7.1 The group of isometries of the plane

In Chapter 3, we took up Euclid's idea of "moving" geometric figures, and we made it precise in the concept of an *isometry* of the plane $\mathbb{R}^2$. An isometry is defined to be a function $f : \mathbb{R}^2 \to \mathbb{R}^2$ that preserves distance; that is,

$$|f(P_1)f(P_2)| = |P_1P_2| \quad \text{for any points } P_1, P_2, \in \mathbb{R}^2,$$

where $|P_1P_2| = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$ denotes the distance between the points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$.

It follows immediately from this definition that, when f and g are isometries, so is their *composite* or *product* fg (the result of applying g, then f). Namely,

$$\begin{aligned} |f(g(P_1))f(g(P_2))| &= |g(P_1)g(P_2)| \quad \text{because } f \text{ is an isometry} \\ &= |P_1P_2| \quad \text{because } g \text{ is an isometry.} \end{aligned}$$

What is less obvious is that *any isometry f has an inverse, f^{-1}, which is also an isometry*. To prove this fact, we use the result from Section 3.7 that any isometry of $\mathbb{R}^2$ is the product of one, two, or three reflections.

First suppose that $f = r_1r_2r_3$, where r_1, r_2, and r_3 are reflections. Then, because a reflection composed with itself is the identity function, we find

$$\begin{aligned} fr_3r_2r_1 &= r_1r_2r_3r_3r_2r_1 \\ &= r_1r_2r_2r_1 \quad \text{because } r_3r_3 \text{ is the identity function} \\ &= r_1r_1 \quad \text{because } r_2r_2 \text{ is the identity function} \\ &= \text{identity function,} \end{aligned}$$

and therefore, $r_3r_2r_1 = f^{-1}$. This calculation also shows that f^{-1} is an isometry, because it is a product of reflections. The proof is similar (but shorter) when f is the product of one or two reflections.

These properties of isometries are characteristic of a *group of transformations*. A *transformation* of a set S is a function from S to S, and a collection G of transformations forms a *group* if it has the two properties:

- If f and g are in G, then so is fg.
- If f is in G, then so is its inverse, f^{-1}.

It follows that G includes the identity function ff^{-1}, which can be written as 1. This notation is natural when we write the composite of two functions f, g as the "product" fg.

What is a geometry?

In 1872, the German mathematician Felix Klein pointed out that various kinds of geometry go with various groups of transformations. For example, the Euclidean geometry of $\mathbb{R}^2$ goes with the group of isometries of $\mathbb{R}^2$. The meaningful concepts of the geometry correspond to properties that are left *unchanged* by transformations in the group. Isometries of $\mathbb{R}^2$ leave distance or length unchanged, so distance is a meaningful concept of Euclidean geometry. It is called an *invariant* of the isometry group of $\mathbb{R}^2$. This invariance is no surprise, because isometries are *defined* as the transformations that preserve distance.

However, it is interesting that other things are also invariant under isometries, such as straightness of lines and circularity of circles. It is not entirely obvious that a length-preserving transformation preserves straightness, but it can be proved by showing first that any reflection preserves straightness, and then using the theorem of Section 3.7, that any isometry is a product of reflections.

An example of a concept without meaning in Euclidean geometry is "being vertical," because a vertical line can be transformed to a nonvertical line by an isometry (for example, by a rotation). We can do without the concept of "vertical" in geometry because we have the concept of "being relatively vertical," that is, perpendicular. A concept that is harder to do without is "clockwise order on the circle." This concept has no meaning in Euclidean geometry because the points $A = (-1,0)$, $B = (0,1)$, $C = (1,0)$, and $D = (0,-1)$ have clockwise order on the circle, but their respective reflections in the x-axis do not.

However, we can define *oriented Euclidean geometry*, in which clockwise order is meaningful, by using a smaller group of transformations. Instead of the group $\mathrm{Isom}(\mathbb{R}^2)$ of all isometries of $\mathbb{R}^2$, take $\mathrm{Isom}^+(\mathbb{R}^2)$, each member of which is the product of an *even* number of reflections. $\mathrm{Isom}^+(\mathbb{R}^2)$ is a group because

- If f and g are products of an even number of reflections, so is fg.
- If $f = r_1 r_2 \cdots r_{2n}$ is the product of an even number of reflections, then so is f^{-1}. In fact, $f^{-1} = r_{2n} \cdots r_2 r_1$, by the argument used above to invert the product of any number of reflections.

And any transformation in $\mathrm{Isom}^+(\mathbb{R}^2)$ preserves clockwise order because any product of two reflections does: The first reflection reverses the order, and then the second restores it.

This example shows how a geometry of $\mathbb{R}^2$ with more concepts comes from a group with fewer transformations. In $\mathbb{R}^3$, one has the concept of "handedness"—which distinguishes the right hand from the left—which is not preserved by all isometries of $\mathbb{R}^3$. However, it is preserved by products of an even number of reflections in planes. Thus, the geometry of $\mathrm{Isom}(\mathbb{R}^3)$ does not have the concept of handedness, but the geometry of $\mathrm{Isom}^+(\mathbb{R}^3)$ does. Restricting the transformations to those that preserve *orientation*—as it is generally called—is a common tactic in geometry.

However, the main goal of this book is to show that there are interesting geometries with *fewer* concepts than Euclidean geometry. These geometries are obtained by taking larger groups of transformations, which we study in the remainder of this chapter.

Exercises

7.1.1 Use the results of Section 3.7 to show that each member of $\mathrm{Isom}^+(\mathbb{R}^2)$ is either a translation or a rotation.

7.1.2 Why does an isometry map any circle to a circle?

We took care to write the inverse of the isometry $r_1r_2r_3$ as $r_3r_2r_1$ because only this ordering of terms will always give the correct result.

7.1.3 Give an example of two reflections r_1 and r_2 such that $r_1r_2 \neq r_2r_1$.

7.2 Vector transformations

In Chapter 4, we viewed the plane $\mathbb{R}^2$ as a *real vector space*, by considering its points to be *vectors* that can be added and multiplied by scalars. If $\mathbf{u} = (u_1, u_2)$ and $\mathbf{v} = (v_1, v_2)$, we defined the *sum* of $\mathbf{u}$ and $\mathbf{v}$ by

$$\mathbf{u} + \mathbf{v} = (u_1 + v_1, u_2 + v_2)$$

and the *scalar multiple* $a\mathbf{u}$ of $\mathbf{u}$ by a real number a by

$$a\mathbf{u} = (au_1, au_2).$$

A transformation f of $\mathbb{R}^2$ *preserves* these two operations on vectors if

$$f(\mathbf{u} + \mathbf{v}) = f(\mathbf{u}) + f(\mathbf{v}) \quad \text{and} \quad f(a\mathbf{u}) = af(\mathbf{u}), \tag{*}$$

and such a transformation is called *linear*.

One reason for calling the transformation "linear" is that it preserves straightness of lines. A straight line is a set of points of the form $\mathbf{a}+t\mathbf{b}$, where $\mathbf{a}$ and $\mathbf{b}$ are constant vectors and t runs through the real numbers. Figure 7.1 shows the role of the vectors $\mathbf{a}$ and $\mathbf{b}$: $\mathbf{a}$ is one point on the line, and $\mathbf{b}$ gives the direction of the line.

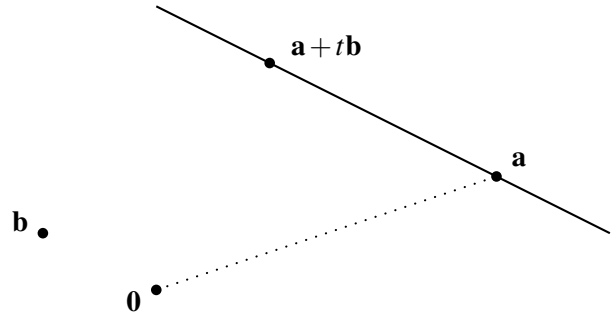

Figure 7.1: Points on a line

If we apply a linear transformation f to this set of points, we get the set of points $f(\mathbf{a}+t\mathbf{b})$. And by the linearity conditions (*), this set consists of points of the form $f(\mathbf{a})+tf(\mathbf{b})$, which is another straight line: $f(\mathbf{a})$ is one point on it, and $f(\mathbf{b})$ gives the direction of the line.

It follows from this calculation that, if $\mathscr{L}_1$ and $\mathscr{L}_2$ are two lines with direction $\mathbf{b}$ and f is a linear transformation, then $f(\mathscr{L}_1)$ and $f(\mathscr{L}_2)$ are two lines with direction $f(\mathbf{b})$. In other words, a linear transformation also preserves parallels.

Matrix representation

Another consequence of the linearity conditions (*) is that each linear transformation f of $\mathbb{R}^2$ can be specified by four real numbers a,b,c,d: any point (x,y) of $\mathbb{R}^2$ is sent by f to the point $(ax+by,cx+dy)$.

Certainly, there are numbers a,b,c,d that give the particular values

$$f((1,0))=(a,c) \quad \text{and} \quad f((0,1))=(b,d).$$

But the value of $f((x,y))$ follows from these particular values by linearity:

$$(x,y)=x(1,0)+y(0,1),$$

and therefore,

$$\begin{aligned} f((x,y)) &= f(x(1,0)+y(0,1)) \\ &= xf((1,0))+yf((0,1)) \\ &= x(a,c)+y(b,d) \\ &= (ax+by,cx+dy). \end{aligned}$$

The linear transformation $(x,y) \mapsto (ax+by,cx+dy)$ is usually represented by the *matrix*

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \text{where} \quad a,b,c,d \in \mathbb{R}.$$

To find where $(x,y) \in \mathbb{R}^2$ is sent by f, one writes it as the "column vector" $\begin{pmatrix} x \\ y \end{pmatrix}$ and multiplies this column on the left by M according to the matrix product rule:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax+by \\ cx+dy \end{pmatrix}.$$

The main advantage of the matrix notation is that it gives the product of two linear transformations, first $(x,y) \mapsto (a_2x+b_2y,c_2x+d_2y)$ and then $(x,y) \mapsto (a_1x+b_1y,c_1x+d_1y)$, by the matrix product rule:

$$\begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix} \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix} = \begin{pmatrix} a_1a_2+b_1c_2 & a_1b_2+b_1d_2 \\ c_1a_2+d_1c_2 & c_1b_2+d_1d_2 \end{pmatrix}.$$

Matrix notation also exposes the role of the *determinant*, $\det(M)$, which must be nonzero for the linear transformation to have an inverse. If

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \text{then} \quad \det(M) = ad-bc,$$

and if $\det(M) \neq 0$, then

$$M^{-1} = \frac{1}{\det M} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Examples of linear transformations

Any 2×2 real matrix M represents a linear transformation, because it follows from the definition of matrix multiplication that

$$M(\mathbf{u}+\mathbf{v}) = M\mathbf{u}+M\mathbf{v} \quad \text{and} \quad M(a\mathbf{u}) = aM\mathbf{u}$$

for any vectors $\mathbf{u}$ and $\mathbf{v}$ (written in column form).

Among the invertible linear transformations are certain isometries, such as rotations and reflections in lines through the origin. Recall from Section 3.6 that a rotation is a transformation of the form

$$(x,y) \mapsto (cx-sy, sx+cy), \quad \text{hence given by the matrix} \quad R = \begin{pmatrix} c & -s \\ s & c \end{pmatrix}.$$

The numbers c and s satisfy $c^2+s^2 = 1$ (they are actually $\cos\theta$ and $\sin\theta$, where θ is the angle of rotation); hence,

$$\det R = 1 \quad \text{and therefore} \quad R^{-1} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix}.$$

Likewise, reflection in the x-axis is the linear transformation

$$(x,y) \mapsto (x,-y), \quad \text{given by the matrix} \quad \overline{X} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

We can reflect $\mathbb{R}^2$ in any line $\mathscr{L}$ through O with the help of the rotation R that sends the x-axis to $\mathscr{L}$:

- First apply R^{-1} to send $\mathscr{L}$ to the x-axis.
- Then carry out the reflection by applying $\overline{X}$.
- Then send the line of reflection back to $\mathscr{L}$ by applying R.

In other words, to reflect the point $\mathbf{u}$ in $\mathscr{L}$, we find the value of $R\overline{X}R^{-1}\mathbf{u}$. Hence, reflection in $\mathscr{L}$ is represented by the matrix $R\overline{X}R^{-1}$.

Thus, the linear transformations of $\mathbb{R}^2$ include the isometries that are products of reflection on lines through O. But this is not all. An example of a linear transformation that is not an isometry is the *stretch by factor k* in the x-direction,

$$(x,y) \mapsto (kx,y), \quad \text{given by the matrix} \quad S = \begin{pmatrix} k & 0 \\ 0 & 1 \end{pmatrix}.$$

It can be shown that any invertible linear transformation of $\mathbb{R}^2$ is a product of reflections in lines through O and stretches in the x-direction (by factors $k \neq 0$).

Affine transformations

Linear transformations preserve geometrically natural properties such as straightness and parallelism, but they also preserve the origin, which really is not geometrically different from any other point. To abolish the special position of the origin, we allow linear transformations to be composed with translations, obtaining what are called *affine* transformations. If we write an arbitrary linear transformation of (column) vectors $\mathbf{u}$ in the form

$$f(\mathbf{u}) = M\mathbf{u}, \quad \text{where } M \text{ is an invertible matrix,}$$

then an arbitrary affine transformation takes the form

$$g(\mathbf{u}) = M\mathbf{u} + \mathbf{c}, \quad \text{where } \mathbf{c} \text{ is a constant vector.}$$

Because translations preserve everything except position, affine transformations preserve everything that linear transformations do, except position. In effect, they allow any point to become the origin.

The geometry of affine transformations is called *affine geometry*. Its theorems include those in the first few sections of Chapter 4, such as the fact that diagonals of a parallelogram bisect each other, and the concurrence of the medians of a triangle. These theorems belong to affine geometry because they are concerned only with quantities, such as the midpoint of a line segment, that are preserved by affine transformations.

Exercises

7.2.1 Compute MM^{-1} for the general 2×2 matrix $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, and verify that it equals the identity matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

7.2.2 Write down the matrix for clockwise rotation through angle $\pi/4$.

7.2.3 Write down the matrix for reflection in the line $y = x$, and check that it equals $R\overline{X}R^{-1}$, where R is the matrix for rotation through $\pi/4$ found in Exercise 7.2.2.

7.2.4 The matrix $M = \begin{pmatrix} k & 0 \\ 0 & k \end{pmatrix}$ represents a dilation of the plane by factor k (also known as a *similarity* transformation). Explain geometrically why this transformation is a product of reflections in lines through O and of stretches by factor k in the x-direction.

7.2.5 Show that the midpoint of any line segment is preserved by linear transformations and hence by affine transformations.

7.2.6 More generally, show that the ratio of lengths of any two segments of the same line is preserved by affine transformations.

7.3 Transformations of the projective line

Looking back at our approach to the projective line in Chapter 5, we see that we were following Klein's idea. First we found the transformations of the projective line, and then a quantity that they leave invariant—the cross-ratio. In this section we look more closely at projective transformations, and show that they too can be viewed as linear transformations.

In Sections 5.5 and 5.6, we showed that the transformations of the projective line $\mathbb{R} \cup \{\infty\}$ are precisely the linear fractional functions

$$f(x) = \frac{ax+b}{cx+d} \quad \text{where} \quad ad-bc \neq 0.$$

We did this by showing:

- Any linear fractional function is a product of functions sending x to $x+l$, x to kx, and x to $1/x$, and that each of the latter functions can be realized by projection of one line onto another.
- Conversely, any projection of one line onto another is represented by a linear fractional function of x, with the understanding that $1/0 = \infty$ and $1/\infty = 0$.

In the exercises to Section 5.6 you were asked to show that linear fractional functions $f(x) = \frac{ax+b}{cx+d}$ behave like the matrices $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, by showing that composition of the functions corresponds to multiplication of the corresponding matrices. In this section, we explain the connection by representing mappings of the projective line directly by linear transformations of the plane.

We begin by defining the projective line in the manner of Section 5.4. There we defined the *real projective plane* $\mathbb{RP}^2$. Its "points" are the lines through O in $\mathbb{R}^3$, and its "lines" are the planes through O. Here we need only one projective line, which we can take to be the real projective line $\mathbb{RP}^1$, whose "points" are the lines through O in the ordinary plane $\mathbb{R}^2$.

We label each line through O, if it meets the line $y = 1$, by the x-coordinate s of the point of intersection (Figure 7.2). The single line that does *not* meet $y = 1$, namely, the x-axis, naturally gets the label ∞.

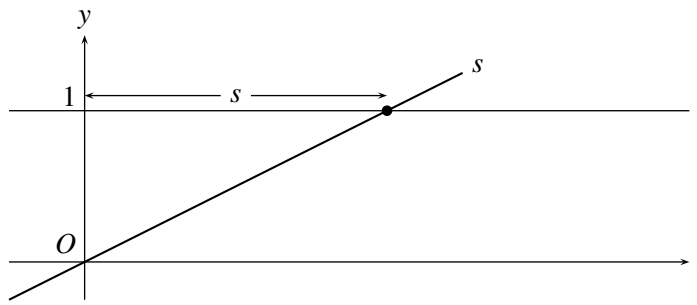

Figure 7.2: Correspondence between lines through O and points on $y = 1$

Figure 7.3 shows some lines through O with their labels.

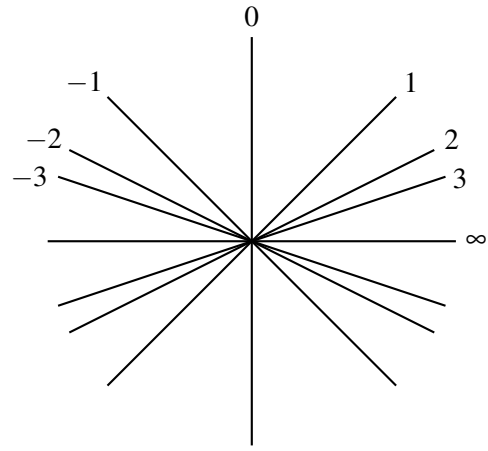

Figure 7.3: Labeling of lines through O

Now a projective map of the ordinary line $y = 1$ sends the point with x-coordinate s to the point with x-coordinate $f(s)$, for some linear fractional function

$$f(s) = \frac{as+b}{cs+d}.$$

This function corresponds to a map of the plane $\mathbb{R}^2$ sending the line with label s to the line with label $\frac{as+b}{cs+d}$. Bearing in mind that the label represents the "reciprocal slope" ("run over rise") of the line, we find that *one such map is the linear map of the plane given by the matrix*

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

To see why, we apply this linear map to a typical point (sx,x) on the line with label s. We find where M sends it by writing (sx,x) as a column vector and multiplying it on the left by M:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} sx \\ x \end{pmatrix} = \begin{pmatrix} asx+bx \\ csx+dx \end{pmatrix}.$$

The column vector $\begin{pmatrix} asx+bx \\ csx+dx \end{pmatrix}$ represents the point $(asx+bx, csx+dx)$, which lies on the line with reciprocal slope

$$\frac{asx+bx}{csx+dx} = \frac{as+b}{cs+d}.$$

The latter line is therefore independent of x and it is the line with label $\frac{as+b}{cs+d}$. Thus, M maps the line with label s to the line with label $\frac{as+b}{cs+d}$, as required. □

Because a "point" of $\mathbb{RP}^1$ is a whole line through O, we care only that the matrix

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

sends the *line* with label s to the *line* with label $\frac{as+b}{cs+d}$. It does not matter how M moves individual points on the line. It is *not* generally the case that M sends the particular point $(s,1)$ on the line with label s to the particular point $(\frac{as+b}{cs+d},1)$ on the line with label $\frac{as+b}{cs+d}$. Indeed, it is clearly impossible when M represents the map $s \mapsto 1/s$ of $\mathbb{RP}^1$. A matrix M sends each point of $\mathbb{R}^2$ to another point of $\mathbb{R}^2$. Hence it cannot send $(0,1)$ to $(1/0,1) = (\infty,1)$, because the latter is not a point of $\mathbb{R}^2$. However, M *can* send the line with label 0 (the y-axis) to the line with label ∞ (the x-axis), and this is exactly what happens (see Exercises 7.3.1 and 7.3.2).

It should also be pointed out that the representation of linear fractional functions by matrices is not unique. The fraction

$$\frac{as+b}{cs+d} \quad \text{is equal to} \quad \frac{kas+kb}{kcs+kd} \quad \text{for any } k \neq 0.$$

Hence, the function $f(s) = \frac{as+b}{cs+d}$ is represented not only by

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{but also by} \quad kM = \begin{pmatrix} ka & kb \\ kc & kd \end{pmatrix} \quad \text{for any } k \neq 0.$$

The linear transformations kM combine the transformation M with dilation by factor k, so they are all different. Thus, the same linear fractional function f is represented by infinitely many different transformations kM of $\mathbb{R}^2$. The message, again, is that we care only that each of these transformations sends the line with label s to the line with label $f(s)$.

Exercises

7.3.1 Write down a matrix M that represents the map $s \mapsto 1/s$ of $\mathbb{RP}^1$.

7.3.2 Verify that your matrix M in Exercise 7.3.1 maps the y-axis onto the x-axis.

7.3.3 Sketch a picture of the lines with labels $1/2$, $-1/2$, $1/3$, and $-1/3$.

The nonuniqueness of the matrix M corresponding to the linear fractional function f raises the question: Is there a natural way to choose *one* matrix for each linear fractional function? Actually, no, but there is a natural way to choose *two* matrices.

7.3.4 Given that

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{and} \quad ad - bc \neq 0,$$

show that the determinant of kM has absolute value 1 for exactly two of the matrices kM, where $k \neq 0$.

7.4 Spherical geometry

The unit sphere in $\mathbb{R}^3$ consists of all points at unit distance from O, that is, all points (x, y, z) satisfying the equation

$$x^2 + y^2 + z^2 = 1.$$

This surface is also called the *2-sphere*, or $\mathbb{S}^2$, because its points can be described by two coordinates—latitude and longitude, for example. Its geometry is essentially *two-dimensional*, like that of the Euclidean plane $\mathbb{R}^2$ or the real projective plane $\mathbb{RP}^2$, and indeed the fundamental objects of spherical geometry are "points" (ordinary points on the sphere) and "lines" (great circles on the sphere).

However, like the projective plane $\mathbb{RP}^2$, the sphere $\mathbb{S}^2$ is best understood via properties of the three-dimensional space $\mathbb{R}^3$. In particular, the "lines" on $\mathbb{S}^2$ are the intersections of $\mathbb{S}^2$ with planes through O in $\mathbb{R}^3$—the great circles—and the isometries of $\mathbb{S}^2$ are precisely the isometries of $\mathbb{R}^3$ that leave O fixed.

By definition, an isometry f of $\mathbb{R}^3$ preserves distance. Hence, if f leaves O fixed, it sends each point at distance 1 from O to another point at distance 1 from O. In other words, an isometry f of $\mathbb{R}^3$ that fixes O also maps $\mathbb{S}^2$ into itself. The restriction of f to $\mathbb{S}^2$ is therefore an isometry of $\mathbb{S}^2$, because f preserves distances on $\mathbb{S}^2$ as it does everywhere else. This statement is true whether one uses the straight-line distance between points of $\mathbb{S}^2$ or, as is more natural, the great-circle distance along the curved surface of $\mathbb{S}^2$ (Figure 7.4). The isometries of $\mathbb{S}^2$ are the maps of $\mathbb{S}^2$ into itself that preserve great circle distance, and we will see next why they are all restrictions of isometries of $\mathbb{R}^3$.

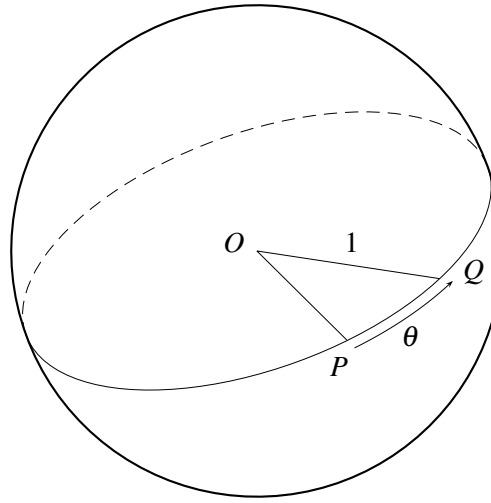

Figure 7.4: Great-circle distance