Chapter 6

Projective Geometry

Computer graphics models are often in 3D. Display devices are 2D, whether LCD screen or printer, so there has to be a projection from 3D to 2D. Projective geometry is the mathematical subject which studies projections. In computer graphics various kinds of projections are possible. In CAD we may want 1-point, 2-point or 3-point projections. In many other applications, we typically want perspective to indicate depth into the scene. In order to understand this properly, we will concentrate first on perspective.

6.1 Perspective projections

What is the key property that tells us we have a perspective projection? Most people think of vanishing points as somehow important for perspective, which is true but they are a side-effect of something more basic. First, think of what happens to the apparent size of an object as it moves further away from us. It becomes smaller. Perspective has the curious property that the scale varies linearly, depending on where the object is in the space. It is this variable-scale property that is fundamental to perspective. Vanishing points are a side-effect of this because the distance between projected parallel lines shrinks as the lines go away from us. Eventually the lines "meet" at a vanishing point.

6.2 Projective geometry

Conventional geometry is called Euclidean geometry. There are certain axioms which define it but it has an inconsistency which even the ancient Greeks who formulated it knew about. One axiom is that any two distinct lines define a point (i.e. where they intersect), *unless* the lines are parallel. In contrast, any two distinct points define a line, with *no* exceptions. Projective geometry copes with these two cases symmetrically: no exceptions.

The Axioms of projective geometry, in essence, are:

Any two distinct points define a line
Any two distinct lines intersect at a point.

The emphasis here is on the word *any*: there are no exceptions for parallel lines; nor indeed for infinite points. Let's try to make some sense of this and see how it relates to computer graphics. We have already mentioned two issues which are at the heart of this: vanishing points are where infinitely long parallel lines in the model appear to "meet" when projected to the image plane; and scale varies linearly with distance. It is points at infinity which Euclidean geometry does not address: it cannot explain parallel lines meeting because they do not meet at a finite point. It cannot explain parallel lines getting closer because Euclidean geometry has no sense of scale.

We get to projective geometry very easily: take Euclidean geometry and add an extra dimension which defines the scale. If we start with Euclidean 3-space (x, y, z), we add an extra dimension to get a 4-space (x, y, z, w). If we put $w = 1$, we get Euclidean space; scale factor 1 everywhere. If we put $w = 2$ (say), we still have a Euclidean space but the space is twice as big. We cannot use $w = 0$ this way because we cannot distinguish distances if the scale is zero. We can have negative w values though.

How does this help? Suppose we want to model a cube "in perspective". We can put the front face of the cube at $w = 1$ and the rear face at $w = 2$. We can then project the rear face to the Euclidean space, at $w = 1$, by dividing all coordinates by w: $(x/w, y/w, z/w, 1)$. This *normalisation* to $w = 1$ has the effect that the new coordinates of the rear face will be half the scale of those of the front face. In between, the scale varies linearly (which is exactly why we can draw the cube edges by joining the coordinates with straight lines and not have to calculate every point on the line). This gives perspective, produced via a central projection (the normalisation process is projecting towards the origin).

By the way, this is not just an abstract idea. For example when making a diorama for a museum display, model builders will sometimes make buildings etc with the perspective built in to the model. Buildings towards the back of the diorama are constructed at a smaller scale than those towards the front. This gives an apparently large scale scene in a small volume. Theatrical sets sometimes exploit this too.

More generally, we can make w a function of x, y and/or z i.e. $f(x, y, z)$. For example suppose we make $w = z$. Any point is now effectively (x, y, z, z), so when we project (normalise on w) we get $(x/z, y/z, 1, 1)$. So the Euclidean point is $(x/z, y/z, 1)$. This is the familiar perspective projection and it shows we have projected to the $z = 1$ plane: our screen.

Of course there is nothing special about z and we could have included perspective in the x or y directions; or all three at once, by a suitable choice of function. (Since perspective maps straight lines to straight lines, this function should be a linear one: it defines a line in Euclidean 1D, a plane in 2D and a volume in 3D.) This is new: we cannot do this in the Euclidean world because we need the "extra" projective dimension to make this work.

To summarise, projective geometry allows us to vary the scale of the space at every point, by choosing suitable w values. We can project from 4-space back to Euclidean 3-space, by dividing by w. This is a central projection, a projection towards the origin.

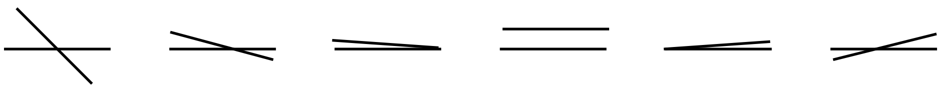
One thing we have not done: we did not project to 2D. We ended in 3D yet we claimed that we had a perspective cube. First notice that we did project from 4D to 3D, so we reduced the dimensionality. To do that projection, we exploited our "new" w dimension. In a real sense, we have used the varying scale to build a 3D model which is already in perspective. We are perhaps used to thinking of perspective as "divide-by-z", to get from 3D to 2D, yet there clearly should be nothing special about the z dimension. Projective geometry treats x, y and z equally: it is the extra w dimension which is different because we use it as a scale factor. The practical benefit is that we can set the perspective in each of the x, y or z dimensions independently, just the thing for computer graphics. The w dimension allows us to define the function which says how these three scales should vary. Perspective is not really a function of projecting to a screen. How do we get from this "perspective" 3D to 2D on screen? Easy, it's just a parallel projection: we simply drop the extra dimension. (In practice we may want to scale and translate as well, to suit the coordinates of our display device.)

We mentioned earlier that projective geometry can cope with vanishing points, corresponding to parallel lines "meeting" at infinity. This might seem a strange claim, given that all we have done is add a scale factor to Euclidean geometry. In fact there is another way of describing projective geometry which brings out this aspect better. Start by thinking of the 4D coordinates like this: the (x, y, z) part defines a *direction* and the w part defines the *depth* or scale along that direction. Now suppose we take all the finite points, which correspond to Euclidean geometry, and add points at infinity in all possible directions. The result is projective geometry.

Let's unpick this. When we said two distinct Euclidean parallel lines do not meet, we meant that no matter how far we go they are still the same distance apart; so they cannot meet. These two Euclidean lines have no point in common precisely because they are parallel. But suppose we ask what they do have in common: the answer is they share the same direction. There are no infinite points in our Euclidean space so we deliberately now add an infinite point in the same direction as the parallel lines. We then define this infinite point as what they have in common. Suppose we do this for all possible directions. We can now say that any two distinct lines have exactly one point in common. For non parallel lines, this is where they intersect. This intersection point has a direction and depth/scale. For parallel lines, it is the corresponding infinite point in their direction. This point also has a direction and depth. This eliminates the axiomatic problem with Euclidean geometry: projective geometry finally solved the ancient Greeks' problem.

An infinite point and a finite point define a line (because we know the direction of the infinite point and a finite point that it goes through). Two distinct infinite points define a line, the *line at infinity*, which contains all the infinite points (and no others). So we see that the Axioms still work, even now we have included infinite points. A set of parallel lines have in common their *direction* and it is this direction that we are recording. The image of this point on the viewing plane is the vanishing point for that direction: even though the point is infinite, it has a finite projection.

There are two further aspects worth mentioning. First, there is only one infinite point for each direction. If we are careful to remember that, we can informally think that the infinite point is where the parallel lines "meet", just to echo what happens with non parallel lines. There is also an easy way to convince yourself that there is only one such point. Consider a pair of lines that intersect, and slowly rotate one line about a fixed point:



The point of intersection of the two lines moves off the the right, until the lines are parallel, when they intersect at a point at infinity. Continuing the rotation, the intersection moves in from the *left*, thus indicating the point at infinity is both to the left and the right. There is only one configuration where the intersection is at infinity, so there can only be one infinite point.

The second aspect is that we seem to have given ourselves a big headache: infinite depths are hard to represent and compute with. A really nice surprise is that our extra w dimension lets us represent the infinite points, with no need for infinite values. The reason is hidden away in my earlier throw-away remark:

"We cannot use $w = 0$ this way because we cannot distinguish distances if the scale is zero."

That's true but if our parallel lines meet "at infinity", there is no distance between them anyway. So we can use $w = 0$ to represent the infinite points. A point like $(x, y, z, 0)$ is the infinite point in the direction (x, y, z). That's all we need: if w is any other value, it represents a finite point at that scale. Something else drops out of this: all points with the same x, y and z, whether finite or infinite, project to the same Euclidean point $(x, y, z, 1)$. They are all in the same direction. Imagine yourself at the origin looking in the direction (x, y, z). You will be looking at the line (wx, wy, wz, w) end-on, so it looks like a single point. No matter how much it is scaled, it still looks the same from the origin. Once again, we do not need to treat the infinite point different from the finite ones.

Finally the origin is special. The point $(0, 0, 0, 0)$ has no direction. In projective geometry we have to exclude the origin (but no other point). This is a small price to pay for including all those infinite points and making the ancient Greeks happy.

6.3 The Mathematics

The mathematics is now quite easy. We will need transformations which are 4 by 4, so we need an extra row and column on our earlier Euclidean transformations. The extra dimension is w; the other three are our familiar x, y and z. Here it is:

$$\begin{pmatrix} r_{1,1} & r_{1,2} & r_{1,3} & v_1 \\ r_{2,1} & r_{2,2} & r_{2,3} & v_2 \\ r_{3,1} & r_{3,2} & r_{3,3} & v_3 \\ t_x & t_y & t_z & s \end{pmatrix}$$

To be a projective transformation, this matrix has to have an inverse. You can see that it simply extends the Euclidean transformation. The top left 3 by 3 matrix is the usual Euclidean compound rotation we met earlier. The three t elements allow translation, which is the standard textbook reason for having the 4 by 4 matrix. The bottom right s is the overall scale factor, if we wish to rescale in the w dimension. Often this is just set to 1. The three v elements are the perspective slopes in the x, y and z directions. (Why?) Since they are slopes, their inverses are the vanishing points. A slope of zero means no perspective, just a parallel projection in that dimension.

If we imagine the Euclidean plane as a disc (or the xyz space as a sphere) with us at the centre, the extended plane is the original disc plus a new line of points at infinity forming a circle around the edge (or a spherical shell of infinite points). Mathematically, these are called *ideal points*, while the points in the Euclidean disc are *affine* points. The ideal points form a line, the *line at infinity*. The whole thing is called the *projective plane*, and is the basis of projective geometry.

6.4 Homogeneous coordinates

As we have just seen, the extra dimension allows us to combine rotation and translation into a single matrix. In many computer graphics text books, projective geometry is introduced via these homogeneous coordinates, as a convenient way to bring rotation and translation into one matrix. As we now understand, that is a side effect of the space being projective.

It is the extra dimension which makes the mathematics look “homogeneous”: all four dimensions appear in the matrices in an identical way. Yet the w dimension does not serve the same purpose as the other dimensions. So what’s homogeneous about homogeneous coordinates? Surely w is special, which sounds very inhomogeneous? Yes it is different, in the sense that we treat it as a scale factor and may divide by it at the end. However if we look at the algebra instead of the geometry, we can see where the term comes from.

The solution of simultaneous linear equations is the link between projective geometry and the homogeneous matrix form. It also shows us why we have a homogeneous appearance even though one dimension is different. Why should simultaneous linear equations be relevant? Linear equations are, as the term says, the equations of lines. Two equations have a simultaneous solution if the lines intersect; and no solution means no intersection. Does this sound familiar?

The equations for two lines can be written

$$p_1x + q_1y + r_1z + s_1 = 0$$

$$p_2x + q_2y + r_2z + s_2 = 0$$

If we solve these simultaneously, we can find the point where they intersect. Suppose though that these lines are parallel. This means that $s_1 \neq s_2$ because we mean *distinct* lines. We must have $p_1 = p_2$, $q_1 = q_2$ and $r_1 = r_2$, to make them have the same slope. But this means our solution needs $s_1 = s_2$, which is the opposite of what we just decided. This contradiction is, algebraically, the equivalent of saying we have no Euclidean place where these lines cross.

[You may have noted that, in 3D, lines which are not parallel may still not intersect: we also have to assume the lines are co-planar because that is the only way we can have parallel lines.]

Things change for the better if we make the equations *homogeneous*, with all terms associated with an axis:

$$p_1x + q_1y + r_1z + s_1w = 0$$

$$p_2x + q_2y + r_2z + s_2w = 0$$

This is where the term "homogeneous" comes from. We note the following interesting features. If $w = 1$, these equations become the Euclidean ones we had a few lines ago. So these new equations will solve all the cases where the lines cross at a finite position. If the two lines are parallel, then we now have

$$(s_1 - s_2)w = 0$$

which has a solution, $w = 0$, which does not require that $s_1 = s_2$. So our parallel lines remain distinct but they "meet" at $w = 0$. Since we already know that we have all the finite solutions at $w = 1$, it follows that $w = 0$ includes the infinite solutions (and in fact only the infinite solutions). The resulting values of p, q and r, being slopes, tell us the *direction* of the point and the $w = 0$ tells us it is infinite. Even the finite solutions can be thought of this way, since (wx, wy, wz, w) effectively tells us the point is at (x, y, z) but scaled by w; so it is in the direction (x, y, z). This is just like using a vector to define a point or to define a direction.

It isn't difficult to see that all values of w except $w = 0$ also produce finite solutions: they are just scaled versions of $w = 1$. In fact this follows directly from the homogeneous equations because

$$kp_1x + kq_1y + kr_1z + ks_1w = 0$$

$$kp_2x + kq_2y + kr_2z + ks_2w = 0$$

will give the same answers for any k, except $k = 0$.

6.5 Some Projection Thought Experiments

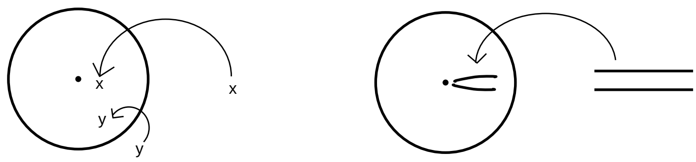
The rest of this Chapter introduces almost no new material. It is provided as **reading material** to help you deepen your understanding.

6.5.1 Projection and Perspective

It helps to get a feel for what we mean by perspective, which is what gives us vanishing points in the first place. Previously we have thought of it as "divide by z", which is not very informative. Let us think for a while about perspective obtained by projecting onto a unit radius sphere (instead of onto a plane). The essence of spherical perspective is that coordinates at Euclidean distance r are transformed to be at $1/r$. Thus we are mapping $r \mapsto 1/r$, which has a stationary point at $r = 1$. Notice however that the transformation $r \mapsto 1/r$ maps *all* of space to *all* of space: by itself, it does not require any 3D to 2D calculation. Let's explore this.

Suppose we have a sphere of radius 1, centred on the origin in the "real" world. For every point outside the sphere at distance r there will be a corresponding point within the sphere at distance $1/r$. The mapping constructs a model of the entire outside world within the sphere.

Points on the surface map to themselves (this is what we mean by a stationary point). Points just outside will map to points just inside. Points at a great distance will map to points close to the centre, and points further and further away in the real world will bunch together progressively in the perspective world. If we were to map pairs of points, unit distance apart, we would find that their separation inside the sphere depended on how far away they were when we started. It is as though the scale of the world, originally nice and linear, is being changed as a function of r. Thus parallel lines in the real world will converge in some way in the sphere.

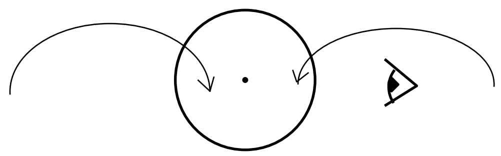


Now there is no preferred direction in this sphere: every direction is equal. If we imagine ourselves outside the whole system but looking into the sphere, we see the world in some kind of perspective. This means that we actually have a perspective model packed into the sphere, it's not just that it *looks* like it's in perspective. The perspective is certainly affected by the choice of origin.

The *steepness* of the perspective is built-in, too. This is the measure of how fast parallel lines converge in the perspective world. With a steep perspective parallel lines meet faster, i.e., at a wider angle. The steepness of the perspective map can be controlled by using p^2/r instead of $1/r$, where p controls the steepness. This is clearly still a perspective map, but the stationary points are now at $r = p$, so we have a sphere of radius p. Changing this *radius of inversion* thus changes the steepness of the perspective.

Moving towards the centre of the sphere takes us out towards infinity in the real world. Looking towards the centre is like looking out to infinity. All outward directions in the real world meet at the centre of the projected world. In particular, points which are infinitely far away, in any direction, will map to the centre of the sphere. (Note however that this is a many-to-one mapping, which is one reason for excluding the origin in our earlier discussion: we cannot tell from the 3-space coordinates from which direction the sphere origin mapped.)

But there's still more to the sphere than that. If we imagine ourselves at the centre of projection, facing in any particular direction, then a section of the sphere corresponds to the part of the real world *behind* you: all directions are mapped into the sphere.



If we look at a point on a diameter then, as we move it towards the centre it corresponds to a point moving further and further away in some direction: let's say it heads off to plus infinity. Once we hit the centre and pass through it, we find a point which corresponds to minus infinity; that is, to infinity in exactly the opposite direction. So points go off to infinity in one direction and come back from another direction.

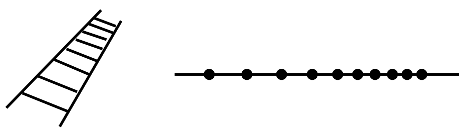
We have squashed most of the real world into a sphere of unit radius. What happened to the part of the real world that the sphere itself occupied? Well, that is mapped to *outside* the sphere. In fact, it is mapped to fill the entire rest of the universe. The single point that was at the centre of the sphere is mapped off to infinity in all directions: it becomes an infinitely distant spherical shell (again, this is not a one-to-one mapping). Thus what we have is not just a mapping on the sphere, it is a mapping of the entire universe to itself such that points on the sphere of radius 1 map to themselves. In this mapped universe, perspective is "built in."

It is however a curved sort of perspective, because of the spherical projection. For example, a line tangential to the sphere will map to a circle inside the sphere, touching the line, the sphere and passing through the centre. Now here is a curio: the projected curve is closed. So the original (infinite) line must also be closed: in fact, it closes at infinity and plus and minus infinity are the same point! If we travel to plus infinity and keep going, we re-emerge from minus infinity. This also works with the projected world we have just constructed: the centre of the sphere, distance zero, corresponds to all the infinite points outside the sphere.

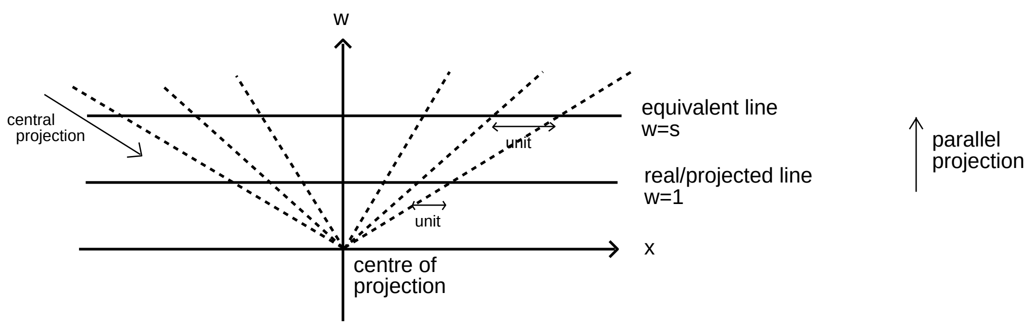
6.5.2 One Dimension

It is not immediately clear how to project in a 1D space: this is just a line. How can we tell if the points on the line are in perspective or not?

We can tell by looking at how the spacing of individual points of the line changes. We take a line in the real world and watch points that are a unit distance apart. Projecting this to a perspective line we expect that the points are no longer uniformly spaced, but cluster together in the distance. Of course, this is true of all points on the line, but it helps just to think of the integer points.



When we talk of projection, we imply that we are projecting from one thing to another, in this case from one line to another.



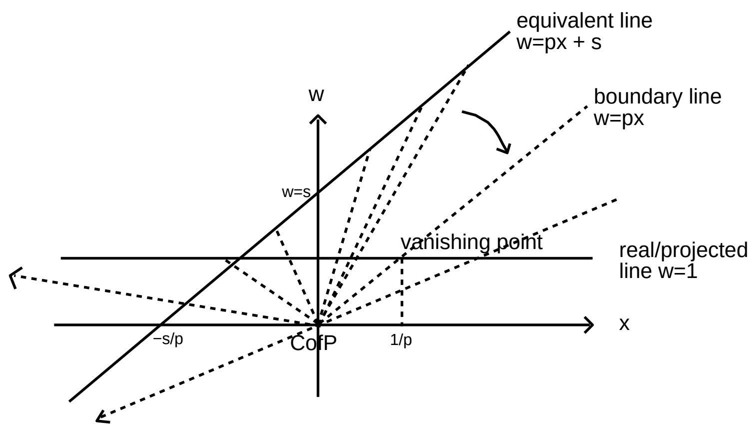
In this picture we have added a dimension – w – that is not one of the normal xyz dimensions, in order to separate the original line and the projected line.

We start with the real line which lives in this diagram at $w = 1$. Parallel to it a little way above is what we will call the *equivalent line*, which is just a copy of the real line but at $w = s$. Thus we *parallel project* the real line to the equivalent line. (We are *embedding* the real line in this higher dimensional projective world.) This is like saying we created our object in a space where $w = s$ instead of at $w = 1$. In other words, the scale is now different.

Now we project the equivalent line back to the real world, but this time use a *central projection*: the normalisation process which is divide by w. In particular, we can imagine how the integral points on the line map, as indicated by the dotted lines. This fan of lines is just another manifestation of the idea that any given segment (between two dashed lines say) can be represented at an infinite number of scales (as we change w). Indeed, the fan should go down as well as up to cover negative values of w, where the order of the points is reversed.

The result is that we end up with a version of our original line, scaled to $1/s$. If we move the equivalent line, i.e., change s, the scale of the space changes. We can choose $s < 1$ for a magnification, $s > 1$ to shrink or even $s < 0$ for an inversion of the line, where plus and minus infinity are swapped. (Note the similarity with our earlier projective sphere thought experiment.)

We have not yet produced any perspective effect, just an overall scaling. To do this, we skew the equivalent line so that is is not parallel to the real line. As we move along any such equivalent line, the scale factor w now varies.



We do the same for this diagram, namely take the real line, project on to the equivalent line, and centrally project back. Now the scale factor of the central projection varies linearly along the projection, which is exactly what we want for perspective.

Notice we have produced a world with the same dimensionality as we started from: we have made a 1D perspective world from a 1D non-perspective one.

This is where homogeneous coordinates come in: the w allows us to perform perspective projections in a higher dimensional space, because the scale factor can vary at each point. Another way to think about this use of w is to say we work with "ordinary" coordinates except that every point in space has an additional scale factor, w, attached to it. The use of an equivalent line in the above is thus a way of saying we want the w scale factor to vary steadily, which is just what we need in perspective.

Understanding this space (1)

Each radial line in wx space corresponds to a single point in our original x space. Put another way, a line in projective space is equivalent to a point in Euclidean space. The direction of the line tells us which Euclidean point generated it. With the extra w dimension, a Euclidean point is represented at all possible scale factors. Any line radial (in wx) from the centre of projection corresponds to a single Euclidean (x) point at all different scale factors (w).

At some stage, we choose the scaling we want. If we choose $w = s$, then we are asking for a constant scaling. But if we choose $w = pq + s$ then we are asking for an oblique slice through the scale space. As this slice is a straight line, we are asking for scale to vary linearly, as it does in perspective.

The scale factor w is the scale factor of the space, not of the object. If we place a Euclidean object in the projective space at *large* w it will appear *small* when projected. As a result, w is the inverse scale factor as far as the finished picture is concerned. There is a simple analogue here. Suppose you are asked to draw a line 100mm long on the surface of an inflated balloon. If the balloon is well inflated, when you let the air out the line shrinks more than it would if the balloon were only slightly inflated. The large balloon is large w and results in a shorter line on the deflated balloon. Letting the air out is projecting back to $w = 1$.

We can also think about what it means to have the *same* length after projection, regardless of w. Imagine yourself at the origin of the projection diagram, looking in the positive w direction. If you look at a line segment which falls between two adjacent dashed lines, you will not be able to tell how far away in the w direction it is. If it is twice as far away, it must be twice as big and thus looks unchanged in size at the origin. Put another way, in "ordinary" xyz space, we cannot tell which w is being used. This has to be the case because we built $xyzw$ space by adding a dimension, not by modifying the original three. So line segments which subtend the same angle at the origin, regardless of w distance, must be bigger when w is bigger.

That allows us to say all of these segments are equivalent, except in scale: they will all project to be the same size at $w = 1$ (or any other plane of constant w). At the origin, they all project to be zero size (and our ability to measure within them collapses). One more way of thinking of this is simply to imagine that w tells us the amount of stretch in the world we create, but we allow it to shrink back to its normal size $w = 1$ when we want to look at it. If we place an object at *large* w, it will shrink to something *smaller* when we go back to $w = 1$. If we place it at small w, it won't shrink as much and, when $w < 1$, it may even grow. Just to hammer it home then, the further away an object is, the more it appears to shrink, so the larger the w we need to produce that effect.

At first glance you might think you could have drawn the diagram with y instead of w but that would not have done the trick: we will shortly want to repeat this in xy space. Do I hear you say, "use z then"? Fine, but I will also want to do this in xyz space. Do I hear you say, "use w then"? I agree: that's what I just did!

We can choose to place our real line at any non-zero value of w. The only change is the amount of overall scaling going on. Thus we usually choose $w = 1$ so the projected line has the same scale as the original. If we use some other value, there will be an *overall* scaling of the projection.

The value $w = 0$ corresponds to the real line passing through the centre of projection, when we cannot get a sensible projection which distinguishes the lines. This is why we earlier required w was not zero. In fact we only have to exclude the origin; all other points in this projective world have a meaning.

Vanishing points (1)

The number p gives the steepness of the perspective projection.

In the above diagram, there is a line which we called the *boundary line*. Imagine how a radial line behaves as it moves to the right along the equivalent line. The real points it corresponds to cluster closer and closer together until the radial line meets the slope of the equivalent line. The point at infinity of the equivalent line projects back to the *vanishing point* on the real line at $x = 1/p$. This is the intersection of the line $w = px$ (passing through the origin, where w is zero and all the lines meet) and the projection plane $w = 1$. We thus have a *one point* perspective (one vanishing point). Note this is exactly what we say informally: a vanishing point is where lines meet "at infinity".

If, now, we continue to move the radial line to the right, the real points now move along the real line from *minus* infinity towards zero. Again we find plus and minus infinity are the same place. Finite parts of the equivalent line have corresponding points of the real line. The vanishing point is also finite but comes from the projected infinite point.

Summary

Summarising, we

- start with a physical line
- embed it in the higher dimensional xw space at $w = 1$
- create the equivalent line at an angle appropriate for the required steepness of perspective, and parallel project the real line on to it
- we centrally project the equivalent line back on to the real line
- we go back to x space by dropping the w coordinate (in general, dividing by w).

Since we like matrix representations for neatness, this process can be written in matrix form.

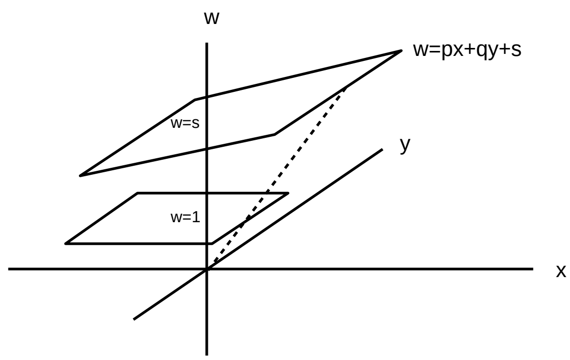
$$(x\ 1)\begin{pmatrix} 1 & p \\ 0 & s \end{pmatrix} = (x\ px + s) \stackrel{\text{normalise}}{\longrightarrow} (x/(px + s)\ 1).$$

Thus the line at $w = 1$ maps to the line $w = px + s$. We can choose $s = 1$ to avoid an overall scaling. If we now centrally project back, this is just normalising w, giving the desired perspective.

We will generalise this idea to 2D, and then 3D.

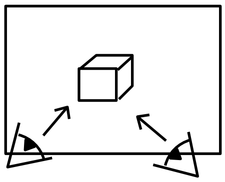
6.5.3 Two Dimensions

We follow precisely the same route as for the 1D case. Now we start with an xy space and add a dimension to get xyw space. The real plane is $w = 1$, which is parallel projected on to the equivalent plane $w = px + qy + r$. Notice that we now have two parameters to control steepness of perspective, one for each of the x and y directions. So we will also get two vanishing points, one for x and one for y. This is the *two point* perspective. If we happen to choose, say, $q = 1$, we only get perspective in the x direction, and only one vanishing point: this is a *one point* perspective. We also have an overall scale factor r.



Notice we have produced a world with the same dimensionality as we started from: we have made a 2D perspective world from a 2D non-perspective one.

There is nothing here to fix the point of view of the viewer, though the centre of projection is the only place where the projection will appear correct. This is quite typical: we often draw perspective pictures of 3D objects on 2D paper, and these pictures will only look technically correct when the viewer is directly in front of the paper and at the correct distance.



The matrix form is a simple extension of the 1D case.

$$(x\ y\ 1)\begin{pmatrix} 1 & 0 & p \\ 0 & 1 & q \\ 0 & 0 & s \end{pmatrix} = (x\ y\ px+qy+s) \stackrel{\text{normalise}}{\longrightarrow} (x/(px+qy+s)\ y/(px+qy+s)\ 1).$$

As before, choose $s = 1$ to avoid overall scaling. The division to normalise is the central projection.

6.5.4 Three Dimensions

We can't draw a picture for this case, but we can proceed with the algebra. The projection plane is $w = px+qy+rz+s$, and we can have up to three vanishing points. As before, we have to choose a viewpoint (the centre of projection) to ensure that the projected object we see is "correct."

Think of what is happening here when we draw a picture of the result: a 3D model is embedded in a 4D space, with the extra dimension w giving us the chance to vary the scale. This is projected back to a 3D "perspectivised" version. Notice we have produced a world with the same dimensionality as we started from: we have made a 3D perspective world from a 3D non-perspective one. To view this, we can parallel project to a 2D representation to view on the screen. (We can choose any viewing direction but we cannot now change the centre of projection because we have built that into our projected world.)

The matrix form is a direct extension of our previous results:

$$(x\ y\ z\ 1)\begin{pmatrix} 1 & 0 & 0 & p \\ 0 & 1 & 0 & q \\ 0 & 0 & 1 & r \\ 0 & 0 & 0 & s \end{pmatrix} = (x\ y\ z\ px+qy+rz+s)$$
$$\stackrel{\text{normalise}}{\longrightarrow}$$
$$(x/(px+qy+rz+s)\ y/(px+qy+rz+s)\ z/(px+qy+rz+s)\ 1).$$

Now p, q and r determine the steepness of the perspective in their respective directions, while s controls the overall scaling.

You can finally see why, much earlier, we showed the standard range of homogeneous rotation, translation and scaling transformations with a unit vector in the right-hand column. This is the (4-space) plane onto which we project, with no perspective effect and unit overall scaling. We finally have an explanation for all 16 of the elements in the matrix: the righthand column gives us full control of the perspective effect.

Understanding this space (2)

Here is an alternative explanation of the space and of vanishing points, a bit more in terms of everyday reality.

A key feature of projection is that parallel lines converge. That is, lines which were parallel before projection converge after projection, to a vanishing point.

We can think of the vanishing point as the image of the point "at infinity" where the parallel lines go off to. The vanishing point is always in the same direction as the parallel lines. If we stand in the middle of a flat infinite plane, we will see the horizon. It will have sky above and ground below. The horizon is all the vanishing points, one point for each radial direction from where we are standing.

Even though we think in (x, y, z) coordinates, there is nothing special about any direction. When we are making pictures however, there are conventions for how to draw. In technical drawing, we talk of one-point perspective, or two-point or three point. These terms refer to the directions of the principal axes x, y and z. Often we do not draw in "full" perspective, partly because it is easier not to but also because it can help understanding.

In our everyday world, all parallel lines appear to converge. We cannot have a one-point perspective, simply because any parallel lines extend in space and so will look closer together at some point. Ideally we would like to control the perspective of our pictures in full, so that we can make a one-point or a two-point or three-point picture. It's not too difficult to understand what we need, if we think of an overhead projector. After all it projects and projection is what we need!

A correctly set-up overhead projector projects a square area (the glass top of the projector) onto a square area of the screen. All that has changed is the *scale*; hopefully, it is bigger. However, if I rotate the projector's overhead mirror, the image moves up (or down) the screen. It will no longer be square. In fact the two "vertical" sides will now slope: they will be converging towards a vanishing point. If instead I had rotated the projector to one side, I would have made the "horizontal" sides converge towards a vanishing point. By making both adjustments I can produce a projected image where there will be both vanishing points: no edges are parallel any more.

So it seems I have everything under control: I can arrange the projection to give me zero, one or two vanishing points. As my object (the surface of the projector) is 2D, that is all I can ask. Surely all I now have to do is imagine some kind of 3D overhead projector and we are done? That's true enough but we have to be careful what we mean by that.

In order to control our 2D vanishing points, we had to position our projector (more accurately, both the Centre of Projection/CoP and its direction) in a 3D world. It was this "extra" dimension that allowed us to achieve control of the vanishing points. We can make a projection purely in 2D and it will indeed show perspective. We can even control the perspective rate (how quickly things get smaller with distance) but we cannot control the vanishing points independently. Just imagine placing a square in a plane, then projecting its corners towards a CoP (still in the plane), with an intervening straight line (again, in the same plane) as the "screen". The points projected on the line will show perspective: edges which are further away will be shorter on the "screen". We can control the perspective rate by moving the "screen" nearer the CoP (or vice versa) but this is only one parameter so it can only make a global change to the perspective. An extra dimension allows us to guarantee to separate the CoP from the image plane and from the object. Without that, the projection is not well-defined. In particular we cannot project the CoP itself, so that point must always be excluded.

If we want to control a 2D projection, we need to do the projection in a 3D coordinate system. If we want to control a 3D projection, we will need a 4D coordinate system (x, y, z, w). As with the 2D case, we can still get a perspective projection from 3D without using an extra dimension: it's the one we see every day. However we will not be able to control the principal axis vanishing points independently.

Vanishing points (2)

At this point it's worth thinking what a projection is doing. If I imagine myself at the centre of projection, then the projection directions come radially in towards me. I will be able to see these directions as lines but I will see each line as a single point: I'm looking at the line end-on. The practical consequence of this is that there are infinitely many points which will project to a single, specific point. These are all the points in a line of sight radiating from/towards the CoP. However, we can note that they are all scaled versions of the same point. That is, if the point in question is (x, y, z), then all the points which project to it are on the line (wx, wy, wz). In fact you, with your global view of the scene, will be able to tell me which value of w a particular point has, whereas I, looking at the projection, can't tell that by observation. So you will be able to identify it as (wx, wy, wz, w). We will both agree that the direction is (x, y, z).

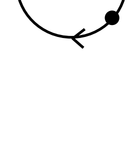
Now suppose we keep the point's 3D coordinates fixed and vary w. That is, we consider (x, y, z, w) as w alone varies. What happens to the point projected onto the $w = 1$ plane? The projected point will be at $(x/w, y/w, z/w, 1)$. As w gets smaller, this projected point goes further away. It moves in the direction (x, y, z). If w approaches zero, then the projected point approaches infinity. We can therefore think of any point $(x, y, z, 0)$ (note the zero) as being the point at infinity in the direction (x, y, z). Such a point is known as an *ideal point*. If we cover all possible directions, we will sweep out a kind of sphere "at infinity", a shell of infinite points which are not in Euclidean geometry. They will all have $w = 0$. How do we know they are not Euclidean? It's easy if we consider a pair of parallel lines. In our projective space, both lines are in the same direction, so they reach the *same* point at infinity (i.e. the same ideal point). In Euclidean space, parallel lines never meet. So these ideal points cannot be present in the Euclidean space.

The projection of an ideal point will be the *vanishing point* in that direction: that's what a vanishing point is, the image of a point at infinity, such as the point where two parallel lines meet. How can we project an ideal point back to $w = 1$? Certainly we cannot do this by dividing by zero but we note an interesting fact. The projection direction must be parallel to the direction of the ideal point, because the projection comes from the ideal point. So its direction is just (x, y, z) and that direction intersects our $w = 1$ plane at $(x, y, z, 1)$. All lines parallel to this line in the 3D space will also reach the same point: parallel lines converge at the vanishing point.

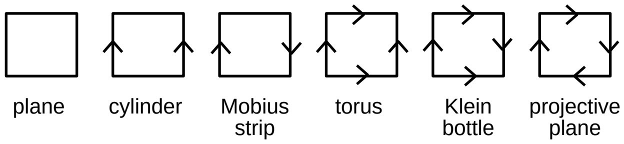
Finally, we have expressed everything so far in 3D or 4D. There is no mention of the 2D space of our screen. What we do in practice is use the 'extra' dimension to establish whatever vanishing points we want. At that point we have a 3D world with the desired perspective built-in. Then we parallel project the 3D version to a 2D plane, possibly scaling, and we are done. That bit at least is easy!

6.5.5 Relation to other Geometries

Here is a small but fun diversion to indicate why the projective plane is related to more familiar "shapes".



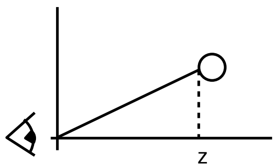
You can in principle make models from paper:



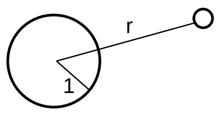
Glue the edges of the piece of paper in such a way that the opposite edges with arrows are brought together and the arrows point in the same direction. You will need to twist the paper for the Möbius strip, Klein bottle and the projective plane. Except that the Klein bottle and projective plane cannot be twisted in 3D space! The projective plane is a bit like a sphere, except that the edge has a twist in it before we glue it together. The twist is essential to ensure that intersections of lines behave as we would expect.

6.6 Perspective Projection Revisited

Let's return briefly to our old version of the perspective calculation. Perspective projection is "divide by distance," which we now know is really "inverse-scale-by-distance" so an object at twice the distance is half the size. Typically graphics books describe how to divide by the z coordinate ("divide by depth") to get this effect. In fact the z depth is not the same as distance, so it is fair to ask if this is correct.



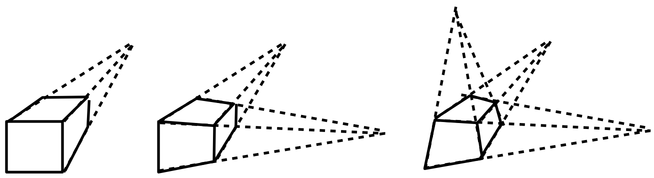
The z value is not the distance to the object, but merely an approximation that is not so bad when the object is close to the z axis. Perhaps the Euclidean distance $\sqrt{x^2 + y^2 + z^2}$ is what we need? This does gives proper perspective but only for a projection on to a unit sphere, not for a projection on to a flat screen.



In general we divide by whatever remains constant on the surface of projection. This will be z for a plane but will be r for a sphere. Normally we are projecting onto a flat screen and hence we should indeed divide by z.

Implicitly, this implies we are viewing the picture at right angles to the screen; that is, our eyes are rigidly staring directly at the screen. When we swivel our eyes or rotate our heads to view the picture, then we are using spherical viewing and there will be a mismatch between what we see here, especially at the edges of the picture, and what we would see in the real scene. As long as the angle is fairly small, the eye will tolerate this error. If it is large, then it starts to look distorted. And if it is one of a stereo pair, then a large angle can cause fusion problems – and headaches!

If you talk about perspective to someone who does technical drawing, you will find that they mention one-point, two-point and three-point perspectives.



Artists have multiple point perspectives, maybe 10 or 20 as the picture seems to require them. So how does this fit in with our notions of perspective? Basically they are setting up vanishing points in every *direction* where they need them. If a set of buildings is constructed around a curve, the Royal Crescent in Bath say, their walls will offer multiple vanishing directions. In terms of our 4D space, this means each will need to be placed with its own perspective correctly formed: there will be a different transformation for each building.

6.7 Constants of Projection

Finally, we know that perspective transformations do not preserve lengths or angles. Is there anything that is invariant? Certainly they transform straight lines to straight lines, which is what allows all computer graphics to work with the coordinates of just the end points of edges. Anything else? Suppose A, B and C are three different collinear points. Define the *division ratio* to be

$$(ABC) = AC/CB$$

where the distances are directed (i.e. AC = -CA). If we have four collinear points, not all ideal, then we can define the *double ratio* to be

$$(ABCD) = (ABC)/(ABD)$$

The double ratio is invariant under a perspective transform. In effect it says the degree of perspective remains unchanged across the whole picture.

6.8 Clipping and Perspective

THIS SECTION IS NOT EXAMINED.

The process in general consists of two transformations. The first (parallel projection) can be any affine transformation. The second is the central projection which introduces the perspective effect.

This permits certain problems which arise to be tackled in new ways. For example, it is possible to clip in between these two transformations.

Refer back to the 2D diagram (only because this is easier to visualise). Remember that any line passing through the origin and cutting the $w = 1$ plane corresponds to a point in the real (Euclidean) world. There are lines passing through the origin which never cut $w = 1$, namely those lines entirely within the $w = 0$ plane. These lines also correspond to points, they are precisely the points at infinity which are not in the Euclidean world but are in the projective one.

Now imagine the tilted plane, shown in that diagram, extending far to the left. It will cut the $w = 0$ plane and carry on into negative w. Suppose we want to draw a single line which, when embedded in this plane, is PQ. Suppose P has a positive w coordinate and Q has a negative one. We can immediately be sure that its central projection will cause a point on the line to go to infinity.

Let's ask what our simple-minded graphics package will do. We complete the projection and arrive at P'Q' say.

```
.....................X...........X....................
                     P'          Q'
```

The package will draw P'Q'.

```
                     X-----------X
```

But this is wrong because the interior of P'Q' is entirely affine: that is, this line does not go to infinity, and it should.

What we need to draw is the line from P' to infinity and then to Q'.

```
---------------------X           X--------------------
```

Of course these lines will have to be clipped against the screen: we cannot really draw to infinity!

We can solve this by performing a *w-clip* in the projective space. Recall that our line crosses the $w = 0$ plane. If we clip the line here, we will have two segments PC (say) and CQ. PC is entirely at positive w and CQ is entirely at negative w. If we clip slightly above and slightly below the plane (to be sure we don't get rounding problems) we will have PC1 and C2Q.

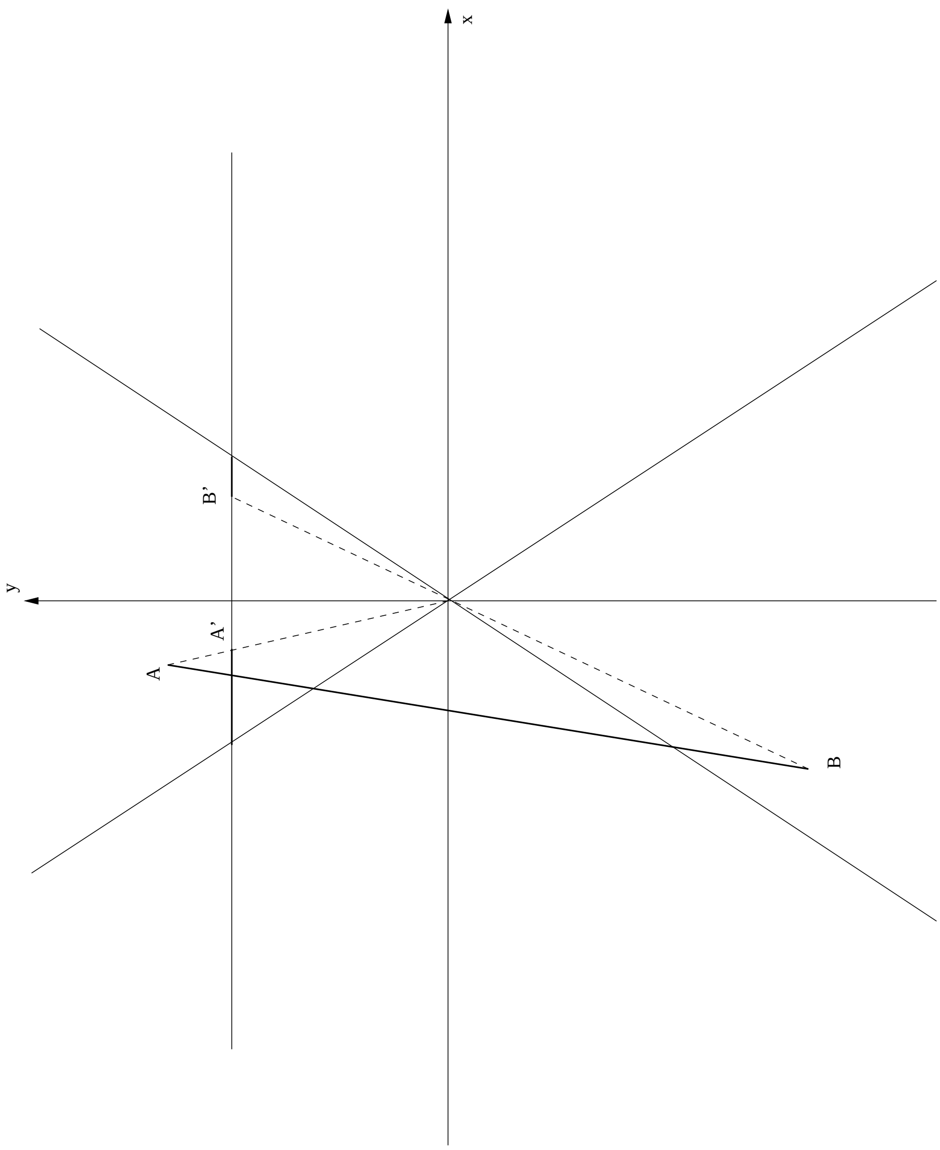
When we project these two lines centrally, C1' heads off in the opposite direction to C2' and we get:

```
..X.................X...........X...............X...
  C1'               P'          Q'              C2'
```

So, when we draw these we will get (before any screen clipping):

```
  X-----------------X           X---------------X
  C1'               P'          Q'              C2'
```

You are probably wondering why we might ever want to draw a line with a hole in the middle! In fact this can happen very easily (and not just to lines). If you move within a rectangular box (a room, say) then the edges which are partly in front, partly behind, will suffer this problem. Here is how it can happen.



The line AB, after central projection, should partly obscure the view to the left and partly to the right, as shown. A' and B', drawn as a line, will wrongly obscure the centre of the view. Of course in practice we should clip edges "behind" the screen and it is possible to do this with a w clip since we know the w value of the projection plane/screen, usually 1.