1 Euclidean geometry

1.1 Euclidean space

Our story begins with a geometry which will be familiar to all readers, namely the geometry of Euclidean space. In this first chapter we study the Euclidean distance function, the symmetries of Euclidean space and the properties of curves in Euclidean space. We also generalize some of these ideas to the more general context of metric spaces, and we sketch the basic theory of metric spaces, which will be needed throughout the book.

We consider Euclidean space $\mathbf{R}^n$, equipped with the standard Euclidean inner-product $(\ ,\ )$, which we also refer to as the dot product; namely, given vectors $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$ with coordinates x_i, y_i respectively, the inner-product is defined by

$$(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i y_i.$$

We then have a *Euclidean norm* on $\mathbf{R}^n$ defined by $\|\mathbf{x}\| = (\mathbf{x}, \mathbf{x})^{1/2}$, and a distance function d defined by

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|.$$

In some books, the Euclidean space will be denoted $\mathbf{E}^n$ to distinguish it from the vector space $\mathbf{R}^n$, but we shall not make this notational distinction.

The Euclidean distance function d is an example of a *metric*, in that for any points P, Q, R of the space, the following three conditions are satisfied:

(i) $d(P, Q) \geq 0$, with equality if and only if $P = Q$.
(ii) $d(P, Q) = d(Q, P)$.
(iii) $d(P, Q) + d(Q, R) \geq d(P, R)$.

The crucial condition here is the third one, which is known as the *triangle inequality*. In the Euclidean case, it says that, for a Euclidean triangle (possibly degenerate) with vertices P, Q and R, the sum of the lengths of two sides of the triangle is at least the length of the third side. In other words, if one travels (along straight line segments)

from P to R via Q, the length of one's journey is at least that of the direct route from P to R.

To prove the triangle equality in the Euclidean case, we use the Cauchy–Schwarz inequality, namely

$$\left(\sum_{i=1}^{n} x_i y_i\right)^2 \leq \left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right),$$

or, in the inner-product notation, that $(\mathbf{x}, \mathbf{y})^2 \leq \|\mathbf{x}\|^2 \|\mathbf{y}\|^2$. The Cauchy–Schwarz inequality also includes the criterion for equality to hold, namely that the vectors $\mathbf{x}$ and $\mathbf{y}$ should be proportional. We may prove the Cauchy–Schwarz inequality directly from the fact that, for any $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$, the quadratic polynomial in the real variable λ,

$$(\lambda\mathbf{x} + \mathbf{y}, \lambda\mathbf{x} + \mathbf{y}) = \|\mathbf{x}\|^2\lambda^2 + 2(\mathbf{x}, \mathbf{y})\lambda + \|\mathbf{y}\|^2,$$

is positive semi-definite. Furthermore, equality holds in Cauchy–Schwarz if and only if the above quadratic polynomial is indefinite; assuming $\mathbf{x} \neq \mathbf{0}$, this just says that for some $\lambda \in \mathbf{R}$, we have $(\lambda\mathbf{x} + \mathbf{y}, \lambda\mathbf{x} + \mathbf{y}) = 0$, or equivalently that $\lambda\mathbf{x} + \mathbf{y} = \mathbf{0}$.

To see that the triangle inequality follows from the Cauchy–Schwarz inequality, we may take P to be the origin in $\mathbf{R}^n$, the point Q to have position vector $\mathbf{x}$ with respect to the origin, and R to have position vector $\mathbf{y}$ with respect to Q, and hence position vector $\mathbf{x} + \mathbf{y}$ with respect to the origin. The triangle inequality therefore states that

$$(\mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y})^{1/2} \leq \|\mathbf{x}\| + \|\mathbf{y}\|;$$

on squaring and expanding, this is seen to be equivalent to the Cauchy–Schwarz inequality.

In the Euclidean case, we have a characterization for equality to hold; if it does, then we must have equality holding in the Cauchy–Schwarz inequality, and hence that $\mathbf{y} = \lambda\mathbf{x}$ for some $\lambda \in \mathbf{R}$ (assuming $\mathbf{x} \neq \mathbf{0}$). Equality then holds in the triangle inequality if and only if $|\lambda + 1| \|\mathbf{x}\| = (|\lambda| + 1) \|\mathbf{x}\|$, or equivalently that $\lambda \geq 0$. In summary therefore, we have equality in the triangle inequality if and only if Q is on the straight line segment PR, in which case the direct route from P to R automatically passes through Q. Most of the metrics we encounter in this course will have an analogous such characterization of equality.

Definition 1.1 A *metric space* is a set X equipped with a *metric* d, namely a function $d : X \times X \to \mathbf{R}$ satisfying the above three conditions.

The basic theory of metric spaces is covered well in a number of elementary textbooks, such as [13], and will be known to many readers. We have seen above that Euclidean

space of dimension n forms a metric space; for an arbitrary metric space (X, d), we can generalize familiar concepts from Euclidean space, such as:

- $B(P, \delta) := \{Q \in X : d(Q, P) < \delta\}$, *the open ball of radius* δ *around a point* P.
- *open sets* U *in* X: by definition, for each $P \in U$, there exists $\delta > 0$ with $B(P, \delta) \subset U$.
- *closed sets in* X: that is, subsets whose complement in X is open.
- *open neighbourhoods of* $P \in X$: by definition, open sets containing P.

Given two metric spaces (X, d_X), (Y, d_Y), and a function $f : X \to Y$, the usual definition of continuity also holds. We say that f is *continuous* at $P \in X$ if, for any $\varepsilon > 0$, there exists $\delta > 0$ such that $d_X(Q, P) < \delta$ implies that $d_Y(f(Q), f(P)) < \varepsilon$. This last statement may be reinterpreted as saying that the inverse image of $B(f(P), \varepsilon)$ under f contains $B(P, \delta)$.

Lemma 1.2 *A map $f : X \to Y$ of metric spaces is continuous if and only if, under f, the inverse image of every open subset of Y is open in X.*

Proof If f is continuous, and U is an open subset of Y, we consider an arbitrary point $P \in f^{-1}U$. Since $f(P) \in U$, there exists $\varepsilon > 0$ such that $B(f(P), \varepsilon) \subset U$. By continuity, there exists an open ball $B(P, \delta)$ contained in $f^{-1}(B(f(P), \varepsilon)) \subset f^{-1}U$. Since this holds for all $P \in f^{-1}U$, it follows that $f^{-1}U$ is open.

Conversely, suppose now that this condition holds for all open sets U of Y. Given any $P \in X$ and $\varepsilon > 0$, we have that $f^{-1}(B(f(P), \varepsilon))$ is an open neighbourhood of P, and hence there exists $\delta > 0$ with $B(P, \delta) \subset f^{-1}(B(f(P), \varepsilon))$. □

Thus, continuity of f may be phrased purely in terms of the open subsets of X and Y. We say therefore that continuity is defined *topologically*.

Given metric spaces (X, d_X) and (Y, d_Y), a *homeomorphism* between them is just a continuous map with a continuous inverse. By Lemma 1.2, this is saying that the open sets in the two spaces correspond under the bijection, and hence that the map is a *topological* equivalence between the spaces; the two spaces are then said to be *homeomorphic*. Thus for instance, the open unit disc $D \subset \mathbf{R}^2$ is homeomorphic to the whole plane (both spaces with the Euclidean metric) via the map $f : D \to \mathbf{R}^2$ given by $f(\mathbf{x}) = \mathbf{x}/(1 - \|\mathbf{x}\|)$, with inverse $g : \mathbf{R}^2 \to D$ given by $g(\mathbf{y}) = \mathbf{y}/(1 + \|\mathbf{y}\|)$.

All the geometries studied in this book will have natural underlying metric spaces. These metric spaces will however have particularly nice properties; in particular they have the property that every point has an open neighbourhood which is homeomorphic to the open disc in $\mathbf{R}^2$ (this is essentially the statement that the metric space is what is called a two-dimensional *topological manifold*). We conclude this section by giving two examples of metric spaces, both of which are defined geometrically but neither of which have this last property.

Example (British Rail metric) Consider the plane $\mathbf{R}^2$ with Euclidean metric d, and let O denote the origin. We define a new metric d_1 on $\mathbf{R}^2$ by

$$d_1(P, Q) = \begin{cases} 0 & \text{if } P = Q, \\ d(P, O) + d(O, Q) & \text{if } P \neq Q. \end{cases}$$

We note that, for $P \neq O$, any small enough open ball round P consists of just the point P; therefore, no open neighbourhood of P is homeomorphic to an open disc in $\mathbf{R}^2$. When the author was an undergraduate in the UK, this was known as the *British Rail* metric; here O represented London, and all train journeys were forced to go via London! Because of a subsequent privatization of the UK rail network, the metric should perhaps be renamed.

Example (London Underground metric) Starting again with the Euclidean plane $(\mathbf{R}^2, d)$, we choose a finite set of points $P_1, \ldots, P_N \in \mathbf{R}^2$. Given two points $P, Q \in \mathbf{R}^2$, we define a distance function d_2 (for $N > 1$, it is not a metric) by

$$d_2(P,Q) = \min\{d(P,Q),\ \min_{i,j}\{d(P,P_i) + d(P_j,Q)\}\}.$$

This function satisfies all the properties of a metric *except* that $d_2(P,Q)$ may be zero even when $P \neq Q$. We can however form a quotient set X from $\mathbf{R}^2$ by identifying all the points P_i to a single point $\bar{P}$ (formally, we take the quotient of $\mathbf{R}^2$ by the equivalence relation which sets two points P, Q to be equivalent if and only if $d_2(P,Q) = 0$), and it is then easily checked that d_2 induces a metric d^* on X. The name given to this metric refers to the underground railway in London; the points P_i represent the idealized stations in this network, idealized because we assume that no walking is involved if we wish to travel between any two stations of the network (even if such a journey involves changing trains). The distance d_2 between two points of $\mathbf{R}^2$ is the minimum distance one has to walk between the two points, given that one has the option of walking to the nearest underground station and travelling by train to the station nearest to one's destination.

We note that any open ball of sufficiently small radius ε round the point $\bar{P}$ of X corresponding to the points $P_1, \ldots, P_N \in \mathbf{R}^2$ is the union of the open balls $B(P_i, \varepsilon) \subset \mathbf{R}^2$, with the points $P_1, \ldots, P_N$ identified. In particular, the punctured ball $B(\bar{P}, \varepsilon)\setminus\{\bar{P}\}$ in X is identified as a disjoint union of punctured balls $B(P_i, \varepsilon)\setminus\{P_i\}$ in the plane. Once we have introduced the concept of connectedness in Section 1.4, it will be clear that this latter space is not connected for $N \geq 2$, and hence cannot be homeomorphic to an open punctured disc in $\mathbf{R}^2$, which from Section 1.4 is clearly both connected and path connected. It will follow then that our open ball in X is not homeomorphic to an open disc in $\mathbf{R}^2$. A similar argument shows that the same is true for any small enough open neighbourhood of $\bar{P}$, and hence no open neighbourhood of $\bar{P}$ in X can be homeomorphic to an open disc in $\mathbf{R}^2$.

1.2 Isometries

We defined above the concept of a *homeomorphism* or *topological equivalence*; the geometries in this course however come equipped with metrics, and so we shall be interested in the stronger notion of an *isometry*.

Definition 1.3 A map $f : (X, d_X) \to (Y, d_Y)$ between metric spaces is called an *isometry* if it is surjective and it preserves distances, that is

$$d_Y(f(x_1), f(x_2)) = d_X(x_1, x_2)$$

for all $x_1, x_2 \in X$.

A few observations are due here:

- The second condition in (1.3) implies that the map is injective. Thus an isometry is necessarily bijective. A map satisfying the second condition without necessarily being surjective is usually called an *isometric embedding*.
- The second condition implies that an isometry is continuous, as is its inverse. Hence isometries are homeomorphisms. However, the homeomorphism defined above between the unit disc and the Euclidean plane is clearly not an isometry.
- An isometry of a metric space to itself is also called a *symmetry* of the space. The isometries of a metric space X to itself form a group under composition of maps, called the *isometry group* or the *symmetry group* of the space, denoted $\mathrm{Isom}(X)$.

Definition 1.4 We say that a group G *acts* on a set X if there is a map $G \times X \to X$, the image of (g, x) being denoted by $g(x)$, such that

(i) the identity element in G corresponds to the identity map on X, and
(ii) $(g_1 g_2)(x) = g_1(g_2(x))$ for all $x \in X$ and $g_1, g_2 \in G$.

We say that the action of G is *transitive* on X if, for all $x, y \in X$, there exists $g \in G$ with $g(x) = y$.

For X a metric space, the obvious action of $\mathrm{Isom}(X)$ on X will not usually be transitive. For the important special cases however of Euclidean space, the sphere (Chapter 2), the locally Euclidean torus (Chapter 3) and the hyperbolic plane (Chapter 5), this action is transitive — these geometries may therefore be thought of as looking the same from every point.

Let us now consider the case of Euclidean space $\mathbf{R}^n$, with its standard inner-product $(\ ,\ )$ and distance function d. An isometry of $\mathbf{R}^n$ is sometimes called a *rigid motion*. We note that any translation of $\mathbf{R}^n$ is an isometry, and hence the isometry group $\mathrm{Isom}(\mathbf{R}^n)$ acts transitively on $\mathbf{R}^n$.

We recall that an $n \times n$ matrix A is called *orthogonal* if $A^t A = A A^t = I$, where A^t denotes the transposed matrix. Since

$$\begin{aligned}(A\mathbf{x}, A\mathbf{y}) &= (A\mathbf{x})^t (A\mathbf{y}) \\ &= \mathbf{x}^t A^t A \mathbf{y} \\ &= (\mathbf{x}, A^t A \mathbf{y}) \\ &= (A^t A \mathbf{x}, \mathbf{y}),\end{aligned}$$

we have that A is orthogonal if and only if $(A\mathbf{x}, A\mathbf{y}) = (\mathbf{x}, \mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$.

Since $(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\{\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2\}$, a matrix A is orthogonal if and only if $\|A\mathbf{x}\| = \|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbf{R}^n$. Thus, if a map $f : \mathbf{R}^n \to \mathbf{R}^n$ is defined by $f(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, for some $\mathbf{b} \in \mathbf{R}^n$, then

$$d(f(\mathbf{x}), f(\mathbf{y})) = \|A(\mathbf{x} - \mathbf{y})\|,$$

and so f is an isometry if and only if A is orthogonal.

Theorem 1.5 *Any isometry $f : \mathbf{R}^n \to \mathbf{R}^n$ is of the form $f(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, for some orthogonal matrix A and vector $\mathbf{b} \in \mathbf{R}^n$.*

Proof Let $e_1, \ldots, e_n$ denote the standard basis of $\mathbf{R}^n$. We set $\mathbf{b} = f(\mathbf{0})$, and $\mathbf{a}_i = f(e_i) - \mathbf{b}$ for $i = 1, \ldots, n$. Then, for all i,

$$\begin{aligned} \|\mathbf{a}_i\| &= \|f(e_i) - f(\mathbf{0})\| = d(f(e_i), f(\mathbf{0})) \\ &= d(e_i, \mathbf{0}) = \|e_i\| = 1. \end{aligned}$$

For $i \neq j$,

$$\begin{aligned} (\mathbf{a}_i, \mathbf{a}_j) &= -\frac{1}{2}\left\{\|\mathbf{a}_i - \mathbf{a}_j\|^2 - \|\mathbf{a}_i\|^2 - \|\mathbf{a}_j\|^2\right\} \\ &= -\frac{1}{2}\left\{\|f(e_i) - f(e_j)\|^2 - 2\right\} \\ &= -\frac{1}{2}\left\{\|e_i - e_j\|^2 - 2\right\} = 0. \end{aligned}$$

Now let A be the matrix with columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$. Since the columns form an orthonormal basis, A is orthogonal. Let $g : \mathbf{R}^n \to \mathbf{R}^n$ be the isometry given by

$$g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}.$$

Then $g(\mathbf{x}) = f(\mathbf{x})$ for $\mathbf{x} = \mathbf{0}, e_1, \ldots, e_n$. Now g has inverse g^{-1}, where

$$g^{-1}(\mathbf{x}) = A^{-1}(\mathbf{x} - \mathbf{b}) = A^t(\mathbf{x} - \mathbf{b});$$

therefore $h = g^{-1} \circ f$ is an isometry fixing $\mathbf{0}, e_1, \ldots, e_n$.

We claim that $h = \mathrm{id}$, and hence $f = g$ as required.

Claim $h : \mathbf{R}^n \to \mathbf{R}^n$ is the identity map.

Proof of Claim For general $\mathbf{x} = \sum x_i e_i$, set

$$h(\mathbf{x}) = \mathbf{y} = \sum y_i e_i.$$

We observe that

$$\begin{aligned} d(\mathbf{x}, e_i)^2 &= \|\mathbf{x}\|^2 + 1 - 2x_i \\ d(\mathbf{x}, \mathbf{0})^2 &= \|\mathbf{x}\|^2 \\ d(\mathbf{y}, e_i)^2 &= \|\mathbf{y}\|^2 + 1 - 2y_i \\ d(\mathbf{y}, \mathbf{0})^2 &= \|\mathbf{y}\|^2. \end{aligned}$$

Since h is an isometry such that $h(\mathbf{0}) = \mathbf{0}$, $h(e_i) = e_i$ and $h(\mathbf{x}) = \mathbf{y}$, we deduce that $\|\mathbf{y}\|^2 = \|\mathbf{x}\|^2$ and $x_i = y_i$ for all i, i.e. $h(\mathbf{x}) = \mathbf{y} = \sum x_i e_i = \mathbf{x}$ for all $\mathbf{x}$. Thus $h = \mathrm{id}$. □

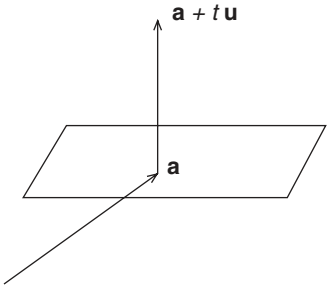
Example (Reflections in affine hyperplanes) If $H \subset \mathbf{R}^n$ is an affine hyperplane defined by

$$\mathbf{u} \cdot \mathbf{x} = c$$

for some unit vector $\mathbf{u}$ and constant $c \in \mathbf{R}$, we define a map R_H, the *reflection* in H, by

$$R_H : \mathbf{x} \mapsto \mathbf{x} - 2(\mathbf{x} \cdot \mathbf{u} - c)\,\mathbf{u}.$$

Note that $R_H(\mathbf{x}) = \mathbf{x}$ for all $\mathbf{x} \in H$. Moreover, one checks easily that any $\mathbf{x} \in \mathbf{R}^n$ may be written uniquely in the form $\mathbf{a} + t\mathbf{u}$, for some $\mathbf{a} \in H$ and $t \in \mathbf{R}$, and that $R_H(\mathbf{a} + t\mathbf{u}) = \mathbf{a} - t\mathbf{u}$.



For any $\mathbf{a}, \mathbf{a}' \in H$ and $t, t' \in \mathbf{R}$, we observe that $(\mathbf{a} - \mathbf{a}') \cdot \mathbf{u} = 0$, and therefore

$$\begin{aligned}\|(\mathbf{a} + t\mathbf{u}) - (\mathbf{a}' + t'\mathbf{u})\|^2 &= \|(\mathbf{a} - \mathbf{a}') + (t - t')\mathbf{u}\|^2 = \|(\mathbf{a} - \mathbf{a}') - (t - t')\mathbf{u}\|^2 \\ &= \|(\mathbf{a} - t\mathbf{u}) - (\mathbf{a}' - t'\mathbf{u})\|^2 = \|R_H(\mathbf{a} + t\mathbf{u}) - R_H(\mathbf{a}' + t'\mathbf{u})\|^2;\end{aligned}$$

hence R_H is an isometry.

Conversely, suppose that S is an isometry fixing H (pointwise) and choose any $\mathbf{a} \in H$; if $T_\mathbf{a}$ denotes translation by $\mathbf{a}$, i.e. $T_\mathbf{a}(\mathbf{x}) = \mathbf{x} + \mathbf{a}$ for all $\mathbf{x}$, then the conjugate $R = T_{-\mathbf{a}} S T_\mathbf{a}$ is an isometry fixing pointwise the hyperplane $H' = T_{-\mathbf{a}} H$ through the origin. If H is given by $\mathbf{x} \cdot \mathbf{u} = c$ (where $c = \mathbf{a} \cdot \mathbf{u}$), then H' is given by $\mathbf{x} \cdot \mathbf{u} = 0$. Therefore, $(R\mathbf{u}, \mathbf{x}) = (R\mathbf{u}, R\mathbf{x}) = (\mathbf{u}, \mathbf{x}) = 0$ for all $\mathbf{x} \in H'$, and so $R\mathbf{u} = \lambda\mathbf{u}$ for some λ.

But $\|R\mathbf{u}\|^2 = 1 \implies \lambda^2 = 1 \implies \lambda = \pm 1$. Since, by (1.5), R is a linear map, we deduce that either $R = \mathrm{id}$ or $R = R_{H'}$.

Therefore either $S = \mathrm{id}$, or $S = T_\mathbf{a} R_{H'} T_{-\mathbf{a}}$:

$$\mathbf{x} \mapsto \mathbf{x} - \mathbf{a} \mapsto (\mathbf{x} - \mathbf{a}) - 2(\mathbf{x} \cdot \mathbf{u} - \mathbf{a} \cdot \mathbf{u})\mathbf{u} \mapsto \mathbf{x} - 2(\mathbf{x} \cdot \mathbf{u} - c)\mathbf{u},$$

i.e. $S = R_H$.

We shall need the following elementary but useful fact about reflections.

Lemma 1.6 *Given points $P \neq Q$ in $\mathbf{R}^n$, there exists a hyperplane H, consisting of the points of $\mathbf{R}^n$ which are equidistant from P and Q, for which the reflection R_H swaps the points P and Q.*

Proof If the points P and Q are represented by vectors $\mathbf{p}$ and $\mathbf{q}$, we consider the perpendicular bisector of the line segment PQ, which is a hyperplane H with equation

$$\mathbf{x} \cdot (\mathbf{p} - \mathbf{q}) = \frac{1}{2}(\|\mathbf{p}\|^2 - \|\mathbf{q}\|^2).$$

An elementary calculation confirms that H consists precisely of the points which are equidistant from P and Q. We observe that $(\mathbf{p} - \mathbf{q})/2$ is normal to H; moreover $(\mathbf{p}+\mathbf{q})/2 \in H$ and hence is fixed under R_H. Noting that $\mathbf{p} = (\mathbf{p}+\mathbf{q})/2+(\mathbf{p}-\mathbf{q})/2$ and $\mathbf{q} = (\mathbf{p}+\mathbf{q})/2-(\mathbf{p}-\mathbf{q})/2$, it follows therefore that $R_H(\mathbf{p}) = \mathbf{q}$ and $R_H(\mathbf{q}) = \mathbf{p}$. □

Reflections in hyperplanes form the building blocks for all the isometries, in that they yield generators for the full group of isometries. More precisely, we have the following classical result.

Theorem 1.7 *Any isometry of $\mathbf{R}^n$ can be written as the composite of at most $(n + 1)$ reflections.*

Proof As before, we let $e_1, \ldots, e_n$ denote the standard basis of $\mathbf{R}^n$, and we consider the $n+1$ points represented by the vectors $\mathbf{0}, e_1, \ldots, e_n$. Suppose that f is an arbitrary isometry of $\mathbf{R}^n$, and consider the images $f(\mathbf{0}), f(e_1), \ldots, f(e_n)$ of these vectors. If $f(\mathbf{0}) = \mathbf{0}$, we set $f_1 = f$ and proceed to the next step. If not, we use Lemma 1.6; if H_0 denotes the hyperplane of points equidistant from $\mathbf{0}$ and $f(\mathbf{0})$, the reflection R_{H_0} swaps the points. In particular, if we set $f_1 = R_{H_0} \circ f$, then f_1 is an isometry (being the composite of isometries) which fixes $\mathbf{0}$.

We now repeat this argument. Suppose, by induction, that we have an isometry f_i, which is the composite of our original isometry f with at most i reflections, which fixes all the points $\mathbf{0}, e_1, \ldots, e_{i-1}$. If $f_i(e_i) = e_i$, we set $f_{i+1} = f_i$. Otherwise, we let H_i denote the hyperplane consisting of points equidistant from e_i and $f_i(e_i)$. Our assumptions imply that $\mathbf{0}, e_1, \ldots, e_{i-1}$ are equidistant from e_i and $f_i(e_i)$, and hence lie in H_i. Thus R_{H_i} fixes $\mathbf{0}, e_1, \ldots, e_{i-1}$ and swaps e_i and $f_i(e_i)$, and so the composite $f_{i+1} = R_{H_i} \circ f_i$ is an isometry fixing $\mathbf{0}, e_1, \ldots, e_i$.

After $n + 1$ steps, we attain an isometry f_{n+1}, the composite of f with at most $n + 1$ reflections, which fixes all of $\mathbf{0}, e_1, \ldots, e_n$. We saw however in the proof of Theorem 1.5 that this is sufficient to imply that f_{n+1} is the identity, from which it follows that the original isometry f is the composite of at most $n+1$ reflections. □

Remark If we know that an isometry f already fixes the origin, the above proof shows that it can be written as the composite of at most n reflections. The above theorem for $n = 2$, that any isometry of the Euclidean plane may be written as the composite of at most three reflections, has an analogous result in both the spherical and hyperbolic geometries, introduced in later chapters.

1.3 The group $O(3, \mathbf{R})$

A natural subgroup of $\mathrm{Isom}(\mathbf{R}^n)$ consists of those isometries fixing the origin, which can therefore be written as a composite of at most n reflections. By Theorem 1.5, this subgroup may be identified with the group $O(n) = O(n, \mathbf{R})$ of $n \times n$ orthogonal matrices, the *orthogonal group*. If $A \in O(n)$, then

$$\det A \, \det A^t = \det(A)^2 = 1,$$

and so $\det A = \pm 1$. The subgroup of $O(n)$ consisting of elements with $\det A = 1$ is denoted $SO(n)$, and is called the *special orthogonal group*. The isometries f of $\mathbf{R}^n$ of the form $f(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, for some $A \in SO(n)$ and $\mathbf{b} \in \mathbf{R}^n$, are called the *direct isometries* of $\mathbf{R}^n$; they are the isometries which can be expressed as a product of an *even* number of reflections.

Example Let us consider the group $O(2)$, which may also be identified as the group of isometries of $\mathbf{R}^2$ fixing the origin. Note that

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in O(2) \iff a^2 + c^2 = 1,\ b^2 + d^2 = 1,\ ab + cd = 0.$$

For such a matrix $A \in O(2)$, we may set

$$\begin{aligned} a &= \cos\theta, & c &= \sin\theta, \\ b &= -\sin\phi, & d &= \cos\phi, \end{aligned}$$

with $0 \leq \theta, \phi < 2\pi$. Then, the equation $ab + cd = 0$ gives $\tan\theta = \tan\phi$, and therefore $\phi = \theta$ or $\theta \pm \pi$.

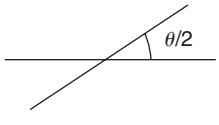
In the first case,

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

is an anticlockwise rotation through θ, and $\det A = 1$; it is therefore the product of two reflections. In the second case,

$$A = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$$

is a reflection in the line at angle $\theta/2$ to the x-axis, and $\det A = -1$.



In summary therefore, the elements of $SO(2)$ correspond to the rotations of $\mathbf{R}^2$ about the origin, whilst the elements of $O(2)$ which are not in $SO(2)$ correspond to reflections in a line through the origin.

In this section, we study in more detail the case $n = 3$. We suppose then that $A \in O(3)$. Consider first the case when $A \in SO(3)$, i.e.

$$\boxed{\det A = 1}$$

Then

$$\begin{aligned} & \det(A - I) = \det(A^t - I) = \det A(A^t - I) = \det(I - A) \\ \Longrightarrow\ & \det(A - I) = 0, \end{aligned}$$

i.e. $+1$ is an eigenvalue. There exists therefore an eigenvector v_1 (where we may assume $\|v_1\| = 1$) such that $Av_1 = v_1$. Set $W = \langle v_1 \rangle^{\perp}$ to be the orthogonal complement to the space spanned by v_1. If $w \in W$, then $(Aw, v_1) = (Aw, Av_1) = (w, v_1) = 0$. Thus $A(W) \subset W$ and $A|_W$ is a rotation of the two-dimensional space W, since it is an isometry of W fixing the origin and has determinant one. If $\{v_2, v_3\}$ is an orthonormal basis for W, the action of A on $\mathbf{R}^3$ is represented with respect to the orthonormal basis $\{v_1, v_2, v_3\}$ by the matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}.$$

This is just rotation about the axis spanned by v_1 through an angle θ. It may be expressed as a product of two reflections.

Now suppose

$$\boxed{\det A = -1}$$

Using the previous result, there exists an orthonormal basis with respect to which $-A$ is a rotation of the above form, and so A takes the form

$$\begin{pmatrix} -1 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{pmatrix}$$

with $\phi = \theta + \pi$. Such a matrix A represents a *rotated reflection*, rotating through an angle ϕ about a given axis and then reflecting in the plane orthogonal to the axis. In the special case $\phi = 0$, A is a pure reflection. The general rotated reflection may be expressed as a product of three reflections.