replace the points $P_1, P_2, P_3, \ldots$ by lines in space, we realize that there are extra lines (the horizontal lines) corresponding to the points on the horizon.

This model of the projective plane nicely captures our intuitive idea of points at infinity, but it also makes the idea clearer. We can see, for example, why it is proper for each line to have only one point at infinity, not two: because the lines $\mathscr{L}$ connecting O to points P along a line $\mathscr{M}$ in the plane $z = -1$ tend toward the *same* horizontal line as P tends to infinity in either direction (namely, the parallel to $\mathscr{M}$ through O).

It is hard to find a surface that behaves like $\mathbb{RP}^2$, but it is easy to find a curve that behaves like any "line" in it, a so-called *real projective line*. Figure 5.11 shows how. The "points" in a "line" of $\mathbb{RP}^2$, namely the lines through O in some plane through O, correspond to points of a circle through O. Each point $P \neq O$ on the circle corresponds to the line through O and P, and the point O itself corresponds to the tangent line at O.

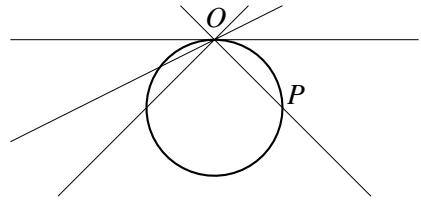

Figure 5.11: Modeling a projective line by a circle

Exercises

To gain more familiarity with calculations in $\mathbb{R}^3$, let us pursue the example of four "points" given above.

5.3.1 Find the plane $ax + by + cz = 0$ through the points $(0,0,1)$ and $(1,1,1)$, and check that it does not contain the points $(1,0,0)$ and $(0,1,0)$.

5.3.2 Show that $\mathbb{RP}^2$ has four "lines," no three of which have a common "point."

Not only does $\mathbb{RP}^2$ contain four "lines," no three of which have a "point" in common; the same is true of *any* projective plane, because this property follows *from the projective plane axioms alone*.

5.3.3 Suppose that A, B, C, D are four "points" in a projective plane, no three of which are in a "line." Consider the "lines" AB, BC, CD, DA. Show that if AB and BC have a common point E, then $E = B$.

5.3.4 Deduce from Exercise 5.3.3 that the three lines AB, BC, CD have no common point, and that the same is true of any three of the lines AB, BC, CD, DA.

5.4 Homogeneous coordinates

Because "points" and "lines" of $\mathbb{RP}^2$ are lines and planes through O in $\mathbb{R}^3$, they are easily handled by methods of linear algebra. A line through O is determined by any point $(x,y,z) \neq O$, and it consists of the points (tx,ty,tz), where t runs through all real numbers. Thus, a "point" is not given by a single triple (x,y,z), but rather by any of its nonzero multiples (tx,ty,tz). These triples are called the *homogeneous coordinates* of the "point."

A plane through O has a linear equation of the form $ax+by+cz=0$, called a *homogeneous equation*. The same plane is given by the equation $tax+tby+tcz=0$ for any nonzero t. Thus, a "line" is likewise not given by a single triple (a,b,c), but by the set of all its nonzero multiples (ta,tb,tc).

If (x_1,y_1,z_1) and (x_2,y_2,z_2) lie on different lines through O, then it is geometrically obvious that they lie in a unique plane $ax+by+cz=0$. The coordinates (a,b,c) of this plane can be found by solving the two equations

$$\begin{aligned} ax_1+by_1+cz_1 &= 0, \\ ax_2+by_2+cz_2 &= 0, \end{aligned}$$

for a, b, and c. Because there are more unknowns than equations, there is not a single solution triple but a whole space of them—in this case, a set of multiples (ta,tb,tc), all representing the same homogeneous equation.

This is the algebraic reason why two "points" lie on a unique "line" in $\mathbb{RP}^2$. There is a similar reason why two "lines" have a unique "point" in common. Two "lines" are given by two equations

$$\begin{aligned} a_1x+b_1y+c_1z &= 0, \\ a_2x+b_2y+c_2z &= 0, \end{aligned}$$

and we find their common "point" by solving these equations for x, y, and z. This problem is the same as above, but with the roles of a,b,c exchanged with those of x,y,z. The solution in this case is a set of multiples (tx,ty,tz) representing the homogeneous coordinates of the common "point."

The practicalities of finding the "line" through two "points" or the "point" common to two "lines" are explored in the next exercise set. But first I want to make a theoretical point. *It makes no algebraic difference if the coordinates of "points" and "lines" are complex numbers.* We can define a *complex projective plane* $\mathbb{CP}^2$, each "point" of which is a set of triples of the form (tx,ty,tz), where x,y,z are particular complex numbers

and t runs through all complex numbers. It remains true that any two "points" lie on a unique "line" and any two "lines" have unique common point, simply because the algebraic properties of complex linear equations are exactly the same as those of real linear equations. Similarly, one can show there are four "points," no three of which are in a "line" of $\mathbb{CP}^2$.

Thus, there is more than one model of the projective plane axioms. Later we shall look at other models, which enable us to see that certain properties of $\mathbb{RP}^2$ are not properties of all projective planes and hence do not follow from the projective plane axioms.

Projective space

It is easy to generalize homogeneous coordinates to *quadruples* (w,x,y,z) and hence to define the three-dimensional *real projective space* $\mathbb{RP}^3$. It has "points," "lines," and "planes" defined as follows (we use vector notation to shorten the definitions):

- A "point" is a line through O in $\mathbb{R}^4$, that is, a set of quadruples $t\mathbf{u}$, where $\mathbf{u} = (w,x,y,z)$ is a particular quadruple of real numbers and t runs through all real numbers.
- A "line" is a plane through O in $\mathbb{R}^4$, that is, a set $t_1\mathbf{u}_1 + t_2\mathbf{u}_2$ where $\mathbf{u}_1$ and $\mathbf{u}_2$ are linearly independent points of $\mathbb{R}^4$ and t_1 and t_2 run through all real numbers.
- A "plane" is a three-dimensional space through O in $\mathbb{R}^4$, that is, a set $t_1\mathbf{u}_1 + t_2\mathbf{u}_2 + t_3\mathbf{u}_3$, where $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$ are linearly independent points of $\mathbb{R}^4$ and t_1, t_2, and t_3 run through all real numbers.

Linear algebra then enables us to show various properties of the "points," "lines," and "planes" in $\mathbb{RP}^3$, such as:

1. Two "points" lie on a unique "line."
2. Three "points" not on a "line" lie on a unique "plane."
3. Two "planes" have unique "line" in common.
4. Three "planes" with no common "line" have one common "point."

These properties hold for any *three-dimensional projective space*, and $\mathbb{RP}^3$ is not the only one. There is also a complex projective space $\mathbb{CP}^3$, and many others. $\mathbb{RP}^3$ has an unexpected influence on the geometry of the sphere, as we will see in Section 7.8.

Exercises

5.4.1 Find the plane $ax+by+cz=0$ that contains the points $(1,2,3)$ and $(1,1,1)$.

5.4.2 Find the line of intersection of the planes $x+2y+3z=0$ and $x+y+z=0$.

5.4.3 You can write down the solution of Exercise 5.4.2 as soon as you have solved Exercise 5.4.1. Why?

5.5 Projection

The three-dimensional Euclidean space $\mathbb{R}^3$, in which the lines through O are the "points" of $\mathbb{RP}^2$ and the planes through O are the "lines" of $\mathbb{RP}^2$, also contains many other planes. Each plane $\mathscr{P}$ *not* passing through O can be regarded as a *perspective view* of the projective plane $\mathbb{RP}^2$, a view that contains all but one "line" of $\mathbb{RP}^2$.

Each point P of $\mathscr{P}$ corresponds to a line ("of sight") through O, and hence to a "point" of $\mathbb{RP}^2$. The only lines through O that do not meet $\mathscr{P}$ are those parallel to $\mathscr{P}$, and these make up the *line at infinity* or *horizon* of $\mathscr{P}$, as we have already seen in the case of the plane $z=-1$ in Section 5.3.

If $\mathscr{P}_1$ and $\mathscr{P}_2$ are any two planes not passing through O we can *project* $\mathscr{P}_1$ *to* $\mathscr{P}_2$ by sending each point P_1 in $\mathscr{P}_1$ to the point P_2 in $\mathscr{P}_2$ lying on the same line through O as P_1 (Figure 5.12). The geometry of $\mathbb{RP}^2$ is called "projective" because it encapsulates the geometry of a whole family of planes related by projection.

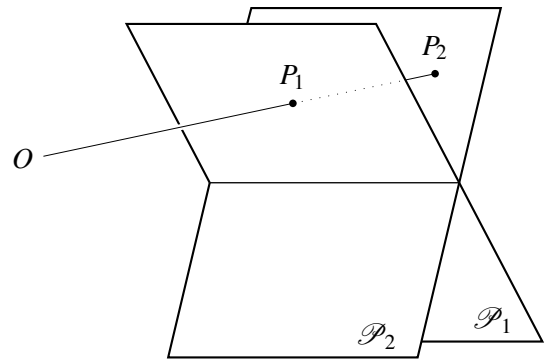

Figure 5.12: Projecting one plane to another

Projections of projective lines

Projection of one plane $\mathscr{P}_1$ onto another plane $\mathscr{P}_2$ produces an image of $\mathscr{P}_1$ that is generally distorted in some way. For example, a grid of squares on $\mathscr{P}_1$ may be mapped to a perspective view of the grid that looks like Figure 5.1. Nevertheless, straight lines remain straight under projection, so there are limits to the amount of distortion in the image. To better understand the nature and scope of projective distortion, in this subsection we analyze the mappings of the projective *line* obtainable by projection.

An effective way to see the distortion produced by projection of one line $\mathscr{L}_1$ onto another line $\mathscr{L}_2$ is to mark a series of equally spaced dots on $\mathscr{L}_1$ and the corresponding image dots on $\mathscr{L}_2$. You can think of the image dots as "shadows" of the dots on $\mathscr{L}_1$ cast by light rays from the point of projection P, except that we have projective lines through P, not rays, so it can seem as though the "shadow" on $\mathscr{L}_2$ comes ahead of the dot on $\mathscr{L}_1$. (See Figure 5.15, but bear in mind that a projective line is really circular, so it is always possible to pass through P, to a point on $\mathscr{L}_1$, then to a point on $\mathscr{L}_2$, in that order.)

In the simplest cases, where $\mathscr{L}_1$ and $\mathscr{L}_2$ are parallel, the image dots are also equally spaced. Figure 5.13 shows the case of *projection from a point at infinity*, where the lines from the dots on $\mathscr{L}_1$ to their images on $\mathscr{L}_2$ are parallel and hence the dots on $\mathscr{L}_1$ are simply translated a constant distance l. If we choose an origin on each line and use the same unit of length on each, then projection from infinity sends each x on $\mathscr{L}_1$ to $x+l$ on $\mathscr{L}_2$.

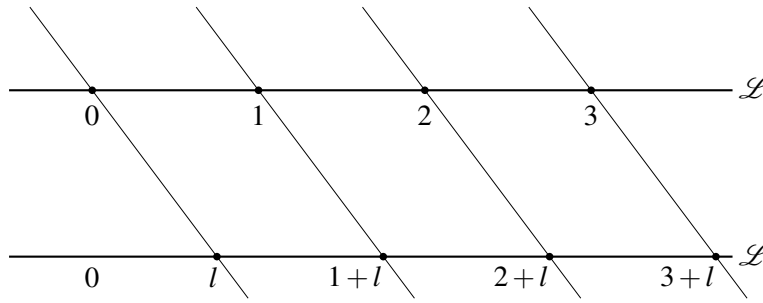

Figure 5.13: Projection from infinity

When $\mathscr{L}_1$ is projected from a finite point P, then the distance between dots is magnified by a constant factor $k \neq 0$. If we take P on a line through

the zero points on $\mathscr{L}_1$ and $\mathscr{L}_2$, then the projection sends each x on $\mathscr{L}_1$ to kx on $\mathscr{L}_2$ (Figure 5.14). Note also that this projection sends x on $\mathscr{L}_2$ to x/k on $\mathscr{L}_1$, so the magnification factor can be *any* $k \neq 0$.

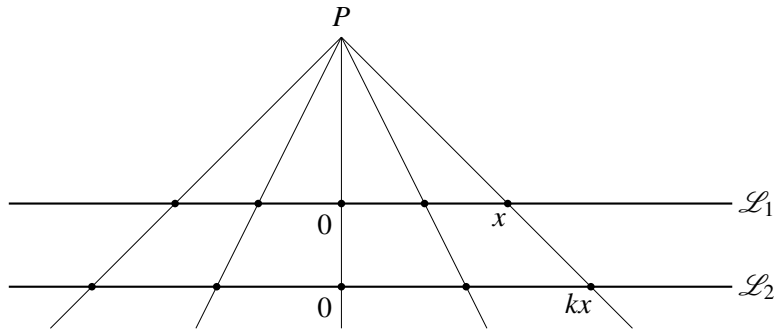

Figure 5.14: Projection from a finite point

When $\mathscr{L}_1$ and $\mathscr{L}_2$ are not parallel the distortion caused by projection is more extreme. Figure 5.15 shows how the spacing of dots changes when $\mathscr{L}_1$ is projected onto a perpendicular line $\mathscr{L}_2$ from a point O equidistant from both. Figure 5.16 is a closeup of the image line $\mathscr{L}_2$, showing how the image dots "converge" to a point corresponding to the horizontal line through O (which corresponds to the point at infinity on $\mathscr{L}_1$).

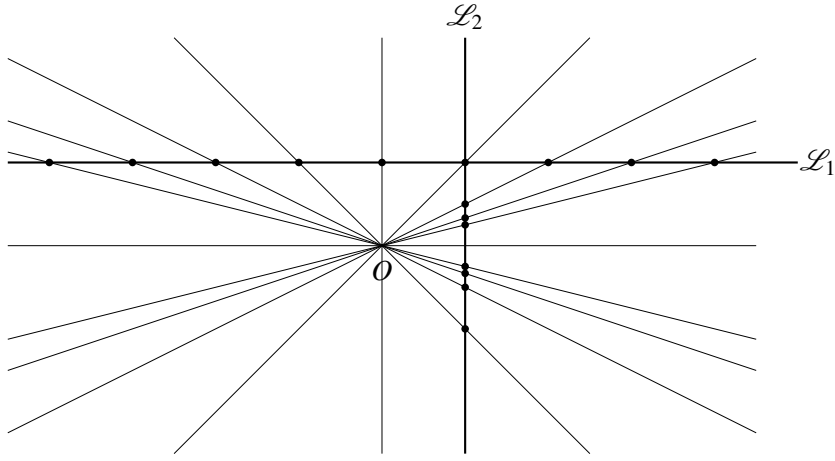

Figure 5.15: Example of projective distortion of the line

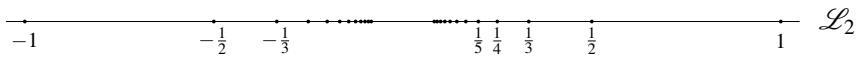

Figure 5.16: Closeup of the image line

We take $O = (0,0)$ as usual, and we suppose that $\mathscr{L}_1$ is parallel to the x-axis, that $\mathscr{L}_2$ is parallel to the y-axis, and that the dots on $\mathscr{L}_1$ are unit distance apart. Then the line from O to the dot at $x = n$ on $\mathscr{L}_1$ has slope $1/n$ and hence it meets the line $\mathscr{L}_2$ at $y = 1/n$. Thus the map from $\mathscr{L}_1$ to $\mathscr{L}_2$ is the function sending x to $y = 1/x$. This map exhibits the most extreme kind of distortion induced by projection, with the point at infinity on $\mathscr{L}_1$ sent to the point $y = 0$ on $\mathscr{L}_2$.

Any combination of these projections is therefore a combination of functions $1/x$, kx, and $x+l$, which are called *generating transformations*. The combinations of generating transformations are precisely the functions of the form

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

that we study in the next section.

Exercises

Before studying all these functions, it is useful to study the (simpler) subclass obtained by composing functions that send x to $x+l$ or kx (for $k \neq 0$). The latter functions obviously include any function of the form $f(x) = ax+b$ with $a \neq 0$, which is the result of multiplying by a, and then adding b.

5.5.1 If $f_1(x) = a_1x+b_1$ with $a_1 \neq 0$ and $f_2(x) = a_2x+b_2$ with $a_2 \neq 0$, show that

$$f_1(f_2(x)) = Ax+B, \quad \text{with} \quad A \neq 0,$$

and find the constants A and B.

5.5.2 Deduce from Exercise 5.5.1 that the result of composing any number of functions that send x to $x+l$ or kx (for $k \neq 0$) is a function of the form $f(x) = ax+b$ with $a \neq 0$.

We know that such functions represent combinations of certain projections from lines to parallel lines, but do they include *any* projection from a line to a parallel line?

5.5.3 Show that projection of a line, from any finite point P, onto a parallel line is represented by a function of the form $f(x) = ax+b$.

5.6 Linear fractional functions

The functions sending x to $1/x$, kx, and $x+l$ are among the functions called *linear fractional*, each of which has the form

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

The condition $ad-bc \neq 0$ ensures that $f(x)$ is not constant. Constancy occurs only if $ax+b = \frac{a}{c}(cx+d)$; in which case, $ad-bc = 0$ because $\frac{ad}{c} = b$.

By writing

$$\frac{ax+b}{cx+d} \quad \text{as} \quad \frac{ax+\frac{ad}{c}+b-\frac{ad}{c}}{cx+d} = \frac{\frac{a}{c}(cx+d)+\frac{1}{c}(bc-ad)}{cx+d}$$

we find that any linear fractional function with $c \neq 0$ may be written in the form

$$f(x) = \frac{a}{c} + \frac{bc-ad}{c(cx+d)}.$$

Such a function may therefore be composed from functions sending x to $1/x$, kx, and $x+l$—the functions that reciprocate, multiply by k, and add l—for various values of k and l:

- first multiply x by c,
- then add d,
- then multiply again by c,
- then reciprocate,
- then multiply by $bc-ad$,
- and finally add $\frac{a}{c}$,

and the result is that x goes to $\frac{a}{c} + \frac{bc-ad}{c(cx+d)} = \frac{ax+b}{cx+d}$. When $c = 0$, the linear fractional function is simply $\frac{ax+b}{cx+d} = \frac{a}{d}x + \frac{b}{d}$, and this can be composed from x by multiplying by a/d and then adding b/d.

Thus, *any linear fractional function is composed from the functions that reciprocate, multiply by k, and add l*, and hence (by the constructions in the previous section) *any linear fractional function on the number line is realized by a sequence of projections of the line.*

We now wish to prove the converse: *Any sequence of projections of the number line realizes a linear fractional function.* From the previous section, we know this is true for projection of a line onto a parallel line, so it suffices to find the function realized by projection of a line onto an intersecting line. We first take the case in which the lines are perpendicular (Figure 5.17). This case generalizes that of Figure 5.15, by allowing projection from an arbitrary point (a,b).

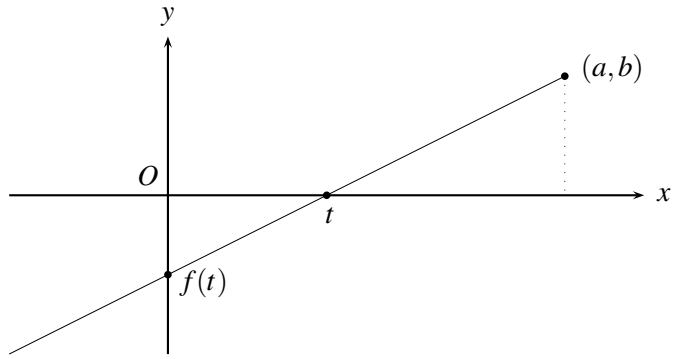

Figure 5.17: Projecting a line onto a perpendicular line

To find where the point t on the x-axis goes on the y-axis, we consider the slope of the line through t and (a,b). Between these points, the rise is b and the run is $a-t$, so the slope is $\frac{b}{a-t}$. Between t and the point $f(t)$ on the y-axis, the run is t and the rise is $-f(t)$; hence,

$$f(t) = \frac{bt}{t-a}, \quad \text{which is a linear fractional function.}$$

For the general case of intersecting lines, we take one line to be the x-axis again, and the other to be the line $y = cx$. Again we project the point t on the x-axis from (a,b) to the other line, and to find where t goes, we first find the equation of the line through t and (a,b). Equating the slope from t to (a,b) with the slope between an arbitrary point (x,y) on the line and (a,b), we find the equation

$$\frac{b}{a-t} = \frac{b-y}{a-x}.$$

This line meets the line $y = cx$ where

$$\frac{b}{a-t} = \frac{b-cx}{a-x},$$

and hence where

$$x = \frac{bt}{ct - ac + b}, \quad \text{which is also a linear fractional function of } t.$$

Thus, any single projection of a line can be represented by a linear fractional function of distance along the line. It is easy to check (Exercise 5.6.2) that the result of composing linear fractional functions is linear fractional. Hence, any finite sequence of projections is represented by a linear fractional function. □

Dividing by zero

You remember from high-school algebra that division by zero is not a valid operation, because it leads from true equations, such as $3 \times 0 = 2 \times 0$, to false ones, such as $3 = 2$. Nevertheless, *in carefully controlled situations*, it is permissible, and even enlightening, to divide by zero. One such situation is in projective mappings of the projective line.

The linear fractional functions $f(x) = \frac{ax+b}{cx+d}$ we have used to describe projective mappings of lines are actually defective if the variable x runs only through the set $\mathbb{R}$ of real numbers. For example, the function $f(x) = 1/x$ we used to map points of the line $\mathscr{L}_1$ onto points of the line $\mathscr{L}_2$ as shown in Figure 5.15 does not in fact map *all* points. It cannot send the point $x = 0$ anywhere, because $1/0$ is undefined; nor can it send any point to $y = 0$, because $0 \neq 1/x$ for any real x. This defect is neatly fixed by extending the function $f(x) = 1/x$ to a new object $x = \infty$, and declaring that $1/\infty = 0$ and $1/0 = \infty$. The new object ∞ is none other than the *point at infinity* of the line $\mathscr{L}_1$, which is supposed to map to the point 0 on $\mathscr{L}_2$. Likewise, if $1/0 = \infty$, the point 0 on $\mathscr{L}_1$ is sent to the point ∞ on $\mathscr{L}_2$, as it should be.

Thus, *the function $f(x) = 1/x$ works properly, not on the real line $\mathbb{R}$, but on the real projective line $\mathbb{R} \cup \{\infty\}$*—a line together with a point at infinity. The rules $1/\infty = 0$ and $1/0 = \infty$ simply reflect this fact.

It is much the same with any linear fractional function $f(x) = \frac{ax+b}{cx+d}$. The denominator of the fraction is 0 when $x = -d/c$, and the correct value of the function in this case is ∞. Conversely, no real value of x gives $f(x)$ the value a/c, but $x = \infty$ does. For this reason, *any function $f(x) = \frac{ax+b}{cx+d}$ with $ad - bc \neq 0$ maps the real projective line $\mathbb{R} \cup \{\infty\}$ onto itself.* The map is also one-to-one, as may seen in the exercises below.

The real projective line $\mathbb{RP}^1$

We can now give an algebraic definition of the object we called the "real projective line" in Section 5.3. It is the set $\mathbb{R} \cup \{\infty\}$ *together with* all the linear fractional functions mapping $\mathbb{R} \cup \{\infty\}$ onto itself. We call this set, with these functions on it, the *real projective line* $\mathbb{RP}^1$.

The set $\mathbb{R} \cup \{\infty\}$ certainly has the points we require for a projective line; the functions are to give $\mathbb{R} \cup \{\infty\}$ the "elasticity" of a line that undergoes projection. The ordinary line $\mathbb{R}$ is not very "elastic" in this sense. Once we have decided which point is 0 and which point is 1, the numerical value of every point on $\mathbb{R}$ is uniquely determined. In contrast, the position of a point on $\mathbb{RP}^1$ is *not* determined by the positions of 0 and 1 alone.

For example, there is a projection that sends 0 to 0, 1 to 1, but 2 to 3. Nevertheless, there is a constraint on the "elasticity" of $\mathbb{RP}^1$. If 0 goes to 0, 1 goes to 1, and 2 goes to 3, say, then the destination of every *other* point x is uniquely determined. In the next two sections, we will see why.

Exercises

The formula $\frac{ax+b}{cx+d} = \frac{a}{c} + \frac{bc-ad}{c(cx+d)}$ gives an inkling why the condition $ad - bc \neq 0$ is part of the definition of a linear fractional function: If $ad - bc = 0$, then $\frac{ax+b}{cx+d} = \frac{a}{c}$ is a constant function, and hence it maps the whole line onto one point.

If we want to map the line onto another line, it is therefore necessary to have $ad - bc \neq 0$. It is also sufficient, because we can solve the equation $y = \frac{ax+b}{cx+d}$ for x in that case.

5.6.1 Solve the equation $y = \frac{ax+b}{cx+d}$ for x, and note where your solution assumes $ad - bc \neq 0$.

5.6.2 If $f_1(x) = \frac{a_1x+b_1}{c_1x+d_1}$ and $f_2(x) = \frac{a_2x+b_2}{c_2x+d_2}$, compute $f_1(f_2(x))$, and verify that it is of the form $\frac{Ax+B}{Cx+D}$.

5.6.3 Verify also that $\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix} \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix}$.

Thus, linear fractional functions behave like 2×2 matrices. Moreover, the condition $ad - bc \neq 0$ corresponds to *having nonzero determinant*, which explains why this is the condition for an inverse function to exist.

5.6.4 It also guarantees that if $a_1d_1 - b_1c_1 \neq 0$ for $f_1(x)$ and $a_2d_2 - b_2c_2 \neq 0$ for $f_2(x)$ in Exercise 5.5.2, then $AD - BC \neq 0$ for $f_1(f_2(x))$. Why?

5.7 The cross-ratio

> You might say it was a triumph of algebra to invent this quantity that turns out to be so valuable and could not be imagined geometrically. Or if you are a geometer at heart, you may say it is an invention of the devil and hate it all your life.
>
> Robin Hartshorne, *Geometry: Euclid and Beyond*, p. 341.

It is visually obvious that projection can change lengths and even the ratio of lengths, because equal lengths often appear unequal under projection. And yet we can recognize that Figure 5.1 is a picture of equal tiles, even though they are unequal in size and shape. Some clue to their equality must be preserved, but what? It cannot be length; it cannot be a ratio of lengths; but, surprisingly, it can be a *ratio of ratios*, called the *cross-ratio*.

The cross-ratio is a quantity associated with four points on a line. If the four points have coordinates p, q, r, and s, then their cross-ratio is the function of the ordered 4-tuple (p,q,r,s) defined by

$$\frac{(r-p)/(s-p)}{(r-q)/(s-q)}, \quad \text{which can also be written as} \quad \frac{(r-p)(s-q)}{(r-q)(s-p)}.$$

The cross-ratio is preserved by projection. To show this, it suffices to show that it is preserved by the three generating transformations from which we composed all linear fractional maps in the previous section:

1. The map sending x to $x+l$.

   Here the numbers p,q,r,s are replaced by $p+l,q+l,r+l,s+l$, respectively. This does not change the cross-ratio because the l terms cancel by subtraction.

2. The map sending x to kx.

   Here the numbers p,q,r,s are replaced by kp,kq,kr,ks, respectively. This does not change the cross-ratio because the k terms cancel by division.

3. The map sending x to $1/x$.

   Here the numbers p,q,r,s are replaced by $\frac{1}{p},\frac{1}{q},\frac{1}{r},\frac{1}{s}$, respectively, so the cross-ratio

   $$\frac{(r-p)(s-q)}{(r-q)(s-p)}$$