Chapter 3

Complex Numbers and Geometry

In this chapter we study familiar geometric objects in the plane, such as lines, circles, and conic sections. We develop our intuition and results via complex numbers rather than via pairs of real numbers. By the end of the chapter we will follow Riemann and think of complex numbers as points on a sphere where the north pole is the point at infinity.

1. Lines, circles, and balls

We often describe geometric objects via equations. The algebra helps the geometry and the geometry guides the algebra. Two kinds of equations, *parametric equations* and *defining equations*, provide different perspectives on the geometry. Given a complex-valued function f, the set of points p for which $f(p) = 0$ is called the *zero-set* of f. A defining function for a set S is thus a function whose zero-set is precisely S. A parametric equation for a set S is a function whose image is the set S. We will use parametric equations when we compute complex line integrals in Chapter 6. Both kinds of equations arise frequently.

We first consider the Euclidean plane as $\mathbf{R}^2$, but we quickly change perspective and think of the plane as $\mathbf{C}$. Let L be a line in $\mathbf{R}^2$. We can regard L as the set of points (x, y) satisfying an equation of the form

$$Ax + By + C = 0, \tag{1}$$

where not both A and B are zero. Equation (1) is called a *defining equation* for L. A point (x, y) lies on L if and only if (x, y) satisfies (1). A defining equation is not unique; we could multiply (1) by a nonzero constant, or even a nonzero function, and the set of solutions would not change. It is natural to seek the simplest possible defining equation; for lines the defining equation should be linear. When $B \neq 0$ in (1), we often solve the defining equation for y and say that the equation of the line

is $y = mx + b$; here $m = \frac{-A}{B}$ is the slope of the line. When $B = 0$, we obtain the special case of the line with infinite slope given by x equals a constant.

Alternatively we can describe L via *parametric equations*. Especially in higher dimensions, the parametric approach has many advantages. A line in the plane through the point (x_0, y_0) is determined by its direction vector (u, v); thus L is the set of points $(x_0, y_0)+t(u, v)$ for $t \in \mathbf{R}$. Here the real number t is called a parameter; it is often useful to regard t as *time* and to think of a particle moving along the line. This formulation works in higher dimensions; the parametric equation $\gamma(t) = \mathbf{p}+t\mathbf{v}$ defines a line containing $\mathbf{p}$ and with direction vector $\mathbf{v}$.

We next consider the same issues for circles. A circle with center at (x_0, y_0) and radius R has the defining equation

$$(x - x_0)^2 + (y - y_0)^2 - R^2 = 0. \tag{2}$$

We could also write the circle using parametric equations:

$$(x(t), y(t)) = (x_0 + R\cos(t), y_0 + R\sin(t)). \tag{3}$$

Now the parameter t lives (for example) in the interval $[0, 2\pi)$; if we let t vary over a larger set, then we cover points on the circle more than once. Another possible parametrization of a circle will be derived in Chapter 8. There we show that the unit circle can be described by the parametric equations

$$(x(t), y(t)) = \left(\frac{1 - t^2}{1 + t^2}, \frac{2t}{1 + t^2}\right), \tag{4}$$

where now $-\infty < t < \infty$. We get all points on the unit circle except for $(-1, 0)$, which we realize by allowing t to take the value infinity. Figure 3.1 indicates the geometric meaning of the parameter t.

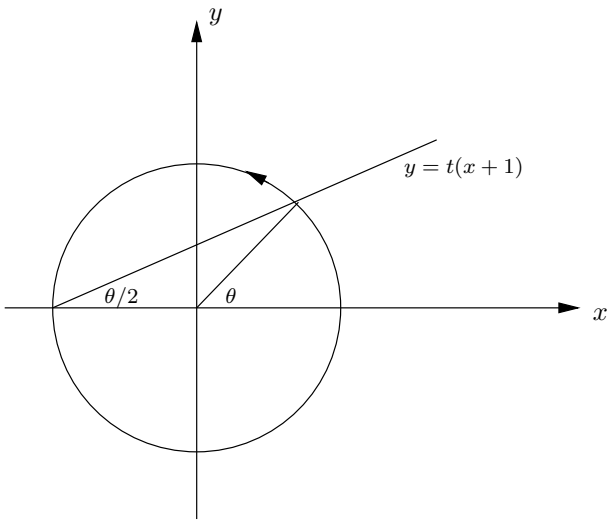

Figure 3.1. Parametrizing the unit circle.

▶ **Exercise 3.1.** Show that (4) parametrizes the unit circle, except for $(-1, 0)$. Show in (4) that $y = t(x + 1)$. What is the geometric meaning of t?

▶ **Exercise 3.2.** If $(x(t), y(t)) = (\cos(\theta), \sin(\theta))$, express the parameter t from (4) in terms of θ. Hint: Look at Figure 3.1. See Chapter 8 for more information.

Things in the plane invariably simplify by using complex variables. A parametric equation for a line in $\mathbf{C}$ is given by $z(t) = z_0 + tv$, where z_0 is a point on the line and v is any nonzero complex number. Note that t is real. The same line has defining equation

$$\mathrm{Re}\,((z - z_0)i\overline{v}) = 0. \tag{5}$$

We can derive (5) with almost no computation. First of all, by inspection, z_0 lies on the solution set to (5). Let us reason geometrically. We know that

$$|\zeta + w|^2 = |\zeta|^2 + |w|^2 + 2\mathrm{Re}(\zeta\overline{w}). \tag{6}$$

Hence, if $\mathrm{Re}(\zeta\overline{w}) = 0$, then (6) says (the converse of the Pythagorean theorem!) that ζ and w are perpendicular. Therefore (5) states that $(z - z_0)$ is perpendicular to the vector $-iv$. Since multiplication by $\pm i$ is a rotation of $\pm 90^{\mathrm{o}}$, we conclude that the vectors $z - z_0$ and v point in the same direction. Thus (5) says both that z_0 is on the line and that v is the direction of the line.

We say a few words about circles and balls. A circle of radius r with center at p is the set of z satisfying $|z - p| = r$. We could also write the circle parametrically as the set of z satisfying $z = p + re^{i\theta}$ for $0 \leq \theta < 2\pi$. The closed ball of radius r about p is given by the set of z satisfying $|z - p| \leq r$ and the open ball $B_r(p)$ is given by the set of z satisfying $|z - p| < r$. Sometimes, we say *disk* instead of ball. The term *unit disk* means $\{z : |z| < 1\}$. Open balls are important because they lead to the more general notion of *open set*.

Definition 1.1. A subset Ω of $\mathbf{C}$ is called *open* if, for each $z \in \Omega$, there is an $r > 0$ such that $B_r(z) \subset \Omega$.

▶ **Exercise 3.3.** Show that the complement of a closed ball is an open set.

▶ **Exercise 3.4.** Show that the collection $\mathcal{F}$ of open subsets of $\mathbf{C}$ satisfies the following properties:

1) The empty set $\varnothing$ and the whole space $\mathbf{C}$ are elements of $\mathcal{F}$.

2) If A, B are elements of $\mathcal{F}$, then so is $A \cap B$.

3) If A_α is any collection of elements of $\mathcal{F}$, then $\bigcup A_\alpha$ is also in $\mathcal{F}$.

We pause to introduce the definition of a topology. Let X be a set, and let $\mathcal{F}$ be a collection of subsets of X. Then $\mathcal{F}$ is called a *topology* on X, and the pair $(X, \mathcal{F})$ is called a *topological space*, if $\mathcal{F}$ satisfies the three properties from the previous exercise. The elements of $\mathcal{F}$ are called *open subsets*. When (X, δ) is a metric space (Section 6 of Chapter 1), we have already given the definition of open set. The collection of open sets in a metric space does satisfy the three properties that make $(X, \mathcal{F})$ into a topological space.

There are many topologies on a typical set X. For example, we could decree that every subset of X is open. At the other extreme we could decree that the only

open subsets of X are the empty set and X itself. The concept of topological space allows one to give the definition and properties of continuous functions solely in terms of the open sets.

We close this section by showing that two specific open subsets of $\mathbf{C}$ can be considered the same from the point of view of complex analysis. The sense in which they are the same is that there is a bijective complex analytic mapping between them. In the next lemma these sets are the open unit ball and the open upper half-plane, defined as the set of z for which $\mathrm{Im}(z) > 0$. See Section 4 of Chapter 8 for these considerations in more generality. While we do not yet wish to develop these ideas, the next lemma also anticipates our study of linear fractional transformations and provides a simple example of conformal mapping.

Lemma 1.1. *Put* $\zeta = i\frac{1-z}{1+z}$. *Then* $|z| < 1$ *if and only if* $\mathrm{Im}(\zeta) > 0$.

Proof. We write $\mathrm{Im}(\zeta)$ as $\frac{\zeta-\overline{\zeta}}{2i}$ and compute:

$$\mathrm{Im}(\zeta) = \frac{\zeta-\overline{\zeta}}{2i} = \frac{1}{2i}\left(i\frac{1-z}{1+z} + i\frac{1-\overline{z}}{1+\overline{z}}\right) = \frac{1}{2}\left(\frac{1-z}{1+z} + \frac{1-\overline{z}}{1+\overline{z}}\right).$$

After clearing denominators, we find that $\mathrm{Im}(\zeta) > 0$ if and only if

$$0 < \frac{1}{2}\big((1-z)(1+\overline{z}) + (1-\overline{z})(1+z)\big) = \frac{1}{2}(2-2z\overline{z}) = 1-|z|^2,$$

that is, if and only if $|z| < 1$. □

The mapping $z \to i\frac{1-z}{1+z}$ is called a linear fractional transformation. Such transformations map the collection of lines and circles in $\mathbf{C}$ to itself. See Section 4.

2. Analytic geometry

We begin by recalling geometric definitions of hyperbolas, ellipses, and parabolas. We also express these objects using defining equations.

Definition 2.1. A *hyperbola* is the set of points in a plane defined by the following condition. Given distinct foci p and q, the hyperbola consists of those points for which the absolute difference in distances to these two foci is a real nonzero constant. A defining equation is

$$|z-p| - |z-q| = \pm c. \tag{7}$$

An *ellipse* is the set of those points for which the sum of the distances to these foci is a positive constant. A defining equation is

$$|z-p| + |z-q| = c. \tag{8}$$

A circle is the set of points in the plane whose distance to a given point is constant. That distance is called the *radius* of the circle. We include the case of a single point as a circle whose radius is 0. A circle of positive radius is the special case of an ellipse when the foci p and q are equal. A defining equation is then $|z-p| = r$.

A parabola is the set of points in a plane that are equidistant from a given point (the focus) and a given line (the directrix). If the focus is p and the line is given by $z_0 + tv$, then we may take

$$(\mathrm{Im}((z - z_0)\overline{v}))^2 = |v|^2|z - p|^2 \tag{9}$$

for a defining equation. See Proposition 3.2. We can simplify the equation (9) slightly, because without loss of generality we may assume that $|v| = 1$.

A variant of the definition of a parabola can also be used to define ellipses and hyperbolas. Given a focus and directrix, one considers the set of points for which the distance to the focus is a constant positive multiple $\mathcal{E}$ of the distance to the directrix. This number $\mathcal{E}$ is called the **eccentricity**. When $\mathcal{E} = 1$, the set is a parabola. When $\mathcal{E} < 1$, the set is an ellipse, and when $\mathcal{E} > 1$, the set is a hyperbola. Most calculus books have lengthy discussions of this matter. See for example [**24**]. We will use complex variables to develop a somewhat different intuition.

We mention also that the defining property (8) of an ellipse helps explain why *whispering galleries* work. In such a gallery, one stands at one focus f_1 and whispers into a wall shaped like an ellipsoid. The sound wave emanating from f_1 reflects off the wall and passes through the other focus f_2. Hence someone located at f_2 can hear the whisper from f_1. This property follows from the law of reflection; imagine placing a mirror tangent to the ellipse at the point p on the ellipse hit by the sound wave. The line segment from f_1 to p and the line segment from p to the reflection of f_2 are part of the same line.

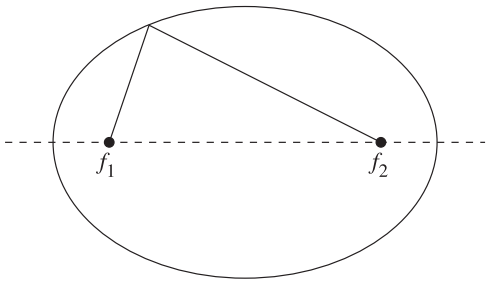

Figure 3.2. The geometric definition of an ellipse.

The definition of a circle allows for a *degenerate* situation consisting of a single point. Degenerate situations for hyperbolas can be more subtle. We pause now to consider such a degenerate situation; Section 3 provides additional examples. Let $H_\pm$ be the set of z such that $|z + 1| - |z - 1| = \pm 2$. Computation (Exercise 3.12) shows that either equation forces z to be real. We can then check that H_+ consists of those real numbers at least 1, and H_- consists of those real numbers at most -1. We can regard these two rays as branches of a degenerate hyperbola.

3. Quadratic polynomials

In this section we develop the relationship between geometry and algebra. For the most part we limit our discussion to quadratic polynomials and the geometry of

their zero-sets. We begin however with a few words about polynomials of arbitrary degree in two real variables.

Let $p(x, y)$ be a polynomial with real coefficients in the two real variables x, y. Thus, there are real coefficients c_{ab} such that

$$p(x, y) = \sum_{a=0}^{m} \sum_{b=0}^{n} c_{ab} x^a y^b. \tag{10}$$

We say that p has degree k if $c_{ab} = 0$ whenever $a + b > k$ and, for some a, b with $a + b = k$, we have $c_{ab} \neq 0$. Let us substitute $x = \frac{z+\overline{z}}{2}$ and $y = \frac{z-\overline{z}}{2i}$ into (10). We obtain a polynomial $\Phi(z, \overline{z})$, defined by

$$\Phi(z, \overline{z}) = \sum_{a=0}^{m} \sum_{b=0}^{n} c_{ab} (\frac{z + \overline{z}}{2})^a (\frac{z - \overline{z}}{2i})^b = \sum d_{ab} z^a \overline{z}^b. \tag{11}$$

The coefficients d_{ab} are not in general real, but the values of $\Phi(z, \overline{z})$ are real. Equating coefficients in $\Phi = \overline{\Phi}$ shows, for all indices a, b, that the Hermitian symmetry condition $d_{ab} = \overline{d_{ba}}$ holds. Conversely suppose we are given a polynomial of the form $\Phi(z, \overline{z}) = \sum d_{ab} z^a \overline{z}^b$ and the Hermitian symmetry condition is satisfied. Then $\Phi(z, \overline{z})$ is real for all z. The following definition and proposition clarify the issues.

Definition 3.1. Let $\Phi(z, \overline{w})$ be a polynomial in the two complex variables z and $\overline{w}$. We say that Φ is Hermitian symmetric if for all z and w we have

$$\Phi(z, \overline{w}) = \overline{\Phi(w, \overline{z})}.$$

Proposition 3.1. *Let Φ be the polynomial in two complex variables defined by*

$$\Phi(z, \overline{w}) = \sum d_{ab} z^a \overline{w}^b.$$

The following statements are equivalent:

- *Φ is Hermitian symmetric.*
- *For all a, b we have $d_{ab} = \overline{d_{ba}}$.*
- *For all z, $\Phi(z, \overline{z})$ is real.*

Proof. This simple proof is left to the reader. □

To each real polynomial in x, y there corresponds a unique Hermitian symmetric polynomial in $z, \overline{w}$, and conversely each Hermitian symmetric polynomial in $z, \overline{w}$ defines a real polynomial after setting $\overline{w}$ equal to $\overline{z}$. It therefore makes little difference whether we study real polynomials or Hermitian symmetric polynomials. The Hermitian symmetric perspective seems easier to understand.

▶ **Exercise 3.5.** Prove Proposition 3.1.

▶ **Exercise 3.6.** Prove that the correspondence going from (10) to (11) between real and Hermitian symmetric polynomials preserves degree.

Our discussion of analytic geometry leads to quadratic polynomials, namely those of (total) degree two. We must distinguish the total degree from the degree in z alone. For example, the Hermitian symmetric polynomial $|z|^2 = z\overline{z}$ is quadratic; its degree is two, but it is of degree one in the z variable alone.

For hyperbolas and ellipses we start with the defining relations $|z-p| = c \pm |z-q|$ and square both sides. Then we isolate the term involving $|z-q|$ and square again. After simplifying, only terms of degree at most two remain. For a parabola, we must use the formula (9), which follows from the following result:

Proposition 3.2. *Let L be the line in $\mathbf{C}$ defined by $z(t) = z_0 + tv$. The (minimum) distance δ from a point z to this line L is given by*

$$\delta = |\frac{1}{\overline{v}} \mathrm{Im}((z-z_0)\overline{v})| = |\frac{(z-z_0)\overline{v} - \overline{(z-z_0)}v}{2\overline{v}}|. \tag{12}$$

Proof. Consider the squared distance $f(t)$ from z to a point on L. Thus $f(t) = |z - z_0 - tv|^2$. We expand and find the minimum of the quadratic polynomial f by calculus. We obtain the equations

$$f(t) = |z-z_0|^2 - 2t\mathrm{Re}((z-z_0)\overline{v}) + t^2|v|^2,$$
$$f'(t) = -2\mathrm{Re}((z-z_0)\overline{v}) + 2t|v|^2.$$

Therefore the minimum occurs when

$$t = \frac{\mathrm{Re}((z-z_0)\overline{v})}{|v|^2}.$$

Plugging in this value for t in f gives the minimum squared distance, namely

$$\delta^2 = \left|(z-z_0) - \frac{\mathrm{Re}((z-z_0)\overline{v})}{\overline{v}}\right|^2 = \left|\frac{2(z-z_0)\overline{v} - \overline{(z-z_0)}v - (z-z_0)\overline{v}}{2\overline{v}}\right|^2$$
$$= \left|\frac{1}{\overline{v}}\mathrm{Im}((z-z_0)\overline{v})\right|^2.$$

□

▶ **Exercise 3.7.** Let L be the line in $\mathbf{C}$ defined by $z(t) = z_0 + tv$. Verify (9) by using Proposition 3.2.

▶ **Exercise 3.8.** Find, in terms of x and y, the equation of a parabola with focus at $(3, 0)$ and directrix the line $x = 1$.

▶ **Exercise 3.9.** Find, in terms of x and y, the equation of any hyperbola with foci at $\pm 3i$.

▶ **Exercise 3.10.** What object is defined by the condition that the eccentricity is 0? What object is defined by the condition that the eccentricity is infinite?

Viewing these objects via eccentricity enables us to conceive of hyperbolas, parabolas, and ellipses in similar fashions. These objects, as well as points, lines, and pairs of lines, are zero-sets for Hermitian symmetric quadratic polynomials.

Proposition 3.3. *Let $z_0 + tv$ define a line in $\mathbf{C}$ and let $p \in \mathbf{C}$. Define a family of Hermitian symmetric polynomials $\Psi_{\mathcal{E}}$ by the formula*

$$\Psi_{\mathcal{E}} = |z-p|^2 - \mathcal{E}^2 |\frac{(z-z_0)\overline{v} - \overline{(z-z_0)}v}{2\overline{v}}|^2.$$

Then the zero-set of $\Psi_{\mathcal{E}}$ is an ellipse for $0 < \mathcal{E} < 1$, a parabola for $\mathcal{E} = 1$, and a hyperbola for $\mathcal{E} > 1$.

Proof. The conclusion follows from (12) and the characterizations of the objects using eccentricities. Alternatively, one can derive the statement from the subsequent discussion by computing the determinant Δ. □

Consider the most general Hermitian symmetric quadratic polynomial:

$$\Phi(z,\overline{z}) = \alpha z^2 + \overline{\alpha z}^2 + \beta z + \overline{\beta}\overline{z} + \gamma z\overline{z} + F = 0. \tag{13}$$

In (13), α and β are complex, whereas γ and F are real. We will analyze the zero-sets of such polynomials, thereby providing a complex variables approach to conic sections. The analysis requires many cases; we first assume $\alpha = 0$ in (13). The zero-set $\mathbf{V}$ of Φ must then be one of the following objects: the empty set, a line, a point, a circle, all of $\mathbf{C}$. This situation arises again in Theorem 4.1 when we study linear fractional transformations.

- Assume $\alpha = \beta = \gamma = 0$ in (13). The equation $\Phi = 0$ becomes $F = 0$. Thus $\mathbf{V}$ is empty if $F \neq 0$ and $\mathbf{V}$ is all of $\mathbf{C}$ if $F = 0$.
- Assume $\alpha = \gamma = 0$ in (13) but $\beta \neq 0$. The equation $\Phi = 0$ becomes

  $$2\mathrm{Re}(z\beta) + F = 0.$$

  Hence in this case $\mathbf{V}$ is a line.
- Assume $\alpha = 0$ and $\gamma \neq 0$. We proceed analogously to the proof of the quadratic formula. We complete the square in the equation $\Phi = 0$ to get

  $$|z + \frac{\overline{\beta}}{\gamma}|^2 = \frac{|\beta|^2 - F\gamma}{\gamma^2}. \tag{14}$$

  Hence $\mathbf{V}$ is either empty, a point, or a circle; the answer depends on whether the right-hand side of (14) is negative, zero, or positive.

It remains to discuss the case where $\alpha \neq 0$, in which case the defining equation must be quadratic. The expression $\Delta = \gamma^2 - 4|\alpha|^2$ governs the geometry. When $\Delta < 0$, we get a (possibly degenerate) hyperbola; when $\Delta = 0$, we get a (possibly degenerate) parabola; when $\Delta > 0$, we get a (possibly degenerate) ellipse. Below we will interpret Δ from the point of view of Hermitian symmetric polynomials.

We first give some examples of degenerate situations arising when $\alpha \neq 0$.

- The equation $\mathrm{Re}(z^2) = 0$, where $\Delta = -1$, defines two lines rather than a hyperbola.
- The equation $(z+\overline{z})^2 = 0$, where $\Delta = 0$, defines a line rather than a parabola.
- The equation $z^2 + \overline{z}^2 + 2|z|^2 - 2(z + \overline{z}) = 0$, where $\Delta = 0$, defines two lines rather than a parabola.
- The equation $|z|^2 = 0$, where $\Delta = 1$, defines a point rather than an ellipse.

Notice that two lines can be a degenerate version of either a parabola or a hyperbola. Another interesting point is that sometimes one line should be regarded as one line, but other times it should be regarded as two lines! The linear equation $x = 0$ defines the line $x = 0$ once, and the zero-set should be regarded as one line. On the other hand, the quadratic equation $x^2 = 0$ defines the single line $x = 0$ twice, and the zero-set should be regarded as two lines.

To determine what kind of an object (13) defines can be a nuisance because of degenerate cases. In the very degenerate case where no quadratic terms are present ($\alpha = \gamma = 0$) the linear terms determine whether the object is a line, the empty set, or all of $\mathbf{C}$. When the quadratic part is not identically zero, the linear terms usually amount to translations and do not matter; the exception comes when the quadratic part itself is degenerate and the linear terms determine whether the object is a parabola. For example compare the equations $x^2 - 1 = 0$ and $x^2 - y = 0$. The first defines two lines while the second defines a parabola.

We consider the pure quadratic case from the complex variable point of view. Thus we let

$$\Phi(z, \overline{z}) = \alpha z^2 + \overline{\alpha z}^2 + \gamma z\overline{z}. \tag{15}$$

For a positive constant c the set $\Phi = c$ defines an ellipse whenever $\Phi(z, \overline{z}) > 0$ for $z \neq 0$. After dividing by $|z|^2$ and introducing polar coordinates, we find that the condition becomes

$$\alpha e^{2i\theta} + \overline{\alpha} e^{-2i\theta} + \gamma > 0. \tag{16}$$

The minimum value of the left-hand side of (16) occurs when $\alpha e^{2i\theta} = -|\alpha|$. Hence the condition for being an ellipse is that $\gamma - 2|\alpha| > 0$.

One can obtain this inequality by other methods. We can write (15) in terms of x, y as

$$P(x, y) = (\gamma + 2\mathrm{Re}(\alpha))x^2 - 4\mathrm{Im}(\alpha)xy + (\gamma - 2\mathrm{Re}(\alpha))y^2,$$

or in terms of matrices as

$$M = \begin{pmatrix} \gamma + 2\mathrm{Re}(\alpha) & -2\mathrm{Im}(\alpha) \\ -2\mathrm{Im}(\alpha) & \gamma - 2\mathrm{Re}(\alpha) \end{pmatrix}.$$

The polynomial P and the matrix M are equivalent ways of defining a quadratic form. The behavior of this quadratic form is governed by the eigenvalues of M. Their product is the determinant Δ, given in (17), and their sum is the trace 2γ.

$$\Delta = \gamma^2 - 4(\mathrm{Re}(\alpha))^2 - 4(\mathrm{Re}(\beta))^2 = \gamma^2 - 4|\alpha|^2. \tag{17}$$

Things degenerate when $\Delta = 0$. When $\Delta > 0$, we also must have $\gamma > \pm 2\mathrm{Re}(\alpha)$, and both eigenvalues are positive. We then obtain an ellipse for the set $\Phi = c$, for $c > 0$. When $\Delta < 0$, we obtain a hyperbola, where we allow the possibility of two crossing lines.

3.1. The situation using real variables. For comparative purposes we recall how this discussion from elementary analytic geometry proceeds when we stay within the realm of real variables. Consider a polynomial P of degree at most two with real coefficients in the variables x and y. Thus there are real numbers A, B, C, D, E, F such that

$$P(x, y) = Ax^2 + 2Bxy + Cy^2 + Dx + Ey + F. \tag{18}$$

The set of points (x, y) for which $P(x, y) = 0$ (called the zero-set of P) must be one of the following geometric objects: the empty set, all of $\mathbf{R}^2$, a point, a line, two lines, a circle, a parabola, a hyperbola, or an ellipse. We may regard a circle as a special case of an ellipse. The reader should be able to solve the following exercise.

▶ **Exercise 3.11.** For each of the geometric objects in the above paragraph, give values of the constants A, B, C, D, E, F such that the zero-set of P is that object. Show that all the above objects, except for the entire plane, are zero-sets of polynomials of degree two. Thus even the *degenerate* cases of empty set, point, line, and two lines are possible for the zero-sets of quadratic polynomials in two real variables.

It is possible to completely analyze the possibilities for the zero-set of P; with the proper background this analysis is concise. Without that background things seem messy. We recall the answer.

Again the difficulty in the analysis arises from the many degenerate cases. If $P(x, y) = F$ is a constant, then the zero-set is either everything or nothing (the empty set), according to whether $F = 0$ or not. If P is of degree one, that is, $P(x, y) = Dx + Ey + F$ and at least one of D and E is not zero, then the zero-set is a line. If P is of degree two, then things depend on the expression $AC - B^2$. If $AC - B^2 < 0$, then the object is a hyperbola, with two lines a possibility. If $AC - B^2 > 0$ and $A > 0$, then the object is either empty, a single point, a circle, or an ellipse. We may regard a single point or a circle as a kind of ellipse. We next analyze the possibilities when $AC - B^2 = 0$, still assuming that P has degree two. Note then that at least one of A and C is nonzero, and they cannot have opposite signs. After multiplying P by -1, we may assume that A and C are nonnegative. We eliminate B and write

$$Ax^2 + 2\sqrt{AC}xy + Cy^2 + Dx + Ey + F = (\sqrt{A}x + \sqrt{C}y)^2 + Dx + Ey + F.$$

Analyzing the possible zero-sets is amusing. If $D = E = 0$, then we get

$$(\sqrt{A}x + \sqrt{C}y)^2 + F = 0,$$

which gives the empty set if $F > 0$, a single line if $F = 0$, and a pair of lines if $F < 0$. If at least one of A and D is nonzero, then we write $u = \sqrt{A}x + \sqrt{C}y$ and $v = Dx + Ey$. First assume that v is not a constant multiple of u. Then the equation $P = 0$ becomes $u^2 + v + F = 0$, which defines a parabola. If v is a multiple of u, then the resulting object becomes two lines, one line, or the empty set.

3.2. Back to complex variables. To complete our discussion, we recall how to rewrite everything in complex notation. We use the formulas for x and y in terms of z and $\overline{z}$ to obtain

$$0 = P(x, y) \tag{19}$$
$$= A(\frac{z+\overline{z}}{2})^2 + 2B(\frac{z+\overline{z}}{2})(\frac{z-\overline{z}}{2i}) + C(\frac{z-\overline{z}}{2i})^2 + D(\frac{z+\overline{z}}{2}) + E(\frac{z-\overline{z}}{2i}) + F.$$

Simplifying (19) gives, for complex numbers α, β, real number γ, and the same real number F, our familiar formula (13).

▶ **Exercise 3.12.** What kind of object does each of the following equations define?

- $iz^2 - i\overline{z}^2 = 4$.
- $|z|^2 + z^2 + \overline{z}^2 = 3$.
- $|z - 1| + |z - 3| = 2$. Also, write this equation in the form (13).
- $\alpha z^2 + \overline{\alpha z}^2 + |z|^2 = 1$. The answer depends on α.

- $z^2 + \overline{z}^2 = 0$.
- $|z+1| - |z-1| = \pm 2$. (Be careful!)

Complex analysis offers much geometric information. Figure 3.3 shows that the level sets of the real and imaginary parts of z^2 form orthogonal (perpendicular) hyperbolas; more generally we shall see that the level sets of the real and imaginary parts of a complex analytic function form orthogonal trajectories. This fact lies at the foundation of various applications to physics and engineering. We close this section by posing some related exercises.

▶ **Exercise 3.13.** Let $f(z) = z^2$. What are the real and imaginary parts of f in terms of x, y? Graph their level sets; show that one gets orthogonal hyperbolas.

▶ **Exercise 3.14.** Let $f(z) = z^3$. What are the real and imaginary parts of f in terms of x, y? Can you prove that the corresponding level sets are orthogonal?

▶ **Exercise 3.15.** What are the real and imaginary parts of $\frac{1}{z}$? (Assume $z \neq 0$.) Show by computation that their level sets form orthogonal trajectories. Do the same for $\frac{1}{z^n}$.

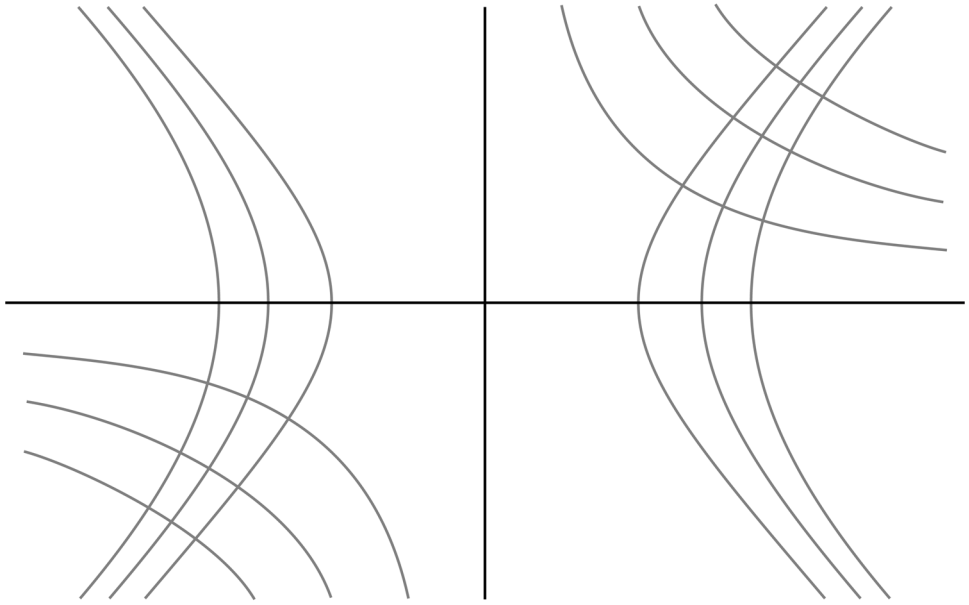

Figure 3.3. Orthogonal hyperbolas.

The close relationship between the trigonometric functions and the unit circle has been useful for us. The next exercise reveals a similar relationship between the hyperbolic functions and the hyperbola $x^2 - y^2 = 1$.

▶ **Exercise 3.16.** Show that

$$\cosh^2(z) - \sinh^2(z) = 1.$$

Find parametric equations for (a branch of) the hyperbola $x^2 - y^2 = 1$.

► **Exercise 3.17.** Find formulas for the multi-valued functions $\cosh^{-1}$ and $\sinh^{-1}$ using logarithms. Do the same for $\tanh^{-1}$.

4. Linear fractional transformations

This section glimpses the interesting and important subject of conformal mapping by way of a natural collection of complex analytic functions. We first consider the rational function f given by

$$f(z) = \frac{az+b}{cz+d}, \tag{20}$$

where a, b, c, d are complex numbers. We cannot allow both c and d to be 0, or else we are dividing by 0. There is another natural condition; we do not want this mapping to degenerate into a constant. Such degeneration occurs if a is a multiple of c and b is the same multiple of d, in other words, if $ad - bc = 0$. By restricting to the case when $ad - bc \neq 0$, we avoid both problems.

Definition 4.1. A *linear fractional transformation* is a rational function of the form (20) where $ad - bc \neq 0$.

We let L denote the collection of linear fractional transformations. We will next show that L can be regarded as the collection of two-by-two complex matrices with determinant one.

Given a linear fractional transformation, we obtain the same function if we multiply the numerator and denominator by the same nonzero constant. We may therefore assume, without loss of generality, that $ad-bc = 1$. Assume f is as in (20) and that $ad - bc = 1$. We can record the information defining f as a two-by-two matrix

$$F = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Thinking of F rather than of f has some advantages. The identity map corresponds to the identity matrix, and composition corresponds to matrix multiplication. If $f, g \in L$, then so is $g \circ f$. Write $g(z) = \frac{Az+B}{Cz+D}$. We compute the composition

$$g(f(z)) = \frac{Af(z)+B}{Cf(z)+D} = \frac{A\frac{az+b}{cz+d}+B}{C\frac{az+b}{cz+d}+D} = \frac{(Aa+Bc)z+(Ab+Bd)}{(Ca+Dc)z+(Cb+Dd)}. \tag{21}$$

We identity f and g with matrices

$$F = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

$$G = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

Then the composition $g \circ f$ is identified with the matrix

$$GF = \begin{pmatrix} Aa+Bc & Ab+Bd \\ Ca+Dc & Cb+Dd \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix}. \tag{22}$$

Note that $ad - bc$ is the determinant of the matrix F corresponding to f and that $AD - BC$ is the determinant of the matrix G corresponding to g. Since the determinant of the product of matrices is the product of the respective determinants, it follows that the determinant of GF is not zero. Had we assumed each determinant was 1, the product would also be 1. We also obtain a formula for the inverse mapping; we take the inverse matrix. See Exercise 3.18. The natural assumption that the determinant equals 1 enables us to compute the inverse easily:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \tag{23}$$

▶ **Exercise 3.18.** Put $w = f(z) = \frac{az+b}{cz+d}$, where $ad - bc = 1$. Solve for z as a function of w. Check that the answer agrees with the inverse matrix from (23).

We may therefore consider L to be the group of two-by-two matrices with complex entries and determinant one. In elementary linear algebra we learn Gaussian elimination, or row operations, as a method for solving a system of linear equations. The effect of row operations is to write a matrix of coefficients as a product of particularly simple matrices. We can do the same thing for linear fractional transformations, and we will obtain a beautiful geometric corollary.

First we discuss the geometric interpretation of the three simplest linear fractional transformations. The mapping $z \to z + \beta = T_\beta(z)$ is a translation. For $\alpha \neq 0$, the map $z \to \alpha z = M_\alpha z$ is a dilation and a rotation; it changes the scale by a factor of $|\alpha|$ and rotates through an angle θ if $\alpha = |\alpha| e^{i\theta}$. The mapping $z \to \frac{1}{z} = R(z)$ is an inversion (taking the reciprocal). For the moment we write T for any translation, M for any multiplication, and R for the reciprocal. We will show that every linear fractional transformation can be written as a composition of these three simpler kinds.

For convenience we write the transformations as matrices:

$$T_\beta = \begin{pmatrix} 1 & \beta \\ 0 & 1 \end{pmatrix},$$

$$M_\alpha = \begin{pmatrix} \alpha & 0 \\ 0 & 1 \end{pmatrix},$$

$$R = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Let $f(z) = \frac{az+b}{cz+d}$. If $c = 0$, there are three possibilities for f. Since $ad - bc \neq 0$, necessarily $d \neq 0$ and $a \neq 0$. In this case $f(z) = \frac{a}{d}z + \frac{b}{d}$. If $b = 0$, then f is a multiplication. If $b \neq 0$, then f is an affine transformation. If $\frac{a}{d} = 1$, f is a translation. Otherwise f is the composition of a translation and a multiplication. Thus, when $c = 0$, the possibilities are $f = M$, $f = T$, and $f = TM$. We may regard the identity mapping as either the translation T_0 or the multiplication M_1.

If $c \neq 0$, we will need to take an inversion. To simplify f, we simply divide $cz + d$ into $az + b$, obtaining $\frac{a}{c}$ with a remainder of $\frac{bc-ad}{c}$. Using the definition of division, we obtain

$$\frac{az+b}{cz+d} = \frac{a}{c} + \frac{bc-ad}{c(cz+d)}. \tag{24}$$

We interpret (24) as a composition of mappings:

$$z \to cz \to cz+d \to \frac{1}{cz+d} \to \frac{bc-ad}{c}\frac{1}{cz+d} \to \frac{bc-ad}{c}\frac{1}{cz+d} + \frac{a}{c} = \frac{az+b}{cz+d}.$$

Using the above geometric language, we have written

$$f = TMRTM, \tag{25}$$

or more specifically

$$\frac{az+b}{cz+d} = T_{\frac{a}{c}} M_{\frac{bc-ad}{c}} R T_d M_c(z). \tag{26}$$

The respective translations in (26) are not needed when $d = 0$ or $a = 0$, but the notation (25) is still valid because T_0 is the identity mapping.

The next theorem summarizes these results and includes a beautiful consequence. Let $\mathcal{S}$ denote the collection of lines and circles in $\mathbf{C}$. We include the special case of a single point (a circle of radius 0) in $\mathcal{S}$, but we will not fully understand this situation until we deal with infinity in the next section.

The statement of Theorem 4.1 means that the image of a line under a linear fractional transformation is either a line or a circle, and the image of a circle under a linear fractional transformation is a line or a circle. The following example, noted earlier in Lemma 1.1, shows that the image of a circle need not be a line, and conversely. The image of a point is a point, but we need to allow the point at infinity.

Example 4.1. Put $f(z) = i\frac{1-z}{1+z}$. Then $|z| = 1$ if and only if $\mathrm{Im}(f(z)) = 0$. Thus the image under f of the unit circle is the real axis. Then f^{-1} maps the real axis to the unit circle.

We can give a simple description of $\mathcal{S}$ in terms of Hermitian symmetric polynomials. Let A, C be real numbers, and let β be a complex number. Consider the Hermitian symmetric polynomial $\Phi(z, \overline{z})$ defined by

$$\Phi(z, \overline{z}) = A|z|^2 + \beta z + \overline{\beta z} + C. \tag{27}$$

Here the coefficients A and C are real and $\beta \in \mathbf{C}$. We assume that not all of A, β, C are zero. If either A or β is not zero, then Φ is nonconstant, and its zero-set of Φ is either a line or a circle, where we allow the special case of a single point. In case $A = \beta = 0$ but $C \neq 0$, we will think of the zero-set of Φ as a single point at infinity. Conversely, each line or circle is the zero-set of some such Φ.

Theorem 4.1. *Each $f \in L$ maps $\mathcal{S}$ to itself.*

Proof. It is evident that each translation T and each multiplication M maps $\mathcal{S}$ to itself. Thus a composition of such does the same. The conclusion therefore holds when $f = TM$. We next check that the inversion (reciprocal) maps $\mathcal{S}$ to itself. Assume that V is a line or a circle. Then V is the zero-set of some Φ as in (27). Note that $0 \in V$ if and only if $C = 0$. Set $z = \frac{1}{w}$ in (27) and clear denominators. We obtain a new Hermitian symmetric polynomial $\Phi^*(w, \overline{w})$ defined by

$$\Phi^*(w, \overline{w}) = |w|^2 \Phi(\frac{1}{w}, \frac{1}{\overline{w}}) = A + \beta\overline{w} + \overline{\beta} w + C|w|^2.$$

The zero-set of Φ^* is a circle when $C \neq 0$, and it is a line if $C = 0$. The special case when $\beta = C = 0$ must be considered. Then $A \neq 0$, and the zero-set of Φ^* becomes the point at infinity. We summarize as follows. If Φ is Hermitian symmetric and its zero-set is a line or circle of positive radius, then the same is true for Φ^*. If the zero-set of Φ is a single point p not the origin, then the zero-set of Φ^* is the single point $\frac{1}{p}$. If the zero-set of Φ is the origin, then the zero-set of Φ^* is the point at infinity. Conversely if the zero-set of Φ is the point at infinity, then the zero-set of Φ^* is the origin.

These remarks show that inversion maps $\mathcal{S}$ to itself. The same is true for translation and multiplication. Hence (25) implies that every linear fractional transformation maps $\mathcal{S}$ to itself. □

▶ **Exercise 3.19.** Find a linear fractional transformation that maps the exterior of a circle of radius 2 with center at 2 to the interior of the unit circle.

▶ **Exercise 3.20.** Let f be a given linear fractional transformation. Determine which lines are mapped to circles under f and which circles are mapped to lines.

▶ **Exercise 3.21.** Given a line in $\mathbf{C}$, describe precisely which linear fractional transformations map this line to a circle. Given a circle in $\mathbf{C}$, describe precisely which linear fractional transformations map this circle to a line.

▶ **Exercise 3.22.** Find all linear fractional transformations mapping the real line to itself.

▶ **Exercise 3.23.** Find all linear fractional transformations mapping the unit circle to itself.

▶ **Exercise 3.24.** Show that the conjugation map $z \to \overline{z}$ maps $\mathcal{S}$ to itself. Conclude that the mapping $z \to \frac{a\overline{z}+b}{c\overline{z}+d}$ maps $\mathcal{S}$ to itself.

5. The Riemann sphere

We are ready to discuss infinity. Riemann had the idea of adding a point to $\mathbf{C}$, called the point at infinity, and then visualizing the result as a sphere. The resulting set is called either the *Riemann sphere* or the *extended complex plane*. The Riemann sphere provides some wonderful new perspectives on complex analysis, and we briefly describe some of these now.

We can realize the Riemann sphere in the following way. Let U_0 be a copy of $\mathbf{C}$. Let U_1 consist of the set of reciprocals of elements in $\mathbf{C}$, where we denote the reciprocal of 0 by ∞. Note that the intersection of these two sets is $\mathbf{C}^*$, standard notation for the nonzero complex numbers. We let $X = U_0 \cup U_1$. When we are working near 0, we work in U_0; when we are working near ∞, we work in U_1. If we are working somewhere and we wish to pass between the two sets, we take a reciprocal. This simple idea leads to the notion of a *Riemann surface* and more generally to that of a *complex manifold*. The subject of *complex geometry* is based upon the study of complex manifolds, but it is far more sophisticated than the geometry we study in this book.

The procedure from the previous section, where we replaced a Hermitian symmetric polynomial Φ with Φ^*, amounts to using the map $z \to \frac{1}{z}$ to pass from U_0

to U_1. The same idea enables us to define limits on the Riemann sphere. Infinity behaves the same as any other point! We have already seen the definition of convergent sequence on $\mathbf{C}$. We recall the definition of limits on $\mathbf{C}$ in order to extend the definition to the Riemann sphere.

Definition 5.1. Fix $a, L \in \mathbf{C}$. Let S be an open set containing a. We say that $\lim_{z\to a} f(z) = L$ if, for all $\epsilon > 0$, there is a $\delta > 0$ such that

$$0 < |z - a| < \delta \quad \text{implies} \quad |f(z) - L| < \epsilon. \tag{28}$$

We can extend the definition of a limit to allow both a and L to be ∞. We can also talk about neighborhoods of infinity.

Definition 5.2. Fix $L \in \mathbf{C}$. Then $\lim_{z\to\infty} f(z) = L$ if and only if $\lim_{z\to 0} f(\frac{1}{z}) = L$. Also, $\lim_{z\to a} f(z) = \infty$ if and only if $\lim_{z\to a} \frac{1}{f(z)} = 0$.

For example, if $k \in \mathbf{N}$, then $\lim_{z\to\infty} z^k = \infty$. Now that we have understood limits, we let X denote $\mathbf{C} \cup \{\infty\}$ together with the topology determined by these limits. We call X the Riemann sphere. By *neighborhood* of a point $p \in \mathbf{C}$ we mean any subset which contains an open ball about p. As we did for limits, we define *neighborhood of infinity* by taking reciprocals.

Definition 5.3. Let S be a subset of the Riemann sphere containing ∞. We say that S is a neighborhood of ∞ if the set of reciprocals of elements of S is a neighborhood of 0.

▶ **Exercise 3.25.** Use Definition 5.2 to show that $\lim_{z\to\infty} \frac{az+b}{cz+d} = \frac{a}{c}$.

▶ **Exercise 3.26.** Use Definition 5.2 to show that $\lim_{z\to \frac{-d}{c}} \frac{az+b}{cz+d} = \infty$.

Once we have understood limits on the Riemann sphere X, we can introduce open sets and make X into a topological space. The resulting space is the basic example of a compact Riemann surface.

The next exercise uses *stereographic projection* to provide a one-to-one correspondence between the unit sphere and the extended complex plane.

▶ **Exercise 3.27.** Consider $\mathbf{R}^3$ with coordinates (x_1, x_2, x_3). Let p be a point on the unit sphere in $\mathbf{R}^3$ other than the north pole $(0, 0, 1)$. Find the line from the north pole to p and see where that point intersects the plane defined by $x_3 = 0$. Call the point of intersection $s(p)$. Define s of the north pole to be infinity. This mapping s is called stereographic projection. Write explicit formulas for $s(p)$ and show that s maps the sphere bijectively onto the extended complex plane. Find a formula for s^{-1}.

▶ **Exercise 3.28.** Find the image of the equator under stereographic projection.

▶ **Exercise 3.29.** Using the notation of Exercise 3.27, assume that $s(p) = z$. Find the point q for which $s(q) = \frac{1}{z}$.

▶ **Exercise 3.30.** Replacing x by $\frac{1}{x}$ is sometimes useful on the real line. For real numbers a, b consider the integral $F(a, b)$ given by

$$F(a, b) = \int_0^\infty \frac{1}{x^2 + 1} \frac{x^b - x^a}{(1 + x^a)(1 + x^b)} dx.$$

First show that $F(a,b) = 0$. Then use this result to show that

$$\int_0^\infty \frac{1}{x^2+1}\frac{1}{(1+x^a)}dx = \frac{1}{2}\int_0^\infty \frac{1}{x^2+1}dx = \frac{\pi}{4}.$$

► **Exercise 3.31.** (Difficult) Suppose that f is a one-to-one continuous mapping from the Riemann sphere onto itself and that f maps $\mathcal{S}$ to itself. Show that f is either a linear fractional transformation or the conjugate of a linear fractional transformation.