that $a$ is an isolated zero of $f$. □

We get a generalisation of the fact that polynomials have finitely many roots:

**Proposition 11.50** *Let $U \subseteq \mathbb{C}$ be a region (open and connected) and let $f : U \to \mathbb{C}$ be analytic. If the set of zeros of $f$ is not discrete then $f = 0$.*

***Proof*** Let $V \subseteq U$ be the collection of non-isolated zeros of $f$. Lemma 11.49 implies that $V$ is open. But $V$ is also closed in $U$. Suppose that $w \in U \setminus V$. If $f(w) \neq 0$ then continuity of $f$ implies that $f \neq 0$ on a neighbourhood of $w$, and this neighbourhood is disjoint from $V$. Otherwise $w$ is an isolated zero of $f$, which again means that it has a neighbourhood disjoint from $V$. □

By considering $f - g$, we see that if $f, g : U \to \mathbb{C}$ are analytic and $U$ is connected, then the points on which $f$ equals $g$ are isolated, or $f = g$.

*Example 11.51* Unlike polynomials, analytic functions may have infinitely many zeros; for example take the function $e^z - 1$ (see Exercise 11.60 below). «

### 11.4.1 Differentiating Power Series

Exercise 11.14 implies the following (again recall that we are only considering piecewise smooth paths):

**Proposition 11.52** *Let $\gamma$ be a path in $\mathbb{C}$, and let $\langle f_n \rangle$ be a sequence of continuous functions which converges uniformly on the image of $\gamma$. Then*

$$\lim_n \int_\gamma f_n \, dz = \int_\gamma \left( \lim_n f_n \right) dz.$$

**Proposition 11.53** *Let $U \subseteq \mathbb{C}$ be a region and let $f_n \colon U \to \mathbb{C}$ be continuously differentiable. Suppose that $f_n$ converge pointwise to $f$ and that $f_n'$ converge locally uniformly to $g$. Then $g = f'$.*

Note that $g$ is continuous (Proposition 11.43) and so $f$ is continuously differentiable.

***Proof*** Let $\gamma$ be a path in $U$. The image of $\gamma$ is compact. By Exercise 11.42, $f_n' \to g$ uniformly on the image of $\gamma$. By Proposition 11.52, $\int_\gamma f_n' \, dz \to \int_\gamma g \, dz$. As mentioned, $g$ is continuous. By Lemma 11.17, $\int_\gamma f_n' \, dz = f_n(\gamma(b)) - f_n(\gamma(a))$ (where $[a, b] = \operatorname{dom} \gamma$). Taking limits (as $f_n \to f$), $\int_\gamma g \, dz = f(\gamma(b)) - f(\gamma(a))$. By Lemma 11.18, $g = f'$. □

**Proposition 11.54** *Let $R$ be the radius of convergence of a power series $\sum c_n (z - a)^n$. Then on $B(a, R)$ the sum $f$ of the series is continuously differentiable, and $f' = \sum n c_n (z - a)^{n-1}$ on $B(a, R)$.*

In fact, the radius of convergence of $\sum n c_n (z-a)^{n-1}$ is precisely $R$. As a result, we see that if $f$ is analytic then it is continuously differentiable and $f'$ is analytic.

To prove Proposition 11.54 we require the following.

*Exercise 11.55* Show that $\lim_n \sqrt[n]{n} = 1$. (There are several ways. Here's one: first show that for $n \geqslant 3$, $(n+1)^n < n^{n+1}$, using the binomial formula for $(n+1)^n$. Conclude that $\sqrt[n+1]{n+1} < \sqrt[n]{n}$. Since $\sqrt[n]{n} > 1$, Exercise 8.88 implies the sequence $\langle \sqrt[n]{n} \rangle$ has a limit $\alpha \geqslant 1$. Next show that for all $q > 1$, for all but finitely many $n$, $q^n > n$; this is because $\sum (1/q)^n$ converges (Exercise 11.28) but the harmonic series does not (Exercise 11.30). Conclude that $\alpha$ cannot be greater than 1, so must equal 1.) «

*Exercise 11.56* Show that if $q > 1$ then $\lim q^n / n = \infty$. «

***Proof of Proposition 11.54*** For all $n$, $\sqrt[n]{|nc_n|} = \sqrt[n]{n} \cdot \sqrt[n]{|c_n|}$; by Exercises 11.45 and 11.55, $\limsup_n \sqrt[n]{n|c_n|} = \lim_n \sqrt[n]{n} \limsup_n \sqrt[n]{|c_n|} = \limsup_n \sqrt[n]{|c_n|}$, so the radius of convergence of the series $\sum n c_n (z-a)^n$ is $R$. However, for all $z$, $\sum n c_n (z-a)^n = (z-a) \sum n c_n (z-a)^{n-1}$, which shows that the series $\sum n c_n (z-a)^{n-1}$ converges if and only if the series $\sum n c_n (z-a)^n$ converges; so the radius of convergence of $\sum n c_n (z-a)^{n-1}$ is $R$ as well.

We know that on $B(a, R)$, $\sum nc_n(z-a)^{n-1}$ is locally uniformly convergent (Lemma 11.44); the result now follows from Proposition 11.53. □

A function $f : U \to \mathbb{C}$ is *infinitely differentiable* if it is continuously differentiable, $f'$ is continuously differentiable, $f''$ is continuously differentiable, ...

**Corollary 11.57** *An analytic function is infinitely differentiable.*

The following is Taylor's formula.

**Proposition 11.58** *Let $a \in \mathbb{C}$; suppose that $f = \sum c_n(z-a)^n$ on some open neighbourhood of $a$. Then $n!c_n = f^{(n)}(a)$, where $f^{(n)}$ is the nth derivative of $f$.*

In particular, this shows that a local power series representation is unique: if $\sum c_n(z-a)^n = \sum d_n(z-a)^n$ on some open neighbourhood of $a$, then $c_n = d_n$ for all $n$.

***Proof*** Differentiating $k$ times (Proposition 11.54), we get that

$$f^{(k)}(z) = \sum_{n \geqslant k} \frac{n!}{(n-k)!} \cdot c_n(z-a)^{n-k}$$

on a neighbourhood of $a$; we evaluate at $z = a$. □

*Exercise 11.59* Let $f = \sum c_n z^n$ around 0. Show that if $f(-z) = -f(z)$ around 0 ($f$ is odd) then $c_n = 0$ for all even $n$. Reach a similar conclusion about even functions ($f(-z) = f(z)$). Show that if $f$ is even then $f'$ is odd, and vice-versa. «

### 11.4.2 The Exponential and Trigonometric Functions

We finally have all the tools necessary to carry out the argument sketched in Chap. 1; however, we extend the exponential function to all of $\mathbb{C}$.

Define, for all $z \in \mathbb{C}$,

$$\exp z = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots$$

*Exercise 11.60* An *entire function* is an analytic function whose domain is all of $\mathbb{C}$. (a) Show that exp is entire. (b) Show that $\exp' = \exp$. (c) Show that if $f$ is an entire function and $f' = f$ then $f = c \cdot \exp$ for some constant $c \in \mathbb{C}$. (d) Conclude that for all $z, w \in \mathbb{C}$, $\exp(z+w) = \exp z \cdot \exp w$. «

Let $e = \exp(1)$.

*Exercise 11.61* (a) Show that for all $n \in \mathbb{Z}$, $\exp(n) = e^n$. (b) Show that for all $n \geqslant 1$, $\exp(1/n) = \sqrt[n]{e}$. «

In light of this, we usually write $e^z$ for $\exp z$.

*Exercise 11.62* (a) Show that if $t \in \mathbb{R}$ then $e^t \in \mathbb{R}$. (b) Show that the real exponential function $t \mapsto e^t$ (for $t \in \mathbb{R}$) is increasing and maps $\mathbb{R}$ onto $(0, \infty)$. (Hint: for $t \geqslant 0$, $f(t) > t$; use the intermediate value theorem and $e^{-t} = 1/e^t$.) (c) Define ln to be the inverse of the real exponential. Show that ln is smooth, and that its derivative is $1/t$. «

Define $E(z)$ to be the even part of $e^z$, namely $E(z) = 1+z^2/2!+z^4/4!+\cdots$, and $O(z) = e^z - E(z)$ to be the odd part of $e^z$, that is, $O(z) = z + z^3/3! + z^5/5! + \cdots$. (See Exercise 11.59). Then, define $\cos z = E(iz)$ and $\sin z = O(iz)/i$. So $e^{iz} = \cos z + i \sin z$.

*Exercise 11.63* (a) Show that both cos and sin are entire, that $\sin' = \cos$ and that $\cos' = -\sin$. (b) Show that $|e^{it}| = 1$ for all $t \in \mathbb{R}$. (c) Show that $\cos 2 < 0$. (Consider the first three terms of the power series, and note that it is alternating.) (d) Follow the rest of the argument from Sect. 1.1 to define $\pi$, and verify that the circumference of the unit circle is $2\pi$. (e) Show that $2\pi i$ is a period of exp: for all $z \in \mathbb{C}$, $\exp(z + 2\pi i) = \exp z$. (f) Show that the range of exp is $\mathbb{C} \setminus \{0\}$. (g) Prove the addition formulas for $\cos(z + w)$ and $\sin(z + w)$. «

*Exercise 11.64* Let $U \subseteq \mathbb{C} \setminus \{0\}$ be simply connected; let $\alpha$ be a continuous choice of argument on $U$ (Proposition 9.30). Show that the map $\ln_\alpha(z) = \ln|z| + i\alpha(z)$ is continuously differentiable on $U$. «

*Example 11.65* Following Example 9.31, for any $\rho \in \mathbb{R}$, for $z \in \mathbb{C} \setminus \{re^{i\rho} : r \geqslant 0\}$, let $\alpha_\rho(z)$ be the unique argument of $z$ in the interval $(t, t + 2\pi)$; then $\ln_\rho(z) = \ln|z| + i\alpha_\rho(z)$ is called a *branch* of the complex logarithm. «

For more on complex logarithms and exponentiation see Exercise 11.87 and some following exercises. A more conceptual understanding of the multi-valued nature of the complex logarithm is given by the *Riemann surface for the logarithm*, which we discuss in Chap. 12.

*Remark 11.66* The course of this book regarding the exponential function has been a bit topsy-turvey. We have started with an informal development in Chap. 1, and then have sporadically used the function $t \mapsto e^{it}$; only now have we given a more formal treatment. A careful inspection though will show that this function has been mostly used in examples, and discussions of the winding number. The formal development of the complex exponential function only relied on results in

this section did not rely on facts deduced using the winding number, and so our development is not, well, (logically) circular. «

### 11.4.3 Continuously Differentiable Functions Are Analytic

The following is the stark difference between real and complex analysis: continuous differentiability implies infinite differentiability, and more.

**Theorem 11.67** *A function $f : U \to \mathbb{C}$ is continuously differentiable if and only if it is analytic.*

***Proof*** The easy direction follows from Corollary 11.57. In the main direction, suppose that $f$ is continuously differentiable.

Given $a \in U$, let $\gamma$ be a parameterised circle with centre $a$ and radius $r$, sufficiently small so that $\overline{B}(a,r) \subset U$. Let $b \in B(a,r)$; let $z$ be a point on the circle. Since $|z-a| = r > |b-a|$, the geometric series $\sum (b-a)^n/(z-a)^n$ converges (Exercise 11.28), indeed

$$\sum_{n\geqslant 0} \frac{(b-a)^n}{(z-a)^n} = \frac{1}{1-\frac{b-a}{z-a}} = \frac{z-a}{z-b},$$

whence

$$\frac{1}{z-b} = \sum_{n\geqslant 0} \frac{(b-a)^n}{(z-a)^{n+1}}.$$

Since $\gamma$ is also a contour around $b$ (Proposition 11.27), by Cauchy's Integral Formula,

$$f(b) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{z-b}\,dz = \int_\gamma \left(\sum_{n\geqslant 0} \frac{f(z)(b-a)^n}{2\pi i(z-a)^{n+1}}\right) dz$$

Let $M = \max\{|f(z)|/2\pi r : z \in \text{range}\,\gamma\}$; then for all $z$ in the image of $\gamma$, $|g_n(z)| \leqslant Mq^n$ where $q = |b-a|/|z-a|$ and

$$g_n(z) = \frac{f(z)(b-a)^n}{2\pi i(z-a)^{n+1}}.$$

The Weierstrass $M$-Test implies that the series $\sum g_n$ converges uniformly on the image of $\gamma$; Proposition 11.52 then implies that we can exchange summation and integration and get

$$f(b) = \int_\gamma \sum_{n\geqslant 0} g_n(z)\,dz = \sum_{n\geqslant 0} \int_\gamma g_n(z)\,dz.$$

Let

$$c_n = \int_\gamma \frac{g_n(z)}{(b-a)^n}\,dz = \int_\gamma \frac{f(z)}{2\pi i(z-a)^{n+1}}\,dz,$$

and note that $c_n$ does not depend on $b$. Then $f(b) = \sum_{n\geqslant 0} c_n(b-a)^n$, giving a power series representation for $f$ on $B(a,r)$. □

*Remark 11.68* Theorem 11.67 and Proposition 11.54 together give us an indirect proof of Exercise 11.48, that the sum of a power series is analytic. «

*Remark 11.69* More is true: the assumption of *continuous* differentiability can be dropped in Theorem 11.67. That is, if a function $f\colon U \to \mathbb{C}$ (with $U \subseteq \mathbb{C}$ open) is (complex) differentiable at every point of $U$, then it is analytic. This is not too difficult to prove, but requires a different argument, and we will not need it. «

In the proof of Theorem 11.67, we could take any radius $r$, as long as $f$ was defined on the image of $\gamma$. Together with the uniqueness of power series representations (Proposition 11.58) we obtain:

**Corollary 11.70** *Let $f\colon U \to \mathbb{C}$ be analytic, let $a \in U$, and suppose that $f = \sum c_n(z-a)^n$ on some neighbourhood of $a$. Let $r > 0$ and suppose that $B(a,r) \subseteq \operatorname{dom} f$. Then the radius of convergence of the series $\sum c_n(z-a)^n$ is at least $r$, and $f = \sum c_n(z-a)^n$ on $B(a,r)$.*

Finally, equating Taylor's formula (Proposition 11.58) with the expression we got for $c_n$ in the proof of Theorem 11.67, we obtain a generalisation of Cauchy's Integral Formula.

**Proposition 11.71** *Let $U$ be simply connected, $f\colon U \to \mathbb{C}$ be analytic, $a \in U$, and let $\gamma$ be a contour in $U$ around $a$. Then*

$$f^{(n)}(a) = \frac{n!}{2\pi i}\int_\gamma \frac{f(z)}{(z-a)^{n+1}}\,dz.$$

***Proof*** As mentioned, the proof of Theorem 11.67 gives the proposition in the case that $\gamma$ is a parameterised circle around $a$ with $\overline{B}(a,r) \subset U$. For other loops $\gamma$,

by Proposition 11.24, we can fix small $r$ such that $\gamma$ is a contour around every $b \in B(a, r)$, and repeat the proof.[5] □

## 11.5 Morera, Weierstrass, Liouville

Here are some corollaries.

**Theorem 11.72 (Morera's Theorem)** *Let $U \subseteq \mathbb{C}$ be simply connected and let $f : U \to \mathbb{C}$ be continuous. Then $f$ is analytic if and only if it has a primitive on $U$.*

***Proof*** One direction follows from Proposition 11.19 and the equivalence between analytic and continuously differentiable functions (Theorem 11.67). In the other direction suppose that $f$ has a primitive $g$ on $U$. Then $g$ is continuously differentiable, hence analytic, and so $f = g'$ is analytic (Corollary 11.57). □

And we can miss a point. We obtain the following proposition by combining Morera's theorem and Lemma 11.20:

**Proposition 11.73** *Let $U$ be open, $a \in U$, and suppose that $f : U \to \mathbb{C}$ is continuous on $U$ and analytic on $U \setminus \{a\}$. Then $f$ is actually analytic on all of $U$.*

Since being continuously differentiable is a local property, we can extend this to more than one "missing" point, as long as they are separated from each other:

**Corollary 11.74** *Let $U$ be open and let $f : U \to \mathbb{C}$ be continuous. Suppose that $F \subset U$ is discrete and closed in $U$, and that $f|_{U \setminus F}$ is analytic. Then $f$ is analytic.*

***Proof*** The assumption means that there is an open cover $\mathcal{O}$ of $U$ such that each $O \in \mathcal{O}$ intersects $F$ in at most one point (Exercise 8.102). By Proposition 11.73, $f$ is analytic on every $O \in \mathcal{O}$. □

### 11.5.1 Liouville's Theorem

The following is known as *Cauchy's estimate*.

[5] One could try to use Proposition 9.29 and Lemma 10.37, but for this, we need to know that if $U \subseteq \mathbb{C}$ is open and simply connected, and $a \in U$, then any two paths in $U \setminus \{a\}$ with the same winding number are homotopic in $U \setminus \{a\}$, rather than merely in $\mathbb{C} \setminus \{a\}$. This is true, but requires some more tools.

**Lemma 11.75** *Suppose that $f = \sum c_n(z-a)^n$ on $B(a,R)$ and that $|f| \leqslant M$ on $B(a,R)$. Then for all $n$, $|c_n| \leqslant M/R^n$.*

***Proof*** Let $r < R$. By Proposition 11.71 and Taylor's formula Proposition 11.58, for all $n$,

$$c_n = \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{(z-a)^{n+1}}\,dz,$$

where $\gamma$ is the parameterised circle of radius $r$ around $a$. By Exercise 11.14,

$$|c_n| \leqslant \frac{1}{2\pi}\int_\gamma \frac{|f(z)|}{r^{n+1}}\,ds \leqslant \frac{1}{2\pi}\ell(\gamma)\frac{M}{r^{n+1}} = \frac{M}{r^n};$$

now let $r \to R$. □

Recall (Exercise 11.60) that an analytic function is called *entire* if its domain is all of $\mathbb{C}$. It is called *bounded* if its range is a bounded subset of $\mathbb{C}$. That is, if for some $M$, $|f(z)| \leqslant M$ for all $z \in \operatorname{dom} f$.

**Theorem 11.76 (Liouville's Theorem)** *A bounded entire function is constant.*

***Proof*** Suppose that $|f(z)| \leqslant M$ for all $z \in \mathbb{C}$. Since $f$ is entire, Corollary 11.70 says that there is a power series $\sum c_n z^n$ with radius of convergence $\infty$ whose sum equals $f$ everywhere. Lemma 11.75 says that for all $n$, $|c_n| \leqslant M/R^n$ for all $R > 0$. If $n \geqslant 1$ this means that $c_n = 0$. □

We settle an old debt: the fundamental theorem of algebra (Theorem 2.30), which says that the field $\mathbb{C}$ is algebraically closed.

***Proof of the Fundamental Theorem of Algebra*** Let $f \in \mathbb{C}[x]$ be a polynomial with no root; we show it is constant. The function $f\colon \mathbb{C} \to \mathbb{C}$ determined by $f$ is analytic and entire (it is a very simple power series!). Since $f$ has no root, $1/f$ is entire (it is analytic by Proposition 11.7). We will show that $1/f$ is bounded; Liouville's Theorem 11.76 will then imply that $1/f$, and so $f$, is constant.

Say $f(z) = a_d z^d + \cdots + a_0$ (with $a_d \neq 0$). Roughly, the reason that $1/f$ is bounded, is that when $|z|$ is large, the term $|a_d z^d|$ dominates the rest of $|f(z)|$, so $|f(z)|$ grows like $|z^d|$. More formally, for $z \neq 0$ we can write

$$f(z) = z^d \cdot \left(a_d + \frac{a_{d-1}}{z} + \frac{a_{d-2}}{z^2} + \cdots + \frac{a_0}{z^d}\right);$$

for brevity let $g(z) = a_{d-1}/z + a_{d-2}/z^2 + \cdots + a_0/z^d$. If $N = d \cdot \max\{|a_0|, |a_1|, \ldots, |a_{d-1}|\}$ then for all $z$ such that $|z| > 1$, $|g(z)| \leqslant N/|z|$. So if we let $R = \max\{1, 2N/|a_d|\}$ then $|g(z)| \leqslant |a_d|/2$ whenever $|z| > R$; so

$|f(z)| \geqslant |z^d| \cdot |a_d|/2$ for such $z$, giving $|1/f| \leqslant 2/(R^d|a_d|)$ outside $\overline{B}(0, R)$. On the other hand, since $1/f$ is continuous and $\overline{B}(0, R)$ is compact, on that closed disc, $1/f$ obtains a maximum. Together, this shows that $1/f$ is indeed bounded. □

**Weierstrass's Theorem**

In contrast with real valued functions (see Exercise 11.100), complex derivatives and uniform limits play nice.

**Theorem 11.77 (Weierstrass's Theorem)** *Let $U \subseteq \mathbb{C}$ be open, and suppose that $f_n : U \to \mathbb{C}$ is a sequence of analytic functions which converge, locally uniformly, to some $f : U \to \mathbb{C}$. Then $f$ is analytic and $\langle f_n' \rangle$ converges, locally uniformly, to $f'$.*

***Proof*** By Proposition 11.53 it suffices to show that the sequence $\langle f_n' \rangle$ converges locally uniformly to some function. Let $a \in U$; by assumption, there is an open neighbourhood $V \subseteq U$ of $a$ on which the $f_n$'s converge uniformly. Let $r > 0$ such that $\overline{B}(a, 2r) \subset V$, and let $\gamma$ be the parameterised circle with centre $a$ and radius $2r$. Let $b \in B(a, r)$.

Let $\varepsilon > 0$; choose some sufficiently large $N$ so that $|f_n - f_m| < \varepsilon$ on $V$ for all $n, m \geqslant N$. Then as $\gamma$ is a contour around $b$, by Proposition 11.71, for such $n$ and $m$, since $|z - b| \geqslant r$ for all $z$ in the image of $\gamma$,

$$|f_n'(b) - f_m'(b)| \leqslant \frac{1}{2\pi} \int_\gamma \frac{|f_n(z) - f_m(z)|}{r^2}\, ds \leqslant \frac{4\pi r}{2\pi} \frac{\varepsilon}{r^2} = \frac{2\varepsilon}{r},$$

so the sequence $\langle f_n' \rangle$ is Cauchy on $B(a, r)$. □

## 11.6 Further Exercises

11.78 We give an alternative proof of Proposition 11.24. Define the loop $\eta$ as in the proof of that proposition; we need to show that $\mathrm{wnd}_\eta(0) = \mathrm{wnd}_\gamma(0)$. Let $z_0$ be the start and end-point of $\gamma$. Let $\mu$ be the loop which travels along the straight line segment from $z_0$ to $z_0 - p$, then along $\eta$, and then back from $z_0 - p$ to $z_0$, again along the straight line segment. (a) Show that $\gamma$ and $\mu$ are homotopic in $\mathbb{C}^*$. (b) Show that $\mathrm{wnd}_\mu(0) = \mathrm{wnd}_\eta(0)$. (c) Conclude that $\mathrm{wnd}_\eta(0) = \mathrm{wnd}_\gamma(0)$.

**The Cauchy-Riemann Equations**

11.79 Let $f: \mathbb{C} \to \mathbb{C}$ be $\mathbb{R}$-linear (linear as a map from $\mathbb{R}^2$ to $\mathbb{R}^2$). Show that $f$ is analytic if and only if $f(z) = az$ for some $a \in \mathbb{C}$.

11.80 At which points $z \in \mathbb{C}$ is the function $z \mapsto |z|^2$ complex differentiable? (For more see [Rem91, Ex.1.4.2].)

11.81 Let $U \subseteq \mathbb{C}$ be a region, and let $f: U \to \mathbb{C}$ be complex differentiable. Suppose that $\alpha f + \beta \bar{f} = 0$ for some constant $\alpha, \beta \in \mathbb{C}$, not both 0. Show that $f$ is constant. (Hint: if $\beta \neq 0$ then $\bar{f}$ is complex differentiable. See also [Rem91, Ex.1.3.1].)

**Harmonic Functions**

Let $U \subseteq \mathbb{C}$ be open. A twice-smooth function $u: U \to \mathbb{R}$ is *harmonic* if it satisfies Laplace's equation $D^{xx}u + D^{yy}u = 0$, that is, if the vector field $F_u = (-D^y u, D^x u)$ is symmetric.

11.82 Let $U \subseteq \mathbb{C}$ be open. Show that if $f = u + iv$ is analytic on $U$ then both $u$ and $v$ are harmonic.

A *harmonic conjugate* of a harmonic function $u$ is a function $v: \mathbb{C} \to \mathbb{R}$ such that $f = u + iv$ is analytic.

11.83 Let $U \subseteq \mathbb{R}^2$ be simply connected. (a) Show that if $u: U \to \mathbb{R}$ is harmonic then it has a harmonic conjugate on $U$. (b) Show that the harmonic conjugate of $u$ is unique up to an additive constant. (c) Show that $u(x, y) = 3x^2y + 2x^2 - y^3 - 2y^2$ is harmonic on $\mathbb{C}$, and find its harmonic conjugate.

11.84 Let $u(z) = \ln|z|$ be defined on $\mathbb{C} \setminus \{0\}$. (a) Show that $u$ is harmonic. (b) Show that $u$ does not have a harmonic conjugate on all of $\mathbb{C} \setminus \{0\}$.[6]

**Double Summation**

Let $\langle a_{n,m} \rangle_{n,m \in \mathbb{N}}$ be an infinite array of complex numbers. We say that $\sum_{n,m} a_{n,m} = c$ if for all $\varepsilon > 0$ there is some $N$ such that for all $n, m \geqslant N$, $|c - \sum_{k \leqslant n, l \leqslant m} a_{k,l}| < \varepsilon$. If there is such $c$ it is unique, and we say that $\sum a_{n,m}$ converges. We say that $\sum_{n,m} a_{n,m}$ converges absolutely if $\sum |a_{n,m}|$ converges.

[6] Thus, the assumption in Exercise 11.83 that $U$ is simply connected is necessary.

11.85 Suppose that $\sum a_{n,m}$ converges absolutely. Show that for all $n$, $\sum_m a_{n,m}$ converges absolutely, and similarly for $\sum_n a_{n,m}$; and that $\sum_n \sum_m a_{n,m}$ and $\sum_m \sum_n a_{n,m}$ converge absolutely and

$$\sum_n \sum_m a_{n,m} = \sum_{n,m} a_{n,m} = \sum_m \sum_n a_{n,m}.$$

Also show that for any bijection $k \mapsto (n_k, m_k)$ from $\mathbb{N}$ to $\mathbb{N}^2$, $\sum_k a_{n_k,m_k}$ converges absolutely to $\sum a_{n,m}$.

11.86 Show that if $\sum_n a_n$ and $\sum_n b_n$ both converge absolutely, then $\sum_{n,m} a_n b_m$ converges absolutely to $(\sum_n a_n) \cdot (\sum_n b_n)$.

**The Exponential and Trigonometric Functions**

The *principal branch* of the complex logarithm is $\ln_{-\pi}$, often denoted by Log. That is, $\operatorname{Log} z = \ln r + i\theta$ where $\theta \in (-\pi, \pi)$ is an argument of $z$. It is defined for all complex numbers except for the negative real numbers (and 0).

For $z, w \in \mathbb{C}$ with $\operatorname{Log} z$ defined, we let

$$z^w = \exp(w \operatorname{Log} z).$$

11.87 Show that: (a) $z^1 = z$; (b) $z^{w_1+w_2} = z^{w_1} z^{w_2}$; (c) $\operatorname{Log}'(z) = 1/z$.

11.88 For all complex $\alpha \in \mathbb{C}$ and $n \in \mathbb{N}$, let

$$\binom{\alpha}{n} = \frac{\alpha(\alpha-1)(\alpha-2)\cdots(\alpha-(n-1))}{n!}.$$

Show that $(1+z)^\alpha = \sum \binom{\alpha}{n} z^n$, with radius of convergence 1 when $\alpha \notin \mathbb{N}$.

11.89 In real analysis, the function $f(x) = e^{-1/x^2}$ (extended with $f(0) = 0$) is the standard example of an infinitely differentiable function which is not analytic. Why is $f(z) = e^{-1/z^2}$ not analytic on $\mathbb{C}$?

11.90 At what points is the function $z \mapsto \sin(\bar{z})$ complex differentiable?

**Contour Integrals, Cauchy's Integral Formula, and Liouville's Theorem**

11.91 Let $\gamma\colon [0, 2\pi] \to \mathbb{C}$ be the parameterised circle of radius 2 and centre $i$. Calculate $\int_\gamma \bar{z}\, dz$.

11.92 Let $\gamma$ be a path in $\mathbb{C} \setminus \{0\}$ from 1 to a point $w = re^{i\theta}$. Show that $\int_\gamma dz/z = \ln r + i(\theta + 2\pi k)$ for some $k \in \mathbb{Z}$.

11.93 Let $f\colon U \to \mathbb{C}$ be analytic, and suppose that $\gamma$ is a path in $U$ from $a$ to $b$ which avoids all $z$ for which $f(z)$ is a nonpositive real number. Show that $\int_\gamma (f'/f)dz = \operatorname{Log} f(b) - \operatorname{Log} f(a) + 2\pi ik$ for some $k \in \mathbb{Z}$.

11.94 [PC] (a) Let $\gamma$ be the parameterised circle of radius 1/2 around 0. Calculate $\int_\gamma 2\, dz/(z^2-1)$. (b) Let $\gamma$ be the parameterised circle of radius 5 around 0. Calculate $\int_\gamma \sin z\, dz/(z+1)^7$.

11.95 Let $f$ be an entire function. (a) Suppose that there is a polynomial $p \in \mathbb{R}[x]$ (with real coefficients) such that for all $z$, $|f(z)| \leqslant p(|z|)$. Show that $f$ is a polynomial. (b) Suppose that $|f(z)| \leqslant \ln(|z| + 1)$ for all $z$. Show that $f$ is constant.

11.96

(a) Show that the function

$$f(z) = \frac{z}{(z-1)(z-2)(z-3)}$$

has a primitive on $\{z \in \mathbb{C} : |z| > 4\}$.

(b) Does the function

$$f(z) = \frac{z^2}{(z-1)(z-2)(z-3)}$$

have a primitive on $\{z \in \mathbb{C} : |z| > 4\}$?

11.97 Let $f$ be a nonconstant entire function. Show that $f[\mathbb{C}]$ is dense in $\mathbb{C}$.[7]

[7] Recall that this means that every open subset of $\mathbb{C}$ contains a point in the image of $f$.

**Convergence, Power Series, Weierstrass's Theorem**

11.98 Find the power series expansion of $f(z) = 1/(3 - z)$ centred at the point $4i$, and calculate the radius of convergence.

11.99 Let $F_n$ be the $n$th Fibonacci number ($F_1 = F_0 = 1$, $F_{n+2} = F_{n+1} + F_n$); let $f(z) = \sum F_n z^n$. Let $R$ be the radius of convergence of this power series. (a) Show that $f(z) = 1/(1 - z - z^2)$ for all $z \in B(a, R)$. (b) Show that $1/R = (1 + \sqrt{5})/2$ (the "golden ratio"). (c) Conclude that $(1 + \sqrt{5})/2 = \lim_n \sqrt[n]{F_n}$.

11.100 For $x \in \mathbb{R}$ let $f_n(x) = \sin(n^2 x)/n$. Show that $\langle f_n \rangle$ converges (locally uniformly) to a smooth function on $\mathbb{R}$, but that $\langle f_n' \rangle$ does not. (Contrast with Proposition 11.53. Why does $\langle f_n \rangle$ not contradict Weierstrass's 11.77? After all, each $f_n$ can be extended to an entire function on $\mathbb{C}$.)

11.101 Suppose that $r_n$ are real numbers and $\sum r_n$ converges but not absolutely. Show that for all $p \in \mathbb{R}$ there is some permutation $\sigma$ of $\mathbb{N}$ such that $\sum r_{\sigma(n)} = p$. (Hint: the sum of the positive $r_n$'s is $\infty$, and of the negative $r_n$'s is $-\infty$. Add the next positive elements until we exceed $p$; then add negative elements until we pass below $p$; repeat.)

The *Lévy-Steinitz* rearrangement theorem implies that if a series $\sum z_n$ of complex numbers converges but not absolutely, then the collection of sums of rearrangements of $\sum z_n$ is either a line in $\mathbb{C}$ or all of $\mathbb{C}$; it is harder to prove.

11.102 (a) Show that $\sum n z^n$ has radius of convergence 1, but converges for no $z$ on the unit circle. (b) Show that $\sum z^n/n^2$ has radius of convergence 1, and converges for all $z$ on the unit circle.

11.103

(a) Let $a_0, \ldots, a_n$ and $b_0, \ldots, b_{n+1}$ be complex numbers. Show that

$$\sum_{k=0}^{n} a_k(b_{k+1} - b_k) + \sum_{k=1}^{n} b_k(a_k - a_{k-1}) = a_n b_{n+1} - a_0 b_0.$$

(b) Show that if $z \in S$ but $z \neq 1$, then for any $n$, $|\sum_{k=0}^{n} z^n| \leqslant 2/|1 - z|$.

(c) Let $\langle a_n \rangle$ be a non-increasing sequence of non-negative real numbers, converging to 0. Show that $\sum a_n z^n$ converges for all $z$ on the unit circle, expect possibly for $z = 1$. (Apply this to $\sum z^n/n$.)

# Riemann Surfaces

# 12

A Riemann surface is a connected 2-dimensional manifold on which we can do complex analysis, at least locally. In this chapter we finally define this notion formally and study its basic properties.

Having defined Riemann surfaces, we first study meromorphic functions, which are holomorphic (read: locally analytic) functions to the Riemann sphere $\mathbb{P}^1(\mathbb{C})$. On the complex plane, this is an extension of the notion of analytic functions. A meromorphic function is allowed to include the point at infinity as a value, which means that meromorphic functions are allowed to have singularities, called *poles*. In turn, this notion can be used to learn new facts about analytic functions: the open mapping theorem says that all analytic functions are open.

The subject of Riemann surfaces is extensive, and we have space for only a few morsels. The rest of this chapter is a kind of potpourri of topics which we will use in the third part of the book. We study compact Riemann surfaces, and show that a holomorphic function from the Riemann sphere to a complex torus is constant; this will be used in the proof of the isomorphism theorem, using the identification of a line in $\mathbb{P}^2(\mathbb{C})$ with the Riemann sphere. We introduce the Riemann surface for the logarithm, and surfaces for the $n$th root functions. These allow us to define uni-valued versions of these functions, and will be used in Chap. 15, when giving a new interpretation of intersection numbers of curves. The process of analytic continuation, extending holomorphic functions to larger domains, will be used in the same chapter. Finally, we study differential forms on Riemann surfaces; we will use them to prove the inversion theorem, which says that every elliptic curve is a complex torus.



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_12

## 12.1 Holomorphic Surfaces

The following definitions mimic Definition 9.61 and Proposition 9.63: a manifold is differentiable if all transition maps are smooth, and a map between differentiable manifolds is smooth if all (equivalently, a cover of) coordinate representations are smooth. Since we did not discuss multivariable complex analysis, we restrict ourselves to one complex dimension.

**Definition 12.1** A *holomorphic surface* is a 2-manifold $X$ for which every transition function (thought of as a function from a subset of $\mathbb{C}$ to $\mathbb{C}$) is analytic.

A *Riemann surface* is a connected holomorphic surface.

Recall that a *coordinate representation* of a function $f\colon X \to Y$ (where $X$ and $Y$ are manifolds) is a map of the form $\varphi \circ f \circ \psi^{-1}$ where $\varphi$ is a chart for $Y$ and $\psi$ is a chart for $X$.

**Proposition 12.2** *Let $X$ and $Y$ be holomorphic surfaces. The following are equivalent for a continuous function $f\colon X \to Y$:*

*(1) Every coordinate representation of $f$ is analytic.*
*(2) For every point $p \in X$ there is a chart $\psi$ for $X$ and a chart $\varphi$ for $Y$ such that $p \in \operatorname{dom}\psi$, $f(p) \in \operatorname{dom}\varphi$, and the coordinate representation $\varphi \circ f \circ \psi^{-1}$ is analytic.*

The proof of Proposition 12.2 is identical to the proof of Proposition 9.63; we just need to know that the composition of analytic functions is analytic (Proposition 11.7). We call a function satisfying the conditions of Proposition 12.2 *holomorphic*. The composition of holomorphic maps is holomorphic.

*Example 12.3* Every open subset $U$ of $\mathbb{C}$ is a holomorphic surface (with the identity map being the only chart). A function $f\colon U \to \mathbb{C}$ is holomorphic if and only if it is analytic. «

*Example 12.4* The transition map between the two charts for the Riemann sphere $\mathbb{P}^1(\mathbb{C})$ is $z \mapsto 1/z$ (Example 9.62); hence the Riemann sphere is a holomorphic surface (indeed a Riemann surface). «

*Example 12.5* The other main example is a torus. We use the notation of Sect. 8.4. Let $\Gamma$ be a lattice in $\mathbb{C}$ (a 2-dimensional discrete subgroup), and recall that $T_\Gamma = \mathbb{C}/\Gamma$ and $\pi_\Gamma\colon \mathbb{C} \to T_\Gamma$ is the quotient map. We showed (Proposition 8.107) that $T_\Gamma$ is a 2-manifold; the charts are functions of the form $(\pi_\Gamma{\restriction_U})^{-1}$, where $U \subset \mathbb{C}$ is a small open ball. We have observed (see the proof of Proposition 8.107) that a transition map $(\pi_\Gamma{\restriction_U})^{-1} \circ (\pi_\Gamma{\restriction_V})$ is the restriction to an open subset of $V$ of the

map $z \mapsto z + a$ where $a$ is some element of $\Gamma$. This map is analytic and so $T_\Gamma$ is a Riemann surface.[1] «

*Exercise 12.6* Show that the quotient map $\pi_\Gamma : \mathbb{C} \to T_\Gamma$ is holomorphic. «

*Exercise 12.7* Show that every change of coordinates of $\mathbb{P}^1(\mathbb{C})$ is holomorphic. (See Exercise 4.75.) «

Isomorphism of holomorphic surfaces is called a *biholomorphism*. Two holomorphic surfaces $X$ and $Y$ are *biholomorphic* if there is a function $f : X \to Y$ which is holomorphic, one to one, and onto, and such that $f^{-1}$ is holomorphic; compare to Definition 8.54. However, we will later see that unlike the continuous case and the smooth case (see Example 8.53), in the analytic category the extra requirement on $f^{-1}$ comes for free; see Proposition 12.37.

*Remark 12.8* Up until now we talked about *the* torus. This is because for any two lattices $\Gamma$ and $\Gamma'$, $T_\Gamma$ and $T_{\Gamma'}$ are homeomorphic, indeed diffeomorphic—we saw that they are both diffeomorphic to $S \times S$. However, it is *not* the case that $T_\Gamma$ and $T_{\Gamma'}$ are always biholomorphic, indeed there are infinitely many (in fact uncountably many) pairwise non-biholomorphic complex tori.[2] «

Holomorphic functions inherit rigidity properties of analytic ones. Here are two examples.

**Proposition 12.9** *If $X$ is a Riemann surface, $Y$ is a holomorphic surface, $f, g : X \to Y$ are holomorphic maps and $\{x \in X : f(x) = g(x)\}$ is not discrete, then $f = g$.*

***Proof*** Suppose that $f(x) = g(x)$ and that $x$ is not isolated in the set of points on which $f = g$. Let $\psi$ be a chart for $X$ such that $x \in \operatorname{dom}\psi$ and let $\varphi$ be a chart for $Y$ such that $f(x) \in \operatorname{dom}\varphi$. Then $\widetilde{f} = \varphi \circ f \circ \psi^{-1}$ and $\widetilde{g} = \varphi \circ g \circ \psi^{-1}$ are coordinate representations of $f$ and $g$, hence are analytic, and $\psi(x)$ is a non-isolated zero of $\widetilde{g} - \widetilde{f}$. By Lemma 11.49, $\widetilde{f} = \widetilde{g}$ on a neighbourhood of $\psi(x)$, which implies that $f = g$ on a neighbourhood of $x$. Now we repeat the argument of Proposition 11.50: the set of non-isolated points of $\{x : f(x) = g(x)\}$ is both closed and open; we are assuming that $X$ is connected. □

[1] Functions on the torus are those induced by *doubly periodic* functions on $\mathbb{C}$, and will be studied in Sect. 14.1.

[2] This is because any linear map taking $\Gamma$ to $\Gamma'$ may fail the Cauchy-Riemann equations, and so fail to be analytic. For more see Exercise 14.42.

Recall that if $f\colon X \to Y$ and $y \in Y$, then we let $f^{-1}[y]$ denote the collection of all $x \in X$ which $f$ maps to $y$. By considering a constant function $g$ we get:

**Corollary 12.10** *Suppose that $X$ is a Riemann surface, $Y$ is a holomorphic surface, and $f\colon X \to Y$ is holomorphic. If for some $y \in Y$, $f^{-1}[y]$ is not discrete, then $f$ is constant.*

The second example is:

**Proposition 12.11** *Let $f\colon X \to Y$ be a continuous map between holomorphic surfaces; suppose that $F \subset X$ is closed and discrete, and that $f \restriction_{X\setminus F}$ is holomorphic. Then $f$ is holomorphic.*

***Proof*** Apply Corollary 11.74 to coordinate representations. □

### 12.1.1 Meromorphic Functions

We have already made use (in the proof of Lemma 11.49) of the following fact: if $f$ is analytic on an open neighbourhood of $a \in \mathbb{C}$ then either $f$ is constant 0 on a neighbourhood of $a$, or there is some $m \geqslant 0$ and some analytic function $g$, defined on an open neighbourhood $W$ of $a$, such that $f = (z-a)^m g$ on $W$ and $g(a) \neq 0$. In fact, Taylor's formula shows that $m$ is the least such that $f^{(m)}(a) \neq 0$. We can extend this definition to negative orders.

*Remark 12.12* In this section, we will assume, without repeatedly stating it, that the domains of functions we consider are open and connected. Thus, if $f$ is analytic and nonconstant, then it is nonconstant on every open subset of the domain. «

A *punctured neighbourhood* of a point $a$ is a set of the form $U \setminus \{a\}$, where $U$ is a neighbourhood of $a$.

**Definition 12.13** Let $f$, defined on a punctured neighbourhood of $a$, be analytic and not constant 0. Let $m \in \mathbb{Z}$. We say that the *order of $f$ at $a$* is $m \in \mathbb{Z}$ if there is a neighbourhood $W$ of $a$ and an analytic function $g\colon W \to \mathbb{C}$ such that $g(a) \neq 0$ and $f = (z-a)^m g$ on $W \setminus \{a\}$. We write $\operatorname{ord}_a(f) = m$.

Note that there is at most one such $m$; if $(z-a)^{m_1} g_1 = (z-a)^{m_2} g_2$ on a punctured neighbourhood of $a$, with $m_2 > m_1$ and $g_1, g_2$ analytic, then by continuity, we will necessarily have $g_1(a) = 0$. It is possible that a function $f$, analytic and nonconstant, does not have any order at $a$. We then call $a$ an *essential singularity of $f$*; see Exercises 12.93 and 12.94.

*Remark 12.14* In Definition 12.13, by shrinking $W$, we may assume that $g \neq 0$ on $W$. Thus, if $f$ has some order at $a$, then there is a punctured neighbourhood of $a$ on which $f$ is nonzero, i.e., $1/f$ is defined on a punctured neighbourhood of $f$. By considering $1/g$, we see that $\operatorname{ord}_a(1/f) = -\operatorname{ord}_a(f)$. «

**Proposition 12.15** *Let $U$ be a neighbourhood of $a$, and suppose that $f : U \setminus \{a\} \to \mathbb{C}$ is analytic and not constant 0. Then $\operatorname{ord}_a(f) \geqslant 0$ if and only $f$ can be extended to an analytic function on $U$.*

***Proof*** One direction follows from the proof of Lemma 11.49 mentioned above. In the other direction, if $f$ is analytic on a punctured neighbourhood $U \setminus \{a\}$ of $a$ and $\operatorname{ord}_a(f) \geqslant 0$, then $\lim_{z\to a} f(z)$ exists, and so by Proposition 11.73, $f$ can be extended to an analytic function on $U$. □

We can do a little better:

**Proposition 12.16** *Suppose that $f$ is analytic and bounded on a punctured neighbourhood $U \setminus \{a\}$ of $a$. Then $f$ can be extended to an analytic function on $U$.*

***Proof*** We may assume that $f$ is nonconstant. The function $(z-a)\cdot f$ is analytic on $U \setminus \{a\}$ and $\lim_{z\to a}(z-a)f(z) = 0$, so $(z-a)f$ can be extended to an analytic function on $U$. The value of this extension at $a$ is 0, so has order $> 0$ at $a$; we can therefore divide by $z-a$ and still get an analytic function. □

It follows that if $\operatorname{ord}_a(f) < 0$ then $f$ is unbounded on any punctured neighbourhood of $a$. In this situation we would still like to extend $f$ to $a$, by letting $f(a) =$ "$\infty$". The notion of Riemann surfaces will allow us to make this meaningful. For $\infty$ we take the point $(0:1) \in \mathbb{P}^1(\mathbb{C})$, the "point at infinity" (under the identification of $z \in \mathbb{C}$ with $(1:z)$); we denote this point by $p_\infty$.

**Definition 12.17** Let $X$ be a Riemann surface. A *meromorphic function* on $X$ is a holomorphic function $f : X \to \mathbb{P}^1(\mathbb{C})$ which is not constant $p_\infty$.

The identification of $z \in \mathbb{C}$ with $\rho_0(z) = (1:z)$ allows us to consider any function to $\mathbb{C}$ also as a function to $\mathbb{P}^1(\mathbb{C})$. The inverse of $\rho_0$ is a chart for $\mathbb{P}^1(\mathbb{C})$. Thus, a function $f : U \to \mathbb{C}$ is analytic if and only if it is holomorphic when thought of as a function to $\mathbb{P}^1(\mathbb{C})$. The new ingredient is that with meromorphic functions we are allowed to take the value $p_\infty$.

**Lemma 12.18** *Let $U$ be a neighbourhood of $a$, and suppose that $f : U\setminus\{a\} \to \mathbb{C}$ is analytic. Then $\operatorname{ord}_a(f) < 0$ if and only if extending $f$ to $U$ by defining $f(a) = p_\infty$ gives a meromorphic function on $U$.*

***Proof*** By Remark 12.14, $\mathrm{ord}_a(f) < 0$ if and only if $\mathrm{ord}_a(1/f) > 0$; by Proposition 12.15, this is equivalent to $1/f$ being extendible to an analytic function on $U$, by mapping $a$ to 0. Consider the chart $\psi(z:1) = z$ for the Riemann sphere (i.e. $\psi = (\rho_1)^{-1}$); then $\psi \circ f = 1/f$ (defined wherever $f \neq 0$). Extending $1/f$ by sending $a$ to 0 is equivalent to extending $f$ by mapping $a$ to $p_\infty$. □

**Corollary 12.19** *Suppose that $U \subseteq \mathbb{C}$ is a region, and that $Z \subset U$ is discrete and closed in $U$. Suppose that $f : U \setminus Z \to \mathbb{C}$ is analytic and not constant 0. Then $f$ can be extended to a meromorphic function $f : U \to \mathbb{P}^1(\mathbb{C})$ if and only if $f$ has an order at every $a \in Z$.*

***Proof*** Working with an open cover, we may assume that $Z$ contains a single point $a$; we then apply Proposition 12.15 and Lemma 12.18. □

If $\mathrm{ord}_a(f) = m > 0$, then (after extending if necessary), $f(a) = 0$, and we say that $a$ is a *zero* of $f$. We sometimes think of $p_\infty$ as the "north pole" of the sphere, and so if $\mathrm{ord}_a(f) = m < 0$ then we say that $a$ is a *pole of order* $-m$ of $f$. Thus $a$ is a zero of order $m$ of $f$ if and only if it is a pole of order $m$ of $1/f$.

*Remark 12.20* If $f$ is a polynomial, then $\mathrm{ord}_a(f)$ is the multiplicity of $a$ as a root of $f$. We will extend this terminology and think of the order $\mathrm{ord}_a(f)$ as some kind of "multiplicity" of zero of $f$; except that we now consider poles as well, and so allow negative multiplicities. A multiset that allows negative multiplicities is called a *divisor*. «

### The Meromorphic Conjugate

Let $\alpha(a:b) = (b:a)$; this is a change of coordinates of the Riemann sphere, and so is a bi-holomorphism from $\mathbb{P}^1(\mathbb{C})$ to itself (Exercise 12.7). Under the identification of $z$ with $(1:z)$, this maps $z \neq 0$ to $1/z$ and exchanges 0 and $p_\infty$. Since the composition of holomorphic functions is holomorphic, for any meromorphic $f : X \to \mathbb{P}^1(\mathbb{C})$ which is not constant 0, we call $\alpha \circ f$ the *meromorphic conjugate* of $f$, and write $1/f$ instead of $\alpha \circ f$. Since $\alpha \circ \alpha$ is the identity, $1/(1/f) = f$.

Suppose that $U \subseteq \mathbb{C}$ is a region. Suppose that $f$ and $g$ are meromorphic on $U$. Let $Z$ be the collection of poles of $f$ and of $g$. By Corollary 12.10, $Z$ is discrete and closed in $U$. The functions $f + g$, $f \cdot g$ and $f'$ are defined on $U \setminus Z$, and are either constant 0, or have an order at every $a \in Z$:

*Exercise 12.21* Suppose that $f$ and $g$ are meromorphic on $V$ and analytic on $V \setminus \{a\}$. Show that: (a) If $fg$ is not constant 0, then $\mathrm{ord}_a(fg) = \mathrm{ord}_a(f) + \mathrm{ord}_a(g)$. (b) If $f+g$ is not constant 0, then $\mathrm{ord}_a(f+g) = \min\{\mathrm{ord}_a(f), \mathrm{ord}_a(g)\}$, except possibly when $\mathrm{ord}_a(f) = \mathrm{ord}_a(g)$ and the "leading coefficients cancel out", in which case $\mathrm{ord}_a(f+g)$ could be any $m \geqslant \mathrm{ord}_a(f)$. (c) If $f'$ is not constant 0, then $\mathrm{ord}_a(f') = \mathrm{ord}_a(f) - 1$, except when $\mathrm{ord}_a(f) = 0$, in which case $\mathrm{ord}_a(f')$ could be any non-negative number. «

We thus can extend $f+g$, $fg$ and $f'$ to meromorphic functions on $U$.

*Exercise 12.22* Let $U$ be a region. Show that the collection of meromorphic functions on $U$, equipped with these operations of addition and multiplication, is a field. «

## 12.2 The Open Mapping Theorem

Our aim now is to show that if $a$ is a zero of order $m$ of $f$, then around $a$, $f$ behaves like $z \mapsto z^m$. A main consequence is the open mapping theorem. We also get a strengthening of the inverse function theorem.

### 12.2.1 The Calculus of Residues

If $f$ is analytic on $U \setminus \{a\}$ (where $U$ is an open neighbourhood of $a$), $\operatorname{ord}_a(f) = m$, and $B(a,r) \subset U$, then on $B(a,r) \setminus \{a\}$,

$$f(z) = c_m(z-a)^m + c_{m+1}(z-a)^{m+1} + \dots$$

with $c_m \neq 0$; this is called a *Laurent series expansion* of $f$ around $a$. Note that $m$ is the order of the formal Laurent series $\sum c_n x^n$ (see Exercise 2.36).

If $m \leqslant -1$, then the coefficient $c_{-1}$ is called the *residue* of $f$ at $a$, and we denote it by $\operatorname{rsd}_a(f)$. If $m \geqslant 0$ then the residue is 0.

*Remark 12.23* The usual notation for the residue is $\operatorname{res}_a(f)$, but we already used this for the resultant. «

For the following lemma, recall that for a loop $\gamma$, $\operatorname{wnd}_\gamma(p)$ denotes the winding number of $\gamma$ around $p$, and that $\gamma$ is a contour around $p$ if $\operatorname{wnd}_\gamma(p) = 1$.

**Lemma 12.24** *Suppose that $W$ is simply connected, $f$ is meromorphic on $W$, and $\gamma$ is a loop in $W$. Then $\gamma$ is a contour around only finitely many poles of $f$.*

***Proof*** Let

$$Q = \operatorname{range}\gamma \cup \left\{p \,:\, \operatorname{wnd}_\gamma(p) \neq 0\right\}.$$

Then $Q$ is bounded (Corollary 11.23) and closed (Proposition 11.24), and so compact. It is contained in $W$ (Proposition 11.22). The set $Z$ of poles of $f$ is closed in $W$. And so $Z \cap Q$ is a closed subset of $Q$, and so is compact. Now $Z$ is discrete (Corollary 12.10) so $Z \cap Q$ is discrete and compact, whence it is finite (Proposition 8.101). □

**Proposition 12.25** *Let $W$ be simply connected. Suppose that $f$ is meromorphic on $W$. Suppose that $\gamma$ is a loop in $W$ which is a contour around every pole of $f$. Then*

$$\int_\gamma f\,dz = 2\pi i \sum_{p\in P} \mathrm{rsd}_p(f)$$

*(where $P$ is the set of poles of $f$).*

***Proof*** For each $p \in P$ let $\sum_{n\geqslant \mathrm{ord}_p(f)} c_n(p)(z-p)^n$ be the Laurent expansion of $f$ around $p$; let

$$g_p = \sum_{n=\mathrm{ord}_p(f)}^{-1} c_n(p)(z-p)^n$$

be the *principal part* of $f$ around $p$. For $k \neq -1$, the function $(z-p)^k$ has a primitive on $W \setminus \{p\}$, namely $(z-p)^{k+1}/(k+1)$. By Lemma 11.16,

$$\int_\gamma c_k(p)(z-p)^k\,dz = 0.$$

On the other hand, by Proposition 11.21,

$$\int_\gamma c_{-1}(p)(z-p)^{-1}\,dz = c_{-1}(p)\cdot 2\pi i = \mathrm{rsd}_p(f)\cdot 2\pi i.$$

Let $g = \sum_{p\in P} g_p$, which is well-defined since $P$ is finite (Lemma 12.24). Now $\mathrm{ord}_p(f-g) \geqslant 0$ for each $p \in P$, so $f-g$ can be extended to an analytic function on $W$. As $W$ is simply connected, $\int_\gamma (f-g)\,dz = 0$ (Lemma 11.16 and Proposition 11.19), whence

$$\int_\gamma f\,dz = \int_\gamma g\,dz = \sum_{p\in P}\int_\gamma g_p\,dz = 2\pi i \sum_{p\in P}\mathrm{rsd}_p(f). \qquad \square$$

*Remark 12.26* If $f$ is meromorphic on $U$ and $\gamma$ is a parameterised circle in $U$ which avoids the poles of $f$ then as the distance between the image of $\gamma$ and the set of poles of $f$ is positive, we can let $W$ be the interior of a circle of slightly larger radius; then the poles of $f$ in $W$ are in the interior of $\gamma$. «

**Lemma 12.27** *Let $f$ be meromorphic on $U$ and not constant 0. For every $a \in U$,*

$$\operatorname{ord}_a(f) = \operatorname{rsd}_a\left(\frac{f'}{f}\right).$$

*Indeed, if* $\operatorname{ord}_a(f) \neq 0$ *then* $\operatorname{ord}_a(f'/f) = -1$*; otherwise* $\operatorname{ord}_a(f'/f) = 0$.

(By Definition 12.21, $f'/f$ is indeed meromorphic on $U$.)

***Proof*** Let $m = \operatorname{ord}_a(f)$. Let $g$ be analytic on a neighbourhood $W$ of $a$ such that $g \neq 0$ on $W$ and $f = (z-a)^m g$ on $W$. Then on $W \setminus \{a\}$,

$$f' = m(z-a)^{m-1}g + (z-a)^m g',$$

and so on $W \setminus \{a\}$,

$$\frac{f'}{f} = \frac{m}{z-a} + \frac{g'}{g}.$$

Since $g \neq 0$ on $W$ and analytic on $W$, $g'/g$ is analytic on $W$, and so the previous equation gives us the Laurent expansion of $f'/f$ around $a$; we see that $\operatorname{rsd}_a(f'/f) = m$. □

Recall that if $f$ is meromorphic on $U$ (and not constant 0) then we consider the collection of zeros and poles of $f$ as a divisor, with multiplicity given by order. The size of a finite divisor is the sum of the multiplicities of its elements. The following corollary, which follows directly from Proposition 12.25 and Lemma 12.27, shows that $\int_\gamma f'/f\,dz$ gives the size of the divisor restricted to the interior of $\gamma$.

**Corollary 12.28** *Let $W$ be simply connected, let $f$ be meromorphic on $W$ and not constant 0, and let $\gamma$ be a loop in $W$ which is a contour around all the zeros and poles of $f$. Then*

$$\int_\gamma \frac{f'}{f}\,dz = 2\pi i \sum_{a \in W} \operatorname{ord}_a(f).$$

(As discussed above, this sum is finite since $f$ has only finitely many poles and zeros around which $\gamma$ is a contour.)

### 12.2.2 The Continuity of Roots of Polynomials

We give an application. We know that the coefficients of a polynomial vary continuously with its roots: if the roots of a polynomial $f \in \mathbb{C}[x]$ of degree $d$

are $c_1, c_2, \ldots, c_d$ (repetitions allowed), then $f = a(x-c_1)(x-c_2)\cdots(x-c_d)$ for some constant $a$, so the coefficients of $f$ can be expressed as continuous functions of the roots. The following proposition says that the roots of polynomials vary continuously with their coefficients.

**Proposition 12.29** *Let $p = a_d x^d + a_{d-1}x^{d-1} + \cdots + a_1 x + a_0 \in \mathbb{C}[x]$ be a nonzero polynomial, and let $c \in \mathbb{C}$ be a root of $p$ of multiplicity $k$. Then for every neighbourhood $U$ of $c$ in $\mathbb{C}$ there is a neighbourhood $\hat{U} \subseteq U$ of $c$ and some $\delta > 0$ such that for every $\boldsymbol{b} = (b_d, b_{d-1}, \ldots, b_0) \in B(\boldsymbol{a}, \delta)$, the polynomial $b_d x^d + \cdots + b_1 x + b_0$ has precisely $k$ roots in $\hat{U}$ (counting multiplicities).*

***Proof*** Given a neighbourhood $U$ of $c$, let $\varepsilon$ be sufficiently small so that $B(c,\varepsilon) \subseteq U$, and $c$ is the only root of $p$ in $\overline{B}(c,\varepsilon)$. Let $\gamma$ be the parameterised circle around $c$ of radius $\varepsilon$; so $|p(z)| > 0$ for all $z$ in the range of $\gamma$, whence by compactness, there is some $\eta > 0$ such that $|p(z)| \geqslant 2\eta$ for all $z \in \operatorname{range}\gamma$.

For $\boldsymbol{b} = (b_d, \ldots, b_0) \in \mathbb{C}^{d+1}$, let $p_{\boldsymbol{b}}$ be the polynomial $b_d x^d + \cdots + b_0$ in $\mathbb{C}[x]$. For $\boldsymbol{w} \in \mathbb{C}^{d+1}$ and $z \in \mathbb{C}$ let $T(\boldsymbol{w}, z) = p_{\boldsymbol{w}}(z)$. The map $T\colon \mathbb{C}^{d+2} \to \mathbb{C}$ is continuous. The set $\overline{B}(\boldsymbol{a}, 1) \times \operatorname{range}\gamma$ is compact, and so $T$ is uniformly continuous on that set. It follows that there is some $\delta > 0$ such that for all $\boldsymbol{b} \in B(\boldsymbol{a}, \delta)$, $|p_{\boldsymbol{b}}(z)| > \eta$ for all $z \in \operatorname{range}\gamma$. What is important for us is that $p_{\boldsymbol{b}}(z) \neq 0$, so we can make the following definition: for all $\boldsymbol{b} \in B(\boldsymbol{a}, \delta)$, let

$$N(\boldsymbol{b}) = \frac{1}{2\pi i}\int_\gamma \frac{p'_{\boldsymbol{b}}}{p_{\boldsymbol{b}}}\,dz.$$

By Corollary 12.28, $N(\boldsymbol{b})$ is the number of zeros of $p_{\boldsymbol{b}}$ in $\hat{U} = B(c,\varepsilon)$, multiplicities counted (note that $p_{\boldsymbol{b}}$ has no poles as it is a polynomial.) In particular, $N(\boldsymbol{a}) = k$. We argue however that $\boldsymbol{b} \mapsto N(\boldsymbol{b})$ is continuous on $B(\boldsymbol{a}, \delta)$. This follows from the uniform continuity of $(\boldsymbol{w}, z) \mapsto (p'_{\boldsymbol{w}}/p_{\boldsymbol{w}})(z)$ on $\overline{B}(\boldsymbol{a}, \delta) \times \operatorname{range}\gamma$, and then using Exercise 11.14. However the range of $N$ is discrete, and $B(\boldsymbol{a}, \delta)$ is connected, and so $N(\boldsymbol{b})$ is constant on $B(\boldsymbol{a}, \delta)$. □

For more, see Exercises 12.121–12.125.

### 12.2.3 Open Mappings and Inverse Functions

Suppose that $f$ is analytic and that $a \in \operatorname{dom} f$. Let $m \geqslant 1$. Let $U \subseteq \operatorname{dom} f$ be an open neighbourhood of $a$. We say that $f$ is *$m$-to-1 on $U$ around $a$* if:

(i) $f[U]$ is open;
(ii) $a$ is the unique $f$-preimage of $f(a)$ in $U$ (that is, $a$ is the only point in $U$ which $f$ maps to $f(a)$); and
(iii) every $b \in f[U]$ other than $f(a)$ has precisely $m$ many $f$-preimages in $U$.

We say that $f$ is $m$-to-1 *arbitrarily close to* $a$ if for every neighbourhood $\hat{U}$ of $a$ there is an open neighbourhood $U \subseteq \hat{U}$ of $a$ such that $f$ is $m$-to-1 on $U$ around $a$.

*Example 12.30* The standard example is the function $z^m$, which is $m$-to-1 around 0 on $B(0, r)$ for each $r > 0$. «

*Remark 12.31* Suppose that $f$ is $m$-to-1 arbitrarily close to $a$. Let $\hat{U}$ be a neighbourhood of $a$ and let $\hat{V}$ be a neighbourhood of $f(a)$. Then there is an open neighbourhood $U \subseteq \hat{U}$ of $a$ such that $f[U] \subseteq \hat{V}$ and $f$ is $m$-to-1 on $U$ around $a$. For we can replace $\hat{U}$ by $\hat{U} \cap f^{-1}[\hat{V}]$. «

**Proposition 12.32** *Let $f$ be analytic and nonconstant on a neighbourhood of $a$. Suppose that $a$ is a zero of order $m$ of $f$. Then $f$ is $m$-to-1 arbitrarily close to $a$.*

***Proof*** We know that the set of zeros of $f$ is discrete (Proposition 11.50). If $f'$ is constant then as $f$ is nonconstant, $f' \neq 0$ near $a$ (Proposition 11.8). Otherwise, we know that the set of zeros of $f'$ on a neighbourhood of $a$ is discrete.

Thus, if $r$ is small then $\overline{B}(a, r) \subset \operatorname{dom} f$ and $f$ and $f'$ have no zeros in $\overline{B}(a, r)$ other than $a$. Fix such small $r$. Let $\gamma$ be the parameterised circle of radius $r$ centred at $a$. Let $\delta = \min\{|f(z)| : z \in \operatorname{range}\gamma\}$, which is positive; let $M = \max\{|f'(z)| : z \in \operatorname{range}\gamma\}$, which is finite.

We will pick some small $\varepsilon \leqslant \delta/2$. Let $b \in B(0, \varepsilon)$. Then for all $z \in \operatorname{range}\gamma$, as $|f(z) - b| \geqslant \delta/2$,

$$|f'(z)| \cdot \left| \frac{1}{f(z)} - \frac{1}{f(z) - b} \right| = |f'(z)| \frac{|b|}{|f(z)| \cdot |f(z) - b|} \leqslant M \frac{2\varepsilon}{\delta^2},$$

and so

$$\left| \int_\gamma \frac{f'}{f}\, dz - \int_\gamma \frac{f'}{f - b}\, dz \right| \leqslant 2\pi r \cdot M \frac{2\varepsilon}{\delta^2}.$$

So we pick $\varepsilon$ small enough so that

$$\left| \int_\gamma \frac{f'}{f}\, dz - \int_\gamma \frac{f'}{f - b}\, dz \right| < 2\pi. \tag{12.1}$$

By Corollary 12.28 (and Remark 12.26),

$$\int_\gamma \frac{f'}{f}\, dz = 2\pi i \cdot m.$$

and similarly, as $f' = (f - b)'$,

$$\int_\gamma \frac{f'}{f-b}\,dz = 2\pi i \cdot k,$$

where $k$ is the number of zeros of $f - b$ in $B(a, r)$, with multiplicities (orders) counted. But since $f'(w) \neq 0$ for all $w \in B(a, r)$ other than possibly $a$, if $b \neq 0$ then $\mathrm{ord}_w(f - b) = 1$ for every zero $w$ of $f - b$ in $B(a, r)$. By Eq. 12.1, $k = m$. So every $b \in B(0, \varepsilon)$ has precisely $m$ many $f$-preimages in $B(a, r)$. We can therefore let $U = B(a, r) \cap f^{-1}[B(0, \varepsilon)]$; $f$ is $m$-to-1 on $U$ around $a$ (note that $f[U] = B(0, \varepsilon)$ is open). And $r$ can be chosen as small as we like. □

**Open Mapping Theorem** *A nonconstant analytic function is open: for all open $U \subseteq \mathrm{dom}\, f$, $f[U]$ is open.*

***Proof*** It suffices to show that for all $a \in \mathrm{dom}\, f$ and every open neighbourhood $W \subseteq \mathrm{dom}\, f$ of $a$, $f[W]$ is a neighbourhood of $f(a)$. But this follows from Proposition 12.32, applied to the function $f - f(a)$. □

We can also conclude that the "bad example" from real analysis (see Remark 9.58) cannot happen in the complex context. If $f$ is analytic and $f'(a) = 0$ then for $m = \mathrm{ord}_a(f - f(a))$, which is greater than 1, $f$ is $m$-to-1 on arbitrarily small neighbourhoods of $a$. Hence:

**Lemma 12.33** *Suppose that $f : U \to \mathbb{C}$ is analytic and 1–1. Then $f'(a) \neq 0$ for all $a \in U$.*

Together with the complex inverse function theorem (Theorem 11.11) we get a strong form of the inverse function theorem (see Exercise 9.57):

**Theorem 12.34 (Analytic Inverse Function Theorem)** *Let $f$ be analytic and 1–1 on $U$. Then $f^{-1}$ is analytic, and for all $b \in f[U]$, $(f^{-1})'(b) = 1/f'(f^{-1}(b))$.*

*Remark 12.35* In fact, with our new tools we can deduce part of Theorem 11.11 without appealing to the full real Inverse Function Theorem. Let $f$ be analytic, let $a \in \mathrm{dom}\, f$, and suppose that $f'(a) \neq 0$. Proposition 12.32 implies that there is a neighbourhood $U \subseteq \mathrm{dom}\, f$ of $a$ on which $f$ is 1-to-1. Since $f \restriction_U$ is open and continuous, it is a homeomorphism between $U$ and an open set $V \subseteq \mathbb{C}$. We then just need to check that the inverse is differentiable, which is not too difficult. «

*Example 12.36* Continuing Example 11.12, the derivative of $z \mapsto z^3$ at 0 is 0, and so unlike the real case, we cannot continuously choose a complex cube root on any neighbourhood of 0. «

### Consequences for Riemann Surfaces

As promised, a holomorphic bijection is a biholomorphism:

**Proposition 12.37** *Let $X$ and $Y$ be holomorphic surfaces, and let $f: X \to Y$ be holomorphic, one to one and onto. Then $f^{-1}$ is holomorphic.*

***Proof*** Let $\psi$ be a chart for $X$ and $\varphi$ be a chart for $Y$. The map $\varphi \circ f \circ \psi^{-1}$ is analytic and one-to-one from an open subset $U$ of the range of $\psi$ to the range of $\varphi$. The analytic inverse function theorem, 12.34, tells us that $(\varphi \circ f \circ \psi^{-1})^{-1}$ is analytic; this is a coordinate representation of $f^{-1}$. □

*Exercise 12.38* Show that the unit sphere is a Riemann surface (Example 8.21), which is biholomorphic with $\mathbb{P}^1(\mathbb{C})$ (see Exercise 8.56). «

The open mapping theorem extends to Riemann surfaces:

**Proposition 12.39** *A nonconstant holomorphic map between Riemann surfaces is open.*

***Proof*** Let $f: X \to Y$ be holomorphic and nonconstant. Let $U \subseteq X$ be open; let $y \in f[U]$; we show that $f[U]$ is a neighbourhood of $y$. Let $x$ be an $f$-preimage of $y$ in $U$. Let $\psi$ be a chart for $X$ and $\varphi$ be a chart for $Y$ such that $x \in \operatorname{dom}\psi$ and $y \in \operatorname{dom}\varphi$. The coordinate representation $g = \varphi \circ f \circ \psi^{-1}$ is analytic; by Corollary 12.10, it is nonconstant, so the Open Mapping Theorem tells us that $g$ is open; so $\varphi \circ f[U \cap \operatorname{dom}\psi]$ is an open subset of $\mathbb{C}$; whence $f[U \cap \operatorname{dom}\psi]$ is an open (in $Y$) subset of $f[U]$ containing $y$. □

*Remark 12.40* Suppose that $f$ is a meromorphic function on a Riemann surface $X$. Let $p \in X$, and let $\psi$ and $\varphi$ be two charts for $X$ such that $p \in \operatorname{dom}\psi \cap \operatorname{dom}\varphi$. The transition function $\psi \circ \varphi^{-1}$ is 1–1 and analytic, and so $(\psi \circ \varphi^{-1})'$ is nonzero at every point. It follows that the order $\operatorname{ord}_{\varphi(p)}(f \circ \varphi^{-1})$ equals $\operatorname{ord}_{\psi(p)}(f \circ \psi^{-1})$, and so the common value is unambiguously denoted by $\operatorname{ord}_p(f)$. «

We also extend our analysis that led to the open mapping theorem. Just like analytic maps, we say that $f: X \to Y$ is *$m$-to-1 arbitrarily close to $p \in X$* if every open neighbourhood $\hat{U}$ of $p$ contains an open neighbourhood $U \subseteq \hat{U}$ of $p$ such that $p$ is the only $f$-preimage of $f(p)$ in $U$, and every $y \in f[U]$ other than $f(p)$ has precisely $m$-many $f$-preimages in $U$. (We already know that $f[U]$ is open.) By applying Proposition 12.32 to coordinate representations, we get:

**Proposition 12.41** *If $f: X \to Y$ is a nonconstant holomorphic map between Riemann surfaces then for every $p \in X$ there is some (unique) $m \geqslant 1$ such that $f$ is $m$-to-1 arbitrarily close to $p$.*

We call this number $m$ the *valency* (or *multiplicity*) of $f$ at $p$ (sometimes it is called the multiplicity of $p$ with respect to $f$). The collection of points with valency greater than one is discrete.

*Example 12.42* If $f$ is meromorphic on $X$ and $p$ is either a zero or a pole of order $m$ of $f$, then the valency of $f$ at $p$ is $m$. «

## 12.3 Compact Riemann Surfaces

We survey a number of consequences of the open mapping theorem for compact Riemann surfaces, in particular for a complex torus $T_\Gamma$.

**Proposition 12.43** *Let $X$ and $Y$ be Riemann surfaces, $f: X \to Y$ holomorphic and nonconstant, and suppose that $X$ is compact. Then $f$ is onto $Y$ (and so $Y$ is compact).*

***Proof*** Since $f$ is continuous, $f[X]$ is compact (Proposition 8.64). Since $Y$ is Hausdorff, $f[X]$ is a closed subset of $Y$ (Proposition 8.71). By Proposition 12.39, $f[X]$ is also an open subset of $Y$. Since $Y$ is connected and $f[X]$ is nonempty, we must have $f[X] = Y$. □

As a result:

**Corollary 12.44** *If $X$ is a compact Riemann surface, then every holomorphic function $f: X \to \mathbb{C}$ is constant.*

**Functions to the Torus**

Recall the notion of a *lifting* of a map to a quotient space $\mathbb{R}^n/G$, such as the torus (Definition 9.18). We fix a torus $T = T_\Gamma$ and quotient map $\pi = \pi_\Gamma: \mathbb{C} \to T$. Recall that liftings of smooth maps are smooth (see for example Exercise 9.109); the same holds in the holomorphic category:

**Lemma 12.45** *Let $X$ be a Riemann surface, and let $f: X \to T$ be holomorphic. Then any lifting of $f$ to a map from $X$ to $\mathbb{C}$ is holomorphic.*

***Proof*** Let $F$ be a lifting of $f$. Let $x \in X$; let $\psi$ be a chart for $X$ with $x \in \operatorname{dom}\psi$. Let $V$ be a neighbourhood of $F(x)$ sufficiently small so that $(\pi \restriction_V)^{-1}$ is a chart for $T$. Since $F$ is continuous, by shrinking, we may assume that $F[\operatorname{dom}\psi] \subseteq V$. Then $F \circ \psi^{-1} = (\pi \restriction_V)^{-1} \circ f \circ \psi^{-1}$ is a coordinate representation of both $F$ and $f$; by assumption on $f$, it is analytic. □

**Corollary 12.46** *If $X$ is a simply connected and compact Riemann surface, then any holomorphic $f\colon X \to T$ is constant.*

***Proof*** Let $f\colon X \to T$ be holomorphic. By Theorem 9.24, there is a (continuous) lifting $F\colon X \to \mathbb{C}$ of $f$. By Lemma 12.45, $F$ is holomorphic. By Corollary 12.44, $F$ is constant. Since $f = \pi \circ F$, $f$ is constant as well. □

In Chap. 14, we will apply this to the Riemann sphere, which is both compact (Theorem 8.91) and simply connected (Proposition 9.16).

**Degrees of Maps**

Let $X$ and $Y$ be Riemann surfaces, $f\colon X \to Y$ be holomorphic and nonconstant, and suppose that $X$ is compact. For every $q \in Y$, $f^{-1}[q]$ is a discrete subset (Corollary 12.10) which is closed in $X$, and so is compact, whence it is finite. Let $\deg_q(f)$ be the sum of the valencies of $p$ with respect to $f$ for all $p \in f^{-1}[q]$.

**Lemma 12.47** *For all $q, q' \in Y$, $\deg_q(f) = \deg_{q'}(f)$.*

***Proof*** We show that the map $q \mapsto \deg_q(f)$ is locally constant: every $q \in Y$ has a neighbourhood $U$ on which this map is constant; connectedness of $Y$ gives the desired result. Fix $q \in Y$. Let $p_1, p_2, \ldots, p_k$ list the points which $f$ maps to $q$. For each $i = 1, \ldots, k$, find some open neighbourhood $V_i$ of $p_i$ such that the sets $V_i$ are pairwise disjoint and $f$ is $m_i$-to-1 on $V_i$ around $p_i$, where $m_i$ is the valency of $p_i$ with respect to $f$.

Let $Q = X \setminus \bigcup_{i \leqslant k} V_i$; so $Q$ is a closed, hence compact, subset of $X$, and so the image $f[Q]$ is a closed subset of $Y$, which does not contain $q$. Let $W = Y \setminus f[Q]$; so $W$ is an open neighbourhood of $q$, and $f^{-1}[W] \subseteq \bigcup_{i \leqslant k} V_i$. Since each $f[V_i]$ is open, by shrinking $W$ we may assume that $W \subseteq f[V_i]$ for each $i$. Then for each $i = 1, \ldots, k$, each $q' \in W$ has $m_i$ many $f$-preimages in $V_i$. These are all the $f$-preimages of $q'$, are all distinct, and each has valency 1 (recall that the collection of points of valency $> 1$ is discrete). Hence $\deg_{q'}(f) = \sum_i m_i = \deg_q(f)$. □

**Definition 12.48** Suppose that $X, Y$ are compact Riemann surfaces and $f\colon X \to Y$ is holomorphic and nonconstant. We denote the common value $\deg_q(f)$ by $\deg(f)$, and call it the *degree* of the map $f$.

**Corollary 12.49** *If $X$ is compact then a nonconstant meromorphic function $f$ on $X$ has the same number of poles and zeros (counted with their orders).*

***Proof*** Both are the degree of the map. □

*Remark 12.50* If $f\colon X \to Y$ is a holomorphic nonconstant map between compact Riemann surfaces, then $f$ has degree 1 if and only if it is a biholomorphism. «

## 12.4 Riemann Surfaces for the Logarithm and Roots

Our inability to continuously define an argument function on $\mathbb{C} \setminus \{0\}$ leaves us with an inability to define an analytic logarithm function or an $n$th root function on the same domain. Riemann surfaces help.

### 12.4.1 The Logarithm

We start with the logarithm function. The idea is simple: since there are "$\mathbb{Z}$-many" choices for a logarithm (equivalently, an argument), we create a new surface on which every nonzero complex number has "$\mathbb{Z}$-many" copies. On that surface we can define a uni-valued logarithm function.

Define

$$\Sigma = \{(z, t) \in \mathbb{C} \times \mathbb{R} : z \neq 0 \text{ and } t \text{ is an argument of } z\},$$

where recall that $t$ is an argument of $z$ if $z = |z|e^{it}$. The "covering map" of $\Sigma$ onto $\mathbb{C} \setminus \{0\}$ is the projection onto the first coordinate: we let $\pi_\Sigma(z, t) = z$. We think of the points $(z, t) \in \Sigma$ as distinct "copies" of $z$, each giving its own choice of argument. See Fig. 12.1.

Suppose that $U \subseteq \mathbb{C} \setminus \{0\}$ is open and that $\theta : U \to \mathbb{R}$ is a continuous choice of argument on $U$. Define $\psi_\theta : U \to \Sigma$ by letting $\psi_\theta(z) = (z, \theta(z))$. Then $\psi_\theta$ is 1–1 and its inverse is the restriction of $\pi_\Sigma$ to the range of $\psi_\theta$ (technically, the range of $\psi_\theta$ is the graph of $\theta$). We let $\psi_\theta^{-1}$ be a chart for $\Sigma$, and let $\mathcal{A}_\Sigma$ be the collection of all such charts.

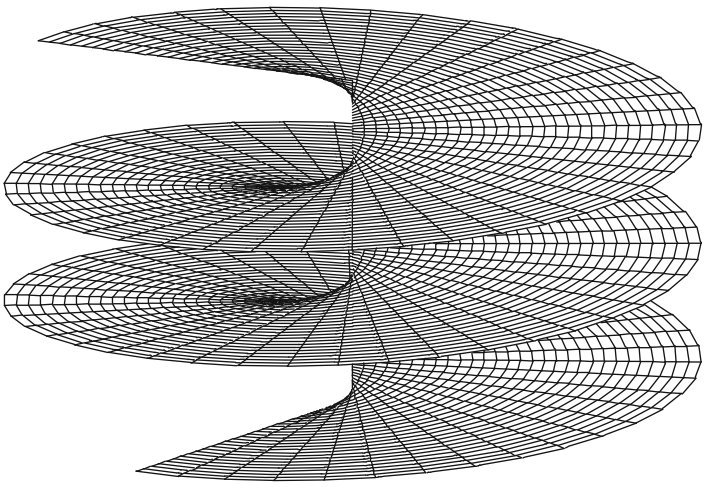


**Fig. 12.1** The Riemann surface for the logarithm. Detail

**Proposition 12.51** *$(\Sigma, \mathcal{A}_\Sigma)$ is a holomorphic surface, and a topological subspace of $\mathbb{C} \times \mathbb{R}$.*

***Proof*** Let $\theta : U \to \mathbb{R}$ be a continuous choice of argument on an open $U$; let $z_0 \in U$; let $t_0 = \theta(z_0)$. Since $\theta$ is continuous, there is some $\varepsilon > 0$ such that $B(z_0, \varepsilon) \subseteq U$ and for all $z \in B(z_0, \varepsilon)$, $|\theta(z) - t_0| < \pi$. Suppose that $(z, t) \in \Sigma$, $|z - z_0| < \varepsilon$ and $|t - t_0| < \pi$. Then $|t - \theta(z)| < 2\pi$, and since both are arguments of $z$, it follows that $t = \theta(z)$. Since $B(0, \varepsilon) \times (t_0 - \pi, t_0 + \pi)$ is an open subset of $\mathbb{C} \times \mathbb{R}$, we see that the range of $\psi_\theta$ is an open subset of $\Sigma$ (when $\Sigma$ is equipped with the subspace topology). Further, since $\theta$ is continuous (and $\pi_\Sigma$ is continuous), we see that $\psi_\theta$ is a homeomorphism from $U$ to its range.

Let $(z_0, t_0) \in \Sigma$. Let $\delta < |z_0|$ be positive; by Proposition 9.30 there is a continuous choice of argument $\theta : B(z_0, \delta) \to \mathbb{R}$, and we can ensure that $\theta(z_0) = t_0$; so $(z_0, t_0) \in \text{range}\, \psi_\theta$. Therefore, by Proposition 8.61, $\mathcal{A}_\Sigma$ is an atlas for $\Sigma$, and $(\Sigma, \mathcal{A}_\Sigma)$ is a 2-manifold which is a topological subspace of $\mathbb{C} \times \mathbb{R}$.

To show that $\Sigma$ is a holomorphic surface, it remains to observe that every transition function is the identity on an open set, and so is analytic. □

*Exercise 12.52* Verify directly from definition that the transition function between two charts in $\mathcal{A}_\Sigma$ is defined on an open set. «

*Exercise 12.53* Show that $\pi_\Sigma : \Sigma \to \mathbb{C}$ is holomorphic. «

**Proposition 12.54** *The function $z \mapsto (e^z, \Im z)$ is a biholomorphism from $\mathbb{C}$ to $\Sigma$.*

(Here $\Im(a + ib) = b$ is the imaginary part of $z = a + ib$).

***Proof*** The main thing to note is that this map is indeed a bijection between $\mathbb{C}$ and $\Sigma$. It is continuous by Proposition 12.51, since it is continuous as a map from $\mathbb{C}$ to $\mathbb{C} \times \mathbb{R}$. A coordinate representation of this map is $z \mapsto e^z$ (restricted to an open set), which is of course analytic. By Proposition 12.37, it is a biholomorphism. □

Let us spell out the inverse of $z \mapsto (e^z, \Im z)$: it is a "global logarithm" on $\Sigma$.

**Proposition 12.55** *The map $(z, t) \mapsto \ln|z| + it$ from $\Sigma$ to $\mathbb{C}$ is holomorphic.*

*Remark 12.56* The surface $\Sigma$ illustrates that to check that a map is holomorphic, we first need to verify that it is continuous. For example, suppose that $\theta$ is a *discontinuous* choice of argument on some open set $U$ (say $\theta(z)$ is the unique argument of $z$ in the interval $[0, 2\pi)$, where $U = \mathbb{C} \setminus \{0\}$). The map $z \mapsto (z, \theta(z))$ from $U$ to $\Sigma$ is discontinuous and so cannot be holomorphic. However any coordinate representation of this map is the identity. So where's the problem? The issue is that the *domain* of the coordinate representation will not be an open set, and so it is not an analytic function. «

Since $\mathbb{C}$ is simply connected, Proposition 12.54 gives:

**Corollary 12.57** *$\Sigma$ is simply connected.*

In particular, it is connected, and so is a Riemann surface.

### 12.4.2 The Surface for the $n$th Root

Let $n \geqslant 1$. The global logarithm allows us to define a global $n$th root. For $(z, t) \in \Sigma$ define

$$\mathrm{rt}_n(z, t) = \sqrt[n]{|z|} e^{it/n}.$$

The point of course is that $(\mathrm{rt}_n(z, t))^n = z$.

**Proposition 12.58** *The map $\mathrm{rt}_n : \Sigma \to \mathbb{C}$ is holomorphic.*

***Proof*** It is the composition of three holomorphic functions:

$$(z, t) \mapsto \ln|z| + it \mapsto \frac{1}{n}(\ln|z| + it) \mapsto \exp(\ln|z|/n + it/n). \qquad \square$$

*Remark 12.59* We can use the global $n$th root to define local ones: suppose that $U \subset \mathbb{C}^*$ is simply connected. Let $\psi$ be a chart for $\Sigma$ whose range is $U$. Then $\mathrm{rt}_n \circ \psi^{-1}$ is an injective analytic function on $U$ whose inverse is the map $z \mapsto z^n$. Compare with Example 11.12. «

The Riemann surface for the $n$th root is a quotient of $\Sigma$, obtained by identifying each $(z, t) \in \Sigma$ with $(z, t + 2\pi n)$: we collapse the $\mathbb{Z}$-many copies of $z$ to just $n$ many. See Fig. 12.2. More formally, consider the quotient of $\mathbb{C} \times \mathbb{R}$ by the cyclic subgroup generated by $(0, 2\pi n)$. This quotient is isomorphic to $\mathbb{C} \times \mathbb{R}/2\pi n\mathbb{Z}$; we let $\Sigma/n$ be the image of $\Sigma$ under this quotient map:

$$\Sigma/n = \{(z, t + 2\pi n\mathbb{Z}) \,:\, (z, t) \in \Sigma\}.$$

The charts for $\mathbb{C} \times \mathbb{R}/2\pi n\mathbb{Z}$ are of the form $(z, t + 2\pi n\mathbb{Z}) \mapsto (z, t)$, defined for $(z, t) \in U \times I$ where $U \subseteq \mathbb{C}$ is open and $I \subset \mathbb{R}$ is an open interval of length $< \pi n$ (so that $t \mapsto t + 2\pi n\mathbb{Z}$ is 1–1 on $I$). We let the charts for $\Sigma/n$ be compositions $\alpha \circ \beta$ where $\beta$ is a chart for $\mathbb{C} \times \mathbb{R}/2\pi n\mathbb{Z}$ and $\alpha$ is a chart for $\Sigma$. In other words, they are the inverses of maps $z \mapsto (z, \theta(z) + 2\pi n\mathbb{Z})$ from $U$ to $\Sigma/n$, where $U \subseteq \mathbb{C} \setminus \{0\}$ is sufficiently small so that the range of $\theta$ (a continuous choice of argument on $U$) is contained in a short interval $I$. The transition map between two such charts $\alpha \circ \beta$ and $\widetilde{\alpha} \circ \widetilde{\beta}$ is the same as the transition map between $\alpha$ and $\widetilde{\alpha}$, and so two such charts are compatible, and we get an atlas on $\Sigma/n$. The fact that the quotient map

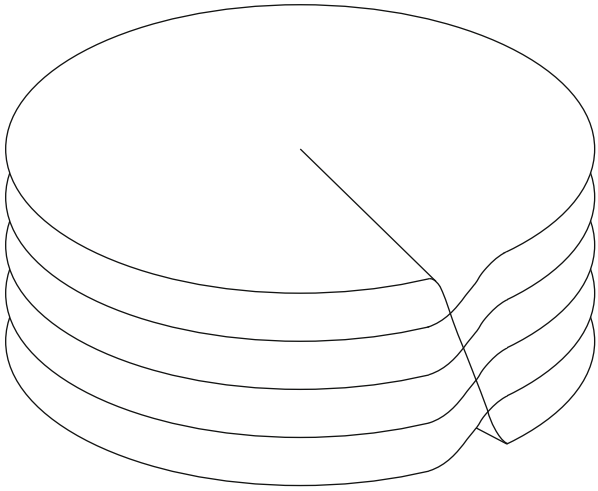


**Fig. 12.2** The Riemann surface for the $n$th root

$(z, t) \mapsto (z, t + 2\pi n\mathbb{Z})$ is locally a homeomorphism between open subsets of $\mathbb{C} \times \mathbb{R}$ and of $\mathbb{C} \times \mathbb{R}/2\pi n\mathbb{Z}$, implies that equipped with this atlas, $\Sigma/n$ is a topological subspace of $\mathbb{C} \times \mathbb{R}/2\pi n\mathbb{Z}$, and is a Riemann surface.

Let $p_n \colon \Sigma \to \Sigma/n$ be the quotient map $(z, t) \mapsto (z, t + 2\pi n\mathbb{Z})$. Then $p_n$ is the restriction to $\Sigma$ of the quotient map from $\mathbb{C} \times \mathbb{R}$ to $\mathbb{C} \times \mathbb{R}/2\pi n\mathbb{Z}$, which is continuous (Proposition 8.107); as $\Sigma$ is a topological subspace of $\mathbb{C} \times \mathbb{R}$, and $\Sigma/n$ is a topological subspace of $\mathbb{C} \times \mathbb{R}/2\pi n\mathbb{Z}$, the map $p_n$ is continuous on $\Sigma$. A coordinate representation of $p_n$ is the identity, so $p_n$ is holomorphic.

The following is a holomorphic analogue of Proposition 8.109:

**Lemma 12.60** *Let $Y$ be a holomorphic surface. A function $f \colon \Sigma/n \to Y$ is holomorphic if and only if the composition $f \circ p_n \colon \Sigma \to Y$ is holomorphic.*

***Proof*** The proof of Proposition 8.109 shows that $f$ is continuous if and only if $f \circ p_n$ is continuous. The equivalence now follows since $f$ and $f \circ p_n$ have the same coordinate representations. □

Here is one example:

*Exercise 12.61* Show that $\pi_{\Sigma/n} \colon \Sigma/n \to \mathbb{C}$ defined by $(z, t + 2\pi n\mathbb{Z}) \mapsto z$ is holomorphic. «

The intended example though is the map $\mathrm{rt}_n$, which induces a well-defined function on $\Sigma/n$. To avoid excessive notation, we also let $\mathrm{rt}_n \colon \Sigma/n \to \mathbb{C}$ be the

induced holomorphic function on $\Sigma/n$. It is a bijection between $\Sigma/n$ and $\mathbb{C} \setminus \{0\}$, and so Proposition 12.37 implies:

**Proposition 12.62** $\mathrm{rt}_n$ *is a biholomorphism between* $\Sigma/n$ *and* $\mathbb{C} \setminus \{0\}$.

Let $\mathrm{pwr}_n : \mathbb{C} \setminus \{0\} \to \Sigma/n$ denote the inverse of $\mathrm{rt}_n$; it maps $z = re^{it}$ to $(z^n, nt + 2\pi n\mathbb{Z})$. We can imagine this map as "wrapping" the punctured complex plane around itself $n$ times, again see Fig. 12.2.

### The Shift on $\Sigma$

Let $\mathrm{sh} : \Sigma \to \Sigma$ be the map

$$(z, t) \mapsto (z, t + 2\pi).$$

This is an "upwards shift" of $\Sigma$; each "sheet" covering $\mathbb{C} \setminus \{0\}$ is mapped to the sheet above it. It is 1–1 and continuous (it is the restriction of a continuous map on $\mathbb{C} \times \mathbb{R}$), and its coordinate representations are the identity, so it is a biholomorphism between $\Sigma$ and itself. For each $k \in \mathbb{Z}$, let $\mathrm{sh}_k$ denote the iteration of this shift $k$ times, that is, the map $(z, t) \mapsto (z, t + 2\pi k)$.

For each $n \geqslant 1$, the shift induces a map $\Sigma/n \to \Sigma/n$.

*Exercise 12.63* Verify that the induced map $\mathrm{sh} : \Sigma/n \to \Sigma/n$ is continuous. «

The coordinate representations of the induced shift are also the identity, and so the induced shift is a biholomorphism from $\Sigma/n$ to itself; the $k$th iteration of the induced map is the map induced by $\mathrm{sh}_k$, so we also denote it as $\mathrm{sh}_k : \Sigma/n \to \Sigma/n$. Note that $\mathrm{sh}_n : \Sigma/n \to \Sigma/n$ is the identity.

*Exercise 12.64* Observe that a function $f : \Sigma \to X$ induces a well-defined function from $\Sigma/n \to X$ if and only if it is invariant under $\mathrm{sh}_n$, that is, if $f \circ \mathrm{sh}_n = f$. «

*Exercise 12.65* Let $\omega_n = e^{2\pi i/n}$ be the primitive $n$th root of unity. Verify that $(\mathrm{rt}_n \circ \mathrm{sh})(q) = \omega_n \cdot \mathrm{rt}_n(q)$ for all $q \in \Sigma$. «

## 12.5 Analytic Continuation

We will later find ourselves in the situation where we have a holomorphic function $f : U \to Y$ where $U$ is an open and connected subset of a Riemann surface $X$, and will want to extend it to a holomorphic function from $X$ to $Y$. To do this we will use the process of *analytic continuation*.

**Definition 12.66** Suppose that $X$ and $Y$ are holomorphic surfaces, $U \subseteq X$ is open, $f : U \to Y$ is holomorphic, and let $\gamma : [a, b] \to X$ be a path with $\gamma(a) \in U$.

An *analytic continuation* of $f$ along $\gamma$ is a continuous function $g\colon [a,b] \to Y$ such that there is a partition $a = t_0 < t_1 < \cdots < t_k = b$ of $[a,b]$, and for $i = 1, 2, \ldots, k$ an open $U_i \subseteq X$ and a holomorphic $f_i\colon U_i \to Y$, such that: (i) $\gamma[t_{i-1}, t_i] \subset U_i$; (ii) $g = f_i \circ \gamma$ on $[t_{i-1}, t_i]$; (iii) for $i < k$, $f_i = f_{i+1}$ on some neighbourhood of $\gamma(t_i)$; and (iv) $U_1 \subseteq U$ and $f_1 = f$ on $U_1$.

For an equivalent definition see Exercise 12.112.

*Example 12.67* Even if $\gamma$ is a loop, the function we get to at the end of the process may not be the same as the one we started with. The standard example is the logarithm. Let $\gamma\colon [0, 2\pi] \to S$ be the usual parameterisation of the unit circle; start with $\mathrm{Log}(re^{it}) = \ln r + it$ where $r > 0$ and $t \in (-\pi, \pi)$ (this is the "principal branch" of the complex logarithm, see page 307). Then $t \mapsto it$ is an analytic continuation of Log along $\gamma$, and note that the value at $t = 2\pi$ is $2\pi i$, which is not $\mathrm{Log}(\gamma(2\pi)) = \mathrm{Log}(1) = 0$. «

**Lemma 12.68** *Let $X$ and $Y$ be holomorphic surfaces, let $U \subseteq X$ be open, $f\colon U \to Y$ be holomorphic, and let $\gamma\colon [a,b] \to X$ be a path with $\gamma(a) \in U$. Then $f$ has at most one analytic continuation along $\gamma$.*

***Proof*** Suppose that $g$ and $h$ are two analytic continuations of $f$ along $\gamma$. By taking a common refinement, we may assume that $g$ and $h$ are witnessed by the same partition: there is a partition $a = t_0 < t_1 < \cdots < t_k = b$ of $[a,b]$, and for $i = 1, 2, \ldots, k$ open sets $U_i, V_i \subseteq X$ and holomorphic $g_i\colon U_i \to Y$ and $h_i\colon V_i \to Y$ with $g = g_i \circ \gamma$ on $[t_{i-1}, t_i]$ and $h = h_i \circ \gamma$ on $[t_{i-1}, t_i]$, $g_i = g_{i+1}$ and $h_i = h_{i+1}$ on a neighbourhood of $\gamma(t_i)$, and $g_1 = f$ on $U_1$ and $h_1 = f$ on $V_1$. Further, each $\gamma[t_{i-1}, t_i]$ has a *connected* open neighbourhood $\subseteq U_i \cap V_i$, so we may assume that $V_i = U_i$ and is connected.

Now by induction on $i$ we show that $g_i = h_i$; this is known for $i = 1$ since $g_i = f = h_i$ on $U_1$. Assuming that $g_i = h_i$ and $i < k$, we know that $g_{i+1} = g_i$ and $h_{i+1} = h_i$ on a neighbourhood of $\gamma(t_i)$, and so $g_{i+1} = h_{i+1}$ on a neighbourhood of $\gamma(t_i)$; since $U_{i+1}$ is connected, we have $g_{i+1} = h_{i+1}$ (Proposition 12.9). □

The following is analogous to Lemma 10.37.

**Lemma 12.69** *Suppose that $H\colon [0,1] \times [a,b] \to X$ is a path homotopy in a holomorphic surface $X$. Suppose that $f\colon U \to X$ is a holomorphic function defined on an open neighbourhood of $H_0(a)$. Suppose that for all $t \in [0,1]$, there is an analytic continuation $g_t$ of $f$ along $H_t$. Then $g_0(b) = g_1(b)$.*

***Proof*** The proof has some resemblance to that of Proposition 9.21. Let $u \in [0,1]$; we show that $u$ has an open neighbourhood in $[0,1]$ on which $t \mapsto g_t(b)$ is constant. The result then follows from the compactness of $[0,1]$.

Let $a = s_0 < s_1 < \cdots < s_k = b$ be a partition of $[a,b]$ and $U_1, \ldots, U_k$ and $f_1, \ldots, f_k$ witness that $g_u$ is an analytic continuation of $f$ along $H_u$. For each

$i = 1, \ldots, k$, $H^{-1}[U_i]$ is open and $\{(u, s) : s \in [s_{i-1}, s_i]\}$ is compact, so there is some $\delta > 0$ such that for all $i \leqslant k$, for all $t \in (u - \delta, u + \delta)$, for all $s \in [s_{i-1}, s_i]$ we have $H_t(s) \in U_i$, and further we may assume that for $i < k$, $H_t(s_i)$ lies in a neighbourhood of $H_u(s_i)$ on which $f_i = f_{i+1}$. For such $t$, $s \mapsto f_i(H_t(s))$ (for $s \in [s_{i-1}, s_i]$) is an analytic continuation of $f$ along $H_t$, and so equals $g_t$ by Lemma 12.68. Hence $g_t(b) = f_k(H_t(b)) = f_k(H_u(b))$ is constant for $t \in (u - \delta, u + \delta)$ as required. □

Our goal is:

**Monodromy Theorem** *Let $X$ and $Y$ be Riemann surfaces, with $X$ simply connected. Let $U \subseteq X$ be open and connected, and suppose that $f : U \to Y$ is holomorphic. Suppose that along any path $\gamma$ in $X$ starting at some point in $U$, there is an analytic continuation of $f$ along $\gamma$. Then there is a (unique) holomorphic function from $X$ to $Y$ extending $f$.*

***Proof*** Uniqueness is by Proposition 12.9. Existence is similar to the proof of Theorem 9.24. Fix some $a \in U$, and define $h : X \to Y$ by letting, for $b \in X$, $h(b)$ be the end value of some analytic continuation along some path $\gamma$ in $X$ from $a$ to $b$. Since $X$ is simply connected, Lemma 12.69 implies that this does not depend on the choice of path. To show that $h$ is holomorphic, fix some $b_0 \in X$ and some path $\gamma_0$ from $a$ to $b_0$; let $g_0$ be the unique analytic continuation of $f$ along $\gamma_0$. Let $h_0$ be a holomorphic function defined on a connected open neighbourhood $V$ of $b_0$ such that $g_0 = h_0 \circ \gamma_0$ on some tail of $\gamma_0$. For any $b \in V$, let $\gamma_b$ be the result of concatenating to $\gamma_0$ a path in $V$ from $b_0$ to $b$. We then let $g_b$ be an extension of $g_0$ along $\gamma_b$ by following $h_0$ on the path from $b_0$ to $b$. Then $g_b$ is an analytic continuation of $f$ along $\gamma_b$, and so $h(b) = h_0(b)$, i.e., $h = h_0$ on $V$; so $h$ is holomorphic on $V$. □

## 12.6 Differential Forms on Surfaces

How can we differentiate and integrate functions on surfaces? The key to this is remembering that fundamentally, the objects that are integrated are not functions but *forms*, and so instead of derivatives we should look for differentials. So how to define forms on a surface? The "proper" way to do it would be to define the tangent bundle, but that would take us too far afield. A low-tech solution is to declare a form on a surface by choosing for each chart a form on the image of the chart, in a pairwise coherent way.

Suppose that $U, V \subseteq \mathbb{C}$ are open, that $\omega$ is a complex linear form on $V$ (see page 285), and that $f : U \to V$ is analytic. We define (compare with Exercise 10.43) a form $f^*\omega$ on $U$ by letting $(f^*\omega)_a(v) = \omega_{f(a)}(f'(a) \cdot v)$ (the product is the product

of complex numbers). Now $\omega = g\,dz$ for some continuous $g\colon V \to \mathbb{C}$; we observe that

$$f^*(g\,dz) = g(f)\cdot f'\,dz. \tag{12.2}$$

*Exercise 12.70* Show that if $f\colon U \to V$ and $h\colon V \to W$ are analytic, and $\omega$ is a form on $W$, then $(h\circ f)^*\omega = f^*(h^*\omega)$. Also show that $\mathrm{id}^*\,\omega = \omega$ where id is the identity function on $W$; conclude that if $f$ is a bijection then $(f^{-1})^* f^*\omega = \omega$. «

**Definition 12.71** Let $X$ be a holomorphic surface. A *form* on $X$ is a collection of pairs $\omega = \{(\omega_i, \psi_i) \,:\, i \in I\}$ where:

(i) each $\psi_i$ is a chart for $X$, and $\{\mathrm{dom}\,\psi_i \,:\, i \in I\}$ is an open cover of $X$;
(ii) each $\omega_i$ is a complex form on range $\psi_i$;
(iii) for any $i, j \in I$,

$$\omega_i \restriction_{\mathrm{dom}\,\psi_{i,j}} = (\psi_{i,j})^*\omega_j.$$

where $\psi_{i,j} = \psi_j \circ \psi_i^{-1}$ is the transition function from $\psi_i$- to $\psi_j$-coordinates.

If $\omega = \{(\omega_i, \psi_i)\}$ is a form on $X$ then we often write $\omega_i = g_i\,dz$ where $g_i$ is continuous. Thus, Eq. 12.2 gives

$$g_i = (g_j \circ \psi_{i,j})\cdot(\psi_{i,j})' \tag{12.3}$$

on $\mathrm{dom}\,\psi_{i,j}$, for all $i$ and $j$. The form is called *holomorphic* if each $g_i$ is analytic. Holomorphic forms are very useful, but we will need to generalise this a bit, using meromorphic functions.

**Definition 12.72** Let $X$ be a holomorphic surface. A *meromorphic form* on $X$ is a collection of pairs $\omega = \{(g_i, \psi_i) \,:\, i \in I\}$ where

1. each $\psi_i$ is a chart for $X$, and $\{\mathrm{dom}\,\psi_i \,:\, i \in I\}$ is an open cover of $X$;
2. each $g_i$ is a meromorphic function on range $\psi_i$; and
3. for any $i, j \in I$,

$$g_i = (g_j \circ \psi_{i,j})\cdot(\psi_{i,j})'.$$

on $\mathrm{dom}\,\psi_{i,j}$.

Here recall that we defined the product of meromorphic functions; see page 317. In this case, since each $\psi_{i,j}$ is an analytic bijection, we know that $(\psi_{i,j})'$ is nonzero everywhere, so $a \in \mathrm{dom}\,\psi_{i,j}$ is a pole of $g_i$ if and only if $\psi_{i,j}(a)$ is a pole of $g_j$; and similarly for zeros.

We say that a meromorphic form $\omega = \{(g_i, \psi_i)\}$ is not constant 0 if for each $i$, $g_i$ is not constant 0. In this case, the order $\mathrm{ord}_a(g_i)$ of $g_i$ at $a$ is defined for all $a \in$

$\operatorname{dom} g_i = \operatorname{range} \psi_i$; by Exercise 12.21, the fact that $(\psi_{i,j})'$ is nonzero everywhere implies:

**Lemma 12.73** *Let $\omega = \{(g_i, \psi_i) : i \in I\}$ be a meromorphic form on $X$ and not constant 0; let $p \in X$. Then for every $i, j \in I$ such that $p \in \operatorname{dom} \psi_i \cap \operatorname{dom} \psi_j$,*

$$\operatorname{ord}_{\psi_i(p)}(g_i) = \operatorname{ord}_{\psi_j(p)}(g_j).$$

As in Remark 12.40, we thus unambiguously define $\operatorname{ord}_p(\omega)$, the order of $\omega$ at $p$, to be this common value $\operatorname{ord}_{\psi_i(p)}(g_i)$. If $\operatorname{ord}_p(\omega) < 0$ then we say that $p$ is a pole of $\omega$, and if $\operatorname{ord}_p(\omega) > 0$ we say that $p$ is a zero of $\omega$.

*Exercise 12.74* (a) Show that the collection of poles of a meromorphic form $\omega$ on $X$ is a discrete and closed subset of $X$. (b) Show that if $\omega$ is not constant 0, then the set of zeros of $\omega$ is discrete and closed as well. (c) Show that if $X$ is a Riemann surface then a meromorphic form $\omega$ on $X$ is constant 0 if and only if every $p \in X$ is a zero of $\omega$.[3] «

If each $g_i$ is analytic, then we can identify $\omega = \{(g_i, \psi_i)\}$ with the form $\{(g_i\, dz, \psi_i)\}$, so we think of meromorphic forms as a generalisation of holomorphic forms.

*Remark 12.75* Strictly speaking, of course, a meromorphic form on $X$ may not be a form on $X$, since $g_i\, dz$ is not defined on the poles of $g_i$. However, removing the poles leaves us with a holomorphic form.

In greater detail, if $\omega = \{(g_i, \psi_i) : i \in I\}$ is a meromorphic form on $X$, and $U \subseteq X$ is open, we let $\omega\restriction_U$ denote the form $\{(g_i\restriction_{\psi_i[U]}, \psi_i\restriction_U) : i \in I\}$. This is a meromorphic form on $U$. If $Z$ is the set of poles of $\omega$, then $\omega\restriction_{X\setminus Z}$ is a holomorphic form on $X \setminus Z$. «

### 12.6.1 Pull-Backs of Meromorphic Forms

Even though it is imprecise, if $U \subseteq \mathbb{C}$ is nonempty and $g$ is meromorphic on $U$, then it is convenient to denote by $g\, dz$ the meromorphic form on $U$ which consists of the unique pair $(g, \mathrm{id}_U)$. If $f : U \to V$ is analytic and $g$ is meromorphic on $V$ then $(g \circ f) \cdot f'$ is meromorphic on $U$ and as above, we let $f^*(g\, dz) = (g \circ f) \cdot f'\, dz$; so

[3] Note that if $\omega$ is not constant 0, then we cannot conclude that we have a form $1/\omega$ on $X$ defined by taking pairs $(1/g_i, \psi_i)$; the coherence condition $1/g_i = ((1/g_j) \circ \psi_{i,j}) \cdot (\psi_{i,j})'$ will fail, as we can have $(\psi_{i,j})' \neq 1$. A meromorphic form does not give a well-defined function on $X$; only the zeros and poles are unambiguous.

condition (iii) of Definition 12.72 becomes $g_i\,dz = (\psi_{i,j})^* g_j\,dz$ on $\operatorname{dom}\psi_{i,j}$, just as in Definition 12.71.

Using this notation, we can extend the pull-back operation to meromorphic forms on surfaces. Suppose that $f\colon X \to Y$ is a holomorphic map between holomorphic surfaces, and that $\omega = \{(\omega_i, \psi_i) : i \in I\}$ is a meromorphic form on $Y$ (where $\omega_i = g_i\,dz$, with $g_i$ a meromorphic function on the range of $\psi_i$). Then we define a form $f^*\omega$ on $X$ by pulling back coordinate representations. We fix a holomorphic atlas $\mathcal{A} = \{\varphi_j : j \in J\}$ on $X$. By breaking up their domains, if necessary, we assume that for every $j \in J$, $\operatorname{dom}\varphi_j$ is sufficiently small so that $f[\operatorname{dom}\varphi_j] \subseteq \operatorname{dom}\psi_{i(j)}$ for some $i(j) \in I$. Then for each $j \in J$, we let

$$\eta_j = \left(\psi_{i(j)} \circ f \circ \varphi_j^{-1}\right)^* \omega_{i(j)}$$

be the pull-back meromorphic form on range $\varphi_j$ using the coordinate representation of $f$ using $\varphi_j$- and $\psi_{i(j)}$-coordinates; we then let

$$f^*\omega = \{(\eta_j, \varphi_j) : j \in J\}.$$

*Exercise 12.76* (a) Show that $f^*\omega$ is a meromorphic form on $X$. (b) Show that if $h\colon Z \to X$ is holomorphic then $(f \circ h)^*\omega = h^*(f^*\omega)$. (c) Let $p \in X$; let $n$ be the valency of $f$ at $p$ (see Proposition 12.41), and let $k = \operatorname{ord}_{f(p)}(\omega)$. Show that $\operatorname{ord}_p(f^*\omega) = nk+n-1$. (In particular, if $f$ is a biholomorphism, then $\operatorname{ord}_p(f^*\omega) = \operatorname{ord}_{f(p)}(\omega)$.) «

*Exercise 12.77* Show that if $\omega = (\omega_i, \psi_i)$ is a meromorphic form on $X$, then for all $i$, $(\psi_i^{-1})^*\omega = \omega_i$. «

**Equivalence of Forms**

For a meromorphic form $\omega = \{(\omega_i, \psi_i) : i \in I\}$ on a surface $X$, temporarily let $\mathcal{A}_\omega = \{\psi_i : i \in I\}$, which is an atlas on $X$ (with analytic transition functions). If $\eta = \{(\eta_j, \varphi_j) : j \in J\}$ is another meromorphic form on $X$ with $\mathcal{A}_\omega \neq \mathcal{A}_\eta$, then strictly speaking, the forms $\omega$ and $\eta$ cannot be identical. However, we do think of them as the same if for all $(i, j) \in I \times J$, $\omega_i = (\varphi_j \circ \psi_i^{-1})^*\eta_j$. That is, if the collection of all pairs $(\omega_i, \psi_i)$ together with all pairs $(\eta_j, \varphi_j)$ (technically speaking, the union $\omega \cup \eta$) is a form on $X$.

### 12.6.2 Quotients of Forms

Recall that if $\omega$ is a form on $U \subseteq \mathbb{C}$ and $h\colon U \to \mathbb{C}$ is a continuous function, then we define the form $h\omega$ on $U$ by letting $(h\omega)_a(v) = h(a)\omega_a(v)$. If $\omega = g\,dz$ then $h\omega = (hg)\,dz$. Since products of meromorphic functions are well-defined, we can extend this to meromorphic forms on $U$.

*Exercise 12.78* Show that if $f\colon U \to V$ is analytic, $\omega = g\,dz$ is a meromorphic form on $V$, and $h$ is a meromorphic function on $V$, then

$$f^*(h\omega) = (h \circ f) \cdot f^*\omega. \qquad \text{«}$$

If $\omega = \{(\omega_i, \psi_i) : i \in I\}$ is a meromorphic form on $X$ and $h$ is a meromorphic function on $X$ (a holomorphic function to $\mathbb{P}^1(\mathbb{C})$), then for $i \in I$ let $h_i = h \circ \psi_i^{-1}$ be the representation of $h$ using $\psi_i$-coordinates; we define

$$h\omega = \{(h_i\omega_i, \psi_i) : i \in I\};$$

Exercise 12.78 implies that this is a meromorphic form on $X$.

*Exercise 12.79* Show that if $f\colon X \to Y$ is holomorphic, $\omega$ is a meromorphic form on $Y$, and $h$ is a meromorphic function on $Y$, then $f^*(h\omega) = (h \circ f)\, f^*\omega$. «

Recall (Remark 12.40) that we can define the order $\operatorname{ord}_p(h)$ of $h$ at a point $p \in X$. Then Exercise 12.21(a) implies:

**Proposition 12.80** *For every $p \in X$, $\operatorname{ord}_p(h\omega) = \operatorname{ord}_p(\omega) + \operatorname{ord}_p(h)$.*

And we can divide forms.

**Proposition 12.81** *Suppose that $\omega$ and $\eta$ are meromorphic forms on $X$, with $\omega$ not constant 0. Then there is a meromorphic function $h$ on $X$ such that $\eta = h\omega$.*

We write $h = \eta/\omega$.

***Proof*** As discussed above, we can take a common refinement and pass to equivalent forms, so we assume that $\omega$ and $\eta$ are defined using the same atlas: $\omega = \{(\omega_i, \psi_i) : i \in I\}$ and $\eta = \{(\eta_i, \psi_i) : i \in I\}$. Say $\omega_i = g_i\,dz$ and $\eta_i = f_i\,dz$. For each $i \in I$ we define $h_i\colon \operatorname{dom}\psi_i \to \mathbb{C}$ by letting $h_i(p) = (f_i(\psi_i(p)))/(g_i(\psi_i(p)))$, which is meromorphic since $g_i$ is not identically zero; a coordinate representation of $h_i$ is $f_i/g_i$. However for $i, j \in I$, the compatibility of $\omega_i$ and $\omega_j$, and of $\eta_i$ and $\eta_j$ shows that $h_i = h_j$ on $\operatorname{dom}\psi_i \cap \operatorname{dom}\psi_j$. So we can unambiguously define $h$ as required. □

Now Proposition 12.80 implies that if both $\eta$ and $\omega$ are holomorphic (they have no poles), and $\omega$ has no zeros as well, then $h$ is in fact holomorphic. We say that $\omega$

is *non-vanishing holomorphic* if it is holomorphic and has no zeros. That is, if $\mathrm{ord}_p(\omega) = 0$ for all $p \in X$. With Corollary 12.44, we get:

**Corollary 12.82** *If $X$ is a compact Riemann surface and $\omega$ is a non-vanishing holomorphic form on $X$, then every holomorphic form on $X$ is a constant multiple of $\omega$.*

### The Differential of a Meromorphic Function

Let $X$ be a holomorphic surface and let $g$ be a meromorphic function on $X$. Let $\mathcal{A} = \{\psi_i : i \in I\}$ be a holomorphic atlas on $X$. We define

$$dg = \left\{ ((g \circ \psi_i^{-1})' \, dz, \psi_i) : i \in I \right\}.$$

*Exercise 12.83* Show that $dg$ is a meromorphic form on $X$. «

*Example 12.84* If $U \subseteq \mathbb{C}$ and $g$ is meromorphic on $U$, then $dg = g' \, dz$. «

*Exercise 12.85* Show that if $f : X \to Y$ is holomorphic and $g$ is meromorphic on $Y$, then $f^*(dg) = d(g \circ f)$. «

We can calculate orders of zeros and poles:

**Proposition 12.86** *For all $p \in X$:*

1. *If $\mathrm{ord}_p(g) \neq 0$ then $\mathrm{ord}_p(dg) = \mathrm{ord}_p(g) - 1$.*
2. *If $\mathrm{ord}_p(g) \geqslant 0$ then $\mathrm{ord}_p(dg) = m - 1$, where $m$ is the valency of $g$ at $p$.*

See Exercise 12.21(c).

*Example 12.87* Let $g = \mathrm{id}_{\mathbb{P}^1(\mathbb{C})}$ be the identity function on $\mathbb{P}^1(\mathbb{C})$. It is a meromorphic function on $\mathbb{P}^1(\mathbb{C})$. The coordinate representations of $g$ are $\rho_0^{-1}$ and $\rho_1^{-1}$, which using our identification of $\mathbb{C}$ with $\rho_0[\mathbb{C}]$, translate to the functions $z \mapsto z$ and $z \mapsto 1/z$. Hence $dg$ is the form consisting of the two pairs $(dz, \rho_0)$ and $((-1/z^2)\, dz, \rho_1)$. We observe then that $dg$ has no zeros, and has a pole of order 2 at $p_\infty$. Since $dg$ is not constant 0, Proposition 12.81 implies that every meromorphic form on the Riemann sphere is $f \, dg$ for some meromorphic function $f$ on the sphere.[4] «

*Exercise 12.88* Fix a complex torus $T = T_\Gamma$. Define a form $\omega$ on $T$ by taking all pairs $(dz, \psi)$ where $\psi$ is a chart for $T$. (a) Verify that $\omega$ is a non-vanishing holomorphic form on $T$. (b) Conclude that every meromorphic form on $T$ is $f\omega$ for

[4] The form $dg$ extends the form $dz$ on $\mathbb{C}$, and so is often also referred to as $dz$.

some meromorphic function $f$ on $T$. (c) If $\pi = \pi_\Gamma$ is the quotient map, show that $dz = \pi^*\omega$. «

### 12.6.3 Integration of Holomorphic Forms

For the rest of the section, we restrict ourselves to piecewise smooth paths in a holomorphic surface $X$ (recall that $X$ is a differentiable 2-manifold, so we have defined the notion of smoothness of maps to and from $X$; see Proposition 9.63). Suppose that $\omega = \{(\omega_i, \psi_i) : i \in I\}$ is a holomorphic form on $X$, and that $\gamma : [a, b] \to X$ is a path in $X$.

Suppose that $i, j \in I$ and that the range $\gamma[a, b]$ is contained in $\operatorname{dom}\psi_i \cap \operatorname{dom}\psi_j$. Then because $\omega_i = (\psi_{i,j})^*\omega_j$ on $\operatorname{dom}\psi_{i,j}$,

$$\int_{\psi_i \circ \gamma} \omega_i = \int_{\psi_j \circ \gamma} \omega_j.$$

[Why? by considering concatenations, we may assume that $\gamma$ is smooth. If $\omega_i = g_i\, dz$ then $\int_{\psi_i\circ\gamma} \omega_i = \int_a^b (g_i \circ \psi_i \circ \gamma)(t) \cdot (\psi_i \circ \gamma)'(t)\, dt$; now use the chain rule. Compare with Exercise 10.43.]

Now dropping this assumption, suppose instead that $a = t_0 < t_1 < \cdots < t_k = b$ is a partition of $[a, b]$ and that for every $j = 1, \ldots, k$, the image $\gamma[t_{j-1}, t_j]$ is contained in the domain of one of the charts $\psi_i$ for some $i = i(j)$. Then we can define

$$\int_\gamma \omega = \sum_{j=1}^{k} \int_{\psi_{i(j)} \circ \gamma\restriction_{[t_{j-1}, t_j]}} \omega_{i(j)},$$

and the point is that such partitions exist, and that the value of the integral does not depend on the choice of partition or the charts $\psi_{i(j)}$. If $X$ is an open subset of $\mathbb{C}$, then this integral is the same as the usual integral we already defined.

*Exercise 12.89* Show that if $f, g : X \to \mathbb{C}$ are holomorphic and $\gamma : [a, b] \to X$ is smooth, then

$$\int_\gamma f\, dg = \int_a^b (f \circ \gamma) \cdot (g \circ \gamma)'(t)\, dt.$$ «

*Exercise 12.90* Show that if $g\colon X \to \mathbb{C}$ is holomorphic and $\gamma\colon [a,b] \to X$ is a path in $X$ then

$$\int_\gamma dg = g(\gamma(b)) - g(\gamma(a)).$$ «

In practice, $\omega$ will be meromorphic rather than holomorphic, and then we just require that $\gamma$ does not pass through any pole of $\gamma$; this reduces to the holomorphic case by Remark 12.75.

*Exercise 12.91* Show that if $f\colon X \to Y$ is holomorphic, $\omega$ is a meromorphic form on $Y$ and $\gamma$ is a path in $X$ avoiding the poles of $f^*\omega$, then

$$\int_\gamma f^*\omega = \int_{f\circ\gamma} \omega.$$ «

*Exercise 12.92* Let $\omega$ be a holomorphic form on $X$. Show that if $\gamma$ and $\delta$ are homotopic in $X$, then $\int_\gamma \omega = \int_\delta \omega$. (Hint: follow the argument of Lemma 10.37. We can require that each $\zeta_i$ (and its "interior") lies in the domain of a chart.) «

## 12.7 Further Exercises

For more problems, a good source is [VLA65].

**Essential Singularities**

12.93 Show that 0 is an essential singularity of the function $z \mapsto e^{1/z}$ (see page 314).

12.94 Let $f$ be analytic, defined on a punctured neighbourhood of $a$, and suppose that $a$ is an essential singularity of $f$. Show that the image of $f$ is dense in $\mathbb{C}$. (This is known as the *Casorati-Weierstrass* theorem. Hint: argue by contradiction; if $f$ avoids a neighbourhood of a point $b \in \mathbb{C}$, then $1/(f-b)$ is bounded. *Picard's Great theorem* extends the Casorati-Weierstrass theorem; it says that the image of $f$ is all of $\mathbb{C}$, except possibly one point.)

12.95 Let $U$ be an open neighbourhood of 0; let $f\colon U \setminus \{0\} \to \mathbb{C}$ be analytic. (a) For $a > 0$ let $\gamma_a$ be the parameterised circle of radius $a$ around 0; for $b > a > 0$ let $\delta_{a,b}$ be the path travelling from $a$ to $b$ along the real line in constant unit speed. Show that the concatenation $\gamma_b - \delta_{a,b} - \gamma_a + \delta_{a,b}$ is homotopic to a constant loop in $U \setminus \{0\}$. (b) Let $w \in U \setminus \{0\}$; let $0 < r < |w| < R$ such that $\overline{B}(0,R) \subset U$. Show that $2\pi i \cdot f(w) = \int_{\gamma_R} f(z)/(z-w)\,dz - \int_{\gamma_r} f(z)/(z-w)\,dz$. (c) Show that $f$ has a *bi-infinite Laurent expansion* around 0: if $B(0,R) \subseteq U$ then $f(z) = \sum_{n=-\infty}^{\infty} c_n z^n$

on $B(0, R) \setminus \{0\}$. (d) Show that 0 is an essential singularity of $f$ if and only if for infinitely many $n \geqslant 0$, $c_{-n} \neq 0$.

**Calculus of Residues**

12.96 Find the poles and the residues of $f(z) = 1/\sin^2 z$.

12.97 Let $\gamma$ be the parameterized unit circle.

(a) Suppose that $f\colon [0, 2\pi] \to \mathbb{R}$ is continuous and $f(0) = f(2\pi)$. Let $\hat{f}(e^{it}) = f(t)$. Show that

$$\int_0^{2\pi} f(t)\, dt = \int_\gamma \frac{\hat{f}(z)}{iz}\, dz.$$

(b) Let $f(t) = 1/(2 + \cos t)$. Show that $\hat{f}(z) = 2z/(z^2 + 4z + 1)$. (Hint: $2\cos t = e^{it} + e^{-it}$.)

(c) Find $\int_0^{2\pi} dt/(2 + \cos t)$. (Hint: find $\operatorname{rsd}_{\sqrt{3}-2}(1/(z^2 + 4z + 1))$.)

12.98 Let $f(z) = (1 + z^2)^{-2}$. For $r > 1$ let $\gamma_r\colon [0, \pi] \to \mathbb{C}$ be $\gamma_r(t) = re^{it}$ be the parameterized *semi*-circle. (a) Show that $\lim_{r\to\infty} \int_{\gamma_r} f(z)\, dz = 0$. (Hint: $f(z)$ behaves like $1/z^4$ when $|z|$ is large.) (b) Calculate $\operatorname{rsd}_i(f)$. (c) For $r > 1$, find $\int_{-r}^{r} f(x)\, dx + \int_{\gamma_r} f(z)\, dz$. (d) Calculate $\int_{-\infty}^{\infty} f(x)\, dx = \lim_{r\to\infty} \int_{-r}^{r} f(x)\, dx$.

12.99 Suppose that $f_n\colon U \to \mathbb{C}$ are analytic and converge locally uniformly to $f\colon U \to \mathbb{C}$. Suppose further that all zeros of each $f_n$ are real. Show that all zeros of $f$ are real.

**Open Mapping and Inverse Function Theorems**

12.100 (a) Prove the *maximum modulus principle*: if $f\colon U \to \mathbb{C}$ is analytic, and $\overline{B}(a, r) \subset U$, then $\max\{|f(z)| : z \in \overline{B}(a, r)\}$ is obtained on the boundary $\overline{B}(a, r) \setminus B(a, r)$. (Hint: the image of $B(a, r)$ is open.) (b) Conclude that if $f\colon U \to \mathbb{C}$ is analytic on a region $U$, and $|f|$ attains a minimum on $U$, then that minimum is 0. (c) Use this to give an alternative proof of the fundamental theorem of algebra 2.30.

12.101 Let $U = B(0, 1)$ be the interior of the unit circle. Let $A, B \subset U$ be finite; let $f\colon (U \setminus A) \to (U \setminus B)$ be an analytic bijection. Show that $A$ and $B$ have the same number of points, and that $f$ can be extended to an analytic bijection $\hat{f}\colon U \to U$ mapping $A$ to $B$. (Hint: $f$ is bounded around every $a \in A$.)

**Holomorphic Surfaces and Meromorphic Functions**

12.102 [PC] Find the Laurent series expansion of $f(z) = (z+4)/(z^2(z^2+3z+2))$ around 0. (Hint: partial fractions.)

12.103 Let $X$ be a Riemann surface. Show that if $f: X \to \mathbb{C}$ is a nonconstant holomorphic map then $|f|$ does not have a maximum.

12.104 (a) Viewing $\mathbb{C}$ as a subset of $\mathbb{P}^1(\mathbb{C})$ via $\rho_0$, show that every polynomial function $f: \mathbb{C} \to \mathbb{C}$ can be extended to a holomorphic function from $\mathbb{P}^1(\mathbb{C})$ to itself. (b) Show that the degree of the resulting holomorphic map (also named $f$) is the degree of the polynomial, and equals the valency of $f$ at $p_\infty$. (c) Show that every meromorphic function on $\mathbb{P}^1(\mathbb{C})$ is defined by a rational function (the ratio of two polynomial functions). (Hint: show that removing finitely many principal parts results in a holomorphic map from $\mathbb{P}^1(\mathbb{C})$ to $\mathbb{C}$.) (d) Conclude that every biholomorphism of $\mathbb{P}^1(\mathbb{C})$ with itself is a change of coordinates of $\mathbb{P}^1(\mathbb{C})$.

12.105 Show that every biholomorphism $f: \mathbb{C} \to \mathbb{C}$ is affine, i.e. of the form $z \mapsto az + b$ for some $a, b \in \mathbb{C}$, $a \neq 0$. (Use the Casorati-Weierstrass theorem (Exercise 12.94) to show that $f$ can be extended to the Riemann sphere.)

12.106 Let $f(z) = (z-1)^3/(z^2+1)$. Compute the degree of the extension of $f$ to a meromorphic function on $\mathbb{P}^1(\mathbb{C})$.

12.107 Let $f: X \to Y$ be holomorphic and let $m$ be the valency of $f$ at $p \in X$. Show that there are charts $\psi$ compatible with $X$ and $\varphi$ compatible with $Y$ such that the coordinate representation $\varphi \circ f \circ \psi^{-1}$ is the restriction of $z \mapsto z^m$ to range $\psi$. (Hint: if $g\restriction_V$ is never zero and $V$ is small then $g$ has an $m$th root on $V$.)

12.108 Show that if $f$ is meromorphic on $\mathbb{C}$ and has a unique simple pole $b$ (a pole of order 1), then $f$ is of the form $z \mapsto a + c/(z-b)$ for some $a, c \in \mathbb{C}$.

12.109 Use Proposition 12.43 to give another proof of the fundamental theorem of algebra. (Extend a polynomial function $f: \mathbb{C} \to \mathbb{C}$ to a function from the Riemann sphere to itself, see Exercise 12.104).

12.110 Use Proposition 12.43 to give another proof of Liouville's Theorem 11.76. (Hint: use Proposition 12.16).

**Analytic Continuation**

12.111 Let $f(z) = \sum_n z^{2^n}$. (a) Show that the radius of convergence of this power series is 1. (b) Suppose that $g$ is an analytic function on a connected domain extending $f$. Show that $g(z) = z + g(z^2)$, $g(z) = z + z^2 + g(z^4)$, $g(z) = z + z^2 + z^4 + g(z^8)$, and so on. (c) Show that if $w \in S$ is an $n$th root on unity

for some $n \geqslant 1$ (i.e., $w^n = 1$), then $g(w)$ is undefined. (d) Conclude that $g = f$. That is, $f$ cannot be extended to any analytic function on a larger domain.[5]

In the following exercises, we give an alternative definition of analytic continuation, which is perhaps a little more "conceptual". The price to pay is a restriction on the paths we can take. Call a path $\gamma : [a, b] \to X$ *prompt* if it is not constant on any sub-interval of $[a, b]$.[6]

12.112 Let $X$ and $Y$ be holomorphic surfaces, $U \subseteq X$ be open, $f : U \to Y$ be holomorphic, and $\gamma : [a, b] \to X$ be a prompt path with $\gamma(a) \in U$. Show that $g : [a, b] \to Y$ is an analytic continuation of $f$ along $\gamma$ if and only if for every $s \in [a, b]$ there is some open $V_s \subseteq X$ and holomorphic $h_s : V_s \to Y$ such that $\gamma(s) \in V_s$, and $g = h_s \circ \gamma$ on some neighbourhood of $s$ in $[a, b]$; and such that $V_a \subseteq U$ and $h_a = f$ on $V_a$. (In the harder direction, observe that for any subinterval $I$ of $[a, b]$, since $\gamma$ is not constant on $I$, $\gamma[I]$ is not discrete (it is a connected set that contains more than one point).)

12.113 Suppose that $\gamma : [a, b] \to X$ is constant on $[a, c]$ (where $a < c < b$), so is not prompt. Suppose that $h : X \to Y$ is a holomorphic function satisfying $h(\gamma(a)) = f(\gamma(a))$. Show that there is a function $g$ satisfying the condition of Exercise 12.112 satisfying $g(b) = h(\gamma(b))$. (That is, if $\gamma$ is not prompt, then the condition of Exercise 12.112 is too weak, in that it allows "continuations" $g$ that have nothing to do with $f$, other than that they agree with $f$ on the starting point.).

12.114 Give a proof of Lemma 12.68 using the characterisation of Exercise 12.112, assuming that $\gamma$ is prompt. (Let $r = \inf\{s \in [a, b] : g(s) \neq h(s)\}$.)

12.115 The following exercise allows us to use the definition of Exercise 12.112 and the proof from Exercise 12.114 to prove the Monodromy Theorem. Let $M$ be a manifold. Show that: (a) If $M$ is connected, then any two points in $M$ are connected by a prompt path. (b) If $M$ is simply connected, then any two prompt paths with the same end-points are homotopic by a homotopy $H$ for which every intermediate path $H_t$ is prompt. (One way to do (a) is to take a path $\gamma$ in $M$, break it up into a finite concatenation of paths $\gamma\restriction_J$, each of which has an image $\gamma[J]$ which is contained in the domain of chart $\psi$ for $M$, indeed such that $\psi[\gamma[J]]$ is contained in an open ball within the range of $\psi$; we then replace $\gamma\restriction_J$ by the pull-back by $\psi$ of a linear path between the end-points of $\psi \circ \gamma\restriction_J$. Similar massaging gives (b).)

[5] We say that the unit circle is the *natural boundary* of $f$.

[6] This is not standard terminology. The path is "prompt" because it doesn't take any rests on the way from $\gamma(a)$ to $\gamma(b)$. Promptness is implied by a common definition of smoothness of a path, which requires $\dot{\gamma} \neq \vec{0}$ at every point.

**Meromorphic Forms**

12.116 Let $\alpha\colon \mathbb{P}^1(\mathbb{C}) \to \mathbb{P}^1(\mathbb{C})$ be the map $(a:b) \mapsto (b:a)$. Find the zeros and poles, and their orders, of $d\alpha$. Do the same for the form $\alpha\, d\alpha$.

12.117 Show that the only holomorphic form on $\mathbb{P}^1(\mathbb{C})$ is the constant 0 form. (Hint: such a form is $f\, dg$ where $g$ is the identity on the sphere, see Example 12.87.)

12.118 Define a form $\omega$ on the Riemann surface $\Sigma$ to be the collection of pairs $(dz, \psi)$, where $\psi$ is a chart for $\Sigma$. (a) Show that $\omega$ is a non-vanishing holomorphic form on $\Sigma$. (b) Let $f\colon \mathbb{C} \to \Sigma$ be the biholomorphism $z \mapsto (e^z, \Im z)$. What is $f^*\omega$? (c) What is $(f^{-1})^* dz$? (d) Do the same for $\Sigma/n$ and the biholomorphism $\mathrm{pwr}_n$.

12.119 Let $\omega$ be the form on a complex torus $T = T_\Gamma$ defined in Exercise 12.88. (a) Let $\alpha$ be one of the generators of $\Gamma$; let $\gamma\colon [0,1] \to \mathbb{C}$ be the path $t \mapsto t\alpha$. Calculate $\int_{\pi\circ\gamma} \omega$. (b) Let $U \subseteq T$ be open and suppose that $(\pi \circ \gamma)[0,1] \subset U$. Show that there is no holomorphic function $g\colon U \to \mathbb{C}$ such that $\omega = dg$ on $U$.

12.120 (a) Show that if $\omega$ is a meromorphic form on a holomorphic surface $X$, then the *residue* of $\omega$ at a point $p \in X$ is well-defined. (That is, if $\omega_i = f_i\, dz$ then the residue of $f_i$ at $\psi_i(p)$ is the same for all $i$ with $p \in \operatorname{dom}\psi_i$. To do this, consider $1/(2\pi i)\int_\gamma \omega$ for a suitable $\gamma$.) (b) Show that if $f$ is meromorphic on $X$ and $p \in X$, then the residue of $df/f$ at $p$ is $\mathrm{ord}_p(f)$.

**Continuity of Roots**

We give two purely topological proofs of Proposition 12.29, not relying on the calculus of residues (or any complex analysis). The proofs use the symmetric power of a quasi-Euclidean space, which is a topological space not immediately seen to be quasi-Euclidean; and so goes a little beyond the tools developed in this book.

12.121 For a quasi-Euclidean space $X$ and $n \geqslant 2$, we let $\mathrm{SP}^n(X)$ be the collection of all multisets of points of $X$ of size $n$. Let $\mu\colon X^n \to \mathrm{SP}^n(X)$ be the "order-forgetting" map: $\mu(x_1, x_2, \ldots, x_n) = [x_1, x_2, \ldots, x_n]$.[7] We put the *quotient topology* on $\mathrm{SP}^n(X)$: a set $U \subseteq \mathrm{SP}^n(X)$ is open if $\mu^{-1}[U]$ is an open subset of $X^n$.

Show that for any space $Y$, a function $f\colon \mathrm{SP}^n(X) \to Y$ is continuous if and only if $f \circ \mu\colon X^n \to Y$ is continuous. (Compare with Proposition 8.109).

12.122 Suppose that $d$ is a metric on $X$ inducing the topology on $X$. Define the following: for $a = [a_1, \ldots, a_n]$, $b = [b_1, \ldots, b_n] \in \mathrm{SP}^n(X)$,

[7] $\mathrm{SP}^n(X)$ is the quotient $X^n/S_n$ where the symmetric group $S_n$ acts on $X^n$ by permuting coordinates.

$$d'(a,b) = \min_{\sigma \in S_n} \max_{i \leqslant n} d(a_i, b_{\sigma(i)}).$$

Show that $d'$ is a metric on $\mathrm{SP}^n(X)$ inducing the topology defined above.

For the following, for $d \geqslant 1$ let $P_d$ be the collection of monic polynomials $f \in \mathbb{C}[x]$ of degree $d$. Using the bijection $(a_{d-1}, \ldots, a_0) \mapsto x^d + a_{d-1}x^{d-1} + \cdots + a_1 x + a_0$, we consider $P_d$ as a topological space homeomorphic to $\mathbb{C}^d$.

12.123 Define the following map $\sigma : \mathrm{SP}^d(\mathbb{C}) \to P_d$ by letting $\sigma([a_1, \ldots, a_n]) = (x - a_1)(x - a_2)\cdots(x - a_n)$. (a) Show that $\sigma$ is a continuous bijection. (b) Show that Proposition 12.29 would follow from $\sigma^{-1}$ being continuous.

12.124

(a) Show that for a bounded set $Q \subseteq P_d$, $(\mu \circ \sigma)^{-1}[Q]$ is bounded. (Hint: show that there is some $R$ such that for all $f \in Q$ and $z \in \mathbb{C}$, $|z| > R$ implies $|f(z)| \geqslant |z|^d/2$.)
(b) Conclude that for every open ball $B$ in $P_d$, $\sigma^{-1}[B]$ is contained in a compact set.
(c) Conclude that $\sigma$ is a homeomorphism. (For every compact $K$, $\sigma \restriction_K$ is a homeomorphism.)

12.125 We give another proof of the continuity of $\sigma^{-1}$, due to Cucker and Gonzalez Corbalan [CGC89].

Let $H_d$ be the collection of homogeneous polynomials $f \in \mathbb{C}[w, x]$ of degree $d$. Just like $P_d$, we topologically identify $H_d$ with $\mathbb{C}^{d+1} \setminus \{\mathbf{0}\}$ by mapping $(a_d, a_{d-1}, \ldots, a_0)$ to $a_d x^d + a_{d-1}x^{d-1}w + \cdots + a_0 w^d$. If we define $f \sim g$ when $f = \lambda g$ for some nonzero $\lambda \in \mathbb{C}$, we obtain an identification of $H_d/\sim$ with $\mathbb{P}^d(\mathbb{C})$. For $f \in H_d$ we let $[f] = [f]_\sim$ be the $\sim$-equivalence class of $f$, and we let $[H_d] = H_d/\sim$.

(a) Show that $f \mapsto [f^\sharp]$ is a topological embedding of $P_d$ into $[H_d]$. We thus identify $P_d$ with the image of this embedding.
(b) Show that the identification $\mathbb{C} \subset \mathbb{P}^1(\mathbb{C})$ extends to an identification of $\mathrm{SP}^d(\mathbb{C})$ as a subspace of $\mathrm{SP}^d(\mathbb{P}^1(\mathbb{C}))$.
(c) Show that if $X$ is compact, so is $\mathrm{SP}^d(X)$.
(d) Define $\tau : H_d \to \mathrm{SP}^d(\mathbb{P}^1(\mathbb{C}))$ to be the map $f \mapsto V_{\mathbb{P}^1}(f)$. Show that the map $\tau$ induces a bijection $[\tau]$ from $[H_d]$ to $\mathrm{SP}^d(\mathbb{P}^1(\mathbb{C}))$.
(e) Show that under the previous identifications, $[\tau]^{-1} \restriction_{\mathrm{SP}^d(\mathbb{C})}$ is $\sigma$ of Exercise 12.123.
(f) Show that $[\tau]^{-1}$ is continuous. Conclude that as $\mathrm{SP}^d(\mathbb{P}^1(\mathbb{C}))$ is compact, $\tau$ is a homeomorphism; conclude that $\sigma$ is a homeomorphism as well.

# Part III

# Curves and Surfaces


# Curves Are Surfaces

# 13

The main result of this chapter is the construction of an atlas which makes a given nonsingular curve in $\mathbb{P}^2(\mathbb{C})$ a Riemann surface (Proposition 13.18). The main tool used is the analytic implicit function theorem (Theorem 13.5). The idea is as follows. Working in affine coordinates, let $D$ be a curve defined by an equation $f(x, y) = 0$; and let $p \in D$. Suppose that $p$ is nonsingular on $D$, and further, that the tangent to $D$ at $p$ is not vertical. Then the implicit function theorem says that on a neighbourhood of $p$, we can express a solution $y$ to the equation $f(x, y) = 0$ as an analytic function of $x$. We will then let the projection $(x, y) \mapsto x$ be a chart for $D$ on a neighbourhood of $p$. Since the inverse of this chart is the pair $z \mapsto (z, g(z))$ with $g$ analytic, the transition functions will be analytic, so we will get a holomorphic surface.

After defining the holomorphic structure of a curve, we revisit intersection numbers of lines and complex curves. We show that we can recover the original intuition for intersection multiplicity: a line $\ell$ intersects a curve $D$ at a point $p$ with multiplicity $m$ if lines close to $\ell$ intersect $D$ close to $p$ at $m$ points (see Proposition 13.47 and Exercise 13.76). To make sense of this, we copy over the topological and holomorphic structures from $\mathbb{P}^2(\mathbb{C})$ to the dual projective plane $\check{\mathbb{P}}^2(\mathbb{C})$.

For much more on Riemann surfaces and algebraic curves, see, for example, [Mir95].

## 13.1 The Implicit Function Theorem

Multivariable complex analysis is a fascinating subject. Many of the results from Chap. 11 generalise to higher dimensions. For example, functions $f : \mathbb{C}^n \to \mathbb{C}^m$ are continuously differentiable if and only if they are analytic, that is, the sum of a multi-variable power series; Cauchy's integral formula generalises as well. A decent treatment of this area is beyond the scope of this book; see, for example, [Gun90,



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_13

Sha92]. Here, we do the bare minimum in order to obtain the implicit function theorem.

The key, as in our development of single-variable complex analysis, is to view functions as multivariable real functions. We identify $\mathbb{C}^n$ with $\mathbb{R}^{2n}$ via $(z_1, z_2, \ldots, z_n) \mapsto (x_1, y_1, x_2, y_2, \ldots, x_n, y_n)$, where $z_i = x_i + iy_i$. If $U \subseteq \mathbb{C}^n$, then a function $f\colon U \to \mathbb{C}^m$ is identified with the corresponding function from $U \subseteq \mathbb{R}^{2n} \to \mathbb{R}^{2m}$. For a linear function $T_A\colon \mathbb{C}^n \to \mathbb{C}^m$ (where $A \in M_{m\times n}(\mathbb{C})$), the corresponding map from $\mathbb{R}^{2n}$ to $\mathbb{R}^{2m}$ is real-linear, defined by the matrix $M_A \in M_{2m\times 2n}(\mathbb{R})$, obtained by replacing each entry $a_{i,j}$ in $A$ by the $2 \times 2$-matrix $M_{a_{i,j}}$ from Lemma 11.1.

**Definition 13.1** Let $U \subseteq \mathbb{C}^n$ be open and let $\boldsymbol{a} \in U$. A function $f\colon U \to \mathbb{C}^m$ is differentiable at $\boldsymbol{a}$ if there is some $A \in M_{m\times n}(\mathbb{C})$ such that for all $\varepsilon > 0$ there is $\delta > 0$ such that for all $\boldsymbol{h} \in \mathbb{C}^n$, if $|\boldsymbol{h}| < \delta$ then $|(f(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a})) - A\boldsymbol{h}| < \varepsilon|\boldsymbol{h}|$.

To differentiate from the real notation, we write $A = f'(\boldsymbol{a})$. Note that $|\boldsymbol{z}|$ is the same if we think of $\boldsymbol{z}$ as an element of $\mathbb{R}^{2n}$ or $\mathbb{C}^n$. Hence $f\colon U \to \mathbb{C}^m$ is differentiable at $\boldsymbol{a}$, with derivative $A = f'(\boldsymbol{a})$, if and only if it is differentiable at $\boldsymbol{a}$ when we think of it as a map from $U$ to $\mathbb{R}^{2m}$, with derivative $Df(\boldsymbol{a}) = M_A$. So $f$ is complex-differentiable at $\boldsymbol{a}$ if and only if it is real-differentiable, and the Cauchy-Riemann Equations hold for each of the $2 \times 2$-sub-matrices of $Df(\boldsymbol{a})$: if $f = (f_1, \ldots, f_m)$ with $f_j\colon U \to \mathbb{C}$, and we write $f_j = f_{j,x} + if_{j,y}$, and the real variables are $(x_1, y_1, \ldots, x_n, y_n)$, then the equations are $D^{x_i} f_{j,x} = D^{y_i} f_{j,y}$ and $D^{y_i} f_{j,x} = -D^{x_i} f_{j,y}$ (at $\boldsymbol{a}$) for all $j \leqslant m$ and $i \leqslant n$.

We can therefore lift theorems from the real realm to the complex one. For example, if $U \subseteq \mathbb{C}^n$ and $f = (f_1, f_2, \ldots, f_m)\colon U \to \mathbb{C}^m$, then $U$ is differentiable at a point $\boldsymbol{a}$ if and only if each $f_j$ is differentiable at $\boldsymbol{a}$. We can similarly define partial derivatives: if $f\colon U \to \mathbb{C}$ and $i \leqslant n$ then $D^{z_i} f$ is the partial derivative in the $i$th complex direction; Proposition 9.52 implies that $f$ is continuously differentiable on $U$ if and only if each $D^{z_i} f$ is defined and continuous on $U$.

If $A \in M_{m\times n}(\mathbb{C})$ and $B \in M_{k\times m}(\mathbb{C})$ then $M_{BA} = M_B M_A$, and so the real chain rule (Proposition 9.42) implies the complex one:

**Proposition 13.2** *Let $U \subseteq \mathbb{C}^n$, $V \subseteq \mathbb{C}^m$, $f\colon U \to V$, $g\colon V \to \mathbb{C}^k$, and $\boldsymbol{a} \in U$. Suppose that $f$ is differentiable at $\boldsymbol{a}$ and that $g$ is differentiable at $f(\boldsymbol{a})$. Then $g \circ f$ is differentiable at $\boldsymbol{a}$ and $(g \circ f)'(\boldsymbol{a}) = g'(f(\boldsymbol{a})) \cdot f'(\boldsymbol{a})$.*

*Exercise 13.3* Show that every rational function on $\mathbb{C}^n$ is continuously differentiable. That is, if $f, g \in \mathbb{C}[z_1, \ldots, z_n]$, then the rational function defined by $f/g$ on $\mathbb{C}^n \setminus V_{\mathbb{A}^n(\mathbb{C})}(g)$ is continuously differentiable, with derivative $(f'g - fg')/g^2$. «

A matrix $A \in M_n(\mathbb{C})$ is invertible if and only if the corresponding real matrix $M_A$ is invertible. Hence the real Inverse Function Theorem implies the complex one:

**Theorem 13.4 (Multivariable Complex Inverse Function Theorem)** *Let $U \subseteq \mathbb{C}^n$ be open, let $f: U \to \mathbb{C}^n$ be continuously differentiable, let $\boldsymbol{a} \in U$ and suppose that $f'(\boldsymbol{a})$ is invertible. Then there is an open neighbourhood $V \subseteq U$ of $\boldsymbol{a}$ such that $f[V]$ is open, the restriction $f\restriction_V$ is a homeomorphism between $V$ and $f[V]$, and its inverse $g = (f\restriction_V)^{-1}$ is differentiable at $f(\boldsymbol{a})$, with $g'(f(\boldsymbol{a})) = (f'(\boldsymbol{a}))^{-1}$.*

As in the real context, the implicit function theorem is a consequence of the inverse function theorem. To simplify notation, we only state the case that we will be using, which is for an implicit definition in two variables. Let $W \subseteq \mathbb{C}^2$ and let $f: W \to \mathbb{C}$; we write $(z, w)$ for the (complex) variables of $f$, and so write $f' = (D^z f, D^w f)$.

**Theorem 13.5** *Let $W \subseteq \mathbb{C}^2$ be open; let $f: W \to \mathbb{C}$ be continuously differentiable, let $(a, b) \in W$ and suppose that $f(a, b) = 0$ but $D^w f(a, b) \neq 0$. Then there are open neighbourhoods $U \subseteq \mathbb{C}$ of $a$ and $V \subseteq \mathbb{C}$ of $b$ and an analytic function $g: U \to V$ such that for all $z \in U$, $g(z)$ is the unique $w \in V$ such that $f(z, w) = 0$.*

***Proof*** Let $F(z, w) = (z, f(z, w))$. Then $F$ is continuously differentiable on $W$. Noting that

$$F' = \begin{pmatrix} 1 & 0 \\ D^z f & D^w f \end{pmatrix},$$

we see that $F'(a, b)$ is invertible. By Theorem 13.4, and since $D^w f$ is continuous, by shrinking $W$ we may assume that $F$ is 1–1 on $W$, $F[W]$ is open, and $D^w f \neq 0$ on $W$. The inverse $G$ of $F$ is differentiable on $F[W]$; write $G = (G_z, G_w)$ (where $G_z(z, w) = z$ on $F[W]$). Since both $W$ and $F[W]$ are open and $F(a, b) = (a, 0)$, we can find open neighbourhoods $U$ of $a$ and $V$ of $b$ such that: (i) $U \times V \subseteq W$; and (ii) $U \times \{0\} \subset F[W]$ (that is, $(z, 0) \in F[W]$ for all $z \in U$).

For $z \in U$ let $g(z) = G_w(z, 0)$. Since $G_w$ is continuously differentiable, by the chain rule, so is $g$, and so $g$ is analytic; by definition of $F$, $f(z, g(z)) = 0$ for all $z \in U$. And since $F$ is 1–1 on $W$, for all $z \in U$ there is at most one $w \in V$ with $f(z, w) = 0$, so $g(z)$ is the unique such $w$. □

*Remark 13.6* Since $G'(F(\boldsymbol{p})) = (F'(\boldsymbol{p}))^{-1}$, a calculation shows that $g'(z) = -D^z f(z, g(z))/D^w f(z, g(z))$. This can also be deduced by the chain rule; see Proposition 13.12 below. «

For a completely different proof of the implicit function theorem, see Exercise 15.108.

## 13.2 Nonsingular Curves Are Riemann Surfaces

Armed with the implicit function theorem, we now see how nonsingular complex algebraic curves are Riemann surfaces.

### 13.2.1 Vertical Parameterisations

Let $D$ be an irreducible algebraic curve in $\mathbb{P}^2(\mathbb{C})$. Since $\mathbb{P}^2(\mathbb{C})$ is a manifold (Example 8.20), we can take the subspace topology for $D$: an open subset of $D$ is one of the form $U \cap D$, where $U$ is an open subset of $\mathbb{P}^2(\mathbb{C})$. We can extend this to reducible curves if we ignore repeated points, that is, if we take the underlying set of the curve, so that we get a subset of $\mathbb{P}^2(\mathbb{C})$. Since we are only interested in the underlying set, from now on, we ignore curves which have repeated components.

*Exercise 13.7* Show that every algebraic curve in $\mathbb{P}^2(\mathbb{C})$ is compact. «

We will work in affine coordinates; recall that $(a, b) \in \mathbb{A}^2(\mathbb{C})$ is identified via $\rho_0$ with $(1 : a : b) \in \mathbb{P}^2(\mathbb{C})$. The map $\rho_0$ is a homeomorphism between $\mathbb{A}^2(\mathbb{C})$ and its range $U_0$; its inverse is a chart for $\mathbb{P}^2(\mathbb{C})$.

*Notation 13.8* Let $D$ be a curve in $\mathbb{P}^2$. We let $D^*$ denote the collection of nonsingular points of $D$. «

Like $D$, we consider $D^*$ as a topological subspace of $\mathbb{P}^2$, and so can talk about neighbourhoods of points in $D^*$ and continuous functions to and from $D^*$. Since we are assuming that $D$ has no repeated components, $D \setminus D^*$ is finite (Corollary 5.41). Hence, $D^*$ is an open subset of $D$. So a neighbourhood of a point in $D^*$ is also a neighbourhood of that point in $D$.

**Definition 13.9** A *vertical parameterisation* of $D$ is a homeomorphism of the form $\eta(z) = (z, \eta_y(z))$ from an open subset of $\mathbb{C}$ to an open subset of $D^* \cap \mathbb{A}^2(\mathbb{C})$, where $\eta_y$ is analytic.[1]

*Example 13.10* Let $D$ be the complex unit circle $x^2 + y^2 = w^2$. There is an analytic square root defined on a neighbourhood of 1 (see Remark 12.59), so we can get a vertical parameterisation $z \mapsto (z, \sqrt{1 - z^2})$ of $D$. The two choices of square root will give two vertical parameterisations, one mapping 0 to $(0, 1)$ and the other to $(0, -1)$. «

[1] To be precise, we should say "local analytic vertical parameterisation", local since it is parameterising only part of the curve; but this is a bit of a mouthful.

Let $f$ be a polynomial which defines $D \cap \mathbb{A}^2$ (a dehomogenisation of a polynomial defining $D$), and let $p = (a, b)$ be a point on $D$. If $p$ is nonsingular on $D$ then the affine tangent to $D$ at $p$ is given by $D^x f(p) \cdot (x - a) + D^y f(p) \cdot (y - b) = 0$ (Eq. (5.3) on page 106, see Exercise 5.17); in particular, the tangent is not vertical if and only if $D^y f(p) \neq 0$.

*Remark 13.11* In the previous section, we used $D^z f$ and $D^w f$ to denote partial derivatives of $f : \mathbb{C}^2 \to \mathbb{C}$; this was to avoid confusion with the real partial derivatives $D^{x_i} f$ and $D^{y_i} f$. Henceforth, we do not need to consider the real derivatives, and so we return to the notation $D^x f$ and $D^y f$ for $f \in \mathbb{C}[x, y]$; as a function on $\mathbb{C}^2$, of course, these are $D^z f$ and $D^w f$. «

**Proposition 13.12** *Suppose that $\eta : U \to V$ is a vertical parameterisation of $D$. Then for all $p = \eta(a) \in V$, the tangent to $D$ at $p$ is not vertical, and its slope is $\eta'_y(a)$.*

***Proof*** Again let $f$ be a polynomial which defines $D \cap \mathbb{A}^2$. By definition, each $p \in V$ is nonsingular on $D$. Say $p = \eta(a)$. By the chain rule (Lemma 13.2), as $f \circ \eta = 0$, $D^x f(p) + \eta'_y(a) \cdot D^y f(p) = 0$; since $(D^x f(p), D^y f(p)) \neq (0, 0)$, we must have $D^y f(p) \neq 0$.

By the equation for the affine tangent (Eq. 5.3) mentioned, we get that the slope of the tangent to $D$ at $p$ is $-D^x f(p)/D^y f(p) = \eta'_y(a)$. □

A neighbourhood of $p = (a, b)$ on $D$ is one which contains $(O \times W) \cap D$, where $O$ is a neighbourhood of $a$ and $W$ is a neighbourhood of $b$ in $\mathbb{C}$. We can thus restate the implicit function theorem:

**Proposition 13.13** *If $D$ is a curve in $\mathbb{P}^2(\mathbb{C})$, $p \in D^* \cap \mathbb{A}^2$ and the tangent to $D$ at $p$ is not vertical, then there is a vertical parameterisation $\eta : U \to V$ of $D$ with $p \in V$.*

*Example 13.14* Let $o = (0, 0)$ be the origin; suppose that $\eta$ is a vertical parameterisation of $D$ with $\eta(0) = o$. Let $f$ define $D \cap \mathbb{A}^2$. There is a power series $\sum c_n z^n$ which defines $\eta_y$ on a neighbourhood of 0. Since $f(z, \eta_y(z)) = 0$, the uniqueness of formal power series (Proposition 11.58) allows us to compute the coefficients $c_n$ recursively.

For example, let $f(x, y) = y^2 - y - x$. Since $\eta_y(0) = 0$, we have $c_0 = 0$. We substitute and get

$$f(z, \eta_y(z)) = -(1+c_1)z + (c_1^2 - c_2)z^2 + (2c_1c_2 - c_3)z^3 + (2c_1c_3 + c_2^2 - c_4)z^4 + \cdots;$$

equating all coefficients to 0, we recursively find that $c_1 = -1$, $c_2 = 1$, $c_3 = -2$, $c_4 = 5, \ldots$ «

The uniqueness part of the implicit function theorem means:

**Proposition 13.15** *Let $\eta_1$ and $\eta_2$ be two vertical parameterisations of a curve $D$. Let $a \in \operatorname{dom}\eta_1 \cap \operatorname{dom}\eta_2$, and suppose that $\eta_1(a) = \eta_2(a)$. Then $\eta_1 = \eta_2$ on a neighbourhood of $a$.*

*Exercise 13.16* Let $D$ be a curve in $\mathbb{P}^2(\mathbb{C})$; let $p = (a_0, b_0) \in D^* \cap \mathbb{A}^2$, and suppose that the tangent to $D$ at $p$ is not vertical. Suppose that $\eta = (\eta_x, \eta_y) \colon U \to D \cap \mathbb{A}^2$ is continuous, where $U$ is a neighbourhood of $a_0$ in $\mathbb{C}$; and suppose that for all $a \in U$, $\eta_x(a) = a$, and that $\eta(a_0) = p$. Show that restricted to some open neighbourhood of $a_0$, $\eta$ is a vertical parameterisation of $D$. (Thus, in Definition 13.9, it is enough to require that $\eta_y$ be continuous, rather than analytic.) «

### 13.2.2 An Atlas for the Nonsingular Part of a Curve

Let $D$ be an algebraic curve in $\mathbb{P}^2(\mathbb{C})$. We will now define an atlas on $D^*$ which makes it a holomorphic surface. The plan is to take the inverse of a vertical parameterisation of $D$ as a chart for $D^*$. By definition, every chart will be a local homeomorphism with $\mathbb{C}$, so the transition functions will be continuous; and we will check that they are analytic. However, not every point of $D^*$ is in the domain of such a chart: we are possibly missing points at infinity, and points at which the tangent is vertical. So we allow changes of coordinates to put points in the right position.

**Definition 13.17** Let $D$ be a curve in $\mathbb{P}^2(\mathbb{C})$. We let $\mathcal{A}_D$ be the collection of maps of the form $\eta^{-1} \circ \alpha$, where $\alpha$ is a change of coordinates of $\mathbb{P}^2(\mathbb{C})$ and $\eta$ is a vertical parameterisation of $\alpha[D]$.

**Proposition 13.18** *Let $D$ be a curve in $\mathbb{P}^2(\mathbb{C})$. The collection $\mathcal{A}_D$ is an atlas for $D^*$, and $(D^*, \mathcal{A}_D)$ is a holomorphic surface, which is a topological subspace of $\mathbb{P}^2(\mathbb{C})$.*

***Proof*** By definition, the range of a vertical parameterisation of a curve $C$ is contained in $C^*$, and is a homeomorphism between an open subset of $\mathbb{C}$ and an open subset of $C^*$. Since changes of coordinates of $\mathbb{P}^2(\mathbb{C})$ are continuous (Exercise 8.60), and map nonsingular points to nonsingular points (Proposition 5.21), it follows that each chart in $\mathcal{A}_D$ is a homeomorphism between an open subset of $D^*$ and an open subset of $\mathbb{C}$.

Any $p \in D^*$ can be moved by a change of coordinates to $\mathbb{A}^2$ with a non-vertical tangent, and so by Proposition 13.13, every point of $D^*$ lies in the domain of some chart in $\mathcal{A}_D$. By Proposition 8.61, $\mathcal{A}_D$ is an atlas for $D^*$, And $(D^*, \mathcal{A}_D)$ is a 2-manifold which is a topological subspace of $\mathbb{P}^2(\mathbb{C})$.

It remains to show that the transition maps are analytic. Let $\psi = \eta^{-1} \circ \alpha$ and $\varphi = \zeta^{-1} \circ \beta$ be two charts in $\mathcal{A}_D$. There are linear homogeneous polynomials $k_w, k_x, k_y \in \mathbb{C}[w, x, y]$ such that

$$(\beta \circ \alpha^{-1})(e:a:b) = (k_w(e,a,b) : k_x(e,a,b) : k_y(e,a,b))$$

for every point $(e:a:b) \in \mathbb{P}^2$. Let $z$ be in the domain of the transition map $\varphi \circ \psi^{-1}$. Then

$$\big(\beta \circ \psi^{-1}\big)(z) = (\beta \circ \alpha^{-1})(1:z:\eta_y(z)) = \\ \big(k_w(1,z,\eta_y(z)) : k_x(1,z,\eta_y(z)) : k_y(1,z,\eta_y(z))\big)$$

(recall that $\eta_y$ is the second coordinate of $\eta$); since this is a point in the range of $\zeta$, it is in $\mathbb{A}^2$, i.e., $k_w(1, z, \eta_y(z)) \neq 0$, so

$$(\varphi \circ \psi^{-1})(z) = (\zeta^{-1} \circ \beta \circ \psi^{-1})(z) = \frac{k_x(1,z,\eta_y(z))}{k_w(1,z,\eta_y(z))}$$

which is an analytic function of $z$. □

*Exercise 13.19* Let $D$ be a curve in $\mathbb{P}^2(\mathbb{C})$, and let $\alpha$ be a change of coordinates of $\mathbb{P}^2(\mathbb{C})$. Show that $\alpha\restriction_{D^*}$ is a biholomorphism from $D^*$ to $\alpha[D^*]$. «

*Remark 13.20* Since a vertical parameterisation of $D$ is the inverse of a chart for $D^*$, it follows that it is a biholomorphism between an open subset of $\mathbb{C}$ and an open subset of $D^*$. «

*Example 13.21* Proposition 13.18 in particular implies that every line in $\mathbb{P}^2(\mathbb{C})$ is a holomorphic surface. In fact, a line in $\mathbb{P}^2(\mathbb{C})$ is biholomorphic with the Riemann sphere $\mathbb{P}^1(\mathbb{C})$ (and so in particular is connected, i.e., is a Riemann surface). By Exercise 13.19 it suffices to show this for one line, say the $x$-axis.

The projective linear parameterisation $(s:t) \mapsto (s:t:0)$ is a bijection between $\mathbb{P}^1(\mathbb{C})$ and the $x$-axis (see Example 4.17). This map is a biholomorphism. By Proposition 12.37, it suffices to show that it is holomorphic. We check two coordinate representations. The map $\eta(z) = (z, 0)$ is a vertical parameterisation of the $x$-axis. For one coordinate representation we choose the chart $(\rho_0)^{-1}$ for $\mathbb{P}^1(\mathbb{C})$ and the chart $\eta^{-1}$ for the $x$-axis; this gives the coordinate representation $z \mapsto (1:z:0) = (z, 0) \mapsto z$, i.e. the identity map on $\mathbb{C}$. The second coordinate representation is obtained by choosing the chart $(\rho_1)^{-1}$ for $\mathbb{P}^1(\mathbb{C})$, and the chart $\eta^{-1} \circ \alpha$ where $\alpha$ is the change of coordinates $\alpha(e:a:b) = (a:e:b)$, which maps the $x$-axis to itself; again the coordinate representation is the identity map. «

### 13.2.3 Rational Functions on Curves

Suppose that $f, g \in \mathbb{C}[w, x, y]$ are homogeneous of the same degree. Then $f$ and $g$ define a function $F\colon \mathbb{P}^2(\mathbb{C}) \setminus V_{\mathbb{P}^2}(g) \to \mathbb{C}$ by letting

$$F(e:a:b) = \frac{f(e,a,b)}{g(e,a,b)};$$

because $\deg f = \deg g$, this is well-defined, in that it does not depend on the choice of particular presentation $(e, a, b)$ of the point $(e:a:b)$. We call $F$ a *rational function*. It is defined on all points outside the curve $g = 0$.

*Exercise 13.22* Show that a rational function is continuous. «

**Proposition 13.23** *Let $F = f/g$ be a rational function, let $D$ be a curve in $\mathbb{P}^2(\mathbb{C})$, and suppose that $D$ and $V_{\mathbb{P}^2}(g)$ have no common component. Then $F\restriction_{D^*}$ extends to a meromorphic function on $D^*$.*

***Proof*** Let $\psi$ be a chart for $D^*$, let $U = \operatorname{range}\psi$. There are analytic functions $h_w, h_x$ and $h_y$ on $U$ such that $\psi^{-1}(z) = (h_w(z):h_x(z):h_y(z))$ for all $z \in U$. Thus, the coordinate representation of $F$ using the chart $\psi$ is the map

$$z \mapsto \frac{f(h_w(z), h_x(z), h_y(z))}{g(h_w(z), h_x(z), h_y(z))}.$$

This is the quotient of two analytic functions. Further, since $D \cap V_{\mathbb{P}^2}(g)$ is finite, $g(h_w, h_x, h_y)$ is not constant zero on $U$, and so this quotient is meromorphic on $U$. □

*Example 13.24* The simplest example are the coordinate maps, for example the map $x/w$ taking $(e:a:b)$ to $a/e$, defined on $\mathbb{A}^2(\mathbb{C})$; in affine coordinates, this is the projection $(a, b) \mapsto a$. «

*Remark 13.25* By taking differentials (see Exercise 12.83) this gives us a way of defining meromorphic differentials on algebraic curves. «

*Exercise 13.26* Let $D$ be a curve in $\mathbb{P}^2$. Let $X$ be a holomorphic surface, and let $\psi = (\psi_x, \psi_y)\colon X \to D^* \cap \mathbb{A}^2$ be a continuous map. Show that $\psi$ is holomorphic (as a map to $D^*$) if and only if both $\psi_x$ and $\psi_y$ are analytic. «

*Exercise 13.27* Let $D$ be a nonsingular cubic curve given by $y^2 = f$ (see Proposition 7.23). (a) Show that the restriction of the rational function $x/w$ to $D$ is a meromorphic function of degree 2. What are the zeros, the poles and their orders? (b) Show that the restriction of the rational function $y/w$ to $D$ is a meromorphic function of degree 3. What are the zeros, the poles and their orders? «

### 13.2.4 Lifting Paths to Curves

Toward showing that (some) algebraic curves are connected, we set up machinery that will also serve us in Chap. 15. Let $D$ be an algebraic curve in $\mathbb{P}^2(\mathbb{C})$. Recall that $o_p(D)$ is the order of $p$ on $D$ (Definition 5.13), and that $(0:0:1)$ is the vertical point at infinity.

**Definition 13.28** We say that $a \in \mathbb{C}$ is a *ramification point* of $D$ if the affine line $x = a$ intersects $D$ in fewer than $\deg D - o_{(0:0:1)}(D)$-many distinct points.

We let $\mathfrak{R} = \mathfrak{R}(D)$ be the collection of ramification points of $D$. Since we are assuming that $D$ has no repeated components, Corollary 5.39 implies:

**Proposition 13.29** *$D$ has only finitely many ramification points.*

Let $a \in \mathbb{K}$. By Theorem 5.34, $i_{(0:0:1)}(D, x = aw) \geqslant o_{(0:0:1)}(D)$ (the projective line $x = wa$ intersects $D$ at the vertical point at infinity with multiplicity at least the order of that point on $D$). Hence by Theorem 5.27, if $a$ is not a ramification point of $D$, then $i_{(0:0:1)}(D, x = aw) = o_{(0:0:1)}(D)$ and for every $p \in D \cap (x = a)$ (a point on $D$ which lies on the *affine* line $x = a$), $i_p(D, x = a) = 1$. Now Theorem 5.34 implies:

**Proposition 13.30** *If $a$ is not a ramification point of $D$, and $p \in D \cap (x = a)$, then $p$ is nonsingular on $D$, and the tangent to $D$ at $p$ is not vertical.*

*Exercise 13.31* Find the ramification points of the projective closures of the following conic curves: (i) $y = x^2$; (ii) $x = y^2$; (iii) $xy = 1$; (iv) $x^2 + y^2 = 1$. «

By Proposition 13.13, for every $a \notin \mathfrak{R}$ and every $p \in D \cap (x = a)$, there is a vertical parameterisation $\eta$ of $D$ on a neighbourhood of $a$ satisfying $\eta(a) = p$. By intersecting the domains of these parameterisations and shrinking their ranges, we obtain the following.

**Lemma 13.32** *We can choose, for every $a \in \mathbb{C} \setminus \mathfrak{R}$, an open disc $U_a$ centered at $a$, disjoint from $\mathfrak{R}$, and for every $p \in D \cap (x = a)$, a vertical parameterisation $\eta_p : U_a \to V_p$ of $D$ satisfying $\eta_p(a) = p$; further, we may assume that if $p, p' \in D \cap (x = a)$ are distinct, then $V_p$ and $V_{p'}$ are disjoint.*

Having fixed such a collection of parameterisations $\eta_p$, we observe the following.

**Lemma 13.33** *Suppose that $a, a' \in \mathbb{C} \setminus \mathfrak{R}$, and that $U_a \cap U_{a'} \neq \varnothing$. Then for every $p \in D \cap (x = a)$ there is a unique $p' \in D \cap (x = a')$ such that $\eta_p$ and $\eta_{p'}$ agree on $U_a \cap U_{a'}$. The map $p \mapsto p'$ is a bijection between $D \cap (x = a)$ and $D \cap (x = a')$.*

***Proof*** Let $\hat{a} \in U_a$. For $p \in D \cap (x = a)$ let $\hat{p} = \eta_p(\hat{a})$ (recall that the range of $\eta_p$ is contained in $D \cap \mathbb{A}^2$). By Proposition 13.15, $\eta_{\hat{p}} = \eta_p$ on a neighbourhood of $\hat{a}$. If $p_1, p_2 \in D \cap (x = a)$ are distinct, then since $V_{p_1}$ and $V_{p_2}$ are disjoint, we must have $\hat{p}_1 \neq \hat{p}_2$. By the definition of ramification points, the number of points in $D \cap (x = a)$ and $D \cap (x = \hat{a})$ is the same. Hence the map $p \mapsto \hat{p}$ is a bijection between $D \cap (x = a)$ and $D \cap (x = \hat{a})$.

As $U_a \cap U_{a'} \neq \varnothing$, choose $\hat{a} \in U_a \cap U_{a'}$, and apply this argument for both $p$ and $p'$. Combining bijections, we get a bijection $p \mapsto p'$ between $D \cap (x = a)$ and $D \cap (x = a')$, with $\eta_p = \eta_{p'}$ on a neighbourhood of $\hat{a}$. Since $U_a \cap U_{a'}$ is connected (it is the intersection of two discs), we have $\eta_p = \eta_{p'}$ on $U_a \cap U_{a'}$. □

Recall the notion of analytic continuation (Definition 12.66). By Remark 13.20, a vertical parameterisation is holomorphic (as a map to $D^*$), and so we can speak of analytic continuations of vertical parameterisations.

**Proposition 13.34** *If $\gamma$ is a path in $\mathbb{C} \setminus \mathfrak{R}$ starting at some $a$, and $\eta$ is a vertical parameterisation of $D$ with $a \in \operatorname{dom} \eta$, then there is an analytic continuation of $\eta$ along $\gamma$.*

***Proof*** This is similar to the argument of Proposition 9.20. Let $p = \eta(a)$; then by Lemma 13.15, $\eta = \eta_p$ on a neighbourhood of $a$. Say $\gamma \colon I \to \mathbb{C} \setminus \mathfrak{R}$. By compactness of $I$, we obtain a partition $t_0 < t_1 < \cdots < t_k$ of $I$ such that for all $i = 1, \ldots, k$, there is some $a_i \in \mathbb{C} \setminus \mathfrak{R}$ such that $\gamma[t_{i-1}, t_i] \subset U_{a_i}$; we may take $a_1 = a$. We recursively choose points $p_i \in D \cap (x = a_i)$. We start with $p_1 = p$. If $p_i$ was already chosen (and $i < k$), then since $\gamma(t_i) \in U_{a_i} \cap U_{a_{i+1}}$, by Lemma 13.33 there is a unique $p_{i+1} \in D \cap (x = a_{i+1})$ such that $\eta_{p_{i+1}} = \eta_{p_i}$ on $U_{a_i} \cap U_{a_{i+1}}$ (and hence on a neighbourhood of $\gamma(t_i)$). This partition, along with the functions $\eta_{p_1}, \eta_{p_2}, \ldots, \eta_{p_k}$, show that $\xi(t) = \eta_{p_i}(\gamma(t))$ for $t \in [t_{i-1}, t_i]$ is an analytic continuation of $\eta_p$ along $\gamma$. □

Let $\gamma \colon I \to \mathbb{C}$ be a path. A *lifting* of $\gamma$ to $D$ is a path $\xi \colon I \to D$ such that for all $t \in I$, $\xi(t)$ lies on the projective line $x = \gamma(t)w$. If $\xi$ maps into $\mathbb{A}^2$, this means that $\gamma$ is the projection of $\xi$ onto the first coordinate; but in general, we also allow the vertical point at infinity to be a value of $\xi$ (if it lies on $D$). The analytic continuation $\eta$ constructed in the proof of Proposition 13.34 is a lifting of $\gamma$ to $D$, starting at $p$, with image entirely within $\mathbb{A}^2$. (This is not a coincidence; see Exercise 13.66.) As a result we get:

**Corollary 13.35** *If $\gamma$ is a path in $\mathbb{C} \setminus \mathfrak{R}$ starting at some $a$, and $p \in D \cap (x = a)$, then there is a lifting of $\gamma$ to a path in $D \cap \mathbb{A}^2$ which starts at $p$.*

To prove the connectedness of curves, we need to consider bad end-points.

**Proposition 13.36** *Let $\gamma$ be a path in $\mathbb{C}$ which avoids all ramification points of $D$, except for its end-point. Let $p \in D \cap (x = a)$, where $\gamma$ starts at $a$. Then $\gamma$ has a lifting to a path in $D$ starting at $p$.*

Alas, to prove this proposition, we need a result which fits better in the next section, so we postpone the proof until a little later; see page 361.

### Algebraic Curves Are Connected

To show that nonsingular algebraic curves are Riemann surfaces, it remains to show that they are connected. This is true; in fact, every algebraic curve in $\mathbb{P}^2(\mathbb{C})$, whether singular or not, is connected. This is not so easy to prove (see for example [Ken11, Thm.4.2] or [Gri89, Thm.II.2.11]). We will present a special case, which makes the proof easier. This case includes nonsingular cubics.

**Theorem 13.37** *Let $D$ be a nonsingular curve in $\mathbb{P}^2(\mathbb{C})$. Suppose that there is some line $\ell$ which intersects $D$ in a single point. Then $D$ is connected.*

***Proof*** Let $\ell$ be a line which intersects $D$ at a unique point $p^*$. We show that for any point $q \in D$, there is a path in $D$ from $q$ to $p^*$; this will show that $D$ is path-connected. Fix a point $q \in D$. By assumption, $q$ is nonsingular on $D$. By Corollary 5.40, there is a line $L$ passing through $q$, which does not pass through $p^*$, and which intersects $D$ at deg $D$-many distinct points. The point of intersection $\ell \cap L$ is not $p^*$, so is not on $D$. We change coordinates by moving $\ell \cap L$ to the vertical point at infinity. After this change of coordinates, $L$, of course, is a vertical line $x = aw$, and $a$ is not a ramification point of (the new) $D$. Also, $\ell$ was moved to a vertical line $x = a^*w$. Since the set of ramification points is finite, we can find a path in $\mathbb{C}$ from $a$ to $a^*$, which avoids all ramification points of $D$ other than the end-point $a^*$. By Proposition 13.36, there is a lifting $\xi$ of $\gamma$ to $D$ starting at $q$. The end-point of $\xi$ is some point in $D \cap (x = a^*w)$; but by assumption, $p^*$ is the only point in $D \cap (x = a^*w)$, so $\xi$ is a path from $q$ to $p^*$. □

**Corollary 13.38** *Every nonsingular cubic curve in $\mathbb{P}^2(\mathbb{C})$ is connected.*

***Proof*** A nonsingular cubic curve has flexes (Proposition 7.10), and the tangent to the curve at a flex $p$ intersects the curve only at $p$. □

*Remark 13.39* The proof of Theorem 13.37 indicates why it is harder to prove the general statement. If $x = a^*$ intersects $D$ in more than one point, then we have no control over what point above $a^*$ the path $\eta$ reaches. The missing part of the proof, then, is showing that if $p$, $p'$ are two points in $D \cap (x = a^*)$, where $a^*$ is a ramification point of $D$, then there is a path in $D$ from $p$ to $p'$. «

## 13.3 Intersections with Lines, Revisited

Recall that the original idea behind the multiplicity of intersection of a curve and a line at some point is that $i_p(D, \ell) = m$ if when we "move the line $\ell$ a little", then "near $p$", the line and the curve intersect (simply) in $m$ distinct points. Working over $\mathbb{C}$, rather than a general field, allows us to formalise this intuition.

Recall (see Sect. 4.6) that $\check{\mathbb{P}}^2$, the dual projective plane, is the collection of lines in $\mathbb{P}^2$. We used the bijection $\iota\colon \mathbb{P}^2 \to \check{\mathbb{P}}^2$, taking $(e:a:b)$ to the line $ew+ax+by = 0$, to give $\check{\mathbb{P}}^2$ the structure of a projective plane; for example, the curves of $\check{\mathbb{P}}^2$ were the images under $\iota$ of curves in $\mathbb{P}^2$, in particular, a linear family of lines $\mathcal{L}$ is the image under $\iota$ of a line in $\mathbb{P}^2$; see Proposition 4.40. Now that we have a topological structure on $\mathbb{P}^2(\mathbb{C})$ and a holomorphic structure on curves in $\mathbb{P}^2(\mathbb{C})$, we can use $\iota$ in exactly the same way to obtain such structures in the dual plane. So:

- $\check{\mathbb{P}}^2(\mathbb{C})$ is a smooth 4-manifold; a chart for $\check{\mathbb{P}}^2(\mathbb{C})$ is a map of the form $\varphi \circ \iota^{-1}$, where $\varphi$ is a chart for $\mathbb{P}^2(\mathbb{C})$.
- Every linear family of lines $\mathcal{L}$ in $\check{\mathbb{P}}^2(\mathbb{C})$ is a Riemann surface, biholomorphic with the Riemann sphere (see Example 13.21); a chart for $\mathcal{L}$ is a map of the form $\psi \circ \iota^{-1}$, where $\psi$ is a chart for the line $\ell$ satisfying $\iota[\ell] = \mathcal{L}$.

We can thus speak of open neighbourhoods of lines in $\check{\mathbb{P}}^2(\mathbb{C})$; and of holomorphic functions to and from linear families of lines.

*Remark 13.40* A change of coordinates $\check{\alpha}$ of $\check{\mathbb{P}}^2(\mathbb{C})$ is a homeomorphism from $\check{\mathbb{P}}^2(\mathbb{C})$ to itself (Exercise 8.60); for any linear family of lines $\mathcal{L}$, the restriction of $\check{\alpha}$ to $\mathcal{L}$ is a biholomorphism from $\mathcal{L}$ to $\check{\alpha}[\mathcal{L}]$ (Exercise 13.19). Note that by Exercise 4.42, changes of coordinates of $\check{\mathbb{P}}^2$ are maps of the form $\ell \mapsto \alpha[\ell]$, where $\alpha$ is a change of coordinates of $\mathbb{P}^2$. «

*Example 13.41* The $x$-axis $y = 0$ is $\iota(0:0:1)$. Using the chart $(\rho_2)^{-1}$ for $\mathbb{P}^2(\mathbb{C})$ (taking $(e:a:1)$ to $(e,a)$), we see that a neighbourhood of the $x$-axis in $\check{\mathbb{P}}^2(\mathbb{C})$ is one which contains the collection of lines $ax + bw + y = 0$ for all $|a|, |b| < \varepsilon$ (for some $\varepsilon > 0$). Exchanging $(a, b)$ with $(-a, -b)$, the restrictions of these lines to $\mathbb{A}^2$ are the lines $y = ax + b$, where $|a|, |b| < \varepsilon$. That is, the lines close to the $x$-axis are the lines whose slope is close to 0, and which intersect the $y$-axis close to the origin. «

*Example 13.42* Let $\mathcal{L}$ be the family of vertical lines: the line at infinity, together with the projective closures of the lines $x = a$ for all $a \in \mathbb{C}$. It is the image under $\iota$ of the $x$-axis. One chart for the $x$-axis (considered as a Riemann surface) is $(1:a:0) \mapsto a$. As in Example 13.41, replacing $a$ by $-a$, we see that one chart for the family of vertical lines is the map taking the line $x = a$ to $a$. Thus, a neighbourhood of a line $x = a_0$ in $\mathcal{L}$ contains a collection of lines $x = a$ where $|a - a_0| < \varepsilon$, for some $\varepsilon > 0$. «

*Exercise 13.43* Show that another chart for the family of vertical lines is the map taking the line at infinity to 0, and lines $x = a$ (for $a \neq 0$) to $1/a$. Conclude that a neighbourhood of the line at infinity $\ell_\infty$ in the family of vertical lines is one which contains $\ell_\infty$, and all lines $x = a$ for $|a| > M$, for some $M > 0$. «

*Exercise 13.44* What are charts for the family of lines which pass through the origin? What are neighbourhoods of the $x$-axis in this family of lines? Of the $y$-axis? «

For disjoint sets $U, V \subset \mathbb{P}^2$, let $\overline{UV} = \{\overline{pq} : p \in U \ \& \ q \in V\}$ be the collection of all lines passing through a point from $U$ and a point from $V$.

*Exercise 13.45* Let $L$ be a line in $\mathbb{P}^2(\mathbb{C})$ and let $p, q \in L$ be distinct points. Show that a set $O \subseteq \check{\mathbb{P}}^2(\mathbb{C})$ is a neighbourhood of $L$ if and only if there are disjoint neighbourhoods $U_p$ and $U_q$ of $p$ and $q$ in $\mathbb{P}^2(\mathbb{C})$ such that $\overline{U_p U_q} \subseteq O$. (To simplify calculations, by Remark 13.40, we may change coordinates so that $L$ is the $x$-axis, $p$ is the origin and $q = (1, 0)$. Note that you need to show two things: (a) $\overline{U_p U_q}$ is open, for any choice of disjoint open $U_p$ and $U_q$; and (b) If $O$ is a neighbourhood of $L$, then $\overline{U_p U_q} \subseteq O$ for sufficiently small $U_p$ and $U_q$.) «

### 13.3.1 Continuous Intersection Multiplicities

**Proposition 13.46** *Let $D$ be an algebraic curve in $\mathbb{P}^2(\mathbb{C})$, let $L$ be a line, let $p \in L$, and let $m = i_p(D, L)$ be the multiplicity of intersection of $D$ and $L$ at $p$.*

*For every neighbourhood $\hat{U}$ of $p$ in $D$ there is an open neighbourhood $U \subseteq \hat{U}$ of $p$ in $D$ and a neighbourhood $O$ of $L$ in $\check{\mathbb{P}}^2(\mathbb{C})$ such that for every $\ell \in O$,*

$$m = \sum_{r \in U} i_r(D, \ell).$$

That is, all lines $\ell$ close to $L$ intersect $D$ close to $p$ in $m$ points, multiplicities counted.

***Proof*** By Proposition 5.29, Exercise 8.60, and Remark 13.40, we may change coordinates so that $p$ is the origin, $L$ is the $x$-axis, and $\ell_\infty$ is not a component of $D$.

Fix a polynomial $f \in \mathbb{C}[x, y]$ which defines $D \cap \mathbb{A}^2$. For each $a$ and $b$, let $\psi_{a,b}(t) = (t, at + b)$ be the affine linear parameterisation of the line $y = ax + b$ (Definition 3.27). Let $f_{a,b} = f_{\psi_{a,b}} = f(t, at + b)$ be the resulting intersection polynomial; Lemma 5.32 says that for a point $q = (\lambda, a\lambda + b)$, the intersection multiplicity $i_q(D, y = ax + b)$ at $q$ of the curve $D$ and the line $y = ax + b$, is the multiplicity of $\lambda$ as a root of $f_{a,b}$. In particular, $m$ is the multiplicity of 0 as a root of $f_{0,0}$.

Since $D$ is a topological subspace of $\mathbb{P}^2(\mathbb{C})$, there is some open neighbourhood $\hat{V}$ of the origin in $\mathbb{A}^2(\mathbb{C})$ such that $\hat{U} = \hat{V} \cap D$.

The map $(\lambda, a, b) \mapsto (\lambda, a\lambda + b)$ is continuous, so there is some $\varepsilon > 0$ such that $(\lambda, a\lambda + b) \in \hat{V}$ whenever $|\lambda|, |a|, |b| < \varepsilon$.

The map taking $(a, b)$ to the coefficients of $f_{a,b}$ is continuous. So by Proposition 12.29, there is a neighbourhood $W \subseteq B(0, \varepsilon)$ of 0 in $\mathbb{C}$ and some $\delta > 0$ such that whenever $|a|, |b| < \delta$, the polynomial $f_{a,b}$ has $m$ roots in $W$, counting multiplicities. We may choose $\delta \leqslant \varepsilon$. We let $O$ be the collection of lines $y = ax + b$ for $|a|, |b| < \delta$; by Example 13.41, this is a neighbourhood of the $x$-axis in $\check{\mathbb{P}}^2(\mathbb{C})$.

Let

$$V = \{(\lambda, a\lambda + b) : \lambda \in W \ \& \ |a|, |b| < \delta\}.$$

So $V \subseteq \hat{V}$, so $U = V \cap D \subseteq \hat{U}$, and $U$ is as required, once we observe that $V$ is an open subset of $\mathbb{A}^2$. Abstractly, this follows from checking the rank of the full derivative of $(\lambda, a, b) \mapsto (\lambda, a\lambda + b)$. More concretely, fixing $a_0$, the map $(\lambda, b) \mapsto (\lambda, a_0\lambda + b)$ is obviously a homeomorphism from $\mathbb{C}^2$ to itself (it is an injective affine map), and this shows that $(\lambda, a, b) \mapsto (\lambda, a\lambda + b)$ is an open map (maps open sets to open sets). □

To characterise intersection multiplicity using topology, we would like to say that $m = i_p(D, L)$ if a "generic line" $\ell$ close to $L$ intersects $D$ close to $p$ in $m$ distinct points. The question is, what do we mean by a generic line? Unfortunately, the statement is not true if by "generic" we mean "all but finitely many", as any open neighbourhood of $L$ may contain tangents at points close to $p$; or $p$ may be singular on $D$. There is a way to make the original intuition precise; see Exercise 13.76. Also see Exercise 15.116, when $L$ is not a tangent to $D$ at $p$. The following is an approximation.

**Proposition 13.47** *Let $D$ be a curve in $\mathbb{P}^2(\mathbb{C})$, let $L$ be a line, and let $p \in D \cap L$. Let $\mathcal{L}$ be a linear family of lines which contains $L$, but which isn't the family of lines passing through $p$. Then there is a neighbourhood $O$ of $L$ in $\mathcal{L}$ and an open neighbourhood $U$ of $p$ in $D$ such that every line $\ell \in O \setminus \{L\}$ intersects $U$ in $i_p(D, L)$-many distinct points.*

***Proof*** Let $q$ be the point such that $\mathcal{L}$ is the family of lines passing through $q$. Let $\hat{U}$ be a neighbourhood of $p$ in $D$ such that $q \notin \hat{U}$. By Proposition 13.46, find neighbourhoods $U \subset \hat{U}$ of $p$ in $D$ and $\hat{O}$ of $L$ in $\check{\mathbb{P}}^2$ such that every $\ell \in \hat{O}$ intersects $U$ in $i_p(D, L)$ many points, multiplicities counted. Let $O = \hat{O} \cap \mathcal{L}$, which is an open neighbourhood of $L$ in $\mathcal{L}$. By Lemma 5.38, since $q \notin U$, for all but finitely many $\ell \in \mathcal{L}$, for all $r \in U$, $i_r(D, \ell) \leqslant 1$. By shrinking $O$ we may avoid the finitely many exceptions, except for $L$ itself. □

In another application, Proposition 13.46 allows us to settle a debt: the proof of Proposition 13.36. We need a lemma.

**Lemma 13.48** *Let $D$ be a curve in $\mathbb{P}^2$, and let $L$ be a line. Let $p_1, \dots, p_k$ be the points of intersection of $D$ and $L$. Let $V_1, \dots, V_k$ be open neighbourhoods of $p_1, \dots, p_k$ in $D$. Then there is a neighbourhood $O$ of $L$ in $\check{\mathbb{P}}^2$ such that for all $\ell \in O$, $\ell \cap D \subseteq \bigcup_j V_j$.*

***Proof*** For $j = 1, \dots, k$, let $m_j = i_{p_j}(D, L)$. By Theorem 5.27, $\sum_j m_j = \deg D$. By Proposition 13.46 we can find neighbourhoods $U_j \subseteq V_j$ of $p_j$ in $D$ and a neighbourhood $O$ of $L$ in $\check{\mathbb{P}}^2$ such that every $\ell \in O$ meets each $V_j$ in $m_j$ points (counted with multiplicities). By Proposition 13.46 again, all points of intersection of $\ell \in O$ with $D$ must be in $\bigcup_j U_j$. □

***Proof of Proposition 13.36*** We are given a path $\gamma$ in $\mathbb{C}$, say with domain $[0, 1]$, from $a = \gamma(0)$ to $b = \gamma(1)$, so that $\gamma(t)$ is not a ramification point of $D$, for all $t \neq 1$. We define a lifting $\xi$ of $\gamma$ to $D$, starting at $p$.

We break $[0, 1)$ into countably many consecutive closed intervals: let $I_n = [1 - 1/n, 1 - 1/(n+1)]$; let $a_n = \gamma(1 - 1/n)$. Recursively, we define $\xi\restriction_{I_n}$. We start with $p_1 = p$. Then, given $p_n = \xi(1 - 1/n) \in D \cap (x = a_n)$, we use Corollary 13.35 to define $\xi\restriction_{I_n}$, starting at $p_n$; we then let $p_{n+1}$ be the end-point of $\xi\restriction_{I_n}$, and note that $p_{n+1} \in \mathbb{A}^2$.

This process defines $\xi\restriction_{[0,1)}$. It remains to show that $\xi$ reaches a limit at 1: there is some $q \in D \cap (x = bw)$ such that $q = \lim_{t \to 1} \xi(t)$, that is, every neighbourhood $V$ of $q$ in $D$ contains $\xi(t)$ for all $t \in (1 - \varepsilon, 1)$ for some $\varepsilon > 0$.

List $D \cap (x = bw)$ as $q_1, q_2, \dots, q_k$. Suppose that $V_1, V_2, \dots, V_k$ are pairwise disjoint neighbourhoods of $q_1, \dots, q_k$ in $D$. By Lemma 13.48, after intersecting with the family of vertical lines, and by Example 13.42, there is some open neighbourhood $W \subseteq \mathbb{C}$ of $b$ such that for all $c \in W$, $D \cap (x = cw) \subseteq \bigcup_j V_j$. Since $\gamma$ is continuous, there is some $\varepsilon > 0$ such that $\gamma[1 - \varepsilon, 1] \subseteq W$. Thus, $\xi[1-\varepsilon, 1) \subseteq \bigcup_j V_j$. Since the sets $V_j$ are pairwise disjoint and the interval $[1-\varepsilon, 1)$ is connected, there is some $j^*$ such that $\xi[1 - \varepsilon, 1) \subseteq V_{j^*}$.

By taking intersections, we observe that the value $j^*$ cannot vary with the choice of pairwise disjoint neighbourhoods $V_j$. Let $q^* = q_{j^*}$. If $V$ is any neighbourhood of $q^*$ in $D$, then we can choose $V_{j^*} = V$ (and choose any $V_j$ for $j \neq j^*$, as long as they are disjoint from $V$); then this analysis shows that a tail of $\xi\restriction_{[0,1)}$ lies in $V$. Thus, setting $\xi(1) = q^*$ makes $\xi$ continuous, and we get the required lifting. □

### 13.3.2 Finding Intersection Points

Suppose that a line $L$ intersects a curve $D$ simply at a point $p$, that is, $i_p(D, L) = 1$. Proposition 13.46 implies that lines $\ell$ close to $L$ have a unique point of intersection with $D$ near $p$.

**Proposition 13.49** *If $i_p(D, L) = 1$ then there is an open neighbourhood $U$ of $p$ in $D$ and an open neighbourhood $O$ of $L$ in $\check{\mathbb{P}}^2(\mathbb{C})$ such that every $\ell \in O$ intersects $U$ in a unique point, and the map taking $\ell$ to this unique intersection point is a continuous function from $O$ to $U$.*

***Proof*** Let $U$ and $O$ be given by Proposition 13.46 when supplied with $D$, $L$ and $p$ (and $\hat{U} = D$). Let $\ell_0 \in O$, let $p_0$ be the point of intersection of $\ell_0$ with $U$. Let $U_0 \subseteq U$ be a neighbourhood of $p_0$. Applying Proposition 13.46 to $D$, $\ell_0$, $p_0$ and $\hat{U} = U_0$, we obtain a neighbourhood $O_0$ of $\ell_0$ such that $\ell \cap D \in U_0$ for all $\ell \in O_0$. So the map taking $\ell$ to $\ell \cap U$ is continuous at $\ell_0$. □

Because we did not properly develop multivariable complex analysis, and so did not define complex manifolds of higher dimensions, we cannot define what it means for a function on an open subset of $\check{\mathbb{P}}^2(\mathbb{C})$ to be holomorphic. But linear families of lines are Riemann surfaces, so we can restrict to those.

Let $\mathcal{L}$ be a linear family of lines; let $D$ be a curve in $\mathbb{P}^2(\mathbb{C})$, let $L \in \mathcal{L}$, and let $p \in L \cap D$. Suppose that $i_p(D, L) = 1$. Then $p \in D^*$ (it is nonsingular on $D$). By intersecting with $\mathcal{L}$, Proposition 13.49 gives us an open neighbourhood $U$ of $p$ in $D$ and an open neighbourhood $O$ of $L$ in $\mathcal{L}$ such that $\ell \mapsto \ell \cap U$ is well-defined and continuous on $O$. By shrinking $U$ and $O$, we may assume that $U \subseteq D^*$. Hence both $U$ and $O$ are holomorphic surfaces, so we can ask if the map $\ell \mapsto \ell \cap U$ is holomorphic on $O$.

The family $\mathcal{L}$ is the family of lines passing through some point $r$. There are two possibilities: either $p = r$, or not. If $p = r$ then every $\ell \in \mathcal{L}$ intersects $U$ at $p$, and so for $\ell \in O$, $p$ is the unique point of intersection $\ell \cap U$; in other words, the map $\ell \mapsto \ell \cap U$ is constant on $O$. In this case, it is certainly holomorphic.

Suppose then that $p \neq r$. We change coordinates (permitted by Exercise 13.19 and Remark 13.40) so that $\mathcal{L}$ is the family of vertical lines and $p$ is the origin. So $L$ is now the $y$-axis; since $i_p(D, L) = 1$, $L$ is not the tangent to $D$ at $p$. By Proposition 13.13, let $\eta\colon W \to V$ be a vertical parameterisation of $D$, where $W$ is an open neighbourhood of 0 in $\mathbb{C}$, $V$ is an open neighbourhood of $p$ in $D$, and $\eta(0) = p$. By shrinking, we may assume that $V \subseteq U$. The map taking $c$ to the line $x = cw$ is a biholomorphism between $W$ and an open neighbourhood of $L$ in $\mathcal{L}$ (Example 13.42). Composing, by shrinking $O$ (to be the image of $W$), we see that $\ell \mapsto \ell \cap V$ is a holomorphic bijection between $O$ and $V$ (hence, a biholomorphism). So we proved:

**Proposition 13.50** *Let $\mathcal{L}$ be a linear family of lines; let $D$ be a curve in $\mathbb{P}^2(\mathbb{C})$, let $L \in \mathcal{L}$, and let $p \in L \cap D$. Suppose that $i_p(D, L) = 1$. Then there is a neighbourhood $U$ of $p$ in $D^*$ and a neighbourhood $O$ of $L$ in $\mathcal{L}$ such that for all $\ell \in O$, $\ell \cap U$ contains a unique point, and the map taking $\ell \in O$ to the unique point $\ell \cap U$ is holomorphic. This map is either constant, or can be taken to be a holomorphic bijection between $O$ and $U$.*

### 13.3.3 Finding Intersecting Lines

We now go the other direction, and consider a function from points to lines. Suppose that $D$ is a nonsingular curve in $\mathbb{P}^2(\mathbb{C})$. Then for each $p, q \in D$, the line $\overline{p\,q}$ is well-defined: when $p = q$, we take $\overline{p\,p} = \ell_p D$, the tangent to $D$ at $p$. Thus $(p, q) \mapsto \overline{p\,q}$ is a well-defined function from $D^2$ to $\check{\mathbb{P}}^2(\mathbb{C})$. We show that it is continuous. Fixing a point $p$, we can also consider the map $q \mapsto \overline{p\,q}$; this is a function from $D$ to the linear family $\mathcal{L}$ of lines passing through $p$. We show that it is holomorphic.

**Proposition 13.51** *Let $D$ be a nonsingular curve in $\mathbb{P}^2(\mathbb{C})$. The function $(p, q) \mapsto \overline{p\,q}$ from $D^2$ to $\check{\mathbb{P}}^2(\mathbb{C})$ is continuous.*

***Proof*** We show that the map $(p, q) \mapsto \overline{p\,q}$ is continuous at every pair $(p_0, q_0)$ of $D^2$. There are two kinds of pairs: $p_0 = q_0$ and $p_0 \neq q_0$.

Suppose that $p_0 \neq q_0$. Let $O$ be a neighbourhood of $\overline{p_0\,q_0}$ in $\check{\mathbb{P}}^2(\mathbb{C})$. By Exercise 13.45, there are disjoint neighbourhoods $V$ of $p_0$ and $W$ of $q_0$ in $\mathbb{P}^2$ such that $\overline{V\,W} \subseteq O$. Since $D$ is a topological subspace of $\mathbb{P}^2(\mathbb{C})$, $V \cap D$ is a neighbourhood of $p_0$ in $D$, and $W \cap D$ is a neighbourhood of $q_0$ in $D$; so $V \times W$ is a neighbourhood of $(p_0, q_0)$ in $D^2$. For every pair $(p, q) \in V \times W$, the line $\overline{p\,q}$ is in $O$. Hence $(p, q) \mapsto \overline{p\,q}$ is continuous at $(p_0, q_0)$.

Now consider $(p_0, p_0)$. We change coordinates so that $p_0$ is the origin and $\overline{p_0\,p_0}$ is the $x$-axis. By Example 13.41, it suffices to show that there is a neighbourhood $U$ of $p_0$ in $D$ such that for $p, q \in U$, the line $\overline{p\,q}$ is (the projective closure of) $y = m(p, q)x + b(p, q)$, and that the functions $(p, q) \mapsto m(p, q)$ and $(p, q) \mapsto b(p, q)$ are both continuous on $U^2$.

Since $p_0$ is nonsingular on $D$ and the tangent $\overline{p_0\,p_0}$ to $D$ at $p_0$ is not the $y$-axis, by Proposition 13.13 let $\eta\colon W \to U$ be a vertical parameterisation of $D$ with $\eta(0) = p_0$.

For $p = \eta(a) = (a, \eta_y(a))$ and $q = \eta(b) \in U$, the slope $m(p, q)$ of the line $\overline{p\,q}$ is $(\eta_y(b) - \eta_y(a))/(b - a)$ if $p \neq q$, and $\eta_y'(a)$ if $p = q$, the latter by Proposition 13.12. By Exercise 11.10, the function $(p, q) \mapsto m(p, q)$ is continuous on $U^2$.

The function $p = \eta(a) \mapsto a$ is continuous on $U$, and $\eta_y$ is continuous, so $(p, q) \mapsto b(p, q) = \eta_y(a) - m(p, q)a$ is also a continuous function on $U^2$. □

**Proposition 13.52** *Let $D$ be a nonsingular algebraic curve in $\mathbb{P}^2(\mathbb{C})$, and let $p \in D$. Let $\mathcal{L}$ be the linear family of lines passing through $p$. The map from $D$ to $\mathcal{L}$ taking $q \in D$ to the line $\overline{p\,q}$ is holomorphic.*

***Proof*** Let $q_0 \in D$; let $\ell_0 = \overline{p\,q_0}$. Suppose that $i_{q_0}(D, \ell_0) = 1$. Then $q_0 \neq p$. By Proposition 13.50, there is an open neighbourhood $U$ of $q_0$ in $D$ and an open neighbourhood $O$ of $\ell_0$ in $\mathcal{L}$ such that $\ell \mapsto \ell \cap U$ is a biholomorphism from $O$ to $U$; we may assume that $p \notin U$. The inverse of this biholomorphism is the map $q \mapsto \overline{p\,q}$ on $U$. Hence the map $q \mapsto \overline{p\,q}$ is holomorphic on a neighbourhood of $q_0$.

By Proposition 5.38, there are only finitely many $q_0 \in D$ such that $i_{q_0}(D, \ell_0) > 1$. By Proposition 13.51, the function $q \mapsto \overline{p\,q}$ is continuous on all of $D$. Hence the proposition follows from Proposition 12.11. □

*Exercise 13.53* By finding coordinate representations, argue directly that $q \mapsto \overline{p\,q}$ is holomorphic on a neighbourhood of $q_0$ when $i_{q_0}(D, \ell_0) > 1$ (without appealing to Propositions 5.38 and 12.11). (Note that there are two cases: either $q_0 = p$, or not.) «

*Exercise 13.54* Check that Proposition 13.52 holds even if $p \notin D$. «

### 13.3.4 An Application to Elliptic Curves

In Chap. 7 we showed that if $D$ is a nonsingular cubic curve in $\mathbb{P}^2$ and $0_D$ is a flex of $D$, then the elliptic curve $(D, 0_D)$ is an abelian group under the operation $p + q = (p * q) * 0_D$, where $p * q$ is the third point of intersection of the line $\overline{p\,q}$ with $D$. Over the complex numbers, everything is continuous and holomorphic.

Recall that a *topological group* is one in which addition and taking inverses is continuous (see page 217).

**Proposition 13.55** *An elliptic curve $(D, 0_D)$ in $\mathbb{P}^2(\mathbb{C})$ is a topological group.*

***Proof*** Since $p + q = (p * q) * 0_D$ and $-p = p * 0_D$, it suffices to show that "star operation" $(p, q) \mapsto p * q$ from $D^2$ to $D$ is continuous.

Let $(p_0, q_0) \in D^2$, let $\ell_0 = \overline{p_0\,q_0}$, and let $r_0 = p_0 * q_0$. We want to show that $(p, q) \mapsto p * q$ is continuous at $(p_0, q_0)$. Let $W$ be a neighbourhood of $r_0$ in $D$. Recall that $\ell_0 \cdot D = [p_0, q_0, r_0]$ is a multiset: some of the points may coincide. By Proposition 13.46, we can choose, for each point $s$ among $p_0$, $q_0$ and $r_0$, a neighbourhood $U_s$ of $s$ in $D$, with $U_{r_0} \subseteq W$, and a neighbourhood $O$ of $\ell_0$ in $\check{\mathbb{P}}^2(\mathbb{C})$, so that each line $\ell \in O$ meets each $U_s$ in $i_s(D, \ell_0)$ many points (multiplicities allowed). We can ensure that if $s$ and $t$ are two *distinct* points among $p_0$, $q_0$ and $r_0$, then $U_s$ and $U_t$ are disjoint.

By Proposition 13.51, there is a neighbourhood $V$ of $(p_0, q_0)$ in $D^2$ such that $\overline{p\,q} \in O$ for all $(p, q) \in V$; we may assume that $V \subseteq U_{p_0} \times U_{q_0}$. Then for all $(p, q) \in V$ we have $p * q \in W$. (This argument is perhaps best understood if you assume that the points $p_0$, $q_0$ and $r_0$ are all distinct; then $U_{p_0}$, $U_{q_0}$ and $U_{r_0}$ are pairwise disjoint, and each $\ell \in O$ meets each one of these sets in exactly one point. Next, consider the other combinations: $p_0 = q_0 \neq r_0$, or $p_0 \neq q_0 = r_0$, or $p_0 = q_0 = r_0$.) □

*Remark 13.56* There are rational functions defining addition on an elliptic curve (Exercise 7.30). Rational functions are continuous, so one could hope to use this to prove Proposition 13.55. The issue though is that we used different rational functions for the two cases $q \neq p$, $q = p$, and so we do not get a single rational function defined on an open neighbourhood of $(p, p)$. It is possible to obtain such rational functions; see, for example, [Sil09, Thm.3.6]; see also [LR85]. «

Restricting to one complex dimension, we can work in the holomorphic category:

**Proposition 13.57** *Let $(D, 0_D)$ be an elliptic curve and let $p \in D$. The map $q \mapsto p + q$ is a biholomorphism from $D$ to itself.*

***Proof*** Since $(D, 0_D)$ is a group, the map $q \mapsto p+q$ is 1–1 and onto; so it suffices to show that it is holomorphic. As in the proof above, it suffices to show that $q \mapsto q*p$ is holomorphic.

Fix $q_0 \in D$. Let $\ell_0 = \overline{p\,q_0}$ and $r_0 = p * q_0$. Let $\mathcal{L}$ be the family of lines passing through $p$.

Suppose first that $q_0$, $p$ and $r_0$ are all distinct. As above we can fix disjoint neighbourhoods $U_{q_0}$, $U_p$ and $U_{r_0}$ of $q_0$, $p$ and $r_0$, and an open neighbourhood $O$ of $\ell_0$ in $\check{\mathbb{P}}^2(\mathbb{C})$, such that each $\ell \in O$ intersects each set $U_{q_0}$, $U_p$ and $U_{r_0}$ exactly once. By Proposition 13.50, we can ensure that the map $\ell \mapsto \ell \cap U_{r_0}$ is holomorphic on $O \cap \mathcal{L}$. By Proposition 13.52, by shrinking $U_{q_0}$, we may ensure that $\overline{p\,q} \in O$ for all $q \in U_{q_0}$, and that the map $q \mapsto \overline{p\,q}$ from $U_{q_0}$ to $O \cap \mathcal{L}$ is holomorphic. Then restricted to $U_{q_0}$, the map $q \mapsto q * p$ is the composition of two holomorphic functions, and so is holomorphic.

By Corollary 5.40, all but finitely many lines $\ell \in \mathcal{L}$ intersect $D$ at three distinct points. It follows that there are only finitely many points $q_0 \in D$ for which $\overline{p\,q_0}$ intersects $D$ at fewer than three points. By Proposition 13.55, $q \mapsto p * q$ is continuous on all of $D$. The result then follows from Proposition 12.11. □

*Exercise 13.58* Show that if $(D, 0_D)$ is an elliptic curve, then the map $p \mapsto -p$ is a biholomorphism from $D$ to itself. «

*Remark 13.59* Proposition 13.57 actually implies Proposition 13.55. Indeed, Hartogs' theorem says that if $f : \mathbb{C}^2 \to \mathbb{C}$ is separately analytic in the two coordinates (for all $a$, $z \mapsto f(z, a)$ is analytic, and $z \mapsto f(a, z)$ is analytic as well), then $f$ is

continuously differentiable. Note that the assumption is much weaker than assuming that $D^z f$ and $D^w f$ are continuous; we are just assuming that for each $a$, separately, $z \mapsto D^z f(z, a)$ is continuous, and the same for $D^w f$. «

## 13.4 Further Exercises

**The Implicit Function Theorem**

13.60 In this exercise, we give a simplified proof of the uniqueness part of the implicit function theorem (Theorem 13.5) when we replace the complex numbers by the reals. Let $U \subseteq \mathbb{R}^2$ be open and let $f : U \to \mathbb{R}$ be smooth; for simplicity suppose that $(0, 0) \in U$, that $f(0, 0) = 0$ and $D^y f(0, 0) \neq 0$. (a) Show that there is some open interval $J = (-\delta, \delta)$ and some $M > 0$ such that for all $a, b \in J$, either $D^y f(a, b) > M$ or $D^y f(a, b) < -M$. (b) Conclude that for all $a, b, c \in J$, $|f(a, c) - f(a, b)| > M \cdot |c - b|$. (c) Conclude that for all $a \in J$, the function $b \mapsto f(a, b)$ is injective on $J$.

13.61 In this exercise we give an algebraic proof of the uniqueness part of Theorem 13.5, when the function $f$ is polynomial. Let $f \in \mathbb{C}[x, y]$; suppose that $f(0, 0) = 0$ and that $D^y f(0, 0) \neq 0$. Define $h \in \mathbb{C}[x, y, t]$ by letting $h(x, y, t) = f(x, t) - f(x, y)$. (a) Show that the polynomial $t - y$ divides $h$. (b) Letting $g = h/(t - y)$, show that $D^y f = g + (y - t)D^y g$; conclude that $D^y f = g(x, y, y)$, so $g(0, 0, 0) \neq 0$. (c) Fix $\delta > 0$ such that $g(z, w, u) \neq 0$ whenever $|z|, |w|, |u| < \delta$. Let $V = U = B(0, \delta)$. Show that for all $a \in U$ there is at most one $b \in V$ such that $f(a, b) = 0$.

**Curves as Surfaces**

13.62 (a) Let $C$ be the projective closure of the complex parabola $y = x^2$. Show that the map $(e : a) \mapsto (e^2 : ea : a^2)$ (a rational parameterisation of the parabola) is a biholomorphism between the Riemann sphere $\mathbb{P}^1(\mathbb{C})$ and $C$. (b) Conclude that every irreducible conic in $\mathbb{P}^2(\mathbb{C})$ is biholomorphic with the Riemann sphere (see Exercise 5.60). (c) Show that the map extending $a \mapsto (a^2 + 1 : 2a : a^2 - 1)$ (Exercise 4.82) is a biholomorphism between the Riemann sphere and the projective closure of the complex unit circle $x^2 + y^2 = 1$.

13.63 Let $D$ be the projective closure of the cuspidal cubic $y^2 = x^3$ (see Exercises 3.47, 4.61 and 5.37 and Example 7.39) in $\mathbb{P}^2(\mathbb{C})$. (a) Show that the map $(e : a) \to (a^3 : e^2 a : e^3)$ is a homeomorphism from $\mathbb{P}^1(\mathbb{C})$ to $D$. (b) Show that $D^*$ is biholomorphic with $\mathbb{C}$. (c) What happens with the nodal cubic $y^2 = x^2 + x^3$ (Exercises 3.48, 4.61 and 5.63)?

13.64 Generalise Exercise 7.2 as follows. Let $f \in \mathbb{C}[x, y]$; suppose that $\deg f > 1$. Let $p$ be the origin, and suppose that $f(p) = 0$, that $p$ is nonsingular on the curve $f = 0$, and that the tangent to that curve at $p$ is not vertical. (a) Let $m$ be the slope of the tangent to $f = 0$ at $p$. Show that $p$ is a flex of $f = 0$ if and only if $D^{xx} f(p) + 2mD^{xy} f(p) + m^2 D^{yy} f(p) = 0$. (b) Suppose that $\eta$ is a vertical parameterisation of $f = 0$ with $\eta(0) = p$. Show that $p$ is a flex of $f = 0$ if and only if $\eta_y''(0) = 0$.

**Lifting Paths to Curves**

13.65 Let $D$ be a curve in $\mathbb{P}^2(\mathbb{C})$, and suppose that the vertical point at infinity $(0:0:1)$ lies on $D$. Let $a \in \mathbb{C}$. (a) Show that if $a \notin \Re(D)$ then the line $x = aw$ is not a tangent to $D$ at $(0:0:1)$. (b) Show that if $x = aw$ is not a tangent to $D$ at $(0:0:1)$ then there is a neighbourhood $W$ of $(0:0:1)$ in $D$ and a neighbourhood $U$ of $a$ in $\mathbb{C}$ such that for all $c \in U$, the line $x = cw$ intersects $W$ only at $(0:0:1)$.

13.66 Let $D$ be a curve in $\mathbb{P}^2(\mathbb{C})$, and let $\gamma : I \to \mathbb{C} \setminus \Re(D)$ be a path avoiding the ramification points of $D$, starting at some $a$; let $p \in D \cap (x = a)$. (a) Show that the range of any lifting of $\gamma$ to $D$ starting at $p$ is within $\mathbb{A}^2$. (b) Let $\eta$ be a vertical parameterisation of $D$, with $p \in \operatorname{range} \eta$. Show that a path $\xi : I \to D$ is a lifting of $\gamma$ to $D$ starting at $p$ if and only if it is an analytic continuation of $\eta$ along $\gamma$.

13.67 Let $D$ be the projective closure of $xy = 1$. What is the end-point of any lifting to $D$ of a path $\gamma$ ending at 0?

13.68 Let $D$ be a nonsingular cubic curve in $\mathbb{P}^2(\mathbb{C})$, the projective closure of $y^2 = f(x)$. (a) Show that the ramification points of $D$ are the three roots of $f$. (b) Let $\gamma$ be any path in $\mathbb{C}$. Show that any lifting of $\gamma$ to a path in $D$ lies within $\mathbb{A}^2$. (Thus, if $\gamma(t) = a$ is a root of $f$, then any lifting $\xi$ of $\gamma$ must satisfy $\xi(t) = (a, 0)$.)

13.69 Let $D$ be a nonsingular cubic curve in $\mathbb{P}^2(\mathbb{C})$, the projective closure of $y^2 = f(x)$. Suppose that $\gamma : [a, b] \to \mathbb{P}^1(\mathbb{C})$ is a path satisfying $\gamma(t) \in \mathbb{C}$ for $t \in [a, b)$ and $\gamma(b) = p_\infty = (0:1)$. Suppose further that $\gamma(t)$ is not a root of $f$ when $a < t < b$. Show that $\gamma$ has a lifting to $D$, meaning a path $\xi : [a, b] \to D$ satisfying $\xi(t) \in D \cap (x = \gamma(t))$ for $t \in [a, b)$ and $\xi(b) = (0:0:1)$.

13.70 Let $D$ be a curve in $\mathbb{P}^2(\mathbb{C})$. Show that if $U \subseteq \mathbb{C} \setminus \Re(D)$ is simply connected, then there is a vertical parameterisation of $D$ whose domain is all of $U$.

**Intersections with Lines**

13.71 Let $D$ be a curve in $\mathbb{P}^2(\mathbb{C})$, and let $\eta\colon U \to V$ be a vertical parameterisation of $D$. Show that for all $p = (a, b) \in V$, $\mathrm{ord}_a(\eta_y) = i_p(D, y = b)$, where recall that $\mathrm{ord}_a(\eta_y)$ is the order $a$ as a zero of $\eta_y$ (Definition 12.13). (Without loss of generality, $p$ is the origin. Let $m = i_p(D, y = 0)$; let $f$ define $D \cap \mathbb{A}^2$. By Example 5.33, write $f = x^m h(x) + yk(x, y)$ where $k(0, 0) \neq 0$. Now $f(x, g(x))$ is the zero formal power series, where $g$ is a power series expansion of $\eta_y$ around 0; compare coefficients.)

13.72 Show that $O \subseteq \check{\mathbb{P}}^2(\mathbb{C})$ is a neighbourhood of the line at infinity if and only if there is some $R > 0$ such that $O$ contains the line at infinity, and the projective closures of all the affine lines whose distance from the origin is at least $R$.

13.73 Let $D$ be a nonsingular curve in $\mathbb{P}^2(\mathbb{C})$, let $p \in D$, and let $\mathcal{L}$ be the family of lines which pass through $p$. What is the degree (Definition 12.48) of the holomorphic map from $D$ to $\mathcal{L}$ taking $q \in D$ to the line $\overline{pq}$?

13.74 Let $D$ be a curve in $\mathbb{P}^2(\mathbb{C})$, let $L$ be a line, and let $p \in D \cap L$. Let $\mathcal{L}$ be the family of lines which pass through $p$. Suppose that $p$ is nonsingular on $D$. Show that there is a neighbourhood $O$ of $L$ in $\mathcal{L}$ and an open neighbourhood $U$ of $p$ in $D$ such that every line $\ell \in O \setminus \{L\}$ intersects $U$ in $i_p(D, L)$-many distinct points.[2]

13.75 The purpose of the next two exercises is to give a topological definition of multiplicity of intersection with lines, similar to Proposition 13.47 but without restricting to linear families of lines. To do this we need to consider "dense and open" as notion of "largeness".

Recall that a subset $Y$ of a quasi-Euclidean space $X$ is dense in $X$ (Exercise 8.44) if every open subset of $X$ intersects $Y$. (a) Show that if $M$ is a manifold and $F \subset M$ is finite then $M \setminus F$ is dense in $M$. (b) Prove the *Baire category theorem*: If $M$ is a manifold, then the intersection of countably many subsets of $M$ which are both dense and open, is dense. (Hint: let $U_1, U_2, \ldots$ be a list of dense open subsets of $M$, and let $U_0$ be an open subset of $M$; we need to show that $\bigcap_n U_n$ is nonempty. By shrinking and taking a chart, we may assume that $U_0 \subset \mathbb{R}^m$. Recursively choose $\boldsymbol{x}_n$ and $\boldsymbol{\varepsilon}_n \leqslant 2^{-n}$ such that $B(\boldsymbol{x}_n, \varepsilon_n) \subseteq U_n$, and $\boldsymbol{x}_k \in B(\boldsymbol{x}_n, \boldsymbol{\varepsilon}_n)$ for all $k \geqslant n$; then use completeness of $\mathbb{R}^m$) (c) Show that the rationals $\mathbb{Q}$ are not the intersection of countably many open subsets of $\mathbb{R}$.[3]

[2] That is, the restriction in Proposition 13.47 that $\mathcal{L}$ is not the family of lines passing through $p$ can be removed, if $p$ is nonsingular on $D$.

[3] Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function. Then the set of $x \in \mathbb{R}$ at which $f$ is continuous is the intersection of countably many open sets. It follows that there is no $f\colon \mathbb{R} \to \mathbb{R}$ which is continuous precisely on the rational numbers. Compare with Exercise 8.121.

13.76 Let $D$ be a curve in $\mathbb{P}^2(\mathbb{C})$. (a) Show that the set of lines $\ell$ which intersect $D$ in $\deg D$ many distinct points is dense and open in $\check{\mathbb{P}}^2(\mathbb{C})$. (b) Let $L$ be a line and let $p$ be a point. Show that $i_p(D, L)$ is the unique integer $m$ such that there is some open neighbourhood $U$ of $p$ in $D$ and an open neighbourhood $O$ of $L$ in $\check{\mathbb{P}}^2(\mathbb{C})$ such that the set of lines which intersect $U$ at precisely $m$ many points is dense and open in $O$.

13.77 This exercise uses the symmetric power of topological spaces introduced in Exercise 12.121. Let $D$ be a curve in $\mathbb{P}^2(\mathbb{C})$ of degree $d$. Show that the map $\ell \mapsto \ell \cdot D$ from $\check{\mathbb{P}}^2(\mathbb{C})$ to $\mathrm{SP}^d(D)$ is continuous. (Recall that $\ell \cdot D$ is $\ell \cap D$ with multiplicity of intersection counted.)

# Elliptic Functions and the Isomorphism Theorem

# 14

Elliptic functions are meromorphic, doubly periodic functions on $\mathbb{C}$. They induce holomorphic functions from complex tori to the Riemann sphere. In this chapter we first prove some general facts about elliptic functions. Then we give an example, namely Weierstrass's function $\wp$. We show that using $\wp$ and its derivative, we can parameterise a nonsingular cubic curve, and obtain the isomorphism theorem between complex tori and elliptic curves (Theorem 14.25). Finally, we prove the inversion theorem (Theorem 14.32), which says that every elliptic curve is obtained this way.

## 14.1 Elliptic Functions

**Definition 14.1** Let $f$ be a function on $\mathbb{C}$. A *period* of $f$ is a complex number $a$ satisfying $f(z+a) = f(z)$ for all $z \in \mathbb{C}$.

For example, $2\pi i$ is a period of the exponential function $z \mapsto e^z$ (see Exercise 11.63)

**Proposition 14.2** *Let $Y$ be a Riemann surface and let $f : \mathbb{C} \to Y$ be holomorphic and nonconstant. Then the set of periods of $f$ is a discrete subgroup of $(\mathbb{C}, +)$.*

***Proof*** Let $G$ be the set of periods of $f$. That $G$ is an additive subgroup of $\mathbb{C}$ is immediate from the definition of periods. Suppose that $G$ is not discrete; let $a \in G$ and let $\langle a_n \rangle$ be a sequence of elements of $G$, distinct from $a$, converging to $a$. By definition, for all $n$, $f(a_n) = f(0) = f(a)$. By Corollary 12.10, $f$ is constant. □

Proposition 8.106 tells us that discrete subgroups of $\mathbb{C}$ are either cyclic, or generated by two elements, linearly independent over $\mathbb{R}$—namely, lattices $\Gamma$. We call a function $f : \mathbb{C} \to Y$ *doubly periodic* if the group of periods of $f$ is of the



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_14

latter kind (not cyclic). For a discrete subgroup $\Gamma$ of $\mathbb{C}$ we say that a function $f$ is $\Gamma$-*periodic* if every element of $\Gamma$ is a period of $f$.

Let $f\colon \mathbb{C} \to Y$ be $\Gamma$-periodic, where $\Gamma$ is a 2-dimensional lattice. Then $f$ induces a well-defined function $\bar{f}\colon T_\Gamma \to Y$ on the complex torus $T_\Gamma = \mathbb{C}/\Gamma$ by letting $\bar{f}(a+\Gamma) = f(a)$. This is the unique function $\bar{f}\colon T_\Gamma \to Y$ satisfying $f = \bar{f} \circ \pi_\Gamma$.

Recall that $T_\Gamma$ is a Riemann surface (Example 12.5).

**Lemma 14.3** *Let $Y$ be a holomorphic surface. A function $g\colon T_\Gamma \to Y$ is holomorphic if and only if the composition $g \circ \pi_\Gamma\colon \mathbb{C} \to Y$ is holomorphic.*

***Proof*** The same as the proof of Lemma 12.60: by Proposition 8.109, $g$ is continuous if and only if $g \circ \pi_\Gamma$ is continuous; both maps have the same coordinate representations. □

It follows that if $f\colon \mathbb{C} \to Y$ is holomorphic and $\Gamma$-periodic, then the induced map $\bar{f}\colon T_\Gamma \to Y$ is holomorphic.

**Definition 14.4** An *elliptic function* is a nonconstant, doubly periodic meromorphic function on $\mathbb{C}$.

Thus, an elliptic function is precisely a composition $f \circ \pi_\Gamma$, where $f$ is meromorphic and nonconstant on $T_\Gamma$.

**Proposition 14.5** *Every elliptic function is onto $\mathbb{P}^1(\mathbb{C})$.*

In particular, every elliptic function has both zeros and poles.

***Proof*** Let $g\colon \mathbb{C} \to \mathbb{P}^1(\mathbb{C})$ be elliptic; let $\bar{g}\colon T_\Gamma \to \mathbb{P}^1(\mathbb{C})$ be the induced holomorphic map. Since $T_\Gamma$ is compact, Proposition 12.43 says that $\bar{g}$ is onto $\mathbb{P}^1(\mathbb{C})$. □

We define the *degree* of an elliptic function to be the degree of the induced holomorphic map $\bar{g}\colon T_\Gamma \to \mathbb{P}^1(\mathbb{C})$ (Definition 12.48).

**Corollary 14.6** *The degree of any elliptic function is at least 2.*

***Proof*** A holomorphic map of degree 1 between compact Riemann surfaces is a biholomorphism (Remark 12.50); by Corollary 12.46, since the sphere is simply connected (Proposition 9.16) and compact, the torus and the Riemann sphere are not biholomorphic.[1] □

[1] In fact, they are not even homeomorphic, as can be seen by examining the fundamental group, see Exercise 9.116.

*Exercise 14.7* Let $G$ be a cyclic subgroup of $\mathbb{C}$. (a) Show that the manifold $\mathbb{C}/G$ (Proposition 8.107) is a non-compact Riemann surface (see Exercise 8.136). (b) Show that a function $g\colon \mathbb{C}/G \to Y$ is holomorphic if and only if $g \circ \pi_G$ is holomorphic. (c) Show that $\mathbb{C}/G$ is biholomorphic with $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$. «

### 14.1.1 The Weierstrass Function $\wp$

Throughout, we fix a 2-dimensional lattice $\Gamma$, and let $T = T_\Gamma$ and $\pi = \pi_\Gamma$. Weierstrass gave an example of an elliptic, $\Gamma$-periodic function $\wp = \wp_\Gamma$. Let $\alpha$ and $\beta$ be a pair of generators for $\Gamma$.

**Lemma 14.8** *The sum*

$$\sum_{u\in\Gamma\setminus\{0\}} \frac{1}{|u|^3}$$

*is finite.*

***Proof*** As usual identifying $\mathbb{C}$ with $\mathbb{R}^2$, let $Q$ be the unique linear transformation from $\mathbb{C}$ to $\mathbb{C}$ mapping 1 to $\alpha$ and $i$ to $\beta$. Since $\alpha$ and $\beta$ are linearly independent, $Q$ is invertible. By Proposition 9.40 (applied to $Q^{-1}$), we get a constant $c > 0$ such that for all $u \in \mathbb{C}$, $c|u| \leqslant |Q(u)|$. Note that $Q\restriction_{\mathbb{Z}[i]}$ is a bijection between $\mathbb{Z}[i]$ (the lattice generated by 1 and $i$) and $\Gamma$.

For $k \in \mathbb{N}$, let

$$A(k) = \{n + im \in \mathbb{Z}[i] \,:\, \max\{|n|, |m|\} = k\};$$

so $\{A(k) \,:\, k \geqslant 0\}$ is a partition of $\mathbb{Z}[i]$. For all $k \geqslant 1$, $|A(k)| = 8k$.

We have

$$\sum_{u\in\Gamma\setminus\{0\}} \frac{1}{|u|^3} = \sum_{k\geqslant 1} \sum_{z\in A(k)} \frac{1}{|Q(z)|^3}. \tag{14.1}$$

For $k \geqslant 1$ and $z \in A(k)$, we have $|z| \geqslant k$, so $|Q(z)| \geqslant ck$. Hence

$$\sum_{z\in A(k)} \frac{1}{|Q(z)|^3} \leqslant \frac{8}{c^3k^2}.$$

As the series $\sum_k 1/k^2$ converges (Exercise 11.28), the result now follows from Eq. (14.1). □

For $u \in \Gamma \setminus \{0\}$ and $z \in \mathbb{C} \setminus \{u\}$, let

$$g_u(z) = \frac{1}{(z-u)^2} - \frac{1}{u^2};$$

for $z \in \mathbb{C} \setminus \{0\}$, let

$$g_0(z) = \frac{1}{z^2}.$$

**Lemma 14.9** *For all $R > 0$ there is some finite $\Lambda \subset \Gamma$ such that*

$$\sum_{u \in \Gamma \setminus \Lambda} g_u$$

*converges absolutely uniformly on $B(0, R)$.*

***Proof*** Given $R > 0$, let $\Lambda = \Gamma \cap \overline{B}(0, 2R)$; $\Lambda$ is finite (Proposition 8.101). Let $z \in B(0, R)$ and let $u \in \Gamma \setminus \Lambda$.

$$\left| \frac{1}{(z-u)^2} - \frac{1}{u^2} \right| = \frac{|2zu - z^2|}{|u|^2 |z-u|^2} = \frac{|z||2u - z|}{|u|^2 |z-u|^2}.$$

We have $|z| < R$ and $|u| > 2R$, so $|u| > 2|z|$; so

$$|2u - z| \leqslant 2|u| + |z| \leqslant 2.5 \cdot |u|;$$

on the other hand,

$$|z - u| \geqslant |u| - |z| \geqslant |u|/2,$$

so overall

$$\left| \frac{1}{(z-u)^2} - \frac{1}{u^2} \right| \leqslant \frac{R \cdot 2.5 \cdot |u|}{|u|^2 |u/2|^2} = \frac{10R}{|u|^3}.$$

The result follows now from Lemma 14.8 (and the Weierstrass $M$-Test). □

**Definition of $\wp$**

Define

$$\wp = \wp_\Gamma = \sum_{u \in \Gamma} g_u = \frac{1}{z^2} + \sum_{u \in \Gamma \setminus \{0\}} \left( \frac{1}{(z-u)^2} - \frac{1}{u^2} \right).$$

For every $u \in \Gamma$, the function $g_u$ is analytic on $\mathbb{C} \setminus \{u\}$. Weierstrass's Theorem 11.77 and Lemma 14.9 imply that $\wp$ is well-defined and analytic on $\mathbb{C} \setminus \Gamma$, and that $\wp'$, which is also analytic on $\mathbb{C} \setminus \Gamma$, is given by the sum

$$\wp' = \sum_{u \in \Gamma} g_u' = -2 \sum_{u \in \Gamma} \frac{1}{(z-u)^3}.$$

Let $u \in \Gamma$. Let $U$ be a bounded open neighbourhood of $u$ such that $\Gamma \cap U = \{u\}$. For $w \in \Gamma \setminus \{u\}$, $g_w$ is analytic on $U$; by Lemma 14.9,

$$\sum_{w \in \Gamma \setminus \{u\}} g_w$$

converges to an analytic function. Hence

$$\wp = g_u + \sum_{w \in \Gamma \setminus \{u\}} g_w$$

gives a Laurent expansion of $\wp$ on $U \setminus \{u\}$. This shows that $\wp$ is meromorphic on $\mathbb{C}$, and that each $u \in \Gamma$ is a pole of order 2 of $\wp$. In particular, we see that $\wp$ is not constant.

Differentiating, we see that $\wp'$ is a meromorphic function and that every $u \in \Gamma$ is a pole of order 3 of $\wp'$ (see Exercise 12.21).

### $\wp$ Is an Elliptic Function

**Lemma 14.10** *For all $z \in \mathbb{C} \setminus \Gamma$, $\wp(z) = \wp(-z)$.*

***Proof*** Let $z \in \mathbb{C} \setminus \Gamma$. For $u \in \Gamma \setminus \{0\}$,

$$g_u(-z) = \frac{1}{(-z-u)^2} - \frac{1}{u^2} = \frac{1}{(z+u)^2} - \frac{1}{(-u)^2} = g_{-u}(z),$$

and similarly, $g_0(-z) = g_0(z)$. Since $u \mapsto -u$ is a permutation of $\Gamma$, we have

$$\wp(-z) = \sum_{u \in \Gamma} g_u(-z) = \sum_{u \in \Gamma} g_{-u}(z) = \sum_{u \in \Gamma} g_u(z) = \wp(z). \qquad \square$$

*Remark 14.11* The proof of Lemma 14.10 relies on Proposition 11.34; in fact the very definition of $\wp$ relies on it, since to apply Theorem 11.77 we need to order $\Gamma$ in some way; Proposition 11.34 implies that the order does not matter. This will be used in the proof of Proposition 14.12 below as well. «

**Proposition 14.12** *$\wp'$ is elliptic; every $u \in \Gamma$ is a period of $\wp'$.*

***Proof*** Let $u \in \Gamma$ and let $z \in \mathbb{C} \setminus \Gamma$. For any $w \in \Gamma$,

$$g'_w(z+u) = \frac{-2}{(z+u-w)^3} = g'_{w-u}(z).$$

Since $w \mapsto w - u$ is a permutation of $\Gamma$,

$$\wp'(z+u) = \sum_{w\in\Gamma} g'_w(z+u) = \sum_{w\in\Gamma} g'_{w-u}(z) = \sum_{w\in\Gamma} g'_w(z) = \wp'(z).$$

Also $\wp'$ is meromorphic on $\mathbb{C}$ and every $u \in \Gamma$ is a pole of $\wp'$, so thought of as a function to $\mathbb{P}^1(\mathbb{C})$, $\wp'$ is $\Gamma$-periodic. □

**Proposition 14.13** *$\wp$ is elliptic; every $u \in \Gamma$ is a period of $\wp$.*

***Proof*** By Proposition 14.2, it suffices to show that $\alpha$ and $\beta$ (the chosen generators of $\Gamma$) are periods of $\wp$. So let $u = \alpha$ or $u = \beta$.

Consider the analytic functions $\wp(z)$ and $\wp(z+u)$, both defined on $\mathbb{C} \setminus \Gamma$. The difference of their derivatives is 0; since $\mathbb{C} \setminus \Gamma$ is connected, it follows that $\wp(z+u) - \wp(z)$ is a constant function $c$ on $\mathbb{C} \setminus \Gamma$ (Proposition 11.8). It remains to show that $c = 0$.

By our choice of $u$, $u/2 \notin \Gamma$. So $\wp(u/2) - \wp(-u/2) = c$. By Lemma 14.10, $\wp(u/2) = \wp(-u/2)$. Hence $c = 0$ as required. □

Let $\bar{\wp}$ be the induced holomorphic function on $T$. Since all points in $\Gamma$ are identified in $T$, $\bar{\wp}$ has a single pole, which we observed has order 2. Hence the degree of $\wp$ is 2. By Corollary 14.6, this is the smallest degree possible for an elliptic function.

### Inverse Images of Points

Proposition 14.5 says that the image of $\wp$ is all of $\mathbb{P}^1(\mathbb{C})$. In fact, since the degree of $\wp$ is 2, every point of $\mathbb{P}^1(\mathbb{C})$ has exactly 2 pre-images by $\bar{\wp}$, multiplicities counted.

Let $c \in \mathbb{C}$, identified with $(1:c) \in \mathbb{P}^1(\mathbb{C})$; let $w \in \mathbb{C}$ such that $\wp(w) = c$. By Lemma 14.10, $\wp(-w) = c$. Hence, if $w + \Gamma \neq -w + \Gamma$, i.e., if $2w \notin \Gamma$, then $w + \Gamma$ and $-w + \Gamma$ are the only $\bar{\wp}$-preimages of $c$, each necessarily with valency (multiplicity) 1. On the other hand, suppose $2w \in \Gamma$ (but $w \notin \Gamma$). There are exactly three such points modulo $\Gamma$, which are $\alpha/2$, $\beta/2$ and $(\alpha+\beta)/2$. (To see this, observe that every $z \in \mathbb{C}$ is equivalent modulo $\Gamma$ to $s\alpha + t\beta$ where $s, t \in [0, 1)$; if $2s\alpha + 2t\beta \in \Gamma$ then we must have both $s$ and $t$ either 0 or $1/2$.) Since $\wp$ is even, $\wp'$ is odd (Exercise 11.59), and so for each such point $w$, $\wp'(w) = -\wp'(-w) = -\wp'(w)$, so $\wp'(w) = 0$. This means that each of these points $w$ has valency greater than one—necessarily 2; and that for each such $w$, $w + \Gamma$ is the unique $\bar{\wp}$-preimage of $c = \wp(w)$. To summarise:

**Proposition 14.14** *For every $w + \Gamma \in T$, the $\bar{\wp}$-preimages of $c = \wp(w)$ are precisely $w + \Gamma$ and $-w + \Gamma$.*

Note that the function $\wp$ depends only on $\Gamma$, and not on the choice of generators $\alpha$ and $\beta$; it follows that the collection of three cosets $\{\alpha/2+\Gamma, \beta/2+\Gamma, (\alpha+\beta)/2+\Gamma\}$ also does not depend on $\alpha$ and $\beta$, only on $\Gamma$.

### 14.1.2 The Differential Equation for $\wp$

Let

$$h_0 = \sum_{u \in \Gamma \setminus \{0\}} g_u.$$

We recall that $h_0$ is analytic in some open neighbourhood $U$ of 0, and that

$$\wp(z) = \frac{1}{z^2} + h_0$$

on $U \setminus \{0\}$. Since both $\wp$ and $1/z^2$ are even functions, so is $h_0$ (also, an inspection of the proof of Lemma 14.10 gives direct verification of this fact, as $u \mapsto -u$ is a permutation of $\Gamma \setminus \{0\}$). Hence on $U$ we may assume that $h_0$ is given by a sum of a power series in which only even powers of $z$ appear (see Exercise 11.59). Also, for all $u \in \Gamma \setminus \{0\}$, $g_u(0) = 0$. Hence $h_0(0) = 0$. We can thus write:

$$\wp(z) = \frac{1}{z^2} + \lambda z^2 + \mu z^4 + \cdots$$

for $z \in U \setminus \{0\}$, where $\ldots$ denotes higher order terms (and in fact, these terms start with $z^6$).

Differentiating, we get

$$\wp'(z) = \frac{-2}{z^3} + 2\lambda z + 4\mu z^3 + \cdots;$$

taking powers, we get

$$(\wp'(z))^2 = \frac{4}{z^6} - \frac{8\lambda}{z^2} - 16\mu + \cdots,$$

and

$$\wp^3(z) = \frac{1}{z^6} + \frac{3\lambda}{z^2} + 3\mu + \cdots.$$

Hence on $U \setminus \{0\}$,

$$(\wp'(z))^2 - 4\wp^3(z) = \frac{-20\lambda}{z^2} - 28\mu + \dots .$$

Let

$$k(z) = (\wp'(z))^2 - 4\wp^3(z) + 20\lambda\wp(z) + 28\mu.$$

It follows that $k$ is analytic on $U$ and satisfies $k(0) = 0$.

Now $k$ is analytic on $\mathbb{C} \setminus \Gamma$; as $\wp$ and $\wp'$ are $\Gamma$-periodic, and $k$ is a polynomial in $\wp$ and $\wp'$, $k$ is also $\Gamma$-periodic. This means that $k$ has limit 0 at every $u \in \Gamma$; so $k$ can be extended to an entire function on $\mathbb{C}$ (an analytic function defined on all of $\mathbb{C}$; in other words, it is a meromorphic function with no poles). As this entire function is also $\Gamma$-periodic, it is constant (Proposition 14.5). As $k(0) = 0$, we see that $k = 0$. In other words, for all $z \in \mathbb{C} \setminus \Gamma$,

$$(\wp'(z))^2 = 4\wp^3(z) - 20\lambda\wp(z) - 28\mu. \tag{14.2}$$

Now we calculate the coefficients $\lambda$ and $\mu$. By Taylor's theorem, we have

$$\lambda = \frac{h_0^{(2)}(0)}{2!}$$

and

$$\mu = \frac{h_0^{(4)}(0)}{4!}.$$

Differentiating the sum

$$h_0 = \sum_{u \in \Gamma \setminus \{0\}} g_u,$$

we get by induction on $k \geqslant 1$,

$$h_0^{(k)}(z) = (-1)^k (k+1)! \sum_{u \in \Gamma \setminus \{0\}} \frac{1}{(z-u)^{k+2}}.$$

Hence

$$\lambda = \frac{3!}{2!} \sum_{u \in \Gamma \setminus \{0\}} \frac{1}{u^4}$$

and

$$\mu = \frac{5!}{4!} \sum_{u \in \Gamma \setminus \{0\}} \frac{1}{u^6}.$$

(These are called the *Eisenstein series* for these coefficients.) Let

$$\gamma_2 = \gamma_2(\Gamma) = 20\lambda = 60 \sum_{u \in \Gamma \setminus \{0\}} \frac{1}{u^4}$$

and

$$\gamma_3 = \gamma_3(\Gamma) = 28\mu = 140 \sum_{u \in \Gamma \setminus \{0\}} \frac{1}{u^6};$$

then

$$(\wp')^2 = 4\wp^3 - \gamma_2\wp - \gamma_3. \tag{14.3}$$

## 14.2 The Curve $E_\Gamma$ and the Isomorphism Theorem

Recall that $(T, +)$ is the quotient group $(\mathbb{C}, +)/(\Gamma, +)$. The points of order 2 in the group $(T, +)$ are the cosets $w + \Gamma$ with $w \notin \Gamma$ but $2w \in \Gamma$; above, we observed that these are precisely the points $\alpha/2 + \Gamma$, $\beta/2 + \Gamma$ and $(\alpha + \beta)/2 + \Gamma$. Together with $\Gamma = 0 + \Gamma$, these form a subgroup of $(T, +)$ isomorphic to $C_2 \times C_2$. (Recall that $C_2$ is the cyclic group of 2 elements.)

Proposition 14.14 implies:

**Lemma 14.15** *The values $\wp(\alpha/2)$, $\wp(\beta/2)$ and $\wp(\alpha/2 + \beta/2)$ are distinct.*

Let $\gamma_2$ and $\gamma_3$ be the associated coefficients defined from $\Gamma$. We let

$$f = f_\Gamma = 4x^3 - \gamma_2 x - \gamma_3.$$

Thus $(\wp')^2 = f(\wp)$.

**Lemma 14.16** *The roots of $f$ in $\mathbb{C}$ are $\wp(\alpha/2)$, $\wp(\beta/2)$ and $\wp(\alpha/2 + \beta/2)$.*

***Proof*** Let $w \in \mathbb{C} \setminus \Gamma$, and suppose that $w + \Gamma$ has order 2 in $(T, +)$. We show that $f(\wp(w)) = 0$. Let $c = \wp(w)$. As noticed in the proof of Proposition 14.14, $w$ is a zero of order 2 of $\wp - c$: $\wp'(w) = 0$, and so $f(c) = (\wp')^2(w) = 0$. □

We let $E = E_\Gamma$ be the projective closure of the affine complex curve defined by the equation $y^2 = f_\Gamma$. This is a cubic curve. Since the three roots of $f$ in $\mathbb{C}$ are distinct, Proposition 7.21 tells us that $E$ is non-singular. By Remark 7.24, $E$ intersects the line at infinity only at the vertical point at infinity $(0:0:1)$, which is a flex of $E$ (with the tangent being the line at infinity). As we did in Chap. 7, we fix the point $(0:0:1)$ to be the identity element $0_E$ of the elliptic curve $(E, 0_E)$, equipped with the chord-and-tangent group structure.

Recall that being $\Gamma$-periodic, $\wp$ and $\wp'$ induce maps on $T$. Define $\varphi\colon T \to \mathbb{P}^2(\mathbb{C})$ by letting

$$\varphi(z+\Gamma) = \begin{cases} (\wp(z), \wp'(z)), & \text{if } z \notin \Gamma; \\ (0:0:1) & \text{otherwise,} \end{cases}$$

(as usual we are identifying $\mathbb{A}^2$ with the subset of $\mathbb{P}^2$ via $\rho_0$).

**Proposition 14.17** *$\varphi$ is a bijection between $T$ and $E$.*

***Proof*** The range of $\varphi$ is contained in $E$: we checked that $(0:0:1) \in E$; for $z \notin \Gamma$, we verified that $(\wp'(z))^2 = f(\wp(z))$.

Next, we show that $\varphi$ is onto $E$. Certainly the vertical point at infinity is in the range of $\varphi$. All other points in $E$ are in $\mathbb{A}^2$; if $(a, b) \in E$ then the other intersection in $\mathbb{A}^2$ of $E$ with the line $x = a$ is $(a, -b)$ (the points of course coincide if $b = 0$, i.e., if $a$ is a root of $f$). Fix some $a \in \mathbb{C}$. By Proposition 14.14, pick $w$ in $\mathbb{C} \setminus \Gamma$ such that $\wp(w) = \wp(-w) = a$. Let $b = \wp'(w)$, so $-b = \wp'(-w)$; so $(a, b) = \varphi(w+\Gamma)$ and $(a, -b) = \varphi(-w + \Gamma)$. The same analysis shows that $\varphi$ is 1–1. □

**Proposition 14.18** *$\varphi\colon T \to E$ is continuous.*

***Proof*** By Proposition 8.109, it suffices to check that the map $\varphi \circ \pi$ is continuous on $\mathbb{C}$. If $w \notin \Gamma$, then on a neighbourhood of $w$, the map $\varphi \circ \pi$ is the map $z \mapsto (\wp(z), \wp'(z))$; both $\wp$ and $\wp'$ are continuous, so this is a continuous map to $\mathbb{A}^2(\mathbb{C})$. Since $E$ is a topological subspace of $\mathbb{P}^2(\mathbb{C})$, we see that $\varphi \circ \pi$ is continuous at $w$ as a map to $E$.

So we need to check that $\varphi \circ \pi$ is continuous at $w \in \Gamma$; since this map is $\Gamma$-periodic, it suffices to show that it is continuous at 0. Let $U$ be a neighbourhood of 0 of small diameter. For all $z \in U \setminus \{0\}$,

$$\wp(z) = \frac{1}{z^2} h_1$$

and

$$\wp'(z) = \frac{1}{z^3} h_2,$$

where $h_1$ and $h_2$ are analytic on $U$ and $h_1, h_2 \neq 0$ on $U$. So for $z \in U \setminus \{0\}$,

$$(\varphi \circ \pi)(z) = (1 : \wp(z) : \wp'(z)) = (z^3 : z h_1 : h_2).$$

However for $z = 0$ we have $(z^3 : z h_1 : h_2) = (0:0:1)$ (as $h_2(0) \neq 0$). Since the functions $z^3$, $z h_1$ and $h_2$ are all continuous on $U$, by Exercise 8.59, $\varphi \circ \pi$ is continuous on $U$. (In greater detail, $\rho_2 \circ \varphi \circ \pi$ is the function $z \mapsto (z^3/h_2, z h_1/h_2)$ which is continuous on $U$, and $\rho_2$ is a homeomorphism.) □

**Proposition 14.19** *$\varphi : T \to E$ is holomorphic.*

***Proof*** Let $w \in \mathbb{C}$, and suppose that $2w \notin \Gamma$. Then $p = (a, b) = (\wp(w), \wp'(w)) = \varphi(w + \Gamma)$ is in $\mathbb{A}^2$, and the tangent to $E$ at that point is not vertical (the tangent is vertical at points $p \in E$ of order 2 in $(E, +)$, where the $x$-coordinate is a root of $f$). So there is a chart $\psi$ for $E$ on a neighbourhood $W$ of $p$ which is the projection onto the first coordinate ($\psi(a, b) = a$; $\psi$ is the inverse of a vertical parameterisation of $E$). Let $U$ be an open neighbourhood of $w$ of small diameter (and disjoint from $\Gamma$), so that $\pi\restriction_U$ is the inverse of a chart for $T$ (see Sect. 8.4), and $\varphi[\pi[U]] \subseteq W$ (as $\varphi$ is continuous at $w + \Gamma$). Then the coordinate representation $\psi \circ \varphi \circ (\pi\restriction_U)$ of $\varphi$ is the restriction of $\wp$ to $U$, which is of course analytic.

The result now follows from Propositions 14.18 and 12.11. □

Proposition 12.37 now implies:

**Corollary 14.20** *$\varphi$ is a biholomorphism between $T$ and $E$.*

### 14.2.1 The Isomorphism Theorem

The isomorphism theorem states that the biholomorphism $\varphi$ from $T$ to $E$ is also a group isomorphism between the quotient group $(T, +)$ and the group structure on $E$ from Chap. 7, which is characterised by $p + q + r = 0_E = (0:0:1)$ if and only if $p, q$ and $r$ are collinear.

To prove this, we define a function $\mu$ from the dual plane $\check{\mathbb{P}}^2(\mathbb{C})$ to $T$. For a line $\ell$, we let $\mu(\ell)$ be the sum, in the group $(T, +)$, of the preimages by $\varphi$ of three points of intersection of $\ell$ with $E$. That is, if $\ell \cdot E = [q_1, q_2, q_3]$ then

$$\mu(\ell) = \varphi^{-1}(q_1) + \varphi^{-1}(q_2) + \varphi^{-1}(q_3),$$

where again the sum is taken in the quotient group $T = \mathbb{C}/\Gamma$.

**Lemma 14.21** $\mu(\ell_\infty) = \Gamma$.

(Recall that $\Gamma = \pi_\Gamma(0)$ is the identity element of the group $T$.)

***Proof*** We know that $E \cdot \ell_\infty = [0_E, 0_E, 0_E]$, and that $\varphi(\Gamma) = 0_E$, so

$$\mu(\ell_\infty) = \Gamma + \Gamma + \Gamma = \Gamma. \qquad \square$$

**Lemma 14.22** *The function $\mu$ is continuous.*

***Proof*** Let $L$ be a line; we show that $\mu$ is continuous at $L$. Let $E \cdot L = [p_1, p_2, p_3]$. Let $W$ be an open neighbourhood of $\mu(L)$ in $T$. Since $\varphi^{-1}$ is continuous, and addition in $T$ is continuous (Exercise 8.113), there are open neighbourhoods $V_1$, $V_2$ and $V_3$ of $p_1$, $p_2$ and $p_3$ in $E$ such that for all $q_i \in V_i$, $\sum_{i \leqslant 3} \varphi^{-1}(q_i) \in W$. If $p_i = p_j$ then we may take $V_i = V_j$; otherwise we can require that $V_i$ and $V_j$ are disjoint. Applying Proposition 13.46 for each $V_i$ we obtain open neighbourhoods $U_i \subseteq V_i$ of $p_i$ in $E$ and an open neighbourhood $O$ of $L$ in $\check{\mathbb{P}}^2(\mathbb{C})$ such that every $\ell \in O$ intersects $U_i$ exactly $i_{p_i}(E, L)$ many times (again we require that if $i = j$ then $U_i = U_j$). Then for $\ell \in O$, $\mu(\ell) \in W$.[2] $\square$

Toward showing that $\mu$ is holomorphic on every linear family of lines, we need:

**Lemma 14.23** *If $X$ is a holomorphic surface, and $f, g \colon X \to T$ are holomorphic, then so is $f + g$ (where the sum is taken in the group $T$).*

***Proof*** Since $T$ is a topological group, $f + g$ is continuous. A coordinate representation of $f + g$ is the sum of coordinate representations of $f$ and of $g$, and so is the sum of analytic functions, which is therefore analytic. $\square$

Recall that a linear family of lines $\mathcal{L}$ is a Riemann surface (see Sect. 13.3).

**Lemma 14.24** *For any linear family of lines $\mathcal{L}$, the function $\mu\restriction_{\mathcal{L}}$ is holomorphic.*

***Proof*** By Corollary 5.40, Lemma 14.22, and Proposition 12.11, it suffices to show that for any $L \in \mathcal{L}$ which intersects $E$ in three distinct points, $\mu\restriction_{\mathcal{L}}$ is holomorphic on a neighbourhood of $L$.

[2] Using Exercise 13.77, this proof essentially says that $\mu$ is continuous since it is the composition of continuous maps $\check{\mathbb{P}}^2(\mathbb{C}) \to \mathrm{SP}^3(E) \to \mathrm{SP}^3(T) \to T$, the last induced by addition (which is well-defined on $\mathrm{SP}^3(T)$ by associativity and commutativity).

Suppose that $L \in \mathcal{L}$ intersects $E$ in three distinct points. We apply Proposition 13.50. Let $O$ be a neighbourhood of $L$ in $\mathcal{L}$, and let $U_1, U_2$ and $U_3$ be three disjoint open subsets of $E$ and let $g_1, g_2$ and $g_3$ be holomorphic functions from $O$ to $U_i$ such that for $i = 1, 2, 3$, for each $\ell \in O$, $g_i(\ell)$ is the unique point of intersection of $\ell$ with $U_i$. Then on $O$, $\mu = \sum_{i \leqslant 3} \varphi^{-1} \circ g_i$, the sum taken in $T$. But each $\varphi^{-1} \circ g_i$ is holomorphic; so by Lemma 14.23, their sum in $T$ is holomorphic as well. □

Since $\mathcal{L}$ is biholomorphic with $\mathbb{P}^1(\mathbb{C})$ (Example 13.21), Corollary 12.46 implies that $\mu \restriction_{\mathcal{L}}$ is constant. But any two lines in $\check{\mathbb{P}}^2$ are elements of a linear family of lines (as any two points in $\mathbb{P}^2$ determine a line). This implies that $\mu$ is constant. Lemma 14.21 implies that this constant value is $\Gamma = \pi(0)$.

**Theorem 14.25 (Abel-Poincaré-Weil)** *$\varphi$ is a group isomorphism between $T$ and $E$.*

***Proof*** Essentially, this follows from the fact that the group structure of $E$ is characterised by the fact that three points add to the identity element if and only if they are collinear. We give the details showing that $\varphi^{-1}$ is a group homomorphism.

Let $0_E = (0:0:1)$ be the identity element of $E$; we know that $\varphi^{-1}(0_E) = \Gamma$.

We first show that for all $p \in E$, $\varphi^{-1}(-p) = -\varphi^{-1}(p)$. Since $\overline{0_E\, p} \cdot E = [0_E, p, -p]$ and $\mu(\overline{0_E\, p}) = \Gamma$, we see that in $T$, $\Gamma + \varphi^{-1}(p) + \varphi^{-1}(-p) = \Gamma$.

Now let $p, q \in E$. Since $\overline{p\, q} \cdot E = [p, q, -(p+q)]$ and $\mu(\overline{p\, q}) = \Gamma$, in $T$ $\varphi^{-1}(p) + \varphi^{-1}(q) - \varphi^{-1}(p+q) = \Gamma$, which shows that $\varphi^{-1}(p) + \varphi^{-1}(q) = \varphi^{-1}(p+q)$. □

## 14.3 Inversion

We have seen that every complex torus is isomorphic to an elliptic curve by an isomorphism which is both algebraic and holomorphic. We now work toward the reverse: every elliptic curve is isomorphic to a complex torus. This is known as the *inversion* (or *uniformisation*) theorem for elliptic curves.

### 14.3.1 A Non-vanishing Form on a Nonsingular Cubic

Let $E$ be a nonsingular cubic curve. Recall (Proposition 13.23) that rational functions extend to meromorphic functions on the curve. We are interested in particular in two such functions, namely $x/w$ and $w/y$. In affine coordinates, these are the functions $(x, y) \mapsto x$ and $(x, y) \mapsto 1/y$. We take the differential of the first

(see page 337) and multiply by the second (see page 335) to obtain the meromorphic form

$$\frac{w}{y} d\left(\frac{x}{w}\right)$$

on $E$, which we denote, using affine coordinates, by $dx/y$. Suppose that $E$ is given by $y^2 = f(x)$ (see Proposition 7.23), and recall that since $E$ is nonsingular, $f$ has three distinct roots (Proposition 7.21).

**Proposition 14.26** *$dx/y$ is a non-vanishing holomorphic form on $E$.*

***Proof*** Recall (see page 337) that this means that the form $dx/y$ has no zeros and no poles.

We now give a solution for Exercise 13.27. There are three kinds of points: (i) the vertical point at infinity $(0:0:1)$; (ii) the three points of intersection of $E$ with the $x$-axis, namely the points $(a, 0)$ where $a$ is one of the three distinct roots of $f$; (iii) all other points.

If $f(a) \neq 0$ then there are two distinct points mapped to $a$ by $x/w$, namely $(a, b)$ and $(a, -b)$ where $b^2 = f(a)$. Hence $x/w$, as a meromorphic function on $E$, has degree 2 (all but finitely many points have valency 1). The vertical point at infinity is the only pole of $x/w$, and so is a pole of order 2. If $f(a) = 0$ then $(a, 0)$ is the unique point in $E$ mapped to $a$ by $x/w$, and so has valency 2. By Proposition 12.86, $\mathrm{ord}_{(0:0:1)}(dx) = -3$ (where $dx$ of course abbreviates $d(x/w)$), $\mathrm{ord}_{(a,0)}(dx) = 1$ where $f(a) = 0$, and $\mathrm{ord}_p(dx) = 0$ at all other points.

Similarly, for all but finitely many $b \in \mathbb{C}$, the polynomial $f(x) - b^2$ has three distinct roots, and so three distinct points are mapped to $1/b$ by $w/y$; hence the degree of this meromorphic map is 3. The map $w/y$ has three distinct poles, at $(a, 0)$ where $f(a) = 0$, and so each pole has order 1. The vertical point at infinity is the only zero of $w/y$, and so is a zero of order 3.

Adding up (using Proposition 12.80), we see that $\mathrm{ord}_p(dx/y) = 0$ for every point $p \in E$. □

Note that since $dx/y$ has no poles, we can integrate it along any path in $E$.

### 14.3.2 Working After the Fact

Let us consider the problem backwards. Suppose that we already know that $E = E_\Gamma$. How can we recover the lattice $\Gamma$ by just looking at $E$? The answer is to integrate the form $dx/y$, as we now explain.

Recall (Exercise 12.88) that we defined a holomorphic form $\omega$ on the torus $T = T_\Gamma$ satisfying $\pi^*\omega = dz$, where $\pi = \pi_\Gamma$ is the quotient map. Let $\varphi\colon T \to E_\Gamma$ be the isomorphism defined in the previous section. Recall (see page 334) that we can pull back meromorphic forms by holomorphic functions.

**Lemma 14.27** $\omega = \varphi^*(dx/y)$.

***Proof*** As above, let $\bar{\wp}$ and $\bar{\wp}'$ denote the induced functions on $T$. Recall that $\varphi = (\bar{\wp}, \bar{\wp}')$ on $T \setminus \{\Gamma\}$. We conclude that $(x/w) \circ \varphi$ is the function $\bar{\wp}$, and $(w/y) \circ \varphi$ is the function $1/\bar{\wp}'$. By Exercises 12.79 and 12.85,

$$\varphi^*(dx/y) = ((w/y) \circ \varphi) \cdot \varphi^*(dx) = ((w/y) \circ \varphi) \cdot d\big((x/w) \circ \varphi\big) = d\bar{\wp}/\bar{\wp}'.$$

On the other hand,

$$\pi^*(d\bar{\wp}/\bar{\wp}') = \wp' dz/\wp' = dz.$$

Thus, the "coordinate representations" of $\varphi^*(dx/y)$ are $(dz, \psi)$ for charts $\psi$ for $T$, the same as for $\omega$. □

This means that integration of $dx/y$ along paths in $E$ can be translated to integration of $\omega$ along paths in $T$. In the following, assume all paths are piecewise smooth.

**Lemma 14.28** $\Gamma$ *is the collection of values* $\int_\gamma \omega$, *where* $\gamma$ *is a loop in* $T$.

***Proof*** Given $q \in \Gamma$, consider the path $\xi(t) = tq$, from $[0, 1]$ to $\mathbb{C}$. Then $\pi \circ \xi$ is a loop in $T$ (it travels from $\pi(0) = \Gamma$ to $\pi(q) = \Gamma$), and by Exercise 12.91,

$$\int_{\pi \circ \xi} \omega = \int_\xi dz = \xi(1) - \xi(0) = q.$$

Hence $q$ is obtained as the value of an integral $\int_\gamma \omega$ for some loop $\gamma$ in $T$.

On the other hand, let $\gamma\colon [a, b] \to T$ be any loop in $T$. By Proposition 9.20, let $\xi$ be a lifting of $\gamma$ to a path in $\mathbb{C}$; by Exercise 9.109, $\xi$ is piecewise smooth. By Exercise 12.91,

$$\int_\gamma \omega = \int_\xi dz = \xi(b) - \xi(a).$$

Since $\pi(\xi(b)) = \gamma(b) = \gamma(a) = \pi(\xi(a))$, we have $\xi(b) - \xi(a) \in \Gamma$. □

By Exercise 12.91 again, we obtain:

**Proposition 14.29** $\Gamma$ *is the collection of values* $\int_\gamma dx/y$, *where* $\gamma$ *is a loop in* $E$.

Thus, we can recover $\Gamma$ by integrating the form $dx/y$ in $E$. In fact, integration also allows us to recover the isomorphism between $T$ and $E$ (see Exercise 14.51), which in turn allows us to interpret the isomorphism theorem as an addition formula for elliptic integrals.

Now it would make sense to try to prove the inversion theorem by first changing coordinates so that $E$ is given in Weierstrass normal form (Proposition 7.28), then *defining* $\Gamma$ as the set of values as described in Proposition 14.29, and then showing that $E = E_\Gamma$. This is possible, but it turns out that to show that the resulting $\Gamma$ is a lattice, it is first necessary to show that $E$ is homeomorphic to the torus (so that the fundamental group is $\mathbb{Z} \times \mathbb{Z}$). This proof is presented in many texts; see, for example, [Kir92, Ex.6.13] or [KJK+06, Thm.5.1.2], or more extensively, [Kna92, Ch.VI]. Another proof of the inversion theorem uses the $j$-invariant, which is related to the theory of modular forms; see, for example, [Lan87, Sec.3.3].

We will present a different proof (see, for example, [KJK+06, Thm.5.2.1]). This proof uses the group structure on $E$.

### 14.3.3 Invariance of the Non-vanishing Holomorphic Form

For the proof of the inversion theorem, we need an *invariance* property of the non-vanishing holomorphic form. Let $(E, 0_E)$ be an elliptic curve. For each $q \in E$, by Proposition 13.57, the map $p \mapsto p + q$ is holomorphic on $E$. Hence, for any meromorphic form $\omega$ on $E$, we can pull back and obtain the form $(p \mapsto p + q)^*\omega$. For ease of notation, let $\mathrm{add}_q$ be the map $p \mapsto p + q$.

**Lemma 14.30** *Let $(E, 0_E)$ be an elliptic curve, with $E$ given by $y^2 = f(x)$ and $0_E = (0:0:1)$. Then for all $q \in E$, $\mathrm{add}_q^*(dx/y) = dx/y$.*

***Proof*** By Corollary 12.82 and Proposition 14.26, for all $q \in E$ there is some $\lambda(q) \in \mathbb{C}$ such that $\mathrm{add}_q^*(dx/y) = \lambda(q)(dx/y)$. Since $\mathrm{add}_{0_E}$ is the identity on $E$, $\lambda(0_E) = 1$. The proposition then follows from Corollary 12.44 (and Exercise 13.7), once we show that $\lambda\colon E \to \mathbb{C}$ is holomorphic.

Let $q_0 \in E \cap \mathbb{A}^2$. Choose any $p_0 \in E \cap \mathbb{A}^2$, distinct from $q_0$ and $-q_0$, such that the tangent to $E$ at $p_0$ is not vertical. Choose some vertical parameterisation $\eta\colon U \to V$ of $E$ with $p_0 \in V$. By shrinking, choose some neighbourhood $W$ of $q_0$ in $E$ so that $p \neq q, -q$ for all $p \in V$ and $q \in W$ (this uses Proposition 13.55).

The key to the proof is that the group operation of $E$ is *algebraic*: by Exercise 7.30, there are rational functions $X$ and $Y$ (in four variables) such that for all $p = (p_x, p_y) \in U$ and $q = (q_x, q_y) \in V$,

$$(q_x, q_y) + (p_x, p_y) = (X(q_x, q_y, p_x, p_y), Y(q_x, q_y, p_x, p_y)).$$

For any $q \in W$, restricted to $V$, $\mathrm{add}_q^*(dx/y) = d(X(q_x, q_y, p_x, p_y))/Y(q_x, q_y, p_x, p_y)$ (where by the differential we mean that $q$ is kept constant and $p$ varies). Combining with the chart $\eta^{-1}$, we get the coordinate representation

$$\eta^*(\mathrm{add}_q^*(dx/y)) = \frac{d(X(q_x, q_y, z, \eta_y(z)))}{Y(q_x, q_y, z, \eta_y(z))} = \frac{(X(q_x, q_y, z, \eta_y(z)))'}{Y(q_x, q_y, z, \eta_y(z))} dz.$$

Similarly we have $\eta^*(dx/y) = (1/\eta_y)\, dz$. By Exercise 12.79,

$$\lambda(q) = \frac{\eta^*(\mathrm{add}_q^*(dx/y))}{\eta^*(dx/y)}.$$

By Exercise 13.3 and the chain rule (Proposition 13.2), $(X(q_x, q_y, z, \eta_y(z)))'$ is a rational function of $q_x$, $q_y$, $z$, $\eta_y(z)$ and $\eta_y'(z)$. Combining, there is a rational function $R$ such that for any $z \in U$,

$$\lambda(q) = R(q_x, q_y, z, \eta_y(z), \eta_y'(z)),$$

and note that this does not depend on $z$, only on $q$. Hence, fixing some $z_0 \in U$, we see that on $W$, $\lambda(q)$ is a rational function of $q$, and so is holomorphic on $W$.

It remains to show that $\lambda$ is holomorphic on a neighbourhood of $0_E$. Before we do so, we observe that as any holomorphic form on $E$ is a constant multiple of $dx/y$, we have shown that for any holomorphic form $\omega$ on $E$, the map $q \mapsto (\mathrm{add}_q^*\omega)/\omega$ is holomorphic on $E \cap \mathbb{A}^2$.

Now, we change coordinates. Since there is more than one flex on $E$ (Exercise 7.35 or Exercise 7.36), by Proposition 7.23 and Remark 7.25 there is a change of coordinates $\beta$ of $\mathbb{P}^2(\mathbb{C})$ so that $\beta[E]$ is given by $y^2 = \widetilde{f}(x)$ but $\beta(0_E) \in \mathbb{A}^2$. The change of coordinates $\beta$ translates $\mathrm{add}_q$ to $\mathrm{add}_{\beta(q)}$: $\beta\restriction_E$ is a group isomorphism from $(E, 0_E)$ to $(\beta[E], \beta(0_E))$ (Exercise 7.38), so the map $\mathrm{add}_{\beta(q)}$, computed in $(\beta[E], \beta(0_E))$, is the map $\beta \circ \mathrm{add}_q \circ \beta^{-1}$, where $\mathrm{add}_q$ is computed in $(E, 0_E)$. Further, the restriction of $\beta$ to $E$ is a biholomorphism between $E$ and $\beta[E]$ (Exercise 13.19). By Exercise 12.79, $\lambda(q) = (\mathrm{add}_{\beta(q)})^*((\beta^{-1})^*(dx/y))/(\beta^{-1})^*(dx/y)$ for all $q$. As observed, the map $p \mapsto (\mathrm{add}_p)^*((\beta^{-1})^*(dx/y))/(\beta^{-1})^*(dx/y)$ is holomorphic on $\beta[E] \cap \mathbb{A}^2$. The map $\lambda$ is the composition of this map with $\beta$, and so is holomorphic on $E \setminus \{\beta^{-1}(0_E)\}$, whence it is holomorphic on a neighbourhood of $0_E$. □

*Exercise 14.31* Let $(E, 0_E)$ be an elliptic curve. (a) Show that there is a non-vanishing holomorphic form $\omega$ on $E$. (b) Show that any such form is invariant under addition in $E$. (Use Remark 7.25.) «

### 14.3.4 Proof of the Inversion Theorem

We can finally prove:

**Theorem 14.32** *Every elliptic curve is isomorphic to a complex torus.*

***Proof*** Let $(E, 0_E)$ be an elliptic curve. By Exercise 14.31, let $\omega$ be a non-vanishing, invariant form on $(E, 0_E)$. The proof has three steps.

**I.** There are open neighbourhoods $U \subseteq V$ of 0 in $\mathbb{C}$ and a biholomorphism $h$ from $V$ to an open neighbourhood of $0_E$ in $E$, such that for all $z, w \in U$, $z + w \in V$ and $h(z + w) = h(z) + h(w)$, where the former addition is in $\mathbb{C}$ and the latter in $(E, 0_E)$.

Let $\psi$ be a chart for $E$ whose domain contains $0_E$. Let $f\,dz = (\psi^{-1})^*\omega$ (in other words $f\,dz$ is the local representation of $\omega$ using $\psi$-coordinates). Since $\omega$ is holomorphic, $f$ is analytic on range $\psi$. By shrinking, we may assume that range $\psi$ is simply connected. Hence (Proposition 11.19), $f$ has a primitive $g$ on range $\psi$; i.e., $f\,dz = dg$. Since $\omega$ is non-vanishing, $f$ does not have a zero; by Theorem 11.11, let $W$ be an open neighbourhood of $0_E$ in $E$ such that $g \circ \psi$ is a biholomorphism from $W$ to an open set $V$. Let $G = g \circ \psi\restriction_W$. By adding a constant to $g$, we may assume that $G(0_E) = 0$.

Since addition on $E$ is continuous (Proposition 13.55), let $O \subseteq W$ be an open neighbourhood of $0_E$ in $E$ such that for all $p, q \in O$, $p + q \in W$; let $U = G[O]$. We will let $h = G^{-1}$, once we show that $G(p+q) = G(p)+G(q)$ for all $p, q \in O$.

To see that $G$ preserves addition on $O$, let $q \in O$. Translating $\omega = \mathrm{add}_q^*\omega$ by the chart $\psi$, we get $dg = \overline{\mathrm{add}}_q^*\, dg$, where $\overline{\mathrm{add}}_q = \psi \circ \mathrm{add}_q \circ \psi^{-1}$ is the coordinate representation of $\mathrm{add}_q$ using $\psi$. By Exercise 12.85, $\overline{\mathrm{add}}_q^*(dg) = d(g \circ \overline{\mathrm{add}}_q)$; which means $g' = (g \circ \overline{\mathrm{add}}_q)'$, whence $g \circ \overline{\mathrm{add}}_q - g$ is constant $c(q)$. Pulling back by $\psi$, we get $G \circ \mathrm{add}_q = c(q) + G$. Evaluating at $0_E$ (and using $G(0_E) = 0$), we get $c(q) = G(q)$. That is, for all $p \in O$, $G(p + q) = (G \circ \mathrm{add}_q)(p) = G(q) + G(p)$, as required.

**II.** There is a holomorphic group homomorphism $H\colon (\mathbb{C}, +) \to (E, 0_E)$.

We will get $H$ by extending $h$ to all of $\mathbb{C}$. To avoid confusion, for $n \in \mathbb{N}$ and $p \in E$ let $[n]p = p+p+\cdots+p$ ($n$ times), addition performed in $E$. For simplicity, replace $U$ by an open ball $B(0, r) \subseteq U$; this allows us to assume that for all $z \in \mathbb{C}$ and $n \in \mathbb{N}$, if $z/n \in U$ then for all $k \geqslant n$, $z/k \in U$ as well. Since $h$ preserves addition, for all $z \in \mathbb{C}$, if $z/n, z/m \in U$ then

$$[n]h(z/n) = [m]h(z/m).$$

[Why? by replacing $m$ by a common multiple, we may assume that $n$ divides $m$. Apply additivity of $h$ to get $h(z/n) = [m/n]h(z/m)$; then multiply by $n$ in the group $(E, 0_E)$.]

On the other hand, for all $z$, for all sufficiently large $n$, we have $z/n \in U$. Thus, we can define $H\colon \mathbb{C} \to E$ by letting $H(z) = [n]h(z/n)$ for some (any) $n \in \mathbb{N}$ such that $z/n \in U$. For a fixed $n$, the map $z \mapsto z/n$ is analytic, and the map $p \mapsto [n]p$ is holomorphic (Proposition 13.57). This implies that the map $H$ restricted to $n \cdot U$ is holomorphic; hence $H$ is holomorphic. Further, for all $z, w \in \mathbb{C}$, if $z/m, w/m, (z+w)/m \in U$, then

$$H(z+w) = [m]h((z+w)/m) = [m](h(z/m) + h(w/m)) = H(z) + H(w),$$

i.e., $H$ is indeed a group homomorphism.

**III.** The kernel $\Gamma$ of $H$ is a lattice in $\mathbb{C}$, and the induced map on the quotient $\bar{H}\colon \mathbb{C}/\Gamma \to E$ is a holomorphic group isomorphism.

Since $H$ is a group homomorphism, the range of $H$ is a subgroup of $(E, 0_E)$. Since $H$ extends $h$, it is not constant. Since $H$ is holomorphic and nonconstant, the range of $H$ is open (Proposition 12.39). However, an open subgroup of a topological group is also closed (Exercise 8.117). Since $E$ is connected (Corollary 13.38), $H$ is onto $E$.

Let $\Gamma$ be the kernel of $H$; it is a subgroup of $\mathbb{C}$. Since $H$ is holomorphic and nonconstant, $\Gamma$ is discrete (Corollary 12.10), and so is generated by $\mathbb{R}$-linearly independent points (Proposition 8.106). Now the induced map on the quotient $\bar{H}\colon \mathbb{C}/\Gamma \to E$ is a bijection. It is holomorphic (Lemma 14.3 and Exercise 14.7). Since $E \cong \mathbb{C}/\Gamma$ is compact, $\Gamma$ is not cyclic (Exercise 14.7), i.e., it is a 2-dimensional lattice. This completes the proof. □

## 14.4 Further Exercises

14.33 Let $r > 2$ be a real number. Show that

$$s_r = \sum_{u \in \Gamma \setminus \{0\}} \frac{1}{u^r}.$$

converges absolutely, and that $s_r = 0$ if $r$ is an odd natural number.

**Singly Periodic Functions**

14.34 (a) Let $N > 0$. Show that

$$\sum_{n \in \mathbb{Z}, |n| > N} \left( \frac{1}{z-n} + \frac{1}{n} \right)$$

converges absolutely uniformly on $B(0, N)$. (b) Show that

$$\frac{1}{z} + \sum_{n \in \mathbb{Z} \setminus \{0\}} \left( \frac{1}{z-n} + \frac{1}{n} \right)$$

converges locally uniformly on $\mathbb{C} \setminus \mathbb{Z}$, and so its sum $f$ is analytic on $\mathbb{C} \setminus \mathbb{Z}$. Calculate $f'$. (As above we use Proposition 11.34.) (c) Show that $f$ is meromorphic on $\mathbb{C}$. (d) Show that $f'$ and $f$ are $\mathbb{Z}$-periodic. (e) Show that the induced meromorphic function on $\mathbb{C}/\mathbb{Z}$ has one simple pole (pole of order 1), but is not a biholomorphism with $\mathbb{P}^1(\mathbb{C})$.

14.35 Let $f : \mathbb{C} \to \mathbb{C}$ be analytic and suppose that $2\pi i$ is a period of $f$. Show that

$$f(z) = \sum_{n=-\infty}^{\infty} c_n e^{nz}$$

converges absolutely and locally uniformly, where

$$c_n = \frac{1}{2\pi i} \int_0^{2\pi i} f(z) e^{-nz},$$

(the integration taking place along any path in $\mathbb{C}$ from 0 to $2\pi i$). (Hint: Exercise 12.95). This series expansion is the *Fourier series* of $f$. (Compare with Exercise 12.97.)

14.36 What are the Fourier series of $\sin iz$ and $\cos iz$? (See [Rem91, Sec.3.12.4] for more complicated examples.)

**Elliptic Functions**

14.37 Use Liouville's theorem to directly show that no elliptic function is entire (it must have poles).

14.38 Let $\Gamma$ be a lattice in $\mathbb{C}$. (a) Show that the collection of $\Gamma$-periodic meromorphic functions is a subfield of the field of meromorphic functions (Exercise 12.22). (b) Show that if $f$ is $\Gamma$-periodic then so is $f'$.

14.39 Let $f$ be an elliptic function with group of periods $\Gamma$, generated by $\alpha$ and $\beta$. For $q \in \mathbb{C}$ let $\gamma_q$ be the loop which circumnavigates the boundary of the parallelogram determined by $q$, $q + \alpha$, $q + \beta$ and $q + \alpha + \beta$. (a) Show that there is some $q \in \mathbb{C}$ such that the image of $\gamma_q$ does not contain any pole of $f$. (b) Show that for such $q$, $\int_{\gamma_q} f\, dz = 0$. (c) Applying this to $f'/f$, show that the sum of the orders of all zeros and poles of $f$ in the interior of $\gamma_q$ is 0. (d) Using the fact that $f$ is not entire (Exercise 14.37), conclude that $f$ does have a zero. (e) By applying this

to $f - c$ for any $c \in \mathbb{C}$, conclude that the range of $f$ is $\mathbb{C}$, and that the sum of the orders of zeros of $f - c$ in the interior of $\gamma_q$ is independent of $c$. (f) Conclude that the induced $\bar{f}\colon T_\Gamma \to \mathbb{P}^1(\mathbb{C})$ has a well-defined degree (that is, give an alternative proof of Lemma 12.47 for $\bar{f}$).

14.40 (a) Show that every elliptic function $f$ (with group of periods $\Gamma$) has a primitive on $\mathbb{C}\backslash\Gamma$. (b) The primitive $g$ may fail to be $\Gamma$-periodic (see Exercise 14.45). Show, however, that it is *quasi-periodic*: there is a linear map $h\colon \mathbb{C} \to \mathbb{C}$ such that for all $z \in \mathbb{C} \setminus \Gamma$ and $u \in \Gamma$, $g(z+u) = g(z) + h(u)$.

14.41 Let $f$ be a nonconstant meromorphic function on $T = T_\Gamma$. Show that

$$\sum_{x \in T} \operatorname{ord}_x(f) \cdot x = 0,$$

with addition taken in the group $(T, +)$. (Hint: one way to do this is integrate the (non-elliptic) function $zf'/f$ around a parallelogram as in Exercise 14.39; evaluate the integral using Exercise 11.93.) Use this to give another proof that an elliptic function cannot have degree 1.

14.42 Let $\Gamma$ and $\Gamma'$ be two lattices in $\mathbb{C}$. Show that $T_\Gamma$ and $T_{\Gamma'}$ are biholomorphic if and only if $\Gamma' = \lambda\Gamma$ for some $\lambda \in \mathbb{C} \setminus \{0\}$. (There are at least two ways to do this. The first is to show, via liftings, that a biholomorphism from $T_\Gamma$ to $T_{\Gamma'}$ is induced by a biholomorphism from $\mathbb{C}$ to itself, and then use Exercise 12.105. Another is to use Lemma 14.28 and Corollary 12.82 (and Exercise 12.91).)

14.43 Let $\Gamma$ be a lattice in $\mathbb{C}$. Define $\Delta(\Gamma) = \gamma_2(\Gamma)^3 - 27\gamma_3(\Gamma)^2$ (which is nonzero by Lemma 7.27) and $J(\Gamma) = \gamma_2(\Gamma)^3/\Delta(\Gamma)$. (a) Show that if $T_\Gamma$ and $T_{\Gamma'}$ are biholomorphic, then $J(\Gamma) = J(\Gamma')$. (b) Show that if $J(\Gamma) = J(\Gamma')$ then there is some $\mu \in \mathbb{C}$ such that $\gamma_2(\Gamma') = \mu^2\gamma_2(\Gamma)$ and $\gamma_3(\Gamma') = \mu^3\gamma_3(\Gamma)$. (Show that $(\gamma_2(\Gamma')/\gamma_2(\Gamma))^3 = (\gamma_3(\Gamma')/\gamma_3(\Gamma))^2$, and then consider the cases $\gamma_i(\Gamma) = 0$ or not.) (c) Show that if $J(\Gamma) = J(\Gamma')$ then there is a change of coordinates of $\mathbb{P}^2(\mathbb{C})$ mapping $E_\Gamma$ to $E_{\Gamma'}$. (d) Conclude that $E_\Gamma$ and $E_{\Gamma'}$ are biholomorphic if and only if one can be mapped to another using a change of coordinates.[3]

**The Weierstrass Function and the Isomorphism Theorem**

In the following exercises, fix a lattice $\Gamma$ and the associated $\wp = \wp_\Gamma$. Fix generators $\alpha$ and $\beta$ of $\Gamma$.

[3] It is possible to show that the $J$-invariant is onto $\mathbb{C}$. In particular, if $E$ is given by Weierstrass normal form $y^2 = 4x^3 - ax - b$ then there is some lattice $\Gamma$ satisfying $J(\Gamma) = a^3/(a^3 - 27b^2)$. From this we can conclude that there is a lattice $\Gamma$ such $\gamma_2(\Gamma) = a$ and $\gamma_3(\Gamma) = b$, i.e., such that $E = E_\Gamma$. It then follows that any two elliptic curves are holomorphically isomorphic if and only if they differ by a change of coordinates. See, for example, [Lan87, Sec.3.3].

14.44 (a) What is the order of the zeros of $\wp'$? (b) By counting orders of zeros and poles, show that

$$\frac{(\wp - \wp(\alpha/2))(\wp - \wp(\beta/2))(\wp - \wp(\alpha/2+\beta/2))}{(\wp')^2}$$

is constant.[4]

14.45 Show that the primitive of the Weierstrass function $\wp$ (Exercise 14.40) is not $\Gamma$-periodic.[5]

14.46 Let $f$ be an elliptic function with group of periods $\Gamma$. Suppose that $f$ is even. Let $a$ be a zero of $f$ and let $b$ be a pole of $f$. Define $g$ according to the following cases:

- If $a, b \notin \Gamma$, $g(z) = f(z)(\wp(z) - \wp(b))/(\wp(z) - \wp(a))$.
- If $a \in \Gamma$, $b \notin \Gamma$, $g(z) = f(z)(\wp(z) - \wp(b))$.
- If $a \notin \Gamma$, $b \in \Gamma$, $g(z) = f(z)/(\wp(z) - \wp(a))$.

(a) Show that $g$ is either constant or elliptic, of degree smaller than the degree of $f$. (b) Show that there is a rational function $R$ such that $f = R(\wp)$. (c) Show that the field of $\Gamma$-periodic meromorphic functions (Exercise 14.38) is generated by $\wp$ and $\wp'$.

14.47 (a) Show that for all $z_1, z_2 \in \mathbb{C} \setminus \Gamma$ such that $z_1 + \Gamma \neq z_2 + \Gamma, -z_2 + \Gamma$,

$$\wp(z_1 + z_2) = \frac{1}{4}\left(\frac{\wp'(z_1) - \wp'(z_2)}{\wp(z_1) - \wp(z_2)}\right)^2 - \wp(z_1) - \wp(z_2). \tag{14.4}$$

(b) Show that if $z \notin \Gamma$ and $\wp'(z) \neq 0$ then

$$\wp(2z) = \frac{1}{4}\left(\frac{\wp''(z)}{\wp'(z)}\right)^2 - 2\wp(z). \tag{14.5}$$

(Use Exercise 7.30 and the isomorphism theorem.)

[4] This is an alternative way to present the argument for the differential equation for $\wp$.

[5] The primitive of $\wp$ is called the *Weierstrass $\zeta$-function*, not to be confused with the more famous Riemann $\zeta$-function.

14.48 Let $z$, $w$ and $u$ be distinct and nonzero modulo $\Gamma$. Show that $z + w + u \in \Gamma$ if and only if

$$\det \begin{pmatrix} 1 & \wp(z) & \wp'(z) \\ 1 & \wp(w) & \wp'(w) \\ 1 & \wp(u) & \wp'(u) \end{pmatrix} = 0.$$

(Use Exercise 4.13 and the isomorphism theorem.)

14.49 In this exercise we show another derivation of Eq. (14.4) avoiding the isomorphism theorem. Let $y = ax + b$ be an affine line which intersects $E_\Gamma$ in three distinct points $(\wp(z_i), \wp'(z_i))$ for $i = 1, 2, 3$. (a) Show that the function $\wp' - (a\wp + b)$ is $\Gamma$-periodic, of degree 3. (b) Use Exercise 14.41 to show that $z_1 + z_2 + z_3 \in \Gamma$. (c) Use this to derive Eq. (14.4) (use continuity considerations when the line is a tangent to $E_\Gamma$).[6]

14.50 Let $(E, 0_E)$ be an elliptic curve. Show that for all $n \geqslant 2$, the collection of points $p \in E$ whose order in the group $(E, 0_E)$ divides $n$, is a subgroup of $(E, 0_E)$ isomorphic to $C_n \times C_n$.

**Elliptic Integrals**

14.51 Let $\Gamma$ be a lattice in $\mathbb{C}$; let $\varphi\colon T_\Gamma \to E_\Gamma$ be as in the proof of the isomorphism theorem. (a) Show that for all $p \in E_\Gamma$, $\varphi^{-1}(p) = \int_\gamma dx/y + \Gamma$, where $\gamma$ is any path in $E$ from $0_E$ to $p$. (b) Conclude that if $p, q, r \in E_\Gamma$ are collinear then

$$\int_{0_E}^{p} dx/y + \int_{0_E}^{q} dx/y + \int_{0_E}^{r} dx/y \in \Gamma,$$

where $\int_{0_E}^{p} dx/y$ denotes the integral of $dx/y$ along a path in $E_\Gamma$ from $0_E$ to $p$.[7]

14.52 Let $E$ be a nonsingular cubic curve defined by $y^2 = f(x)$, and let $\xi$ be a (piecewise smooth) path in $E$. Show that if $\xi$ is a lifting of a path $\gamma$, which avoids the roots of $f$, then

$$\int_\xi \frac{dx}{y} = \int_\gamma \frac{dz}{\sqrt{f(z)}},$$

[6] This can be used to give another proof that $\varphi$ is a group isomorphism; see, for example, [Lan87, Sec.1.3] or [Was08, Thm.9.10].

[7] This is a presentation of the isomorphism theorem as an *addition formula* for integrals.

where by $\sqrt{f(z)}$ we mean the continuous choice of a square root on the image of $f \circ \gamma$ which determines the lifting $\xi$.[8]

14.53 Again let $E$ be a nonsingular cubic curve defined by $y^2 = f(x)$. Let $a \in \mathbb{C}$. Let $\gamma : [0, 1] \to \mathbb{P}^1(\mathbb{C})$ be a path from $a$ to $p_\infty$ with $\gamma(t) \in \mathbb{C}$ and not a root of $f$ for $t \in (0, 1)$, and let $\xi$ be a lifting of $\gamma$ to a path in $E$ from $p$ to $0_E$ (see Exercise 13.69). Show that

$$\int_\xi \frac{dx}{y} = \int_\gamma \frac{dz}{\sqrt{f(z)}} = \lim_{t \to 1} \int_{\gamma|_{[0,t]}} \frac{dz}{\sqrt{f(z)}},$$

where by $\int_\gamma dz/\sqrt{f(z)}$ we mean the integral in $\mathbb{P}^1(\mathbb{C})$ using the form on $\mathbb{P}^1(\mathbb{C})$ extending $dz$ (see Example 12.87). We denote this integral by $\int_a^\infty dz/\sqrt{f(z)}$.

Conclude that if $E = E_\Gamma$ then for all $z \notin \Gamma$,

$$z = \int_\infty^{\wp(z)} \frac{dz}{\sqrt{4z^3 - \gamma_2 z - \gamma_3}}$$

modulo $\Gamma$.[9]

[8] See Exercise 13.68. Such an integral is called an *elliptic integral*, since the arc-length of an ellipse can be expressed as such an integral.

[9] That is, the elliptic function $\wp$ can be defined as the inverse of the elliptic integral $\int dz/\sqrt{f(z)}$. As we shall mention in Chap. 16, it was Abel and Jacobi's fundamental insight that the inverses of such elliptic integrals can be extended to the complex numbers and become doubly periodic functions.


# Puiseux Theory

# 15

Our very first definition of the intersection multiplicity of a line with a curve (Definition 5.25) relied on a parameterisation of the line: in affine coordinates, we let $t \mapsto \alpha(t)$ be a parameterisation of the line; we then define the intersection multiplicity of the line with the curve $f = 0$ at a point $p = \alpha(\lambda)$ to be the multiplicity of the root $\lambda$ of the polynomial $f(\alpha(t))$. In this chapter we show how we can extend this idea to intersections between any two curves. The issue, of course, is that most curves will not have rational parameterisations.

When we deal with nonsingular points, the implicit function theorem gives us vertical analytic parameterisations of curves, and the results of this chapter show that we could similarly use such parameterisations to compute intersection multiplicities. Say $f \in \mathbb{C}[x, y]$ defines the affine part of a curve $D$, and $\eta(z) = (z, \eta_y(z))$ is a vertical parameterisation of a curve $C$. Let $p = \eta(a)$. The composition $f \circ \eta = f(z, \eta_y(z))$ is analytic and so has an order at $a$ (Definition 12.13); we will see that this order equals $i_p(C, D)$.

At singular points, we cannot have vertical parameterisations, but we will show that there will always be analytic parameterisations $\psi = (\psi_x, \psi_y)$ of the curve close to a point. But such parameterisations will sometimes not cover all of the curve near the point. Rather, we will show that the curve near a singular point $p$ can consist of several separate *branches*, each with their own analytic parameterisations. The standard example are the two branches of the nodal cubic $y^2 = x^3 + x^2$ at the origin, where the curve intersects itself.

We will show how to define these branches (more precisely, their "germs" at the singular point, which are called *places*). We will show how to give implicit definitions of each place separately, using convergent power series. We use parameterisations and these implicit definitions to define intersection multiplicity between places. We then finally show that intersections of curves reduce to counting the intersections of their various places, and adding up the results. This will allow us to give much more intuitive proofs of some of our results about intersection multiplicity from Chap. 6.



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_15

## 15.1 Fractional Power Series and Their Holomorphic Functions

In Chap. 2 we introduced the ring of formal power series $\mathbb{C}[\![x]\!]$. In Chap. 11 we discussed the sums of power series. We now consider the relationship between these. We then extend this to power series which allow fractional exponents.

### 15.1.1 Formal and Informal Power Series

One obvious difference between formal power series and analytic functions is that formal power series in $\mathbb{C}[\![x]\!]$ have no convergence criteria. In other words, many power series in that ring do not define analytic functions because their radius of convergence is 0. For $f \in \mathbb{C}[\![x]\!]$ let $R(f)$ be the radius of convergence of $f$. For all $f \in \mathbb{C}[\![x]\!]$ such that $R(f) > 0$ we let $\mathfrak{f}$ be the analytic function on $B(0, R(f))$ defined by $f$. The following lemma says that the map $f \mapsto \mathfrak{f}$ is injective, and preserves the ring operations of $\mathbb{C}[\![x]\!]$.

**Proposition 15.1** *Let $f, g \in \mathbb{C}[\![x]\!]$, and suppose that $R(f), R(g) > 0$. Let $R = \min\{R(f), R(g)\}$.*

1. *If there is a neighbourhood of 0 on which $\mathfrak{f} = \mathfrak{g}$ then $f = g$.*
2. *If $s = f + g$ is the formal sum of $f$ and $g$ in $\mathbb{C}[\![x]\!]$ then $R(s) \geqslant R$ and $\mathfrak{s} = \mathfrak{f} + \mathfrak{g}$ on $B(0, R)$.*
3. *If $p = fg$ is the formal product of $f$ and $g$ in $\mathbb{C}[\![x]\!]$ then $R(p) \geqslant R$ and $\mathfrak{p} = \mathfrak{f}\mathfrak{g}$ on $B(0, R)$.*

***Proof*** (a) follows from Proposition 11.58. (b) follows from Exercise 8.78.

For a direct proof of (c) see Exercise 15.102. We give a "reverse" proof. We know that $\mathfrak{f}\mathfrak{g}$ is defined on $B(0, R)$ and analytic there (Proposition 11.7 and Theorem 11.67). By Corollary 11.70, there is a power series $h$ with $R(h) \geqslant R$ and $\mathfrak{h} = \mathfrak{f}\mathfrak{g}$ on $B(0, R)$; we just need to show that $h = fg$. Iterating the product rule for derivatives, by induction on $n$ we see that $(\mathfrak{f}\mathfrak{g})^{(n)} = \sum_{k \leqslant n} \binom{n}{k} \mathfrak{f}^{(k)} \mathfrak{g}^{(n-k)}$ (where $\mathfrak{f}^{(0)} = \mathfrak{f}$); this uses $\binom{n+1}{k+1} = \binom{n}{k+1} + \binom{n}{k}$. Letting $f = \sum a_n x^n$, $g = \sum b_n x^n$, $h = \sum c_n x^n$ and $fg = \sum d_n x^n$, by Proposition 11.58, for all $n$,

$$c_n = \frac{(\mathfrak{f}\mathfrak{g})^{(n)}(0)}{n!} = \sum_{k \leqslant n} \frac{\binom{n}{k}}{n!} \mathfrak{f}^{(k)}(0) \mathfrak{g}^{(n-k)}(0) =$$

$$\sum_{k \leqslant n} \frac{\mathfrak{f}^{(k)}(0)}{k!} \frac{\mathfrak{g}^{(n-k)}(0)}{(n-k)!} = \sum_{k \leqslant n} a_k b_{n-k} = d_n$$

as required. □

### Germs

Proposition 15.1 implies that the collection of $f \in \mathbb{C}[\![x]\!]$ with $R(f) > 0$ is a subring of $\mathbb{C}[\![x]\!]$, the ring of *convergent power series*. We would like the map $f \mapsto \mathfrak{f}$ to be a ring isomorphism between the ring of convergent power series and a ring of analytic functions (with pointwise addition and multiplication). However, the definition of this ring is a little awkward, since different convergent power series have different convergence radii: the domains of the functions $\mathfrak{f}$ vary. And it is possible, for example, that $h = f + g$ but $R(h) > R(f), R(g)$, in which case we cannot get all the values of $\mathfrak{h}$ by adding the values of $\mathfrak{f}$ and $\mathfrak{g}$.

A solution to this problem is the notion of a *germ* of an analytic function. Let $\varphi, \psi$ be two analytic functions which are defined on open neighbourhoods of 0. We say that $\varphi$ and $\psi$ are *essentially equal* if $\varphi = \psi$ on some neighbourhood of 0. This is an equivalence relation on analytic functions defined at 0, and the equivalence classes are called *germs of analytic functions at 0*. The idea is that the germ tells us what happens arbitrarily close to 0.

Germs of analytic functions can be added and multiplied. Let $\varphi_0, \varphi_1, \psi_0, \psi_1$ be analytic functions defined at 0; suppose that $\varphi_0$ and $\varphi_1$ are essentially equal, and so are $\psi_0$ and $\psi_1$. Then $\varphi_0 + \psi_0$ and $\varphi_1 + \psi_1$ are both defined on open neighbourhoods of 0, and are essentially equal. So if $\varphi$ and $\psi$ are *germs* of analytic functions at 0, then we can define their sum $\varphi + \psi$ to be the germ of $\varphi_0 + \psi_0$, where $\varphi_0$ and $\psi_0$ are analytic functions whose germs are $\varphi$ and $\psi$; the germ does not depend on the choice of representatives $\varphi_0$ and $\psi_0$. We can similarly define the product $\varphi\psi$. These operations make the collection of germs into a ring; indeed, it is an integral domain, as a nonzero germ only contains functions which are nonzero on a punctured neighbourhood of 0 (Proposition 11.50). Now Proposition 15.1 says that the map taking a convergent $f \in \mathbb{C}[\![x]\!]$ to the germ of $\mathfrak{f}$ is a ring isomorphism between the ring of convergent formal power series and the ring of germs.

The idea of germs of functions, measuring the "local behaviour" at a point, can be extended to other contexts, for example, germs of smooth functions on neighbourhoods of 0 in $\mathbb{R}$. For analytic functions, this concept is somewhat less crucial, because of their rigidity: essentially equal analytic functions are actually equal on any connected neighbourhood of 0 on which they are both defined; this is not so for merely smooth functions. Later in this chapter we will revisit the idea of germs, when considering parameterisations of curves and their branches. For the time being, we will not need them.

## 15.1.2 Substitutions into Power Series

The fact that $\mathbb{C}[\![x]\!]$ is a ring means that we can substitute power series in polynomials: if $g \in \mathbb{C}[\boldsymbol{y}]$ and $\boldsymbol{f}$ is a tuple of power series in $\mathbb{C}[\![x]\!]$ then $g(\boldsymbol{f})$ is well-defined. Proposition 15.1 shows:

**Proposition 15.2** *If $g \in \mathbb{C}[y_1, \ldots, y_m]$ is a polynomial and $f_1, \ldots, f_m \in \mathbb{C}[\![x]\!]$ then for $R = \min_{i \leqslant m} R(f_i)$, letting $h = g(f_1, \ldots, f_m)$ and assuming that $R > 0$, we have that $R(h) \geqslant R$ and $\mathfrak{h} = g(\mathfrak{f}_1, \ldots, \mathfrak{f}_m)$ on $B(0, R)$.*

In general, it is meaningless to substitute power series into power series. However, it is possible to do when constant terms are missing. Let $f \in \mathbb{C}[\![x]\!]$ and suppose that $x \mid f$, that is, that the constant coefficient of $f$ is 0; in the notation of Exercise 2.21, $\operatorname{ord}(f) > 0$. For all $n$, $x^n \mid f^n$ (i.e., $\operatorname{ord}(f^n) \geqslant n$), so we can add $\sum c_n f^n$ for any choice of coefficients $c_n$; the coefficient of $x^m$ in this sum is determined by $\sum_{n \leqslant m} c_n f^n$. For any $g = \sum c_n x^n \in \mathbb{C}[\![x]\!]$ we let $g(f) = \sum c_n f^n$. This agrees with previous notation when $g$ is a polynomial.

*Exercise 15.3* Let $f, g \in \mathbb{C}[\![x]\!]$, and suppose that $\operatorname{ord}(g) > 0$. (a) Show that $\operatorname{ord}(f(g)) = \operatorname{ord}(f) \cdot \operatorname{ord}(g)$. (b) Let $h \in \mathbb{C}[\![x]\!]$, and suppose that $\operatorname{ord}(h) > 0$. Show that $(f(g))(h) = f(g(h))$. (c) Fix $g \in \mathbb{C}[\![x]\!]$ with $\operatorname{ord}(g) > 0$. Show that $f \mapsto f(g)$ is a ring homomorphism from $\mathbb{C}[\![x]\!]$ to itself. «

*Exercise 15.4* Recall (Exercise 2.2) that $\sum_{n \geqslant 0} x^n = 1/(1-x)$ in $\mathbb{C}[\![x]\!]$. Generalise this to show that for all $g \in \mathbb{C}[\![x]\!]$, if $\operatorname{ord}(g) > 0$ then $\sum_{n \geqslant 0} g^n = 1/(1-g)$. «

In more than one variable this works in a similar way. If $g \in \mathbb{C}[\![y_1, \ldots, y_m]\!]$, $f_1, \ldots, f_m \in \mathbb{C}[\![x]\!]$ and $\operatorname{ord}(f_i) > 0$ for each $i$, then $g(f_1, \ldots, f_n)$ is well-defined.

*Remark 15.5* Substitutions of formal power series, and the results mentioned so far, are valid over any integral domain $R$, not only $\mathbb{C}$. «

**Proposition 15.6** *Let $f, g \in \mathbb{C}[\![x]\!]$, and suppose that $\operatorname{ord}(f) > 0$. Let $h = g(f)$. Suppose that $R(f), R(g) > 0$. Then $R(h) > 0$, and $\mathfrak{h} = \mathfrak{g} \circ \mathfrak{f}$ on a neighbourhood of 0.*

We will not use Proposition 15.6 below, so we relegate the proof to Exercise 15.104. We will use a special case, which is easy: when $f = at$ (for nonzero $a \in \mathbb{C}$).

### 15.1.3 Fractional Power Series

A *fractional power series* is like a formal power series, except that we allow non-negative rational numbers as exponents of the variables. We want to multiply such series just like formal power series; this means that we have to restrict which such formal objects we allow. For example, if we allowed a series such as $g = \sum_{q \geqslant 0} x^q$ (we take all rational $q \geqslant 0$, each $x^q$ with coefficient 1) then it is not clear what the coefficient of any $x^q$ for $q > 0$ should be in the "series" $g^2$: there are infinitely many pairs of rationals $r, s \geqslant 0$ such that $r + s = q$, so summing the coefficients over all such pairs $(r, s)$ is impossible.

This is no longer a problem if we require a constant denominator for all fractions admitting nonzero coefficients: for example, if

$$f = 1 + x^{1/2} + x + x^{3/2} + x^2 + x^{5/2} + \cdots$$

and

$$g = 1 + x^{1/3} + x^{2/3} + x + x^{4/3} + \cdots$$

then we can unambiguously let

$$fg = 1 + x^{1/3} + x^{1/2} + x^{2/3} + x^{5/6} + 2x + x^{7/6} + 2x^{4/3} + 2x^{3/2} + \cdots$$

which is a series of the correct form since all denominators divide 6; the same holds for $f + g$. In other words, we allow series of the form $g(x^{1/n})$, the result of substituting the "variable" $x^{1/n}$ into a formal power series $g$. For an integral domain $R$, we let $R[\![x]\!]$ denote the ring of fractional power series (in indeterminate $x$) with coefficients from $R$.

*Exercise 15.7* Do all this more formally. That is: let $R$ be an integral domain. Define a *fractional power series* with coefficients from $R$ to be a formal object $\sum_{q \geqslant 0} a_q x^q$, for rational numbers $q \geqslant 0$, where $a_q \in R$, such that for some $n \geqslant 1$, $a_q = 0$ if $q \neq k/n$ for some integer $k \geqslant 0$. Define addition and multiplication of fractional power series, and show that $R[\![x]\!]$, equipped with these operations, is an integral domain; show that $R[x]$ is a subring of $R[\![x]\!]$. «

We let $R[x^{1/n}]$ be the collection of fractional power series $f \in R[\![x]\!]$ of the form $g(x^{1/n})$ for a power series $g$. It is a subring of $R[\![x]\!]$; $R[x^{1/n}] \subseteq R[x^{1/m}]$ if and only if $n$ divides $m$, and $R[\![x]\!] = \bigcup_n R[x^{1/n}]$.

*Exercise 15.8* (a) Show that $R[x^{1/n}] \cap R[x^{1/m}] = R[x^{1/\gcd(n,m)}]$ (where $\gcd(n, m)$ is the greatest common divisor of $n$ and $m$). (b) Conclude that for $f \in R[\![x]\!]$, if $n$ is the least such that $f \in R[x^{1/n}]$, then for all $m$, $f \in R[x^{1/m}]$ if and only if $n$ divides $m$. «

*Exercise 15.9* For nonzero $f = \sum a_q x^q \in R[\![x]\!]$ let $\operatorname{ord}(f) = q$ for the least $q$ such that $a_q \neq 0$ (note that there is such a least $q$); as usual, let the order of the zero series be $\infty$. (a) Show that Exercise 2.21(a) holds for $R[\![x]\!]$ as well. (b) What are the units of $R[\![x]\!]$? (c) Suppose that $F$ is a field. Describe the association classes of $F[\![x]\!]$. (d) Show that in $F[\![x]\!]$ there are no irreducible elements. «

### 15.1.4 The Holomorphic Function Defined by a Fractional Power Series

Like formal power series, fractional power series also define functions—but to make them uni-valued, we need a choice of an $n$th root of the input. The Riemann surfaces $\Sigma$ and $\Sigma/n$ (see Sect. 12.4) allow us to do so, via the uni-valued root function $\mathrm{rt}_n$.

Recalling that $\pi_\Sigma\colon \Sigma \to \mathbb{C}\setminus\{0\}$ is the projection map $(z,t)\mapsto z$, for $r>0$ we let

$$B^*_\Sigma(0,r) = \{q\in\Sigma \,:\, 0<|\pi_\Sigma(q)|<r\}\,;$$

a *punctured neighbourhood of 0 in* $\Sigma$ is a subset of $\Sigma$ containing $B^*_\Sigma(0,r)$ for some $r>0$. We use the starred notation to indicate that $\Sigma$ does not contain a "copy" of 0. Now for $f=g(x^{1/n})\in\mathbb{C}[\![x^{1/n}]\!]$ (where $g\in\mathbb{C}[\![x]\!]$) we let $R(f)=R(g)^n$ and

$$\mathfrak{f} = \mathfrak{g}\circ \mathrm{rt}_n,$$

defined on $B^*_\Sigma(0,R(f))$. The following exercise shows that this does not depend on the choice of $n$:

*Exercise 15.10* Suppose that $n$ divides $m$. Let $f\in\mathbb{C}[\![x^{1/n}]\!]$; so $f=g(x^{1/n})=h(x^{1/m})$ (where $h(x)=g(x^{m/n})$). Show that $R(f)=R(g)^n=R(h)^m$ and that $\mathfrak{g}\circ\mathrm{rt}_n=\mathfrak{h}\circ\mathrm{rt}_m$ on $B^*_\Sigma(0,R(f))$. «

Since $\mathfrak{g}$ is analytic and $\mathrm{rt}_n$ is holomorphic (Proposition 12.58), $\mathfrak{f}$ is holomorphic on $B^*_\Sigma(0,R(f))$. Proposition 15.1 implies its extension to fractional power series; the map $f\mapsto\mathfrak{f}$ is injective on the collection of $f\in\mathbb{C}[\![x]\!]$ with $R(f)>0$, and preserves addition and multiplication, and hence substitution into polynomials (Proposition 15.2):

**Proposition 15.11**

*(a) Let $f,g\in\mathbb{C}[\![x]\!]$, and suppose that $R(f),R(g)>0$. If there is a punctured neighbourhood of 0 in $\Sigma$ on which $\mathfrak{f}=\mathfrak{g}$, then $f=g$.*

*(b) If $g\in\mathbb{C}[y_1,\dots,y_m]$ is a polynomial, $f_1,\dots,f_m\in\mathbb{C}[\![x]\!]$, and $R=\min_{i\leqslant m}R(f_i)>0$, then for $h=g(f_1,\dots,f_m)$ we have $R(h)\geqslant R$ and $\mathfrak{h}=g(\mathfrak{f}_1,\dots,\mathfrak{f}_m)$ on $B^*_\Sigma(0,R)$.*

*Example 15.12* The series $x$ defines the function $\pi_\Sigma$ on $\Sigma$, as $\mathrm{rt}_1=\pi_\Sigma$. More generally, every convergent $g\in\mathbb{C}[\![x]\!]$ defines the function $\mathfrak{g}\circ\pi_\Sigma$ on $B^*_\Sigma(0,R(g))$.«

### The Induced Function on the Root Surface

The $n$th root function $\mathrm{rt}_n$ is invariant under the $n$th iteration $\mathrm{sh}_n$ of the shift map on $\Sigma$ (see page 330), and so for any $f \in \mathbb{C}[\![x^{1/n}]\!]$, the map $\mathfrak{f}$ is also invariant under $\mathrm{sh}_n$, that is, $\mathfrak{f} = \mathfrak{f} \circ \mathrm{sh}_n$. By Lemma 12.60, $\mathfrak{f}$ induces a holomorphic function on $B^*_{\Sigma/n}(0, R(f))$ (where as expected, $B^*_{\Sigma/n}(0, r) = \{q \in \Sigma/n : 0 < |\pi_{\Sigma/n}(q)| < r\}$ is the image of $B^*_\Sigma(0, r)$ under the quotient map $(z, t) \mapsto (z, t + 2\pi n\mathbb{Z})$ from $\Sigma$ to $\Sigma/n$; recall that $\pi_{\Sigma/n}$ is the induced projection map $(z, t + 2\pi n\mathbb{Z}) \mapsto z$). We denote this function also by $\mathfrak{f}$, and note that in this sense as well, if $f = g(x^{1/n})$ then $\mathfrak{f} = \mathfrak{g} \circ \mathrm{rt}_n$, where we now take $\mathrm{rt}_n : \Sigma/n \to \mathbb{C} \setminus \{0\}$ to be the induced $n$th root function on $\Sigma/n$.

Every analytic function is defined by a power series; indeed, by Corollary 11.70, if $\psi : B(0, r) \to \mathbb{C}$ is analytic, then there is some $g \in \mathbb{C}[\![x]\!]$ such that $R(g) \geqslant r$ and $\psi = \mathfrak{g}$ on $B(0, r)$. When considering holomorphic functions on punctured neighbourhoods of 0 in $\Sigma/n$, we need to be mindful that they are not defined "at zero" and so may have poles, or even essential singularities there, in which case we would need negative exponents in a fractional Laurent series defining these functions. We do have the following:

**Proposition 15.13** *Let $n \geqslant 1$ and $r > 0$; suppose that $\psi : B^*_{\Sigma/n}(0, r) \to \mathbb{C}$ is holomorphic and bounded. Then there is a fractional power series $f \in \mathbb{C}[\![x^{1/n}]\!]$ such that $R(f) \geqslant r$ and $\psi = \mathfrak{f}$ on $B^*_{\Sigma/n}(0, r)$.*

***Proof*** Let $\varphi = \psi \circ \mathrm{pwr}_n$, where recall that $\mathrm{pwr}_n : \mathbb{C} \setminus \{0\} \to \Sigma/n$ is the inverse of $\mathrm{rt}_n$. Then $\varphi$ is defined on the punctured neighbourhood $B^*(0, \sqrt[n]{r})$, and is analytic and bounded there. By Proposition 12.16, $\varphi$ can be extended to an analytic function on $B(0, \sqrt[n]{r})$; by Corollary 11.70, there is some $g \in \mathbb{C}[\![x]\!]$ such that this extension of $\varphi$ equals $\mathfrak{g}$ on its domain. Then $f = g(x^{1/n})$ is as required. □

*Remark 15.14* Note that a restricted converse holds: if $f \in \mathbb{C}[\![x]\!]$, $R(f) > 0$ and $r < R(f)$, then $\mathfrak{f}$ is bounded on $B^*_\Sigma(0, r)$, as $\mathfrak{g}$ is continuous on the closed disc $\overline{B}(0, \sqrt[n]{r})$ (where $f = g(x^{1/n})$). «

*Remark 15.15* We can add a "zero point" to $\Sigma/n$, and extend the bijections $\mathrm{rt}_n$ and $\mathrm{pwr}_n$ to match 0 with 0; this results in a Riemann surface, biholomorphic with $\mathbb{C}$. We didn't do this, since the resulting surface is not a topological subspace of $\mathbb{C} \times \mathbb{R}/2\pi n\mathbb{Z}$. We could similarly attach a zero point in the middle of $\Sigma$, but this is a topological space which is not a 2-manifold (and not a subspace of $\mathbb{C} \times \mathbb{R}$). «

### The Shift of a Fractional Power Series

If $f \in \mathbb{C}[\![x]\!]$ and $R(f) > 0$ then $\mathfrak{f} \circ \mathrm{sh}$ is holomorphic on $\Sigma$, and has the same range as $\mathfrak{f}$, so is defined by a fractional power series. We will now identify this series. For

$n \geqslant 1$, let $\omega_n = e^{2\pi i/n}$ be the "primitive" $n$th root of unity. For $f = \sum_k a_k x^{k/n} \in \mathbb{C}[\![x^{1/n}]\!]$ let

$$\mathrm{sh}(f) = \sum_k a_k \omega_n^k x^{k/n}.$$

That is, if $g = \sum_k a_k x^k$ and $f = g(x^{1/n})$, then $\mathrm{sh}(f) = h(x^{1/n})$, where $h = g(\omega_n x)$ is the result of substituting the formal power series $\omega_n x$ into $g$ (see page 397; note that $\mathrm{ord}(\omega_n x) = 1$). The following exercise implies that this is a well-defined operation on $f \in \mathbb{C}[\![x]\!]$:

*Exercise 15.16* Suppose that $n$ divides $m$. Let $f \in \mathbb{C}[\![x^{1/n}]\!]$; say $f = g(x^{1/n}) = \bar{g}(x^{1/m})$. Let $h = g(\omega_n x)$ and $\bar{h} = \bar{g}(\omega_m x)$. Show that $h(x^{1/n}) = \bar{h}(x^{1/m})$. «

Note that $\mathrm{sh}(f)$ itself is not the result of any substitution $f(ax)$, indeed, it is not clear what such a substitution should mean.

*Exercise 15.17* Show that $\mathrm{sh}\colon \mathbb{C}[\![x]\!] \to \mathbb{C}[\![x]\!]$ is a ring homomorphism. «

**Proposition 15.18** *For all $f \in \mathbb{C}[\![x]\!]$, $f = \mathrm{sh}(f)$ if and only if $f \in \mathbb{C}[x]$.*

***Proof*** If $f \in \mathbb{C}[x]$ then $\mathrm{sh}(f) = f$ since $\omega_1 = 1$. Suppose that $\mathrm{sh}(f) = f$, and say $f \in \mathbb{C}[\![x^{1/n}]\!]$. Write $f = \sum a_k x^{k/n}$; so $\mathrm{sh}(f) = \sum a_k \omega_n^k x^{k/n}$, so $a_k = \omega_n^k a_k$ for all $k$. If $n$ does not divide $k$ then $\omega_n^k \neq 1$, so $a_k = 0$. □

As promised:

**Proposition 15.19** *If $f \in \mathbb{C}[\![x]\!]$ and $R(f) > 0$, then $\mathrm{sh}(f)$ defines the map $\mathfrak{f} \circ \mathrm{sh}$.*

***Proof*** Say $f = g(x^{1/n})$ with $g \in \mathbb{C}[\![x]\!]$. By Exercise 12.65, for all $q \in B^*_\Sigma(0, R(f))$, $(\mathfrak{f} \circ \mathrm{sh})(q) = \mathfrak{g}(\omega_n \cdot \mathrm{rt}_n(q))$. Let $h = g(\omega_n x)$; then $\mathfrak{h}$ is the map $z \mapsto \mathfrak{g}(\omega_n z)$ (this is the easy case of Proposition 15.6 mentioned above). By Proposition 15.11, $\mathfrak{f} \circ \mathrm{sh}$ is defined by $h(x^{1/n})$. □

## 15.2 Parameterisations of a Curve

In this section we fix a curve $D$ in $\mathbb{P}^2(\mathbb{C})$. As in Chap. 13, we are really thinking of the underlying set of $D$, so we assume that $D$ has no repeated components. Recall that $D^*$ denotes the collection of nonsingular points of $D$, which is a holomorphic surface (Proposition 13.18).

**Definition 15.20** A *parameterisation* of $D$ is an injective continuous map $\psi : U \to D$ where $U \subseteq \mathbb{C}$ is an open neighbourhood of 0, such that the restriction of $\psi$ to $U^* = U \setminus \{0\}$ is a holomorphic function to $D^*$.

We call the point $\psi(0)$ the *centre* of the parameterisation $\psi$.

Thus, the centre of a parameterisation is allowed to be singular on $D$ (and this will be the interesting case). The restriction that all other points are nonsingular is easy to achieve, since there are only finitely many singular points on $D$ (Corollary 5.41); so if $\psi : U \to D$ is continuous and injective, there is a neighbourhood $V \subseteq U$ of 0 such that $\psi[V^*] \subseteq D^*$.

If the range of $\psi$ is contained in $D \cap \mathbb{A}^2$, then we write $\psi = (\psi_x, \psi_y)$.

**Lemma 15.21** *Let $\psi : U \to D \cap \mathbb{A}^2$ be continuous and injective, with $\psi[U^*] \subseteq D^*$. Then $\psi$ is a parameterisation of $D$ if and only if both $\psi_x$ and $\psi_y$ are analytic.*

***Proof*** By Exercise 13.26, $\psi\restriction_{U^*}$ is holomorphic if and only if both $\psi_x\restriction_{U^*}$ and $\psi_y\restriction_{U^*}$ are holomorphic. In that case, since $\psi$ is continuous at 0, by Proposition 11.73, $\psi_x$ and $\psi_y$ are analytic on $U$. □

*Example 15.22* An affine or projective linear parameterisation (Definitions 3.27 and 4.16) is a parameterisation. A rational parameterisation of an affine curve (Definition 3.29) can usually be restricted to a parameterisation as in Definition 15.20, provided that it is defined at 0 (finitely many points may need to be removed). A vertical parameterisation of a curve (Definition 13.9) is a parameterisation, provided that it is defined at 0. «

Changes of coordinates are injective, continuous (Exercise 8.60), map nonsingular points to nonsingular points (Proposition 5.21) and are holomorphic on $D^*$ (Exercise 13.19). This implies:

**Proposition 15.23** *If $\psi$ is a parameterisation of $D$ and $\alpha$ is a change of coordinates of $\mathbb{P}^2(\mathbb{C})$ then $\alpha \circ \psi$ is a parameterisation of $\alpha[D]$.*

**Proposition 15.24** *Let $\psi$ be a parameterisation of a curve $D$. There is a neighbourhood $U \subseteq \operatorname{dom}\psi$ of 0 and three analytic functions $\psi_w, \psi_x, \psi_y : U \to \mathbb{C}$ such that $\psi = (\psi_w : \psi_x : \psi_y)$ on $U$.*

***Proof*** By choosing one of the affine covers $\rho_0$, $\rho_1$ or $\rho_2$, we may assume that $\psi(0) \in \mathbb{A}^2$; apply Lemma 15.21. □

We call such a triple $\boldsymbol{\psi} = (\psi_w, \psi_x, \psi_y)$ an *analytic presentation* of $\psi$.

*Remark 15.25* If $\boldsymbol{\psi}$ and $\boldsymbol{\varphi}$ are two analytic presentations of $\psi$, then $\boldsymbol{\varphi} = h \cdot \boldsymbol{\psi}$ on some neighbourhood of 0, where $h$ is analytic and $h(0) \neq 0$: for example, if $\psi_w(0) \neq 0$, take $h = \varphi_w/\psi_w$. (Compare with Exercise 4.15.) «

### 15.2.1 $n$-Fold Parameterisations

**Definition 15.26** An *$n$-fold parameterisation* of $D$ is a parameterisation $\eta\colon U \to D \cap \mathbb{A}^2$ such that $\eta_x(z) = z^n$ for all $z \in U$.

We refer to $n$ as the *valency* of the parameterisation. The centre of an $n$-fold parameterisation must lie on the affine line $x = 0$.

*Example 15.27* A vertical parameterisation of $D$ is a 1-fold parameterisation of $D$ (provided it is defined at 0). «

*Example 15.28* (a) the map $z \mapsto (z^2, z)$ is a 2-fold parameterisation of the "sideways parabola" $y^2 = x$. (b) the map $z \mapsto (z^2, z^3)$ is a 2-fold parameterisation of the cuspidal cubic $y^2 = x^3$ (verify that it is injective). «

*Example 15.29* Let $D$ be the nodal cubic $y^2 = x^3 + x^2$. We write $y = \pm x\sqrt{x+1}$. There is an analytic choice of square root on a neighbourhood of 1 (see Remark 12.59), so we get two 1-fold parameterisations of $D$: the maps $z \mapsto (z, z\sqrt{z+1})$ and $z \mapsto (z, -z\sqrt{z+1})$. «

### 15.2.2 Fractional Parameterisations

**Definition 15.30** A *fractional parameterisation* of $D$ is a holomorphic injective function $\zeta$ from a punctured open neighbourhood $W^*$ of 0 in $\Sigma/n$ (for some $n \geqslant 1$) to $D^* \cap \mathbb{A}^2$, such that for all $q \in W^*$, the point $\zeta(q)$ lies on the line $x = \pi_{\Sigma/n}(q)$. We refer to $n$ as the valency of $\zeta$.

The map $\mathrm{rt}_n$ is a biholomorphism between $\Sigma/n$ and $\mathbb{C}^*$. It maps punctured neighbourhoods of 0 in $\Sigma/n$ to punctured neighbourhoods of 0 in $\mathbb{C}$, and vice-versa; and for all $q \in \Sigma/n$, $(\mathrm{rt}_n(q))^n = \pi_{\Sigma/n}(q)$. Thus:

**Proposition 15.31** *If $\eta$ is an $n$-fold parameterisation of $D$, then $\eta \circ \mathrm{rt}_n$ is a fractional parameterisation of $D$.*

A partial converse holds:

**Proposition 15.32** *If $\zeta$ is a fractional parameterisation of $D$, of valency $n$, then $\zeta \circ \mathrm{pwr}_n$ can be extended to a parameterisation of $D$.*

The *centre* of a fractional parameterisation $\zeta$ is defined to be the centre of the continuous extension of $\zeta \circ \mathrm{pwr}_n$. Continuity shows that the centre of $\zeta$ must lie on the projective line $x = 0w$.[1]

***Proof*** We need to show that there is some $q \in D \cap (x = 0w)$ such that extending $\eta = \zeta \circ \mathrm{pwr}_n$ by mapping 0 to $q$ makes a continuous function. The idea is the same as in the proof of Proposition 13.36. Let $q_1, \ldots, q_k$ list $D \cap (x = 0w)$. By Lemma 13.48 (and Example 13.42), if we choose pairwise disjoint neighbourhoods $V_j$ of $q_j$ in $D$, for small enough $\varepsilon > 0$, for all $|a| < \varepsilon$, $D \cap (x = aw) \subset \bigcup_j V_j$. Hence $\zeta[B^*_{\Sigma/n}(0, \varepsilon)] \subseteq \bigcup_j V_j$.

For all $\varepsilon > 0$, the punctured neighbourhood $B^*_{\Sigma/n}(0, \varepsilon)$ is connected: it is the image of the connected set $B^*(0, \sqrt[n]{\varepsilon})$ under the homeomorphism $\mathrm{pwr}_n$. Hence, for any choice of neighbourhoods $V_j$, there is some $j^* \leqslant k$ such that $\zeta[B^*_{\Sigma/n}, (0, \varepsilon)] \subseteq V_{j^*}$; and as in the proof of Proposition 13.36, we get the same $j^*$ no matter the choice of neighbourhoods $V_j$. Hence setting $\eta(0) = q_{j^*}$ works. □

Let $\zeta\colon W^* \to D$ be a fractional parameterisation of $D$ of valency $n$; let $\eta\colon U \to D$ be the continuous extension of $\zeta \circ \mathrm{pwr}_n$. Then for all $z \in U^*$, $\eta_x(z) = z^n$. If the centre of $\zeta$ is in $\mathbb{A}^2$, then $\eta$ is an $n$-fold parameterisation of $D$.

*Example 15.33* Let $D$ be the hyperbola $xy = w^2$. Then $z \mapsto (z, 1/z)$ (defined on $\mathbb{C}^*$) is a fractional parameterisation of $D$ of valency 1 (recall that $\Sigma/1 = \mathbb{C}^*$). However the continuous extension of $\zeta = \zeta \circ \mathrm{pwr}_1$ has a centre the vertical point at infinity, so is not a 1-fold parameterisation. «

**Proposition 15.34** *Let $W^* = B^*_\Sigma(0, r)$ for some $r > 0$. Suppose that $\theta\colon W^* \to D^* \cap \mathbb{A}^2$ is holomorphic, and that $\theta(q)$ lies on the line $x = \pi_\Sigma(q)$ for all $q \in W^*$. Then $\theta$ induces a fractional parameterisation of $D$.*

To prove Proposition 15.34, we will need the following lemma. Recall that $\mathrm{sh}_k$ is the $k$th iteration of the shift map on $\Sigma$.

**Lemma 15.35** *Let $W^* = B^*_\Sigma(0, r)$ for some $r > 0$. Suppose that $\theta_1$ and $\theta_2$ are holomorphic maps from $W^*$ to $D^* \cap \mathbb{A}^2$ satisfying $\theta_1(q), \theta_2(q) \in D \cap (x = \pi_\Sigma(q))$ for all $q \in W^*$. Let $q_0 \in W^*$ and $k \in \mathbb{Z}$, and suppose that $\theta_2(q_0) = \theta_1(\mathrm{sh}_k(q_0))$. Then $\theta_2 = \theta_1 \circ \mathrm{sh}_k$.*

***Proof*** Let $\varphi$ be a chart for $\Sigma$ with $q_0 \in \mathrm{dom}\,\varphi$; we assume $\mathrm{dom}\,\varphi \subset W^*$. Both $\theta_2 \circ \varphi^{-1}$ and $\theta_1 \circ \mathrm{sh}_k \circ \varphi^{-1}$ are vertical parameterisations of $D$ defined on the range of $\varphi$, and they agree on $\pi_\Sigma(q_0)$; so they agree on a neighbourhood of $\pi_\Sigma(q_0)$

[1] As above, we use $x = a$ to denote the vertical affine line and $x = aw$ to denote its projective closure. This notation for $a = 0$ is ambiguous, so we sometimes write $x = 0w$ to emphasise that we mean the projective line, including the vertical point at infinity.

(Proposition 13.15). So $\theta_2$ and $\theta_1 \circ \mathrm{sh}_k$ agree on a neighbourhood of $q_0$. Since they are both holomorphic, and $W^*$ is connected (see Proposition 15.36 shortly), they agree on all of $W^*$ (Proposition 12.9). □

***Proof of Proposition 15.34*** Fix any $a \in \pi_\Sigma[W^*] = B^*(0, r)$. Since $D \cap (x = a)$ is finite, there are distinct $q_1, q_2 \in W^*$ such that $\pi_\Sigma(q_1) = \pi_\Sigma(q_2) = a$ and $\theta(q_1) = \theta(q_2)$. Let $n \geqslant 1$ such that $q_2 = \mathrm{sh}_n(q_1)$. Apply Lemma 15.35 with $\theta_2 = \theta_1 = \theta$ and $q_1$ to conclude that $\theta = \theta \circ \mathrm{sh}_n$.

We take $n$ minimal such that $\theta = \theta \circ \mathrm{sh}_n$. By Lemma 12.60 (and Exercise 12.64), $\theta$ induces a holomorphic map $\zeta$ on $V^* = B^*_{\Sigma/n}(0, r)$. The map $\zeta$ maps each $q \in V^*$ to a point on the line $x = \pi_{\Sigma/n}(q)$. It remains to check that $\zeta$ is injective. Suppose that $\zeta(q_2) = \zeta(q_1)$ for some points $q_1, q_2 \in V^*$. Then $\pi_{\Sigma/n}(q_1) = \pi_{\Sigma/n}(q_2)$, and so there is some $k$ such that $q_2 = \mathrm{sh}_k(q_1)$. Considering the pull-back to $\Sigma$, by Lemma 15.35, $\theta = \theta \circ \mathrm{sh}_k$; by minimality of $n$, we must have $n$ dividing $k$, so $q_2 = q_1$. □

### 15.2.3 Existence of Parameterisations

Our next step is to show the existence of parameterisations. We will use analytic continuation, so we will need:

**Proposition 15.36** *For all $r > 0$, $B^*_\Sigma(0, r)$ is simply connected.*

***Proof*** $B^*_\Sigma(0, r)$ is the image of the open half-plane $\{a + ib \in \mathbb{C} : a < \ln r\}$ under the homeomorphism $z \mapsto (e^z, \Im z)$ (Proposition 12.54), and the half-open plane is simply connected (it is convex; see Example 9.15). □

Recall the notion of a ramification point of a curve (Definition 13.28). By Proposition 13.29, for some $r > 0$, $B^*(0, r)$ contains no ramification points of $D$.

**Proposition 15.37** *Suppose that $B^*(0, r)$ contains no ramification points of $D$. Then for every $a \in B^*(0, r)$, for every $p \in D \cap (x = a)$, there is a fractional parameterisation $\zeta : B^*_{\Sigma/n}(0, r) \to D$ with $p \in \mathrm{range}\, \zeta$.*

***Proof*** Let $W^* = B^*_\Sigma(0, r)$. Fix $a_0 \in B^*(0, r)$ and $p_0 \in D \cap (x = a_0)$ (again, note that this is the affine line $x = a_0$, i.e., $p_0 \in \mathbb{A}^2$). By Proposition 15.34, it suffices to show that there is a holomorphic map $\theta : W^* \to D^* \cap \mathbb{A}^2$ with $p_0 \in \mathrm{range}\, \theta$ and $\theta(q) \in D \cap (x = \pi_\Sigma(q))$ for all $q \in W^*$.

We start with some chart $\varphi$ for $\Sigma$ whose range contains $a_0$; we suppose that $\mathrm{dom}\, \varphi \subseteq W^*$. Since the range of $\varphi$ avoids the ramification points of $D$, by shrinking $\mathrm{dom}\, \varphi$ if necessary, let $g : \mathrm{range}\, \varphi \to D^*$ be a vertical parameterisation of $D$ with $g(a_0) = p_0$ (Proposition 13.13).

Let $\gamma$ be a path in $W^*$, starting at some point in $\operatorname{dom}\varphi$. Then $\pi_\Sigma \circ \gamma$ is a path in $B^*(0,r)$, and so avoids all ramification points of $D$. It starts in $\operatorname{range}\varphi = \operatorname{dom} g$, and so by Proposition 13.34, there is an analytic continuation of $g$ along $\pi_\Sigma \circ \gamma$; this is also an analytic continuation of $g \circ \pi_\Sigma$ along $\gamma$. By Proposition 15.36, we can apply the Monodromy Theorem (with $X = W^*$, $Y = D^* \cap \mathbb{A}^2$, $U = \operatorname{dom}\varphi$ and $f = g \circ \pi_\Sigma$), and we obtain a holomorphic map $\theta\colon W^* \to D^* \cap \mathbb{A}^2$ extending $g \circ \pi_\Sigma$. In particular, $p_0 \in \operatorname{range}\theta$.

We need to show that $\theta(q)$ lies on the line $x = \pi_\Sigma(q)$ for all $q \in W^*$. Let $Q = \{q \in W^* : \theta(q) \text{ lies on the line } x = \pi_\Sigma(q)\}$. Then $\operatorname{dom}\varphi \subseteq Q$, as $g(a)$ lies on $x = a$ for all $a \in \operatorname{range}\varphi$. Further, $q \in Q$ if and only if $(P \circ \theta)(q) = \pi_\Sigma(q)$, where $P(x,y) = x$ is the projection from $\mathbb{A}^2$ onto the first coordinate. The restriction of $P$ to $D^* \cap \mathbb{A}^2$ is holomorphic (Example 13.24). Also, $\pi_\Sigma$ is holomorphic (Exercise 12.53). Hence $Q$ is the set of points on which two holomorphic functions agree; since it contains a nonempty open set, and $W^*$ is connected, we must have $Q = W^*$ (Proposition 12.9). □

**Corollary 15.38** *Every point $q \in D \cap (x = 0) \cap \mathbb{A}^2$ is the centre of a fractional parameterisation of $D$.*

***Proof*** Let $q_1, \ldots, q_k$ be the points on $D \cap (x = 0w)$ (this includes the vertical point at infinity, if it is on $D$). Choose pairwise disjoint neighbourhoods $V_j$ of $q_j$ in $D$. Let $m_j = i_{q_j}(D, x = 0w)$. By Proposition 13.46, For small enough $s$, for all $a \in B^*(0,s)$, for all $j$, $(x = aw) \cap V_j$ contains $m_j$ points (multiplicities counted); in particular, this intersection is nonempty. Let $j$ such that $q_j \in \mathbb{A}^2$; we assume that $V_j \subset B(0,s) \times \mathbb{C}$. Choose any $p \in V_j$ other than $q_j$. By Proposition 15.37, let $\zeta$ be a fractional parameterisation with $p \in \operatorname{range}\zeta$. By connectedness, $\zeta[B^*_{\Sigma/n}(0,s)] \subseteq V_j$. By continuity, the centre of $\zeta$ must be $q_j$.

For an alternative proof see Exercise 15.112. □

**Corollary 15.39** *Every point $p \in D$ is the centre of some parameterisation of $D$.*

***Proof*** By Proposition 15.23, we may change coordinates so that $p \in \mathbb{A}^2$ and lies on $x = 0$; then apply Corollary 15.38. □

## 15.3 Branches and Places

**Definition 15.40** Let $\psi_1$ and $\psi_2$ be parameterisations of curves in $\mathbb{P}^2$.

(a) We say that $\psi_1$ and $\psi_2$ are *essentially equal* if $\psi_2 = \psi_1$ on a neighbourhood of 0.
(b) We say that $\psi_1$ and $\psi_2$ are *equivalent* if there is an injective analytic function $h$ mapping 0 to 0 such that $\psi_2 = \psi_1 \circ h$ on a neighbourhood of 0.

The analytic inverse function theorem (Theorem 12.34) implies that equivalence of parameterisations is indeed an equivalence relation. Two essentially equal parameterisations are equivalent. Two equivalent parameterisations have the same centre.

**Definition 15.41**

(a) A *germ* of parameterisations is an equivalence class of parameterisations under essential equality.
(b) A *place* is an equivalence class of parameterisations under parameterisation equivalence.

The *centre* of a place is the centre of the parameterisations of the place.

These notion are invariant under changes of coordinates:

**Proposition 15.42** *Let $\psi_1$ and $\psi_2$ be parameterisations, and let $\alpha$ be a change of coordinates of $\mathbb{P}^2$. Then:*

*(a) $\psi_1$ and $\psi_2$ are essentially equal if and only if $\alpha \circ \psi_1$ and $\alpha \circ \psi_2$ are essentially equal; and*
*(b) $\psi_1$ and $\psi_2$ are equivalent if and only if $\alpha \circ \psi_1$ and $\alpha \circ \psi_2$ are equivalent.*

Note that these notions apply to parameterisations of any curve, not a fixed curve $D$. However, any place determines an irreducible curve.

**Proposition 15.43** *If $\psi_1$ and $\psi_2$ are equivalent parameterisations, then there is a unique irreducible curve $D$ such that both $\psi_1$ and $\psi_2$ are essentially equal to parameterisations of $D$.*

***Proof*** For uniqueness, suppose that $\psi_1$ and $\psi_2$ are essentially equal to parameterisations of irreducible curves $D_1$ and $D_2$, and are equivalent, witnessed by injective $h\colon U_2 \to U_1$. Let $B = \psi_1[U_1] = \psi_2[U_2]$. By shrinking $U_1$ and $U_2$, we get $B \subseteq D_1 \cap D_2$. Since the $\psi_i$ are injective, $B$ is infinite; so $D_1 \cap D_2$ is infinite, whence $D_1 = D_2$ (Proposition 6.6).

For existence, let $\psi$ be a parameterisation of a curve $D$. Let $f$ be a polynomial defining $D$, and let $\boldsymbol{\psi}$ be an analytic presentation of $\psi$ (Proposition 15.24). Write $f = f_1 \cdots f_m$, where $f_i$ are irreducible. The product of the analytic functions $f_i \circ \boldsymbol{\psi}$ is the analytic function $f \circ \boldsymbol{\psi}$, which is constant zero; hence one of the functions $f_i \circ \boldsymbol{\psi}$ is constant 0. Then $\psi$ is a parameterisation of the irreducible component $f_i = 0$ of $D$. □

We say that a place $P$ is a place *of* a curve $D$ if every parameterisation $\psi$ of $P$ is essentially equal to a parameterisation of $D$. Note that if $P$ is a place of $D$ and $D$ is a component of $E$ then $P$ is also a place of $E$. If $P$ is a place, and $\alpha$ is a change of coordinates, we let $\alpha[P]$ denote the place with parameterisations $\alpha \circ \psi$, where $\psi$ are parameterisations of $P$. If $P$ is a place of $D$ then $\alpha[P]$ is a place of $\alpha[D]$.

### 15.3.1 Central Places

**Definition 15.44** A *central place* is a place whose centre lies on the *affine* line $x = 0$, but is not a place of the line $x = 0$.

**Proposition 15.45** *Every central place is parameterised by some n-fold parameterisation.*

***Proof*** This is essentially Exercise 12.107. Let $\psi$ be a parameterisation of a central place. By Lemma 15.21, $\psi_x$ is analytic; let $n = \operatorname{ord}_0(\psi_x)$ be the order of $\psi_x$ at 0 (Definition 12.13). Since $P$ is not a place of $x = 0$, $\psi_x$ is nonconstant; hence $n \neq \infty$ (where $\infty$ is the order of the constant 0 function). Also $\psi_x(0) = 0$ so $n > 0$.

Write $\psi_x = z^n g$ where $g$ is analytic and $g(0) \neq 0$. Choose an analytic $n$th root on a neighbourhood of $g(0)$; composing with $g$, we get an analytic $\bar{g}$ such that $g = (\bar{g})^n$ on a neighbourhood of 0. Since $\bar{g}(0) \neq 0$, the map $z \mapsto z\bar{g}$ has an analytic inverse $h$. Then $\psi \circ h$ is an $n$-fold parameterisation, equivalent to $\psi$. □

**Lemma 15.46** *Let $n \geqslant 1$. Two n-fold parameterisations $\eta$ and $\widetilde{\eta}$ are equivalent if and only if on some neighbourhood of 0, $\widetilde{\eta}(z) = \eta(\omega z)$ for some nth root of unity $\omega \in \mathbb{C}$.*

Note that since the $x$-coordinate map is $z \mapsto z^n$ for both $\eta$ and $\widetilde{\eta}$, and $(\omega z)^n = z^n$, we have $\widetilde{\eta}(z) = \eta(\omega z)$ if and only if $\widetilde{\eta}_y(z) = \eta_y(\omega z)$.

***Proof*** The map $z \mapsto \omega z$ is analytic and 1-1 (and maps 0 to 0), so one direction is immediate. For the other direction, suppose that $\widetilde{\eta}$ and $\eta$ are equivalent $n$-fold parameterisations, say $\widetilde{\eta} = \eta \circ h$ on a neighbourhood of 0, with $h$ analytic and injective. Since $h(0) = 0$, write $h(z) = z \cdot \omega(z)$ with $\omega$ analytic. Looking at the $x$-coordinate, we get $z^n = (h(z))^n$, so $(\omega(z))^n = 1$ on a punctured neighbourhood of 0. There are only $n$-many $n$th roots of unity, and $\omega(z)$ is continuous, so it is constant on a neighbourhood of 0. □

*Remark 15.47* If $n \neq m$ then an $n$-fold parameterisation and an $m$-fold parameterisation cannot be equivalent. Say $\eta$ and $\widetilde{\eta}$ are $n$- and $m$-fold parameterisations; suppose that $\widetilde{\eta} = \eta \circ h$ with $h$ analytic and injective. Then on a punctured neighbourhood of 0 we have $h(z)^n = z^m$; since $\operatorname{ord}_0(h(z))^n = n$ and $\operatorname{ord}_0 z^m = m$ we get $n = m$. «

Let $P$ be a central place of a curve $D$. By Proposition 15.45, $P$ has an $n$-fold parameterisation for some $n$; by Remark 15.47, this $n$ is determined by the place, so we can refer to $n$ as the valency of the place.

Each place consists of infinitely many germs of parameterisations. But a central place has only finitely many germs of $n$-fold parameterisations:

**Proposition 15.48** *A central place of valency $n$ has $n$-many germs of $n$-fold parameterisations.*

***Proof*** Let $P$ be a central place of valency $n$; let $\eta$ be an $n$-fold parameterisation of $P$. There are $n$ many $n$th roots of unity in $\mathbb{C}$, namely $\omega_n^k$ for $k = 0, \ldots, n-1$. Let $\eta_k(z) = \eta(\omega_n^k z)$. For distinct $k, k' \in \{0, \ldots, n-1\}$, the maps $z \mapsto \omega_n^k z$ and $z \mapsto \omega_n^{k'} z$ disagree on every nonzero $z$; since $\eta$ is injective, it follows that the $\eta_k$ are pairwise not essentially equal. By Lemma 15.46, these give all of the germs of $n$-fold parameterisations of $P$. □

*Example 15.49* (a) The 2-fold parameterisation $z \mapsto (z^2, z)$ of $y^2 = x$ (Example 15.28) is equivalent to $z \mapsto (z^2, -z)$. Similarly, the 2-fold parameterisation $z \mapsto (z^2, z^3)$ of $y^2 = x^3$ is equivalent to $z \mapsto (z^2, -z^3)$. (b) Two 1-fold parameterisations are equivalent if and only if they are essentially equal. Thus, the two parameterisations of $y^2 = x^3 + x^2$ given in Example 15.29 are not equivalent. «

We say that two fractional parameterisations $\zeta$ and $\widetilde{\zeta}$ (of valencies $n$ and $\widetilde{n}$) are equivalent if the continuous extensions of $\zeta \circ \mathrm{pwr}_n$ and $\widetilde{\zeta} \circ \mathrm{pwr}_{\widetilde{n}}$ are equivalent.

**Lemma 15.50** *Two fractional parameterisations $\zeta$ and $\widetilde{\zeta}$ are equivalent if and only if they have the same valency, and $\widetilde{\zeta} = \zeta \circ \mathrm{sh}_k$ for some $k$.*

***Proof*** If $\zeta$ and $\widetilde{\zeta}$ have centres in $\mathbb{A}^2$ then this is implied by Lemma 15.46, Remark 15.47, and Exercise 12.65; but the argument holds even if the centre of $\eta = \zeta \circ \mathrm{pwr}_n$ and $\widetilde{\eta} = \widetilde{\zeta} \circ \mathrm{pwr}_{\widetilde{n}}$ is the vertical point at infinity. □

*Exercise 15.51* Give a direct proof of Lemma 15.50. «

### 15.3.2 Branches of a Curve

In the following definition, we consider the range of a parameterisation as a topological subspace of $\mathbb{P}^2(\mathbb{C})$.

**Definition 15.52**

(a) A parameterisation is *tidy* if it is a homeomorphism between its domain and its range.
(b) A *branch* is the range of a tidy parameterisation.

The *centre* of the branch $B = \operatorname{range}\psi$ is the centre of $\psi$. The definition thus is not quite precise, because by shifting, we could have two (tidy) parameterisations with different centres but the same range; so a formal definition would be that a branch is the pair $(\operatorname{range}\psi, \psi(0))$. We will continue to be imprecise with our notation.

The restriction to tidy parameterisations is immaterial:

**Lemma 15.53** *Every parameterisation is essentially equal to a tidy one.*

***Proof*** Let $\psi : U \to D$ be a parameterisation of $D$. There is some open disc $V$ such that $0 \in V$ and $\overline{V} \subset U$. Since $\overline{V}$ is compact and $\psi$ is injective, $\psi\restriction_{\overline{V}}$ is a homeomorphism from $\overline{V}$ to $\psi[\overline{V}]$; so $\psi\restriction_V$ is a homeomorphism from $V$ to $\psi[V]$. □

Like parameterisations, we are only interested in what happens close to the centre, so we define:

**Definition 15.54** Two branches $B_1$ and $B_2$ with the same centre $c$ are *essentially equal* if there is a neighbourhood $W$ of $c$ in $\mathbb{P}^2$ such that $B_1 \cap W = B_2 \cap W$. An equivalence class of branches under essential equality is called a *germ of branches*.

**Proposition 15.55** *If $\psi_1$ and $\psi_2$ are equivalent tidy parameterisations then* $\operatorname{range}\psi_1$ *and* $\operatorname{range}\psi_2$ *are essentially equal.*

Thus, every place determines a germ of branches.

***Proof*** First, we observe that if $\psi : U \to D$ is a tidy parameterisation and $V \subseteq U$ is an open neighbourhood of 0, then $\psi[V]$ is essentially equal to $\psi[U]$: as $\psi$ is a homeomorphism from $U$ to $\psi[U]$, $\psi[V]$ is an open subset of $\psi[U]$ (and it contains the centre $\psi(0)$); by definition of the subspace topology, there is an open neighbourhood $W$ of the centre in $\mathbb{P}^2$ such that $\psi[V] = W \cap \psi[U]$, so $\psi[V] \cap W = \psi[U] \cap W$. Thus, two essentially equal parameterisations parameterise essentially equal branches.

To extend this to equivalent parameterisations, let $U_1$ and $U_2$ be open neighbourhoods of 0 and let $h : U_2 \to U_1$ be an analytic bijection such that $\psi_2 = \psi_1 \circ h$ on $U_2$. Then $\psi_2[U_2] = \psi_1[U_1]$. □

*Example 15.56* Really, untidy parameterisations are an anomaly; they are parameterisations which are defined on domains which are too large, so they allow another part of the curve to "curl" back and approach the centre. For example, let $D$ be the nodal cubic $y^2 = x^3 + x^2$; let $\psi(z) = (z^2 + 2z, z^3 + 3z^2 + 2z)$ be a shift of the rational parameterisation of $D$ discussed in Exercise 4.61 (shifted so that the centre is at 0). Then $\psi(0) = \psi(-2) = o$. But $\psi\restriction_{B(0,2)}$ is injective, so strictly speaking, it is a parameterisation of $D$; but it is not tidy, since using it we can approach the

origin in two ways, by getting close to 0 or to $-2$. And indeed $\psi[B(0,2)]$ is not essentially equal to say $\psi[B(0,1)]$. «

We say that two branches $A$ and $B$ are *essentially disjoint* if they have different centres; or if they have the same centre $c$, but there is an open neighbourhood $W$ of $c$ in $\mathbb{P}^2$ such that $A \cap B \cap W = \{c\}$. This notion is well-defined for germs of branches: if $A_1$ and $A_2$ are essentially equal, and so are $B_1$ and $B_2$, then $A_1$ and $B_1$ are essentially disjoint if and only if $A_2$ and $B_2$ are essentially disjoint. Two branches which are essentially equal are not essentially disjoint.

**Theorem 15.57**

*(a) Distinct germs of branches are essentially disjoint.*
*(b) Two distinct places determine distinct germs of branches.*

We thus often identify a place and the germ of branches it determines, and refer to the germ as a place as well.

***Proof*** Let $\psi_1$ and $\psi_2$ be tidy parameterisations. Suppose that $B_1 = \operatorname{range}\psi_1$ and $B_2 = \operatorname{range}\psi_2$ are not essentially disjoint; we need to show that $\psi_1$ and $\psi_2$ are equivalent.

Since $B_1$ and $B_2$ are not essentially disjoint, they have the same centre $c$. By Proposition 15.43, let $D_i$ be the unique irreducible curve parameterised by $\psi_i$. If $D_1 \neq D_2$ then there is an open neighbourhood $W$ of $c$ in $\mathbb{P}^2$ such that $D_1 \cap D_2 \cap W = \{c\}$; this would make $B_1$ and $B_2$ essentially disjoint. Hence $D_1 = D_2$, call it $D$.

Change coordinates so that the centre $c$ lies on the affine line $x = 0$ (say the origin), and $D$ is not the $y$-axis $x = 0$; so the places of $\psi_1$ and $\psi_2$ are central. By Proposition 15.45, let $\eta_1$ be an $n_1$-fold parameterisation equivalent to $\psi_1$, and $\eta_2$ be an $n_2$-fold parameterisation equivalent to $\psi_2$. Considering $\mathrm{rt}_{n_1}$ and $\mathrm{rt}_{n_2}$ as maps on $\Sigma$ (rather than $\Sigma/n_i$), let $\theta_1 = \eta_1 \circ \mathrm{rt}_{n_1}$ and $\theta_2 = \eta_2 \circ \mathrm{rt}_{n_2}$. Fix some sufficiently small $r > 0$ such that $B^*_\Sigma(0,r) \subseteq \operatorname{dom}\theta_1, \operatorname{dom}\theta_2$.

Since $B_1$ and $B_2$ are not essentially disjoint, we can find two points $q_1, q_2 \in B^*_\Sigma(0,r)$ such that $\theta_1(q_1) = \theta_2(q_2)$. Since $\theta_i(q_i)$ lies on the line $x = \pi_\Sigma(q_i)$, we must have $\pi_\Sigma(q_1) = \pi_\Sigma(q_2)$. So there is some $k \in \mathbb{Z}$ such that $q_2 = \mathrm{sh}_k(q_1)$. By Lemma 15.35, $\theta_2 = \theta_1 \circ \mathrm{sh}_k$ on $B^*_\Sigma(0,r)$. Since $\theta_1 = \theta_1 \circ \mathrm{sh}_{n_1}$, this shows that $\theta_2 = \theta_2 \circ \mathrm{sh}_{n_1}$. Since $\eta_2$ is injective, $n_2$ is the least $n$ such that $\theta_2 = \theta_2 \circ \mathrm{sh}_n$; hence $n_2$ divides $n_1$. By symmetry, $n_2 = n_1$. By Lemma 15.50, $\eta_1$ and $\eta_2$ are equivalent. □

## 15.4 Puiseux Expansions and Factorisation into Places

We show that places of a curve have implicit definitions; a curve can be presented as the sum of some of its places. We start by counting germs of $n$-fold parameterisations of a curve.

**Proposition 15.58** *Suppose that the vertical point at infinity* $(0:0:1)$ *does not lie on* $D$*. Then there are precisely* $\deg D$*-many germs of* $n$*-fold parameterisations of* $D$ *(for all* $n$*).*

In particular, there are only finitely many central places of $D$, indeed at most $\deg D$-many.

***Proof*** Let $r > 0$ be sufficiently small so that $B^*(0,r)$ contains no ramification points of $D$. Further, by Theorem 15.57, by shrinking $r$, ensure that if $A$ and $B$ are central branches of $D$ (branches of central places) which are not essentially equal, then $A \cap B \cap (B^*(0,r) \times \mathbb{C}) = \emptyset$. (In fact, the proof of Theorem 15.57 shows that we do not need to shrink $r$ to achieve that.)

Choose any $a \in B^*(0,r)$. Since $a$ is not a ramification point of $D$, and the vertical point at infinity does not lie on $D$, the affine line $x = a$ intersects $D$ at $\deg D$-many distinct points. By Proposition 15.37, each one of these points lies on a branch of $D$ with centre on $x = 0$; by assumption on $r$, this branch is unique up to essential equality. If the branch has valency $n$, then it contains $n$ of the points on $x = a$; and by Proposition 15.48, is parameterised by precisely $n$-many germs of $n$-fold parameterisations. Adding up, we see that the number of germs of $n$-fold parameterisations (for all $n$) is the same as the number of points on $D \cap (x = a)$, i.e., $\deg D$. □

**Corollary 15.59** *For any curve* $D$ *and any* $p \in D$*, there are finitely many places of* $D$ *with centre* $p$*.*

***Proof*** Change coordinates so that $(0:0:1) \notin D$ and $p$ lies on $x = 0$. After the change, since $x = 0$ is not a component of $D$, every place of $D$ with centre on $x = 0$ is central. Apply Proposition 15.58 and Proposition 15.45. □

### 15.4.1 Puiseux Expansions

In this section, again fix a curve $D$ with no repeated components.

Let $\zeta$ be a fractional parameterisation of $D$, whose centre is in $\mathbb{A}^2$. Write $\zeta = (\pi_\Sigma, \zeta_y)$. By Exercise 13.26, $\zeta_y$ is holomorphic. By assumption, $\zeta_y$ is bounded on a punctured neighbourhood of 0 in $\Sigma/n$. Hence, by Proposition 15.13, $\zeta_y = \mathfrak{g}_\zeta$ (on a punctured neighbourhood of 0 in $\Sigma/n$) for some $g = g_\zeta \in \mathbb{C}[\![x^{1/n}]\!]$.

**Definition 15.60** A *Puiseux expansion* of $D$ is a fractional power series $g = g_\zeta \in \mathbb{C}[\![x]\!]$ for some fractional parameterisation $\zeta$ of $D$ (with centre in $\mathbb{A}^2$).

Thus, if $h \in \mathbb{C}[\![x]\!]$ defines the second coordinate of an $n$-fold parameterisation $z \mapsto (z^n, \mathfrak{h}(z))$ of $D$, then $g = h(x^{1/n})$ is a Puiseux expansion of $D$. Now Proposition 15.11 (and Example 15.12) imply:

**Proposition 15.61** *If $g$ is a Puiseux expansion of $D$ and $f$ is a polynomial in $\mathbb{C}[x, y]$ defining $D \cap \mathbb{A}^2$, then $f(x, g) = 0$ (in the ring $\mathbb{C}[\![x]\!]$).*

*Example 15.62* (a) $g = x^{1/2}$ is a Puiseux expansion of $y^2 = x$. (b) $g = x^{3/2}$ is a Puiseux expansion of $y^2 = x^3$ (Example 15.28). (c) One of the analytic square roots around 1 is given by $\sqrt{z+1} = 1 + z/2 - z^2/8 + z^3/16 - \cdots$ (see Exercise 11.88). Hence a Puiseux expansion of $y^2 = x^3 + x^2$ is $x + x^2/2 - x^3/8 + x^4/16 - \cdots$ (see Example 15.29). «

By Proposition 15.11, two fractional parameterisations are defined by the same Puiseux expansion if and only if they are essentially equal. With Proposition 15.58, this gives:

**Proposition 15.63** *If $(0:0:1) \notin D$, then $D$ has precisely* $\deg D$*-many Puiseux expansions.*

### 15.4.2 The Implicit Definition of a Place

Let $P$ be a central place of $D$, of valency $n$. A *Puiseux expansion of $P$* is a Puiseux expansion which corresponds to an $n$-fold parameterisation of $P$. By Proposition 15.48, let $g_1, g_2, \ldots, g_n$ be the (distinct) Puiseux expansions of $P$. We let

$$f_P = (y - g_1)(y - g_2) \cdots (y - g_n);$$

it is an element of the polynomial ring $\mathbb{C}[\![x]\!][y]$.

**Proposition 15.64** $f_P \in \mathbb{C}[x][y]$.

That is, the coefficients of $f_P$ (thought of as a polynomial in $y$) are formal power series with integer exponents; they have no proper fraction exponents, even though the $g_i$'s may have proper fraction exponents.

***Proof*** The map $\mathrm{sh}\colon \mathbb{C}[\![x]\!] \to \mathbb{C}[\![x]\!]$ is a ring homomorphism (Exercise 15.17); we extend it to a ring homomorphism $\mathrm{sh}\colon \mathbb{C}[\![x]\!][y] \to \mathbb{C}[\![x]\!][y]$, by mapping $\sum h_j y^j \to \sum \mathrm{sh}(h_j) y^j$ (check that this indeed respects sums and products of polynomials). Then

$$\mathrm{sh}(f_P) = (y - \mathrm{sh}(g_1))(y - \mathrm{sh}(g_2)) \cdots (y - \mathrm{sh}(g_n)).$$

But $\{\mathrm{sh}(g_1), \mathrm{sh}(g_2), \ldots, \mathrm{sh}(g_n)\} = \{g_1, g_2, \ldots, g_n\}$, as by Proposition 15.19 and Lemma 15.50, they both equal $\{\mathrm{sh}_1(g_i), \mathrm{sh}_2(g_i), \ldots, \mathrm{sh}_n(g_i)\}$ (for any $g_i$). Here of course $\mathrm{sh}_k$ denotes the $k$th iteration of the map $g \mapsto \mathrm{sh}(g)$ defined on $\mathbb{C}[\![x]\!]$; note

that $\mathrm{sh}_n(g_i) = g_i$ for all $i$. Therefore, $f_P = \mathrm{sh}(f_P)$. Write $f_P = \sum_{j \leqslant n} f_j y^j$; so by our definition of the extension of sh to a map on polynomials, we have $\mathrm{sh}(f_j) = f_j$ for each $j$. The result then follows from Proposition 15.18. □

Each Puiseux expansion $g_i$ of $P$ is convergent: $R(g_i) > 0$ (in fact, the root test shows that all $g_i$ have the same radius of convergence). Fix some $R_P > 0$ and fractional parameterisations $\zeta_i : B^*_{\Sigma/n}(0, R_P) \to D$ with $g_i = g_{\zeta_i}$. Let $\eta_i$ be the corresponding $n$-fold parameterisations (the continuous extensions of $\zeta_i \circ \mathrm{pwr}_n$); let $B = \operatorname{range} \eta_i$ (which does not depend on $i$)

Each coefficient $f_j$ of $f_P$ is obtained from the $g_i$'s by addition and multiplication, and so by Proposition 15.11, $R(f_j) \geqslant R_P$ for each $j$. The functions $\mathfrak{f}_j$ are defined on $B^*_\Sigma(0, R_P)$ and induces functions on $B^*(0, R_P)$, also denoted $\mathfrak{f}_j$; we extend these continuously at 0.

For $a \in B(0, R_P)$ and $b \in \mathbb{C}$, let

$$\mathfrak{f}_P(a, b) = \sum_{j \leqslant n} \mathfrak{f}_j(a) b^j;$$

this is the function on $B(0, R_P) \times \mathbb{C}$ defined by $f_P$. Its zero set is $B$:

**Lemma 15.65** $B = \{(a, b) \in B(0, R_P) \times \mathbb{C} : \mathfrak{f}_P(a, b) = 0\}$.

***Proof*** Let $a \in B^*(0, R_P)$. Let $q$ be any point in $\Sigma/n$ above $a$ ($\pi_{\Sigma/n}(q) = a$); Let $b \in \mathbb{C}$. By definition, $\mathfrak{f}_P(a, b) = \prod_{i \leqslant n}(b - \zeta_{i,y}(q))$, and so $\mathfrak{f}_P(a, b) = 0$ if and only if $(a, b) \in \operatorname{range} \zeta_i$ for some $i \leqslant n$. For $a = 0$ use continuity, as both $\mathfrak{f}_P$ and $\eta_i$ are continuous. □

Combining with a change of coordinates we conclude that for any place $P$ of $D$ (whether central or not), there is a complex continuously differentiable $f$, defined on a neighbourhood of the centre of $P$ in $\mathbb{P}^2$, such that $f = 0$ defines a branch whose place is $P$. Note though that $f$ depends on the choice of change of coordinates which moves $P$ to a central place.

### A Factorisation of the Defining Polynomial

Every polynomial in $\mathbb{C}[x, y]$ can be considered also as an element of $\mathbb{C}[\![x]\!][y]$. Suppose that $(0:0:1) \notin D$; let $d = \deg D$. Then the monomial $y^d$ appears in the polynomials defining $D$ (Remark 6.1), as well as their dehomogenisations (which define $D \cap \mathbb{A}^2$). Fix $f_D$ defining $D \cap \mathbb{A}^2$ for which the coefficient of $y^d$ is 1.

**Theorem 15.66** *Suppose that* $(0:0:1) \notin D$. *If* $P_1, P_2, \ldots, P_m$ *are the central places of* $D$, *then* $f_D = f_{P_1} f_{P_2} \cdots f_{P_m}$.

So even if $D$ is irreducible, we can present it, in some sense, as the sum of its central places.

***Proof*** By Proposition 15.63, let $g_1, g_2, \ldots, g_d$ be the Puiseux expansions of $D$. The main point is that $\mathbb{C}[\![x]\!]$ is an integral domain, and by Proposition 15.61, each $g_i$ is a root of the polynomial $f_D$, viewed as a polynomial in $y$ with coefficients from $\mathbb{C}[\![x]\!]$: $f_D(x, g_i) = 0$. So by Theorem 2.16, each $y - g_i$ divides $f_D$ in the polynomial ring $\mathbb{C}[\![x]\!][y]$; repeatedly dividing, since $f_D$ has $d = \deg_y f_D$ many distinct roots, we get $f \sim \prod_{i \leqslant d}(y - g_i)$. Since we chose $f_D$ to be monic (viewed as a polynomial in $y$), comparing coefficients we get

$$f_D = (y - g_1)(y - g_2) \cdots (y - g_d);$$

now group the $g_i$'s into their places to get the polynomials $f_{P_i}$. □

*Example 15.67* Theorem 15.66 holds for some curves which contain the vertical point at infinity (but we may get fewer than deg $D$-many germs of $n$-fold parameterisations); see Exercise 15.110.

For example, the cuspidal cubic $y^2 = x^3$ has a single place at the origin, of valency 2; the two Puiseux expansions are $x^{3/2}$ and $-x^{3/2}$, and $f_P = f_D = (y - x^{3/2})(y + x^{3/2})$. The nodal cubic $y^2 = x^3 + x^2$ has two places at the origin, of valency 1; the Puiseux expansions are the formal power series $g_1$ and $g_2$ defining the analytic functions $z\sqrt{z+1}$ and $-z\sqrt{z+1}$; and $f_D = f_{P_1} f_{P_2} = (y - g_1)(y - g_2)$.«

**Remark on Newton Polygons**

We presented the theory developed by Victor Puiseux in the mid-nineteenth century. However, a lot earlier, Newton gave a purely algebraic treatment. Seen from a modern point of view, Newton showed that if $\mathbb{K}$ is an algebraically closed field of characteristic 0, then the fraction field of $\mathbb{K}[\![x]\!]$ (the field of fractional Laurent series with coefficients from $\mathbb{K}$) is algebraically closed. This implies that not only curves defined by polynomials in $\mathbb{K}[x, y]$ have Puiseux expansions, but in fact curves in $\mathbb{K}[\![x]\!][y]$ have such expansions as well (but they may need negative fractional exponents).

Newton's method does not show that the resulting Puiseux expansions are convergent ($R(g) > 0$), so does not show how to define branches of curves and does not provide for their implicit definitions. However, his method of "Newton polygons" gives us an algorithm for calculating the coefficients of Puiseux expansions (compare with Exercise 13.14). For details, see, for example, [Wal50, BK86, Bix06, Kir92].

## 15.5 Intersection Multiplicities Using Places

Recall that the general idea is to use parameterisations to define intersection multiplicity, generalising Definition 5.25 (and Exercise 6.47): we substitute a

parameterisation of a curve into the polynomial defining another curve, and define the order of zero to be the intersection number. We can do this for places of curves.

Let $P$ be a central place; let $\psi$ be a parameterisation of a place $Q$. If the centre of $Q$ lies on the affine line $x = 0$, then $\psi(0) \in \operatorname{dom} \mathfrak{f}_P$, and the composition $\mathfrak{f}_P \circ \psi$ is analytic on a neighbourhood of 0.

**Lemma 15.68** *Let $\psi_1$ and $\psi_2$ be two parameterisations of $Q$. Then* $\operatorname{ord}_0(\mathfrak{f}_P \circ \psi_1) = \operatorname{ord}_0(\mathfrak{f}_P \circ \psi_2)$.

***Proof*** Take $h$ witnessing that $\psi_1$ and $\psi_2$ are equivalent: $\psi_2 = \psi_1 \circ h$ on a neighbourhood of 0, and $\operatorname{ord}_0(h) = 1$. Then $\mathfrak{f}_P \circ \psi_2 = (\mathfrak{f}_P \circ \psi_1) \circ h$, so $\operatorname{ord}_0(\mathfrak{f}_P \circ \psi_2) = \operatorname{ord}_0(\mathfrak{f}_P \circ \psi_1) \cdot \operatorname{ord}_0(h)$. □

We therefore define:

**Definition 15.69** Let $P$ be a central place, and let $Q$ be a place whose centre lies on the affine line $x = 0$. We let $i(P, Q)$ be the order $\operatorname{ord}_0(\mathfrak{f}_P \circ \psi)$, where $\psi$ is any parameterisation of $Q$.

We omitted a subscript from the notation $i(P, Q)$; the point is understood to be the centre of the places:

**Proposition 15.70** *$i(P, Q) > 0$ if and only if $P$ and $Q$ have the same centre.*

***Proof*** Let $\psi$ be a parameterisation of $Q$. Since the centre $\psi(0)$ of $Q$ lies on $x = 0$, by Lemma 15.65 it equals the centre of $P$ if and only if $\mathfrak{f}_P(\psi(0)) = 0$, i.e., if and only if $\operatorname{ord}_0(\mathfrak{f}_P \circ \psi) > 0$. □

Therefore, if $Q$ is a place whose centre does not lie on $x = 0$, then we define $i(P, Q) = 0$. Thus, $i(P, Q)$ is defined for any central place $P$ and any place $Q$.

**Proposition 15.71** *$i(P, Q) = \infty$ if and only if $P = Q$.*

***Proof*** Let $\psi$ be a parameterisation of $Q$. Then $i(P, Q) = \infty$ if and only if $\mathfrak{f}_P \circ \psi$ is the zero function on a neighbourhood of 0, which by Lemma 15.65 holds if and only if $\psi[U] \subseteq B$ where $B$ is a branch of the place $P$ and $U$ is a small open neighbourhood of 0. If $\psi[U] \subseteq B$ then $\psi[U]$ and $B$ are not essentially disjoint. By Theorem 15.57, $\psi[U] \subseteq B$ if and only if $P = Q$. □

We give an algebraic characterisation of the intersection number of two central places. Recall that we extended orders of power series to fractional power series (see Exercise 15.9).

**Lemma 15.72** *Let $P$ and $Q$ be central places. Let $n$ be the valency of $Q$. Then $i(P, Q) = n \cdot \operatorname{ord}(f_P(x, g))$, where $g$ is any Puiseux expansion of $Q$.*

***Proof*** Let $g = g_\zeta$ be a Puiseux expansion of $Q$; let $\eta = \zeta \circ \mathrm{pwr}_n$. Then $\eta$ is defined by the pair of $(x^n, g(x^n))$ of formal power series. So $i(P, Q)$ is the order of the formal power series $f_P(x^n, g(x^n))$, which is the result of substituting $x^n$ into the fractional power series $f_P(x, g)$. For any fractional power series $h$, we have $\operatorname{ord} h(x^n) = n \cdot \operatorname{ord} h$. □

Let $P$ and $Q$ be central places, of valencies $m$ and $n$, respectively. Let $g_1, \ldots, g_n$ be the Puiseux expansions of $Q$, and $h_1, \ldots, h_m$ be the Puiseux expansions of $P$. Applying Lemma 15.72 for each $i$, and summing over all $i \leqslant n$, we get

$$n \cdot i(P, Q) = n \cdot \sum_{i \leqslant n} \operatorname{ord}(f_P(x, g_i)).$$

By Exercise 15.9, $\sum_{i \leqslant n} \operatorname{ord}(f_P(x, g_i))$ is the order of the series

$$\prod_{i \leqslant n} f_P(x, g_i) = \prod_{i \leqslant n, j \leqslant m} (g_i - h_j).$$

So dividing by $n$, we obtain:

**Lemma 15.73** *If $P$ and $Q$ are central places, then the intersection number $i(P, Q)$ equals the order of the fractional power series*

$$\prod_{i \leqslant n, j \leqslant m} (g_i - h_j),$$

*where $g_1, \ldots, g_n$ are the Puiseux expansions of $Q$, and $h_1, \ldots, h_m$ are the Puiseux expansions of $P$.*

We immediately get symmetry:

**Proposition 15.74** *For any central places $Q$ and $P$, $i(P, Q) = i(Q, P)$.*

Also note that Lemma 15.73 gives another proof of Proposition 15.71 if $Q$ is central: $i(P, Q) = \infty$ if and only if $g_i = h_j$ for some $i$ and $j$, i.e., if and only if $P$ and $Q$ share a parameterisation.

### 15.5.1 Intersections of Curves and Places

We define the intersection multiplicity of a curve and a place. Let $D$ be a curve, and let $Q$ be a place. Let $f$ define $D$, and let $\boldsymbol{\psi}$ be an analytic presentation of a parameterisation $\psi$ of $Q$ (Proposition 15.24). Then $f \circ \boldsymbol{\psi}$ is an analytic function (note that $f$ does not define a function on $\mathbb{P}^2$, so $f \circ \psi$ is meaningless.)

By Remark 15.25, the order $\operatorname{ord}_0 f \circ \boldsymbol{\psi}$ does not depend on the choice of analytic presentation $\boldsymbol{\psi}$ of $\psi$. The argument of Lemma 15.68 shows that it also does not depend on the choice of parameterisation $\psi$ of $Q$. And it certainly does not depend on the choice of $f$ defining $D$. Hence we define:

**Definition 15.75** Let $D$ be a curve, and let $Q$ be a place. Let $f$ define $D$. We define $i(D, Q)$ to be the order $\operatorname{ord}_0(f \circ \boldsymbol{\psi})$, where $\boldsymbol{\psi}$ is an analytic presentation of a parameterisation of $Q$.

The following is immediate from the definition and our earlier analysis of branches:

**Proposition 15.76** *$i(D, Q) > 0$ if and only if the centre of $Q$ lies on $D$; $i(D, Q) = \infty$ if and only if $Q$ is a place of $D$.*

As usual, we like working in affine coordinates:

**Lemma 15.77** *If the centre of $Q$ is in $\mathbb{A}^2$, and $g$ defines $D \cap \mathbb{A}^2$, then $i(D, Q) = \operatorname{ord}_0 g \circ \psi$, for any parameterisation $\psi$ of $Q$.*

***Proof*** $g = f^\flat$ where $f$ defines $D$; by definition, $i(D, Q) = \operatorname{ord}_0 f(1, \psi_x, \psi_y) = \operatorname{ord}_0 g(\psi_x, \psi_y)$. □

The reason to use projective coordinates to begin with, is to deal with changes of coordinates. Recall that for a place $Q$ and change of coordinates $\alpha$, we let $\alpha[Q]$ be the place of the parameterisations $\alpha \circ \psi$, where $\psi$ are parameterisations of $Q$.

**Proposition 15.78** *Let $D$ be a curve and $Q$ be a place. For any change of coordinates $\alpha$, $i(\alpha[D], \alpha[Q]) = i(D, Q)$.*

***Proof*** Let $\boldsymbol{\alpha}$ be a linear presentation of $\alpha$. The curve $\alpha[D]$ is defined by the polynomial $\boldsymbol{\alpha}^*(f)$, which defines the function $f \circ \boldsymbol{\alpha}^{-1}$ on $\mathbb{C}^3$ (Proposition 4.30); On the other hand, if $\boldsymbol{\psi}$ is an analytic presentation of a parameterisation of $Q$, then $\boldsymbol{\alpha} \circ \boldsymbol{\psi}$ is an analytic presentation of a parameterisation of $\alpha[Q]$. □

*Remark 15.79* We would like to define the intersection number $i(P, Q)$ of any two places, even when neither is central (and we would like it to be invariant under changes of coordinates). The missing ingredient is Study's lemma: for each place $P$ we would like to choose $f_P$ defining $P$ near its centre, and we want to choose them in a way invariant under changes of coordinates. Since $P$ is not an algebraic curve (rather, it is analytic), Study's lemma does not apply, and the choice of $f_P$ is difficult. A judicious choice of a class of possible $f_P$'s, which results in a well-defined and invariant notion of multiplicity numbers, can be done using tools of multivariable complex analysis, in particular, the Weierstrass preparation theorem. «

**Proposition 15.80** *Suppose that* $(0:0:1) \notin D$*, and that* $Q$ *is a place whose centre lies on the affine line* $x = 0$*. Then* $i(D, Q) = \sum_P i(P, Q)$*, where* $P$ *ranges over the places of* $D$ *which share a centre with* $Q$*.*

***Proof*** Let $P_1, \ldots, P_k$ be the places of $D$ with the same centre as $Q$ (they are all central); and let $R_1, \ldots, R_m$ be the central places of $D$ with other centres. By Theorem 15.66, $f_D = f_{P_1} \cdots f_{P_k} f_{R_1} \cdots f_{R_m}$. Composing with a parameterisation of $Q$, we see that by Lemma 15.77, $i(D, Q) = i(P_1, Q) + \cdots + i(P_k, Q) + i(R_1, Q) + \cdots + i(R_m, Q)$. By Proposition 15.70, $i(R_1, Q) = \cdots = i(R_m, Q) = 0$. □

*Example 15.81* Let $Q_1$ be the place of parameterisation $z \mapsto (z, z + z^2/2 - z^3/8 + \cdots)$ of the nodal cubic, and let $D$ be the folium of Descartes, given by $f = x^3 + y^3 - 3xy$. Then

$$f(z, z+z^2/2-z^3/8+\cdots) = z^3+(z+z^2/2-z^3/8-\cdots)^3-3z(z+z^2/2-z^3/8-\cdots).$$

The lowest order term is $-3z^2$, so $i(D, Q_1) = 2$. «

### 15.5.2 Intersections of Curves

We can finally relate our work to intersection multiplicity of curves, as defined in Chap. 6. Recall that we can take the resultant $\mathrm{res}_y(f, g)$ of polynomials in any polynomial ring $R[y]$, in our case, $R = \mathbb{C}[\![x]\!]$; we get a resultant in $\mathbb{C}[\![x]\!]$.

**Lemma 15.82** *Let* $P$ *and* $Q$ *be central places. Then* $i(P, Q) = \mathrm{ord}(\mathrm{res}_y(f_P, f_Q))$.

***Proof*** By Corollary 6.31,

$$\mathrm{res}_y(f_P, f_Q) = \prod_{i \leqslant n, j \leqslant m} (g_i - h_j),$$

where $g_1, \ldots, g_n$ are the Puiseux expansions of $Q$ and $h_1, \ldots, h_m$ are the Puiseux expansions of $P$; apply Lemma 15.73. □

**Lemma 15.83** *Let* $D$ *be a curve,* $Q$ *be a central place, and suppose that* $(0:0:1) \notin D$*. Then* $i(D, Q) = \mathrm{ord}(\mathrm{res}_y(f_D, f_Q))$.

***Proof*** Let $P_1, \ldots, P_m$ be the central places of $D$, so $f_D = f_{P_1} \cdots f_{P_m}$. By Lemma 6.32, $\mathrm{res}_y(f_D, f_Q) = \mathrm{res}_y(f_{P_1}, f_Q) \cdots \mathrm{res}_y(f_{P_m}, f_Q)$; now apply Proposition 15.80 and 15.82 and Lemma 15.70. □

**Lemma 15.84** *Let $D$ and $E$ be curves. Suppose that $(0:0:1) \notin D, E$ and suppose that $p \in \mathbb{A}^2$ is the unique point on $D \cap E \cap (x = 0)$. Then $i_p(D, E) = \sum_Q i(D, Q)$, where $Q$ ranges over the places of $E$ at $p$.*

***Proof*** If $p$ lies on a common component of $D$ and $E$, then $i_p(D, E) = \infty$ and $D$ and $E$ share a place $Q$ at $p$, whence $i(D, Q) = \infty$ (Proposition 15.76). As before, since $(0:0:1) \notin E$, any place of $E$ with centre on $x = 0$ is central.

Suppose then that $D$ and $E$ do not share a common component. Since $p$ is the unique point on $D \cap E \cap (x = 0)$, if $q$ is the common centre of central places $P$ of $D$ and $Q$ of $E$, then $q = p$. So by Proposition 15.76, it suffices to show that $i_p(D, E)$ is the sum of $i(D, Q)$ where $Q$ ranges over all central places of $E$, regardless of centre.

Let $Q_1, \ldots, Q_n$ be the central places of $E$. By Lemma 15.83, $i(D, Q_j)$ is the order of $\mathrm{res}_y(f_D, f_{Q_j})$. By Theorem 15.66 and Lemma 6.32, $\prod_j \mathrm{res}_y(f_D, f_{Q_j}) = \mathrm{res}_y(f_D, f_E)$. By Remark 6.26, $i_p(D, E)$ is the order of $\mathrm{res}_y(f_D, f_E)$ (as this is a polynomial, this is the multiplicity of 0 as its root). □

**Corollary 15.85** *For any two curves $D$ and $E$, for any point $p$, $i_p(D, E) = \sum_Q i_p(D, Q)$, where $Q$ ranges over the places of $E$ at $p$.*

Note that this implies $\sum_Q i(D, Q) = \sum_P i(P, E)$.

***Proof*** Change coordinates so that the hypotheses of Lemma 15.84 hold; this is permitted by Propositions 15.78 and 6.21. □

*Example 15.86* Continuing Example 15.81, the other place $Q_2$ of the nodal cubic at the origin (call it $E$) has parameterisation $z \mapsto (z, -z - z^2/2 + z^3/8 + \cdots)$. A similar calculation shows that $i(D, Q_2) = 2$ as well. Hence $i_o(D, E) = i(D, Q_1) + i(D, Q_2) = 4$. An argument using Proposition 6.40 will give the same result (try it!). «

*Example 15.87* We compute $i_o(y^3 + 2xy + x^6, y - x^2)$, which we first did in Example 6.41. The curve $y = x^2$ has a single place $Q$ at the origin, with parameterisation $z \mapsto (z, z^2)$, which we plug into the defining polynomial $y^3 + 2xy + x^6$, and get $2z^3 + 2z^6$; so the multiplicity of intersection is 3. «

Reducing to places on both sides, we get:

**Corollary 15.88** *Let $D$ and $E$ be curves. Suppose that $(0:0:1) \notin D \cap E$, and let $p \in \mathbb{A}^2 \cap (x = 0)$. Then $i_p(D, E) = \sum_{P,Q} i(P, Q)$, where $P$ ranges over the places of $D$ at $p$ and $Q$ ranges over the places of $E$ at $p$.*

***Proof*** Combine Corollary 15.85 with Proposition 15.80. □

### 15.5.3 Orders and Tangents of Places

**Definition 15.89** Let $P$ be a place; let $\boldsymbol{\psi} = (\psi_w, \psi_x, \psi_y)$ be an analytic presentation of a parameterisation of $P$ (Proposition 15.24). Let $(p_w, p_x, p_y) = (\psi_w, \psi_x, \psi_y)(0)$ (a presentation of the centre of $P$). We let the *order* of the place $P$, denoted by $o(P)$, be $\min\{\mathrm{ord}_0(\psi_w - p_w), \mathrm{ord}_0(\psi_x - p_x), \mathrm{ord}_0(\psi_y - p_y)\}$.

Remark 15.25 and the definition of equivalence of parameterisations imply that this is a good definition: the value $o(P)$ does not depend on the choice of parameterisation $\psi$ of $P$ or of analytic presentation $\boldsymbol{\psi}$ of $\psi$. By definition, the order cannot be 0; since not all components can be constant, the order is finite.

We get geometric invariance:

**Proposition 15.90** *If $P$ is a place and $\alpha$ is a change of coordinates, then $o(\alpha[P]) = o(P)$.*

***Proof*** Let $\boldsymbol{\psi}$ be an analytic presentation of a parameterisation of $P$, and let $\boldsymbol{\alpha}$ be a linear presentation of $\alpha$. Then $\boldsymbol{\varphi} = \boldsymbol{\alpha} \circ \boldsymbol{\psi}$ is an analytic presentation of a parameterisation of $\alpha[P]$. Let $\boldsymbol{p} = \boldsymbol{\psi}(0)$ and $\boldsymbol{q} = \boldsymbol{\alpha}(\boldsymbol{p}) = \boldsymbol{\varphi}(0)$. Since $\boldsymbol{\alpha}$ is linear, $\boldsymbol{\alpha} \circ (\boldsymbol{\psi} - \boldsymbol{p}) = \boldsymbol{\varphi} - \boldsymbol{q}$. So every component of $\boldsymbol{\varphi} - \boldsymbol{q}$ is a linear combination of the components of $\boldsymbol{\psi} - \boldsymbol{p}$. Linear combinations cannot introduce lower-order terms (they can only increase the order if there is some cancellation of lowest-order terms). This shows that $o(\alpha[P]) \geqslant o(P)$. Since $\alpha$ is invertible, we get equality. □

And in affine:

**Lemma 15.91** *If $P$ is an affine place (its centre $p = (p_x, p_y)$ is in $\mathbb{A}^2$), and $\psi$ is a parameterisation of $P$, then $o(P) = \min\{\mathrm{ord}_0(\psi_x - p_x), \mathrm{ord}_0(\psi_y - p_y)\}$.*

***Proof*** Choose $\boldsymbol{\psi} = (1, \psi_x, \psi_y)$; then $\mathrm{ord}_0(\psi_w - 1) = \infty$. □

**Proposition 15.92** *Let $P$ be a place, with centre $p$, and let $k = o(P)$. There is a unique line $\ell$ passing through $p$ such that $i(\ell, P) > k$. For all other lines $\ell$ passing through $p$, we have $i(\ell, P) = k$.*

***Proof*** By Propositions 15.78 and 15.90, we may change coordinates, so that the centre of $P$ is the origin. The lines through the origin are defined by $ay - bx$ for $(a:b) \in \mathbb{P}^1$. Let $\psi$ be a parameterisation of $P$. If $\ell$ is the line $ay = bx$ then $i(\ell, P) = \mathrm{ord}_0(a\psi_y - b\psi_x)$. There is precisely one choice of $(a:b)$ which would make the lowest-order term in $a\psi_y - b\psi_x$ vanish: if $\mathrm{ord}_0 \psi_x = \mathrm{ord}_0 \psi_y$ then we choose nonzero $a, b$ that would cause cancellation of the lowest-order terms; if $\mathrm{ord}_0 \psi_x < \mathrm{ord}_0 \psi_y$ then we choose $b = 0$, similarly in the other case. All other choices give $\mathrm{ord}_0(a\psi_y - b\psi_x) = \min\{\mathrm{ord}_0(\psi_x), \mathrm{ord}_0(\psi_y)\} = o(P)$. □

**Definition 15.93** The *tangent* of a place $P$ is the unique line $\ell$ satisfying $i(\ell, P) > o(P)$. We write $\ell(P)$ for the tangent of $P$.

Every line $\ell$ passing through the origin has a unique place $Q_\ell$ with centre $o$, given by a parameterisation $\psi_\ell(z) = (az, bz)$ (where $\ell$ is the line $ay = bx$).

**Lemma 15.94** *For any line $\ell$ passing through the origin and any central place $P$ at the origin, $i(\ell, P) = i(P, Q_\ell)$.*

***Proof*** If $\ell$ is not vertical then $Q_\ell$ is central and $(0:0:1) \notin \ell$; in this case the result follows from Propositions 15.74 and 15.80. Suppose that $\ell$ is the $y$-axis $x = 0w$. Let $\eta$ be an $n$-fold parameterisation of $P$. Then $i(\ell, P) = n$ (the order of $\eta_x = z^n$), and $\mathrm{ord}_0(\mathfrak{f}_P \circ \psi_\ell)$ is the order of the formal power series $f_P(0, y) = y^n$. □

Let $P$ be a central place, with centre the origin. Then $f_P \in \mathbb{C}[\![x, y]\!]$ has no constant term. Write $f_P = \sum_{k>0} f_{P,k}$, where each $f_{P,k} \in \mathbb{C}[x, y]$ is homogeneous of degree $k$.

**Proposition 15.95** *Let $P$ be a central place with centre the origin. Then $o(P)$ is the least $k$ such that $f_{P,k} \neq 0$, and $V_{\mathbb{P}^2}(f_{P,o(P)})$ consists of $o(P)$-many copies of $\ell(P)$.*

***Proof*** By Lemma 15.94, $i(ay = bx, P) = \mathrm{ord}_0(\mathfrak{f}_P(az, bz)) = \mathrm{ord}(f_P(at, bt))$. Now $f_P(at, bt) = \sum_{k>0} f_{P,k}(at, bt)$. For each $k$, if $f_{P,k} \neq 0$ then the curve $V_{\mathbb{A}^2}(f_{P,k})$ is the sum of $k$-many lines through the origin (repetitions allowed), and $f_{P,k}(at, bt) = 0$ if and only if $ay = bx$ is one of these lines. For $k < o(P)$, since $i(\ell, P) > k$ for all $\ell$, $f_{P,k}(at, bt) = 0$ for any $(a:b)$, so $f_{P,k} = 0$. And $f_{P,o(P)}(at, bt) = 0$ if and only if $i(ay = bx, P) > o(P)$ if and only if $ay = bx$ is the tangent $\ell(P)$. □

**Proposition 15.96** *Let $D$ be a curve and let $p \in D$.*

*(a) $o_p(D) = \sum_P o(P)$, where $P$ ranges over the places of $D$ at $p$.*
*(b) A line is tangent to $D$ at $p$ if and only if it is tangent to some place $P$ of $D$ at $p$; in fact, $\ell_p D$ is the multiset sum of $o(P)$ copies of $\ell(P)$, for all places $P$ of $D$ at $p$.*

***Proof*** By changing coordinates, we assume that $(0:0:1) \notin D$, and that $p$ is the origin. By Theorem 15.66, $f_D = f_{P_1} \cdots f_{P_m} f_{R_1} \cdots f_{R_n}$, where the $P_i$'s are the central places of $D$ at the origin, and the $R_j$'s are the other central places of $D$. Each $f_{R_j}$ has nonzero constant term. Using the notation above, the least $k$ with $f_{D,k} \neq 0$ is a constant multiple of $f_{P_1,o(P_1)} \cdots f_{P_m,o(P_m)}$. The proposition then follows from Proposition 15.95 and Proposition 5.16. □

*Example 15.97* The two places of the nodal cubic $y^2 = x^3 + x^2$ at the origin both have order 1, and their tangents are $y = x$ and $y = -x$; the sum of these are the tangents to the cubic at the origin, and the order of the origin on the cubic is 2. «

*Exercise 15.98* Use Proposition 15.96 to give a proof of Theorem 5.34 for curves in $\mathbb{P}^2(\mathbb{C})$ with no repeated components. «

We call a place $P$ *singular* if $o(P) > 1$, *nonsingular* otherwise.

**Corollary 15.99** *A point $p$ is nonsingular on $D$ if and only if $D$ has a single, nonsingular place at $p$.*

### 15.5.4 Some Nifty Consequences

We can now give more transparent proofs of some of the results of Chap. 6, for complex curves with no repeated components. The original proofs relied on manipulations of matrices and resultants. For example, the symmetry property $i_p(D, E) = i_p(E, D)$ follows from Proposition 15.74 and Corollary 15.88; and the additivity property Proposition 6.29, $i_p(C, D + E) = i_p(C, D) + i_p(C, E)$ follows from Corollary 15.88 and Proposition 15.43, as the places of $D + E$ are the places of $D$ together with the places of $E$. This is not that surprising, as the proofs of both Proposition 6.29 and Corollary 15.88 both rely on Lemma 6.30; but this still gives us some insight as to what's really going on. We give three deeper applications.

#### Intersections with Shifted Curves

One of the more mysterious properties of intersection multiplicity is the invariance under shifting curves. In affine coordinates, this is Proposition 6.40(5): $i_p(f, g) = i_p(f, fh + g)$. We give another proof.

***Proof of Proposition 6.40(5)*** By Corollary 15.85, it suffices to show that for every place $P$ of the curve $f = 0$, we have $i(g, P) = i(fh + g, P)$. Let $\psi$ be a parameterisation of $P$. Then $i(g, P) = \mathrm{ord}_0(g \circ \psi)$ and $i(fh + g, P) = \mathrm{ord}_0((fh + g) \circ \psi)$. Since $P$ is a place of $f = 0$, $f \circ \psi$ is the zero function, so $g \circ \psi = (fh + g) \circ \psi$. □

#### Intersection Multiplicity, Orders, and Shared Tangents

We give another proof of Proposition 6.43.

***Proof of Proposition 6.43*** As in Chap. 6, we may assume that $p \in D$, so $i_p(C, \ell) > 1$. We change coordinates so that $p$ is the origin and $\ell$ is the line $y = 0$. Since $p$ is nonsingular on $C$, there is a single place $P$ of $C$ with centre at $p$, and $o(P) = 1$ (Corollary 15.99). By Corollary 15.85, $i_p(C, \ell) = i(\ell, P)$. Let $\psi$

be a parameterisation of $P$. Then $i(\ell, P) = \mathrm{ord}_0(\psi_y)$, whence $\mathrm{ord}_0(\psi_y) > 1$; so $1 = o(P) = \mathrm{ord}_0(\psi_x)$.

Let $k = i_p(D, \ell)$; let $f$ define $D \cap \mathbb{A}^2$. Since $f(0,0) = 0$, write $f = x^k u(x) + yv(x, y)$, where $u$ and $v$ are polynomials and $u(0) \neq 0$. We need to show that $k = i_p(C, D)$ and the assumption is that $k < i_p(C, \ell)$, i.e., that $k < \mathrm{ord}_0(\psi_y)$. Since $P$ is the unique place of $C$ with centre $P$, by Corollary 15.85, $i_p(C, D) = i(D, P)$, i.e., $i_p(C, D) = \mathrm{ord}_0(f(\psi_x, \psi_y))$. However $f(\psi_x, \psi_y) = \psi_x^k u(\psi_x) + \psi_y v(\psi_x, \psi_y)$; since $u(0) \neq 0$, $\mathrm{ord}_0(\psi_y) > k$ and $\mathrm{ord}_0(\psi_x) = 1$, the order of $f(\psi_x, \psi_y)$ is $k$. □

In Chap. 6, we stated Theorem 6.42, but did not give a full proof. We can now provide one.

**Proposition 15.100** *Let $P$ and $Q$ be central places with the same centre. Then $i(P, Q) \geqslant o(P) \cdot o(Q)$; equality holds if and only if $P$ and $Q$ do not share a tangent.*

***Proof*** For notational simplicity, suppose that the common centre is the origin; for the general case we only need to add a constant to the $y$-values everywhere (note that we cannot rely on changes of coordinates when considering the intersection of two places).

Let $\psi$ be a parameterisation of $Q$. For all $k$, $\mathrm{ord}_0(f_{P,k}(\psi_x, \psi_y)) \geqslant k \cdot o(Q)$, so $i(P, Q) = \mathrm{ord}_0(\mathfrak{f}_P \circ \psi) \geqslant o(P) \cdot o(Q)$. Equality holds if and only if $\mathrm{ord}_0(f_{P,o(P)}(\psi_x, \psi_y)) = o(P) \cdot o(Q)$. If $\ell(P)$ is the line $ay = bx$ then $f_{P,o(P)} = (ay - bx)^{o(P)}$, so $i(P, Q) > o(P) \cdot o(Q)$ if and only if $o(Q) < \mathrm{ord}_0(a\psi_y - b\psi_x)$. But $\mathrm{ord}_0(a\psi_y - b\psi_x) = i(\ell(P), Q)$ and $i(\ell, Q) > o(Q)$ if and only if $\ell = \ell(Q)$. □

***Proof of Theorem 6.42*** Let $C$ and $D$ be curves. We may assume they have no repeated components. Let $p \in C \cap D$. We want to show that $i_p(C, D) \geqslant o_p(C) \cdot o_p(D)$ and equality holds if and only if $C$ and $D$ have no shared tangent at $p$.

After a change of coordinates we assume that $(0:0:1) \notin C, D$ and that $p$ is the origin. Let $P_1, \ldots, P_m$ be the places of $C$ at $p$; let $Q_1, \ldots, Q_n$ be the places of $D$ at $p$; all are central. By Proposition 15.96, $o_p(C) \cdot o_p(D) = \sum_{i,j} o(P_i) o(Q_j)$, and $C$ and $D$ share a tangent at $p$ if and only if $\ell(P_i) = \ell(Q_j)$ for some pair $i$ and $j$. By Corollary 15.88, $i_p(C, D) = \sum_{i,j} i(P_i, Q_j)$. The result then follows from Proposition 15.100. □

*Example 15.101* Let $E$ be the nodal cubic $y^2 = x^3 + x^2$, and let $D$ be the folium of Descartes $y^3 + x^3 = 3xy$. The origin has order 2 on both curves; the tangents to $E$ at the origin are $y = x$ and $y = -x$, while the tangents to $D$ at the origin are $y = 0$ and $x = 0$. So they don't share a tangent at the origin; Theorem 6.42 implies that $i_o(E, D) = 4$. Compare with Example 15.86. «

## 15.6 Further Exercises

**Formal and Informal Power Series**

15.102 In this exercise we give a "direct" proof of Proposition 15.1(c). Let $f = \sum a_n x^n$ and $g = \sum b_n x^n$ be formal power series in $\mathbb{C}[\![x]\!]$, and suppose that $R(f), R(g) > 0$. Let $R = \min\{R(f), R(g)\}$. Let $p = fg$. (a) Let $c_n = \sum_{i+j=n} |a_i b_j|$. Use Proposition 11.46 to show that for all $s > 1/R$ there is some $M$ such that for all but finitely many $n$, $c_n \leqslant (M+n)s^n$. Conclude that the radius of convergence of $\sum c_n z^n$ is at least $R$. Conclude that $R(p) \geqslant R$. (b) Let $f_m = \sum_{k \leqslant m} a_k x^k$ be the degree $m$ part of $f$, and similarly define $g_m$ and $p_m$. Show that $\operatorname{ord}(f_m g_m - p_m) \geqslant m+1$. (c) Let $z \in B(0,R)$. Show that $|f_m g_m(z) - p_m(z)| \leqslant \sum_{n>m} c_n |z|^n$. (Show that for $n > m$, $|d_n - e_{m,n}| \leqslant c_n$, where $p = \sum d_n x^n$ and $f_m g_m = \sum e_{m,n} x^n$.) (d) Using $\mathfrak{f}_m(z) \mapsto \mathfrak{f}(z)$ (and similarly for $g$ and $p$), show that $\mathfrak{p}(z) = \mathfrak{f}\mathfrak{g}(z)$.

15.103 We give yet another proof of Proposition 15.1(c). (a) Suppose that $\sum a_n$ and $\sum b_n$ both converge absolutely. Let $c_n = \sum_{i \leqslant n} a_i b_{n-i}$. Show that $\sum c_n$ converges absolutely to $\sum_{n,m} a_n b_m$ (see Exercises 11.85 and 11.86.) (b) Let $d_n = (\sum_{i \leqslant n} a_i) \cdot (\sum_{i \leqslant n} b_i)$. Show that $\langle d_n \rangle$ converges absolutely to $\sum_{n,m} a_n b_m$. (c) Use this to prove Proposition 15.1(c).

15.104 In this exercise we give a proof of Proposition 15.6. Let $f, g \in \mathbb{C}[\![x]\!]$; suppose that $R(f), R(g) > 0$. Suppose that $\operatorname{ord}(f) > 0$; let $h = g(f)$. Write $f = \sum a_i x^i$, $g = \sum c_i x^i$, and $h = \sum d_i x^i$. For $k \geqslant 1$ write $f^k = \sum b_{k,i} x^i$. Also, let $\hat{f} = \sum |a_i| x^i$ and write $(\hat{f})^k = \sum \hat{b}_{k,i} x^i$. (a) Show that there is some positive $R \leqslant R(f)$ such that $\hat{\mathfrak{f}}[B(0,R)] \subseteq B(0, R(g))$. (b) Show that $|b_{k,n}| \leqslant \hat{b}_{k,n}$ for all $k$ and $n$. (c) Let $z \in B(0,R)$. Show that $\sum_{n,k} c_k b_{k,n} z^n$ converges absolutely. (d) Let $\varphi = \mathfrak{g} \circ \mathfrak{f}$. Show that $\varphi(z) = \sum_{k=0}^{\infty} \sum_{n=0}^{k} c_k b_{k,n} z^n$ for all $z \in B(0,R)$. (e) Conclude that $\varphi = \mathfrak{h}$ on $B(0,R)$.

15.105 Let $F$ be a field; let $g \in F[\![x]\!]$, and suppose that $\operatorname{ord}(g) = 1$. (a) Show that there is a (unique) $h \in F[\![x]\!]$ with $\operatorname{ord}(h) > 0$ such that $g(h) = x$. (b) Conclude that the map $f \mapsto f(g)$ is a ring automorphism of $F[\![x]\!]$.

15.106 (a) Show that the formal chain rule holds for substitution into power series: if $R$ is an integral domain, $h \in R[\![y_1, y_2, \ldots, y_m]\!]$, $g_1, g_2, \ldots, g_m \in R[\![x]\!]$ and $\operatorname{ord}(g_i) > 0$ for $i \leqslant m$, then $D^x(h(\boldsymbol{g})) = \sum_{i \leqslant k} D^{y_i} h(\boldsymbol{g}) D^x g_i$ (see Remark 5.4). (b) Use this to prove the multivariable complex chain rule (Proposition 13.2) when $n = k = 1$.

**Vertical Parameterisations and the Implicit Function Theorem**

15.107 (a) Let $f(x, y) = 2x - x^2 + y - xy + y^2$. Let $g(x) = a_1x + a_2x^2 + \cdots$ satisfy $f(x, g) = 0$ in the ring $\mathbb{C}[\![x]\!]$. Find $a_1$, $a_2$, $a_3$ and $a_4$. (b) Generalise this to show that if $h \in \mathbb{C}[\![x, y]\!]$ has a zero constant term and a term linear in $y$, then there is a unique $g \in \mathbb{C}[\![x]\!]$ with $\operatorname{ord}(g) > 0$ and $h(x, g) = 0$.[2]

In the following exercise we give an alternative proof of the implicit function theorem, without relying on the inverse function theorem. For more details, see, for example, [Gri89, Ex.9.4]. Since we have not explored the connection between multivariable power series and the multivariable analytic functions that they define, we restrict ourselves to a polynomial implicit equation. The uniqueness of the implicit function theorem for this case was proved in Exercise 13.61, so we prove existence.

15.108 Let $f = \sum a_{i,j}x^i y^j \in \mathbb{C}[x, y]$. Suppose that $f(0, 0) = 0$ and $D^y f(0, 0) \neq 0$. Without loss of generality, assume that $a_{0,1} = 1$. Let $h = a_{1,0}x - y - f$; write $h = \sum \hat{a}_{i,j}x^i y^j$.

(a) Let $g = \sum b_n x^n \in \mathbb{C}[\![x]\!]$ satisfy $f(x, g) = 0$ (Exercise 15.107). So $h(x, g) = g$. Show that there are polynomials $P_n \in \mathbb{N}[t_{i,j}]_{i,j\leqslant n}$ such that $c_n = P_n(\hat{a}_{i,j})_{i,j\leqslant n}$. The point is that the coefficients of the $P_n$ are nonnegative.

(b) Let $M = \max\{|\hat{a}_{i,j}|\}$. For ease of notation, suppose that $M = 1$ (the argument is the same in the general case). Let $H = c_{i,j}x^i y^j \in \mathbb{C}[\![x, y]\!]$ where $c_{0,0} = c_{0,1} = 0$ and $c_{i,j} = 1$ for all other $(i, j)$. Show that

$$H = \frac{1}{(1-x)(1-y)} - y - 1$$

in $\mathbb{C}[\![x, y]\!]$. Thus, we define $H(a, b)$ for $a, b \in B(0, 1)$. (In the general case, we replace $H$ by $MH$.)

(c) Show that there is an analytic function $\psi$, defined on a neighbourhood of 0, such that $\psi(0) = 0$ and $H(z, \psi(z)) = 0$ on a neighbourhood of 0. (Hint: let $F(x, y) = (1-x)(1-y)(2y+1) - 1$. Find the discriminant of $F$ with respect to $y$, and show that it is nonzero on a neighbourhood of 0. Define $\psi$ using the quadratic formula and an analytic choice of the square root.)

(d) Let $\sum d_n x^n$ be the formal power series defining $\psi$ around 0. Show that $|c_n| \leqslant d_n$ for all $n$. (Use the polynomials $P_n$ to compute both $c_n$ and $d_n$, and use $|\hat{a}_{i,j}| \leqslant c_{i,j}$). Conclude that $R(g) > 0$, and so $\mathfrak{g}$ satisfies $f(z, \mathfrak{g}(z)) = 0$ on a neighbourhood of 0.

[2] Note that with this method we can compute the coefficients of a power series defining a vertical parameterisation of $f = 0$.

**Fractional Parameterisations**

15.109 Let $D$ be a curve in $\mathbb{P}^2$, and suppose that $(0:0:1) \in D$. Show that there is a fractional parameterisation of $D$ whose centre is $(0:0:1)$ if and only if the line $x = 0w$ is a tangent to $D$ at $(0:0:1)$. (Hint: in one direction you can use Exercise 13.65. In the other, let $P$ be a place of $D$ at $(0:0:1)$ whose tangent is $x = 0w$, and consider a parameterisation of $P$.)

15.110 Let $D$ be the projective closure of $y^k = f(x, y)$, where $\deg_y f < k$ (but note that possibly $\deg f > k$). (a) Show that $(0:0:1) \in D$ if and only if $\deg f > k$, in which case the only tangent to $D$ at $(0:0:1)$ is $\ell_\infty$. (b) Show that $D$ has $k$-many germs of $n$-fold parameterisations. (c) Show that $f_D = f_{P_1} \cdots f_{P_k}$, where $P_1, \ldots, P_k$ are the central places of $D$.

15.111 In this exercise we give an alternative proof of Proposition 15.32 in the case that $(0:0:1) \notin D$. Assuming this, let $\zeta$ be a fractional parameterisation of $D$. (a) Show that $\zeta$ is bounded on a punctured neighbourhood of 0 in $\Sigma/n$. (Let $f(x, y)$ be a polynomial defining $D \cap \mathbb{A}^2$. Consider $f$ as a polynomial in $y$ with coefficients in $\mathbb{C}[x]$; we may assume that $y^d$ (for $d = \deg D$) appears in $f$. The coefficients of $f(a, y)$ are bounded for $a \in B(0, r)$. Hence if $|b|$ is large, the term $b^d$ dominates the other terms in $f(a, b)$ (see the proof of the fundamental theorem of algebra, page 304). So for sufficiently large $M$, for all $a \in B(0, r)$, for all $b$ such that $(a, b) \in D$, we have $|b| < M$.) (b) Use Proposition 12.16 to prove Proposition 15.32.

15.112 We give an alternative proof of Corollary 15.39, using Theorem 15.66 (which, note, uses Propositions 15.37 and 15.58 but not Corollary 15.38).

By changing coordinates, we may assume that $(0:0:1) \notin D$ and $p = (0, b)$ for some $b$. Show that $\mathfrak{f}_P(0, b) = 0$ for some central place $P$ of $D$; conclude that $p$ is the centre of $P$.

15.113 Let $D$ be a curve; let $\zeta : W^* \to D$ be a continuous function, where $W^*$ is a punctured neighbourhood of 0 in $\Sigma/n$. Suppose that for all $q \in W^*$, $\zeta(q)$ lies on the line $x = \pi_\Sigma(q)$. Show that $\zeta$ is holomorphic on a punctured neighbourhood $V^*$ of 0 in $\Sigma/n$ (and thus $\zeta \restriction_{V^*}$ is a fractional parameterisation of $D$).

**An Application to Intersection with Lines**

15.114 Let $\psi$ be an affine parameterisation of a curve $D$. Let $\lambda \in \operatorname{dom} \psi$ and suppose that $\psi'(\lambda) = (\psi_x', \psi_y')(\lambda) \neq (0, 0)$. Suppose also that $p = \psi(\lambda)$ is nonsingular on $D$. Show that the tangent to $D$ at $p$ is parallel to the vector $\psi'(\lambda)$.[3]

[3] Compare to the original motivation for defining tangents, at the very beginning of Chap. 5.

15.115 Let $P$ be a place of a curve $D$, with centre $p$. Suppose that $B$ is a branch of $P$, sufficiently small so that every $q \in B$ other than $p$ is nonsingular on $D$. For such $q$ let $\ell_q(P) = \ell_q(D)$; let $\ell_p(D) = \ell(P)$. Show that the map $q \mapsto \ell_q(P)$ is continuous on a neighbourhood of $p$ in $B$ (as a map to $\check{\mathbb{P}}^2$). (Use Exercise 15.114. You may change coordinates (explain why!) so that $p$ is the origin and $\ell(P)$ is the $x$-axis, so that $\operatorname{ord}_0 \psi_y > \operatorname{ord}_0 \psi_x$ for a parameterisation $\psi$ of $P$. Show that $\psi_y'/\psi_x' \to 0$ at 0.)

15.116 Let $D$ be a curve in $\mathbb{P}^2$, let $p \in D$, and let $L$ be a line which passes through $p$ but is not a tangent to $D$ at $p$. (a) Show that there is a neighbourhood $U$ of $p$ in $D$ and a neighbourhood $O$ of $L$ in $\check{\mathbb{P}}^2$ such that for all $\ell \in O$ and $q \in U$, $\ell$ is not a tangent to $D$ at $q$. (Hint: by changing coordinates suppose that $p$ is the origin and $L$ is the $x$-axis. Use Example 13.41 and Exercise 15.114; consider each place at the origin separately.) (b) Conclude that there is a neighbourhood $U$ of $p$ in $D$ and a neighbourhood $O$ of $L$ in $\check{\mathbb{P}}^2$ such that for all $\ell \in O$, either $\ell \cap U = \{p\}$, or $\ell \cap U$ consists of $o_p(D)$ many distinct points.

**Some Examples**

We give some examples in which Theorem 6.42 cannot be used to compute intersection multiplicities, as the curves have shared tangents. Observe that this is reflected in cancellation of lowest-order terms.

15.117 Let $D$ be the cardioid (Fig. 3.1) and let $E$ be the cuspidal cubic $y^2 = x^3$. (a) Compute $o_o(D) \cdot o_o(E)$. (b) Let $P$ be the unique place of $E$ at the origin $o$. Compute $i(D, P)$ and hence $i_o(D, E)$.

15.118 Let $D$ be the eight curve (Fig. 5.5) and let $E$ be the nodal cubic $y^2 = x^3 + x^2$. (a) Compute $o_o(D) \cdot o_o(E)$. (b) Let $Q_1$ and $Q_2$ be the two places of $E$ at the origin; compute $i(D, Q_1)$ and $i(D, Q_2)$, and so $i_o(D, E)$. (c) Find parameterisations for the two places $P_1$ and $P_2$ of $D$ at the origin. Compute $i(E, P_1)$ and $i(E, P_2)$ and so $i_o(D, E)$.

15.119 Let $D$ be the folium of Descartes (Fig. 3.4).

(a) Find parameterisations of the two places of $D$ at the origin. (See Exercise 3.49; try both $y = tx$ and $x = ty$.)
(b) Use these to compute the intersection of $D$ at the origin with the cuspidal cubic $y^2 = x^3$, and with the quadrifolium (Fig. 5.1).

**Miscellaneous Exercises**

15.120 Let the origin $o$ be nonsingular on a curve $D$, and suppose that the $x$-axis is the tangent to $D$ at $o$. Let $P$ be the place of $D$ at $o$, and let $\psi$ be a parameterisation of $P$. What is the valency of $P$? What is $f_P$?

15.121 Let $C$, $D$ and $E$ be three curves, and let $p$ be a point, which is nonsingular on each of these curves. Show that two of $i_p(C, D)$, $i_p(C, E)$ and $i_p(D, E)$ are equal, and less than or equal to the third.

15.122 A polynomial $f = \sum c_{i,j} x^i y^j \in \mathbb{C}[x, y]$ is *quasi-homogeneous* of degree $d$ if there are natural numbers $p, q > 0$ such that $pi + qj = d$ whenever $c_{i,j} \neq 0$. (This holds if and only if $f(t^p x, t^q y) = t^d f(x, y)$.) The pair $(p, q)$ is the pair of *weights* of $f$. (a) Show that if $f$ is quasi-homogeneous with weights $(p, q)$ then the curve $f = 0$ has an $p$-fold parameterisation $z \mapsto (z^p, cz^q)$ for some constant $c \in \mathbb{C}$. (b) Find a Puiseux expansion for the curve $y^4 = x^6 - 2x^3y^2$.[4]

[4] This is the first step in the *Newton polygon* method for computing Puiseux expansions.


# A Brief History of Elliptic Functions

# 16

The theory of elliptic functions and curves led to a profound unification of much of nineteenth century mathematics. It was the first arena in which topology, geometry, number theory, analysis, and algebra met in a significant and mutually revealing way. This book was an exposition for undergraduates of that mid-nineteenth century synthesis. To grasp elliptic function theory is to grasp the unity of mathematics.

## 16.1 A History of Circles and Ellipses

The Sumerians (3000–2000 BCE) developed base 60 arithmetic. They had efficient algorithms for both the four rational fundamental operations and for the extraction of square roots. The Babylonian star catalogues (1200 BCE) were tables exhibiting the periodicities of heavenly events such as the rising and setting of the sun, the waxing and waning of the moon, the four seasons. These tables correlated heavenly events with earthly events such as plantings, harvests, festivals, the crowning and death of kings, elections, battles, athletic and religious ceremonies. Time was measured in hours of the day, day of the lunar month, sidereal year, but had different values in different localities. This locality was hardly helpful to travelers on the seas and along caravan routes: simple instruments for measuring time and space, such as the theodolite, astrolabe, water clock, and gnomon, were developed to help remedy this.

The Babylonians (1894–539 BCE) developed the Zodiac. Their celestial sphere had earth at its center and all the fixed stars mounted on the boundary of the very distant celestial sphere. The axis of rotation of the celestial sphere was a straight line through the center of the earth perpendicular to the equatorial plane terminating at the north polestar. The celestial sphere appears to rotate around the earth once a day. The ecliptic plane was defined as the plane of the apparent circular path of the sun on the celestial sphere over the course of the year. The Zodiac is a circular band on the celestial sphere extending about eight degrees north and south of the ecliptic.



A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2_16

The Zodiac contained the orbits of the known planets, the moon, and the sun. The orbits appear as circles on the celestial sphere.

The late Babylonian Zodiac (Seleucid Empire, 312–63 BCE) and early Greek Zodiac (Eudoxus, 395–340 BCE) was divided into twelve sectors, or signs, each sign identified by a constellation in the Zodiac band: Aries, Taurus, Gemini, Cancer, Leo, Virgo, Libra, Scorpio, Sagittarius, Capricorn, Aquarius and Pisces. The Greeks, unlike predecessors, tried to design the sectors of the zodiac to be of equal length so that dates of entry of the sun into Cancer, Libra, Capricorn, and Aries were the dates of the summer solstice, autumnal equinox, winter solstice, and vernal equinox.

No geometric model of the motions of the sun and planets on the celestial sphere seems to have survived from the period before the Greece of Pythagoras (570–495 BCE) and Plato (424–348 BCE). Perhaps one never previously existed. The Greeks of classical Greece modelled the planets and the sun as rotating at uniform but distinct velocities, with their circles of rotation situated in different planes passing through or near the center of the earth. This was the geocentric world-view. By 300 BCE, the Babylonian astronomical tables had been acquired by the Greeks. In his *Phaenomena*, Euclid (c. 300 BCE) used his ruler and compass geometry in three dimensions to write formulas for calculating the length of day anywhere on earth.

The modern definition of angle as a real number was absent in Greek theoretical astronomy because the Greeks did not have our conception of a real number. The only numbers, or in Euclid's terms the only points, were those that can be constructed by ruler and compass from a chosen unit line segment. So angles had to be represented in terms of the points and figures that could be constructed by ruler and compass.

Suppose that a regular $n$-gon can be inscribed in a circle by ruler and compass. It will meet the circle in $n$ evenly spaced points. Connect the center to each of these $n$ points. The regular $n$-gon is then divided into $n$ congruent triangles. The bounding circle is divided into into $n$ congruent arcs, the area of the circle into $n$ congruent sectors. A *chord* is by definition the length of any of the $n$ sides of the inscribed $n$-gon. In our modern notation, if $\theta$ is the central angle of one of the $n$ congruent triangles, then the length of the chord is $2r \sin \theta/2$. Euclid and predecessors measured angles in fractions of a half chord.

In the generation after Euclid, Apollonius of Perga (262–190 BCE) defined a conic section to be the intersection of a right circular cone and a plane, and classified them as hyperbolas, parabolas, and ellipses. This is a three dimensional space definition not possible in Euclid's ruler and compass geometry. There was no apparent practical reason to develop conic sections. Apollonius wrote an astronomical treatise that did not survive, but he seems not to have considered elliptical orbits of planets as a possibility even though the hypothesis of circular uniform velocity orbits for planets was marred by observational anomalies such as retrograde motion. A retrograde motion is one in which to the earthbound observer a planet seems to reverse course for a short time and then proceed forward as before.

Ptolemy of Alexandria (100–170 CE) was the greatest and most influential ancient astronomer. Just as all geometries before Euclid disappeared, so also did all astronomical works before Ptolemy's masterwork, the *Almagest*. This work

develops a three dimensional model of the circular motions of the sun and planets. The orbits of the planets are identified with circular rotations of constant velocities around axes through the center of the earth. He shows how to use epicycles to model the anomaly of retrograde motion, but expresses no confidence that they represent anything physical.

The Almagest is a deductive mathematical text. Some form of three dimensional "trigonometry" is indispensable in astronomical computation. Euclid in his *Phaenomena* used his geometry to pass from equatorial coordinates to ecliptic coordinates. Ptolemy gives strict Euclidean ruler and compass constructions for changes of coordinate systems such as this. He computes tables of chords and proves theorems in three dimensional geometry by a myriad of rotation and ruler and compass constructions. He cites Euclid, Apollonius, Hipparchus of Nicea (190–120 BCE), and Menelaus of Alexandria (70–140 CE) as his predecessors and gives them full credit for their work. He makes direct comparison of actual and predicted observations.

To a modern eye, using ruler and compass proofs may seem incongruous. But Babylonian sexagesimal arithmetic happened to provide the algorithms for his table of chords and for his many astronomical calculations. Ptolemy used chords of a regular polygon to approximate arclength of the corresponding arc on the circle in the manner of Euclid and Archimedes (287–212 BCE). He bounded errors made by this approximation by introducing both the inscribed and circumscribed regular $n$-gons, with arclength caught in between. The numerical evaluations for these constructions only require rational operations and extraction of positive square roots.

Someone may know who first put Euclid's ratios on the line as real numbers or who defined angle by radian measure. We do not. Perhaps Aryabhatta (476–550 CE) in India. Notations for trigonometric functions of arclength, $\sin x, \cos x, \tan x, \sec x,$ were used by Abu al-Wafa' Buzjani (940–998 CE). The *Almagest* the latter wrote has survived and was widely disseminated in the Arab world. He also used negative numbers. He compiled tables of sines, cosines, and tangents.

Astronomy based on chords of the circle and sexagesimal arithmetic survived through Copernicus (1473–1543). His *De revolutionibus orbium coelestium* is, like Ptolemy, based on circular orbits. He just changed the center of the coordinate systems from the earth to the sun, and recomputed motions in that coordinate system. Our use of the term "coordinate system" here is an anachronism. Descartes (1596–1650) had not yet introduced coordinate geometry. Arclength of curves, generally the rectification of curves by finding formulas for the length of segments of curves, became a serious object of study only after the introduction of the differential triangle of Barrow (1630–1677).

Here is an intriguing question. In high school you are taught that a circle in a plane is the set of all points $P$ on that plane equidistant from a fixed point $C$ in that plane. This formal definition of circle goes back at least to Pythagoras. Similarly, an ellipse in a plane is defined as the set of all points $P$ on that plane such that the sum of the distances of $P$ to two fixed points $F_1, F_2$ in that plane is

constant. These two points are called its foci. Many ancient philosophers felt that circles were perfect while other closed curves such as ellipses were not. Why did the circle occupy a special place in Greek geometry when a device as simple as a string with a peg at each end can be used to construct an ellipse? Possibly the circle was so regarded because all fixed stars in the night sky seem to rotate in circles around the earth. Possibly because using a model based on circles, and using the gnomon and the astrolabe as measuring instruments, they could construct tables and tools for navigation on land and sea. Possibly because the measurements they could make could not distinguish between a circular and an elliptical orbit. The deviation (eccentricity) of the actual orbits of the planets from perfect circular orbits is quite small.

Here is another intriguing question. Why did Euclid's *Elements* survive transmission through many cultures, republished almost as many times as the Bible? Was it because of a respect for mathematics or for learning? Perhaps not. The *Almagest* survived because it is a guide to how to do useful astronomy. But the *Almagest* is probably not comprehensible without reading Euclid first. Is it possible that Euclid was transmitted for the most part as a "pony" for Ptolemy?

Galileo (1564–1642) and Huygens (1629–1695) reintroduced Archimedes' methodology that physics is based on experiments. Galileo's *Two New Sciences* defined, with precision but without calculus, the concepts of velocity and acceleration and resolution of forces into vertical and horizontal components. He deduced that a fired cannon ball follows a parabolic path using a characterisation of parabolas due to Apollonius. Tycho Brahe (1546–1601) made observations of orbits precise enough so that Johannes Kepler (1571–1630) could formulate his three laws for planetary motion.

Newton (1642–1717) deduced the inverse square law of universal gravitational attraction between two bodies from these three laws. The role of the ellipse in Newton's masterwork *Principia Mathematica*, which replaced the role that the circular functions played in the Ptolemaic-Copernican theory of the motion of celestial bodies, motivated mathematicians to introduce elliptical analogues of the chords and trigonometric functions. The crucial astronomical need was to compute the deviations from perfect elliptical orbits which are due to the gravitational forces of all the other planets and the sun. The last part of Newton's master work was devoted to precisely that problem for three bodies such as the sun, moon, and earth. It is valuable to have explicit and computationaly feasible formulas about elliptic arclength for use in astronomy. Expanding the integral for arclength of a segment of an ellipse in a power series or a quotient of infinite products is not enough. No way was found to express this arclength integral in terms of the standard rational, exponential, and trigonometric functions. That is, the integration rules of freshman calculus do not give you a formula for arclength of a segment of an ellipse. This is no different from the situation with the ordinary trigonometric functions. The sine and cosine were introduced precisely because these circular functions are periodic, while rational functions, logarithms, and exponentials are not periodic and no composition of them even allowing inverses is periodic either. We also had to introduce logarithms and exponentials because the integrals defining them either

decrease or increase more rapidly than rational and trigonometric functions and their compositions, even allowing inverses. That is why these useful functions have all entered our basic toolbox. So why is it that we cannot express elliptic arclength in terms of rational functions, circular functions, logarithms and exponentials, their inverses, and compositions of these functions? We will answer that shortly.

## Interlude: The Circular Functions

We will describe the way that arclength and the corresponding trigonometry on the ellipse were developed by presenting a fable, an "alternate world" history of the sin function. Suppose, like Ptolemy, that we had never heard of the trigonometric functions. Also suppose that algebra and calculus had both already been invented along with the formula for arclength of a curve $y = g(t)$ from 0 to $x$:

$$\int_0^x \sqrt{1 + (g')^2}\, dt.$$

Then the arc length from (0, 1) to $(x, y)$ on the unit circle $x^2 + y^2 = 1$ is the integral

$$f(x) = \int_0^x \frac{dt}{\sqrt{1 - t^2}}.$$

Then $f$ maps $[-1, 1]$ onto an interval $[-a, a]$, and we define $\pi = 2a$ to be the circumference of half a circle. We then *define* $\sin x$ on $[-\pi/2, \pi/2]$ to be the inverse $f^{-1}$.

Ptolemy constructed his tables using Babylonian sexagesimal arithmetic plus the ruler and compass equivalents of trigonometric addition formulas, one of which is

$$\sin(\theta + \eta) = \sin\theta \cos\eta + \sin\eta \cos\theta. \tag{16.1}$$

Let us show how to derive this addition formula in our alternate world. Let $z \in [-1, 1]$, and suppose that we define $y$ to be a function of $x$, so that $f(x) + f(y) = f(z)$. That is,

$$\int_0^x \frac{dt}{\sqrt{1 - t^2}} + \int_0^{y(x)} \frac{dt}{\sqrt{1 - t^2}} = \int_0^z \frac{dt}{\sqrt{1 - t^2}}. \tag{16.2}$$

Differentiating with respect to the variable $x$, we get

$$\frac{1}{\sqrt{1 - x^2}} + y' \frac{1}{\sqrt{1 - y^2}} = 0. \tag{16.3}$$

Let $h(x) = x\sqrt{1-y^2} + y\sqrt{1-x^2}$. Then

$$h'(x) = \sqrt{1-y^2} - y'\frac{xy}{\sqrt{1-y^2}} - \frac{xy}{\sqrt{1-x^2}} + y'\sqrt{1-x^2}.$$

The second and third together are the left hand side of Eq. (16.3), multiplied by $-xy$; the first and the fourth together are the left hand side of Eq. (16.3) multiplied by $\sqrt{1-x^2}\sqrt{1-y^2}$. It follows that $h'(x) = 0$. We conclude that $h(x)$ is a constant $c$. Substituting $x = 0$ we get $c = y(0)$. However from Eq. (16.2) and the injectivity of $f$ on $[-1, 1]$ we get $y(0) = z$. That is, we conclude that

$$f(x) + f(y) = f\left(x\sqrt{1-y^2} + y\sqrt{1-x^2}\right).$$

Inverting, and using $\cos\theta = \sqrt{1-\sin^2\theta}$, we get the addition formula for the sine, Eq. (16.1). The algebraic relation $z = x\sqrt{1-y^2} + y\sqrt{1-x^2}$ is built up from the rational operations and positive square roots, and so by Descartes can be executed as a ruler and compass construction. So there is a ruler and compass construction for adding two arcs of the circle, allowing Ptolemy of our fable to build tables for sine and cosine.

## Inverting Elliptic Integrals

This fable gives the flavor of what actually happened when attention turned to the ellipse. A remarkable algebraic relation between upper limits of integrals for the arclength of the lemniscate $r^2 = \cos 2\theta$ was discovered by Count Fagnano in 1714. He used the integral formula for arc length to prove that, like for the circle in Euclid, the circumference of the lemniscate could be cut by ruler and compass into $2^m$, $3\times 2^m$, and $5\times 2^m$ equal parts. When Fagnano's result was finally published in 1750, Euler (1707–1783) immediately investigated the general equation relating *any* two arcs in that curve. Unlike the circle,where the upper limits are quadratically related, for the lemniscate the algebraic relation between the upper limits is given by a *fourth* degree algebraic relation.

Cardano (1501–1576) solved fourth degree equations using both quadratic *and* cubic radicals and so goes beyond ruler and compass constructions. This example fits into Abel's later investigation of those division problems for elliptic integrals which can be solved by radicals.

Euler established several such algebraic relations for other curves. Arclength for the lemniscate and ellipse can both be expressed as integrals of the form $\phi(x) = \int_c^x R(x, \sqrt{P(x)})dx$, where $R$ is a rational function and $P$ is a polynomial of degree three or four with distinct roots. This more general class of integrals became named the elliptic integrals.

Following Euler, Legendre (1752–1833) spent much of his career investigating elliptic integrals regarded as a function of their upper limits. Legendre entered the scene with his *Memoire sur les integrations par d'arcs d'ellipse* (1788), establishing a theorem on divisions of the ellipse identical with that for the lemniscate. His *Memoire sur les Transcendantes elliptiques* (1792) introduced the definition of elliptic integrals we still use.

Legendre's book includes applications to the rotation of solids, the motion of a body attracted to two fixed bodies, the attraction of a homogeneous ellipsoid, motions under central forces, surface area of ellipsoids. He succeeded in giving very complete tables of values of his elliptic trigonometric functions for use in astronomy and applied mathematics. Legendre computed tables of values of elliptic integrals, solved their differential equations, and investigated those parts of mechanics in which such integrals arise. Legendre, like his predecessors, considered elliptic integrals solely as real valued functions of a real variable.

Complex numbers had been used formally by the Italian algebraists such as Tartaglia (1499–1557) and Cardano for solving second, third, and fourth degree algebraic equations. Euler used formal expansions in complex power series to "prove" $e^{i\theta} = \cos\theta + i\sin\theta$. Then, a little before 1800, an interpretation of complex numbers as points on the Euclidean plane with real and imaginary axis was offered by Wessel (1745–1818) and Argand (1768–1822). The theory of functions of a complex variable was then developed over many years primarily by Cauchy (1789–1857).

But Legendre failed to discover the role of complex numbers in elliptic integrals. He was both surprised and gratified when around 1826 the young mathematicians Abel (1802–1829) and Jacobi (1804–1851) first inverted the elliptic integrals (mirroring the passage from arcsin to sin in the fictional development above) and then extended the resulting *elliptic functions* to the complex numbers. Neither Abel's nor Jacobi's treatment of this extension bears rigorous scrutiny. Rather, our admiration for this work is based on their plowing ahead anyway based on intuition and the algebra of infinite series and products in the tradition of Euler.

In his *Disquisitiones Mathematicae* (1801) Gauss (1777–1855) constructed the first new regular $n$-gon constructible by ruler and compass since Euclid. This was the 17-gon. More generally he proved that a regular $n$-gon is constructible by ruler and compass if and only if $n$ is a product of a power of 2 and prime factors of the form $2^{2^n} + 1$. Complex numbers and Euler's formula for roots of unity are fundamental to his proof.

Gauss added a suggestive remark about his proof methods:

"Not only can they be applied to the theory of circular functions, but also many other transcendental functions, e.g., those which depend on the integral $\int \frac{dt}{\sqrt{1-t^4}}$."

In his unpublished notebooks he developed the necessary formal theory of functions of a complex variable.

In 1826 Abel visited Paris and heard about Cauchy's theory of functions of a complex variable. He followed this hint of Gauss about division of the arclength of the lemniscate $\int_0^y \frac{dx}{\sqrt{1-x^4}}$. In *Recherches sur les fonctions elliptique* (1827)

he completed Count Fagnano's investigation and proved that the lemniscate can be divided into $n$ equal parts by ruler and compass for precisely the same $n$ as determined by Gauss for the circle.

The most fundamental discovery by Abel and Jacobi was that, just as the ordinary trigonometric functions are periodic functions over the complex numbers with all periods real multiples of a fixed real period, the elliptic functions are doubly periodic functions over the complex numbers with two complex periods with a non-real ratio, such that every period is an integral linear combination of these two periods. This is the reason that the elliptic integral could not be expressed using compositions of rational operations, roots, logs, exponentials, sin, cos, etc. These functions do not have two independent complex periods.

Jacobi (1804–1851) had a much longer career than Abel, and went deeply into applied mathematics. He demonstrated that equations of motion are integrable for the pendulum and for planetary motion in a central gravitational field using elliptic functions. He developed the theory of theta functions. He used methods evolved from elliptic functions to prove many new theorems in number theory, such as the number of representations of an integer as the sum of four squares. His book *Fundamenta nova theoriae functionum ellipticarum* (1829) established elliptic function theory and its generalizations as a principal subject of study for the rest of the nineteenth century.

The blurry issue left open by both of these brilliant mathematicians was that the functions being inverted are two valued due to the square root in the denominator of the integrand. The foundations of elliptic function theory were made firm by the introduction of path integration, Riemann surfaces, and analytic continuation by Cauchy (1789–1857), Weierstrass (1815–1897), Riemann (1826–1866), and Puiseux (1820–1883). In analytic continuation, complex power series convergent in the interior of a circle of convergence were pasted together if they coincided as functions on a common sub-circle so that double valued square roots in the complex numbers became single valued on the resulting pasted-together Riemann surface. Some would say that this construction of a Riemann surface was only made completely rigorous by Hermann Weyl (1885–1955) in his 1913 book *Die Idee der Riemannschen Fläche* [Wey97]. It is well worth reading to this day.

**Further Reading**

For a brief history of both elliptic functions and curves, see [RB12]. An account of the rise of complex analysis in the nineteenth century, including elliptic functions, is given in [BG13]. For a short account of ancient astronomy and the remarkable Antikythera astronomical calculator, see [Jon17].

# Bibliography


[Art91] Artin, M. (1991). *Algebra*. Prentice Hall, Inc.

[BG13] Bottazzini, U., & Gray, J. (2013). *Hidden harmony—geometric fantasies*. Sources and Studies in the History of Mathematics and Physical Sciences. Springer, The rise of complex function theory.

[Bix06] Bix, R. (2006). *Conics and cubics* (2nd ed.). Undergraduate Texts in Mathematics. Springer. A concrete introduction to algebraic curves.

[BK86] Brieskorn, E., & Knörrer, H. (1986). *Plane algebraic curves*. Birkhäuser Verlag. Translated from the German by John Stillwell.

[Ces58] Cesari, L. (1958). Rectifiable curves and the Weierstrass integral. *The American Mathematical Monthly*, *65*, 485–500.

[CGC89] Cucker, F., & Gonzalez Corbalan, A. (1989). An alternate proof of the continuity of the roots of a polynomial. *The American Mathematical Monthly*, *96*(4), 342–345.

[Die69] Dieudonné, J. (1969). *Foundations of modern analysis*. Pure and Applied Mathematics (Vol. 10-I). Academic Press. Enlarged and corrected printing.

[Fis01] Fischer, G. (2001). *Plane algebraic curves*, volume 15 of Student Mathematical Library. American Mathematical Society. Translated from the 1994 German original by Leslie Kay.

[Ful69] Fulton, W. (1969). *Algebraic curves. An introduction to algebraic geometry*. W. A. Benjamin, Inc. Notes written with the collaboration of Richard Weiss, Mathematics Lecture Notes Series.

[Gib98] Gibson, C. G. (1998). *Elementary geometry of algebraic curves: an undergraduate introduction*. Cambridge University Press.

[Gri89] Griffiths, P. A. (1989). *Introduction to algebraic curves*, volume 76 of Translations of mathematical monographs. American Mathematical Society. Translated from the Chinese by Kuniko Weltin.

[Gun90] Gunning, R. C. (1990). *Introduction to holomorphic functions of several variables. Vol. III*. The Wadsworth & Brooks/Cole Mathematics Series. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA. Homological theory.

[Jon17] Jones, A. (2017). *A portable cosmos: revealing the Antikythera mechanism, scientific wonder of the ancient world*. Oxford University Press.

[Ken11] Kendig, K. (2011). *A guide to plane algebraic curves*, volume 46 of The Dolciani mathematical expositions. Mathematical Association of America. MAA Guides, 7.

[Kir92] Kirwan, F. (1992). *Complex algebraic curves*, volume 23 of London mathematical society student texts. Cambridge University Press.

[KJK+06] Kurja, R. V., Joshi, K., Mohan Kumar, N., Raranape, K. H., Ramanathan, A., Shorey, T. N., Simha, R. R., & Srinivas, V. (2006). *Elliptic curves*. Narosa Pub House.

[Kna92] Knapp, A. W. (1992). *Elliptic curves*, volume 40 of Mathematical notes. Princeton University Press.




A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2

[Kun05] Kunz, E. (2005). *Introduction to plane algebraic curves*. Birkhäuser Boston, Inc. Translated from the 1991 German edition by Richard G. Belshoff.

[Lan87] Lang, S. (1987). *Elliptic functions* (2nd ed.), volume 112 of Graduate texts in mathematics. Springer-Verlag. With an appendix by J. Tate.

[LR85] Lange, H., & Ruppert, W. (1985). Complete systems of addition laws on abelian varieties. *Inventiones Mathematicae*, *79*(3), 603–610.

[Mir95] Miranda, R. (1995). *Algebraic curves and Riemann surfaces*, volume 5 of Graduate studies in mathematics. American Mathematical Society.

[Mun91] Munkres, J. R. (1991). *Analysis on manifolds*. Addison-Wesley Publishing Company, Advanced Book Program.

[PC] Ponce Campuzano, J. C. Complex analysis: Problems with solutions. Available at https://faculty.ksu.edu.sa/sites/default/files/2016_complex_analysis_problems_solutions.pdf

[RB12] Rice, A., & Brown, E. (2012). Why ellipses are not elliptic curves. *Mathematics Magazine*, *85*(3), 163–176.

[Rem91] Remmert, R. (1991). *Theory of complex functions*, volume 122 of Graduate texts in mathematics. Springer-Verlag. Translated from the second German edition by Robert B. Burckel, Readings in Mathematics.

[Rot00] Rotman, J. J. (2000). *A first course in abstract algebra* (2nd ed.). Prentice Hall, Inc.

[Sha92] Shabat, B. V. (1992). *Introduction to complex analysis. Part II*, volume 110 of Translations of mathematical monographs. American Mathematical Society. Functions of several variables, Translated from the third (1985) Russian edition by J. S. Joel.

[Sil09] Silverman, J. H. (2009). *The arithmetic of elliptic curves*, volume 106 of Graduate texts in mathematics (2nd ed.). Springer.

[SK59] Semple, J. G., & Kneebone, G. T. (1959). *Algebraic curves*. Oxford University Press.

[Spi65] Spivak, M. (1965). *Calculus on manifolds. A modern approach to classical theorems of advanced calculus*. W. A. Benjamin, Inc.

[Sti94] Stillwell, J. (1994). *Elements of algebra*. Undergraduate texts in mathematics. Springer-Verlag. Geometry, numbers, equations.

[Str76] Strang, G. (1976). *Linear algebra and its applications* (4th ed.). Brooks/Cole/Cengage, 2006. 1st edition: Academic Press.

[Tuc97] Tucker, T. W. (1997). Rethinking rigor in calculus: the role of the mean value theorem. *The American Mathematical Monthly*, *104*(3), 231–240.

[vdW45] van der Waerden, B. L. (1945). *Einführung in die algebraische Geometrie*. Dover Publications.

[VLA65] Volkovyskiĭ, L. I., Lunts, G. L., & Aramanovich, I. G. (1965). *A collection of problems on complex analysis*. Pergamon Press/Addison-Wesley Publishing Co. Translated from the Russian by J. Berry, Translation edited by T. Kovari.

[Wal50] Walker, R. J. (1950). *Algebraic curves* (vol. 13). Princeton Mathematical Series. Princeton University Press.

[Was08] Washington, L. C. (2008). *Elliptic curves* (2nd ed.). Discrete mathematics and its applications (Boca Raton). Chapman & Hall/CRC. Number theory and cryptography.

[Wey97] Weyl, H. (1997). *Die Idee der Riemannschen Fläche*, volume 5 of *Teubner-Archiv zur Mathematik. Supplement [Teubner Archive on Mathematics. Supplement]*. B. G. Teubner Verlagsgesellschaft mbH, Stuttgart. Reprint of the 1913 German original, With essays by Reinhold Remmert, Michael Schneider, Stefan Hildebrandt, Klaus Hulek and Samuel Patterson, Edited and with a preface and a biography of Weyl by Remmert.

# Index





A. Nerode, N. Greenberg, *Algebraic Curves and Riemann Surfaces for Undergraduates*, https://doi.org/10.1007/978-3-031-11616-2