M2YAh Geometry Notes Ver. 2.0

Konstantinos Drakakis
Lecturer
School of Mathematics
University of Edinburgh

Contents

CHAPTER 1

Introduction

The present set of notes is aimed to be a broad introduction to the field of Computational Geometry. More precisely, we cover Vector, Analytical, and Projective Geometry, as well as some fundamentals of Fractals.

Initially, we provide some background in the definition and the classification of functions (as we will later study some geometric transformations, i.e. functions from one vector space to another), in vector spaces, and in particular the vector spaces $\mathbb{R}^2$ and $\mathbb{R}^3$ (as we will be dealing with geometry on the plane and in space), in the geometric interpretation of differentiation (so that we can easily find tangent lines and planes to curves and surfaces, respectively), and in change of basis in a vector space (which is intimately connected with geometric transformations, such as rotations and reflections).

Subsequently, we focus in two-dimensional geometry: we examine different ways to locate a point on the plane (i.e. different coordinate systems), the general equation of the tangent of a curve at a particular point, and the notion of the curvature of a curve, before we proceed to the study of specific curves. The first curve we study is of course the straight line: we see how to find a line along a vector through a point, a line through two points, a line perpendicular to a given line through a point etc. Then, we proceed to examine the conic sections: the circle, the ellipse, the parabola, and the hyperbola.

We come afterwards to the three-dimensional geometry, whose development follows the plan of the two-dimensional one: we examine different coordinate systems, we see various ways to specify a line and a plane in space, we see the formula for the tangent plane at a point of a surface, and then proceed to study some particular surfaces, the *quadratic* surfaces.

Throughout the study of two- and three-dimensional geometry, we emphasize geometrical transformations, and in particular rotations, rescalings, and translations (or displacements). The goal is to become accustomed to "moving things around".

Next comes the study of projections. We start with orthogonal projection on a sphere and a plane. Then we focus on Projective Geometry and perspective projection: we learn here how to depict three-dimensional objects on a plane, and some of the properties of this kind of projection, in particular the cross-ratio invariance and Pappus's Theorem.

Subsequently, we turn again to two-dimensional geometry to study some more, but less "trendy" special curves: the cycloid, the Quadratrix of Hippias, and spirals.

Finally, we conclude with an introduction to Fractals.

The emphasis of the development is not on rigor, but rather on the computational aspect and the applications in Computer Graphics.

These notes are not intended as a substitute to course attendance, but rather as a study aid.

CHAPTER 2

Basics

1. Mappings

1.1. Definitions. A *function* (or *mapping*, or *map*) is the mathematical formalization of anything that takes some input and produces an output. If the set of inputs is X and the set of possible outputs is contained in Y, then we write $f : X \mapsto Y$ for a function f that maps X into Y. The output corresponding to input $x \in X$ is written $f(x)$. The function could also have been specified by writing $f : x \mapsto f(x),\ \forall x \in X$. The essential feature of a function $f : X \mapsto Y$ is that *for every $x \in X$ there is one and only one output $f(x)$*. But functions can be classified further. A function f is:

- *one-to-one* (1:1), or *injective*, or an *injection*, if f maps distinct elements of X to distinct elements of Y: $x \neq x' \Rightarrow f(x) \neq f(x')$.
- *onto*, or *surjective*, or a *surjection*, if f hits every element of Y: $\forall y \in Y,\ \exists x \in X :\ f(x) = y$.
- *bijective*, or a *bijection*, if f is one-to-one and onto (injective and surjective).

In the context of geometry, a function $f : X \to X$ is often also called a *transformation*.

If $S \subset X$ then the *image* of S through f is $f(S) = \{y \in Y | \exists x \in S : y = f(x)\} \subset Y$. The *range* of f is defined to be $f(X) \subset Y$. Note that f is onto Y if and only if $f(X) = Y$.

1.2. Composition and inverse. If we have another function g whose inputs are the outputs of the function f, then one can do f and then g. This is the *composite function* $g \circ f$. Notice carefully that 'do f then g' is written symbolically as $g \circ f$—we parse the expression from right to left! More formally, suppose that $f : X \to Y$, $g : Y \to Z$ and $h : Z \to W$ are three functions. The composite function $g \circ f : X \to Z$ is defined by

$$g \circ f(x) = g(f(x))$$

The operation of composition is *associative*

$$[h \circ g] \circ f = h \circ [g \circ f]$$

This is very convenient, since it means we don't need to worry about putting in the brackets, and can unambiguously write $h \circ g \circ f$. Composition of functions is *not* commutative, so $f \circ g \neq g \circ f$. In fact, $f \circ g$ is not usually defined. But even if $X = Y = Z$, so all composites are defined, we have, in general, $f \circ g \neq g \circ f$.

For every set X there is a special function Id, or, if necessary, Id_X, called the *identity function*, which sends x to x for all $x \in X$. A function $f : X \to Y$ is called *invertible* if there exists a function $g : Y \to X$ such that

$$g \circ f = \mathrm{Id}_X \text{ and } f \circ g = \mathrm{Id}_Y$$

In this case g is unique and is written f^{-1}, and is called the *inverse* of f. A function is *invertible* if and only if it is *bijective*. There is an order reversal when a composite function is inverted:

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}$$

1.3. Graphs. If $f : \mathbb{R} \to \mathbb{R}$ is some function, then the graph of f is the subset $y = f(x)$ of the usual (x, y)-plane. If f is reasonably well-behaved (continuously differentiable, perhaps) then this graph will be a curve in the plane. Thus functions can be used to represent (some) curves. Recall that the derivative

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f'(x)$$

of f represents the slope of the line tangent to the graph through the point $(x, f(x))$.

Consider now a function $F : \mathbb{R}^2 \mapsto \mathbb{R}$ and consider the special equation $F(x, y) = 0$. The graph of this equation is the set of all $(x, y) \in \mathbb{R}^2$ which satisfy $F(x, y) = 0$. This is a generalized definition of a graph. Indeed, every function graph is a graph in this sense also: just choose $F(x, y) = y - f(x)$. However, the equation $x^2 + y^2 - 1 = 0$ is a graph in the generalized sense without being the graph of any function (as y cannot be expressed as a (single) function of x).

Example. Let $f : \mathbb{R} \mapsto \mathbb{R} : f(x) = 2x + 1$ and $g : \mathbb{R} \mapsto \mathbb{R}_+ : g(x) = x^2$. Then, $(f \circ g)(x) = 2(x^2) + 1$ and $(g \circ f)(x) = (2x + 1)^2$. Both f and g are surjective, and f is also injective, but not g. Hence, the inverses of $f \circ g$ and $g \circ f$ are not defined. The inverse of f is $f^{-1}(x) = 0.5(x - 1)$. If, though, we modify the definition of g as $g : \mathbb{R}_+ \mapsto \mathbb{R}_+ : g(x) = x^2$, then g becomes injective. Under the new definition, $g^{-1}(x) = \sqrt{x}$, and hence $(f \circ g)^{-1}(x) = \sqrt{0.5(x - 1)}$ and $(g \circ f)^{-1}(x) = (2x + 1)^2 = 0.5(\sqrt{x} - 1)$.

2. Vectors

2.1. Definitions. Vectors are the elements of a set with special properties, called a *Vector Space*. The usual rigorous definition of a vector space starts more or less like this: "A vector space $\mathbb{V}$ over a field $\mathbb{F}$ is...". You do not need to know what a field is: for our purposes, we will exclusively be using the real numbers, i.e. for us $\mathbb{F} = \mathbb{R}$.

What are the properties of a vector space? First of all, two functions are defined on it: $+ : \mathbb{V}^2 \mapsto \mathbb{V}$ and $\cdot : \mathbb{R} \times \mathbb{V} \mapsto \mathbb{V}$. Then, we ask for them to have the following properties:

- (Neutral element) $\exists 0 \in \mathbb{V} : \forall a \in \mathbb{V}, a + 0 = 0 + a = a$
- (Negative element) $\forall a \in \mathbb{V}, \exists b \in \mathbb{V} : a + b = b + a = 0$. We denote b by $-a$.
- (Commutative property) $\forall a, b \in \mathbb{V} : a + b = b + a$.
- (Associative property) $\forall a, b, c \in \mathbb{V} : a + (b + c) = (a + b) + c$.

We ask $\mathbb{F}$ to be equipped with two functions $+$ and $\cdot$ (addition and multiplication), having the properties they have in $\mathbb{R}$. Finally, we ask for some combined properties:

- $\forall a, b \in \mathbb{F}, x \in \mathbb{V} : (ab)x = a(bx)$
- $\forall x \in \mathbb{V} : 1 \cdot x = x, 0 \cdot x = 0$
- $\forall a, b \in \mathbb{F}, x \in \mathbb{V} : (a + b)x = ax + bx$
- $\forall a \in \mathbb{F}, x, y \in \mathbb{V} : a(x + y) = ax + ay$

2.2. Overview of properties and definitions. We will now proceed to give some intuitively obvious properties of and definitions about vectors and vector spaces.

- Let $\{a_i\} \in \mathbb{R}$ and $\{x_i\} \in \mathbb{V}$, $i = 1, \ldots, m$. If $\sum_{i=1}^m a_i x_i = 0 \Rightarrow a_i = 0,\ i = 1, \ldots, m$, then the vectors $\{x_i\}$ are called *linearly independent*, otherwise *linearly dependent*.
- There may be an infinite number of linearly independent vectors in the vector space, or a finite one, which means $\exists n > 0 : \forall m > n, \sum_{i=1}^m a_i x_i = 0 \Rightarrow \exists 0 < i \leq m : a_i \neq 0$. In that case, it can be proved that there exist many collections of n linearly independent vectors, but all collections of $n + 1$ vectors or more are linearly dependent. The former spaces are called *infinite dimensional*, the former *finite dimensional*. We will be dealing exclusively with the latter from now on.
- It is clear that n characterizes the vector space, not a specific collection of vectors; it is called the *dimension* of the vector space. Any collection of n linearly independent vectors is called a *basis* of the vector space. Any other vector then can be expressed as a linear combination of the basis. Indeed, take $\{x_i\} \in \mathbb{V}$, $i = 1, \ldots, n$ to be a basis, take a random $y \in \mathbb{V}$, and form the equation $by + \sum_{i=1}^m a_i x_i = 0$. If $b = 0$, then necessarily all $a_i = 0$ by linear independence, hence we have a set of $n + 1$ linearly independent vectors, which is a contradiction. Thus, $b \neq 0$, and we can write: $y = \sum_{i=1}^{n} \left(-\frac{a_i}{b}\right) x_i$ Q.E.D.
- Vector spaces can be many different things: the set of all polynomials with real coefficients, the set of all $m \times n$ matrices with real elements etc. The former is an example of an infinite dimensional space. A fundamental theorem states that all finite dimensional real vector spaces are *isomorphic* to $\mathbb{R}^n$ over $\mathbb{R}$, i.e. there exists a bijective mapping which preserves the operations of vector addition and multiplication

of a vector by a scalar. In what follows, we will study exclusively the vector space $\mathbb{R}^n$, i.e. the points of the n-dimensional space.

Example. The set P_n of all polynomials with real coefficients of degree at most $n \in \mathbb{N}$ is a vector space. Its dimension is $n+1$, as the polynomials $1, x, \ldots, x^n$ form a basis: every polynomial of degree at most n can be written as a linear combination of these, and clearly no power of x is expressible as a linear combination of other powers. This space is then isomorphic to $\mathbb{R}^{n+1}$.

Example. The set C of all continuous functions $f : \mathbb{R} \mapsto \mathbb{R}$ is a vector space, as a linear combination of continuous functions is continuous. The dimension, though, of this vector space is infinite: we need infinitely many functions in order to be sure that any continuous function can be expressed as a linear combination of these functions.

2.3. The vector space $\mathbb{R}^n$. The vectors are ordered n-tuples of real numbers, and we define addition and scalar multiplication by:

- $(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n),\ x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{R}$
- $t \cdot (x_1, \ldots, x_n) = (tx_1, \ldots, tx_n),\ x_1, \ldots, x_n, t \in \mathbb{R}$

It is customary to represent the vectors in $\mathbb{R}^n$ as $n \times 1$ matrices $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ (notice that the attentive reader won't be helpless by the different representations of vectors: when we write them horizontally we use parentheses, not square brackets!); this allows us to manipulate vectors using matrices, as we are about to see.

We can easily check whether a collection of n vectors $x_1, \ldots, x_n$ are linearly independent or not in $\mathbb{R}^n$: the defining equation of linear independence leads to a homogeneous square linear system, whose matrix contains the vectors of the collection as its columns: $Aa = 0$, where $A = [x_1, \ldots, x_n]$, so by Linear Algebra we know that $a = 0$ iff $|A| \neq 0$ (here $|\cdot|$ denotes the *determinant* of the matrix).

The simplest, most obvious, and also most useful basis available in $\mathbb{R}^n$ is $\{e_i\},\ i = 1, \ldots, n$, where $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, the 1 coming at the ith position. For example, for $n = 3$ this gives the 3 triples $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. We can now write any vector as follows: $(x_1, \ldots, x_n) = \sum_{i=1}^n x_i e_i$; the numbers x_i are called the *coordinates* of the vector with respect to the basis $\{e_i\},\ i = 1, \ldots, n$, and in this particular case they coincide with the elements of the vector. Needless to say, we can write the coordinates of the vector with respect to any basis we like, but in this case the coordinates, although uniquely defined, will not necessarily coincide with the elements of the vector. For example, if $\{e_i\}$ is another basis: $(x_1, \ldots, x_n) = \sum_{i=1}^n x'_i e'_i$. More specifically, we call these coordinates *Cartesian* (see Fig. 1). We will see other types of coordinates later on.

Example. In the case of the vector space P_n (see above), the isomorphism between P_n and $\mathbf{R}^{n+1}$ is obvious: $x^i \leftrightarrow e_{i+1},\ i = 0, \ldots, n$. Thus, in P_3,

2.4. Remarks. In order to add some visual information to vectors and be able to recognize them as such without explicitly declaring them as elements of a vector space, we overline them, underline them, or mark them as bold. So, $x \in \mathbb{V}$ is usually written as $\overline{x}$, $\underline{x}$, or $\mathbf{x}$. Be ready to recognize these different representations when you see them.

Traditionally, we tend to think of the elements of $\mathbb{R}^n$ as "points" instead of vectors; we think of vectors as joining points. This is because we think of vectors, which are an abstract construction, "materializing" as oriented linear segments beginning and ending somewhere. In particular, we choose a point of the n-dimensional space and call it the *origin* $\mathbf{O}$ (we commonly use capital/small letters to denote points/vectors), and associate with it the zero vector $\mathbf{0} = (0, \ldots, 0)$, which is the vector that begins and ends at the same point, according to this interpretation. We then associate bijectively vectors and points, by thinking of vectors as oriented linear segments starting at the origin and ending on the point associated with them. Thus, we assign coordinates to points: they are the coordinates of their associated vectors.

The vector beginning at point A and ending at point B is denoted by $\mathbf{AB}$. The vector associated with a point P is the one beginning at O and ending at P, hence it is the vector $\mathbf{OP}$. As the origin is fixed and known, we

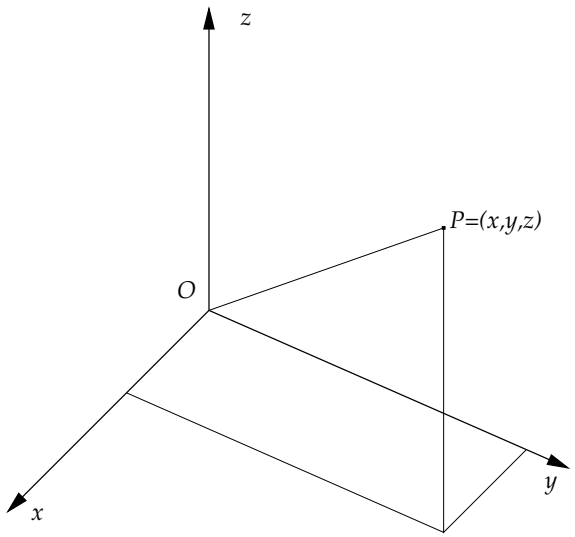

FIGURE 1. Cartesian coordinates with respect to the usual basis for $n = 3$.

omit it very often from our notation; also, we tend to blur the distinction between a point and a vector, so that, for example, we may write $\mathbf{OP} = (2, 5)$ as $\mathbf{P} = (2, 5)$, or even as $P = (2, 5)$.

But we can also use vectors, as we saw, to connect two points: $\mathbf{AB} = \mathbf{OB} - \mathbf{OA}$. Are there then *two* kinds of vectors? No, but the association of vectors and points forces us to artificially separate them into two groups: those who start at the origin are called *position vectors*, while the rest are called *displacement vectors*. So, for example, the vector $\mathbf{x} = (1, 1)$ can be a position vector, corresponding to a point X, or be a displacement vector between the points $(0, 1)$ and $(1, 2)$. It should be clear from the context whether we are viewing a vector as a displacement or as a position.

Moreover, you will probably have noticed by now some *operator overloading* taking place here. Namely, $+$ is used for the addition of either vectors or numbers, whereas $\cdot$ is used for the multiplication of either numbers or a number and a vector. Similarly, 0 can be either the number 0 or the zero vector. There is no danger in confusing all these operations: the symbol is the same, but the operands are different. We will introduce later a product between two vectors, and denote it again by $\cdot$. When even later, however, we will introduce another product between vectors, we will need a different symbol, as these latter products will no longer be distinguishable.

Example. Consider the vector $\mathbf{AB} = \mathbf{OB} - \mathbf{OA}$. The points X lying on the straight line through A and B will be given by the vectors $\mathbf{AX}$ parallel to $\mathbf{AB}$: $\mathbf{AX} = t\mathbf{AB},\ t \in \mathbf{R} \Leftrightarrow \mathbf{OX} = \mathbf{OA} + t\mathbf{AB}$.

The point R so that $AR/RB = m/n$ can be found as follows: $n\mathbf{AR} = m\mathbf{RB} \Leftrightarrow n\mathbf{OR} - n\mathbf{OA} = m\mathbf{OB} - m\mathbf{OR} \Leftrightarrow (n + m)\mathbf{OR} = m\mathbf{OB} + n\mathbf{OA} \Leftrightarrow \mathbf{OR} = \dfrac{m\mathbf{OB} + n\mathbf{OA}}{n + m}$

2.5. The inner product of two vectors. The inner product of two vectors is a function[1] $\cdot : \mathbb{V}^2 \mapsto \mathbb{R}$, with the following properties:

- $\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x},\ \forall \mathbf{x}, \mathbf{y} \in \mathbb{V}$
- $t(\mathbf{x} \cdot \mathbf{y}) = (t\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (t\mathbf{y}),\ \forall t \in \mathbb{R}$
- $\mathbf{x} \cdot (\mathbf{y} + \mathbf{y}) = \mathbf{x} \cdot \mathbf{y} + \mathbf{x} \cdot \mathbf{z}$
- $\mathbf{x} \cdot \mathbf{x} = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$
- $\mathbf{x} \cdot \mathbf{x} \geq 0$

Inner products are intimately connected to both *distance functions* $d(\mathbf{x}, \mathbf{y})$ and *norms* $|\mathbf{x}|$. Distance functions allow us to find the distance between two points, while norms give us the "length" of a vector. Both functions

[1]It is also called *dot product*, because of the symbol we use to denote it

have to be always non-negative, while a distance function has to obey the *triangle inequality*: for any three vectors, $d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y})$. Other properties are $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$, $d(\mathbf{x},\mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$, for $t \in \mathbb{R}$: $|t\mathbf{x}| = |t||\mathbf{x}|$, and $|\mathbf{x}| = 0 \Leftrightarrow \mathbf{x} = 0$. Notice the overloading of the absolute value!

Given the inner product, we define $|\mathbf{x}| = \sqrt{\mathbf{x}\cdot\mathbf{x}}$ and $d(\mathbf{x},\mathbf{y}) = |\mathbf{x}-\mathbf{y}|$. You can check that all properties are satisfied.

Another very important property is the Cauchy-Schwartz inequality. Observe that:

$$|\mathbf{x} + t\mathbf{y}|^2 = |\mathbf{x}|^2 + 2t\bar{x}\bar{y} + t^2|\mathbf{x}| \geq 0$$

hence the discriminant of this trinomial in $t \in \mathbb{R}$ must be nonnegative. It follows that

$$|\mathbf{x}\mathbf{y}| \leq |\mathbf{x}||\mathbf{y}| \tag{1}$$

This implies that $-1 \leq \dfrac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}||\mathbf{y}|} \leq 1$, so it can be viewed as the sine or the cosine of an angle. We use this formula to define the angle between the two vectors; because the inner product is symmetric in its arguments and because the angle of a vector with itself should be 0, while the inner product of a vector with itself is the square of its norm by definition, hence the formula has value 1, we use this formula to define the cosine of the angle between two vectors:

$$\frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}||\mathbf{y}|} = \cos\left(\widehat{\mathbf{x},\mathbf{y}}\right) \Leftrightarrow \mathbf{x}\mathbf{y} = |\mathbf{x}||\mathbf{y}|\cos\left(\widehat{\mathbf{x},\mathbf{y}}\right) \tag{2}$$

Example. The angle between the vectors $(3,0,4)$ and $(0,6,8)$ is found by $\cos(\theta) = \dfrac{32}{5\cdot 10} = 0.64 \Rightarrow \theta \approx 50.21^o$

Although not important in the context of the inner product, it will prove important later to have a more precise definition of the angle between two vectors: the angle then between $\mathbf{x}$ and $\mathbf{y}$ is defined as the angle in $(-\pi,\pi]$ by which $\mathbf{x}$ has to turn so that aligns itself with $\mathbf{y}$. By convention, the angles opening clockwise are taken to be positive.

Example. Consider the vector space C of continuous functions on $\mathbf{R}$. We know that continuous functions are bounded, so the expression $P(f,g) = \displaystyle\int_{-\infty}^{+\infty} f(x)g(x)e^{-|x|}dx$ is well defined (i.e. it converges to a finite real number for any f, g). We can verify the properties of the inner product one by one: $P(f,g) = P(g,f)$, $P(af+bg,h) = aP(f,h) + bP(g,h)$, $P(f,f) \geq 0$ obviously, while $P(f,f) = 0$ implies clearly that $f = 0$ everywhere. Thus, this expression is a valid dot product for the continuous functions, and we can use it to define geometry: it is now valid to talk about the "angle between f and g", or the "length of f". Of course, there are many other valid dot products, e.g. $Q(f,g) = \int_{-\infty}^{+\infty} f(x)g(x)e^{-x^2}dx$, which will create rather different geometries.

So far we have not given a formula for the inner product in $\mathbb{R}^n$; we will now give one by specifying the norm first. Consider the two-dimensional plane, which is the vector space $\mathbb{R}^2$. Here $\mathbf{e}_1 = (1,0)$ and $\mathbf{e}_2 = (0,1)$ are the vectors of the usual basis. The former lies on the x-axis, while the latter on the y-axis, hence they are perpendicular to each other, and therefore, by Pythagoras's theorem, $|(x,y)| = \sqrt{x^2+y^2}$. In three dimensions, elementary geometry gives again $|(x,y,z)| = \sqrt{x^2+y^2+z^2}$. So, we generalize in n dimensions by defining $|(x_1,\ldots,x_n)| = \sqrt{x_1^2+\ldots+x_n^2}$, which is the same as saying that we demand the usual basis vectors to be perpendicular to each other and of unit norm. Use now the equality $4\mathbf{x}\cdot\mathbf{y} = |\mathbf{x}+\mathbf{y}|^2 - |\mathbf{x}-\mathbf{y}|^2$ to obtain

$$\mathbf{x}\cdot\mathbf{y} = \sum_{i=1}^{n} x_i y_i = \mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x} \tag{3}$$

We see that, according to this definition, the vectors of the usual basis have unit norm and are perpendicular to each other. Bases with the former property are called *normalized*; with the latter *orthogonal*; with both *orthonormal*.

If $\mathbf{x}$ is a vector, we define $\hat{\mathbf{x}} = \dfrac{\mathbf{x}}{|\mathbf{x}|}$ to be the unit vector corresponding to it. It is customary in two (and three) dimensions to write the basis as $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ (and $\hat{\mathbf{z}}$), or as $\mathbf{i}$, $\mathbf{j}$ (and $\mathbf{k}$).

To conclude, the existence of the inner product is really important: it allows us to define angles, hence geometry, within a vector space. From now on, we will confine our study in familiar settings, i.e. the two-dimensional plane and the three-dimensional space. Notice also that the transformation $\mathbf{x} \mapsto A\mathbf{x}$ leaves the inner product of two vectors unchanged if $A^TA = I$:

$$(A\mathbf{x}) \cdot (A\mathbf{y}) = (A\mathbf{x})^T(A\mathbf{y}) = \mathbf{x}^TA^TA\mathbf{y} = \mathbf{x}^T\mathbf{y} = \mathbf{x} \cdot \mathbf{y}$$

This means that many operations whose matrices satisfy this property leave the distance between to vectors, the angle between two vectors, and the vector norm unchanged!

Example. The inner product of $(1, 1, 1, 1)$ and $(1, -1, 1, -1)$ is $1 - 1 + 1 - 1 = 0$; the two vectors are perpendicular.

Example. Consider the vectors $(0, 0, 1, 1)$ and $(1, -1, 0, a)$, and suppose that we seek a so that the angle between them is 60^o. Then, $(0, 0, 1, 1) \cdot (1, -1, 0, a) = a = \sqrt{2}\sqrt{a^2 + 2}\cos(60^o) = \sqrt{\dfrac{a^2 + 2}{2}} \Leftrightarrow 2a^2 = a^2 + 2 \Leftrightarrow a = \sqrt{2}$, as $a > 0$.

3. Some elements of linear algebra useful in geometry

3.1. Determinants. Determinants come up quite frequently in geometry, especially when it comes to finding the equation of a plane or a line, so it is useful to know how to compute them. The computation of a determinant is most of the times tedious but fairly systematic, because it is recursive. The determinant of the matrix A is denoted by $|A|$ or $\det(A)$. Then, the following end conditions for the recursion apply:

- The determinant of a number is the number itself: $\det(x) = x, \forall x \in \mathbb{R}$.
- The determinant of a 2×2 matrix is $\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}$
- The determinant of a *triangular* matrix, i.e. a matrix whose elements lying above or below the key diagonal are 0, is the product of the elements of its main diagonal.

In order to compute the determinant of an $n \times n$ matrix, where $n \geq 3$, follow this algorithm:

(1) Choose a particular row or column of the matrix. You can choose anyone you like, but it really helps to choose one with many zeros in it!
(2) Associate to every element a_{ij} of the row or column you chose the sign $(-1)^{i+j}$. This is equivalent to saying that the top left element is associated with $+$, and then all neighboring elements (up, down, left, right) of a given element have the opposite sign than the sign it has itself.
(3) Set $\det(A) = 0$ and repeat for every element a_{ij} of the row or column you chose:
   (a) Find the matrix A_{ij} resulting from the matrix A if you remove entirely the row and column this element lies in (i.e. the ith row and the jth column).
   (b) Set $\det(A) \leftarrow \det(A) + a_{ij}(-1)^{i+j}\det(A_{ij})$

If $\det(A) = 0$ then the vectors corresponding to the rows of A are linearly dependent. The same is true for the vectors corresponding to the columns of A.

Example.

$$\begin{vmatrix} 1 & 2 & 1 \\ 1 & 1 & -1 \\ 2 & 3 & 1 \end{vmatrix} = 1\begin{vmatrix} 1 & -1 \\ 3 & 1 \end{vmatrix} - 2\begin{vmatrix} 1 & -1 \\ 2 & 1 \end{vmatrix} + 1\begin{vmatrix} 1 & 1 \\ 2 & 3 \end{vmatrix} = 4 - 6 + 1 = -1$$

3.2. Introduction to orthogonal transformations and isometries.

3.2.1. *Orthogonal Matrices.* In several instances we will need to perform some linear operations on vectors. All linear operations on vectors in $\mathbb{R}^n$, which leave the origin fixed, can be described by means of square $n \times n$ matrices, which describe how the base is transformed under the operation. In particular, we will be interested in operations that leave the inner product, and thus the angle between vectors, the distance between two points, the norm of a vector etc. unchanged. As we saw before, this is true if the matrix satisfies the property $A^TA = I$, i.e. it is *orthogonal*. Conversely, if the equation $\mathbf{x}^TA^TA\mathbf{y} = \mathbf{x}^T\mathbf{y}$ holds, by assigning successively $\mathbf{x} = \mathbf{e}_i,\ \mathbf{y} = \mathbf{e}_j,\ i, j = 1, \ldots, n$ we obtain the matrix equality $A^TA = I$. Observe that the interpretation of this equation is that a matrix

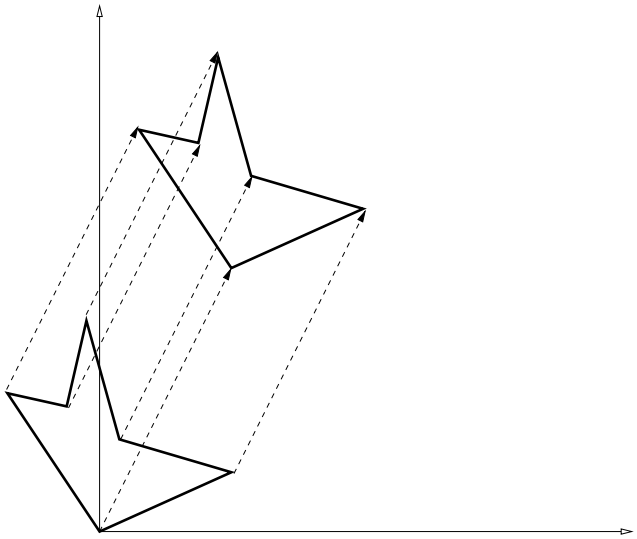

FIGURE 2. The effect of translation in three dimensions.

is orthogonal iff its rows (or columns) form an orthonormal basis for the space. Moreover, it is obvious that *the inverse of an orthogonal matrix is its transpose.*

3.2.2. *Vectors and coordinates.* Having chosen the base $E = [\mathbf{e}_1 \dots \mathbf{e}_n]$ of our vector space, we can write the vector $\mathbf{x}$ as $\mathbf{x} = EC(\mathbf{x})$, which is just a matrix notation for the fact that the vector is a linear combination of the basis vectors, the coefficients of this expansion being the vector's coordinates $C(\mathbf{x})$. Notice here one more case of overloading: we call two different things by the name "vector", i.e. what we already defined as vector, but also the set of its coordinates! There is a way to avoid the ambiguity, by distinguishing between *covariant* and *contravariant* quantities, but we will not get into this. We will instead reserve the notation $C(\mathbf{x})$ for coordinate vectors, whenever we need to make the distinction. Note that the distinction is not useful, unless we work with two or more bases; otherwise, as we work with the usual basis most of the times, for which $E = I$, it holds that $\mathbf{x} = C(\mathbf{x})$!

3.2.3. *Linear Translation.* All matrix transformations leave the origin fixed. But there exists a transformation which preserves inner products, does not leave the origin fixed, and is not given by a matrix: this is the *linear translation* by a fixed vector $\mathbf{a}$, i.e. the mapping $\mathbf{OA} \mapsto \mathbf{OA} + \mathbf{a}$, for any point A (see Fig. 2). This mapping leaves the measure of a vector unchanged (actually it leaves the vector itself unchanged), because for any two points A, B, mapping to A', B': $\mathbf{A'B'} = \mathbf{OB'} - \mathbf{OA'} = (\mathbf{OB} + \mathbf{a}) - (\mathbf{OA} + \mathbf{a}) = \mathbf{OB} - \mathbf{OA} = \mathbf{AB}$.

3.2.4. *Isometry.* The mappings which leave the vector norm unchanged are called *isometries.* It can be shown that the general isometry consist of one operation whose matrix is orthogonal and one translation: $\mathbf{x} \mapsto A\mathbf{x} + \mathbf{a}$.

3.2.5. *Change of basis.* In general, there are two ways we can view the action of a transformation, i.e. either acting on the coordinates or acting on the basis. Imagine we are already using a certain orthonormal basis E (not necessarily the usual one) and we want to use the new orthonormal basis E', connected to the initial one with the transformation matrix P:

$$E = E'P \tag{4}$$

(notice that the ith column of P contains the coordinates of the ith vector of the old basis with respect to the new basis). Let $\mathbf{u}$ be a vector and let its coordinate vector be $C(\mathbf{u})$, so that $\mathbf{u} = EC(\mathbf{u}) = E'PC(\mathbf{u}) = E'C(\mathbf{u}')$. We see then that the coordinates get transformed as

$$C(\mathbf{u}') = PC(\mathbf{u}) \tag{5}$$

i.e. inversely with respect to the basis (P is used to express the old basis with respect to the new, but the new coordinates with respect to the old).

Imagine now that we want to apply an isometry, whose orthogonal matrix we happen to know for the case of a particular basis (most of the times the usual one), to our vector space which is now equipped with a different basis. How are we going to do it? We know that the mapping is $C(\mathbf{v}) = AC(\mathbf{u})$ for the basis we are comfortable with. Applying the rule for transforming the coordinates we just stated: $C(\mathbf{v}') = P\mathbf{c}_v = PAC(\mathbf{u}) = PAP^{-1}PC(\mathbf{u}) = PAP^{-1}C(\mathbf{u}')$, and the transformation matrix we seek is

$$A' = PAP^{-1} = PAP^T \tag{6}$$

Notice that P is orthogonal, so that A' is also orthogonal.

3.3. An alternative approach to orthogonal transformations. There is an alternative approach to writing down the matrix corresponding to a transformation, which may turn out to be easier sometimes. Consider the transformation $\mathbf{x} \mapsto A\mathbf{x}$, and assume that $\mathbf{x}$ is expressed in the usual basis: $\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{e}_i$. Then, $A\mathbf{x} = \sum_{i=1}^{n} x_i A\mathbf{e}_i = [A\mathbf{e}_1 \dots A\mathbf{e}_n]C(A\mathbf{x})$, and the transformation matrix A can be written as:

$$A = [A\mathbf{e}_1 \dots A\mathbf{e}_n] \tag{7}$$

This equality is almost trivial, but it reveals a simple way of writing the transformation matrix: its columns are the images of the basis vectors through the transformation.

3.4. Rotations and reflections. Two special types of isometry are rotations and translations. Observe that $A^T A = I \Rightarrow |A| = \pm 1$ (another case of overloading: the absolute value used with a matrix denotes the determinant!). Rotations correspond to determinant $+1$, reflections to -1. According to this definition, any number of successive rotations is a rotation, while an even number of successive reflections is also a rotation. Reflections and rotations do not commute in general, but for every pair of a reflection followed by a rotation there exist a pair of a (possibly different) rotation followed by a (possibly different) reflection leading to the same result. This implies that, if choose and fix one matrix with determinant -1, all possible reflections will be given by all possible rotations times this matrix.

Notice that repeated rotations and reflections which keep the origin fixed can be written as the product of the vector with the corresponding rotation or reflection matrices. Pay attention to the order: if we apply to $\mathbf{x}$ sequentially the operations with matrices $A_1, \dots, A_m,\ m \in \mathbb{N}$, then the result will be $A_m \dots A_1 \mathbf{x}$ so that we apply first A_1, then A_2, and so on.

3.4.1. *Two dimensions: rotations.* In the two dimensional world, imposing the orthogonality conditions on the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ yields the relations:

$$a^2 + c^2 = 1 = b^2 + d^2,\ ab + cd = 0$$

Imposing in addition that $ad - bc = 1$, so that the matrix represents a rotation, we can write $a = \cos(u)$, $c = \sin(u)$, $b = \cos(v)$, $d = \sin(v)$, because of the first two equations. The other two then give $\cos(u)\cos(v) + \sin(u)\sin(v) = 0 = \cos(u - v)$ and $\cos(u)\sin(v) - \sin(u)\cos(v) = 1 = \sin(v - u)$, i.e. $v - u = \dfrac{\pi}{2}$. Set then $u = \theta$ and $v = \theta + \frac{\pi}{2}$, and the rotation matrix has the form:

$$R(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{8}$$

for some parameter (angle) $\theta \in [-\pi, \pi)$. Actually, θ is a well chosen parameter, as $\theta = 0$ gives the identity matrix, which results to no rotation at all.

Alternatively, rotation by θ will transform the basis vectors $(1, 0)$ and $(0, 1)$ into the vectors $(\cos(\theta), \sin(\theta))$ and $(-\sin(\theta), \cos(\theta))$, respectively. Putting these vectors as columns in the transformation matrix, we obtain (8) again!

3.4.2. *Two dimensions: reflections.* The simplest reflection we can choose, which will generate all the rest, is the reflection along the x-axis, i.e. turning (x, y) into $(x, -y)$. The transformation matrix is $S(0) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ (this strange notation will be explained shortly). How can we obtain the general reflection matrix, describing the reflection along the line forming an angle θ with the x-axis and going through the origin? We are going to use (6).

The unit vector along the line of reflection is $(\cos(\theta), \sin(\theta))$, and the vector perpendicular to that (forming an angle of 90^o CCW) is $(-\sin(\theta), \cos(\theta))$. Changing the basis to this new one, we see that the reflection now takes place along the x-axis of the new basis, whose matrix we just gave above. So, (4) gives: $I = R(\theta)P \Rightarrow P = R(-\theta)$, and finally, if we denote by $S(\theta)$ the reflection matrix, (6) gives: $S(0) = R(-theta)S(\theta)R(\theta) \Rightarrow S(\theta) = R(theta)S(0)R(-\theta)$, so that:

$$S(\theta) = \begin{bmatrix} \cos(2\theta) & \sin(2\theta) \\ \sin(2\theta) & -\cos(2\theta) \end{bmatrix} \tag{9}$$

Example. Consider the vector $\mathbf{v} = (-2, 1)$: rotation by 30^o maps it to $R(30^o)\mathbf{v} = (-\sqrt{3} - 0.5, -1 + 0.5\sqrt{3})$; reflection along the line forming an angle of 30^o with the x-axis and going through the origin maps the vector to $S(30^o)\mathbf{v} = (-1 + 0.5\sqrt{3}, -\sqrt{3} - 0.5)$.

Example. Consider again the vector $\mathbf{v} = (-2, 1)$, but this time imagine we want to rotate it by 30^o around the point $A = (1, 1)$. We will follow the same procedure essentially, but we need to translate the origin to A before we start, and back after we finish. The result is then $\mathbf{OA} + R(30^o)(\mathbf{v} - \mathbf{OA}) = (1 - 1.5\sqrt{3}, -0.5)$.

3.4.3. *Three dimensions.* It is not easy to write down the general equation for a rotation matrix in space. The matrix has 9 parameters, and we impose 3 normality conditions, 3 orthogonality conditions, and 1 more condition about the determinant. Hence, we would need 2 parameters to write it. In a rotation in space, there is always a line, called the *axis of rotation*, which remains fixed during the transform. If the rotation matrix A is given, the axis of rotation is the vector $\mathbf{u}$ with the property $A\mathbf{u} = \mathbf{u}$, i.e. the eigenvector of the matrix corresponding to the eigenvalue 1 (it can be shown that a rotation matrix always has an eigenvalue equal to 1). We will see more about eigenvectors and eigenvalues later on.

We will not try to find the general form of the matrix; instead, we will use the trick of writing the desired rotation using the rotation matrix around a special axis, which we know how to write, and then change the basis. Step by step we do the following (some of the steps may not be clear yet, but we will explain them later in detail):

(1) We know how the rotation matrix would look if we performed the rotation around the z-axis:
$$R_z(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$
(2) We embed the axis of rotation we are given into a new basis: it is straightforward to find the normalized vector parallel to the axis of rotation, a normalized vector perpendicular to it, and finally a normalized vector perpendicular to both.
(3) We change the basis and find the transformation matrix P which maps $\hat{\mathbf{z}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{x}}$ to the vectors we found, following the order by which we introduced them.
(4) The rotation matrix we ask for is going to be: $R(\theta) = P^T R_z(\theta) P$.

3.5. Non-isometric useful transformations. The most important non-isometric transformation in $\mathbb{R}^n$ is *scaling*, i.e. the multiplication of the coordinates by some positive number fixed for each coordinate (*scaling factors*). In other words, the scaling which, for $i = 1, \ldots, n$, scales the ith coordinate by $s_i > 0$ has as its transformation matrix the diagonal matrix $\text{diag}(s_1, \ldots, s_n)$ and its effect on the vector $\mathbf{x}$ is: $(c_{\mathbf{x},1}, \ldots, c_{\mathbf{x},n}) \mapsto (s_1 c_{\mathbf{x},1}, \ldots, s_n c_{\mathbf{x},n})$.

Fig. 3 shows the effect some of the transformations discussed above have on a closed curve in the plane.

3.6. Extended vectors — Unifying rotations and translations. Consider a transformation $\mathbf{x} \mapsto A\mathbf{x} + \mathbf{b}$. We saw earlier that translations do not correspont to an orthogonal matrix; this is the case because they are not really operations that we can perform on vectors, but rather on points (i.e. only on vectors whose start is fixed at

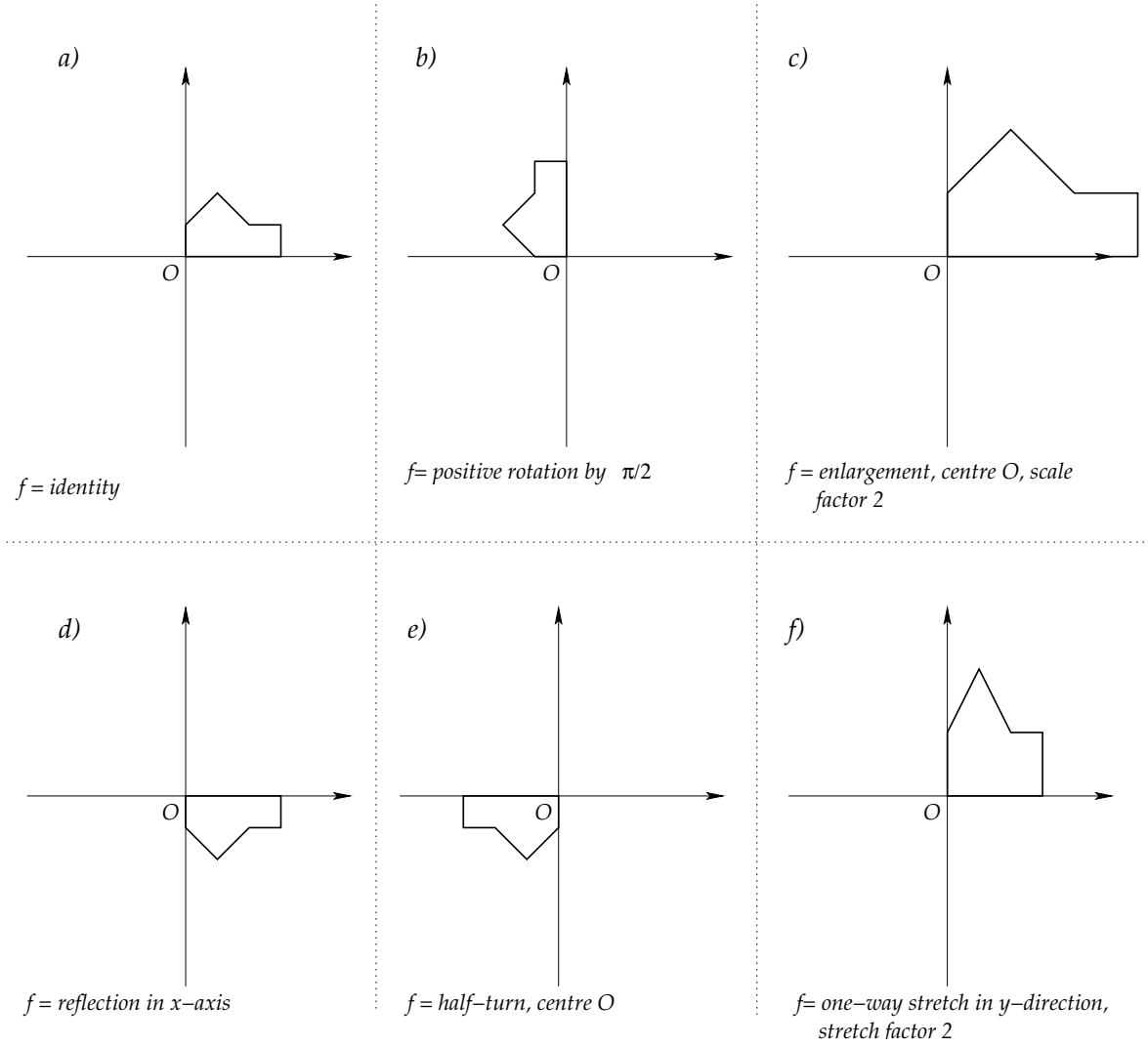

FIGURE 3. The effect of various transformations in two dimensions.

the origin). Matrix notation is more convenient though, so we would like to be able to express this transformation by means of a matrix. This is indeed possible by using extended vectors $\tilde{x} = (\mathbf{x}, 1)$, i.e. by adding the element 1 at the end of the vector. Then, the above mapping becomes:

$$\mathbf{x} \mapsto A\mathbf{x} + \mathbf{b} = [A\ \mathbf{b}] \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}$$

In order to make the matrix square, we use an extended vector for the result as well:

$$\tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ 1 \end{bmatrix} = \begin{bmatrix} A & \mathbf{b} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}$$

CHAPTER 3

The two-dimensional world

1. Polar coordinates

Polar coordinates is an alternative way to specify the location of a point on the plane. They can be more useful than Cartesian coordinates in specifying the equation of a plane curve, especially when the curve is closed.

Take a point P with Cartesian coordinates (x, y). Instead of specifying its position on the plane by means of the vector $\mathbf{OP} = x\mathbf{i} + y\mathbf{j}$, we specify it using the polar coordinates (r, θ), where $r = |\mathbf{OP}|$, and θ the oriented counter-clockwise angle by which $\mathbf{i}$, i.e. the x-axis, has to turn in order to coincide with $\mathbf{OP}$. We can obviously limit the values of θ in $(-\pi, \pi]$ (see Fig.1).

Since Cartesian and polar coordinates are completely equivalent, we can convert from one to the other. The conversion from polar to Cartesian is simple:

$$x = r\cos(\theta),\ y = r\sin(\theta) \tag{10}$$

The conversion from Cartesian to polar needs some attention in finding θ. The previous equations suggest that $\tan(\theta) = \frac{y}{x}$, but concluding that $\theta = \tan^{-1}\left(\frac{y}{x}\right)$ is *wrong*, because then (x, y) and $(-x, -y)$ get assigned the same θ. Instead we use:

$$r = \sqrt{x^2 + y^2},\ \theta = \text{atan2}(x, y) \tag{11}$$

where

$$\text{atan2}(x, y) = \begin{cases} \tan^{-1}\left(\frac{y}{x}\right), & x > 0 \\ \tan^{-1}\left(\frac{y}{x}\right) - \pi, & x < 0, y < 0 \\ \tan^{-1}\left(\frac{y}{x}\right) + \pi, & x < 0, y \geq 0 \end{cases}$$

and $\tan^{-1}$ is defined to return values in $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

The problem is that $\tan^{-1}$ is periodic of period π, and we want a function that returns a value in the whole range of $(-\pi, \pi]$. The name atan2 is also used by major computer languages, such as Fortran, C, Java, and Matlab for this function.

Example. Let us convert $(3, 1)$ in polar coordinates: $r = \sqrt{3^2 + 1^2} = \sqrt{10}$, and $\theta = \text{atan2}(3, 1) = \tan^{-1}(1/3) = 18.435^o$.

2. Complex numbers

Complex numbers are a very big subject by themselves; we will here just introduce them and show how they can be useful in plane geometry.

Complex numbers can be put into one-to-one correspondence with the points of the plane, and therefore with vectors in two dimensions. The mapping is $(a, b) \mapsto a + ib$, where i, corresponding to $(0, 1)$, is the *imaginary unit*. Complex numbers can be added: $(a + ib) + (c + id) = (a + b) + i(c + d)$, but the big difference with vectors is that they can also be multiplied, under the rule that $i^2 = -1$: $(a + ib)(c + id) = (ac - bd) + i(ad + bc)$.

A priceless formula for geometric applications is *Euler's formula*: $e^{ix} = \cos(x) + i\sin(x),\ x \in \mathbb{R}$. Since $a + ib = \sqrt{a^2 + b^2}\left(\frac{a}{\sqrt{a^2 + b^2}} + i\frac{b}{\sqrt{a^2 + b^2}}\right)$, we get the polar form of a complex number:

$$a + ib = re^{i\theta},\ r = \sqrt{a^2 + b^2} = |a + ib|,\ \cos(\theta) = \frac{a}{\sqrt{a^2 + b^2}},\ \sin(\theta) = \frac{b}{\sqrt{a^2 + b^2}} \Rightarrow \theta = \text{atan2}(a, b) = \arg(a + ib)$$

where r is the norm and θ the argument of the number (see Fig. 2).

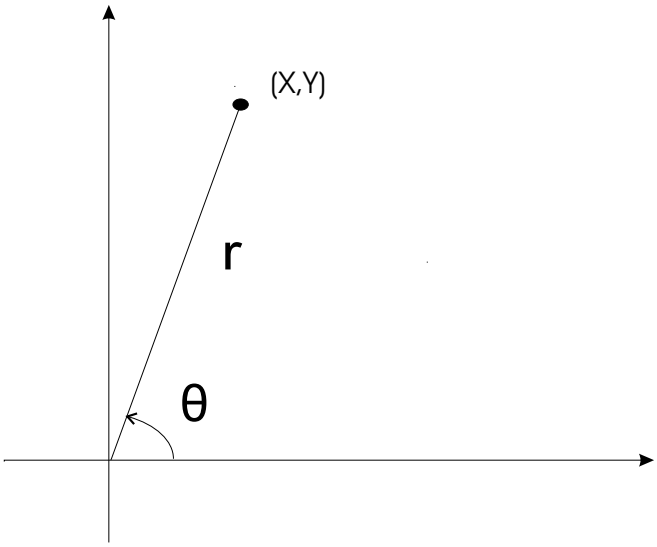

FIGURE 1. Polar coordinates.

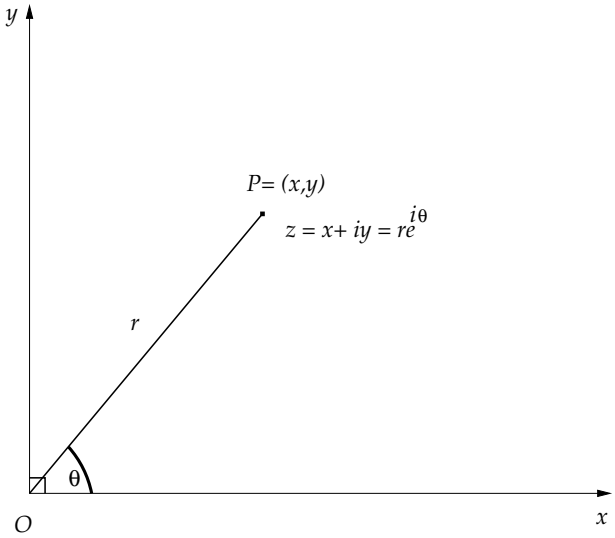

FIGURE 2. Representations of points on the plane using complex number notation.

Complex numbers make rotations really easy: $e^{i\theta}$ is the unit vector towards the direction of angle θ with respect to the x-axis; multiplication of a complex number by i preserves its norm but turns the number by 90^o CCW on the plane, so (back in vector language) it produces a vector perpendicular to the original one; multiplication of $r_1e^{i\theta_1}$ with $r_2e^{i\theta_2}$ gives $r_1r_2e^{i(\theta_1+\theta_2)}$, i.e. angles get added while norms are multiplied.

We see then that it is easy to turn and scale a vector: instead of multiplying by matrices, we can rotate by simple multiplication with $e^{i\theta}$, and scale by multiplying with a real positive number.

Finally, it is easy to reflect a vector: if $z = x + iy,\ x, y \in \mathbb{R}$, then $\bar{z} = x - iy$ is defined to be the *complex conjugate* of z (notice the notation overloading here, and do not confuse conjugates with vectors!). The conjugate of a number in vector language is the reflection of the vector with respect to the x-axis!

The set of complex numbers is symbolized by $\mathbb{C}$.

Example. Let us rotate $\mathbf{v} = (3, -2)$ around $A = (1, 2)$ by 45^o, by using complex numbers: to $\mathbf{v}$ corresponds the complex number $v = 3-2i$, to $\mathbf{OA}$ the number $a = 1+2i$, so that the result is $(1+2i)+e^{i45^o}[(3-2i)-(1+2i)] = (1+2i)+2e^{i45^o}(1-2i) = (1+2i)+\sqrt{2}(1+i)(1-2i) = (1+2i)+\sqrt{2}(3-i) = (1+3\sqrt{2})+i(2\sqrt{2}-1)$.

3. About the representation of curves and surfaces

Smooth curves and surfaces can be described in equations in various different ways:

- As $F(x, y) = 0$, where $F : \mathbb{R}^2 \to \mathbb{R}$, i.e. in Cartesian coordinates.
- As $F(r, \theta) = 0$, where $F : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$, i.e. in polar coordinates.
- As the set $\{(x(t), y(t)) : t \in I \subset \mathbb{R}\}$, i.e. in *parametric form*.

Usually one expects that all of the above functions will be well behaved, i.e. piecewise continuous and differentiable, at least for simple and usual curves. Also, one expects that at least one of the representations above will be easy to find.

4. The line on the plane

Let $\mathbf{w} = (u, v)$ be a vector and (X, Y), (x_1, y_1) and (x_2, y_2) points on the plane. A straight line can be defined as:

(1) parallel to a vector and passing through a point:

$$x = X + \lambda u,\ y = Y + \lambda v,\ \lambda \in \mathbb{R} \Rightarrow v(x - X) - u(y - Y) = 0 \tag{12}$$

(2) perpendicular to a vector and passing through a point:

$$(u, v) \cdot (x - X, y - Y) = 0 \Leftrightarrow u(x - X) + v(y - Y) = 0 \tag{13}$$

.

(3) passing through two points: the displacement vector from the first to the second point is $(x_2 - x_1, y_2 - y_1)$, so we fall back to the case where the line goes through a point along a vector. We can write the line equation as a determinant:

$$(y_2 - y_1)(x - x_1) - (x_2 - x_1)(y - y_1) = \begin{vmatrix} x - x_1 & y - y_1 \\ x_2 - x_1 & y_2 - y_1 \end{vmatrix} = 0 = \begin{vmatrix} x_1 & y_1 & 1 \\ x - x_1 & y - y_1 & 0 \\ x_2 - x_1 & y_2 - y_1 & 0 \end{vmatrix} \Rightarrow \begin{vmatrix} x & y & 1 \\ x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \end{vmatrix} = 0 \tag{14}$$

Example. The line going through the points $(1, 3)$ and $(2, 1)$ is $(1 - 3)(x - 1) - (2 - 1)(y - 3) = 0 \Leftrightarrow -y + 3 - 2x + 2 = 0 \Leftrightarrow y + 2x - 5 = 0$.

5. Angle between two lines

Suppose a line passing through (x_0, y_0) forms an angle θ with the x-axis: then, the unit vector of the line is $(\cos(\theta), \sin(\theta))$. The perpendicular vector is $(\cos(\theta + \pi/2), \sin(\theta + \pi/2)) = (-\sin(\theta), \cos(\theta))$, so that the line equation is $\cos(\theta)(y - y_0) - \sin(\theta)(x - x_0) = 0 \Leftrightarrow y - y_0 = \tan(\theta)(x - x_0)$; therefore, if a line equation is given as $y = \lambda x + a$, then $\lambda = \tan(\theta)$, so that $\theta = \tan^{-1}(\lambda)$.

Suppose then that two lines form angles θ_1 and θ_2 with the x-axis. Then, the angle between them is $\theta = \theta_2 - \theta_1$, and we can use the well known formula (try to show it!):

$$\tan(\theta_2 - \theta_1) = \frac{\tan(\theta_2) - \tan(\theta_1)}{1 + \tan(\theta_2)\tan(\theta_1)} = \frac{\lambda_2 - \lambda_1}{1 + \lambda_2\lambda_1} \tag{15}$$

to find it. In particular, if this angle is 90^o, then the above formula gives that

$$\tan(\theta_2)\tan(\theta_1) = \lambda_2\lambda_1 = -1 \tag{16}$$

Example. What is the angle between the lines $y+2x-5=0$ and $y-3x+2=0$? Rewrite them as $y=-2x+5$ and $y=3x-2$, so that $\tan(\theta)=\dfrac{3-(-2)}{1+3(-2)}=-1\Rightarrow\theta=135^o$.

6. Distance of a point from a line

Consider the line L: $Ax+By+C=0$ and the point P: $\mathbf{P}=(X,Y)$. The line perpendicular to L through P is $B(x-X)-A(y-Y)=0$, or parametrically $x=X+At, y=Y+Bt,\ t\in\mathbb{R}$, and plugging these equations into the equation for L we get $t=-\dfrac{AX+BY+C}{A^2+B^2}$. The intersection then of the two perpendicular lines is the point

$$(X,Y)-\frac{AX+BY+C}{A^2+B^2}(A,B)$$

hence the distance between the two points, i.e. the distance of the point from the line, is:

$$d(P,L)=\frac{|AX+BY+C|}{\sqrt{A^2+B^2}} \tag{17}$$

A nice application of this formula is the calculation of the area of a triangle with vertices $\mathbf{Z}_1=(x_1,y_1)$, $\mathbf{Z}_2=(x_2,y_2)$, and $\mathbf{Z}_3=(x_3,y_3)$. The idea is to compute the height of the triangle from Z_1 as the distance between Z_1 and the line Z_2Z_3. The line Z_2Z_3 is parallel to the vector $(x_3-x_2,y_3-y_2)=(-B,A)$, which is also the base of the triangle. Therefore, by the well known formula for the triangle area:

$$E=\frac{1}{2}|\text{height}||\text{base}|\Rightarrow E=\frac{1}{2}|\begin{vmatrix} x_1 & y_1 & 1\\ x_2 & y_2 & 1\\ x_3 & y_3 & 1\end{vmatrix}| \tag{18}$$

Example. Let us find the distance of $(1,1)$ from the line $x+y=5$: according to (17), the result is $\dfrac{|1+1-5|}{\sqrt{1^2+1^2}}=\dfrac{3}{\sqrt{2}}$

7. Tangent on a plane curve at a point

Consider a plane curve given by the equation $F(x,y)=0$. Take now two particular pairs $\mathbf{Q}=(X,Y)$ and $\mathbf{Q}'=(X+dX,Y+dY)$ satisfying the equation, i.e. belonging in the curve, so that dX is "very small". Finally, let $P=(x,y)$ be a point of the tangent. The equation of the tangent can be found by the observation that $\mathbf{QP}$ and $\mathbf{QQ}'$ must be parallel, i.e. $\exists k\in\mathbb{R}:\mathbf{QP}=k\mathbf{QQ}'$, which leads to the equation

$$dY(x-X)-dX(y-Y)=0$$

Assume now that $F(x,y)=0$ can be solved in a neighborhood of (X,Y) with respect to either x or y, so that we get an expression of the form $y=f(x)$ or $x=g(y)$, for some f or g. In the former case, the Implicit Function Theorem guarantees that if F is differentiable at (X,Y), then f is also, and $\dfrac{dY}{dX}=f'=-\dfrac{F_x}{F_y}$. Substituting into the equation, we get:

$$F_x(x-X)+F_y(y-Y)=0\Leftrightarrow\frac{\partial F}{\partial x}(x-X)+\frac{\partial F}{\partial y}(y-Y)=0 \tag{19}$$

depending on your preference of partial derivative symbol! The latter case ultimately leads to the same expression, so this equation is valid as long as any variable can be expressed as a function of the other.

Finding the tangent to a curve parametrically represented is straightforward:

$$f(t)=\begin{bmatrix} x(t)\\ y(t)\end{bmatrix}\Rightarrow f'(t)=\begin{bmatrix} x'(t)\\ y'(t)\end{bmatrix}$$

and the tangent at the point corresponding to $t=t_0$ is the line:

$$l(s)=(x(t_0),y(t_0))+s(x'(t_0),y'(t_0))),\ s\in\mathbb{R} \tag{20}$$

Example. Let us find the tangent line on the curve $x^2+y^4=10$ at $(3,1)$ (verify that the point belongs in the curve). $F(x,y)=x^2+y^4-10$, so that $F_x(x,y)=2x \Rightarrow F_x(3,1)=6$ and $F_y(x,y)=4y^3 \Rightarrow F_y(3,1)=4$. So, by (19), the tangent line is: $6(x-3)+4(y-1)=0 \Leftrightarrow 3x+2y-11=0$.

8. Curvature, osculating circle and evolute

It would be useful to have a measure of how much a curve "curves" at a point. The way to measure this is by the *osculating circle*. The curvature is defined to be inverse of the radius of the osculating circle, and its derivation is a fine exercise in differentiation. We will follow here the exposition given in [**3**].

One way to measure the radius r of a circle is to express it as $ds=rd\phi$, where ds is a small arc length and $d\phi$ the angle corresponding to that arc length. It is easy to see that $d\phi$ can be measured also as the derivative of the angle that the tangent at this point forms with the x-axis.

Let us be a bit more concrete now: consider the curve given in parametric form by $(f(t),g(t))$ for t in some range. The tangent vector is $(f'(t),g'(t))$, and the angle of the tangent with the x-axis is given by $\tan(\phi)=\dfrac{g'}{f'} \Rightarrow \phi=\tan^{-1}\left(\dfrac{g'}{f'}\right)$, so the derivative is:

$$d\phi=\left(\frac{g'}{f'}\right)'\frac{dt}{1+\left(\frac{g'}{f'}\right)}=\frac{g''f'-f''g'}{f'^2+g'^2}dt$$

The arc length variation, on the other hand, is $ds=\sqrt{dx^2+dy^2}=\sqrt{f'^2+g'^2}dt$, so that the curvature is finally:

$$K=\frac{1}{r}=\frac{g''f'-f''g'}{(f'^2+g'^2)^{\frac{3}{2}}} \tag{21}$$

So, for a straight line, the radius of curvature is ∞ and the curvature 0. Now, the line perpendicular to the tangent at the point of tangency is $f'(x-f)+g'(y-g)=0$, and in parametric form:

$$x=f\pm\frac{g'}{\sqrt{f'^2+g'^2}}\lambda,\ y=g\mp\frac{f'}{\sqrt{f'^2+g'^2}}\lambda,\ \lambda\in\mathbb{R}$$

The center of the circle therefore can lie in one of two points, for $\lambda=r$:

$$\left(f\pm\frac{g'(f'^2+g'^2)}{g''f'-f''g'},g\mp\frac{f'(f'^2+g'^2)}{g''f'-f''g'}\right)$$

How do we know which one? Take a look at Fig. 3. We see that the center of the circle (X,Y) is at $X-f=r\cos\delta,\ Y-g=r\sin\delta$, and that $\delta=\pi-\phi$, and we know that $(\cos\phi,\sin\phi)=\dfrac{(f',g')}{\sqrt{f'^2+g'^2}}$. We see then that δ and ϕ must have equal sines and opposite cosines, hence $X-f=-r\cos(\phi),\ Y-g=r\sin(\phi)$, which gives:

$$\left(f-\frac{g'(f'^2+g'^2)}{g''f'-f''g'},g+\frac{f'(f'^2+g'^2)}{g''f'-f''g'}\right) \tag{22}$$

Observe that these points, considered as a curve for all values of t, form a new curve called the *evolute* of the original curve.

Example. Let us find the evolute of $y^2=x$, which, as we are about to see a bit later, is a parabola. This equation can be expressed parametrically as $(f(t),g(t))=(t,t^2),\ t\in\mathbb{R}$, so (22) gives the evolute in parametric form as: $\left(t-\dfrac{2t(1^2+(2t)^2)}{2-0},t^2+\dfrac{(1^2+(2t)^2)}{2-0}\right)=\left(t-t(1+4t^2),t^2+\dfrac{1+4t^2}{2}\right)=\left(-4t^3,3t^2+\dfrac{1}{2}\right)$. We can also find the Cartesian form of the equation: just set $x=-4t^3$ and $y=3t^2+0.5$, and observe that $\left(\dfrac{x}{-4}\right)^2=\left(\dfrac{y-0.5}{3}\right)^3=t^6 \Rightarrow y=0.5+3\left(\dfrac{x}{4}\right)^{\frac{2}{3}}$

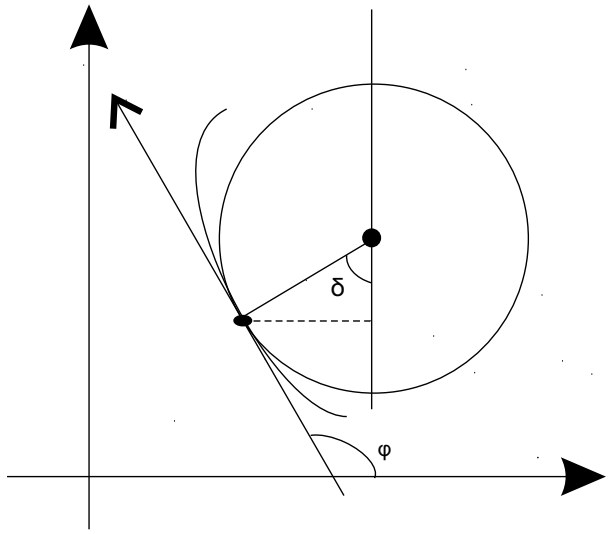

FIGURE 3. Diagram explaining the derivation of the radius of curvature and the osculating circle.

9. The conic sections

The first curves we are going to examine are the conic sections. The "conic sections" is a term referring collectively to four curves: the circle, the ellipse, the parabola, and the hyperbola. There is nothing really special about these curves, except that they are traditionally the first curves one examines in geometry, and we will follow too this honorable tradition! These curves, as any other curve actually, are defined as the set of points of the plane that have a particular property.

9.1. The circle.

9.1.1. *Definition.* Given a point A of the plane (called the"center") and a positive real r, the circle with center $A = (x_A, y_A)$ and radius r is the set of points of the plane whose distance from A is r:

$$C_{A,r} = \{P : |\mathbf{AP}| = r\}$$

The equation of the circle is straightforward to find:

$$(x - x_A)^2 + (y - y_A)^2 = r^2 \tag{23}$$

9.1.2. *Alternative representations.* The expanded form of (23) is:

$$x^2 + y^2 - 2x_A x - 2y_A y + x_A^2 + y_A^2 - r^2 = 0 \tag{24}$$

On the other hand, every equation of the form:

$$Ax^2 + Ay^2 + Bx + Cx + D = 0$$

describes either a circle, or a single point, or the empty set, depending on the parameter values.

Finally, the parametric equations for the circle are:

$$x(t) = x_A + r\cos(t),\ y(t) = y_A + r\sin(t),\ t \in (-\pi, \pi] \tag{25}$$

In the case of the simple circle $x^2 + y^2 = R^2$, the equation in polar coordinates is $r = R$. Polar coordinates are unfortunately not translation invariant, so the polar coordinates do not yield such a pretty formula for the general circle.

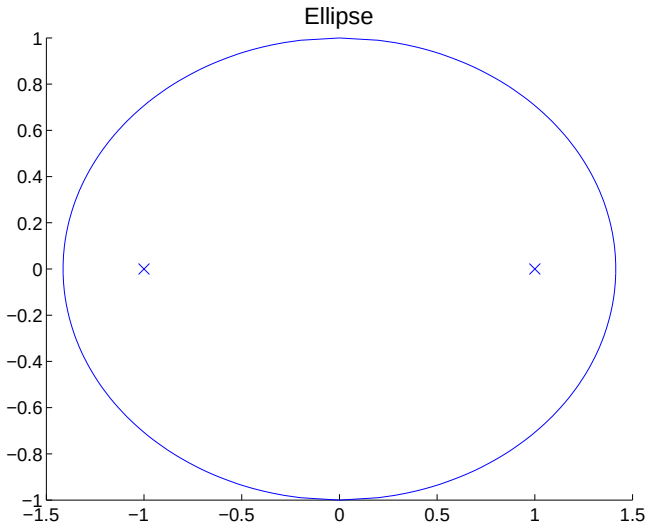

FIGURE 4. The ellipse.

9.1.3. *Tangent.* The tangent of the circle at the point (X, Y) can be found by means of the general formula (19). In our case, $F(x, y) = (x - x_A)^2 + (y - y_A)^2 - r^2$, hence $F_x(x, y) = 2(x - x_A)$, $F_y(x, y) = 2(y - y_A)$, so the tangent is:

$$(X - x_A)(x - X) + (Y - y_A)(y - Y) = 0 \Leftrightarrow (x - x_A)(X - x_A) + (y - y_A)(Y - y_A) = r^2$$

9.1.4. *Perimeter and Area.* For reasons of completeness, we mention that the area of a circle is πr^2, and its perimeter $2\pi r$.

9.2. The ellipse.

9.2.1. *Introduction.* Given two points A and B on the plane, and a positive real a, the ellipse with focal points (or focii) A and B and parameter a is the set of the points of the plane whose distances from A and B sum to $2a$ (see Fig.4):

$$E_{A,B,a} = \{P : |\mathbf{AP}| + |\mathbf{BP}| = 2a\}$$

9.2.2. *Definition.* In order to find a simple equation for the ellipse, notice first that the ellipse can be rotated and translated so that the linear segment AB gets positioned on the x-axis, and that its center coincides with O. Moreover, since for such an arrangement $A = (-r, 0)$, $B = (r, 0)$, so that $0 \leq r \leq a$, it is convenient to write $A = (-ae, 0)$, $B = (ae, 0)$, $0 \leq e \leq 1$. If then $P = (x, y)$ lies on the ellipse:

$$\sqrt{(x + ae)^2 + y^2} + \sqrt{(x - ae)^2 + y^2} = 2a \Leftrightarrow \sqrt{(x + ae)^2 + y^2} = 2a - \sqrt{(x - ae)^2 + y^2} \Leftrightarrow$$
$$(x + ae)^2 + y^2 = 4a^2 + (x - ae)^2 + y^2 - 4a\sqrt{(x - ae)^2 + y^2} \Leftrightarrow 4aex = 4a^2 - 4a\sqrt{(x - ae)^2 + y^2} \Leftrightarrow$$
$$(x - ae)^2 + y^2 = (a - ex)^2 \Leftrightarrow x^2(1 - e^2) + y^2 = a^2(1 - e^2) \Leftrightarrow \frac{x^2}{a^2} + \frac{y^2}{a^2(1 - e^2)} = 1$$

It is customary to set $b = a\sqrt{1 - e^2}$, so that the ellipse equation becomes:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{26}$$

- The parameter e is called the *eccentricity* of the ellipse. An eccentricity close to 0 implies that the ellipse is close to a circle; an eccentricity close to 1 implies that the ellipse is close to the linear segment AB. So, the ellipse is a "squashed" circle, and the higher the eccentricity, the more squashed it is.
- a and b are called the (lengths of the) *large* and *small semiaxis* of the ellipse.

An alternative definition of the ellipse is that the ellipse is the image of a circle through a map which dilates the unit vectors of two vertical directions unequally. Again, by translations and rotations we can bring the center of the circle to O and the two vertical directions in question to coincide with the axes. Let f map the xy-plane to the uv-plane, so that $u = ax$, $v = by$, $a, b > 0$. Then, the equation $x^2 + y^2 = 1$ becomes $\frac{u^2}{a^2} + \frac{v^2}{b^2} = 1$, i.e. the circle becomes an ellipse under the mapping.

9.2.3. *Alternative equations.* The general equation of the ellipse is of the form:

$$Ax^2 + By^2 + Cx + Dy + E = 0,\ A \neq B,\ A, B > 0 \tag{27}$$

which can actually represent an ellipse, a single point, or nothing, according to the parameter values.

The ellipse has a simple parametric definition: $\frac{(x-x_A)^2}{a^2} + \frac{(y-y_A)^2}{b^2} = 1$ can be represented as

$$x = x_A + a\cos(t),\ y = y_A + b\sin(t),\ t \in [-\pi, \pi] \tag{28}$$

It is also simple to express $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ in polar coordinates. First, add and subtract the term $\frac{y^2}{a^2}$. Subsequently, simple algebraic manipulations yield the equation

$$r = \frac{b}{\sqrt{1 - e^2\cos^2(\theta)}} \tag{29}$$

9.2.4. *Tangent.* The tangent to the ellipse at the point (X, Y) can once more be determined by the general formula (19): $F(x, y) = \frac{(x-x_A)^2}{a^2} + \frac{(y-y_A)^2}{b^2} - 1$, $F_x(x, y) = 2\frac{x-x_A}{a^2}$, $F_y(x, y) = 2\frac{y-y_A}{b^2}$, so the tangent is:

$$(x - X)\frac{X - x_A}{a^2} + (y - Y)\frac{Y - y_A}{b^2} = 0 \Leftrightarrow (x - x_A)\frac{X - x_A}{a^2} + (y - y_A)\frac{Y - y_A}{b^2} = 1 \tag{30}$$

The area of the ellipse can be seen immediately to be πab, just by rescaling the axes and converting it to a circle. Its perimeter length p though cannot be computed exactly by means of elementary functions; it is given by what is called *an elliptic integral of the second kind*. However, there exists an amazingly accurate approximation, found by the legendary Indian mathematician Srinivasa Ramanujan:

$$p = \pi(a + b)\left(1 + \frac{3h}{10 + \sqrt{4 - 3h}}\right), h = \left(\frac{a - b}{a + b}\right)^2 \tag{31}$$

whose error is bounded by $3 \cdot 2^{-17}h^5\pi(a + b)$.

9.3. The parabola.

9.3.1. *Introduction.* Given a straight line on the plane (*directrix*) and a point not contained in it (*focus*), the parabola is the set of the points of the plane whose distances from the directrix and the focus are equal. If l denotes the directrix and F the focus (see Fig. 5):

$$PB_{l,F} = \{P : |\mathbf{PF}| = \min_{Q \in l} |\mathbf{QP}|\}$$

9.3.2. *Equation.* As usual, the equation of the parabola is simple once we position it appropriately. By means of a rotation and a translation, we can bring the directrix to be the line $x = -a$, and the focus to be $F = (a, 0)$. Then, if $P = (x, y)$ is a point of the parabola, the distance from l is $x+a$, and the distance from F is $\sqrt{(x - a)^2 + y^2}$.

$$\sqrt{(x - a)^2 + y^2} = x + a \Leftrightarrow (x - a)^2 + y^2 = (x + a)^2 \Leftrightarrow y^2 = 4ax \tag{32}$$

9.3.3. *Alternative Representations.* A parametric expression for the parabola is:

$$y(t) = 2at,\ x(t) = at^2,\ t \in \mathbb{R}$$

while the polar equation is:

$$r = \frac{2a}{1 - \cos(\theta)} - \frac{2a}{1 + \cos(\theta)},\ \theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] - \{0\}$$

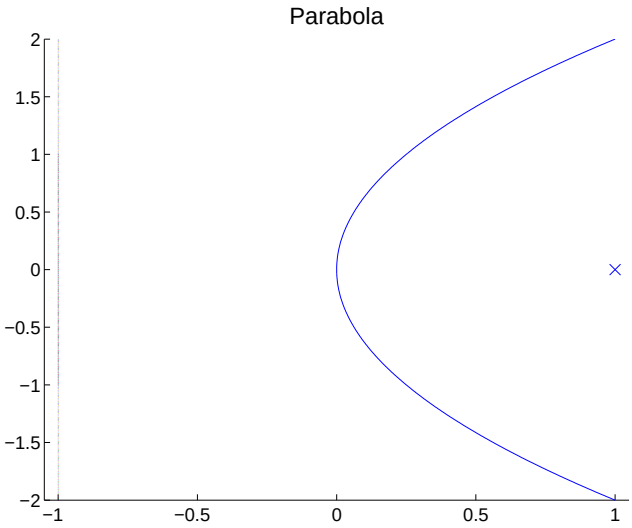

FIGURE 5. The parabola.

The point $(0,0)$ coincides with the *vertex* of the parabola. If we wish to translate the vertex to (u,w), the equation becomes $(y-w)^2 = 4a(x-u)$, where now the focus is at $(u+a,w)$, while the directrix is $x = u-a$. The general equation of the parabola is then of the form:

$$Ax^2 + Bx + Cy + D = 0 \text{ or } Ay^2 + Bx + Cy + D = 0,\ A \neq 0 \tag{33}$$

9.3.4. *Tangent.* The tangent at a point (X,Y) can be found as usual: $F(x,y) = (y-w)^2 - 4a(x-u) = 0$, $F_x(x,y) = -4a$, $F_y(x,y) = 2(y-w)$, so that the tangent is:

$$-2a(x-X) + (Y-w)(y-Y) = 0 \Leftrightarrow (Y-w)(y-w) = 2a(x+X-2u) \tag{34}$$

9.3.5. *An interesting property.* A nice property of the parabola is that any ray parallel to its axis and reflected on the interior of its surface passes through the focus. This is why the paraboloid, a surface produced by rotating the parabola around its axis of symmetry, is used as an antenna.

9.4. The hyperbola.

9.4.1. *Introduction.* Given two points A and B on the plane and a positive real a, the hyperbola is the set of points of the plane whose distances from A and B differ by $2a$ (see Fig. 6):

$$HB_{A,B,a} = \{P : ||\mathbf{AP}| - |\mathbf{BP}|| = 2a\}$$

9.4.2. *Equation.* Once more, by rotating and translating, we can bring the points A and B on the y-axis, so that $A = (-ae,0)$, $B = (ae,0)$, where e is a positive constant (sometimes called the *eccentricity*, just like in the case of the ellipse); a little calculation shows that actually the condition $e \geq 1$ must hold. In order to derive the equation of the hyperbola, let the point $P = (x,y)$ belong to it, and let us assign a sign to the expression in absolute value of the definition arbitrarily:

$$\sqrt{(x+ae)^2+y^2} - \sqrt{(x-ae)^2+y^2} = 2a \Leftrightarrow (x+ae)^2 + y^2 = 4a^2 + (x-ae)^2 + y^2 + 4a\sqrt{(x-ae)^2+y^2} \Leftrightarrow$$
$$(xe-a)^2 = (x-ae)^2 + y^2 \Leftrightarrow y^2 - x^2(e^2-1) = a^2(e^2-1) \Leftrightarrow \frac{y^2}{b^2} - \frac{x^2}{a^2} = 1$$

where $b = a\sqrt{e^2-1}$. If we had taken the opposite sign, we would have reached the equation $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$. Therefore, the hyperbola equation is

$$\left|\frac{y^2}{b^2} - \frac{x^2}{a^2}\right| = 1 \tag{35}$$

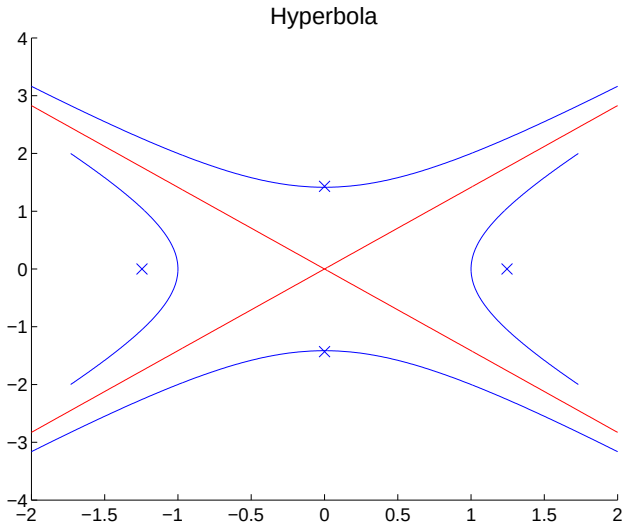

FIGURE 6. The hyperbola.

which consists of 4 individual branches, not connected to each other. In practice, it is customary to consider the 2 possible equations without the absolute value individually, and we call each one of them a hyperbola (and each one is the *conjugate* of the other). So, each hyperbola has 2 branches.

9.4.3. *Alternative representations.* The general equation is of the form:

$$Ax^2 - By^2 + Cx + Dy + E = 0,\ AB > 0 \tag{36}$$

For the parametric representation, we use the identity between the hyperbolic sine and cosine: $\cosh^2(x) - \sinh^2(x) = 1$. Hence:

$$x(t) = u + a\sinh(t),\ y(t) = w \pm b\cosh(t),\ t \in \mathbb{R} \tag{37}$$

The polar representation can be found easily after adding and subtracting the term $\frac{y^2}{a^2}$. For all 4 branches, it is:

$$r = \frac{ab}{\sqrt{|a^2\sin^2(\theta) - b^2\cos^2(\theta)|}} \tag{38}$$

For each individual conjugate parabola, just remove the absolute value, take the two sign cases, and restrict the formula to the values of θ it makes sense.

9.4.4. *Asymptotes.* The hyperbola has two straight line asymptotes: this means that for large values of x and y the branches of the hyperbola behave more or less linearly. This can be seen from:

$$\frac{y^2}{b^2} - \frac{x^2}{a^2} = 1 \Leftrightarrow \frac{y^2}{b^2} = \frac{x^2}{a^2} + 1 \approx \frac{x^2}{a^2} \Leftrightarrow y \approx \pm\frac{b}{a}x$$

Hence O is the point of intersection of the asymptotes. If we translate this to (u, w), the equation becomes

$$\frac{(y-w)^2}{b^2} - \frac{(x-u)^2}{a^2} = 1 \tag{39}$$

Notice that the two conjugate hyperbolas have the same asymptotes!

9.4.5. *Tangent.* The tangent at (X,Y) is now routine to compute — use (19): $F(x,y)=\dfrac{(y-w)^2}{b^2}-\dfrac{(x-u)^2}{a^2}-1$, $F_x(x,y)=-2\dfrac{x-u}{a^2}$, $F_y(x,y)=2\dfrac{y-w}{b^2}$, so the tangent is:

$$\frac{Y-w}{b^2}(y-Y)-\frac{X-u}{a^2}(x-X)=0\Leftrightarrow\frac{Y-w}{b^2}(y-w)-\frac{X-u}{a^2}(x-u)=1 \tag{40}$$

10. Recognizing a conic section

From (24), (27), (33), and (36) you can see that the first part of the general equation of a conic section is enough to identify its type: Ax^2+By^2. If A and B have the same sign, it is an ellipse (or a circle if they are also equal); if they are of opposite sign, it is a hyperbola; if one of them is 0, it is a parabola.

A complication arises when we need to recognize a rotated curve, in which case a cross-product term (Fxy) appears in the equation. We then need to rotate the curve back, which in mathematical terms means to eliminate this cross-product term by a linear substitution of variables; once we eliminate this term, we can use the previous paragraph to identify the conic section. In order to carry out this elimination, though, you will need to know how to diagonalize a matrix; if you don't, just attempt to substitute manually equations of the form $x=Pu+Qv, y=Ru+Sv$, and solve a system for the four parameters, or try to complete squares. We will see the general method below, when we talk about quadratic surfaces in three dimensions.

11. A unified treatment of the conic sections

It is possible to express all four conic sections (ellipse, parabola, hyperbola) by means of a single equation with one parameter, the eccentricity. Consider a straight line d (the *directrix*) and a point F not contained in it (the *focus*). By rotating and translating we can bring d to coincide with the y-axis ($x=0$), while $F=(p,0)$ for some $p\in\mathbb{R}$. We seek the set of all points $P=(x,y)$ on the plane with the property that

$$\frac{\mathrm{d}(F,P)}{\mathrm{d}(d,P)}=e>0$$

($\mathrm{d}(\cdot,\cdot)$ denotes the distance function defined in (17) a bit earlier). Observing that $\mathrm{d}(d,P)=|x|$ and $\mathrm{d}(F,P)=\sqrt{y^2+(x-p)^2}$, we finally get the equation

$$(1-e^2)x^2+y^2-2px+p^2=0$$

which assumes the forms:

$$\begin{aligned}
0<e<1:&\quad y^2+(1-e^2)\left(x-\frac{p}{1-e^2}\right)^2=\frac{p^2e^2}{1-e^2}\\
e=1:&\quad y^2=2p\left(x-\frac{p}{2}\right)\\
e>1:&\quad (e^2-1)\left(x+\frac{p}{e^2-1}\right)^2-y^2=\frac{p^2e^2}{e^2-1}
\end{aligned}$$

We see that as the eccentricity crosses the value 1, the curve changes from an ellipse into a parabola and then into a hyperbola.

CHAPTER 4

The three-dimensional world

Much of what we said already about two dimensions remain still valid for three. The differences are mainly that:

- There exist no more tangent lines but rather tangent planes,
- We study both curves and two-dimensional surfaces,
- Rotations are slightly more complicated, as we can now rotate around infinitely many axes instead of one,
- Reflections take place through planes, not lines.

1. Coordinate systems

Besides the Cartesian coordinates (x, y, z) there are three more useful and interesting coordinate systems (remember that $\hat{()}$ denotes a unit vector):

- *Cylindrical coordinates*:

$$r = \sqrt{x^2 + y^2},\ \theta = \operatorname{atan2}(x, y),\ z = z \tag{41}$$

  These are very convenient for studying surfaces generated by a full rotation of a two-dimensional curve around an axis. We can invert them just like the polar coordinates (see (11) for details).
- *Spherical coordinates*: Set $\mathbf{r} = (x, y, z)$, $\mathbf{r}_{xy} = (x, y)$. Then,

$$r = \sqrt{x^2 + y^2 + z^2},\ \theta = \angle(\hat{r}, \hat{z}),\ \phi = \angle(\hat{r}_{xy}, \hat{x}) \tag{42}$$

  where $\theta \in [0, \pi]$, $\phi \in [0, 2\pi]$. These are very convenient for studying the sphere and ellipsoids. We invert them as follows:

$$z = r\cos(\theta),\ x = r\sin(\theta)\cos(\phi),\ y = r\sin(\theta)\sin(\phi) \tag{43}$$

- *Parametric representation*: A surface needs two parameters to be described, so a parametric representation looks like:$x = x(u, v), y = y(u, v), z = z(u, v)$.

Example. Consider a curve (helix) in parametric form: $(\cos(t), \sin(t), t),\ t \in \mathbb{R}$; let us convert that in spherical coordinates with t as a parameter. (42) gives that $r = \sqrt{1 + t^2},\ \theta = \cos^{-1}\left(\frac{t}{\sqrt{t^2+1}}\right) = \tan^{-1}(t^{-1})$. *It is important to note that we need to keep the solution in* $[0, \pi]$. Finally, $\phi = \cos^{-1}(\cos(t)) = t$.

2. The line

The equation of the line passing through the point (X, Y, Z) along the vector (A, B, C) is given parametrically by:

$$x = X + At,\ y = Y + Bt,\ z = Z + Ct,\quad t \in \mathbb{R} \tag{44}$$

while the Cartesian equation is:

$$\frac{x - X}{A} = \frac{y - Y}{B} = \frac{z - Z}{C} \tag{45}$$

Example. The Cartesian equation of the parametric line $(x,y,z)=(1,2,1)+t(3,0,-4)$ is $\dfrac{x-1}{3}=\dfrac{y-2}{0}=-\dfrac{z-1}{4}$. Notice here that 0 is allowed to appear in the denominator, as long as we agree to interpret this as $\dfrac{x-1}{3}=-\dfrac{z-1}{4}$ and $y=2$.

3. The plane

3.1. Definition. A plane can be defined in several equivalent ways, just like the line:

- *A point and a vector*: the plane is the collection of all vectors starting at (X,Y,Z) and perpendicular to (u,v,w). If (x,y,z) belongs to the plane, then:

$$(u,v,w)(x-X,y-Y,z-Z)=0\Leftrightarrow u(x-X)+v(y-Y)+w(z-Z)=0 \tag{46}$$

- Two linearly independent vectors and a point: if (x,y,z) belongs in the plane, and we want the plane to pass through (X,Y,Z) in the direction of the two vectors $\mathbf{u_1}=(x_1,y_1,z_1)$ and $\mathbf{u_2}=(x_2,y_2,z_2)$, then $(x-X,y-Y,z-Z)$ will have to be a linear combination of these two vectors. Remembering from linear algebra that a square matrix has determinant 0 iff its rows and/or columns are linearly dependent vectors, it is easy to write the plane equation as a determinant:

$$\begin{vmatrix} x-X & y-Y & z-Z \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{vmatrix}=0\Leftrightarrow x(y_1z_2-z_1y_2)+y(z_1x_2-z_2x_1)+z(x_1y_2-y_1x_2)=0 \tag{47}$$

- Three points (x_1,y_1,z_1), (x_2,y_2,z_2), (x_3,y_3,z_3): they define two vectors, so we fall back into the previous case:

$$\begin{vmatrix} x-x_3 & y-y_3 & z-z_3 \\ x_1-x_3 & y_1-y_3 & z_1-z_3 \\ x_2-x_3 & y_2-y_3 & z_2-z_3 \end{vmatrix}=0=-\begin{vmatrix} x_3 & y_3 & z_3 & 1 \\ x-x_3 & y-y_3 & z-z_3 & 0 \\ x_1-x_3 & y_1-y_3 & z_1-z_3 & 0 \\ x_2-x_3 & y_2-y_3 & z_2-z_3 & 0 \end{vmatrix}\Leftrightarrow\begin{vmatrix} x & y & z & 1 \\ x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \end{vmatrix}=0 \tag{48}$$

- General equation: it is clear from the above that the general plane equation is:

$$Ax+By+Cz+D=0 \tag{49}$$

Example. The plane going through the points $(1,1,1)$, $(1,1,0)$, and $(1,2,3)$ is given by:

$$\begin{vmatrix} x & y & z & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 2 & 3 & 1 \end{vmatrix}=0\Rightarrow\begin{vmatrix} x & y & z & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 3 & 0 \end{vmatrix}=0\Rightarrow\begin{vmatrix} x & y & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{vmatrix}=0\Rightarrow\begin{vmatrix} x & 1 \\ 1 & 1 \end{vmatrix}=0\Rightarrow x-1=0$$

Example. The plane perpendicular to $(1,2,-1)$ through the point $(1,1,-1)$ is $(x-1)+2(y-1)-(z+1)=0\Leftrightarrow x+2y-z=4$.

3.2. Distance of a point from a plane. Consider the plane Π be given by $Ax+By+Cz+D=0$, and the point $P=(u,v,w)$. The straight line passing through P and perpendicular to Π has the parametric equation

$$x=u+At,\ y=v+Bt,\ z=w+Ct,\quad t\in\mathbb{R}$$

and upon substitution in the plane equation we immediately find:

$$t=-\frac{Au+Bv+Cw+D}{A^2+B^2+C^2}$$

so that the intersection of Π and the line is

$$\mathbf{Op}=(u,v,w)-\frac{Au+Bv+Cw+D}{A^2+B^2+C^2}(A,B,C)$$

This is also the projection of P on Π.

As the projection is orthogonal,

$$|\mathbf{Pp}| = \min_{x\in\Pi} |\mathbf{Px}|$$

and therefore the distance between the point and the plane, which is defined precisely as above, is:

$$d(P,\Pi) = \frac{|Au+Bv+Cw+D|}{\sqrt{A^2+B^2+C^2}} \tag{50}$$

Example. The distance of $(1,2,3)$ from $2x-3y+z=10$ is $\dfrac{|2\cdot 1-3\cdot 2+3-10|}{\sqrt{2^2+3^2+1^2}} = \dfrac{11}{\sqrt{14}}$.

4. Tangential plane on a surface

Consider a smooth surface given by $F(x,y,z)=0$. The gradient vector $\nabla F = (F_x, F_y, F_z)$ points towards the direction of fastest increase: indeed, consider Taylor's formula

$$F(\mathbf{r}+\Delta\mathbf{r}) - F(\mathbf{r}) = \nabla F\cdot\Delta\mathbf{r} + O(|\Delta\mathbf{r}|^2)$$

For a $\Delta\mathbf{r}$ of fixed measure, we see that the increase is maximal iff $\Delta\mathbf{r} \uparrow\uparrow \nabla F$. Given that on the surface the value of F is constant, i.e. it does not increase or decrease at all, the direction of maximal increase is to move perpendicular to the surface, so ∇F is perpendicular to the surface. Hence, the tangential plane through $\mathbf{s}$ is

$$\nabla F(\mathbf{s})\cdot(\mathbf{r}-\mathbf{s}) = 0 \tag{51}$$

Example. Let us find the tangent plane at $(1,1,1)$ on the surface $xyz=1$. $F(x,y,z)=xyz-1$, and thus $F_x(x,y,z)=yz \Rightarrow F_x(1,1,1)=1$, $F_y(x,y,z)=xz \Rightarrow F_y(1,1,1)=1$, $F_z(x,y,z)=xy \Rightarrow F_z(1,1,1)=1$. According to (51), the tangent we ask is: $(x-1)+(y-1)+(z-1)=0 \Leftrightarrow x+y+z=3$

5. The outer (or cross) product

5.1. Introduction. The outer product[1] is defined *in three dimensions only.* In this case, choosing two vectors of a basis leaves only one possibility for the third one if we ignore its multiplicative coefficient; in other words, if $\mathbf{x}$ is a vector perpendicular to both of the chose basis vectors, all other vectors with this property are of the form $t\mathbf{x},\ t\in\mathbb{R}$.

The outer product of two vectors returns a vector perpendicular to its arguments, and chooses its norm in an interesting way: it is the area of the parallelogram the two vectors define! It is denoted by $\times$.

5.2. Formulas for the cross product. Let the two vectors be $\mathbf{u_1}=(x_1,y_1,z_1)$ and $\mathbf{u_2}=(x_2,y_2,z_2)$. By (47), we see that the vectors perpendicular to the plane are $t(y_1z_2-z_1y_2, z_1x_2-z_2x_1, x_1y_2-y_1x_2),\ t\in\mathbb{R}$, and the one corresponding to $t=1$ can be written symbolically as:

$$\mathbf{u}_1\times\mathbf{u}_2 = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{vmatrix} \tag{52}$$

We should add a multiplicative constant to the formula, but we shall see that this is the right value.

The norm of this can be seen to be (do it as an exercise, to see an alternative proof of the Cauchy-Schwartz inequality!):

$$|\mathbf{u}_1\times\mathbf{u}_2|^2 = |\mathbf{u}_1|^2|\mathbf{u}_2|^2 - (\mathbf{u_1}\cdot\mathbf{u}_2)^2 = |\mathbf{u}_1|^2|\mathbf{u}_2|^2\left(1-\cos^2\left(\widehat{\mathbf{u_1},\mathbf{u_2}}\right)\right) = |\mathbf{u}_1|^2|\mathbf{u}_2|^2\sin^2\left(\widehat{\mathbf{u_1},\mathbf{u_2}}\right) \Rightarrow$$
$$|\mathbf{u_1}\times\mathbf{u_2}| = |\mathbf{u}_1||\mathbf{u}_2|\left|\sin\left(\widehat{\mathbf{u_1},\mathbf{u_2}}\right)\right|$$

Incidentally, the outer product, as we defined it, gives some additional orientation information: *the outer product points towards the direction that the thumb of our right hand will point, if the rest of the fingers curl in the direction that* $\mathbf{u}_1$ *has to move in order to align itself with* $\mathbf{u}_2$*, tracing the angle between the two vectors* (as already defined in our discussion of the inner product). See Fig. 1.

[1] It is also called *cross product*, because of the symbol "$\times$" we use to denote it

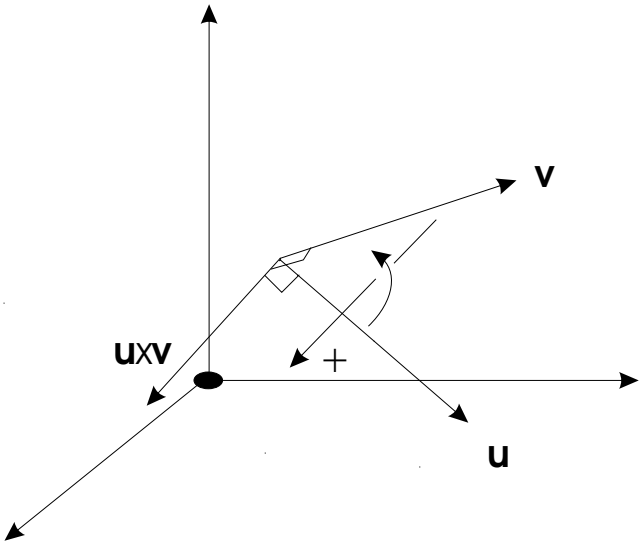

FIGURE 1. Outer product of two vectors: as the fingers of your right hand curl along the arc, your thumb points to the + direction, which is the direction of the outer product.

5.3. Properties. The outer product has some interesting properties immediately checked by the definition:

- $\mathbf{x} \times \mathbf{x} = 0$
- $(\mathbf{x} \times \mathbf{y}) \cdot \mathbf{y} = (\mathbf{x} \times \mathbf{y}) \cdot \mathbf{x} = 0$
- $\mathbf{x} \times \mathbf{y} = -\mathbf{y} \times \mathbf{x}$
- $\mathbf{i} \times \mathbf{j} = \mathbf{k},\ \mathbf{j} \times \mathbf{k} = \mathbf{i},\ \mathbf{k} \times \mathbf{i} = \mathbf{j}$

5.4. Some further results.

(1) *Volume of a parallelepiped — Scalar Triple Product*: If we combine the cross with the dot product, we can find the volume of a parallelepiped. Let us consider the parallelepiped whose sides are parallel to the vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$. There are two sides with edges parallel to $\mathbf{v}, \mathbf{w}$; $\mathbf{v} \times \mathbf{w}$ provides a vector perpendicular to these, with norm equal to the area of one of them. The distance between them is then the projection of $\mathbf{u}$ on $\mathbf{v} \times \mathbf{w}$, hence the volume is

$$V = |\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})| \tag{53}$$

But since we could have started with any side, it is true that:

$$\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = \mathbf{w} \cdot (\mathbf{u} \times \mathbf{v}) = \mathbf{v} \cdot (\mathbf{w} \times \mathbf{u})$$

Since the cross product is given by (5), we get:

$$\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w} = \begin{vmatrix} \mathbf{u_1} & \mathbf{u_2} & \mathbf{u_3} \\ \mathbf{v_1} & \mathbf{v_2} & \mathbf{v_3} \\ \mathbf{w_1} & \mathbf{w_2} & \mathbf{w_3} \end{vmatrix} \tag{54}$$

This is the scalar triple product; the fact that it expresses a volume immediately suggests an application: *Three vectors are coplanar iff their triple scalar product is 0*, because in that case the parallelepiped they define is flat, so its volume is 0!

(2) *Vector Triple product*: $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$ is perpendicular to $\mathbf{v} \times \mathbf{w}$, i.e. perpendicular to the vector perpendicular to $\mathbf{v}, \mathbf{w}$, i.e. a linear combination of $\mathbf{v}, \mathbf{w}$: $\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = a\mathbf{v} + b\mathbf{w}$. Inner product with $\mathbf{u}$ gives that $a(\mathbf{v} \cdot \mathbf{u}) + b(\mathbf{w} \cdot \mathbf{u}) = 0$, so a, b are related: $a = t(\mathbf{w} \cdot \mathbf{u}),\ b = -t(\mathbf{v} \cdot \mathbf{u})$, so that:

$$\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = t(\mathbf{w} \cdot \mathbf{u})\mathbf{v} - t(\mathbf{v} \cdot \mathbf{u})\mathbf{w}$$

But $\mathbf{u}$ can be taken in the plane defined by the other two, as any component it may have perpendicular to this plane will not affect the result; for the same reason, $\mathbf{v}$ can be taken perpendicular to $\mathbf{w}$. Also, by dividing both sides with the product of the norms of the three vectors, we can proceed assuming that the vectors have norm one. Then, we may write $\mathbf{u} = (\mathbf{w}\cdot\mathbf{u})\mathbf{w} + (\mathbf{u}\cdot\mathbf{v})\mathbf{v}$ and substituting into the previous equality we find:

$$(\mathbf{w}\cdot\mathbf{u})\mathbf{w}\times(\mathbf{v}\times\mathbf{w}) + (\mathbf{v}\cdot\mathbf{u})\mathbf{v}\times(\mathbf{v}\times\mathbf{w}) = t(\mathbf{w}\cdot\mathbf{u})\mathbf{v} - t(\mathbf{v}\cdot\mathbf{u})\mathbf{w} = (\mathbf{w}\cdot\mathbf{u})\mathbf{v} - (\mathbf{v}\cdot\mathbf{u})\mathbf{w}$$

so $t = 1$ and

$$\mathbf{u}\times(\mathbf{v}\times\mathbf{w}) = (\mathbf{w}\cdot\mathbf{u})\mathbf{v} - (\mathbf{v}\cdot\mathbf{u})\mathbf{w} \tag{55}$$

Beware! We cannot remove the parentheses, because $(\mathbf{u}\times\mathbf{v})\times\mathbf{w} = -(\mathbf{w}\cdot\mathbf{v})\mathbf{u} + (\mathbf{w}\cdot\mathbf{u})\mathbf{v}$! Associativity does not work!

Example. By using (52), we find that a vector perpendicular to $(1,2,1)$ and $(2,0,-1)$ is $(1,2,1)\times(2,0,-1) = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ 1 & 2 & 1 \\ 2 & 0 & -1 \end{vmatrix} = (-2,3,-4)$.

Example. The volume of the parallelepiped with the vectors $(1,0,0)$, $0,1,1$, and $(1,2,3)$ as sides, is $V = \left|\begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 2 & 3 \end{vmatrix}\right| = \left|\begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{vmatrix}\right| = 1$, by (54).

6. Special surfaces

We are going to examine here *quadratic* surfaces, i.e. those surfaces whose equation is given by a polynomial of the second degree in x, y, z. We will give a procedure by which the general equation can be reduced to a canonical form, and then proceed to classify the different canonical forms we can obtain. Note that this procedure will work in any dimension, although we will use it here specifically for three dimensions.

The general equation of a quadratic surface is $Ax^2+By^2+Cz^2+2Dxy+2Eyz+2Fzx+Gx+Hy+Iz+J=0$, which, in matrix notation, can be written as:

$$\begin{bmatrix} x & y & z \end{bmatrix}\begin{bmatrix} A & D & F \\ D & B & E \\ F & E & C \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} G & H & I \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} + J = 0 \Leftrightarrow \mathbf{r}^T U\mathbf{r} + V\mathbf{r} + W = 0 \tag{56}$$

Here we will need some results about *eigenvectors* and *eigenvalues* in order to proceed. For a given matrix U, consider the equation $U\mathbf{s} = \lambda\mathbf{s}$: the values of λ that satisfy this equation are called eigenvalues of U; the vectors $\mathbf{s}$ that satisfy it are called eigenvectors. Without loss of generality, since any scalar multiple of an eigenvector is still an eigenvector, we can suppose that all eigenvectors are normalized to having unit norm. Obviously, $\mathbf{s} = \mathbf{0}$ is a trivial and obvious eigenvector. In order for nontrivial ones to exist the matrix $U - \lambda I$ must be non-invertible, i.e. its determinant must be zero: $|U - \lambda I| = 0$. This proves that there are as many eigenvalues as the dimension of the matrix, possibly non-different and possibly complex. But our matrix U is symmetric, and the following facts are true (given without proof):

- The eigenvalues are real.
- There are as many eigenvectors as the dimension of the matrix.
- Any two different eigenvectors are orthogonal.

In vector notation then we can write:

$$U\mathbf{s}_1 = \lambda_1\mathbf{s}_1,\ U\mathbf{s}_2 = \lambda_2\mathbf{s}_2,\ U\mathbf{s}_3 = \lambda_3\mathbf{s}_3 \Leftrightarrow U\begin{bmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \mathbf{s}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \mathbf{s_3} \end{bmatrix}\operatorname{diag}(\lambda_1,\lambda_2,\lambda_3) \Leftrightarrow$$
$$US = S\Lambda \Leftrightarrow U = S\Lambda S^{-1} = S\Lambda S^T$$

since S is orthogonal.

By the substitution $\mathbf{q} = S^T\mathbf{r}$, (56) becomes:

$$\mathbf{q}^T\Lambda\mathbf{q} + VS\mathbf{q} + W = 0 \Leftrightarrow \lambda_1 q_1^2 + \lambda_2 q_2^2 + \lambda_3 q_3^2 + \mu_1 q_1 + \mu_2 q_2 + \mu_3 q_3 + W = 0 \tag{57}$$

The next step is to apply, for all non-zero eigenvalues, the following procedure (we omit the index for simplicity). Therefore, the substitutions:

(1) Observe that, by completing the square:

$$\lambda q^2 + \mu q = \lambda\left(q^2 + \frac{\mu}{\lambda}q\right) = \lambda\left(q^2 + 2\frac{\mu}{2\lambda}q\right) = \lambda\left(q^2 + \frac{\mu}{\lambda}q + \frac{\mu^2}{\lambda^2} - \frac{\mu^2}{\lambda^2}\right) = \lambda\left(q + \frac{\mu}{\lambda}\right)^2 - \frac{\mu^2}{\lambda}$$

(2) $t = q + \dfrac{\mu}{\lambda}$

(3) $y = \sqrt{|\lambda|}t$

turn the eigenvalue/coefficient λ into $\text{sign}(\lambda) = \pm 1$, and eliminate the linear term. Having performed this step for all non-zero eigenvalues, (57) turns into:

$$\sum_{\{i:\lambda_i \neq 0\}} \text{sign}(\lambda_i) y_i^2 + \sum_{\{i:\lambda_i = 0\}} \mu_i q_i + C = 0 \tag{58}$$

where

$$C = W - \sum_{\{i:\lambda_i \neq 0\}} \frac{\mu^2}{\lambda}$$

The final step is the following:

- If $|\{i : \lambda_i = 0\}| > 0$, substitute:

$$y = \sum_{\{i:\lambda_i = 0\}} \mu_i q_i + C$$

which corresponds to a translation, scaling, and possibly a rotation by 180^o. The final form of (58) becomes then:

$$\sum_{\{i:\lambda_i \neq 0\}} \text{sign}(\lambda_i) y_i^2 + y = 0 \tag{59}$$

- If $|\{i : \lambda_i = 0\}| = 0$, and $C \neq 0$ substitute $y_i \to \sqrt{|C|}y_i$ for all indexes remaining in (58), which, by elimination of $|C|$, becomes:

$$\sum_{\{i:\lambda_i \neq 0\}} \text{sign}(\lambda_i) y_i^2 + \text{sign}(C) = 0 \tag{60}$$

Clearly, there are many trivial cases: If $\lambda_i = 0,\ i = 1, \ldots, n$, then the surface is just a plane; if all λ and C are of the same sign in (60), then the equation is impossible and represents nothing, whereas if C=0 the equation is just a point. Let us now go through the (interesting and non-trivial) different cases (the ones not listed become similar to one of those in the list after reordering the axes):

- $y_1^2 + y_2^2 + y_3^2 = 1$ is the *sphere*.
- $y_1^2 + y_2^2 = y_3^2$ is the *cone* with the y_3-axis as its axis of symmetry.
- $y_1^2 + y_2^2 = y_3$ is the *paraboloid* symmetric around the y_3-axis.
- $y_1^2 + y_2^2 - y_3^2 = 1$ is the *one-sheeted hyperboloid* symmetric around the y_3-axis.
- $y_3^2 - y_1^2 - y_2^2 = 1$ is the *two-sheeted hyperboloid* symmetric around the y_3-axis.
- $y_2^2 - y_1^2 = y_3$ is the *hyperbolic paraboloid*.

The following forms can be obtained from the ones above by undoing the normalization of the eigenvalues, and by switching the notation to (x, y, z) coordinates. Because they are extremely common, they are listed separately:

- Ellipsoid (un-normalized sphere):

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 + \left(\frac{z}{c}\right)^2 = 1$$

- Sphere:

$$x^2 + y^2 + z^2 = r^2$$

- (Elliptic) Paraboloid:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=z$$

- (Elliptic) Cone:

$$\left(\frac{x}{a}\right)^2+\left(\frac{y}{b}\right)^2=z^2$$

- Two-Sheeted Hyperboloid:

$$\left(\frac{z}{c}\right)^2-\left(\frac{x}{a}\right)^2-\left(\frac{y}{b}\right)^2=1$$

- One-Sheeted Hyperboloid:

$$\left(\frac{x}{a}\right)^2+\left(\frac{y}{b}\right)^2-\left(\frac{z}{c}\right)^2=1$$

- Hyperbolic paraboloid:

$$\left(\frac{y}{b}\right)^2-\left(\frac{x}{a}\right)^2=z$$

Example. Let us try to classify the surface given by: $3x^2+y^2+2xz+x+2z+10=0$. By writing it in the form of (56) we find that $U=\begin{bmatrix}3&0&2\\0&1&0\\2&0&0\end{bmatrix}$, whose eigenvalues are $\lambda=1$, $\lambda=-1$, and $\lambda=4$, corresponding to the eigenvectors $(0,1,0)$, $\left(\frac{1}{\sqrt{5}},0,-\frac{2}{\sqrt{5}}\right)$, and $\left(\frac{2}{\sqrt{5}},0,\frac{1}{\sqrt{5}}\right)$. We also find that $V=\begin{bmatrix}1&0&2\end{bmatrix}$, and $W=10$. S contains the eigenvectors as columns in the order we gave them.

Subsequently, (57) gives: $q_1^2-q_2^2+4q_3^2-\frac{3}{\sqrt{5}}q_2+\frac{4}{\sqrt{5}}q_3+10=0$, and (58) gives: $y_1^2-y_2^2+y_3^2+11=0$. Therefore, (60) gives $y_2^2-y_1^2-y_3^2=1$, which is a two-sheeted hyperboloid.

6.1. Cylinders. Consider a curve on the plane given by $F(x,y)=0$. The same equation in space describes a *cylinder* generated by the translation of this curve along the z-axis.

6.2. Surfaces generated by rotations of curves. Consider a curve on the plane given by $F(x,y)=0$. We can rotate it:

- around the y-axis, to get a surface given by the equation

$$F\left(\sqrt{x^2+z^2},y\right)=0$$

- around the x-axis, to get a surface given by the equation

$$F\left(x,\sqrt{y^2+z^2}\right)=0$$

For more generality, dilate the z-axis by $c>0$ to get:

- Rotation around the y-axis:

$$F\left(\sqrt{x^2+\left(\frac{z}{c}\right)^2},y\right)=0 \tag{61}$$

- Rotation around the x-axis:

$$F\left(x,\sqrt{y^2+\left(\frac{z}{c}\right)^2}\right)=0 \tag{62}$$

6.2.1. *The torus.* Consider the ellipse $\left(\frac{x-u}{a}\right)^2+\left(\frac{y}{b}\right)^2=1$. If we rotate it around the y-axis, we get:

$$\left(\frac{\sqrt{x^2+z^2}-u}{a}\right)^2+\left(\frac{y}{b}\right)^2=1$$

This generates a "doughnut" shape, like the wheel of the car, with an elliptic cross-section (or circular iff $a=b$).

7. Examples of rotation in 3D

Rotations in 3D are in general quite complicated; rather than providing unwieldy general formulas, we will explain the procedure of

a) writing the rotation matrix for rotation around a given vector, and
b) finding the equation of a rotated curve or surface

in two detailed examples.

Example. Given $\hat{\mathbf{n}}=\frac{1}{\sqrt{3}}(1,1,1)$ and $\mathbf{v}=(1,2,-1)$, we will find the vector $\mathbf{v}'$ resulting from the rotation of $\mathbf{v}$ around $\hat{\mathbf{n}}$ by $\theta=30^o$. We will divide the solution in stages:

(1) Make sure that $\hat{\mathbf{n}}$ is normalized to unit norm (it is here).
(2) Set $\hat{\mathbf{z}}'=\hat{\mathbf{n}}$. Our goal is to incorporate this vector in a new orthonormal basis, so we still need $\hat{\mathbf{x}}'$ as $\hat{\mathbf{y}}'$.
(3) Choose any vector of unit norm perpendicular to $\hat{z}'$ as $\hat{y}'$. This is simple: $(-b,a,0)$ is always perpendicular to (a,b,c)!. Choose then $(-1,1,0)$, and normalize to get $\hat{\mathbf{y}}'=\frac{1}{\sqrt{2}}(-1,1,0)$.
(4) Finally, in order to have a right-handed system, we need to make sure that the properties at the end of section 5 are satisfied. These give us a way to compute $hat\mathbf{x}'$:

$$\hat{\mathbf{x}}'=\hat{\mathbf{y}}'\times\hat{\mathbf{z}}'=\frac{1}{\sqrt{6}}(1,1,-2)$$

You do not need to worry about normalization here: if you were careful to normalize $\hat{\mathbf{y}}'$ and $\hat{\mathbf{z}}'$, $\hat{\mathbf{x}}'$ will be automatically normalized.
(5) The problem has now come down to the determination of the rotation matrix around the desired axis. We have no clue what this matrix looks like, but we do know how to rotate around the z-axis: indeed, this is just a rotation on the xy-plane, leaving the z coordinate intact, and we know how to rotate vectors on the plane! The rotation matrix is:

$$R_z(\theta)=\begin{bmatrix}\cos(\theta) & -\sin(\theta) & 0\\ \sin(\theta) & \cos(\theta) & 0\\ 0 & 0 & 1\end{bmatrix}$$

(6) Now we can change the basis from $[\hat{\mathbf{x}},\hat{\mathbf{y}},\hat{\mathbf{z}}]$ to $[\hat{\mathbf{x}}',\hat{\mathbf{y}}',\hat{\mathbf{z}}']$; in the latter basis the rotation matrix we seek is just R_z. Remember the formulas $E=E'P\Leftrightarrow P=(E')^TE$, $\mathbf{r}'=P\mathbf{r}$, and $R_z=PR_nP^T\Leftrightarrow R_n=P^TR_zP$, where E and E' (along with P) are orthogonal matrices and have as columns the old and the new basis vectors, respectively. In our case $E=I$, so $P=(E')^T$. In these formulas we made extensive use of the fact that the inverse of an orthonormal matrix is its transpose.
(7) We are ready now to finish the computation:

$$\begin{bmatrix}x'\\y'\\z'\end{bmatrix}=P^TA_zP\begin{bmatrix}x\\y\\z\end{bmatrix}=\begin{bmatrix}\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}}\\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}}\\ -\frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{3}}\end{bmatrix}\begin{bmatrix}\frac{\sqrt{3}}{2} & -\frac{1}{2} & 0\\ \frac{1}{2} & \frac{\sqrt{3}}{2} & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}}\\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0\\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}}\end{bmatrix}\begin{bmatrix}x\\y\\z\end{bmatrix}=$$

$$=\begin{bmatrix}0.9107 & -0.2441 & 0.3333\\ 0.3333 & 0.9107 & -0.2441\\ -0.2441 & 0.3333 & 0.9107\end{bmatrix}\begin{bmatrix}x\\y\\z\end{bmatrix}$$

hence $\mathbf{v}' = (0.0893, 2.3987, -0.4880)$.

Example. Consider the cone $x^2 + y^2 = z^2$, whose symmetry axis is the z-axis. We will now rotate it so that its symmetry axis becomes $\hat{n}$ of the previous example, and give its new formula. We will rely on the computations carried out in the previous example.

Choose $\hat{\mathbf{z}}' = \hat{\mathbf{n}}'$, $\hat{\mathbf{y}}'$, and $\hat{\mathbf{x}}' = \hat{\mathbf{y}}' \times \hat{\mathbf{z}}'$ as before, thus finding a new basis, in which the equation of the cone is simply $x'^2 + y'^2 = z'^2$. But the transformation formula is $\mathbf{r}' = P\mathbf{r}$, where $P = (E')^T E = (E')^T$, as $E = I$. We computed P in the previous example, and found it:

$$P = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix} \Longrightarrow \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{6}}(x + y - 2z) \\ \frac{1}{\sqrt{2}}(x - y) \\ \frac{1}{\sqrt{3}}(x + y + z) \end{bmatrix}$$

so the cone becomes:

$$\frac{(x + y - 2z)^2}{6} + \frac{(x - y)^2}{2} = \frac{(x + y + z)^2}{3} \Leftrightarrow x^2 + y^2 + z^2 - 4xy - 4yz - 4zx = 0$$

CHAPTER 5

Orthogonal projections

In many occasions we wish to reduce the complexity of a curve or surface. An obvious way to do that is to reduce its dimensions, so that a three-dimensional surface is represented on the plane, or a two-dimensional curve as an one-dimensional one. If you think about it, you will see that this is what painting is all about: we represent the real three-dimensional world on a sheet of paper! We create the illusion of perspective, the third-dimension, by some optical tricks, e.g. by drawing more distant objects smaller.

A different kind of projection, but still very useful, is the one we use to create a wall map of the world. We now project a sphere on the cylinder, but the projection is no more orthogonal!

1. Orthogonal projection of a point on a plane

We have already derived this formula in (50), in finding the distance of a point from a plane. Let the plane Π be given by $Ax + By + Cz + D = 0$, and the point P be given by $\mathbf{OP} = (u, v, w)$. The projection p of P on Π is

$$\mathbf{Op} = (u, v, w) - \frac{Au + Bv + Cw + D}{A^2 + B^2 + C^2}(A, B, C) \tag{63}$$

The above formula can be used to find the projection of a curve or a surface on the plane if the parametric representation of the curve or the surface is given.

Example. The projection of $(1, 2, 3)$ on the plane $x + y + z = 1$ is $(1, 2, 3) - \frac{1+2+3-1}{1^2+1^2+1^2}(1, 1, 1) = (1, 2, 3) - \frac{5}{3}(1, 1, 1) = \frac{1}{3}(-2, 1, 4)$

2. Orthogonal projection on a sphere

Suppose we want to project the point P, given by $\mathbf{OP} = (u, v, w)$, on the sphere with center (X, Y, Z) and radius r. The equation of the line passing through the point and the center of the sphere is:

$$x = X + t(u - X),\ y = Y + t(v - Y),\ x = Z + t(w - Z),\quad t \in \mathbb{R}$$

The intersection of this line with the sphere is the projection. We immediately see that t can have two possible values:

$$t = \pm\frac{r}{\sqrt{(u - X)^2 + (v - Y)^2 + (w - Z)^2}}$$

which is reasonable, as a straight line going through the center of the sphere intersects with it at two points. The intersection point we are interested in is the one which is closer to the original. Now, at $t = 0$ is the center of the sphere, whereas at $t = 1$ the point P, hence we must have $t > 0$ at the projection! The projection therefore is:

$$\mathbf{Op} = (X, Y, Z) + \frac{r}{\sqrt{(u - X)^2 + (v - Y)^2 + (w - Z)^2}}(u - X, v - Y, w - Z) \tag{64}$$

CHAPTER 6

Projective geometry and perspective

In *affine geometry*, which could be regarded as mainly concerned with properties of points, lines and planes in $\mathbb{R}^3$ it is often difficult to make general statements because objects can be parallel to each other. For example, two lines in $\mathbb{R}^2$ need not meet in a point: they do meet in a point if and only if they are not parallel. In projective geometry, one studies objects that are a lot like ordinary lines and planes, but in a setting where *any two* lines in a plane, meet in a point. There is no such thing as 'parallel' in projective geometry. Very roughly speaking, the projective plane $\mathbb{P}^2$ is obtained from the ordinary plane $\mathbb{R}^2$ by adding "ideal points" aka "points at infinity". These are added in such a way that lines parallel in $\mathbb{R}^2$ meet in a 'point at infinity'. (There is one point at infinity for each possible *direction* of parallel lines.

It turns out that projective geometry is the right setting for a study of *perspective*, the art/science of accurate depiction of three-dimensional objects as seen by the human eye. (It is not at all clear at first what this has to do with parallel lines!) In understanding perspective, the key observation is that we see objects through the light rays that enter our eyes having been scattered off the object we're looking at. To a good approximation, such light-rays are straight lines through a point (our pupil). (We ignore the fact that we have two eyes for the moment.) Thus our vision is bound up with the study of the set of straight lines (light-rays) in $\mathbb{R}^3$ through a given point (our pupil). In a moment we shall *define* $\mathbb{P}^2$ to be exactly the set of all lines through the origin in $\mathbb{R}^3$.

There is now a relation with parallel lines, for consider the classic photograph of a straight railway track receding into the distance. The rails are pretty much parallel lines, yet in the photo, there is a definite point on the horizon where they appear to meet.

Using projective geometry, we can construct mathematical projections of 3D objects onto a plane (think of it as a photograph) which give an accurate depiction of that object as it would appear to us. A related question is how different people view the same 3D object. It turns out that there is a nice simple answer: any two such views are related by a projective transformation and these can be written in terms of matrices, even though they are strictly more general than the linear or affine transformations that we have met so far.

As we shall see (or as you can see if you think about photos of objects from different angles and distances) a lot of standard euclidean geometry goes out of the window. Angles, lengths, and even ratios of lengths, can change, in contrast to the rigid motions of euclidean geometry. However, there are more subtle quantities that *cannot* be altered! We shall look at the simplest: the cross-ratio of 4 points on a line.

1. Definition of projective space

By a *ray* through the origin, we simply mean an (undirected) line through the origin, usually of $\mathbb{R}^3$. The set of all rays through O in $\mathbb{R}^3$ is called the *(real) projective plane* and denoted by $\mathbb{RP}^2$, or just $\mathbb{P}^2$. Equivalently, $\mathbb{P}^2$ is the set of all 1-dimensional (vector) subspaces of $\mathbb{R}^3$. More generally, projective n-space $\mathbb{P}^n$ is defined to be the set of all 1-dimensional (vector) subspaces of $\mathbb{R}^{n+1}$. In particular, the (real) projective line $\mathbb{P}^1$ the set of 1-dimensional subspaces of $\mathbb{R}^2$ (or rays through the origin of $\mathbb{R}^2$).

1.1. Homogeneous coordinates. Let L be a ray through the origin of $\mathbb{R}^3$. Then L is completely determined by any non-zero point (X, Y, Z) on it. However, if (X, Y, Z) is on L then so is any multiple $(\lambda X, \lambda Y, \lambda Z)$. Therefore it is only the ratios of X to Y to Z that are important in determining L. We write $(X : Y : Z)$ for the *homogeneous coordinates* of the point of $\mathbb{P}^2$ corresponding to the ray L through (X, Y, Z). NB $(0 : 0 : 0)$ are not allowable homogeneous coordinates and do not represent any point of $\mathbb{P}^2$.

Similarly, homogeneous coordinates $(X_0 : X_1 : \ldots : X_n)$ are used to represent points of $\mathbb{P}^n$.

Example. $(1:0) = (4:0) \neq (9:1)$ in $\mathbb{P}^1$. $(9:1:2) = (-3:-1/3:-2/3)$ in $\mathbb{P}^2$.

2. Perspective mapping of $\mathbb{P}^2$

If we pick a plane π not through O, say

$$ax + by + cz = 1, \tag{65}$$

then $\mathbb{P}^2$ gets divided into two disjoint subsets V_π and I_π called the set of *visible points* and the set of *ideal points* of π. Geometrically, V_π consists of those rays through O in $\mathbb{R}^3$ which meet π; I_π consists of the rays through O that do not meet π. If a ray L is in I_π, then it must lie entirely in the plane π' through O parallel to π. In fact, I_π is the set of all rays through O contained in π', so it is a real projective line.

2.1. Definition. The *perspective projection* with centre O and view-plane π is the mapping $f_\pi : V_\pi \to \pi$ which maps the ray OP to the intersection point $OP \cap \pi$.

We note that f_π is a 1:1 correspondence (bijection) between the set of visible points (rays) V_π and the points of π itself. To sum up, we see that $\mathbb{P}^2 = V_\pi \cup I_\pi$ can be thought of the plane π, together with a projective line of 'ideal points' aka points at infinity.

2.2. Formulae for perspective projection. The ray through (X, Y, Z) is given by the parametric equation

$$x = \lambda X,\ y = \lambda Y,\ z = \lambda Z \ \ (\lambda \in \mathbb{R}).$$

The value of λ for which (x, y, z) satisfies (65) is given by

$$\lambda(aX + bY + cZ) = 1.$$

This has a solution for λ iff $aX + bY + cZ \neq 0$, so

$$V_\pi = \{(X:Y:Z) \text{ such that } aX + bY + cZ \neq 0\},\ \ I_\pi = \{(X:Y:Z) \text{ such that } aX + bY + cZ = 0\}.$$

This confirms the geometric reasoning above that a point is in I_π if the corresponding ray lies in the plane π' parallel to π through O. (For the equation of π' is $aX + bY + cZ = 0$.)

If $(X:Y:Z) \in V_\pi$, then we have

$$f_\pi(X, Y, Z) = \left(\frac{X}{aX + bY + cZ}, \frac{Y}{aX + bY + cZ}, \frac{Z}{aX + bY + cZ} \right). \tag{66}$$

2.3. Remark. We shall use the term "perspective projection" both for the map from $\mathbb{R}^3$ to π and from $\mathbb{P}^2$ to π. The above formula makes sense in either case, it's just a question of whether you think of (X, Y, Z) as homogeneous coordinates or not.

Example. Determine the homogeneous coordinates of the set of points visible from $4x + 5y + 6z = 1$ and the ideal points of this plane.

Solution: The point $(X:Y:Z)$ determines the ray

$$x = \lambda X, y = \lambda Y, z = \lambda Z,$$

and this meets the given plane provided the equation

$$\lambda(4X + 5Y + 6Z) = 1$$

can be solved. Hence the set of visible points is

$$V_\pi = \{(X:Y:Z) \in \mathbb{P}^2 : 4X + 5Y + 6Z \neq 0\}.$$

The set of ideal points is the complement of this,

$$I_\pi = \{(X:Y:Z) \in \mathbb{P}^2 : 4X + 5Y + 6Z = 0\}.$$

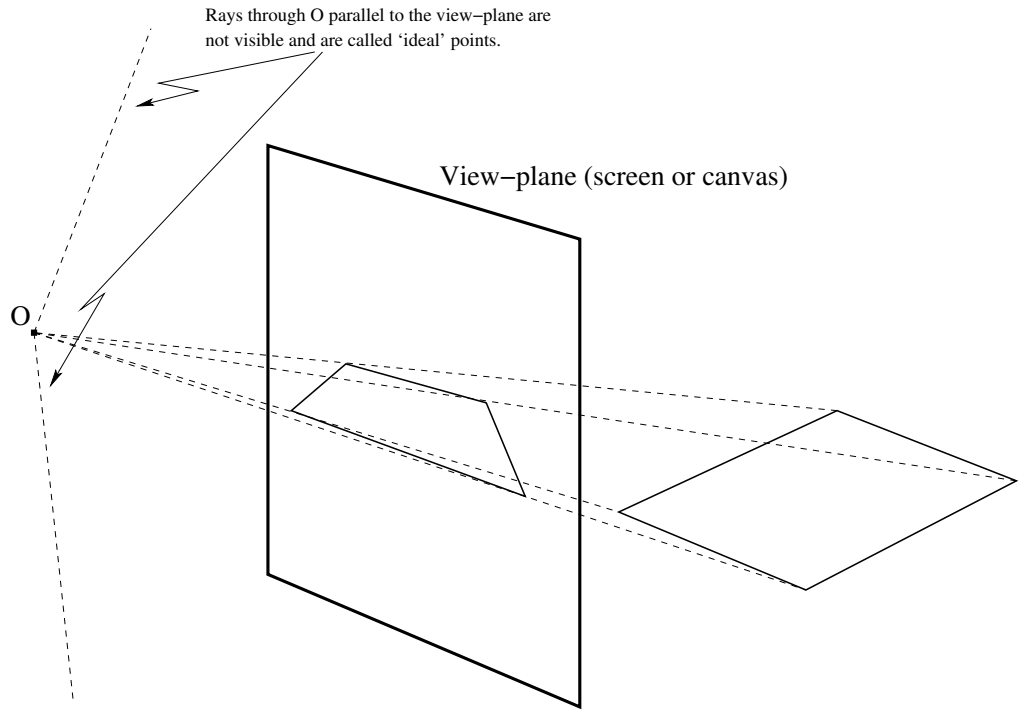

FIGURE 1. Perspective projection with centre O to a view-plane π.

Example. Suppose a sign lies on the plane $y = -d$, and that its borders lie on the lines $z = -e$, $z = a$, $x = b$, and $x = c$. Compute its perspective projection on the plane $x = 1$ through the origin.

Solution:

The given border lines project onto the plane $x = 1$ as $(1, -d/x, -e/x)$, $(1, -d/x, a/x)$, $(1, -d/b, z/b)$, and $(1, -d/c, z/c)$. If you plot this, you will find out that it corresponds exactly to what you would draw on the paper by experience.

2.4. Remark. Perspective projections were used by artists since the Renaissance (and for computer graphics applications) to help give accurate representations of 3D scenes in a 2D plane. In this case the centre of perspective O is the viewer's (or artist's) eye and the canvas (or computer screen) is the plane π. The rays through O are the light rays that enter the eye after having been scattered by objects in the 3D scene. (Fig. 1.)

3. Properties of perspective projection

If we take π to be the plane $z = 1$ (so $a = b = 0$, $c = 1$ in (65)) then we can identify the point $(x, y, 1)$ of π with the point (x, y) of $\mathbb{R}^2$, and this is helpful in trying to understand properties of perspective projection. We denote this projection by f, so that

$$f(X, Y, Z) = (X/Z, Y/Z, 1), \text{ and } (x, y) = (X/Z, Y/Z). \tag{67}$$

In particular, f is not a *linear* map. It is intuitively clear that f does not generally preserve lengths or angles (see Fig. 1).

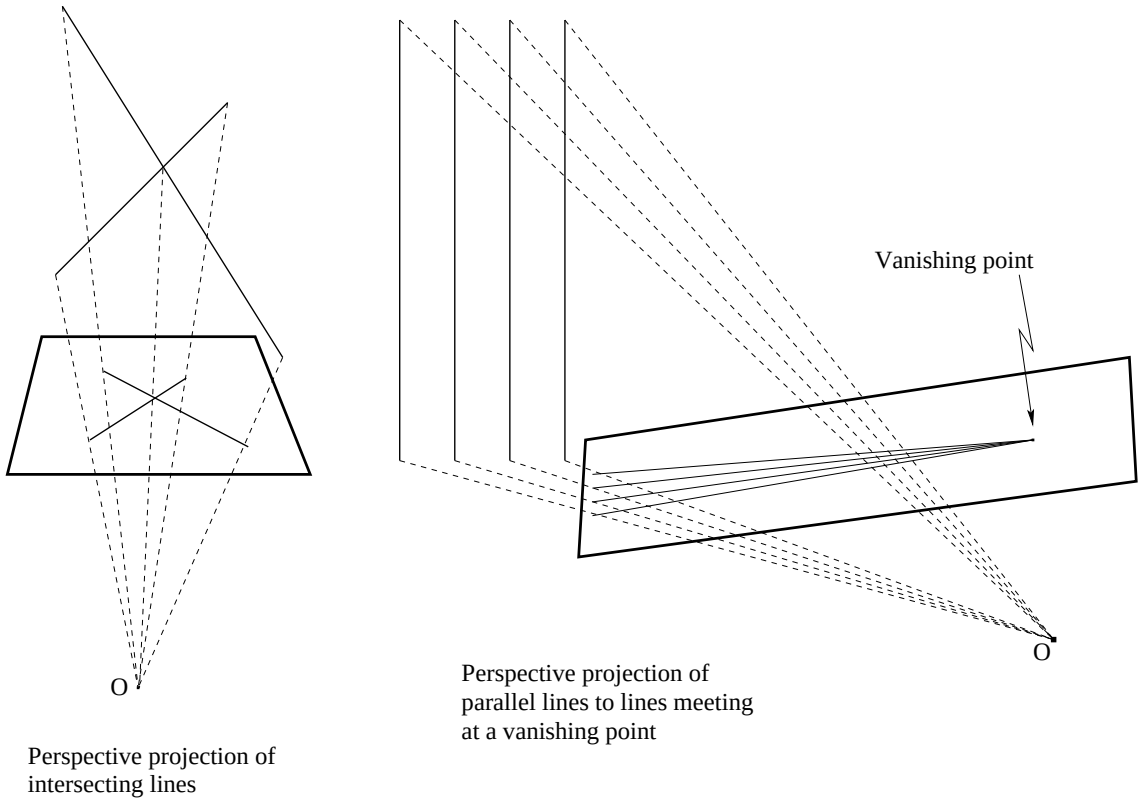

FIGURE 2. Perspective projection of intersecting lines and of parallel lines

However, f does map straight lines in 3D space to straight lines in π. This can be seen geometrically or algebraically. Geometrically, a line ℓ determines a unique plane σ containing O and ℓ. The image $f(\ell)$ by the perspective projection is just the intersection $\pi \cap \sigma$, which is a line.

It is clear that if lines ℓ and ℓ' meet in a point P in $\mathbb{R}^3$, then their projections $f(\ell)$ and $f(\ell')$ meet in $f(P)$. (See Figure 2.)

Alternatively, we can use our formula (67) for the perspective projection to π to obtain an explicit formula for $f(\ell)$, as a parameterized line.

To do this, suppose that ℓ passes through $\mathbf{A} = (A_1, A_2, A_3)$ and is in the direction of the vector $\mathbf{U} = (U_1, U_2, U_3)$. Then a parametric representation of ℓ is given by

$$\ell : \ X = A_1 + U_1 t, \ Y = A_2 + U_2 t, \ Z = A_3 + U_3 t, \ \ (t \in \mathbb{R}). \tag{68}$$

To find $f(\ell)$, we simply substitute these values of (X, Y, Z) into (67), to get the parametric equation

$$f(\ell) : \ x = \frac{A_1 + U_1 t}{A_3 + U_3 t}, \ y = \frac{A_2 + U_2 t}{A_3 + U_3 t}, \ \ (t \in \mathbb{R}).$$

This is a disguised version of the parametric equation of a line in the (x, y) plane. The disguise is thinnest if $U_3 = 0$, which corresponds to ℓ being parallel to π. Then we get

$$x = \frac{A_1}{A_3} + \frac{U_1}{A_3} t, \ y = \frac{A_2}{A_3} + \frac{U_2}{A_3} t,$$

which is certainly the equation of a straight line in the (x, y) plane. (This goes haywire if $A_3 = 0$ as well, but this is precisely the case that all points of ℓ are ideal points of π (i.e. not visible from π.)

If, on the other hand, $U_3 \neq 0$, we may put

$$s = \frac{U_3 t}{A_3 + U_3 t}, \text{ so that } 1 - s = \frac{A_3}{A_3 + U_3 t}. \tag{69}$$

Then after some calculation,

$$f(\ell): \ x = \frac{A_1}{A_3}(1-s) + \frac{U_1}{U_3}s, y = \frac{A_2}{A_3}(1-s) + \frac{U_2}{U_3}s.$$

Hence we see that

$$f(\ell) \text{ is the straight line joining } (A_1/A_3, A_2/A_3) \text{ to } (U_1/U_3, U_2/U_3). \tag{70}$$

4. Vanishing points

If $U_3 \neq 0$ (so that ℓ is not parallel to π) there is a point on $f(\ell)$ that is not on ℓ. This is the so-called *vanishing point* of ℓ. It corresponds to the illegal parameter value $t = \infty$ which translates to the perfectly ordinary parameter value $s = 1$ in the change of variables (69).

$$\text{the vanishing point of } \ell \text{ has coordinates } (U_1/U_3, U_2/U_3). \tag{71}$$

4.1. Projection of parallel lines. By varying $\mathbf{A}$ in (68), while keeping (U_1, U_2, U_3) fixed, we generate a family of parallel lines in $\mathbb{R}^3$ (all in the direction $\mathbf{U}$). Since the coordinates of the vanishing point of ℓ are independent of $\mathbf{A}$, we see that these parallel lines all have the same vanishing point in π. In other words, *parallel lines in 3D space project to lines that meet in a particular point of* π. The only exceptions are families of parallel lines with $U_3 = 0$. Such a family of lines is parallel to π and projects to a family of parallel lines in π.

Note that the vanishing point 'explains' the familiar photo of parallel railway lines meeting at a point on the horizon! See Fig. 2.

4.2. Lines of vanishing points. Consider next the vanishing points of lines that lie in a plane σ in 3D space. If the plane is perpendicular to a vector $\mathbf{E} = (E_1, E_2, E_3)$, then the direction vectors (U_1, U_2, U_3) of lines in σ all satisfy $\mathbf{E} \cdot \mathbf{U} = 0$, i.e.

$$E_1 U_1 + E_2 U_2 + E_3 U_3 = 0.$$

Dividing by U_3, we obtain

$$E_1(U_1/U_3) + E_2(U_2/U_3) + E_3 = 0$$

so that the vanishing points $(U_1/U_3, U_2/U_3)$ of such a family of lines lie on the straight line $E_1 x + E_2 y + E_3 = 0$ in π. (The exception is the case $E_1 = E_2 = 0$, $E_3 = 1$. Then we get the family of lines parallel to the view-plane π and these project to parallel lines in π.)

5. Perspective projection of a coordinate grid

We have seen how straight lines are mapped by perspective projection and that parallel lines (generally) map to lines passing through a vanishing point of π. We can also use our formula (67) to determine the image of an entire coordinate grid on a given plane σ. This has practical applications: suppose we want an accurate representation on the screen π of the facade of a building or a wall covered in wall-paper. If this is to be done with a computer, then one would describe the pattern in terms of a coordinate system 'stuck to the wall'. Mathematically, this corresponds to parameterizing the plane σ by

$$\sigma = \{\mathbf{A} + s\mathbf{U} + t\mathbf{V} : (s,t) \in \mathbb{R}^2\}$$

where $\mathbf{A}$, $\mathbf{U}$ and $\mathbf{V}$ are some fixed vectors. (In this way, (s,t) become coordinates on the plane σ.)

Taking the view-plane to be $z = 1$ as before, we apply (67) to the space-coordinates

$$X = A_1 + U_1 s + V_1 t, Y = A_2 + U_2 s + V_2 t, Z = A_3 + U_3 s + V_3 t$$

of the point with parameter value (s,t) in the plane, and get

$$x = \frac{A_1 + U_1 s + V_1 t}{A_3 + U_3 s + V_3 t}, \ y = \frac{A_2 + U_2 s + V_2 t}{A_3 + U_3 s + V_3 t}. \tag{72}$$

This formula gives an answer to our question: it gives the mapping, in terms of the vectors **A**, **U**, **V** specifying the plane σ (and the coordinate system within it) to the plane of the canvas (with coordinates (x, y).) Such formulae are used several times per second in the graphics associated with arcade games.

Example. Find the images of the lines $s =$ constant, $t =$ constant in the parameterized plane

$$\sigma = \{s\mathbf{i} + 10\mathbf{j} + t\mathbf{k}\}$$

Solution The general point is $X = s$, $Y = 10$, $Z = t$, so $x = X/Z = s/t$, $y = Y/Z = 10/t$. Consider for example $s = 0$, $s = \pm 1$ and $s = \pm 2$. If $s = 0$, we get $x = 0$ (the y-axis). If $s = \pm 1$, we get

$$x = \pm 1/t, \;\; y = 10/t, \text{ so that } y = \pm 10x.$$

Similarly if $s = \pm 2$, we get $y = \pm 5x$. And so on. Since all these go through the origin, this must be the vanishing point for this family of parallel lines.

Next, consider $t = 0$. Our formulae blow up for this line, so this line consists entirely of 'ideal' points. Then if $t = \pm 1$, we find $y = \pm 10$ (and x is free). If $t = \pm 2$, we find $y = \pm 5$. And so on.

6. Geometric construction

Although the formula (72) is practical for computers, it is not practical for an artist on location. There are, however, relatively simple *geometric constructions* for the perspective projection of a coordinate grid in a given plane. The process is illustrated in Figs. 3–6.

The input is a grid of congruent rectangles in a plane σ. We shall refer to the lines determining the grid as 'horizontal' and 'vertical', to avoid unnecesarily complicated notation. We label the intersection points of the grid $P(m, n)$, where (m, n) are integers. (So $P(m, n)$ is the point of intersection of the m-th vertical line and the n-th horizontal line.)

Notation The image of $P(m, n)$ by the projection f will be called $Q(m, n)$.

In addition to this grid, we must decide where to place on π the images of four points. One possibility is to choose where to place the vanishing point H, say, of the horizontal lines, the vanishing point V, say, of the vertical lines, and the images $Q(0, 0)$ and $Q(1, 1)$ of $P(0, 0)$ and $P(1, 1)$ respectively. In the illustrations, we have chosen the vanishing point of the vertical lines in σ to be 'ideal', so that they project to parallel lines in π.

Anyhow, having chosen these four points, we determine $Q(0, 1)$ and $Q(1, 0)$ as in Fig. 3. Thus we have the image of one of our rectangles on π.

The next step is to determine the line v of vanishing points for lines in π. This must go through H and V, so that is simple. Because V is 'ideal' in our example, v is the line through H, parallel to the line through $Q(0, 0)$ and $Q(0, 1)$. Now we can determine the vanishing points L and R, say, in π of the two families of diagonal lines of the grid. Thus L is the intersection of v with the line joining $Q(0, 1)$ and $Q(1, 0)$ and R is the intersection of v with the line joining $Q(0, 0)$ and $Q(1, 1)$. (Fig. 4)

Now we can determine further points of the grid as follows. We can draw the diagonals through $Q(1, 1)$ and $Q(0, 0)$, by joining these points to the vanishing points L and R. They cross the horizontal and vertical lines at $Q(2, 0)$, $Q(2, 1)$, $Q(0, -1)$ and $Q(0, 2)$. We can now draw more grid lines by connecting these to the vanishing points V and H. The process continues, successively drawing diagonals, determining more of the $Q(m, n)$, then determining more of the grid lines.

7. Cross-ratio

Perspective projections do not preserve length or angle, though they do map straight lines to straight lines. As we've seen they do not take parallel lines to parallel lines, so they are apparently very different from affine transformations. From this, you might think that perspective projections are so 'flexible' that all notions of distance are lost. In fact, this is not the case.

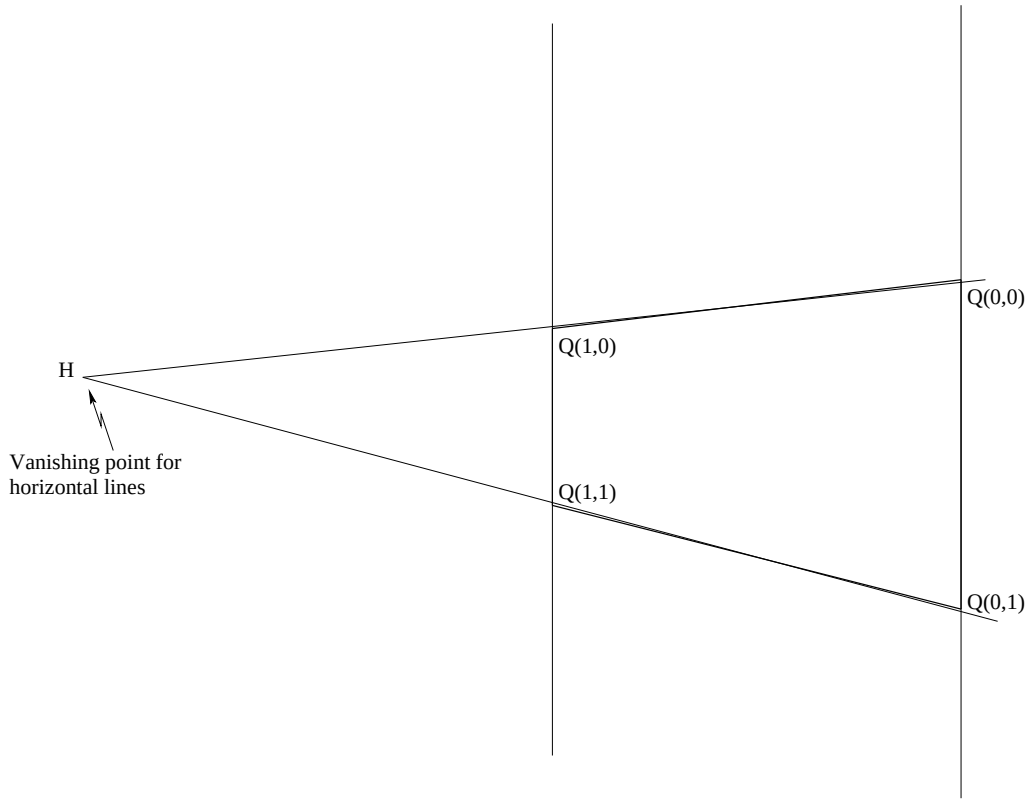

FIGURE 3. Perspective projection of a rectangular grid: first step

7.1. Definition. Let P_1, P_2, P_3, P_4 be four distinct points on a line ℓ. The *cross-ratio* of the ordered set (P_1, P_2, P_3, P_4) is the quantity:

$$\mathfrak{R}(P_1, P_2, P_3, P_4) = \frac{P_1P_3 \cdot P_2P_4}{P_1P_4 \cdot P_2P_3}.$$

Here P_jP_k denotes the *directed length* from P_j to P_k; so $P_kP_j = -P_jP_k$.

The first point to be made about this is that the order is important: in general

$$\mathfrak{R}(P_1, P_2, P_3, P_4) \neq \mathfrak{R}(P_2, P_1, P_3, P_4).$$

In fact,

$$\mathfrak{R}(P_1, P_2, P_3, P_4) = \frac{1}{\mathfrak{R}(P_2, P_1, P_3, P_4)}.$$

7.2. Invariance of the cross-ratio. The importance of the cross-ratio is that it is unaffected by perspective transformations: let f be the perspective projection to the plane π and let $Q_1 = f(P_1)$, $Q_2 = f(P_2)$, $Q_3 = f(P_3)$, $Q_4 = f(P_4)$. Then

$$\mathfrak{R}(Q_1, Q_2, Q_3, Q_4) = \mathfrak{R}(P_1, P_2, P_3, P_4).$$

Proof We can return to the formulae we had when we were studying the perspective projection of ℓ. In terms of the parameterization (68), let P_j correspond to parameter value $t = t_j$. Then $P_jP_k = |\mathbf{U}|(t_k - t_j)$ (this is the signed distance) and so

$$\mathfrak{R}(P_1, P_2, P_3, P_4) = \frac{(t_3 - t_1)(t_4 - t_2)}{(t_4 - t_1)(t_3 - t_2)}. \tag{73}$$

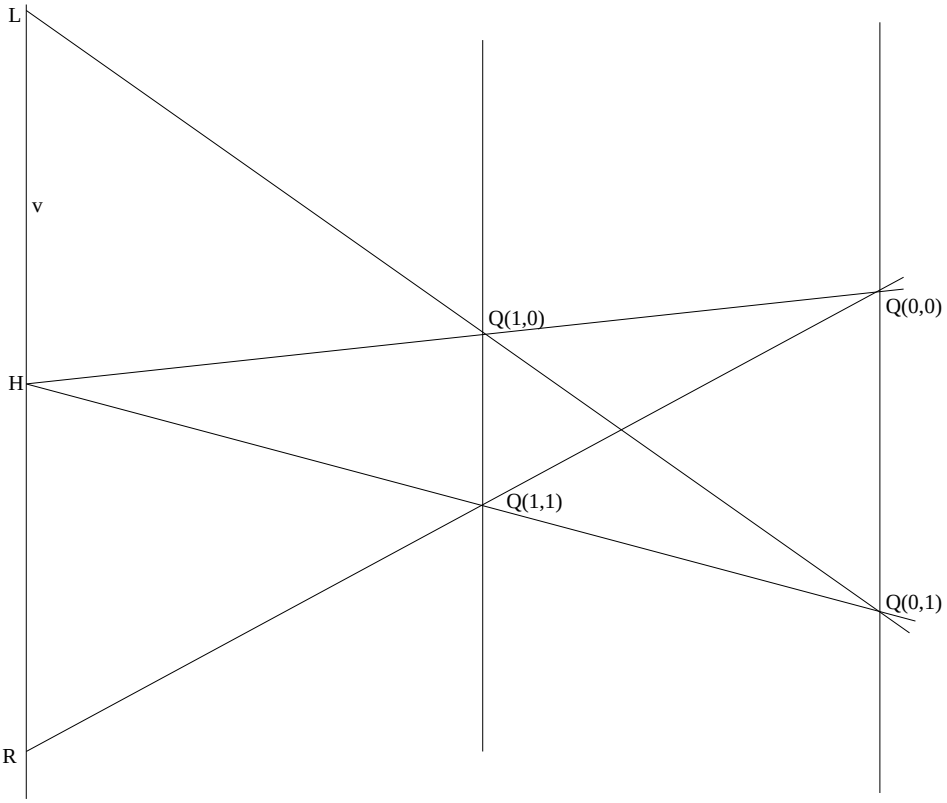

FIGURE 4. Perspective projection of a rectangular grid: step 2. The line of vanishing points v is located, and the vanishing points L and R for the two families of diagonals have been found.

But

$$Q_j = \left(\frac{A_1}{A_3} + \left(\frac{U_1}{U_3} - \frac{A_1}{A_3}\right) s_j, \frac{A_2}{A_3} + \left(\frac{U_2}{U_3} - \frac{A_2}{A_3}\right) s_j\right)$$

where

$$s_j = \frac{U_3 t_j}{A_3 + U_3 t_j}.$$

So $Q_j Q_k = |\mathbf{U}'|(s_k - s_j)$, where $\mathbf{U}' = \left(\frac{U_1}{U_3} - \frac{A_1}{A_3}, \frac{U_2}{U_3} - \frac{A_2}{A_3}\right)$ is the direction vector of $f(\ell)$. Hence

$$\mathfrak{R}(Q_1, Q_2, Q_3, Q_4) = \frac{(s_3 - s_1)(s_4 - s_2)}{(s_4 - s_1)(s_3 - s_2)}. \tag{74}$$

To complete the proof, observe that

$$s_j - s_k = \frac{U_3 t_j}{A_3 + U_3 t_j} - \frac{U_3 t_k}{A_3 + U_3 t_k} = \frac{A_3 U_3 (t_j - t_k)}{(A_3 + U_3 t_j)(A_3 + U_3 t_k)}.$$

When this is inserted in (74) we get the RHS of (73), which completes the proof that the cross-ratio is invariant.

7.3. Permutations. Although the cross-ratio depends on the order of the points, the cross-ratios of the different possible permutations are related to each other in quite a simple way:

$$\mathfrak{R}(P_1, P_2, P_3, P_4) = \frac{1}{\mathfrak{R}(P_2, P_1, P_3, P_4)} = \frac{1}{\mathfrak{R}(P_1, P_2, P_4, P_3)} \tag{75}$$

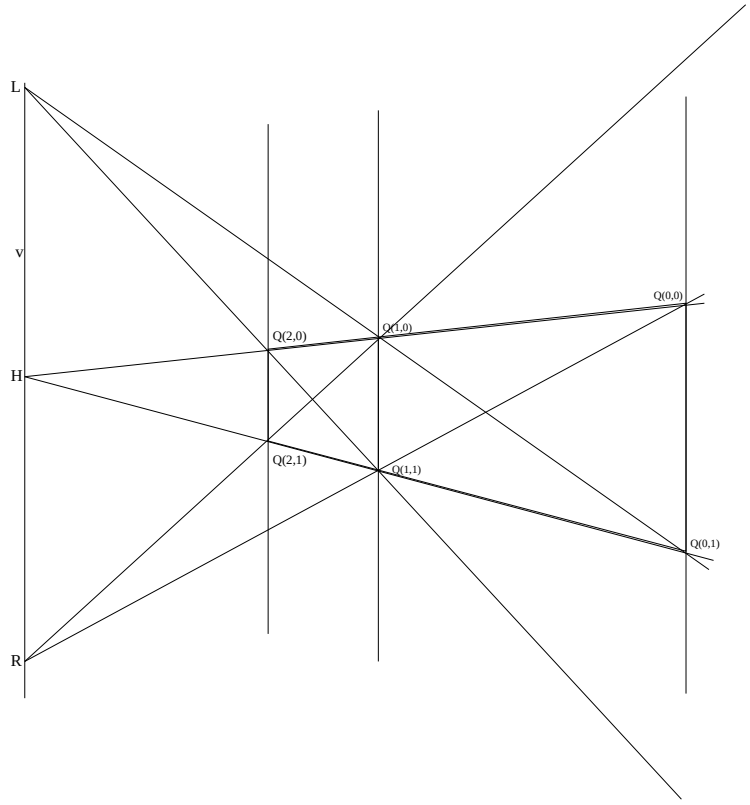

FIGURE 5. Perspective projection of a rectangular grid: step 3

and

$$\mathfrak{R}(P_1, P_2, P_3, P_4) + \mathfrak{R}(P_1, P_3, P_2, P_4) = 1. \tag{76}$$

From this it is possible to determine the cross-ratio of any permutation of the P_j in terms of the cross-ratio of (P_1, P_2, P_3, P_4).

8. More projective geometry: lines in $\mathbb{P}^2$

In general, a plane V through O will meet a view-plane π. In fact, this is the case unless the rays in V are all ideal for π.

The Cartesian equation $AX+BY+CZ = 0$ of V gives the *homogeneous equation* this line, simply by interpreting X, Y and Z as homogeneous coordinates $(X : Y : Z)$ in $\mathbb{P}^2$. Notice that this is OK because (X, Y, Z) satisfies the equation if and only if $(\lambda X, \lambda Y, \lambda Z)$ satisfies the equation. This is a non-trivial check: notice that the equation $AX + BY + CZ = D$ *cannot* be interpreted 'projectively' (i.e. thinking of X, Y, Z as homogeneous coordinates) if $D \neq 0$.

8.1. Intersection of lines in $\mathbb{P}^2$. Two lines in $\mathbb{P}^2$ correspond precisely to two planes V_1 and V_2, say, through O. These *always* meet in a one-dimensional subspace L, which represents a point of $\mathbb{P}^2$. Hence *any pair of lines in* $\mathbb{P}^2$ *meets in a point.* There are no parallel lines in $\mathbb{P}^2$.

More explicitly, if the two lines have the homogeneous equations

$$A_1X + B_1Y + C_1Z = 0 \ \ ; A_2X + B_2Y + C_2Z = 0$$

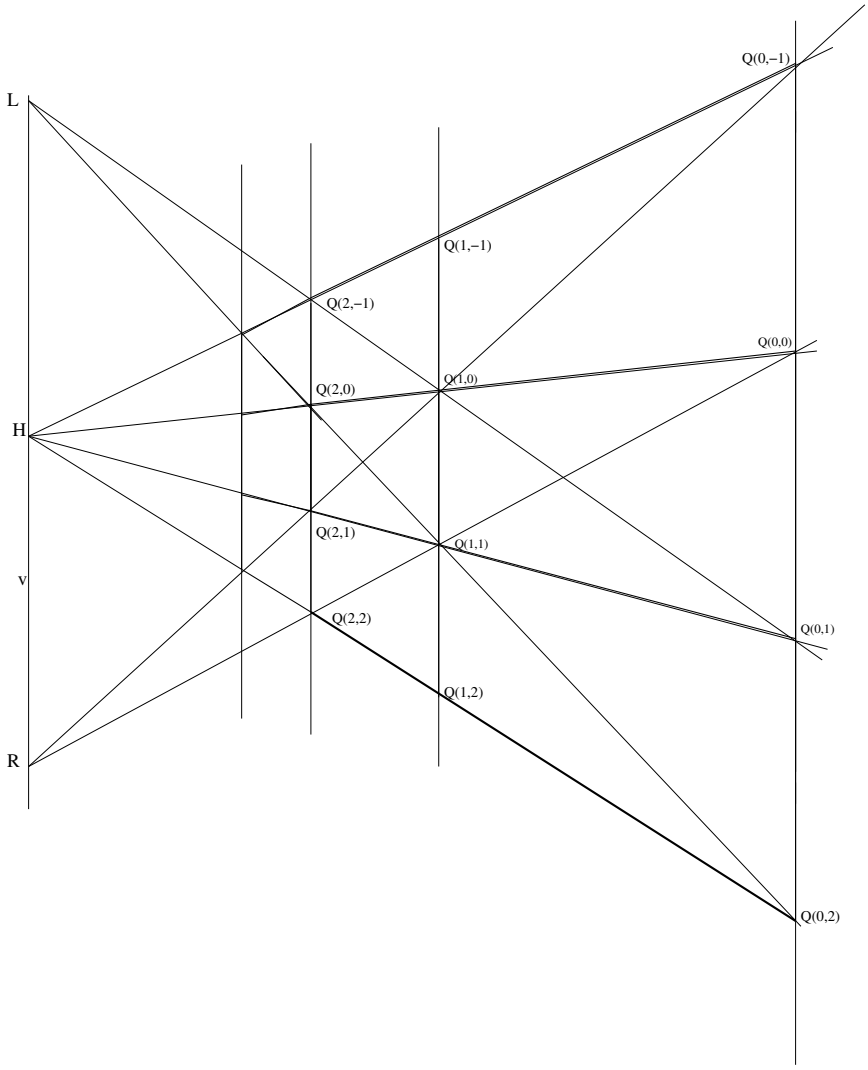

FIGURE 6. Perspective projection of a rectangular grid: step 4. The process now continues indefinitely

then their intersection point has homogeneous coordinates

$$(B_1C_2 - B_2C_1 : C_1A_2 - C_2A_1 : A_1B_2 - A_2B_1)$$

as is easily checked by substituting this into the equation. (The resemblance to the cross product is not accidental!)

8.2. The projective line through two given points. Suppose that $P_1 = (a_1 : b_1 : c_1)$ and $P_2 = (a_2 : b_2 : c_2)$ are two given points of $\mathbb{P}^2$. The line joining them can be written as the determinant equation

$$\begin{vmatrix} X & Y & Z \\ a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{vmatrix} = 0.$$

8.3. View-planes revisited. The discussion of visible and ideal points of a view-plane can now be streamlined. Given a view-plane π with equation $ax + by + cz = 1$, the visible points of $\mathbb{P}^2$ are those whose homogeneous coordinates satisfy $aX + bY + cZ \neq 0$, while the ideal points form the line with equation $aX + bY + cZ = 0$. This is the equation of the line at infinity of π. Note that, conversely, any given line in $\mathbb{P}^2$ is the line at infinity for a certain

view-plane π. In fact, if the equation of the given line is $aX + bY + cZ = 0$, then the view-plane $ax + by + cz = 1$ has this as its line at infinity.

9. Projective transformations

The group of projective transformations of $\mathbb{P}^n$ is the set of all maps $\mathbb{P}^n \to \mathbb{P}^n$ determined by *invertible linear transformations* of $\mathbb{R}^{n+1}$.

For example, the general projective transformation of $\mathbb{P}^1$ is map of the form

$$(X : Y) \mapsto (AX + BY : CX + DY), \text{ where } AD - BC \neq 0. \tag{77}$$

Note we can see the necessity for invertibility here. If $AD - BC = 0$, then $(B : -A) \mapsto (0 : CB - AD) = (0 : 0)$ which is not allowed.

Similarly, the general projective transformation of $\mathbb{P}^2$ is given by

$$(X : Y : Z) \mapsto (A_1X + B_1Y + C_1Z : A_2X + B_2Y + C_2Z : A_3X + B_3Y + C_3Z)$$

where

$$\begin{vmatrix} A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \\ A_3 & B_3 & C_3 \end{vmatrix} \neq 0.$$

Notice that if we restrict this transformation to the subset $\{Z \neq 0\}$, then in terms of $x = X/Z$, $y = Y/Z$, we get

$$(x, y) \mapsto \left(\frac{A_1x + B_1y + C_1}{A_3x + B_3y + C_3}, \frac{A_2x + B_2y + C_2}{A_3x + B_3y + C_3} \right).$$

Comparing with the formula we found when mapping by perspective projection a coordinate grid in a given plane σ to a view-plane π, we see that perspective projection is an example of a projective transformation!

9.1. Affine transformations. The point $(X : Y : 0)$ is mapped by this transformation to

$$(A_1X + B_1Y : A_2X + B_2Y : A_3X + B_3Y).$$

In particular, the line $Z = 0$ is mapped to itself if and only if $A_3 = B_3 = 0$. Thinking of this as the line at infinity in the usual way, we see that such a projective transformation also maps π to itself, provided we rescale so that $C_3 = 1$.

10. Cross-ratio revisited

Let P_1, P_2, P_3, P_4 be four distinct points on $\mathbb{P}^1$, so that P_j has homogeneous coordinates $(X_j : Y_j)$. The cross-ratio of (P_1, P_2, P_3, P_4) can be defined as

$$\mathfrak{R}(P_1, P_2, P_3, P_4) = \frac{(X_1Y_3 - X_3Y_1)(X_2Y_4 - X_4Y_2)}{(X_1Y_4 - X_4Y_1)(X_2Y_3 - X_3Y_2)}.$$

This needs some justification. First of all, if (X_j, Y_j) is replaced by $(\lambda_j X_j, \lambda_j Y_j)$, this ratio does not change, so it really does depend on the homogeneous coordinates $(X_j : Y_j)$. If we introduce the coordinate $t = Y/X$, then the expression reduces to

$$\mathfrak{R}(P_1, P_2, P_3, P_4) = \frac{(t_3 - t_1)(t_4 - t_2)}{(t_4 - t_1)(t_3 - t_2)}$$

in agreement with (73). On the other hand it is easier to show that the cross-ratio is invariant using this new definition. Since perspectivities are examples of projective transformations, it is sufficient to apply (77) to the $(X_j : Y_j)$. Now if we put

$$(X_j', Y_j') = (AX_j + BY_j, CX_j + DY_j)$$

then it is straightforward to check that

$$X_j'Y_k' - Y_j'X_k' = (AD - BC)(X_jY_k - Y_jX_k).$$

From this the invariance of the cross-ratio follows at once.

10.1. Remark. Thinking of P_1, P_2, P_3 as fixed and $P_4 = (X : Y)$ as variable, we can view the cross-ratio as a projective transformation

$$T : (X : Y) \mapsto ((X_1Y_3 - X_3Y_1)(X_2Y - XY_2) : (X_1Y - XY_1)(X_2Y_3 - X_3Y_2)).$$

this has the nice property that

$$T(X_1 : Y_1) = (1 : 0), T(X_2 : Y_2) = (0 : 1), T(X_3, Y_3) = (1 : 1). \tag{78}$$

In other words, T maps any given three points on $\mathbb{P}^1$ to the three simplest ones, $(1 : 0)$, $(0 : 1)$ and $(1 : 1)$. Conversely, if T is the projecive transformation of $\mathbb{P}^1$ with the property (78), then if $T(X_4 : Y_4) = (u : v)$,

$$\mathfrak{R}(P_1, P_2, P_3, P_4) = u/v.$$

11. Harmonic division

Consider the figure shown in the upper part of Figure 7. Given A, B and D, we construct C by picking P arbitrarily, drawing the line out of D arbitrarily, and then marking in the diagonals AZ and BX of the quadrilateral $ABZX$. Finally C is constructed by joining P to the intersection point Q of the diagonals. We can determine $\mathfrak{R}(D, B, C, A)$ by mapping the line DP to infinity and simultaneously mapping B to $B' = 0$, and C to $C' = 1$. The lines through P become the sloping parallel lines while the lines through A become the horizontal parallel lines shown in the lower part of the figure. It is clear that in this second picture, C' is the mid-point of $A'B'$, so $A' = 2$. But by the above description of cross-ratio, it follows that

$$\mathfrak{R}(D, B, C, A) = \mathfrak{R}(D', B', C', A') = T(A') = 2.$$

It follows that $\mathfrak{R}(A, B, C, D) = -1$ and one says that (A, B, C, D) are in harmonic division. This construction is closely related to the perspective mapping of the coordinate grid considered in §6 where we had to find how equispaced points on a line in one plane are mapped by a perspective transformation.

12. Pappus and Desargues

These are two fundamental and very old theorems of projective geometry. In axiomatic approaches to the subject they are sometimes taken as axioms. For us, however, they are theorems.

12.1. Pappus's Theorem. Let ℓ and ℓ' be any two lines in $\mathbb{P}^2$ and let points A, B, C on ℓ and A', B', C' on ℓ' be given, all distinct, and distinct from the point of intersection of ℓ and ℓ'. Then the three points

$$P = BC' \cap B'C, \quad Q = AC' \cap A'C, \quad R = AB' \cap A'B$$

are collinear. (See Fig. 8)

Remark: here we have written for example $AB' \cap A'B$ to mean 'the point of intersection of the line through A to B' and the line through A' and B'. AB' does *not* mean the line segment between A and B' here.

We shall prove Pappus's Theorem by reducing it to a theorem in ordinary linear geometry by sending points to infinity! Indeed, let ℓ_∞ be the line joining P and Q. If we decompose $\mathbb{P}^2 = \mathbb{R}^2 \cup \ell_\infty$ (a division into visible points and points at infinity), then P and Q will by construction go to infinity, so that BC' is parallel to $B'C$ and AC' is parallel to $A'C$. What we have to show is that R is on ℓ_∞, in other words, in $\mathbb{R}^2$, AB' and $A'B$ are parallel.

To do this, distinguish two cases. First, suppose that ℓ and ℓ' intersect in a visible point O of $\mathbb{R}^2$. There is a $\lambda \in \mathbb{R}$ so that $\mathbf{OC} = \lambda\mathbf{OA}$, and a $\mu \in \mathbb{R}$ so that $\mathbf{OB} = \mu\mathbf{OC}$. But by examining the pairs of similar triangles $\triangle OAC'$ and $\triangle OCA'$, and also $\triangle OBC'$ and $\triangle OCB'$, we get that $\mathbf{OA}' = \lambda\mathbf{OC}'$, $\mathbf{OC}' = \mu\mathbf{OB}'$, and that $\mathbf{CA}' = \lambda\mathbf{AC}'$. All this gives that

$$\mathbf{BA}' = \mathbf{OA}' - \mathbf{OB} = \mathbf{OC} + \mathbf{CA}' - \mu\mathbf{OC} = (1-\mu)\lambda\mathbf{OA} + \lambda\mathbf{AC}' = (1-\mu)\lambda\mathbf{OA} + \lambda\mathbf{OC}' - \lambda\mathbf{OA} =$$
$$= \lambda\mathbf{OC}' - \mu\lambda\mathbf{OA} = \mu\lambda\mathbf{OB}' - \mu\lambda\mathbf{OA} = \mu\lambda\mathbf{AB}'$$

which proves that $\mathbf{BA}' || \mathbf{AB}'$.

If ℓ and ℓ' are parallel, the argument is the same, but instead of dilations S and T, we use translations. QED.

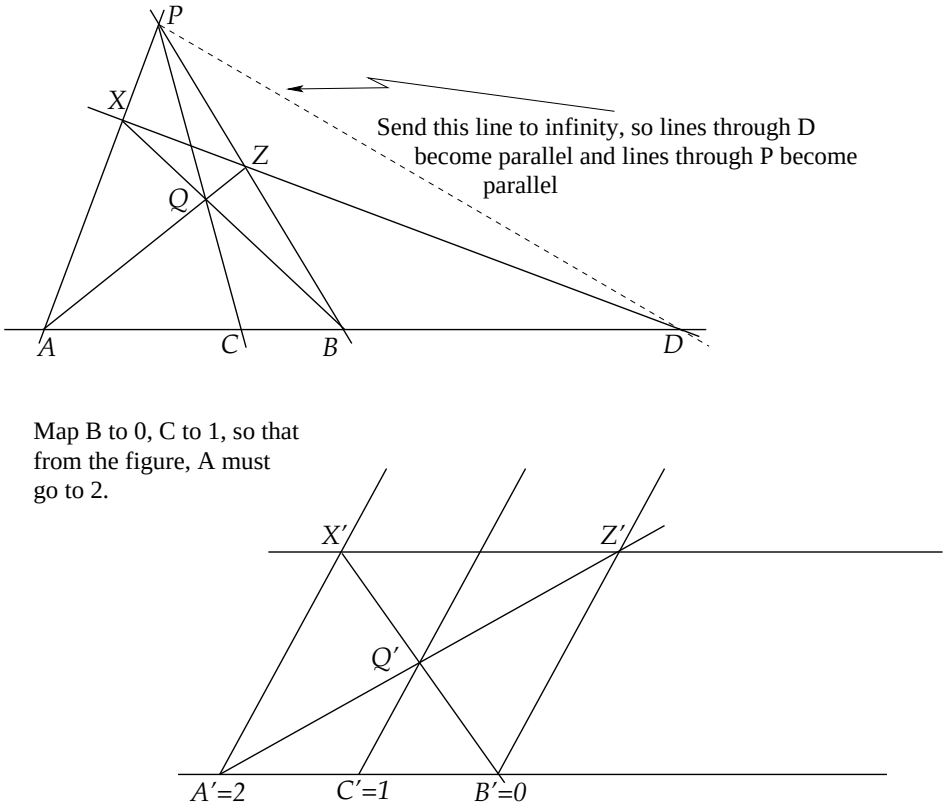

FIGURE 7. Cross-ratio of four points in harmonic division. It follows by mapping A, B, C and P as shown, that D maps to 2.

12.2. Desargues's Theorem. Given six points in $\mathbb{P}^2$, A, B, C, A', B', C' such that the triangles ABC and $A'B'C'$ are in perspective from some point U. (In other words, U, A, A' are collinear, U, B, B' are collinear and U, C, C' are collinear.) Then the points

$$P = BC \cap B'C', \;\; Q = CA \cap C'A', \;\; R = AB \cap A'B'$$

are collinear (Fig. 9).

There are many possible proofs of this. One, in the same style as the above proof of Pappus, is to send P and Q off to infinity, and then show that AB must be parallel to $A'B'$.

An alternative is to obtain the result from an analogous result in 3 dimensions. To do so, consider the Fig. 9 to be lying in a view-plane π in $\mathbb{R}^3$. Let W be the 2D subspace through O and containing UCC'. Replace this line by a line ℓ' not in π, but lying in W, and passing through U. Pick c and c' on ℓ' so that they map by perspective projection from O to C and C'.

Now consider the triangles ABc and $A'B'c'$. These determine two planes σ and σ'. These planes meet, because AB lies in σ, $A'B'$ lies in σ' and $AB \cap A'B' = R$ is a point of π. The intersection of two plane in $\mathbb{R}^3$ is a line (if it's non empty) and we see that $\sigma \cap \sigma'$ also contains $Bc \cap B'c'$ and $Ac \cap A'c'$. These three points are therefore collinear. But these project (by the same perspectivity that maps Ucc' to UCC') to P and Q, respectively, so that P, Q and R are collinear.

12.3. Applications. Using Pappus's Theorem, one can construct a straight line joining two points, using only a ruler that isn't quite long enough to connect the points.

Using Desargues's Theorem, one can construct the line joining a given point P to $\ell \cap \ell'$, even if the latter is off one's piece of paper.

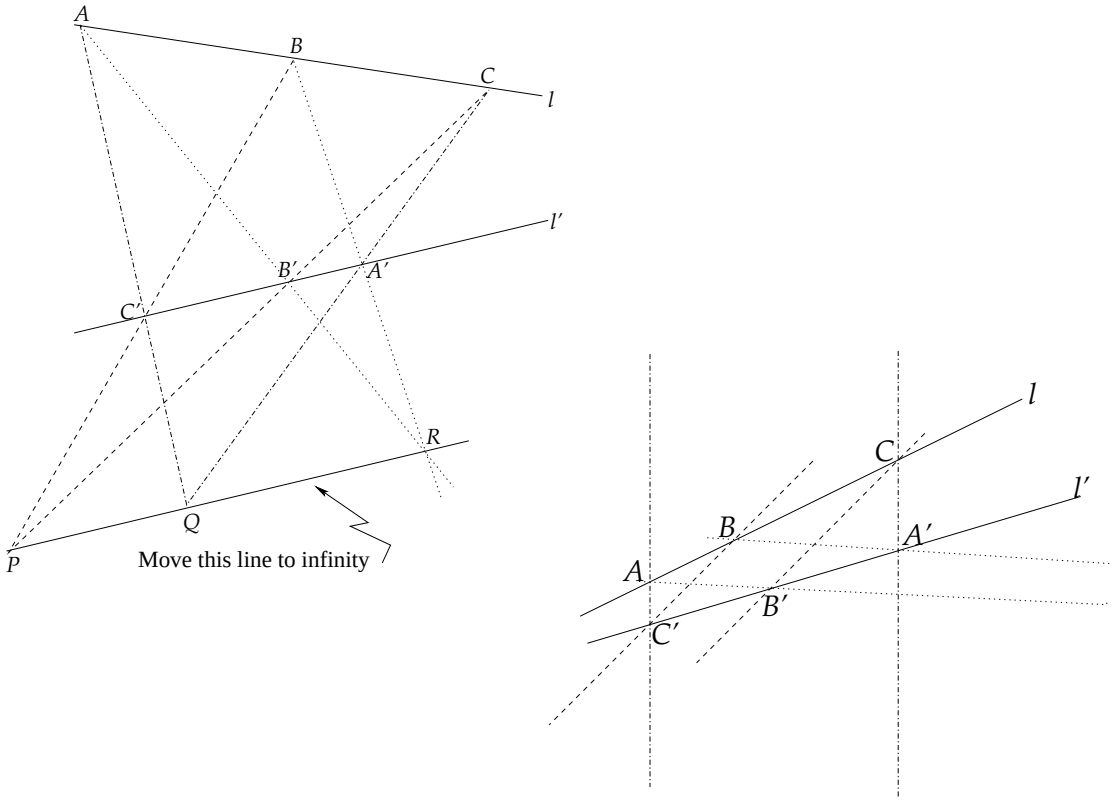

FIGURE 8. Pappus's Theorem

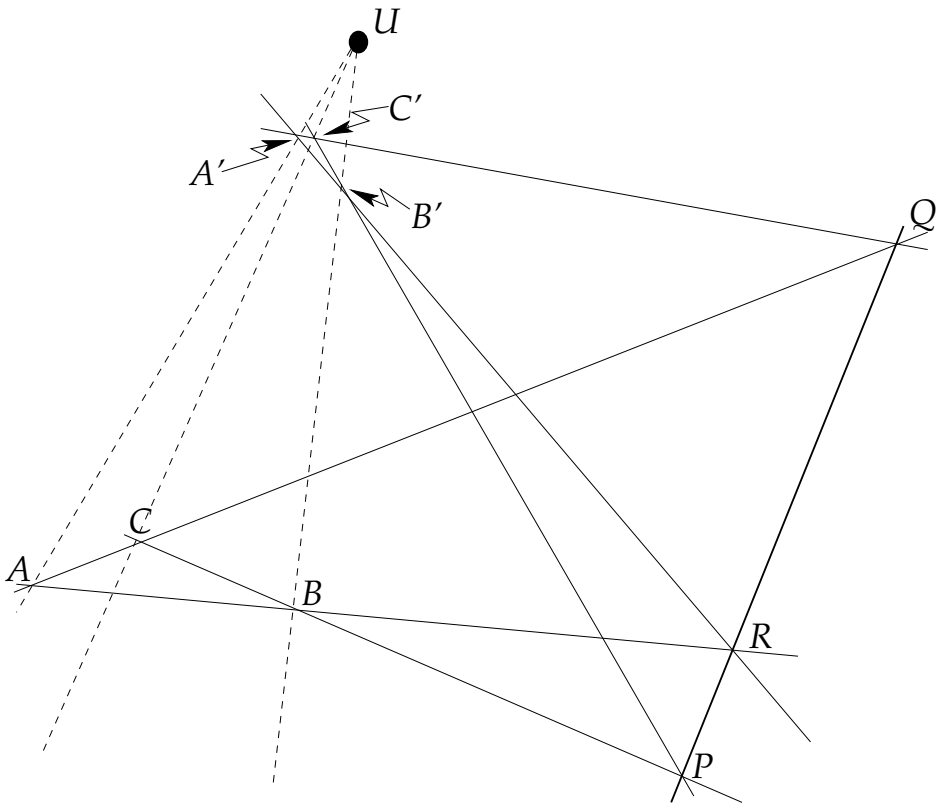

FIGURE 9. Desargues

CHAPTER 7

More special curves

The term "Special Curves" can be very misleading, as there is really nothing special about these curves, except the fact that they happened to be useful to some people for some specific purposes at some point in time. Therefore, this term refers collectively to a number of curves which have wildly different properties, but which share the feature that they have been studied and documented extensively.

We have already met four of them, i.e. the conic sections: the circle, the ellipse, the hyperbola, and the parabola. In what follows we will focus our attention on some other famous curves, some of which will give us the opportunity to appreciate the usefulness of other coordinate systems. A nice Web reference for all the curves we will mention, and have mentioned, and about mathematics in general is [**1**].

1. The cycloid

The cycloid is the curve describing the motion of a point of the wheel of a car, as the car moves with constant speed on a straight line without slipping. Let us find the equation of the curve in parametric form, assuming that the radius of the wheel is r, its angular velocity is ω, and that at time $t = 0$ the point of interest is the point of contact of the wheel and the ground (see Fig. 1).

The key to finding the equation is to think that, as the wheel does not slip, a turn by angle θ implies also a horizontal translation of the wheel by $r\theta$. Let us take a coordinate system of the plane so that the ground coincides with the x-axis and the point of interest at $t = 0$ lies at the origin, so that the center of the wheel lies at $(0, r)$. In our case, the center of the wheel at time $t > 0$ will be at $(r\omega t, r)$, while the point of interest will have traced an angle of $-\omega t$, its initial angle being $-\pi/2$. Adding the two vectors, we find that:

$$(x(t), y(t)) = (r\omega t, r) + r\left(\cos\left(-\frac{\pi}{2} - \omega t\right), \sin\left(-\frac{\pi}{2} - \omega t\right)\right)$$

which gives

$$\begin{aligned} x(t) &= r\omega t - r\sin(\omega t) \\ y(t) &= r - r\cos(\omega t) \end{aligned} \tag{79}$$

Observe that it is possible to find a closed expression for the curve (of the form $F(x, y) = 0$), but it will not be pretty at all! Here, the parametric form serves as much better!

The cycloid is also famous from some applications in Physics: it is the *Brachistochrone* and the *Tautochrone*, the solution of two very old and important problems:

- Consider any two points on the plane $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$, so that $x_1 \neq x_2$ and $y_1 > y_2$. Let a mass at P_1 move down to P_2 by means of gravity alone without friction. What is the curve along which the mass will move from one point to the other in the least time possible? This curve is called the *Brachistochrone*(=shortest-time curve), and we can prove that it is the cycloid through the two points.
- Consider a reflected cycloid with respect to the x-axis, and put a mass at any of its points; allow it to slide down to the bottom without friction. The time this will take is independent of the initial point; this is why it is called the *Tautochrone*(=same-time curve).

2. Spirals

Spirals are curves most easily expressed in polar coordinates, where they have the general form $r = f(\theta)$, where f is a strictly monotone function. There exist several types of spirals, each with a distinct name. Let us see some:

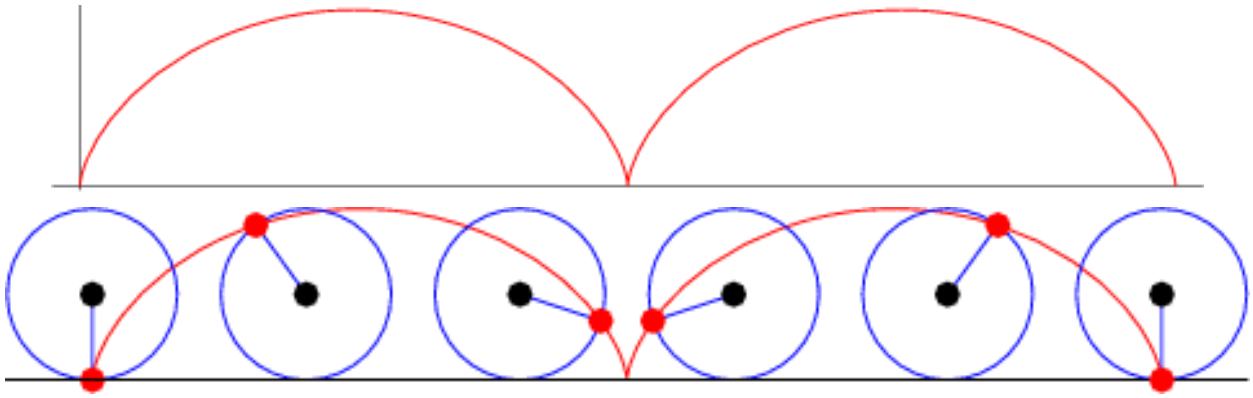

FIGURE 1. The cycloid.

- *Logarithmic*: $r = ae^{b\theta}$
- *Archimedean*: This is an entire family of spirals having the form $r = a\theta^{\frac{1}{n}}$, where $n \in \mathbb{N}$. *Lituus, Hyperbolic, Archimedes's* and *Fermat's* spirals correspond to $n = -2, -1, 1, 2$, respectively.

3. The trisectrix (or quadratrix) of Hippias

One of the major problems in geometry in antiquity was the trisection of a random angle. This was proved to be impossible by means of ruler and compass only, but possible if some especially constructed curves were available. The Quadratrix of Hippias is one of them.

Consider a line parallel to the x-axis, initially at $y = 0$ and moving at a constant speed towards the line $y = \dfrac{\pi}{2}$. Consider also a line through the origin, initially coinciding with the x-axis, rotating around the origin towards the y-axis at a constant rate. The rotation and the translation rates are such so that when the rotating line coincides with the y-axis, the translating line coincides with $y = \frac{\pi}{2}$. The intersection of the two lines traces the Quadratrix of Hippias (see Fig.2).

The two lines, with $t \in \left[0, \dfrac{\pi}{2}\right]$ as a parameter, have the equations $y = t$ and $y = \tan(t)x$, therefore the Cartesian equation is (do it as an exercise)

$$x = y\cot(y) \Leftrightarrow r = \frac{\theta}{\sin(\theta)}$$

As an application, let us see how we can trisect a random angle using this curve. Observe that, without loss of generality, we can assume that the line is acute (less than 90^o), as the angles of 90^o, 180^o, and 270^o can be easily trisected (the angles of 30^o, 60^o, and 90^o can be constructed in a straightforward way). Let us refer now to Fig. 3. In order to trisect the angle $\widehat{AOB}$, we must find $|\mathbf{OE}|$, and take D so that $|\mathbf{OD}| = \dfrac{1}{3}|\mathbf{OE}|$ (trisection of a segment can be done with ruler and compass). Let the intersection of the curve and the horizontal line at D be P, then $\widehat{AOP} = \dfrac{1}{3}\widehat{AOB}$. This follows because, by the definition of the curve, $\dfrac{|\mathbf{OE}|}{|\mathbf{OC}|} = \dfrac{\widehat{AOB}}{\frac{\pi}{2}}$. As $|\mathbf{OC}| = \frac{\pi}{2}$, it follows that $|\mathbf{OE}| = \widehat{AOB}$.

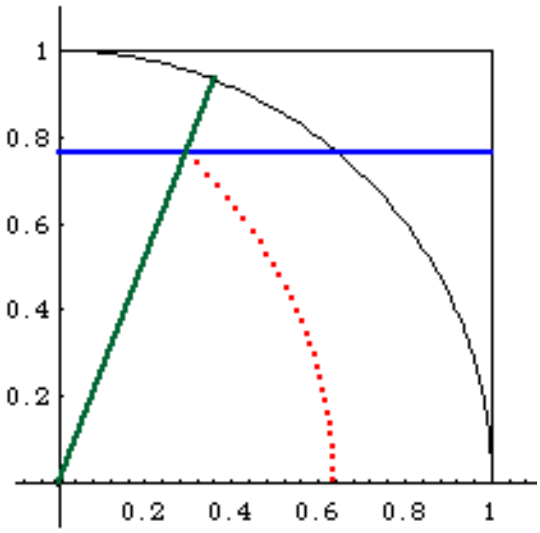

FIGURE 2. Diagram (obtained from [**2**]) explaining the derivation of the Quadratrix of Hippias (both axes have been rescaled by $\frac{\pi}{2}$).

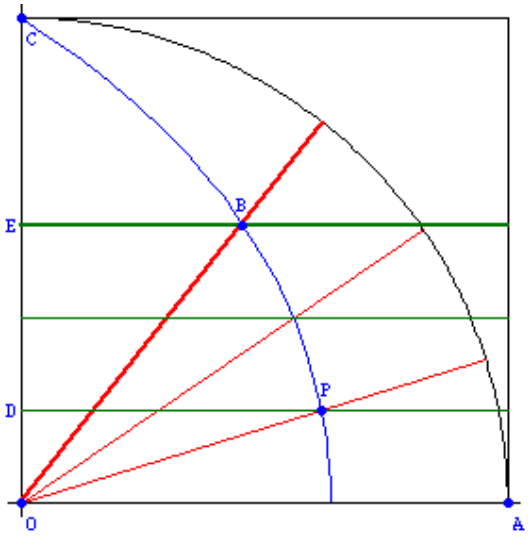

FIGURE 3. Trisection of an acute angle using the Trisectrix of Hippias (obtained from [**2**]).

CHAPTER 8

Fractals

1. Introduction

There is no universally agreed upon definition of a fractal; fortunately, there is a set of properties agreed upon that a curve or a surface must have in order to be called a fractal:

- *Self-similarity:* Any part of a fractal "looks like" the entire fractal. The degree of similarity can be quantified in various ways.
- *Non-integral dimension:* The "dimension" (see below) of a fractal is not an integer. In simpler terms, a fractal curve is very "dense" or very "sparse", and can be sometimes considered something more (a surface) or something less (a discrete set of points) than a curve.

These properties are often the result of a *recursive* construction method. Indeed, the construction of most of the "famous" fractals follows the steps below:

- We start by a simple curve (or surface).
- We substitute a part of it by another curve (or surface), or we append to it another curve (or surface).
- After the substitution we are able to identify one or more smaller parts similar to the original.
- We repeat the previous two steps for every part similar to the original we identified.

The curve (surface) thus generated is certainly self-similar with a recursive structure.

2. The plus-sign example [5]

Let us see an example (see Fig. 1). Suppose we want to construct a fractal using as a basic initial curve two linear segments of equal length (say unit length) intersecting perpendicularly and at their centers, i.e. a plus-sign. At each step we will be placing copies of this plus-sign on the curve generated so far, so that

- the new plus-sign is scaled down with respect to the plus-sign of the previous step by 0.5,
- their linear segments are oriented the same way,
- the center of the new plus-sign is an endpoint of the curve generated in the previous step.

If we repeat this process for an infinite number of times, we will produce a fractal. Let us do some numerics on that:

- The curve is obviously symmetric. What is the side length of the smallest square containing the entire curve? As every plus-sign we add is half the size of the previous one, the side length we ask for is
$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots = 2$$
- How many endpoints does the curve have at the nth step? At the initial (0th) step we only have the plus sign, which has obviously 4 endpoints. At the $(n+1)$th step, we will have placed a plus-sign on every one of the $E(n)$ endpoints of the curve of the nth step; each plus-sign has four endpoints, exactly one of which lies already in the curve and thus is not an endpoint for the curve; so, only 3 remain, and we reach the formula $E(n+1) = 3E(n)$. Along with the initial condition, this gives:

$$E(n) = 4 \cdot 3^n, \ n \geq 0 \tag{80}$$

The recursive structure and the self-similarity are obvious in this example. What about the dimension? We proceed to say a few words about it.

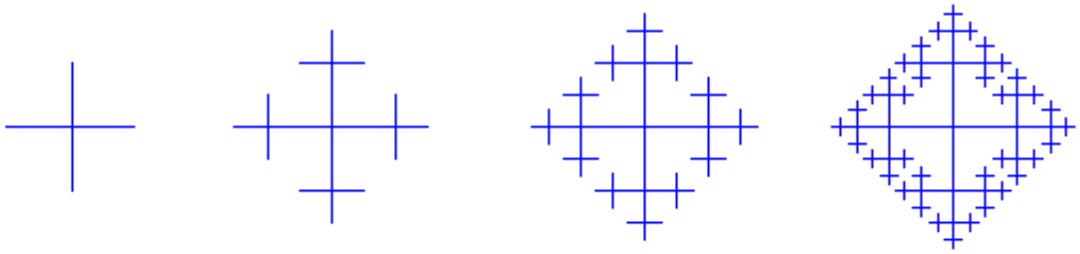

FIGURE 1. A fractal generated by a plus-sign (as found in [**5**]).

3. Hausdorff dimension

Hausdorff dimension measures in how many dimensions a shape lives: a curve lives typically in one dimension, a surface in two, a volume in three etc. But, as we mentioned before, there are cases of shapes which have a non-integer dimension. Hausdorff dimension is in practice very difficult to calculate, therefore we will follow the scientific community in using an approximation known as *box dimension*, which may underestimate the Hausdorff dimension, but in simple cases yields accurate results.

Imagine then you have a linear segment of length 1 on the plane, and you cover it completely by as few balls as possible (in this case adjacent non-overlapping balls of diameter ϵ whose centers lie on the segment). How many do we need? The answer is $\left\lceil \frac{1}{\epsilon} \right\rceil$. Disposing of the ceiling function, we can write the number of balls N we need to cover the segment as a function of their diameter D as:

$$N(\epsilon) = \frac{1}{\epsilon} = D(\epsilon)^{-1}$$

The (negative of the) exponent here reveals the dimension of the shape (the segment). It is 1, therefore the segment is a line as we know it.

What about a square of side 1? Let us try to cover it with little squares of side ϵ this time. We are going to need $\left\lceil \frac{1}{\epsilon^2} \right\rceil$ this time, so the number of squares as a function of the diameter becomes:

$$N(\epsilon) = \frac{1}{\epsilon^2} = D(\epsilon)^{-2}$$

and the exponent is now correctly 2, implying that the square is a surface.

What about the orthogonal parallelogram with side lengths a and b?

$$N(\epsilon) = \frac{ab}{\epsilon^2} = abD(\epsilon)^{-2}$$

and the exponent is still 2.

It is customary to use a sequence of discrete values for ϵ, and more precisely $\epsilon(n) = b \cdot a^{-n},\ n \in \mathbb{N}$, where $a > 1$ is a number chosen according to the problem at hand. Then, typically, the dimension equation can be expressed as:

$$N(n) = C(n)D(n)^{-d},\ C(n) \to C > 0,\ |\log(D(n))| \to \infty$$

so that

$$d = -\lim \frac{\log(N(n))}{\log(D(n))} \tag{81}$$

This is the definition of the box dimension. The shape of the box is immaterial: it can be spherical, orthogonal, etc. All that matters is for it to have a characteristic dimension (diameter, side length, etc.) which we can vary by keeping the shape the same. Then, the results we will obtain by using different box shapes will be identical.

These being said, let us find the box dimension of our plus-sign fractal. Let us cover the curve produced at the nth step by (the least possible number of) discs of diameter 2^{-n-1}. Then, the initial plus-sign requires only four discs; assume that the nth step requires $N(n)$ discs. How many will we need for the $(n+1)$th step? In the $(n+1)$th step we will use disks of half the size of what we used before, so we need $2N(n)$ to cover the curve of the previous step. The new plus-signs we add have segments of length 2^{-n-1}, thus each one requires three additional balls, and we have $E(n+1)$ of them (see (80)). Therefore:

$$N(n+1) = 2N(n) + 3E(n) = 2N(n) + 4 \cdot 3^{n+1},\ N(0) = 4$$

whose solution is:

$$N(n) = 4 \cdot 3^{n+1} - 2^{n+3},\ n \geq 0$$

and introducing $D(n) = 2^{-n-1}$:

$$N(n) = 4\left[1 - D(n)^{\log_2(3)-1}\right] D(n)^{-\log_2(3)}$$

we get that:

$$d = -\lim \frac{\log(N(n))}{\log(D(n))} = \log_2(3) \approx 1.585$$

We see then that this curve is somewhere in between a traditional curve and a surface: its dimension is larger than 1 but less than 2.

4. Plotting fractals

Plotting fractals, which have a recursive structure, is a process also done recursively. As we saw in our plus-sign example earlier, we start with a basic curve, which we subsequently refine more and more in an infinity of steps.

Suppose we want to use a programming language to draw the fractal, and that this programming language can draw a linear segment given its endpoints. We will then have to use a coordinate system and find the endpoints' coordinates. Let us start by choosing as the origin the center of the initial plus-sign, and take one segment of the plus-sign lying on $y = 0$, so that the other lies on $x = 0$. The endpoints of the segments will be:

$$\left(0, \frac{1}{2}\right), \left(0, -\frac{1}{2}\right) \text{ and } \left(\frac{1}{2}, 0\right), \left(-\frac{1}{2}, 0\right)$$

respectively.

Let us agree on a numbering scheme for the endpoints: number the above four points, in the order they appear from left to right, as $1, -1, 0, 2$. The number then denotes what angle (in multiples of 90^o) each arm of the plus-sign forms with the x-axis. How are we going to number the endpoints of the nth step? We are going to use strings of these 4 symbols, but now interpreting the numbers as angles formed not with respect to the x-axis, but with respect to the direction indicated by the previous characters of the string (and with the x-axis if no previous characters exist).

Observe that all strings having 2 as their second or later character do not correspond to endpoints, because they imply turning 180^o with respect to the direction we were already moving towards, i.e. turning back into an already existing segment. The first then character can have 4 values, the rest of them 3, so that all possible strings at the nth step (of length $n+1$) are $4 \cdot 3^n$ in number, i.e. exactly as many as the endpoints.

How can we find the coordinates of an endpoint by using the representation above? It is best to use complex numbers here. If the string representing the endpoint is $x_0x_1 \ldots x_n$, then the complex number representing its location is:

$$\sum_{m=0}^{n} \left(\frac{1}{2}\right)^{m+1} i^{\sum_{k=0}^{m} x_k} \tag{82}$$

Finally, given of the endpoints of a linear curve, how can we find the other? Suppose $x_0x_1 \ldots x_n$ is the endpoint given. We start at the end of the string (the rightmost character) moving towards the beginning, skipping all 0s. If we encounter an 1 or -1, we negate it and stop. If we reach the beginning of the string, and it is 0, we change it into 2.

The above process gives us all the information we need: we know where the curve endpoints lie and how to pair them into linear segments, so that we can draw these segments, and thus draw the nth step approximation to the fractal.

Bibliography

[1] *Mathworld* http://mathworld.wolfram.com

[2] Xah Lee. *A visual dictionary of famous plane curves* http://www.xahlee.org/SpecialPlaneCurves_dir/specialPlaneCurves.html

[3] J.D. Lawrence. *A catalog of special plane curves* Dover, 1972

[4] (Previous class notes prepared by Prof. T. A. Gillespie and Dr. M. A. Singer)

[5] *Fractals — An introduction* http://ejad.best.vwh.net/java/fractals/intro.shtml

Index