

A course in low-dimensional geometry

Mark Steinberger

CONTENTS

Preface	7
Terminology	7
Functions	7
Relations	8
Acknowledgements	10
1. Some linear algebra	11
1.1. Vector spaces and linear maps	11
1.2. Spans and linear independence	13
1.3. Matrices	16
1.4. Block matrices and their multiplication	19
1.5. Dimension	20
1.6. Rank	23
1.7. Direct sums	24
1.8. Base change	26
1.9. Exercises	32
2. Basic Euclidean geometry	34
2.1. Lines in \mathbb{R}^n	34
2.2. Lines in the plane	38
2.3. Inner products and distance	39
2.4. Euclidean isometries are affine	44
2.5. Affine functions and linearity	47
2.6. Affine automorphisms of \mathbb{R}^n	49
2.7. Similarities	50
2.8. Convex and affine hulls; affine subspaces and maps	51
2.8.1. Convex and affine hulls	52
2.8.2. Joins	56
2.8.3. Affine maps	58
2.8.4. Affine and convex hulls of infinite sets	62
2.8.5. Convex subsets of lines	63
2.9. Affine independence, interiors and faces	64
2.9.1. Affine independence	64
2.9.2. Interiors	68
2.9.3. Faces	74
2.9.4. Examples	80
2.10. Exercises	82
3. Groups	83
3.1. Definition and examples	83

3.2.	Orders of elements	86
3.3.	Conjugation and normality	87
3.4.	Homomorphisms	89
3.5.	A matrix model for isometries and affine maps	94
3.6.	G -sets	96
3.7.	Direct products	98
4.	Linear isometries	99
4.1.	Orthonormal bases and orthogonal matrices	99
4.2.	Gramm–Schmidt	105
4.3.	Orthogonal complements	106
4.4.	Applications to rank	109
4.5.	Invariant subspaces for linear isometries	110
5.	Isometries of \mathbb{R}^2	112
5.1.	Reflections	112
5.2.	Trigonometric functions	121
5.3.	Linear isometries of \mathbb{R}^2 : calculation of $O(2)$	128
5.4.	Angles in \mathbb{R}^2 and \mathbb{R}^n ; the cosine law; orientation in \mathbb{R}^2	132
5.4.1.	Angles in \mathbb{R}^2	132
5.4.2.	Angles in \mathbb{R}^n	133
5.4.3.	Orientation in \mathbb{R}^2	135
5.5.	Calculus of isometries of \mathbb{R}^2	137
5.5.1.	Glide reflections	138
5.5.2.	Calculating composites of isometries	140
5.5.3.	Calculus of reflections	145
5.6.	Classical results from Euclidean geometry	149
5.7.	Exercises	149
6.	Groups of symmetries: planar figures	151
6.1.	Symmetry in \mathbb{R}^n ; congruence and similarity	151
6.1.1.	The group of symmetries of $X \subset \mathbb{R}^n$	151
6.1.2.	The subgroups $\mathcal{T}(X)$ and $\mathcal{O}(X)$ of $\mathcal{S}(X)$	152
6.1.3.	Congruence and similarity	152
6.2.	Symmetries of polytopes	154
6.2.1.	Generalities	154
6.2.2.	Centroids	157
6.2.3.	Symmetries of the n -cube	158
6.2.4.	Symmetries of the regular n -gon in \mathbb{R}^2	159
6.3.	Geometry meets number theory: the golden mean	159
6.4.	Symmetries of points and lines in \mathbb{R}^2	161

6.5.	Dihedral groups	162
6.6.	Index 2 subgroups	165
6.7.	Left cosets; orbit counting; the first Noether theorem	168
6.8.	Leonardo's theorem	172
6.9.	Orbits and isotropy in the plane	174
6.10.	Frieze groups	176
6.11.	Fundamental regions and orbit spaces	184
6.12.	Translation lattices in \mathbb{R}^n	188
6.13.	Orientation-preserving wallpaper groups	199
6.13.1.	Groups admitting 4-centers	201
6.13.2.	Groups admitting 6-centers	204
6.13.3.	Groups admitting 3-centers but not 6-centers	207
6.13.4.	The remaining cases	210
6.14.	General wallpaper groups	213
6.14.1.	Wallpaper groups with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_4$	218
6.14.2.	Wallpaper groups with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_6$	224
6.14.3.	Wallpaper groups with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_3$	227
6.14.4.	Wallpaper groups with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_2$	232
6.14.5.	Wallpaper groups with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_1$	239
6.15.	Exercises	246
7.	Linear isometries of \mathbb{R}^3	262
7.1.	Linear orientations of \mathbb{R}^n	262
7.2.	Rotations	264
7.3.	Cross products	269
7.4.	Reflections	271
7.5.	Rotation-reflections	274
7.6.	Symmetries of the Platonic solids	276
7.6.1.	The cube and the regular tetrahedron	276
7.6.2.	The regular tetrahedron	277
7.6.3.	Calculation of $\mathcal{O}(\mathbf{C})$	278
7.6.4.	The dodecahedron	279
7.6.5.	Duality	290
7.6.6.	The octahedron	291
7.6.7.	Dual of the tetrahedron	292
7.6.8.	The icosahedron	293
7.7.	Exercises	296
8.	Spheres and other manifolds	297
8.1.	Some advanced calculus	297

8.2.	Orientation properties of nonlinear mappings in \mathbb{R}^n	302
8.3.	Topological manifolds; \mathbb{S}^{n-1}	303
8.4.	Smooth manifolds	306
8.5.	Products of manifolds	311
8.6.	Oriented atlases	312
8.7.	Exercises	313
9.	Spherical geometry	315
9.1.	Arc length and distance in \mathbb{S}^n ; isometries of \mathbb{S}^n	315
9.2.	Lines and angles in \mathbb{S}^2	326
9.3.	Spherical triangles	329
9.4.	Isometries of \mathbb{S}^2	330
9.5.	Perpendicular bisectors	332
9.6.	Exercises	333
10.	Tangent bundles	335
10.1.	The local model	335
10.2.	The tangent bundle of a smooth manifold	336
10.3.	Tangent bundles of products	341
10.4.	Immersions and embeddings; submersions	342
10.5.	Orientation of manifolds	345
10.6.	Vector fields	346
11.	Riemannian manifolds	347
11.1.	Riemannian metrics	347
11.2.	Arc length, distance and angles	350
11.3.	Geodesics	355
11.3.1.	Geodesics in the local model	355
11.3.2.	Geodesics in general Riemannian manifolds	359
11.4.	The exponential map	361
12.	Hyperbolic geometry	367
12.1.	Boundary of \mathbb{H} and compactification of \mathbb{C} .	368
12.2.	Möbius transformations	371
12.3.	Isometric properties of Möbius transformations	377
12.4.	Hyperbolic lines and geodesics	379
12.5.	Incidence relations and transitivity properties	390
12.6.	Hyperbolic line segments	392
12.7.	Parallels and perpendiculars	394
12.8.	Reflections	397
12.9.	Generalized Möbius transformations	401
12.10.	Calculus of isometries	405

12.11. Exercises	406
Appendix A. Spaces with identifications	408
A.1. Metric topology	408
A.2. Subspace topology	412
A.3. Quotient topology	413
A.4. Group actions and orbit spaces	418
A.5. Basis for a topology	421
A.6. Properly discontinuous actions	424
A.6.1. Product topology	425
A.6.2. Disjoint unions	429
A.7. Topology of the orbit space	429
Appendix B. Compactness	434
B.1. Heine–Borel	434
B.2. Maps out of compact spaces	438
B.3. Cones and convex bodies	439
B.3.1. Cones	439
B.3.2. Convex bodies	440
References	444

Preface

This book has grown out of two courses the author has taught at the University at Albany. The first course investigates the rigid motions (isometries) of the Euclidean plane and develops rosette, frieze and wallpaper groups. In particular, we study discrete subgroups of the group of isometries of the plane (i.e., two-dimensional crystallographic groups).

The second course develops Euclidean, spherical and hyperbolic geometry in dimension two from an analytic point of view, providing realizations of Euclidean and non-Euclidean geometry from an analytic, rather than axiomatic, point of view, and develops tools like the cosine law and the Gauss–Bonnet theorem that provide a deeper insight into the geometry than is provided by the axioms alone. This has particular value for prospective high school teachers, as the cosine law is actually used in the high school curriculum, and a unified development of both that and Euclidean geometry puts things in a modern perspective that might be very useful to the high school students themselves.

In particular, we show how isometries provide clean and direct proofs of the basic theorems in Euclidean geometry and their analogues on the sphere and in hyperbolic space.

Terminology. We shall be primarily interested in subspace of Euclidean n -space, \mathbb{R}^n , defined as the space of all n -tuples of real numbers:

$$\mathbb{R}^n = \{(x_1, \dots, x_n) : x_1, \dots, x_n \in \mathbb{R}\}.$$

Thus, formally, \mathbb{R}^n is the cartesian product of n copies of the real line. More generally, if X_1, \dots, X_n are sets, their cartesian product is given by

$$X_1 \times \cdots \times X_n = \{(x_1, \dots, x_n) : x_i \in X_i \text{ for } i = 1, \dots, n\}.$$

If X is a finite set, we write $|X|$ for the number of elements in X . An elementary counting argument gives:

Lemma 0.0.1. *Let X_1, \dots, X_n be finite sets. Then $X_1 \times \cdots \times X_n$ is finite and the number of elements in it is the product of the numbers of elements in the individual sets:*

$$(0.0.1) \quad |X_1 \times \cdots \times X_n| = |X_1| \cdots |X_n|.$$

Given subset $X, Y \subset Z$, we write $X \setminus Y$ for the set-theoretic difference:

$$\begin{aligned} X \setminus Y &= \{x \in X : x \notin Y\} \\ &= X \setminus (X \cap Y). \end{aligned}$$

Functions. A function

$$f : X \rightarrow Y$$

is sometimes called a *map*, a *mapping*, or a *transformation*. We call X its *domain* and Y its *codomain*.

We say that f is *one-to-one* or *injective* if

$$f(x) = f(y) \Rightarrow x = y.$$

An injective function is called an *injection*.

The *image* or *range* of f is

$$\text{im } f = f(X) = \{y \in Y : y = f(x) \text{ for some } x \in X\}.$$

For $Z \subset X$, we write $f(Z) = \{f(z) : z \in Z\}$ for the image of $f|_Z : Z \rightarrow Y$, the restriction of f to Z .

We say that $f : X \rightarrow Y$ is *onto* or *surjective* if its range is all of Y , i.e., if the range is equal to the codomain. A surjective function is called a *surjection*.

If f is both injective and surjective, it is said to be *bijective*. A bijective function is called a *bijection* or *one-to-one correspondence*.

If $f : X \rightarrow Y$ is bijective, its inverse function

$$f^{-1} : Y \rightarrow X$$

is defined by setting $f^{-1}(y)$ to be the unique $x \in X$ with $f(x) = y$: x exists because f is surjective and is unique because f is injective. The inverse function f^{-1} is easily seen to be bijective.

Relationships between functions are often expressed via a diagram. We say

$$\begin{array}{ccc} X & \xrightarrow{h} & Z \\ & \searrow f & \nearrow g \\ & & Y \end{array}$$

commutes if $h = g \circ f$. A more general diagram commutes if any two ways of getting from one node to another give the same function. Thus,

$$\begin{array}{ccc} X & \xrightarrow{f} & Y \\ g \downarrow & & \downarrow h \\ Z & \xrightarrow{k} & W \end{array}$$

commutes if $k \circ g = h \circ f$.

Relations. Formally, a relation R on a set X is a subset $R \subset X \times X$. It is customary to write xRy if the ordered pair (x, y) is in R (and usually a symbolic operator such as \sim is used rather than a letter like R). Relations express a relationship between x and y . One generally does not talk about the subset R at all, but gives a criterion for xRy to hold.

Examples 0.0.2.

- (1) The usual ordering \leq on \mathbb{R} is a relation. It is an example of an *order relation* to be discussed below.

- (2) Let X be a set and let $\mathcal{P}(X)$ be the set of all subsets of X . ($\mathcal{P}(X)$ is called the power set of X .) Then the standard inclusion of subsets $S \subset T$ gives a relation on $\mathcal{P}(X)$. It is also an order relation.
- (3) Among the integers, there is a relation, let's call it \equiv , defined by setting $a \equiv b$ if they have the same parity (i.e., either both are odd or both are even). Note that we can test for parity by taking powers of (-1) : n is even if $(-1)^n = 1$ and is odd if $(-1)^n = (-1)$. Thus, $a \equiv b$ if and only if $(-1)^a = (-1)^b$. This relation is generally called congruence mod 2, and is written $a \equiv b \pmod{2}$. It is an example of what's known as an *equivalence relation* to be discussed below.
- (4) Let $f : X \rightarrow Y$ be a function. Define the relation \sim on X by setting $x_1 \sim x_2$ if $f(x_1) = f(x_2)$. We call this the relation on X induced by f . This will also be seen to be an equivalence relation.

With these examples in mind, we give some potential properties of relations.

Definition 0.0.3. Let R be a relation on X

- R is *reflexive* if xRx for all $x \in X$.
- R is *symmetric* if $xRy \Rightarrow yRx$.
- R is *antisymmetric* if $(xRy \text{ and } yRx) \Rightarrow x = y$.
- R is *transitive* if $(xRy \text{ and } yRz) \Rightarrow xRz$.

We say R is an *order relation* or *partial ordering* if it is reflexive, antisymmetric and transitive. We say it is a *total ordering* if in addition each pair of elements in X is comparable, i.e., for $x, y \in X$, either xRy or yRx .

We say R is an *equivalence relation* if it is reflexive, symmetric and transitive.

Remark 0.0.4.

- The relation \leq on \mathbb{R} is a total ordering.
- The relation \subset on $\mathcal{P}(X)$ is a partial ordering, but not a total ordering if X has more than one element: if $x \neq y \in X$, then neither $\{x\}$ nor $\{y\}$ is contained in the other, so these subsets are not related in $\mathcal{P}(X)$.
- For a set X , let $\mathcal{P}_f(X) \subset \mathcal{P}(X)$ be the finite subsets of X . Then there is an equivalence relation on $\mathcal{P}_f(X)$ given by setting $S \sim T$ if S and T have the same number of elements.

Equivalence relations are about grouping similar things together, while order relations are about comparing different things. In the former case the notion of equivalence classes is valuable.

Definition 0.0.5. If \sim is an equivalence relation on X we define the equivalence class of $x \in X$ by

$$[x] = \{y \in X : x \sim y\} \subset X.$$

The set of all equivalence classes, X/\sim , is

$$X/\sim = \{[x] : x \in X\}$$

and the canonical map

$$\pi : X \rightarrow X/\sim$$

is defined by $\pi(x) = [x]$ for all $x \in X$.

Example 0.0.6. Let $f : X \rightarrow Y$ be a function and let \sim be the relation on X induced by f (Example 0.0.2(4)). Then the equivalence classes are the subsets $f^{-1}(y)$ with y in the image of f .

The key result about equivalence relations is the following.

Lemma 0.0.7. *Let \sim be an equivalence relation on X . Then every element of X belongs to exactly one equivalence class. A given element x lies in $[x]$, and if $[x] \cap [y] \neq \emptyset$, then $[x] = [y]$. Finally, if $\pi : X \rightarrow X/\sim$ is the canonical map, then π is onto and*

$$(0.0.2) \quad [x] = \pi^{-1}([x]).$$

Of course, here the $[x]$ on the right is a point in the set X/\sim (which consists of subsets of X), while the $[x]$ on the left is a subset of X .

Proof. If $z \in [x] \cap [y]$, then $x \sim z \sim y$ by symmetry, so $x \sim y$ by transitivity. Transitivity then shows that $[y] \subset [x]$. By symmetry, $y \sim x$, so $[x] \subset [y]$, and both must coincide with $[z]$ by the argument just given. In particular,

$$y \in [x] \quad \Leftrightarrow \quad [y] = [x],$$

and (0.0.2) follows. □

Acknowledgements. Many mathematicians have been generous with their thoughts and insights. I wish to thank Noam Elkies, Charles Frohman, Greg Kuperberg, John Randall, Peter Shalen and Deane Yang.

I have also learned from my students, both past and present. In addition to contributing ideas and asking interesting and useful questions, they have caught mistakes and typos for which I am grateful. Special thanks go to Gabe Holmes and Aaron Wolff.

1. Some linear algebra

We review some basic linear algebra. The reader may wish to start with Chapter 2 and use this chapter as a reference work.

1.1. Vector spaces and linear maps. We will make significant use of the

standard vector operations in \mathbb{R}^n : for $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ and $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ in \mathbb{R}^n and $a \in \mathbb{R}$ we have the operations of vector addition and of scalar multiplication:

$$x + y = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix},$$

$$ax = \begin{bmatrix} ax_1 \\ \vdots \\ ax_n \end{bmatrix}.$$

These operations satisfy the following properties:

- (1) Vector addition is associative and commutative with additive identity element $0 = (0, \dots, 0)$, i.e.:
 - (a) $(x + y) + z = x + (y + z)$ for all $x, y, z \in \mathbb{R}^n$.
 - (b) $x + y = y + x$ for all $x, y \in \mathbb{R}^n$.
 - (c) $0 + x = x$ for all $x \in \mathbb{R}^n$.
- (2) The “distributivity laws” hold:

$$a(x + y) = ax + ay,$$

$$(a + b)x = ax + bx,$$

for all $x, y \in \mathbb{R}^n$ and $a, b \in \mathbb{R}$.

- (3) Scalar multiplication is “associative” and unital:
 - (a) $a(bx) = (ab)x$ for all $a, b \in \mathbb{R}$ and $x \in \mathbb{R}^n$.
 - (b) $1x = x$ for all $x \in \mathbb{R}^n$.

We sometimes write $a \cdot x$ for ax .

A set V with operations of addition and scalar multiplication satisfying (1)–(3) is called a vector space over \mathbb{R} and forms the basic object of study in an elementary linear algebra course. We shall assume the student is familiar with the material from such a course, though we will review some of it here. The reader may consult [1] for most of the omitted proofs. For determinant theory, we suggest looking at [17]. One of the central objects of study is the linear functions.

Definition 1.1.1. Let V and W be vector spaces over \mathbb{R} . A function $f : V \rightarrow W$ is said to be linear if

$$(1.1.1) \quad f(x + y) = f(x) + f(y),$$

$$(1.1.2) \quad f(ax) = af(x),$$

for all $x, y \in V$ and $a \in \mathbb{R}$. A linear function $f : V \rightarrow W$ is an isomorphism if it is one-to-one and onto. We then say V and W are isomorphic. We write

$$f : V \xrightarrow{\cong} W$$

when f is an isomorphism.

Since an isomorphism $f : V \rightarrow W$ is one-to-one and onto, it has an inverse function $f^{-1} : W \rightarrow V$. As the reader may easily check, f^{-1} is linear, and hence an isomorphism as well. Since f preserves the vector operations, an isomorphism allows us to identify the vector spaces V and W .

Definition 1.1.2. Let $f : V \rightarrow W$ be linear. The kernel of f is

$$\ker f = \{v \in V : f(v) = 0\}.$$

The following is elementary and standard.

Lemma 1.1.3. *Let $f : V \rightarrow W$ be linear. Then f is one-to-one if and only if $\ker f = \{0\}$.*

An important subject in mathematics is finding solutions of $f(x) = y$ for a function $f : X \rightarrow Y$. One of the major values of linear algebra is that such problems can be solved more easily than analogous nonlinear problems. As a result, linear approximation is often used to study nonlinear problems. This is one of the main motivations for differential calculus, as the derivative determines the best linear approximation for a function at a particular point.

Definition 1.1.4. Let $f : V \rightarrow W$ be linear. The problem

$$f(x) = y$$

is said to be homogeneous if $y = 0$ and inhomogeneous if $y \neq 0$. If $y \neq 0$, the problem

$$f(x) = 0$$

is the associated homogeneous problem to $f(x) = y$. If $y \neq 0$ and if x_0 is some specific solution of $f(x) = y$, then we call it a *particular solution* of $f(x) = y$.

The solutions of $f(x) = 0$ are precisely the kernel of f . Another standard result is the following.

Lemma 1.1.5. *Let $f : V \rightarrow W$ be linear and let v_0 be a particular solution of $f(v) = w$. Then the set of all solutions of $f(v) = w$ is*

$$(1.1.3) \quad v_0 + \ker f = \{v_0 + v : v \in \ker f\}.$$

Note that the elements of $v_0 + \ker f$ are the set of all possible particular solutions of $f(v) = w$. Thus, if $v \in \ker f$ and if we replace v_0 with $v_1 = v_0 + v$, then $v_0 + \ker f = v_1 + \ker f$.

When $V = \mathbb{R}^n$ we shall see that $v_0 + \ker f$ is an example of what's known as an affine subspace of \mathbb{R}^n .

Even to restrict attention to studying the geometry of \mathbb{R}^n , the properties of abstract vector spaces become important, as we shall make use of the subspaces of \mathbb{R}^n . For instance, in studying the rotations of \mathbb{R}^3 (and hence also of the unit sphere \mathbb{S}^2 , as studied in spherical geometry, below), it becomes valuable to take seriously the linear structure of the plane perpendicular to the axis of rotation.

Definition 1.1.6. Let V be a vector space. A subspace $W \subset V$ of V is a subset with the properties that:

- (1) For $w_1, w_2 \in W$, the sum $w_1 + w_2 \in W$.
- (2) For $w \in W$ and $a \in \mathbb{R}$, $aw \in W$.

In particular, W is a subspace if and only if the vector operations of V (addition and scalar multiplication) restrict to well-defined operations on W . In this case, it is easily seen that W is a vector space under the operations inherited from V .

Example 1.1.7. Let $f : V \rightarrow W$ be linear. Then $\ker f$ is a subspace of V and the image $\operatorname{im} f$ is a subspace of W .

As we saw above, $\ker f$ detects whether f is one-to-one, while f is onto if and only if $\operatorname{im} f = W$.

1.2. Spans and linear independence. We shall also need the notion of a basis.

Definition 1.2.1. Let V be a vector space over \mathbb{R} and let $v_1, \dots, v_k \in V$. A linear combination of v_1, \dots, v_k is a sum

$$a_1v_1 + \dots + a_kv_k$$

with $a_1, \dots, a_k \in \mathbb{R}$. We write $\operatorname{span}(v_1, \dots, v_k)$ for the set of all linear combinations of v_1, \dots, v_k :

$$\operatorname{span}(v_1, \dots, v_k) = \{a_1v_1 + \dots + a_kv_k : a_1, \dots, a_k \in \mathbb{R}\}.$$

If $\operatorname{span}(v_1, \dots, v_k) = V$ we say that v_1, \dots, v_k span V .

We say that v_1, \dots, v_k are linearly independent if

$$a_1v_1 + \dots + a_kv_k = 0 \quad \Rightarrow \quad a_1 = \dots = a_k = 0.$$

We say that v_1, \dots, v_k form a basis for V if

- (1) v_1, \dots, v_k are linearly independent.
- (2) v_1, \dots, v_k span V .

Example 1.2.2. Let $e_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \in \mathbb{R}^n$ with the 1 in the i th coordinate. Then e_1, \dots, e_n is a basis for \mathbb{R}^n , as

$$(1.2.1) \quad \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = x_1 e_1 + \dots + x_n e_n,$$

so e_1, \dots, e_n span \mathbb{R}^n , and if $x_1 e_1 + \dots + x_n e_n = 0$, then $x = 0$, so each $x_i = 0$, giving linear independence. We call $\mathcal{E} = e_1, \dots, e_n$ the standard or canonical basis of \mathbb{R}^n .

Lemma 1.2.3. *Let $v_1, \dots, v_k \in V$ be linearly independent. Then any element of their span can be written uniquely as a linear combination of v_1, \dots, v_k , i.e., if $a_1 v_1 + \dots + a_k v_k = b_1 v_1 + \dots + b_k v_k$, then $a_i = b_i$ for all i .*

Proof. If $a_1 v_1 + \dots + a_k v_k = b_1 v_1 + \dots + b_k v_k$, then

$$\begin{aligned} 0 &= (a_1 v_1 + \dots + a_n v_n) - (b_1 v_1 + \dots + b_n v_n) \\ &= (a_1 - b_1) v_1 + \dots + (a_n - b_n) v_n. \end{aligned}$$

By linear independence $a_i - b_i = 0$ for all i , hence $a_i = b_i$. \square

Proposition 1.2.4. *Let $\mathcal{B} = v_1, \dots, v_n$ be a basis for V . Then there is an isomorphism*

$$\Phi_{\mathcal{B}} : \mathbb{R}^n \rightarrow V$$

given by

$$\Phi_{\mathcal{B}} \left(\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \right) = a_1 v_1 + \dots + a_n v_n$$

Thus, if V has a basis with n elements, it is isomorphic to \mathbb{R}^n .

Moreover, $\Phi_{\mathcal{B}}$ satisfies

$$\Phi_{\mathcal{B}}(e_i) = v_i$$

for $i = 1, \dots, n$.

Notation 1.2.5. The inverse function $\Phi_{\mathcal{B}}^{-1}$ has the special notation

$$\Phi_{\mathcal{B}}^{-1}(v) = [v]_{\mathcal{B}}$$

for all $v \in V$. We call $[v]_{\mathcal{B}}$ the \mathcal{B} -coordinates of v . Explicitly, if $\mathcal{B} = v_1, \dots, v_n$

and $v = a_1v_1 + \dots + a_nv_n$, then $[v]_{\mathcal{B}} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$. We emphasize that the \mathcal{B} -

coordinates depend strongly on the basis \mathcal{B} . If we change even one element of the basis, it will change the coordinates significantly.

Proof of Proposition 1.2.4. $\Phi_{\mathcal{B}}$ is linear because

$$\begin{aligned} (a_1v_1 + \dots + a_nv_n) + (b_1v_1 + \dots + b_nv_n) &= (a_1 + b_1)v_1 + \dots + (a_n + b_n)v_n, \\ a(a_1v_1 + \dots + a_nv_n) &= (aa_1)v_1 + \dots + (aa_n)v_n, \end{aligned}$$

by the basic vector identities.

$\Phi_{\mathcal{B}}$ is onto because v_1, \dots, v_n span V . $\Phi_{\mathcal{B}}$ is one-to-one by Lemma 1.2.3: if $a_1v_1 + \dots + a_nv_n = b_1v_1 + \dots + b_nv_n$, Then $a_i = b_i$ for all i .

The last statement is immediate from the definitions of $\Phi_{\mathcal{B}}$ and the canonical basis elements e_i . \square

The following is now immediate from (1.2.1).

Lemma 1.2.6. *Let $\mathcal{E} = e_1, \dots, e_n$ be the canonical basis of \mathbb{R}^n . Then $\Phi_{\mathcal{E}} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the identity map, hence $[x]_{\mathcal{E}} = x$ for all $x \in \mathbb{R}^n$.*

The following result makes use of similar arguments to those in Proposition 1.2.4.

Lemma 1.2.7. *Let V and W be vector spaces and let v_1, \dots, v_n be a basis for V . Let $w_1, \dots, w_n \in W$ be arbitrary. Then there is a unique linear function $f : V \rightarrow W$ with $f(v_i) = w_i$ for all i . It satisfies*

$$(1.2.2) \quad f(a_1v_1 + \dots + a_nv_n) = a_1w_1 + \dots + a_nw_n.$$

In particular:

- (1) *The range of f is $\text{span}(w_1, \dots, w_n)$, so f is onto if and only if w_1, \dots, w_n span W .*
- (2) *f is one-to-one if and only if w_1, \dots, w_n are linearly independent.*
- (3) *f is an isomorphism if and only if w_1, \dots, w_n is a basis for W .*

Proof. If $f : V \rightarrow W$ is linear and $f(v_i) = w_i$ for $i = 1, \dots, n$ then

$$\begin{aligned} f(a_1v_1 + \dots + a_nv_n) &= a_1f(v_1) + \dots + a_nf(v_n) \\ &= a_1w_1 + \dots + a_nw_n \end{aligned}$$

with the first equality following from linearity, so f is uniquely determined by linearity and knowing the values of f on the basis elements.

Conversely, given arbitrary elements w_1, \dots, w_n of W , we use (1.2.2) to define a function $f : V \rightarrow W$. This gives a well-defined function by Proposition 1.2.4: every element of V may be written uniquely as a linear combination of v_1, \dots, v_n . The linearity of f now follows by the properties

defining a vector space. So it remains to verify (1)–(3). Of these, (1) is obvious, while (3) follows from (1) and (2).

For (2), note that if w_1, \dots, w_n are linearly dependent there are scalars a_1, \dots, a_n , not all 0, with $a_1w_1 + \dots + a_nw_n = 0$. So $a_1v_1 + \dots + a_nv_n \in \ker f$. But since a_1, \dots, a_n are not all 0 and since v_1, \dots, v_n are linearly independent, $a_1v_1 + \dots + a_nv_n \neq 0$ and $\ker f \neq 0$. So f is not one-to-one. The converse is clear. \square

Corollary 1.2.8. *Let $\mathcal{B} = v_1, \dots, v_n$ be a basis of V . Then $\Phi_{\mathcal{B}} : \mathbb{R}^n \rightarrow V$ is the unique linear function such that*

$$\Phi_{\mathcal{B}}(e_i) = v_i$$

for $i = 1, \dots, n$. We obtain a one-to-one correspondence between the ordered bases $\mathcal{B} = v_1, \dots, v_n$ of V and the linear isomorphisms $f : \mathbb{R}^n \rightarrow V$ given by

$$\mathcal{B} \mapsto \Phi_{\mathcal{B}}.$$

Lemma 1.2.7 is a key tool in understanding linear maps. We may use it to study the linear maps from \mathbb{R}^n to \mathbb{R}^m and to show they are induced by matrices. We may then use matrix manipulations to study the linear maps from \mathbb{R}^n to \mathbb{R}^m in greater detail. We can then use bases to apply these more detailed results to linear maps between more general vector spaces.

1.3. Matrices. We write $A = (a_{ij})$ for the $m \times n$ matrix

$$(1.3.1) \quad A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ & & \ddots & \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

with mn real entries a_{ij} , $1 \leq i \leq m$, $1 \leq j \leq n$. This matrix induces a linear function $T_A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ via the matrix product $T_A(x) = Ax$. Here we use column vectors for the elements of both \mathbb{R}^n and \mathbb{R}^m , and write

$$(1.3.2) \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

for the generic element of \mathbb{R}^n . Thus, x is an $n \times 1$ column matrix. So we regard Ax as the product of two matrices, and, as usual, if A is as above and if $B = (b_{ij})$ is any $n \times k$ matrix, AB is the $m \times k$ matrix whose ij th coordinate is $\sum_{k=1}^n a_{ik}b_{kj}$.

The matrix product satisfies an important property.

Lemma 1.3.1. *Let $A = (a_{ij})$ be $m \times n$ and let $B = (b_{ij})$ be $n \times k$. Then*

$$T_A \circ T_B = T_{AB} : \mathbb{R}^k \rightarrow \mathbb{R}^m,$$

i.e., matrix multiplication corresponds to composition of functions (in the usual order).

Proof. This follows from the associativity of matrix multiplication: for A and B as above and for a $k \times \ell$ matrix $C = (c_{ij})$, we have

$$(AB)C = A(BC).$$

We refer the reader to [1] or [17] for the basic properties of matrix addition and multiplication. \square

Note that for an $m \times n$ matrix A , the function $T_A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ as defined above is linear because matrix multiplication satisfies a distributive property: for any two $n \times k$ matrices $B = (b_{ij})$ and $C = (c_{ij})$, we have

$$A(B + C) = AB + AC,$$

and also

$$A \cdot aB = aAB$$

for all $a \in \mathbb{R}$. In the above, $B + C$ is the $n \times k$ matrix whose ij th entry is $b_{ij} + c_{ij}$ and aB is the $n \times k$ matrix whose ij th entry is ab_{ij} .

Definition 1.3.2. Let A be an $m \times n$ matrix. The nullspace, $N(A)$, of A is the kernel of T_A :

$$N(A) = \{x \in \mathbb{R}^n : Ax = 0\}.$$

Phrased in terms of matrices, Lemma 1.1.5 becomes:

Lemma 1.3.3. Let A be an $m \times n$ matrix and let $b \in \mathbb{R}^m$. Let x_0 be a solution of $Ax = b$. Then the set of all solutions of $Ax = b$ is

$$(1.3.3) \quad x_0 + N(A) = \{x + v : v \in N(A)\}.$$

Notation 1.3.4. We write $M_{m,n}(\mathbb{R})$ for the set of all $m \times n$ matrices with coefficients in \mathbb{R} . We abbreviate $M_n(\mathbb{R}) = M_{n,n}(\mathbb{R})$ the set of square, $n \times n$ matrices with real coefficients.

Given column vectors $v_1, \dots, v_n \in \mathbb{R}^m$ we write $[v_1 | \dots | v_n]$ for the $m \times n$ matrix whose i th column is v_i . Thus, $A = (a_{ij})$ may be written as $[a_1 | \dots | a_n]$

where $a_i = \begin{bmatrix} a_{1i} \\ \vdots \\ a_{ni} \end{bmatrix}$ for $i = 1, \dots, n$.

Straightforward calculation gives the following.

Lemma 1.3.5. Let $A = [v_1 | \dots | v_n]$ be $m \times n$ and let $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$. Then Ax

may be given explicitly as the linear combination

$$(1.3.4) \quad Ax = x_1v_1 + \dots + x_nv_n,$$

Thus:

- (1) $Ae_i = v_i$, the i th column of A , for $i = 1, \dots, n$.

- (2) T_A is the unique linear function from \mathbb{R}^n to \mathbb{R}^m with $T_A(e_i) = v_i$ for $i = 1, \dots, n$.

Note that the uniqueness in (2) comes from Lemma 1.2.7. From this, we immediately obtain that the linear functions from \mathbb{R}^n to \mathbb{R}^m are in one-to-one correspondence with the $m \times n$ matrices:

Corollary 1.3.6. *Let $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be linear. Then there is a unique $m \times n$ matrix $A = [T]$ such that $T = T_A$:*

$$[T] = [T(e_1) | \dots | T(e_n)].$$

Lemma 1.3.5 allows us to translate Lemma 1.2.7 into a statement about matrices:

Corollary 1.3.7. *Let $A = [v_1 | \dots | v_n]$ be $m \times n$ and let $T_A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be the induced linear transformation.*

- (1) T_A is onto if and only if v_1, \dots, v_n span \mathbb{R}^m .
- (2) T_A is one-to-one if and only if v_1, \dots, v_n are linearly independent.
- (3) T_A is an isomorphism if and only if v_1, \dots, v_n is a basis of \mathbb{R}^m .

We can now apply the technique of Gauss elimination to study linear transformations from \mathbb{R}^n to \mathbb{R}^m . First we review the notion of invertibility of matrices.

Definition 1.3.8. The $n \times n$ matrix A is invertible if there is an $n \times n$ matrix B such that

$$AB = BA = I_n$$

where I_n is the $n \times n$ identity matrix

$$I_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & 1 \end{bmatrix},$$

i.e., the $n \times n$ matrix whose diagonal entries are all 1 and whose off-diagonal entries are all 0.

The identity matrix is the unique matrix with the property that $T_{I_n} = I$, the identity map of \mathbb{R}^n . If A is invertible, there is a unique matrix B with $AB = BA = I_n$ as shown in Lemma 3.1.3(1) below. We write A^{-1} for the inverse matrix B . Invertibility is important for the following reason.

Lemma 1.3.9. *The $n \times n$ matrix A is invertible if and only if the linear mapping T_A is an isomorphism. If A is invertible, then $T_A^{-1} = T_{A^{-1}}$.*

Proof. If A is invertible, then

$$T_{A^{-1}}T_A = T_{A^{-1}A} = T_{I_n} = I.$$

Since I is one-to-one, this forces T_A to be one-to-one. Similarly, $T_AT_{A^{-1}} = I$, and since I is onto, this forces T_A to be onto. Thus, if A is invertible, T_A is

an isomorphism, and the identities $T_{A^{-1}}T_A = T_AT_{A^{-1}} = I$ display $T_{A^{-1}}$ as the inverse function of T_A .

Conversely, if T_A is an isomorphism, the inverse function T_A^{-1} is linear, and hence is equal to T_B for some $n \times n$ matrix B by Corollary 1.3.6. By the calculations just given, the identities $T_{A^{-1}}T_A = T_AT_{A^{-1}} = I$ together with Corollary 1.3.6 force $AB = BA = I_n$. \square

Notation 1.3.10. We write $\text{GL}_n(\mathbb{R})$ for the set of invertible $n \times n$ matrices with coefficients in \mathbb{R} .

1.4. Block matrices and their multiplication. It is valuable in several contexts to subdivide matrices into blocks.

Definition 1.4.1. Let $A = (a_{ij})$ be a $k \times \ell$ matrix, $B = (b_{ij})$ a $k \times r$ matrix, $C = (c_{ij})$ an $s \times \ell$ matrix and $D = (d_{ij})$ and $s \times r$ matrix. We write

$$(1.4.1) \quad M = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right]$$

for the $(k+r) \times (\ell+s)$ matrix whose ij -th entry is given by

$$(1.4.2) \quad m_{ij} = \begin{cases} a_{ij} & \text{if } i \leq k \text{ and } j \leq \ell, \\ b_{i,j-\ell} & \text{if } i \leq k \text{ and } j > \ell, \\ c_{i-k,j} & \text{if } i > k \text{ and } j \leq \ell, \\ d_{i-k,j-\ell} & \text{if } i > k \text{ and } j > \ell. \end{cases}$$

The following is immediate.

Lemma 1.4.2. *Let M be a $(k+r) \times (\ell+s)$ matrix. Then there are unique matrices A , B , C and D with*

$$M = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right].$$

Remark 1.4.3. Indeed, there are other block structures of significant interest. For instance, in (1.4.1) we shall denote the first ℓ columns of M by $\left[\begin{array}{c} A \\ C \end{array} \right]$, and denote the first k rows by $[A|B]$. In fact, we can dice a matrix into an arbitrary grid of blocks, and it is often valuable to do so. An important point is that these grids multiply gridwise. That has important consequences.

Proposition 1.4.4. *Let B_1, \dots, B_n be matrices with k rows and arbitrary numbers of columns. Let A be a matrix with k columns. Then*

$$(1.4.3) \quad A[B_1 | \dots | B_n] = [AB_1 | \dots | AB_n].$$

Let A_1, \dots, A_m be matrices with k columns and arbitrary numbers of rows. Let B be a matrix with k rows. Then

$$(1.4.4) \quad \left[\begin{array}{c} A_1 \\ \vdots \\ A_m \end{array} \right] B = \left[\begin{array}{c} A_1 B \\ \vdots \\ A_m B \end{array} \right].$$

Now let's reverse things, with the block "columns" on the left and the block "rows" on the right: let A_1, \dots, A_n be matrices with r rows and let B_1, \dots, B_n be matrices with s columns. Suppose the number of columns of A_i is equal to the number of rows of B_i for $i = 1, \dots, n$. Then

$$(1.4.5) \quad [A_1 | \dots | A_n] \left[\begin{array}{c} B_1 \\ \vdots \\ B_n \end{array} \right] = A_1 B_1 + \dots + A_n B_n.$$

The result is $r \times s$.

Putting these together in the case of matrices with a 2×2 block structure, we obtain the following: if the number of columns of A and C is equal to the number of rows of X and Y , and the number of columns of B and D is equal to the number of rows of Z and W , then

$$(1.4.6) \quad \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right] \left[\begin{array}{c|c} X & Y \\ \hline Z & W \end{array} \right] = \left[\begin{array}{c|c} AX + BZ & AY + BW \\ \hline CX + DZ & CY + DW \end{array} \right].$$

Proof. The identities (1.4.3) and (1.4.4) are straightforward from the definition of matrix multiplication. For (1.4.5), let v_{ij} be the i -th row of A_j and let w_{jk} be the k -th column of B_j . then the i -th row of $[A_1 | \dots | A_n]$ is obtained by horizontally concatenating v_{i1}, \dots, v_{in} and the k -th column

of $\left[\begin{array}{c} B_1 \\ \vdots \\ B_n \end{array} \right]$ is obtained by vertically concatenating w_{1k}, \dots, w_{nk} . Their matrix product is precisely $\sum_{j=1}^n v_{ij} w_{jk}$, which is precisely the ik -th entry of $A_1 B_1 + \dots + A_n B_n$.

The formula (1.4.6) follows from the others. \square

1.5. Dimension. We can now use Gauss elimination to study matrices. Complete proofs of the following assertions can be found in [1].

Proposition 1.5.1. Let $A = [v_1 | \dots | v_n]$ be an $m \times n$ matrix and let

$$T_A : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

be the induced linear function.

- (1) If T_A is one-to-one, then $n \leq m$, and if $n = m$, then A is invertible.
- (2) If T_A is onto, then $n \geq m$, and if $n = m$, then A is invertible.

Sketch of proof. If T_A is one-to-one, then there are no free variables, so A reduces to a matrix with a pivot in every column. Since there is at most one pivot per row, the number of columns is less than or equal to the number of rows. If there is the same number of rows as columns, then the reduced matrix must be I_n . Any matrix that reduces to I_n is invertible.

If T_A is onto, then A reduces to a matrix with a pivot in every row. We repeat the argument above, reversing the role of rows and columns. \square

Corollary 1.5.2. *Let $\mathcal{B} = v_1, \dots, v_n$ be a basis for the vector space V . Then any other basis also has n elements. Moreover, we have the following:*

- (1) *If $w_1, \dots, w_k \in V$ are linearly independent, then $k \leq n$, and if $k = n$, then w_1, \dots, w_k form a basis for V .*
- (2) *If w_1, \dots, w_k span V , then $k \geq n$, and if $k = n$ then w_1, \dots, w_k form a basis for V .*

Proof. The uniqueness of the number of elements in a basis follows from (1) and (2). The proofs of (1) and (2) follow by applying the isomorphism $\Phi_{\mathcal{B}}^{-1} : V \rightarrow \mathbb{R}^n$. Since $\Phi_{\mathcal{B}}^{-1}$ is an isomorphism, it is easy to see that if $w_1, \dots, w_k \in V$, then w_1, \dots, w_k are linearly independent if and only if $\Phi_{\mathcal{B}}^{-1}(w_1), \dots, \Phi_{\mathcal{B}}^{-1}(w_k)$ are linearly independent, and w_1, \dots, w_k span V if and only if $\Phi_{\mathcal{B}}^{-1}(w_1), \dots, \Phi_{\mathcal{B}}^{-1}(w_k)$ span \mathbb{R}^n .

Thus, we may assume $V = \mathbb{R}^n$ and we may apply Corollary 1.3.7 and Proposition 1.5.1 to $A = [w_1 | \dots | w_k]$. \square

This results in a significant strengthening of Corollary 1.3.7.

Corollary 1.5.3. *\mathbb{R}^n and \mathbb{R}^m are isomorphic if and only if $n = m$. Moreover, if A is an $n \times n$ matrix, then the following conditions are equivalent.*

- (1) *T_A is one-to-one, i.e., the columns of A are linearly independent.*
- (2) *T_A is onto, i.e., the columns of A span \mathbb{R}^n .*
- (3) *T_A is an isomorphism, i.e., the columns of A form a basis of \mathbb{R}^n .*

Proof. If $T_A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is an isomorphism, then the columns of A form a basis of \mathbb{R}^m . Since there are n columns, this forces $m = n$. Now apply Proposition 1.5.1. \square

Since the number of elements in a basis is unique, the following makes sense.

Definition 1.5.4. If the vector space V has a basis with n elements we say V has dimension n and write $\dim V = n$. We adopt the convention that the zero vector space $0 = \{0\}$ has the empty set as its basis, so that $\dim 0 = 0$.

Note that every n -dimensional vector space is isomorphic to \mathbb{R}^n .

The concept of span is useful in studying subspaces. The following lemma is elementary.

Lemma 1.5.5. *Let $v_1, \dots, v_k \in V$. Then $\text{span}(v_1, \dots, v_k)$ is a subspace of V , and any subspace of V containing v_1, \dots, v_k must contain $\text{span}(v_1, \dots, v_k)$. Thus, $\text{span}(v_1, \dots, v_k)$ is the smallest subspace of V containing v_1, \dots, v_k .*

Of course v_1, \dots, v_k are linearly dependent if they are not linearly independent. In other words, v_1, \dots, v_k are linearly dependent if there are real numbers a_1, \dots, a_k , not all 0, such that

$$a_1v_1 + \dots + a_kv_k = 0.$$

A key observation is the following:

Lemma 1.5.6. *Suppose that $a_1v_1 + \dots + a_kv_k = 0$ with $a_i \neq 0$. Then*

$$v_i \in \text{span}(v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_k).$$

Conversely, if $v_i \in \text{span}(v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_k)$, then there are real numbers a_1, \dots, a_k with $a_i \neq 0$ such that $a_1v_1 + \dots + a_kv_k = 0$. In particular, the following properties hold:

- (1) *The vectors $v_1, \dots, v_k \in V$ are linearly dependent if and only if one of the v_i is in the span of the others.*
- (2) *If $v_i \in \text{span}(v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_k)$, then*

$$\text{span}(v_1, \dots, v_k) = \text{span}(v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_k).$$

Thus, v_1, \dots, v_k are linearly dependent if and only if $\text{span}(v_1, \dots, v_k)$ is the span of a proper subset of $\{v_1, \dots, v_k\}$.

- (3) *Let v_1, \dots, v_k be linearly independent and suppose*

$$v_{k+1} \notin \text{span}(v_1, \dots, v_k).$$

Then v_1, \dots, v_{k+1} are linearly independent.

Proof. If $a_1v_1 + \dots + a_kv_k = 0$ with $a_i \neq 0$, then

$$a_iv_i = -a_1v_1 - \dots - a_{i-1}v_{i-1} - a_{i+1}v_{i+1} - \dots - a_kv_k, \quad \text{so}$$

$$v_i = -\frac{a_1}{a_i}v_1 - \dots - \frac{a_{i-1}}{a_i}v_{i-1} - \frac{a_{i+1}}{a_i}v_{i+1} - \dots - \frac{a_k}{a_i}v_k$$

$$\in \text{span}(v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_k).$$

Conversely, if $v_i \in \text{span}(v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_k)$, we have

$$v_i = c_1v_1 + \dots + c_{i-1}v_{i-1} + c_{i+1}v_{i+1} + \dots + c_kv_k$$

for real numbers $c_1, \dots, c_{i-1}, c_{i+1}, \dots, c_k$. hence

$$c_1v_1 + \dots + c_kv_k = 0$$

with $c_i = -1$, establishing the desired dependence relation.

(1) is now immediate, and (2) follows from Lemma 1.5.5. For (3), if $v_{k+1} \notin \text{span}(v_1, \dots, v_k)$ and if

$$a_1v_1 + \dots + a_{k+1}v_{k+1} = 0,$$

then a_{k+1} cannot be nonzero. But linear independence of v_1, \dots, v_k then forces the other coefficients to be 0 as well. \square

Definition 1.5.7. A vector space is finite-dimensional if it is the span of a finite set of vectors. If a vector space is not finite-dimensional we write $\dim V = \infty$.

The following is immediate from Lemma 1.5.6(2).

Corollary 1.5.8. *Let V be a finite dimensional vector space, say*

$$V = \text{span}(v_1, \dots, v_k).$$

then some subset of $\{v_1, \dots, v_k\}$ forms a basis for V . In particular, V has a basis and $\dim V \leq k$.

When V is a subspace of \mathbb{R}^n , one can actually solve for the subset in question using Gauss elimination. Just set $A = [v_1 | \dots | v_k]$ and reduce A to the reduced row echelon matrix $B = [w_1 | \dots | w_k]$. Suppose that $w_{i_1}, \dots, w_{i_\ell}$ are the pivot columns of B . Then $v_{i_1}, \dots, v_{i_\ell}$ can be shown to give a basis for $V = \text{span}(v_1, \dots, v_k)$.

Example 1.5.9. If $f : V \rightarrow W$ is linear with V finite-dimensional, then the image $f(V)$ is finite-dimensional, as if v_1, \dots, v_k is a basis for V , then $f(V) = \text{span}(f(v_1), \dots, f(v_k))$.

We obtain a very useful tool from the results above.

Corollary 1.5.10. *Let V be an n -dimensional vector space. If $v_1, \dots, v_k \in V$ are linearly independent, then they may be extended to a basis v_1, \dots, v_n for V .*

Proof. We argue by induction on $n - k$. If $n - k = 0$, then v_1, \dots, v_k is already a basis for V by Corollary 1.5.2(1). Otherwise, since v_1, \dots, v_k are linearly independent, $\text{span}(v_1, \dots, v_k)$ must be a proper subspace of V , so there exists $v_{k+1} \in V - \text{span}(v_1, \dots, v_k)$. But then v_1, \dots, v_{k+1} are linearly independent by Lemma 1.5.6(3). This is also the inductive step, and the result follows. \square

We also obtain the following.

Corollary 1.5.11. *Let W be a subspace of the finite-dimensional vector space V . Then W is finite-dimensional, with $\dim W \leq \dim V$. If $\dim W = \dim V$ then $W = V$.*

Proof. We construct a basis for W using the inductive procedure given in the proof of Corollary 1.5.10. We start with a nonzero element $w_1 \in W$ and continue until we have a basis w_1, \dots, w_k of W . This must eventually occur, as there can be at most $\dim V$ elements in any linearly independent subset of V by Corollary 1.5.2(1).

Suppose then that we have obtained a basis w_1, \dots, w_k of W . If $k = \dim V$ then w_1, \dots, w_k is also a basis of V by Corollary 1.5.2(1), and hence $W = V$. \square

1.6. Rank. We can now apply dimension to the study of linear functions.

Definition 1.6.1. Let $f : V \rightarrow W$ be linear with V finite-dimensional. Then the rank, $\text{rank } f$, of f is the dimension of the image, $f(V)$, of f . If A is an $m \times n$ matrix, then the rank, $\text{rank } A$, of A is the rank of $T_A : \mathbb{R}^n \rightarrow \mathbb{R}^m$.

Proposition 1.6.2. *Let $f : V \rightarrow W$ be linear with V finite-dimensional. Let w_1, \dots, w_r be a basis for the image of f and let $v_1, \dots, v_r \in V$ with $f(v_i) = w_i$ for $i = 1, \dots, r$. Let y_1, \dots, y_m be a basis for $\ker f$. Then*

$$(1.6.1) \quad \mathcal{B} = v_1, \dots, v_r, y_1, \dots, y_m$$

is a basis for V . Thus,

$$(1.6.2) \quad \text{rank } f + \dim \ker f = \dim V.$$

Proof. We first show $v_1, \dots, v_r, y_1, \dots, y_m$ are linearly independent. Suppose $c_1v_1 + \dots + c_rv_r + d_1y_1 + \dots + d_my_m = 0$. Then

$$\begin{aligned} 0 &= c_1f(v_1) + \dots + c_rf(v_r) + d_1f(y_1) + \dots + d_mf(y_m) \\ &= c_1f(v_1) + \dots + c_rf(v_r) \\ &= c_1w_1 + \dots + c_rw_r, \end{aligned}$$

as $f(y_1) = \dots = f(y_m) = 0$. But w_1, \dots, w_r are linearly independent, so $c_1 = \dots = c_r = 0$. This leaves

$$d_1y_1 + \dots + d_my_m = 0.$$

But y_1, \dots, y_m are linearly independent, so $d_1 = \dots = d_m = 0$.

Now we show $v_1, \dots, v_r, y_1, \dots, y_m$ span V . Let $v \in V$. Then $f(v) \in f(V) = \text{span}(w_1, \dots, w_r)$. Say $f(v) = c_1w_1 + \dots + c_rw_r$. Then,

$$f(v - (c_1v_1 + \dots + c_rv_r)) = f(v) - (c_1w_1 + \dots + c_rw_r) = 0,$$

so $v - (c_1v_1 + \dots + c_rv_r) \in \ker f = \text{span}(y_1, \dots, y_m)$. Say

$$v - (c_1v_1 + \dots + c_rv_r) = d_1y_1 + \dots + d_my_m.$$

But then $v = c_1v_1 + \dots + c_rv_r + d_1y_1 + \dots + d_my_m$. □

The following lemma is easy but valuable.

Lemma 1.6.3. *Let V be finite dimensional and let $f : V \rightarrow W$ be linear. Let $g : W \xrightarrow{\cong} W'$ and $h : V' \xrightarrow{\cong} V$ be isomorphisms. Then*

$$\text{rank}(g \circ f \circ h) = \text{rank } f.$$

Proof. Since h is onto, $f(V) = f \circ h(V')$. So $g \circ f \circ h(V')$ is the image of $f(V)$ under the isomorphism g . So the dimension of the image $g \circ f \circ h$ equals the dimension of the image of f . □

1.7. Direct sums. There is a useful operation on vector spaces that reflects the way \mathbb{R}^{m+n} is obtained from \mathbb{R}^m and \mathbb{R}^n .

Definition 1.7.1. Let V and W be vector spaces. The direct sum $V \oplus W$ is the set $V \times W$ endowed with the vector operations

$$\begin{aligned} (v_1, w_1) + (v_2, w_2) &= (v_1 + v_2, w_1 + w_2) \\ c(v_1, w_1) &= (cv_1, cw_1) \end{aligned}$$

for all $v_1, v_2 \in V$, $w_1, w_2 \in W$ and $c \in \mathbb{R}$.

This is just the expected structure, and there is an obvious isomorphism

$$\mathbb{R}^m \oplus \mathbb{R}^n \rightarrow \mathbb{R}^{m+n}$$

$$\left(\begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}, \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \right) \mapsto \begin{pmatrix} x_1 \\ \vdots \\ x_m \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

It is useful to view this as an external operation on vector spaces. The reader may verify the following propositions.

Proposition 1.7.2. *Let V and W be finite-dimensional with bases v_1, \dots, v_m and w_1, \dots, w_n , respectively. Then*

$$(v_1, 0), \dots, (v_m, 0), (0, w_1), \dots, (0, w_n)$$

is a basis for $V \oplus W$, hence $\dim(V \oplus W) = \dim V + \dim W$.

Proposition 1.7.3. *Let V, W be vector spaces. Then the maps*

$$\begin{aligned} \iota_1 : V &\rightarrow V \oplus W, \\ \iota_2 : W &\rightarrow V \oplus W \end{aligned}$$

given by $\iota_1(v) = (v, 0)$, $\iota_2(w) = (0, w)$ are linear and if Z is a vector space and $f : V \rightarrow Z$ and $g : W \rightarrow Z$ are linear, there is a unique linear map $h : V \oplus W \rightarrow Z$ such that the following diagram commutes:

$$\begin{array}{ccccc} V & \xrightarrow{\iota_1} & V \oplus W & \xleftarrow{\iota_2} & W \\ & \searrow f & \downarrow h & \swarrow g & \\ & & Z & & \end{array}$$

Specifically, $h(v, w) = f(v) + g(w)$.

Proposition 1.7.4. *Let V, W be vector spaces. Then the maps*

$$\begin{aligned} \pi_1 : V \oplus W &\rightarrow V, \\ \pi_2 : V \oplus W &\rightarrow W \end{aligned}$$

given by $\pi_1(v, w) = v$, $\pi_2(v, w) = w$ are linear and if Z is a vector space and $f : Z \rightarrow V$ and $g : Z \rightarrow W$ are linear, there is a unique linear map $h = (f, g) : Z \rightarrow V \oplus W$ such that the following diagram commutes:

$$\begin{array}{ccccc} & & Z & & \\ & \swarrow f & \downarrow h & \searrow g & \\ V & \xleftarrow{\pi_1} & V \oplus W & \xrightarrow{\pi_2} & W \end{array}$$

Specifically, $h(z) = (f(z), g(z))$. In other words, f and g are the coordinate functions of h in terms of the ordered pairs in $V \oplus W$.

1.8. Base change. We review the relationship between bases and matrices. Since we are using matrices, we will write vectors in \mathbb{R}^n as column vectors as in (1.3.2). Thus, if $\mathcal{B} = v_1, \dots, v_n$ is a basis of V , then the isomorphism

$$\Phi_{\mathcal{B}} : \mathbb{R}^n \xrightarrow{\cong} V$$

is given by

$$\Phi_{\mathcal{B}} \left(\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \right) = x_1 v_1 + \cdots + x_n v_n.$$

Recall from Notation 1.2.5, that the inverse isomorphism $\Phi_{\mathcal{B}}^{-1} : V \rightarrow \mathbb{R}^n$ has a special notation: if $v = x_1 v_1 + \cdots + x_n v_n \in V$, we write

$$(1.8.1) \quad [v]_{\mathcal{B}} = \Phi_{\mathcal{B}}^{-1}(v) = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = x_1 e_1 + \cdots + x_n e_n.$$

We call $[v]_{\mathcal{B}}$ the \mathcal{B} -coordinates of v .

Now suppose given a linear function $f : V \rightarrow W$ and bases $\mathcal{B} = v_1, \dots, v_n$ of V and $\mathcal{B}' = w_1, \dots, w_m$ of W . Write $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ for the composite

$$\mathbb{R}^n \xrightarrow{\Phi_{\mathcal{B}}} V \xrightarrow{f} W \xrightarrow{\Phi_{\mathcal{B}'}} \mathbb{R}^m.$$

Definition 1.8.1. With the notations above, the matrix $[f]_{\mathcal{B}'\mathcal{B}}$ of f with respect to the bases \mathcal{B}' , \mathcal{B} is the matrix of T as given by Corollary 1.3.6:

$$\begin{aligned} [f]_{\mathcal{B}'\mathcal{B}} &= [T] \\ &= [T(e_1) | \cdots | T(e_n)] \\ &= [\Phi_{\mathcal{B}'}^{-1} f \Phi_{\mathcal{B}}(e_1) | \cdots | \Phi_{\mathcal{B}'}^{-1} f \Phi_{\mathcal{B}}(e_n)] \\ &= [\Phi_{\mathcal{B}'}^{-1} f(v_1) | \cdots | \Phi_{\mathcal{B}'}^{-1} f(v_n)] && \text{(Proposition 1.2.4)} \\ &= [[f(v_1)]_{\mathcal{B}'} | \cdots | [f(v_n)]_{\mathcal{B}'}] && (1.8.1). \end{aligned}$$

This generalizes what we have done for linear functions from \mathbb{R}^n to \mathbb{R}^m :

Lemma 1.8.2. Let $\mathcal{E} = e_1, \dots, e_n$ and $\mathcal{E}' = e_1, \dots, e_m$ be the canonical bases of \mathbb{R}^n and \mathbb{R}^m , respectively. Let $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be linear. Then

$$[T]_{\mathcal{E}'\mathcal{E}} = [T] = [T(e_1) | \cdots | T(e_n)],$$

the matrix of T as given in Corollary 1.3.6.

Proof. By Lemma 1.2.6, $\Phi_{\mathcal{E}}$ and $\Phi_{\mathcal{E}'}$ are the identity maps of \mathbb{R}^n and \mathbb{R}^m , respectively. \square

There is a nice relationship between $[f]_{\mathcal{B}'\mathcal{B}}$ and the coordinate functions given by the bases \mathcal{B} and \mathcal{B}' via (1.8.1).

Lemma 1.8.3. *With the notations above,*

$$(1.8.2) \quad [f]_{\mathcal{B}'\mathcal{B}}[v]_{\mathcal{B}} = [f(v)]_{\mathcal{B}'}$$

for all $v \in V$. Here, the left-hand side is the product of the matrix $[f]_{\mathcal{B}'\mathcal{B}}$ with the \mathcal{B} -coordinates of v .

Proof. Let $v = a_1v_1 + \cdots + a_nv_n$. Then $[v]_{\mathcal{B}} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$, so

$$\begin{aligned} [f]_{\mathcal{B}'\mathcal{B}}[v]_{\mathcal{B}} &= [[f(v_1)]_{\mathcal{B}'} : \cdots : [f(v_n)]_{\mathcal{B}'}] \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \\ &= a_1[f(v_1)]_{\mathcal{B}'} + \cdots + a_n[f(v_n)]_{\mathcal{B}'} \\ &= [a_1f(v_1) + \cdots + a_nf(v_n)]_{\mathcal{B}'}, \end{aligned}$$

as $w \mapsto [w]_{\mathcal{B}'}$ is linear (as the inverse to $\Phi_{\mathcal{B}'}$). But linearity of f gives

$$a_1f(v_1) + \cdots + a_nf(v_n) = f(a_1v_1 + \cdots + a_nv_n) = f(v),$$

and the result follows. \square

Note that all three of the terms in (1.8.2) depend strongly on the bases chosen. In fact, we will see that the matrices $[I]_{\mathcal{B}'\mathcal{B}}$ (with I the identity map of a vector space V) play a very important role in understanding the relationship between two bases and how it affects coordinatizing matrices. The following is very important in developing that investigation.

Proposition 1.8.4. *Let $f : V \rightarrow W$ and $g : W \rightarrow Z$ be linear. Let $\mathcal{B} = v_1, \dots, v_n$ be a basis of V , let $\mathcal{B}' = w_1, \dots, w_m$ be a basis of W and let $\mathcal{B}'' = z_1, \dots, z_k$ be a basis of Z . Then*

$$[g]_{\mathcal{B}''\mathcal{B}'}[f]_{\mathcal{B}'\mathcal{B}} = [g \circ f]_{\mathcal{B}''\mathcal{B}},$$

where the left-hand side is the matrix product.

Proof. It suffices to show that both sides have the same effect on an arbitrary vector

$\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$. But this is easy, as, if $v = a_1v_1 + \cdots + a_nv_n$, then

$$\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = [v]_{\mathcal{B}}. \text{ So}$$

$$\begin{aligned} [g]_{\mathcal{B}''\mathcal{B}'}[f]_{\mathcal{B}'\mathcal{B}} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} &= [g]_{\mathcal{B}''\mathcal{B}'}[f]_{\mathcal{B}'\mathcal{B}}[v]_{\mathcal{B}} \\ &= [g]_{\mathcal{B}''\mathcal{B}'}[f(v)]_{\mathcal{B}'} \end{aligned}$$

$$\begin{aligned}
&= [g \circ f(v)]_{\mathcal{B}'} \\
&= [g \circ f]_{\mathcal{B}'\mathcal{B}} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}.
\end{aligned}$$

Here, we have used Lemma 1.8.3 twice. \square

Proposition 1.8.4 is very flexible and has many important consequences. Of particular importance is the analysis of linear functions $f : V \rightarrow V$ where we use the same basis for both copies of V .

Notation 1.8.5. Let $f : V \rightarrow V$ be linear and let $\mathcal{B} = v_1, \dots, v_n$ be a basis of V . We write $[f]_{\mathcal{B}}$ for the matrix $[f]_{\mathcal{B}\mathcal{B}}$:

$$[f]_{\mathcal{B}} = [[f(v_1)]_{\mathcal{B}} | \dots | [f(v_n)]_{\mathcal{B}}].$$

Using the same basis for the domain and codomain gives a consistent coordinatization of f , and we can ask what geometric effect f has with respect to these coordinates. This is of value even when $V = \mathbb{R}^n$ as we may think of the basis \mathcal{B} as providing a linear change of variables. We have seen the value of changing variables in calculus, and it also has significant value in both linear algebra and geometry. In geometry, we shall be most interested in the change of variables given by an orthonormal basis of \mathbb{R}^n . These are the bases that will arise in studying linear isometries. For more general linear functions, arbitrary bases become important for base change.

The following is immediate from Proposition 1.8.4.

Corollary 1.8.6. Let $f, g : V \rightarrow V$ be linear and let $\mathcal{B} = v_1, \dots, v_n$ be a basis of V . Then

$$[g \circ f]_{\mathcal{B}} = [g]_{\mathcal{B}}[f]_{\mathcal{B}}.$$

Proposition 1.8.4 also shows how to convert between different coordinatizations of f .

Definition 1.8.7. Let $\mathcal{B} = v_1, \dots, v_n$ and $\mathcal{B}' = w_1, \dots, w_n$ be two different bases of the vector space V . The transition matrix from \mathcal{B} to \mathcal{B}' is

$$[I]_{\mathcal{B}'\mathcal{B}} = [[v_1]_{\mathcal{B}'} | \dots | [v_n]_{\mathcal{B}'}],$$

where I is the identity map of V .

A useful example is the following, whose proof is immediate from Lemma 1.2.6.

Lemma 1.8.8. Let $\mathcal{B} = v_1, \dots, v_n$ be a basis of \mathbb{R}^n . Then the transition matrix from \mathcal{B} to the standard basis is given by

$$[I]_{\mathcal{E}\mathcal{B}} = [v_1 | \dots | v_n].$$

Lemma 1.8.9. *Let $\mathcal{B} = v_1, \dots, v_n$ and $\mathcal{B}' = w_1, \dots, w_n$ be two different bases of the vector space V . Then the transition matrices between them in opposite directions are inverse to one another:*

$$[I]_{\mathcal{B}'\mathcal{B}} = [I]_{\mathcal{B}\mathcal{B}'}^{-1}.$$

Proof.

$$[I]_{\mathcal{B}'\mathcal{B}}[I]_{\mathcal{B}\mathcal{B}'} = [I \circ I]_{\mathcal{B}'\mathcal{B}'} = [[w_1]_{\mathcal{B}'} | \dots | [w_n]_{\mathcal{B}'}] = [e_1 | \dots | e_n] = I_n.$$

A similar calculation gives $[I]_{\mathcal{B}\mathcal{B}'}[I]_{\mathcal{B}'\mathcal{B}} = I_n$. \square

The following is now immediate from Proposition 1.8.4 and Lemma 1.8.9.

Corollary 1.8.10. *Let $\mathcal{B} = v_1, \dots, v_n$ and $\mathcal{B}' = w_1, \dots, w_n$ be two different bases of the vector space V . Let $f : V \rightarrow V$ be linear. Then*

$$[f]_{\mathcal{B}'} = P[f]_{\mathcal{B}}P^{-1}$$

where $P = [I]_{\mathcal{B}'\mathcal{B}}$.

Corollary 1.8.11. *Let \mathcal{B} and \mathcal{B}' be bases for the finite-dimensional vector space V and let $f : V \rightarrow V$ be linear. Then*

$$\det[f]_{\mathcal{B}} = \det[f]_{\mathcal{B}'}.$$

Proof.

$$\begin{aligned} \det(P[f]_{\mathcal{B}}P^{-1}) &= \det(P) \det([f]_{\mathcal{B}}) \det(P^{-1}) \\ &= \det P \det[f]_{\mathcal{B}} (\det P)^{-1} \\ &= \det[f]_{\mathcal{B}}, \end{aligned}$$

as real numbers commute. \square

Thus, the following is well-defined.

Definition 1.8.12. Let $f : V \rightarrow V$ be linear with V a finite-dimensional vector space. We define $\det f$ by

$$\det f = \det[f]_{\mathcal{B}}$$

for any basis \mathcal{B} of V .

Proposition 1.8.13. *Let $f, g : V \rightarrow V$ be linear with V a finite-dimensional vector space. Then:*

- (1) $\det(f \circ g) = \det f \det g$.
- (2) $\det I = 1$.
- (3) $\det(af) = a^{\dim V} \det f$ for $a \in \mathbb{R}$.
- (4) f is invertible if and only if $\det f \neq 0$.

Proof. These follow from the basic properties of determinants of matrices. Here we use that if $\dim V = n$, then $[aI]_{\mathcal{B}} = aI_n$ for any basis \mathcal{B} and any $a \in \mathbb{R}$, and $af = (aI) \circ f$. \square

The following is more delicate, as it requires the development of determinant theory for matrices with coefficients in a commutative ring (see [17]). Recall that the characteristic polynomial $\text{ch}_A(x)$ of an $n \times n$ matrix is given by

$$\text{ch}_A(x) = \det(xI_n - A).$$

Here, $xI_n - A$ is an $n \times n$ matrix with coefficients in the commutative ring $\mathbb{R}[x]$ of polynomials (with variable x) with coefficients in \mathbb{R} . So xI_n is the matrix whose diagonal entries are all x and whose off-diagonal entries are all 0, and A is considered to be a matrix of constant polynomials.

Characteristic polynomials are important, as their roots are real numbers c such that $\det(cI_n - A) = 0$, meaning that the real $n \times n$ matrix $cI_n - A$ is not invertible. But that in turn means there is a nonzero vector $v \in \mathbb{R}^n$ with $(cI_n - A)v = 0$, i.e., $cv = Av$. This says c is an eigenvalue of A and that v is an eigenvector for (A, c) (Definition 4.1.22). We obtain:

Lemma 1.8.14. *The eigenvalues of A are the roots of the characteristic polynomial $\text{ch}_A(x)$.*

We bring this in now to extend it to linear functions on arbitrary finite-dimensional vector spaces.

Lemma 1.8.15. *Let $f : V \rightarrow V$ be linear with V finite-dimensional. Let \mathcal{B} and \mathcal{B}' be bases for V . Then the matrices $[f]_{\mathcal{B}}$ and $[f]_{\mathcal{B}'}$ have the same characteristic polynomial.*

Proof.

$$(P[f]_{\mathcal{B}}P^{-1}) - I_n = P([f]_{\mathcal{B}} - I_n)P^{-1}.$$

Now use the fact that determinants of matrices with coefficients in a commutative ring are product preserving (see [17]). \square

So the following is well-defined.

Definition 1.8.16. Let $f : V \rightarrow V$ be linear with V finite-dimensional. Then the characteristic polynomial $\text{ch}_f(x)$ is defined to be $\text{ch}_{[f]_{\mathcal{B}}}(x)$ for any basis \mathcal{B} of V .

As above, we define an eigenvalue c of f to be a real number for which there exists a nonzero $v \in V$ with $f(v) = cv$. For an eigenvalue c of f , the eigenspace of (f, c) is the subspace consisting of all vectors $v \in V$ with $f(v) = cv$. Such vectors are called eigenvectors of (f, c) .

Proposition 1.8.17. *Let $f : V \rightarrow V$ be linear with V finite-dimensional. Let \mathcal{B} be a basis of V and let $c \in \mathbb{R}$. Then*

$$f(v) = cv \quad \Leftrightarrow \quad [f]_{\mathcal{B}}[v]_{\mathcal{B}} = c[v]_{\mathcal{B}}.$$

Thus, the eigenvalues of f are the eigenvalues of $[f]_{\mathcal{B}}$, which, in turn are the roots of $\text{ch}_f(x)$. Moreover, $\Phi_{\mathcal{B}}$ maps the eigenspace of $([f]_{\mathcal{B}}, c)$ isomorphically onto the eigenspace of (f, c) .

Proof. $[f(v)]_{\mathcal{B}} = [f]_{\mathcal{B}}[v]_{\mathcal{B}}$. \square

A very important topic in base change is invariant subspaces. Recall that if $f : V \rightarrow V$ is linear, then a subspace W is f -invariant, or an invariant subspace of f , if $f(W) = W$.

Lemma 1.8.18. *Let $\mathcal{B} = v_1, \dots, v_n$ be a basis of V and let $f : V \rightarrow V$ be linear. Let $W = \text{span}(v_1, \dots, v_k)$. Then W is f -invariant if and only if $[f]_{\mathcal{B}}$ has the following form:*

$$(1.8.3) \quad [f]_{\mathcal{B}} = \left[\begin{array}{c|c} A & X \\ \hline 0 & B \end{array} \right],$$

where A is $k \times k$, X is $k \times (n - k)$, 0 is the $(n - k) \times k$ zero-matrix and B is $(n - k) \times (n - k)$. This simply means that the ij th entry of $[f]_{\mathcal{B}}$ is 0 if $i > k$ and $j \leq k$, and we identify the remaining entries in matrix blocks.

In this case, $A = [f|_W]_{\mathcal{B}'}$ with $\mathcal{B}' = v_1, \dots, v_k$ and $f|_W : W \rightarrow W$ is the restriction of f to have both domain and codomain W .

Proof. $[f]_{\mathcal{B}}$ has the stated form if and only if $f(v_j) \in \text{span}(v_1, \dots, v_k)$ whenever $j \leq k$. \square

Recall the expansion of the determinant by the j th column. See [17, Corollary 10.2.12] for a proof.

Lemma 1.8.19. *Let $A = (a_{ij})$ be $n \times n$ and let $j \in \{1, \dots, n\}$. Then*

$$\det A = \sum_{i=1}^n (-1)^{i+j} a_{ij} \det A_{ij}$$

where A_{ij} is the $(n - 1) \times (n - 1)$ matrix obtained by deleting the i th row and j th column of A .

Corollary 1.8.20. *Let $C = \left[\begin{array}{c|c} A & X \\ \hline 0 & B \end{array} \right]$ as in (1.8.3). Then*

$$\det C = \det A \det B.$$

Proof. We argue by induction on k and take the decomposition with respect to the first column. Then the only nonzero entries c_{i1} are for $i \leq k$. Moreover, when $i \leq k$, we have $c_{i1} = a_{i1}$. Thus,

$$\det C = \sum_{i=1}^k (-1)^{i+1} a_{i1} \det C_{i1}.$$

When $k = 1$, this is just $a_{11} \det B = \det A \det B$ and we are done. For $k > 1$ we get

$$\det C = \sum_{i=1}^k (-1)^{i+1} a_{i1} \det \left[\begin{array}{c|c} A_{i1} & X_i \\ \hline 0 & B \end{array} \right],$$

where X_i is obtained from X by deleting the i th row. Now A_{i1} is $\ell \times \ell$ for $\ell = k - 1$, and inductively we may assume

$$\det \left[\begin{array}{c|c} A_{i1} & X_i \\ \hline 0 & B \end{array} \right] = \det A_{i1} \det B.$$

We obtain that

$$\det C = \left(\sum_{i=1}^k (-1)^{i+1} a_{i1} \det A_{i1} \right) \det B = \det A \det B,$$

by the expansion of $\det A$ by its first column. \square

In line with our discussion of conjugacy in a group, we make the following definition.

Definition 1.8.21. The $n \times n$ matrices A and B are conjugate if there is an invertible $n \times n$ matrix P with $B = PAP^{-1}$.

In particular, the matrices of $f : V \rightarrow V$ with respect to two different bases are conjugate by Corollary 1.8.10. We also have a converse:

Corollary 1.8.22. Let $f : V \rightarrow V$ be linear and let $\mathcal{B} = v_1, \dots, v_n$ be a basis of V . Let $A = [f]_{\mathcal{B}}$ and let B be an $n \times n$ matrix conjugate to A . Then there is a basis \mathcal{B}' of V with $B = [f]_{\mathcal{B}'}$.

Proof. It suffices to show that $P = [I]_{\mathcal{B}'\mathcal{B}}$ for a basis \mathcal{B}' of V . By Corollary 1.3.7, the columns of P form a basis $\mathcal{B}'' = z_1, \dots, z_n$ of \mathbb{R}^n . Let $g : \mathbb{R}^n \rightarrow V$ be the composite

$$\mathbb{R}^n \xrightarrow{T_{P^{-1}}} \mathbb{R}^n \xrightarrow{\Phi_{\mathcal{B}}} V.$$

$\overset{g}{\curvearrowright}$

Then, g is an isomorphism, so $\mathcal{B}' = g(e_1), \dots, g(e_n)$ is a basis of V . In particular, then $g = \Phi_{\mathcal{B}'}$. So $\Phi_{\mathcal{B}'}^{-1} = T_{P^{-1}}^{-1} \Phi_{\mathcal{B}}^{-1} = T_P \Phi_{\mathcal{B}}^{-1}$, hence

$$\begin{aligned} [v_i]_{\mathcal{B}'} &= \Phi_{\mathcal{B}'}^{-1}(v_i) \\ &= T_P(\Phi_{\mathcal{B}}^{-1}(v_i)) \\ &= T_P(e_i) \\ &= z_i \end{aligned}$$

\square

1.9. Exercises.

1. Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and let $B = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.
 - (a) Show that $AB = BA = \begin{bmatrix} ad - bc & 0 \\ 0 & ad - bc \end{bmatrix}$.
 - (b) Deduce that A is invertible if and only if $ad - bc \neq 0$ and that if $ad - bc \neq 0$, then $A^{-1} = \frac{1}{ad - bc} B$.
2. Prove Proposition 1.7.2.

3. Prove Proposition [1.7.3](#).
4. Prove Proposition [1.7.4](#).

2. Basic Euclidean geometry

2.1. Lines in \mathbb{R}^n .

Definition 2.1.1. A line through the origin in \mathbb{R}^n is a one-dimensional subspace:

$$\ell = \text{span}(v) = \{tv : t \in \mathbb{R}\}$$

for some nonzero vector $v \in \mathbb{R}^n$. (Note that a singleton v is linearly independent if and only if $v \neq 0$, as then $tv = 0 \Rightarrow t = 0$.) In general, a line in \mathbb{R}^n has the form

$$\ell = x + \text{span}(v) = \{x + tv : t \in \mathbb{R}\}$$

for $x, v \in \mathbb{R}^n$ with $v \neq 0$.

Of course, $y = x + tv \Leftrightarrow y - x = tv \in \text{span}(v)$. Thus:

Lemma 2.1.2. $x + \text{span}(v) = \{y \in \mathbb{R}^n : y - x \in \text{span}(v)\}$.

We may think of $x + \text{span}(v)$ as the translation of $\text{span}(v)$ by x :

Definition 2.1.3. Let $x \in \mathbb{R}^n$. Then the translation by x , $\tau_x : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is given by

$$\tau_x(y) = x + y$$

for all $y \in \mathbb{R}^n$.

So translation by x is just vector addition with x . This is our first example of an isometry of \mathbb{R}^n (Lemma 2.4.2, below). It enters the picture here, as

$$x + \text{span}(v) = \tau_x(\text{span}(v)).$$

Some basic properties of translations are:

Lemma 2.1.4.

- (1) $\tau_x \circ \tau_y = \tau_{x+y}$ for all $x, y \in \mathbb{R}^n$. Here, \circ denotes composition of functions. Thus, $\tau_x \circ \tau_y = \tau_y \circ \tau_x$.
- (2) $\tau_0 = \text{id}$, where id is the identity function and 0 is the origin.
- (3) τ_{-x} is the inverse function for τ_x , i.e., $\tau_x \circ \tau_{-x} = \tau_{-x} \circ \tau_x = \text{id}$. In particular, τ_x is one-to-one and onto.

In any case, any line is a translation of a line through the origin. But what if we translate $\text{span}(v)$ to a different point in $x + \text{span}(v)$? Do we get the same line?

Lemma 2.1.5. Let $y \in x + \text{span}(v)$. Then $x + \text{span}(v) = y + \text{span}(v)$.

Proof. Write $y = x + cv$. Then $y + tv = x + (t + c)v$, giving

$$y + \text{span}(v) \subset x + \text{span}(v),$$

while $x + tv = y + (t - c)v$, giving

$$x + \text{span}(v) \subset y + \text{span}(v). \quad \square$$

We next characterize lines through the origin. This is basic linear algebra.

Lemma 2.1.6. *Let v and w be nonzero elements of \mathbb{R}^n . Then the following conditions are equivalent.*

- (1) $\text{span}(v) = \text{span}(w)$.
- (2) $w = tv$ for some $t \neq 0$.
- (3) v, w is a linearly dependent set.

Proof. Since v and w are nonzero, (1) \Leftrightarrow (2) is immediate. If $w = tv$ with $t \neq 0$, then $tv - w = 0$, so (2) \Rightarrow (3). If $av + bw = 0$ with a, b not both 0, then $b \neq 0$, as $v \neq 0$. So $w = -\frac{a}{b}v$, so (3) \Rightarrow (2). \square

We obtain the following.

Proposition 2.1.7. *Let v and w be nonzero. Then*

$$(2.1.1) \quad x + \text{span}(v) = y + \text{span}(w) \quad \Leftrightarrow \quad y - x \in \text{span}(v) = \text{span}(w).$$

In particular, if $x \neq y$, this gives $\text{span}(y - x) = \text{span}(v) = \text{span}(w)$.

Proof. Clearly, we may assume $x \neq y$.

Suppose $x + \text{span}(v) = y + \text{span}(w)$. Since $y \in y + \text{span}(w)$, $y - x \in \text{span}(v)$ by Lemma 2.1.2. By similar reasoning, $x - y \in \text{span}(w)$. Since $x \neq y$, this forces $\text{span}(y - x) = \text{span}(v) = \text{span}(w)$ by Lemma 2.1.6.

Conversely, suppose $y - x \in \text{span}(v) = \text{span}(w)$. Since $\text{span}(v) = \text{span}(w)$, we may, without loss of generality, assume $v = w$ (recall our discussion of translation). $y - x \in \text{span}(v)$, so $y \in x + \text{span}(v)$ by Lemma 2.1.2. Now apply Lemma 2.1.5. \square

We now recover one of Euclid's axioms:

Corollary 2.1.8. *Let $x \neq y \in \mathbb{R}^n$. Then there is a unique line containing both points:*

$$\overleftrightarrow{xy} = x + \text{span}(y - x) = \{x + t(y - x) : t \in \mathbb{R}\} = \{(1 - t)x + ty : t \in \mathbb{R}\}.$$

Proof. Since $x \neq y$, $x + \text{span}(y - x)$ is a line, and it certainly contains both x and y . Conversely, any line containing x must have the form $x + \text{span}(v)$ for some v , and similarly for y . Now apply Proposition 2.1.7 to see $\text{span}(v) = \text{span}(y - x)$. \square

We wish now to prove Euclid's parallel postulate. But a word of caution is in order first: there are two different notions of being parallel, which coincide in \mathbb{R}^2 but are different in \mathbb{R}^n for $n > 2$.

Definition 2.1.9.

- (1) The lines $\ell = x + \text{span}(v)$ and $m = y + \text{span}(w)$ in \mathbb{R}^n are parallel, written $\ell \parallel m$, if $\text{span}(v) = \text{span}(w)$ (i.e., if they are translates of each other).
- (2) Lines ℓ and m in \mathbb{R}^2 are two-dimensionally parallel if either $\ell = m$ or $\ell \cap m = \emptyset$.

Note that we allow a line to be parallel to itself. Two-dimensional parallelism is the notion used by Euclid and also in classical non-Euclidean geometry, which is realized by hyperbolic geometry, to be studied below. But it is the wrong notion to use in \mathbb{R}^3 : the lines $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \text{span} \left(\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right)$ and $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \text{span} \left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right)$ do not intersect, but one is a translate of the z -axis and the other of the x -axis. (Lines in \mathbb{R}^n that are not parallel but do not intersect are called skew lines.)

The definition of parallel makes the parallel postulate immediate.

Theorem 2.1.10 (Parallel postulate). *Let $\ell = x + \text{span}(v)$ be a line in \mathbb{R}^n and let $y \in \mathbb{R}^n$. Then there is a unique line through y parallel to ℓ : the line $y + \text{span}(v)$.*

Proof. Lemma 2.1.5. □

But what is really meant by the parallel postulate in Euclidean geometry is that there is a unique line through y two-dimensionally parallel to ℓ . So it's important that the two notions of parallelism coincide in \mathbb{R}^2 :

Proposition 2.1.11. *Lines $\ell = x + \text{span}(v)$ and $m = y + \text{span}(w)$ in \mathbb{R}^2 are two-dimensionally parallel if and only if they are parallel. Nonparallel lines in \mathbb{R}^2 intersect in exactly one point.*

Proof. If $\text{span}(v) = \text{span}(w)$ and ℓ and m have a point of intersection, then $\ell = m$ by Lemma 2.1.5.

If ℓ and m have two points of intersection, they must be equal by Corollary 2.1.8. Thus, it suffices to show that if $\text{span}(v) \neq \text{span}(w)$, then $\ell \cap m$ is nonempty. Thus, suppose $\text{span}(v) \neq \text{span}(w)$. We wish to find $z \in \ell \cap m$, i.e., $z = x + sv = y + tw$, i.e.,

$$(2.1.2) \quad sv + (-t)w = y - x.$$

Because $\text{span}(v) \neq \text{span}(w)$, v, w are linearly independent by Lemma 2.1.6. Since \mathbb{R}^2 has dimension 2, they form a basis of \mathbb{R}^2 , so we can solve (2.1.2) for s and t .

Less abstractly, write v, w, x and y as column vectors, with $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ and $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$. Let $A = \begin{bmatrix} v_1 & w_1 \\ v_2 & w_2 \end{bmatrix}$. Then (2.1.2) is equivalent to saying that

$$(2.1.3) \quad A \cdot \begin{bmatrix} s \\ -t \end{bmatrix} = y - x.$$

Since the columns of A are linearly independent, A is invertible, and we may solve (2.1.3) either by Gauss elimination or by multiplying through by A^{-1} . □

Note that all we have used is that \mathbb{R}^2 is 2-dimensional. So the above argument also holds in a plane through the origin in \mathbb{R}^3 .

Remark 2.1.12. We can extend these same ideas to affine planes in \mathbb{R}^3 . (An affine plane in \mathbb{R}^n is a translate of a plane through the origin, i.e., a translate of a two-dimensional linear subspace.) Two affine planes in \mathbb{R}^n are defined to be parallel if they are translates of one another. One may then show that two distinct affine planes in \mathbb{R}^3 are parallel if and only if they do not intersect.

There are higher dimensional analogues as well.

We next wish to address the question of which translations preserve a particular line. The following definition is useful.

Definition 2.1.13. Given a vector $0 \neq w \in \mathbb{R}^n$ and a line $\ell = x + \text{span}(v)$ in \mathbb{R}^n we say that w is parallel to ℓ ($w \parallel \ell$) if $w \in \text{span}(v)$.

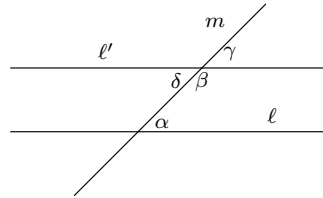
Proposition 2.1.14. Let $\ell = x + \text{span}(v)$ be a line in \mathbb{R}^n and let $0 \neq w \in \mathbb{R}^n$. Then $\tau_w(\ell) = \ell$ if and only if $w \parallel \ell$.

Proof. $\tau_w(\ell) = (x + w) + \text{span}(v)$. By Proposition 2.1.7, this is equal to ℓ if and only if $w = (x + w) - x \in \text{span}(v)$. \square

We can recover more of Euclid's standard results.

Definition 2.1.15. Let ℓ and ℓ' be lines in \mathbb{R}^2 . A transversal to ℓ and ℓ' is a line m that intersects both. In the following diagram, α and γ are called corresponding angles, while α and δ are called alternate interior angles. γ and δ are vertical angles.

(2.1.4)



We shall not discuss angle measure until Section 5.4 below. But all we shall need about it here is that a straight angle has measure π and that translations preserve angle measure (see Proposition 5.4.9). We use the unsigned notion of angle measure here.

Proposition 2.1.16. Let ℓ and ℓ' be parallel lines in \mathbb{R}^2 and let m be transversal to them. Then corresponding angles have equal measure, as do alternate interior angles.

Proof. Alternate interior angles clearly have the same measure, as, in (2.1.4), the measures of γ and β add up to π , as do the measures of δ and β . Thus, it suffices to show that the measures of corresponding angles are equal.

Let $x = m \cap \ell$ and let $y = m \cap \ell'$. Then $\tau_{y-x}(\ell)$ is the unique line through y parallel to ℓ , and hence is equal to ℓ' . But $(y-x) \parallel m$, so $\tau_{y-x}(m) = m$. In particular, τ_{y-x} carries α onto γ . Since translations preserve angle measure, the result follows. \square

Corollary 2.1.17. *The measures of the three interior angles of a triangle in \mathbb{R}^2 add up to π .*

Proof. Let A , B and C be the vertices of the triangle and let α , β and γ be the interior angles at A , B and C , respectively. Let $\ell' = \tau_{(B-A)}(\ell)$. Then we obtain the following diagram.



By Proposition 2.1.16, α and δ have the same measure, as do γ and ϵ . But the measures of δ , β and ϵ add up to π , as their angle sum forms a straight line. \square

Just as the parallel postulate is false in hyperbolic space, so is the angle sum theorem.

2.2. Lines in the plane. We should discuss the relationship between our definition of lines in \mathbb{R}^n and the more usual definitions of lines in the plane. We have defined lines to be translates of one-dimensional linear subspaces of \mathbb{R}^n . This coincides with the notion of one-dimensional affine subspaces as discussed in Section 2.8 below (Definition 2.8.6).

Linear algebra gives another way to construct affine subspaces: by Lemma 1.3.3, the solution set of a linear system $Ax = b$ for an $m \times n$ matrix A is either \emptyset or an affine subspace of \mathbb{R}^n of dimension $n - \text{rank } A$. With a little work, one can show that every affine subspace can be obtained in this way. There are obviously a lot of variables here, and this description of a particular affine subspace is not unique.

When $n = 2$ and the subspace is one-dimensional, this latter description becomes simpler, and the most common description of a line in the plane is as the solutions of a linear system

$$(2.2.1) \quad ax + by = c,$$

where a and b are not both 0. This is precisely the solutions of $Ax = [c]$ where $A = [a \ b]$ and $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$. That a and b are not both zero says that A is a nonzero matrix and hence has rank one (its single row is linearly independent). The nullspace $N(A)$ obviously contains $\begin{bmatrix} -b \\ a \end{bmatrix}$ and hence also contains its span. By (1.6.2), $N(A)$ has dimension one. Thus,

$$(2.2.2) \quad N(A) = \text{span} \left(\begin{bmatrix} -b \\ a \end{bmatrix} \right).$$

(This could also be verified via Gauss elimination with a lot less theory.) Lemma 1.3.3 then gives the following:

Proposition 2.2.1. *Let v be a particular solution of (2.2.1). Then set of all solutions is precisely the line*

$$v + \text{span} \left(\begin{bmatrix} -b \\ a \end{bmatrix} \right).$$

Of course, the line $\begin{bmatrix} a \\ b \end{bmatrix} + \text{span}(e_2)$ is the line $x = a$ and if $c \neq 0$, the line $\begin{bmatrix} a \\ b \end{bmatrix} + \text{span}(\begin{bmatrix} c \\ d \end{bmatrix})$ may be written in slope-intercept form as $y = \frac{d}{c}x + \frac{bc-ad}{c}$. In particular, the slope of the line $\begin{bmatrix} a \\ b \end{bmatrix} + \text{span}(\begin{bmatrix} c \\ d \end{bmatrix})$ is $\frac{d}{c}$. The following is immediate from the slope-intercept form.

Proposition 2.2.2. *Two lines in the plane are parallel if and only if they have the same slope.*

The point-slope formula is also important. Here, the line containing $\begin{bmatrix} a \\ b \end{bmatrix}$ with slope m has point-slope formula

$$(2.2.3) \quad \frac{y - b}{x - a} = m.$$

This may be immediately converted to slope-intercept form by multiplying both sides by $x - a$. We obtain

$$(2.2.4) \quad y = mx + (b - ma).$$

This line does have slope m and contains the point $\begin{bmatrix} a \\ b \end{bmatrix}$. Substituting for x and y , we see it is the unique such line:

Proposition 2.2.3. *$y = mx + (b - ma)$ is the unique line through $\begin{bmatrix} a \\ b \end{bmatrix}$ with slope m .*

2.3. Inner products and distance. All the geometric properties we study in this book are based on inner products. The inner product determines distance and angles. The simplest case is the Euclidean case, where the inner product is the same at every point. We study that here. More complicated geometries are obtained by allowing the inner product to vary from point to point, providing what's called a Riemannian metric on the space one is studying. Distance and angle are then obtained by applying this metric to pairs of tangent vectors at the point in question. That is the defining property of the geometry of hyperbolic space, for instance. We shall expand on this later.

The standard inner product on \mathbb{R}^n is called the dot product: for vectors $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ and $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ in \mathbb{R}^n , the dot product $\langle x, y \rangle$ is given by

$$\langle x, y \rangle = x_1y_1 + \cdots + x_ny_n.$$

As the reader may easily verify, the dot product satisfies the following properties:

Lemma 2.3.1. *The dot product is:*

(1) *Bilinear: for $x, y, z \in \mathbb{R}^n$ and $a \in \mathbb{R}$,*

$$(2.3.1) \quad \langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle \quad \langle x, ay \rangle = a\langle x, y \rangle,$$

$$(2.3.2) \quad \langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle \quad \langle ax, y \rangle = a\langle x, y \rangle.$$

(2) *Symmetric: for $x, y \in \mathbb{R}^n$,*

$$(2.3.3) \quad \langle x, y \rangle = \langle y, x \rangle.$$

(3) *Positive-definite:*

$$(2.3.4) \quad \langle x, x \rangle \geq 0 \quad \text{for all } x \in \mathbb{R}^n,$$

$$(2.3.5) \quad \langle x, x \rangle = 0 \quad \text{if and only if } x = 0.$$

Property (3) follows because if $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, then $\langle x, x \rangle = \sum_{i=1}^n x_i^2$. Since each x_i^2 is nonnegative,

$$\begin{aligned} \sum_{i=1}^n x_i^2 = 0 &\Leftrightarrow x_i^2 = 0 \text{ for all } i \\ &\Leftrightarrow x_i = 0 \text{ for all } i \\ &\Leftrightarrow x = 0. \end{aligned}$$

Note that (2.3.1) says that if $x \in \mathbb{R}^n$ then the function $f_x : \mathbb{R}^n \rightarrow \mathbb{R}$ given by

$$f_x(y) = \langle x, y \rangle$$

is linear. (2.3.2) shows the same for $g_x(y) = \langle y, x \rangle$.

The positive-definite property allows us to define distance in \mathbb{R}^n . First we define the norm.

Definition 2.3.2. The norm function on \mathbb{R}^n is given by

$$\|x\| = \sqrt{\langle x, x \rangle}.$$

For $x, y \in \mathbb{R}^n$ we define the distance from x to y to be

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

We wish to show that distance is well-behaved. The next theorem will help show this.

Theorem 2.3.3 (Cauchy–Schwarz Inequality). *Let $x, y \in \mathbb{R}^n$. Then*

$$(2.3.6) \quad |\langle x, y \rangle| \leq \|x\| \|y\| \quad \text{with equality} \Leftrightarrow x, y \text{ are linearly dependent.}$$

Removing the absolute value, $\langle x, y \rangle = \|x\| \|y\|$ if and only if either $x = 0$ or $y = sx$ for some $s \geq 0$.

Proof. If either x or y is 0, both sides of the inequality are 0, and the result follows, so assume $x, y \neq 0$. Consider the inner product

$$(2.3.7) \quad \langle tx + y, tx + y \rangle = t^2 \langle x, x \rangle + 2t \langle x, y \rangle + \langle y, y \rangle.$$

as a quadratic in t . By positive-definiteness, the quadratic has a root if and only if $tx + y = 0$ for some t , which may occur if and only if x, y is linearly dependent. If (2.3.7) has no roots, then the discriminant

$$(2.3.8) \quad 4 \langle x, y \rangle^2 - 4 \langle x, x \rangle \langle y, y \rangle < 0,$$

hence

$$|\langle x, y \rangle| = \sqrt{\langle x, y \rangle^2} < \sqrt{\langle x, x \rangle \langle y, y \rangle} = \|x\| \|y\|.$$

If (2.3.7) does have a root it has only one root: the value of t for which $tx = -y$. There is only one such t since $x \neq 0$. For a quadratic with only one root, the discriminant is 0, and (2.3.6) follows.

In particular for $x \neq 0$, we have equality in (2.3.6) if and only if $y = sx$ for some s . In this case, $\langle x, y \rangle = \langle x, sx \rangle = s\langle x, x \rangle$, in which case $\langle x, y \rangle$ is nonnegative if and only if $s \geq 0$. Thus, $\langle x, y \rangle = \|x\| \|y\|$ if and only if either $x = 0$ or $y = sx$ for $s \geq 0$. \square

Proposition 2.3.4. *The Euclidean norm $\|x\| = \sqrt{\langle x, x \rangle}$ satisfies:*

- (1) $\|x\| \geq 0$ for all x , with equality if and only if $x = 0$.
- (2) $\|cx\| = |c|\|x\|$ for $x \in \mathbb{R}^n$, $c \in \mathbb{R}$.
- (3) *The triangle inequality holds: $\|x + y\| \leq \|x\| + \|y\|$ with equality if and only if either $x = 0$ or $y = sx$ for some $s \geq 0$.*

Proof. (1) and (2) follow from the positive-definiteness and bilinearity of the inner product, respectively. (3) follows from the Cauchy–Schwarz inequality:

$$\begin{aligned} \langle x + y, x + y \rangle &= \langle x, x \rangle + 2\langle x, y \rangle + \langle y, y \rangle \\ &\leq \langle x, x \rangle + 2|\langle x, y \rangle| + \langle y, y \rangle \\ &\leq \langle x, x \rangle + 2\|x\| \|y\| + \langle y, y \rangle \\ &= (\|x\| + \|y\|)^2, \end{aligned}$$

with equality if and only if $\langle x, y \rangle = \|x\| \|y\|$. \square

Recall from Corollary 2.1.8 that if $x \neq y$, the unique line containing x and y is

$$\overleftrightarrow{xy} = \{(1-t)x + ty : t \in \mathbb{R}\}.$$

As t goes from 0 to 1, this traces out the line segment from x to y :

Definition 2.3.5. For $x \neq y \in \mathbb{R}^n$, the line segment from x and y is

$$(2.3.9) \quad \overline{xy} = \{(1-t)x + ty : 0 \leq t \leq 1\}.$$

We shall also record the following here.

Definition 2.3.6. Let $x \neq y \in \mathbb{R}^n$. The ray emanating at x and containing y is

$$(2.3.10) \quad \overrightarrow{xy} = \{(1-t)x + ty : t \geq 0\}.$$

The following measure of distance is important. Recall $d(x, y) = \|y - x\|$.

Lemma 2.3.7. *Let $z = (1-t)x + ty$ for $t \in \mathbb{R}$.*

- (1) *If $t \geq 0$, then $d(x, z) = td(x, y)$.*
- (2) *If $t \leq 1$, then $d(z, y) = (1-t)d(x, y)$.*
- (3) *If $0 \leq t \leq 1$ then $d(x, z) + d(z, y) = d(x, y)$.*

Proof. (1) is a direct calculation:

$$d(x, z) = \|(1-t)x + ty - x\| = \|-tx + ty\| = \|t(y-x)\| = |t|\|y-x\|.$$

The result follows, as we've assumed $t \geq 0$. The proof of (2) is similar, or we could deduce it by exchanging x with y and $1-t$ with t . (3) follows by adding (1) and (2). \square

Proposition 2.3.8. *The Euclidean distance function satisfies the following, for all $x, y, z \in \mathbb{R}^n$:*

- (1) $d(x, y) \geq 0$ with equality if and only if $x = y$.
- (2) $d(x, y) = d(y, x)$.
- (3) $d(x, y) \leq d(x, z) + d(z, y)$ with equality if and only if z is on the line segment \overline{xy} .

Proof. These follow from the analogous properties in Proposition 2.3.4, with (3) being the only one requiring a proof. Note that (3) is trivial if $x = y$, where the line segment degenerates to the point x . Thus, we assume $x \neq y$. Now,

$$\begin{aligned} d(x, y) &= \|y - x\| = \|(y - z) + (z - x)\| \\ &\leq \|y - z\| + \|z - x\| = d(z, y) + d(x, z) \end{aligned}$$

with equality if and only if either $y - z = 0$ or $z - x = s(y - z)$ for some $s \geq 0$. In the former case, $z = y$. In the latter, we solve for z , getting

$$\begin{aligned} z + sz &= x + sy \\ (1+s)z &= x + sy \\ z &= \frac{1}{1+s}x + \frac{s}{1+s}y \\ &= (1-t)x + ty \end{aligned}$$

for $t = \frac{s}{1+s}$. Since $s \geq 0$, both t and $1-t$ are nonnegative, and hence $t \in [0, 1]$. \square

Our goal in this part of the book is to study the geometry of \mathbb{R}^n , and in doing so it will be valuable to study its linear subspaces. A subspace inherits the inner product from \mathbb{R}^n , and it will be useful in this context to study it using abstract tools for studying a vector space with an inner product.

Definition 2.3.9. An inner product space is a vector space over \mathbb{R} together with a function

$$\begin{aligned} V \times V &\rightarrow \mathbb{R} \\ (v, w) &\mapsto \langle v, w \rangle, \end{aligned}$$

satisfying the properties of bilinearity, symmetry and positive-definiteness given in Lemma 2.3.1 for the dot product.

As was the case for the standard inner product on \mathbb{R}^n , we may make the following definitions.

Definition 2.3.10. Let V be an inner product space. Then the induced norm on V is given by setting $\|v\| = \sqrt{\langle v, v \rangle}$. The distance function induced by this norm is given by setting $d(v, w) = \|w - v\|$ for $v, w \in V$.

The reader should verify the following.

Theorem 2.3.11. *Let V be an inner product space. Then the Cauchy–Schwarz inequality holds for the inner product and its induced norm, precisely as stated in Theorem 2.3.3. The norm then satisfies the properties listed in Proposition 2.3.4. In consequence, the induced distance function satisfies the three properties listed in Proposition 2.3.8.*

Norms are important in analysis, and do not always come from inner products, so we give a general definition:

Definition 2.3.12. A norm on a vector space V is a function $\|\cdot\| : V \rightarrow \mathbb{R}$ such that:

- (1) $\|\cdot\|$ is positive definite: $\|v\| \geq 0$ for all $v \in V$, with equality if and only if $v = 0$.
- (2) $\|cv\| = |c|\|v\|$ for $c \in \mathbb{R}$ and $v \in V$.
- (3) $\|\cdot\|$ satisfies the following triangle inequality: $\|v + w\| \leq \|v\| + \|w\|$ for all $v, w \in V$.

The distance function induced by this norm is given by setting

$$d(v, w) = \|w - v\|$$

for $v, w \in V$.

Note that the triangle inequality in Definition 2.3.12 is weaker than that in Proposition 2.3.4. Thus, the following example shows that not every norm comes from an inner product.

Lemma 2.3.13. *There is a norm on \mathbb{R}^2 given by setting $\left\| \begin{bmatrix} a \\ b \end{bmatrix} \right\| = |a| + |b|$. Since $\|e_1 + e_2\| = \|e_1\| + \|e_2\|$ this norm is not induced by an inner product on \mathbb{R}^2 .*

Proof. All three properties in Definition 2.3.12 are immediate from the properties of the absolute value function on \mathbb{R} , which is the norm coming from the standard inner product on $\mathbb{R} = \mathbb{R}^1$. \square

From now on we will only consider norms induced by inner products.

Remark 2.3.14. There are a number of different inner products we could put on \mathbb{R}^n . For instance, if c_1, \dots, c_n are positive constants and if $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$

and $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, then

$$(2.3.11) \quad \langle\langle x, y \rangle\rangle = c_1 x_1 y_1 + \cdots + c_n x_n y_n$$

gives an inner product satisfying the properties in Lemma 2.3.1, and inducing a different notion of distance from the usual one. The set of points of norm one in this new norm is an ellipsoid. (Ellipsoids are important in studying multinormal distributions in probability and statistics, and this new inner product could be a useful tool in such a study.)

We shall see using Gram–Schmidt orthogonalization that any two inner products on \mathbb{R}^n differ by a linear change of variables.

2.4. Euclidean isometries are affine. We will make heavy use of isometries in studying Euclidean, spherical and hyperbolic geometry. Isometries provide the congruences studied by Euclid.

Definition 2.4.1. An isometry of \mathbb{R}^n is a function $\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that

$$(2.4.1) \quad d(\alpha(x), \alpha(y)) = d(x, y) \quad \text{for all } x, y \in \mathbb{R}^n,$$

where d is the standard Euclidean distance function on \mathbb{R}^n : $d(x, y) = \|y - x\|$, defined using the standard norm. We write \mathcal{I}_n for the set of all isometries of \mathbb{R}^n .

The subsets $X, Y \subset \mathbb{R}^n$ are said to be congruent if there is an isometry $\alpha \in \mathcal{I}_n$ with $\alpha(X) = Y$. We then say that α provides a congruence from X to Y .

Note that (2.4.1) implies an isometry is one-to-one, as if $\alpha(x) = \alpha(y)$, then $d(\alpha(x), \alpha(y)) = 0$. It is customary to also require that α be onto, as that then shows that \mathcal{I}_n is a group (as discussed below). But we shall see that (2.4.1) implies that α is onto, and that is one of the goals of this section.

We've seen one infinite family of isometries already: the translations. Recall that for $x \in \mathbb{R}^n$, the translation $\tau_x : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is defined by

$$\tau_x(y) = (x + y).$$

Lemma 2.4.2. For $x \in \mathbb{R}^n$, the translation τ_x is an isometry.

Proof. For $y, z \in \mathbb{R}^n$,

$$\begin{aligned} d(\tau_x(y), \tau_x(z)) &= \|\tau_x(y) - \tau_x(z)\| \\ &= \|(x + y) - (x + z)\| = \|y - z\| = d(y, z). \quad \square \end{aligned}$$

We shall see that all isometries of \mathbb{R}^n are composites of translations and linear isometries. The following lemma is key.

Lemma 2.4.3. Let $\alpha \in \mathcal{I}_n$ and let x, y in \mathbb{R}^n . Then

$$(2.4.2) \quad \alpha((1 - t)x + ty) = (1 - t)\alpha(x) + t\alpha(y) \quad \text{for } 0 \leq t \leq 1.$$

Proof. We may and shall assume $x \neq y$. Write $z = (1 - t)x + ty$. We have

$$\begin{aligned} d(\alpha(x), \alpha(z)) + d(\alpha(z), \alpha(y)) &= d(x, z) + d(z, y) \quad (\alpha \text{ is an isometry}) \\ &= d(x, y) \quad (\text{Lemma 2.3.7}) \\ &= d(\alpha(x), \alpha(y)). \end{aligned}$$

By Proposition 2.3.8(3), $\alpha(z)$ is on the line segment from $\alpha(x)$ to $\alpha(y)$, say

$$\alpha(z) = (1 - s)\alpha(x) + s\alpha(y)$$

for $s \in [0, 1]$. By Lemma 2.3.7(1),

$$d(\alpha(x), \alpha(z)) = sd(\alpha(x), \alpha(y)) = sd(x, y),$$

but since α is an isometry,

$$d(\alpha(x), \alpha(z)) = d(x, z) = td(x, y).$$

Since $x \neq y$, $s = t$ and the result follows. \square

The following is immediate.

Corollary 2.4.4. *Isometries of \mathbb{R}^n preserve line segments: if $\alpha \in \mathcal{I}_n$ and $x, y \in \mathbb{R}^n$, then*

$$\alpha(\overline{xy}) = \overline{\alpha(x)\alpha(y)}.$$

We wish to remove the condition that $0 \leq t \leq 1$ in Lemma 2.4.3. Some definitions may be helpful. Note that Lemma 2.4.3 says that an isometry $\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is an affine function as defined in the following:

Definition 2.4.5.

- (1) A subset $C \subset \mathbb{R}^n$ is convex if $\overline{xy} \subset C$ for all $x, y \in C$.
- (2) Let $C \subset \mathbb{R}^n$ be convex. A function $f : C \rightarrow \mathbb{R}^m$ is affine if

$$f((1 - t)x + ty) = (1 - t)f(x) + tf(y) \quad \text{for all } x, y \in C \text{ and } t \in [0, 1].$$

Lemma 2.4.3 may now be restated as follows.

Lemma 2.4.6. *Isometries of \mathbb{R}^n are affine.*

We currently have only one family of isometries:

Corollary 2.4.7. *Translations of \mathbb{R}^n are affine.*

For some examples of convex sets, we have the following.

Lemma 2.4.8. *Let $-\infty \leq a < b \leq \infty$. Then the interval (a, b) is convex. So are $[a, b]$, $[a, b)$ and $(a, b]$ whenever they are defined.*

Proof. We treat the case $[a, b]$. The others are similar. Let $r, s \in [a, b]$, and $t \in [0, 1]$. Then $a \leq r, s \leq b$. Since t and $1 - t$ are nonnegative, we obtain

$$a = (1 - t)a + ta \leq (1 - t)r + ts \leq (1 - t)b + tb = b. \quad \square$$

And more examples come as follows:

Example 2.4.9. A linear subspace $V \subset \mathbb{R}^n$ is convex, as convexity is expressed in terms of the vector operations. Indeed,

$$(2.4.3) \quad (1 - t)x + ty \in V \quad \text{for all } x, y \in V \text{ and } t \in \mathbb{R}.$$

Thus, for $x \neq y \in V$, $\overline{xy} \subset V$.

Affine functions are very important in piecewise linear topology, or, on a more basic level, in studying simplicial complexes. In that context, they are important in developing modern algebraic topology. Lemma 2.4.3 says that isometries of \mathbb{R}^n are affine maps. We wish to show they automatically satisfy a seemingly stronger property, based on the algebraic closure properties of \mathbb{R}^n :

Proposition 2.4.10. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be affine. Then*

$$(2.4.4) \quad f((1-t)x + ty) = (1-t)f(x) + tf(y) \quad \text{for all } t \in \mathbb{R}.$$

Indeed, if C is convex and if $f : C \rightarrow \mathbb{R}^m$ is affine, then

$$(2.4.5) \quad f((1-t)x + ty) = (1-t)f(x) + tf(y) \\ \text{whenever } x, y, (1-t)x + ty \in C.$$

Proof. Let $x, y \in C$, and again we may assume $x \neq y$. Let $t \in \mathbb{R}$, and suppose $z = (1-t)x + ty \in C$. We wish to show $f(z) = (1-t)f(x) + tf(y)$.

We already know this for $t \in [0, 1]$. Assume $t > 1$. Then we can solve for y as an element of \overline{xz} :

$$y = \frac{t-1}{t}x + \frac{1}{t}z \\ = (1-u)x + uz,$$

for $u = \frac{1}{t}$. Since $t > 1$, $u \in (0, 1)$. Since f is affine,

$$f(y) = (1-u)f(x) + uf(z).$$

Now solve for $f(z)$:

$$f(z) = \frac{u-1}{u}f(x) + \frac{1}{u}f(y) \\ = (1-t)f(x) + tf(y),$$

as $t = \frac{1}{u}$.

Thus, the desired result holds for $t > 1$, and it suffices to consider $t < 0$. Let $z = (1-t)x + ty$ and let $s = 1-t$. Then $s > 1$ and $z = (1-s)y + sx$. Thus, exchanging the roles of x and y , the preceding case gives

$$f(z) = (1-s)f(y) + sf(x) \\ = tf(y) + (1-t)f(x),$$

as desired. □

The following is immediate from (2.4.4).

Corollary 2.4.11. *Isometries preserve lines: if $\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is an isometry and $x \neq y \in \mathbb{R}^n$, then*

$$(2.4.6) \quad \alpha(\overrightarrow{xy}) = \overleftarrow{\alpha(x)\alpha(y)}.$$

Indeed, if $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is affine and $f(x) \neq f(y)$, then

$$(2.4.7) \quad f(\overleftrightarrow{xy}) = \overleftrightarrow{f(x)f(y)}.$$

2.5. Affine functions and linearity. Not every affine map from \mathbb{R}^n to itself is an isometry, but the affine property will help us understand the isometries.

Proposition 2.5.1. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be affine with $f(0) = 0$. Then f is linear. More generally, if $C \subset \mathbb{R}^n$ is convex with $0 \in C$ and if $f : C \rightarrow \mathbb{R}^m$ is affine with $f(0) = 0$, then*

- (1) $f(x + y) = f(x) + f(y)$ whenever x, y and $x + y$ lie in C .
- (2) $f(ax) = af(x)$ whenever x and ax lie in C , $a \in \mathbb{R}$.

In particular, if $V \subset \mathbb{R}^n$ is a linear subspace and $f : V \rightarrow \mathbb{R}^m$ is affine with $f(0) = 0$, then f is linear.

Proof. We first show (2). Here $ax = (1 - a)0 + ax$, so

$$f(ax) = (1 - a)f(0) + af(x)$$

by (2.4.5). (2) follows, as $f(0) = 0$.

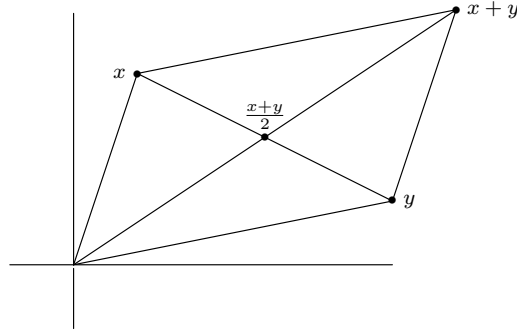


FIGURE 2.5.1. The parallelogram law

For (1), we use (2) with $a = \frac{1}{2}$, noting that the average,

$$\frac{x + y}{2} = \left(1 - \frac{1}{2}\right)x + \frac{1}{2}y,$$

of x and y lies in C .

$$\begin{aligned} \frac{1}{2}f(x + y) &= f\left(\frac{x + y}{2}\right) && \text{by (2)} \\ &= f\left(\left(1 - \frac{1}{2}\right)x + \frac{1}{2}y\right) \\ &= \left(1 - \frac{1}{2}\right)f(x) + \frac{1}{2}f(y) && \text{by (2.4.5)} \end{aligned}$$

$$= \frac{1}{2}(f(x) + f(y)).$$

Now multiply through by 2. □

Corollary 2.5.2. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be an affine map. Then $f = \tau_x \circ g$, where $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is linear and τ_x is the translation of \mathbb{R}^m by $x = f(0)$. Thus,*

$$g(v) = f(v) - f(0)$$

for all $v \in \mathbb{R}^n$.

Proof. Let $g = \tau_{-x} \circ f$ with $x = f(0)$. Since translations are isometries, they are affine. The composite of affine maps is clearly affine. And

$$g(0) = \tau_{-x}(f(0)) = f(0) - x = 0.$$

So g is linear. Now $\tau_x \circ g = \tau_x \circ \tau_{-x} \circ f = f$. □

We obtain the following.

Theorem 2.5.3. *Let $\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be an isometry. Then $\alpha = \tau_x \circ \beta$, where $\beta : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a linear isometry and $x = \alpha(0)$. Thus,*

$$\beta(v) = \alpha(v) - \alpha(0)$$

for all $v \in \mathbb{R}^n$.

Proof. Just apply the preceding proof and note that $\tau_{-x} \circ \alpha$ is the composite of two isometries, and hence an isometry. □

This allows a nice refinement of Corollary 2.4.11.

Corollary 2.5.4. *Let $\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be an isometry. Write $\alpha = \tau_y \circ \beta$, with β a linear isometry. Then the effect of α on lines is given by*

$$(2.5.1) \quad \alpha(x + \text{span}(v)) = \alpha(x) + \text{span}(\beta(v)).$$

Expressed purely in terms of α this gives

$$\alpha(x + \text{span}(v)) = \alpha(x) + \text{span}(\alpha(v) - \alpha(0)).$$

Proof. Let $\ell = x + \text{span}(v)$. We have $\beta(x + tv) = \beta(x) + t\beta(v)$, so

$$\beta(\ell) = \beta(x) + \text{span}(\beta(v)).$$

Translating this by y , we get

$$\alpha(\ell) = (\beta(x) + y) + \text{span}(\beta(v)) = \alpha(x) + \text{span}(\beta(v)). \quad \square$$

Linear algebra now gives us the missing piece in showing isometries of \mathbb{R}^n are onto and invertible.

Lemma 2.5.5. *Let A be an $n \times n$ matrix such that the induced linear function $T_A : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is an isometry. Then A is invertible and $T_A^{-1} = T_{A^{-1}}$ is a linear isometry.*

Thus if $\beta : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a linear isometry, it is bijective and its inverse function is a linear isometry.

Proof. By (2.4.1), isometries are one-to-one. So the columns of A are linearly independent. Since there are n of them, they form a basis of \mathbb{R}^n , hence A is invertible. Thus T_A is onto. But if $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is an onto isometry, then its inverse function is clearly an isometry. $T_A^{-1} = T_{A^{-1}}$ so the inverse is a linear isometry. \square

For a linear isometry β , the inverse function of $\tau_x \circ \beta$ is $\beta^{-1} \circ \tau_{-x}$, the composite of two isometries. We obtain:

Corollary 2.5.6. *Every isometry $\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is bijective and its inverse function is an isometry.*

The above argument depended on the fact that isometries are one-to-one. But linear functions, and hence affine functions, are neither one-to-one nor onto in general.

We also have the following.

Corollary 2.5.7. *The decomposition of Theorem 2.5.3 is unique: if*

$$\tau_x \circ \beta = \tau_y \circ \gamma$$

with β, γ linear isometries, then $x = y$ and $\beta = \gamma$.

Proof. By Lemma 2.5.5, the inverse function β^{-1} of β is a linear isometry.

$$\begin{aligned} \tau_y^{-1} \circ \tau_x \circ \beta \circ \beta^{-1} &= \tau_y^{-1} \circ \tau_y \circ \gamma \circ \beta^{-1} \\ \tau_{x-y} &= \gamma \circ \beta^{-1}. \end{aligned}$$

The right-hand side is a linear isometry, so τ_{x-y} is linear, and hence preserves 0. But that forces $x - y = 0$, so $x = y$, and hence

$$\text{id} = \tau_0 = \gamma \circ \beta^{-1},$$

so $\gamma = \beta$. \square

2.6. Affine automorphisms of \mathbb{R}^n . There is an useful generalization Euclidean isometries.

Definition 2.6.1. A one-to-one, onto affine map $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is called an affine automorphism. We write \mathcal{A}_n for the collection of all affine automorphisms of \mathbb{R}^n .

Since isometries $\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^n$ are bijective (Corollary 2.5.6) and are affine, we have:

Lemma 2.6.2. *Isometries of \mathbb{R}^n are affine automorphisms: $\mathcal{I}_n \subset \mathcal{A}_n$.*

In general, an automorphism of a given mathematical object is a bijective function f from that object to itself, such that both f and its inverse function, f^{-1} preserve the mathematical structure we are studying. This idea occurs in many areas of mathematics, e.g., in differential topology, an automorphism is called a diffeomorphism. We shall study these in the context of smooth manifolds, below. The simplest instance of this is the

following: a diffeomorphism $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a differentiable bijection whose inverse function is also differentiable. Thus, the function $f : \mathbb{R} \rightarrow \mathbb{R}$ given by $f(x) = x^3$, while bijective, is not a diffeomorphism, because its inverse function, $f^{-1}(x) = \sqrt[3]{x}$ is not differentiable at 0.

Since linear isomorphisms have linear inverses, the following is justified.

Definition 2.6.3. A linear automorphism of a vector space V is a linear isomorphism $g : V \rightarrow V$.

We shall now show that our definition of “affine automorphism” is also justified, i.e., that the inverse function of a bijective affine map is affine.

Proposition 2.6.4. *Let f be an affine automorphism of \mathbb{R}^n . Then f can be written uniquely in the form*

$$(2.6.1) \quad f = \tau_x \circ g$$

with g a linear automorphism of \mathbb{R}^n (and hence $x = f(0)$). The inverse function f^{-1} is also an affine automorphism.

Proof. The decomposition (2.6.1) comes from Corollary 2.5.2. Since f is one-to-one, so is g . So g is a linear isomorphism by the first part of the argument for Lemma 2.5.5. Uniqueness follows precisely as in Corollary 2.5.7.

The collection of affine automorphisms is closed under composition and contains both the isometries and the linear isomorphisms. So $f^{-1} = g^{-1} \circ \tau_{-x}$ is an affine automorphism. \square

2.7. Similarities. Similarities are important in Euclidean geometry.

Definition 2.7.1. A function $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a similarity with scaling factor $s > 0$ if

$$(2.7.1) \quad d(f(x), f(y)) = s \cdot d(x, y)$$

for all $x, y \in \mathbb{R}^n$. We write \mathcal{S}_n for the set of all similarities of \mathbb{R}^n .

Note that an isometry is a similarity with scaling factor 1. Another family of examples is as follows.

Example 2.7.2. Let $0 \neq s \in \mathbb{R}$. Define $\mu_s : \mathbb{R}^n \rightarrow \mathbb{R}^n$ by $\mu_s(x) = sx$. Then for $s > 0$, μ_s is a similarity with scaling factor s , as

$$(2.7.2) \quad d(\mu_s(x), \mu_s(y)) = \|sy - sx\| = s\|y - x\| = sd(x, y).$$

Similarly, for $s < 0$, μ_s is a similarity with scaling factor $|s|$.

Since the functions μ_s , $s \neq 0$, are linear, they are affine. They are automorphisms with inverse functions $\mu_{\frac{1}{s}}$.

The following is immediate.

Lemma 2.7.3. *Let f and g be similarities of \mathbb{R}^n with scaling factors s and t , respectively. Then $f \circ g$ is a similarity with scaling factor st .*

Corollary 2.7.4. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a similarity with scaling factor s . Then $f = \mu_s \circ \alpha$ for α an isometry of \mathbb{R}^n . Thus, f is a bijection, and $f^{-1} = \alpha^{-1} \circ \mu_{\frac{1}{s}}$ is a similarity with scaling factor $\frac{1}{s}$. In particular, f is an affine automorphism of \mathbb{R}^n . We have inclusions*

$$(2.7.3) \quad \mathcal{I}_n \subset \mathcal{S}_n \subset \mathcal{A}_n.$$

Finally, if $f = \mu_s \circ \alpha$, then f is linear if and only if α is linear.

Proof. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a similarity with scaling factor s . Then

$$\alpha = \mu_{\frac{1}{s}} \circ f$$

is a similarity with scaling factor 1, and hence is an isometry. And $f = \mu_s \circ \alpha$.

The rest follows, as μ_t is linear and hence affine for all $t > 0$. \square

Corollary 2.7.5. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a similarity with scaling factor s . Then f may be written uniquely in the form*

$$(2.7.4) \quad f = \tau_x \circ \mu_s \circ \beta$$

With β a linear isometry of \mathbb{R}^n . Here $x = f(0)$.

Proof. Since f is an affine automorphism, we may apply Proposition 2.6.4 to obtain

$$f = \tau_x \circ g$$

with $x = f(0)$ and g a linear automorphism of \mathbb{R}^n . Since translations are similarities with scaling factor 1, g is a similarity with scaling factor s . Now apply Corollary 2.7.4 to g . \square

Another consequence of Corollary 2.7.4 is the following.

Corollary 2.7.6. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a similarity and let $\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be an isometry. Then $f \circ \alpha \circ f^{-1}$ is an isometry.*

Proof. $f \circ \alpha \circ f^{-1}$ is a similarity with scaling factor $s \cdot 1 \cdot \frac{1}{s} = 1$, hence an isometry. \square

Note the same proof shows the following.

Corollary 2.7.7. *Let $f, g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be similarities. Then $f \circ g \circ f^{-1}$ is a similarity whose scaling factor is the same as that of g .*

2.8. Convex and affine hulls; affine subspaces and maps. Since we are studying convex sets in a very general manner we should dispose of the following immediately. We shall use it without discussion

Lemma 2.8.1. *The intersection of an arbitrary family of convex sets is convex.*

Note that the empty set is convex, and may occur as such an intersection. We regard this as a pathological example, and wish to study nonempty convex sets.

2.8.1. Convex and affine hulls.

Definition 2.8.2. An affine combination of $x_1, \dots, x_k \in \mathbb{R}^n$ is a sum

$$a_1x_1 + \cdots + a_kx_k$$

with $\sum_{i=1}^k a_i = 1$. A convex combination of $x_1, \dots, x_k \in \mathbb{R}^n$ is an affine combination $a_1x_1 + \cdots + a_kx_k$ in which the coefficients a_i are all nonnegative.

The affine hull (or affine span) $\text{Aff}(x_1, \dots, x_k)$ of x_1, \dots, x_k is the set of all affine combinations of x_1, \dots, x_k . The convex hull $\text{Conv}(x_1, \dots, x_k)$ of x_1, \dots, x_k is the set of all convex combinations of x_1, \dots, x_k .

Note that the order of x_1, \dots, x_k is irrelevant to the definitions of affine and convex hulls. A convenient alternative notation is given as follows: if $X = \{x_1, \dots, x_k\}$ we may write $\text{Aff}(X)$ and $\text{Conv}(X)$ for the affine and convex hulls of x_1, \dots, x_k , respectively.

This permits the following definition:

Definition 2.8.3. A polytope is the convex hull of some finite set of points in \mathbb{R}^n for some n .¹

Polytopes are important in several areas of topology and geometry, including applied topics such as linear programming.

Note that for a single point, $\text{Aff}(x) = \text{Conv}(x) = \{x\}$. We have also seen the affine and convex hulls of two points:

Example 2.8.4. Let $x_1 \neq x_2 \in \mathbb{R}^n$ and let $a_1 + a_2 = 1$. Then $a_1 = 1 - a_2$, so

$$a_1x_1 + a_2x_2 = (1 - a_2)x_1 + a_2x_2 = (1 - t)x_1 + tx_2$$

for $t = a_2$. Moreover, a_1 and a_2 are both nonnegative if and only if $t \in [0, 1]$. Thus, $\text{Conv}(x_1, x_2)$ is the line segment $\overline{x_1x_2}$ and $\text{Aff}(x_1, x_2)$ is the line $\overleftrightarrow{x_1x_2}$.

In particular, for $a < b \in \mathbb{R}$, $\text{Conv}(a, b) = \overline{ab}$ is just the closed interval $[a, b]$, and $\text{Aff}(a, b) = \mathbb{R}$.

In particular, we have a solid understanding of the convex and affine hulls of any two distinct points in \mathbb{R}^n . We also have some important explicit examples of convex and affine hulls of multiple points:

Example 2.8.5. The convex and affine hulls of the canonical basis vectors of \mathbb{R}^n play important roles in multiple mathematical contexts. Since

$$a_1e_1 + \cdots + a_n e_n = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix},$$

¹Some authors would call this a (compact) convex polyhedron. There are numerous variations in naming, with the same name being used for different concepts in some cases.

we obtain

$$(2.8.1) \quad \text{Aff}(e_1, \dots, e_n) = \left\{ \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} : \sum_{i=1}^n a_i = 1 \right\} \\ = \{ \alpha \in \mathbb{R}^n : \langle \alpha, \xi \rangle = 1 \}, \quad \xi = e_1 + \dots + e_n.$$

Equivalently, one can view $\text{Aff}(e_1, \dots, e_n)$ as solution space of the matrix equation

$$\xi^T \cdot x = [1].$$

Here, the transpose, ξ^T , of ξ is the $1 \times n$ matrix $[1 \cdots 1]$, i.e., each entry of the matrix is 1.

The convex hull of e_1, \dots, e_n is known as the standard $(n-1)$ -simplex Δ^{n-1} :

$$(2.8.2) \quad \Delta^{n-1} = \left\{ \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} : \sum_{i=1}^n a_i = 1 \text{ and } a_i \geq 0 \text{ for } i = 1, \dots, n \right\},$$

the set of all $\alpha \in \mathbb{R}^n$ whose coordinates are all nonnegative with $\langle \alpha, \xi \rangle = 1$. Such α are sometimes called probability vectors and figure prominently in the study of Markov chains and other phenomena in probability theory.

The indexing comes from the fact that the topological dimension of Δ^{n-1} is $n-1$. Δ^1 is the line segment $\overline{e_1 e_2}$, and hence has dimension 1. Δ^2 is the triangle (including its interior) in \mathbb{R}^3 with vertices e_1, e_2 and e_3 and therefore is 2-dimensional. Δ^3 is a solid (i.e., including its interior) tetrahedron in \mathbb{R}^4 , and for $k > 3$, Δ^k is a k -dimensional analogue of a tetrahedron. One can show that topologically, Δ^k is homeomorphic to the closed unit disk $\mathbb{D}^k = \{x \in \mathbb{R}^k : \|x\| \leq 1\}$.

Note that the line segment Δ^1 has two vertices, the triangle Δ^2 has three 1-dimensional edges, and the tetrahedron Δ^3 has four 2-dimensional faces. There is much to study and generalize here.

For $x \neq y \in \mathbb{R}^n$, $\text{Aff}(x, y) = \overleftrightarrow{xy}$ is a line in \mathbb{R}^n , and hence is a translate of a one-dimensional linear subspace of \mathbb{R}^n . To study affine combinations of more than two vectors, it is useful to generalize the notion of lines in \mathbb{R}^n .

Definition 2.8.6. An affine subspace of \mathbb{R}^n is a translate

$$\tau_x(V) = \{v + x : v \in V\}$$

of a linear subspace V of \mathbb{R}^n . The dimension of $\tau_x(V)$ is defined to be the dimension of V . We shall refer to V as the *linear base* of $H = \tau_x(V)$.

Examples 2.8.7. 0 is the only linear subspace of \mathbb{R}^n of dimension 0, so the affine subspaces of dimension 0 are those of the form $\tau_x(0) = \{x\}$ for $x \in \mathbb{R}^n$.

Note that our definition of a line is precisely that of an affine subspace of dimension 1. A line has the form $\tau_x(\text{span}(v))$, where $x, v \in \mathbb{R}^n$ with $v \neq 0$. But every one-dimensional linear subspace of \mathbb{R}^n has the form $\text{span}(v)$ for

some $v \neq 0$, as every one-dimensional vector space has a basis with one element.

We shall refer to the two-dimensional affine subspaces of \mathbb{R}^n as affine planes.

The following is a direct generalization of Proposition 2.1.7.

Proposition 2.8.8. *The affine subspaces $\tau_x(V)$ and $\tau_y(W)$ are equal if and only if $V = W$ and $y - x \in V$ (i.e., $y \in \tau_x(V)$). In particular, $\tau_x(V)$ is a linear subspace if and only if $x \in V$. Moreover, if $H = \tau_x(V)$, then its linear base V is equal to $\tau_{-y}(H)$ for all $y \in H$.*

Proof. $\tau_x(V) = \tau_y(W)$ if and only if $\tau_{x-y}(V) = W$. Since $0 \in V$, This implies $x - y \in W$. Similarly, $y - x \in V$. Since these are subspaces, $x - y \in V \cap W$, so $\tau_{x-y}(V) = V$ and $\tau_{y-x}(W) = W$. \square

Corollary 2.8.9. *Let $H \subset K$ be affine subspaces of \mathbb{R}^n . Then $\dim H \leq \dim K$, and if $\dim H = \dim K$, then $H = K$.*

Proof. Let $x \in H$, then the linear bases of H and K are $\tau_{-x}(H)$ and $\tau_{-x}(K)$, respectively. The result now follows from standard properties of linear subspaces. \square

There is a nice characterization of affine subspaces coming from linear algebra. We borrow from Chapter 4 for a clean proof.

Proposition 2.8.10. *A subset $H \subset \mathbb{R}^n$ is an affine subspace if and only if there is a linear map $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ for some m and an element $y \in \mathbb{R}^m$ such that $H = f^{-1}(y)$.*

Proof. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be linear and suppose $f^{-1}(y)$ is nonempty. Let $x_0 \in f^{-1}(y)$. Then x_0 is a ‘‘particular solution’’ of the equation $f(x) = y$. The general solution is then given by all the elements of the affine subspace $(\ker f) + x_0 = \tau_{x_0}(\ker f)$.

Conversely, let H be an affine subspace of \mathbb{R}^n and let V its linear base. Say $H = \tau_x(V)$. Corollary 4.3.7 constructs a linear map $\pi_{V^\perp} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ whose kernel is V . But then $H = \pi_{V^\perp}^{-1}(\pi_{V^\perp}(x))$. In fact, since $\text{Im } \pi_{V^\perp} = V^\perp$, we may replace $\pi_{V^\perp} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ with the composite

$$\mathbb{R}^n \xrightarrow{\pi_{V^\perp}} V^\perp \xrightarrow{\cong} \mathbb{R}^m,$$

where $m = n - \dim V$ and the isomorphism $V^\perp \xrightarrow{\cong} \mathbb{R}^m$ maybe be induced by an orthonormal basis of V^\perp (making it an isometry). \square

We may identify $\text{Aff}(x_1, \dots, x_n)$ as an affine subspace.

Proposition 2.8.11. *Let $x_1, \dots, x_k \in \mathbb{R}^n$. Then*

$$(2.8.3) \quad \text{Aff}(x_1, \dots, x_k) = \tau_{x_1}(\text{span}(x_2 - x_1, \dots, x_k - x_1)).$$

$\text{Aff}(x_1, \dots, x_k)$ is the smallest affine subspace containing x_1, \dots, x_k : if H is any affine subspace of \mathbb{R}^n containing x_1, \dots, x_k , then

$$\text{Aff}(x_1, \dots, x_k) \subset H.$$

In particular, affine subspaces are closed under taking affine combinations of their elements (but not linear combinations, unless they are actually linear subspaces).

In consequence, if $x \neq y \in H$, then $\overleftrightarrow{xy} = \{(1-t)x + ty : t \in \mathbb{R}\} \subset H$.

Proof. Let H be an affine subspace containing $x_1, \dots, x_k \in H$. Then the linear subspace $V = \tau_{-x_1}(H)$ contains $x_2 - x_1, \dots, x_k - x_1$, and hence contains their span. It suffices to verify (2.8.3)

Now,

$$\begin{aligned} \tau_{x_1}(a_2(x_2 - x_1) + \dots + a_k(x_k - x_1)) \\ &= x_1 + (a_2x_2 + \dots + a_kx_k) - \left(\sum_{i=2}^k a_i \right) x_1 \\ &= \left(1 - \sum_{i=2}^k a_i \right) x_1 + a_2x_2 + \dots + a_kx_k. \end{aligned}$$

Since the coefficients now add up to 1, this shows

$$\tau_{x_1}(\text{span}(x_2 - x_1, \dots, x_k - x_1)) \subset \text{Aff}(x_1, \dots, x_k).$$

Conversely, if $\sum_{i=1}^k a_i = 1$, then

$$\begin{aligned} \tau_{-x}(a_1x_1 + \dots + a_kx_k) &= (a_1x_1 + \dots + a_kx_k) - (a_1x_1 + \dots + a_kx_1) \\ &= a_2(x_2 - x_1) + \dots + a_k(x_k - x_1), \end{aligned}$$

so $\tau_{-x}(\text{Aff}(x_1, \dots, x_n)) \subset \text{span}(x_2 - x_1, \dots, x_k - x_1)$. \square

Convex hulls have an analogous property:

Proposition 2.8.12. *Let $x_1, \dots, x_k \in \mathbb{R}^n$. Then $\text{Conv}(x_1, \dots, x_k)$ is the smallest convex subset of \mathbb{R}^n containing x_1, \dots, x_k .*

Proof. We first show $\text{Conv}(x_1, \dots, x_k)$ is convex. Let $x = \sum_{i=1}^k a_i x_i$ and $y = \sum_{i=1}^k b_i x_i$ be convex combinations of x_1, \dots, x_k and let $t \in [0, 1]$. Then

$$(1-t)x + ty = [(1-t)a_1 + tb_1]x_1 + \dots + [(1-t)a_k + tb_k]x_k.$$

Since $a_i, b_i \in [0, \infty)$, each coefficient $(1-t)a_i + tb_i \in [0, \infty)$ by Lemma 2.4.8. These coefficients add up to 1 because

$$\sum_{i=1}^k [(1-t)a_i + tb_i] = (1-t) \left(\sum_{i=1}^k a_i \right) + t \left(\sum_{i=1}^k b_i \right) = (1-t)1 + t \cdot 1 = 1.$$

Now let C be a convex subset of \mathbb{R}^n containing x_1, \dots, x_k . We show by induction on k that $\text{Conv}(x_1, \dots, x_k) \subset C$. When $k = 2$, this is just the definition of convexity. Assume now that $x = a_1x_1 + \dots + a_kx_k$ is a convex

combination of x_1, \dots, x_k and that the result is true for fewer than k points. We wish to show $x \in C$. If $a_k = 1$, $x = x_k$, and $x \in C$ by hypothesis. Otherwise,

$$(2.8.4) \quad x = (1 - a_k) \left(\frac{a_1}{1 - a_k} x_1 + \dots + \frac{a_{k-1}}{1 - a_k} x_{k-1} \right) + a_k x_k.$$

By the case $k = 2$, it suffices to show that $\frac{a_1}{1 - a_k} x_1 + \dots + \frac{a_{k-1}}{1 - a_k} x_{k-1} \in C$. But this follows from the induction hypothesis as $\sum_{i=1}^{k-1} a_i = 1 - a_k$. \square

The following is now immediate from Proposition 2.4.10.

Corollary 2.8.13. *Affine subspaces are convex. Moreover, if H is an affine subspace of \mathbb{R}^n and if $f : H \rightarrow \mathbb{R}^m$ is affine, then*

$$(2.8.5) \quad f((1 - t)x + ty) = (1 - t)f(x) + tf(y) \quad \text{for all } x, y \in H \text{ and } t \in \mathbb{R}.$$

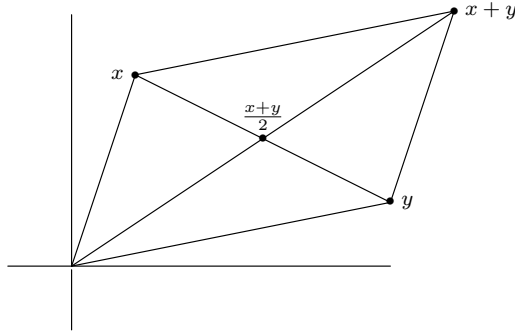
2.8.2. Joins. The proof of Proposition 2.8.12 motivates a useful notion we will call the *linear join*. In some contexts this is simply called the join, but there is a related, but different, notion we shall call the topological (or external) join that coincides with this one in special circumstances. Topological joins are important in both homotopy theory and geometry.

Definition 2.8.14. Let X and Y be subsets of \mathbb{R}^n . The linear join, $X \cdot Y$, of X and Y is the union of the line segments from points in X to points in Y

$$(2.8.6) \quad \begin{aligned} X \cdot Y &= \{(1 - t)x + ty : x \in X, y \in Y, t \in [0, 1]\} \\ &= \bigcup_{(x,y) \in X \times Y} \overline{xy}. \end{aligned}$$

If $Y = \{y\}$, a single point, we refer to $X \cdot \{y\}$ as the linear cone on X , with cone point y .

Note that the line segments in a join can intersect. For instance, for any distinct points $x, y \in \mathbb{R}^n$, the point $\frac{1}{2}(x + y)$ lies on both of the segments \overline{xy} and $\overline{0(x + y)}$ of $\{0, x\} \cdot \{y, x + y\}$:



This ambiguity is the primary distinction between linear joins and topological joins of *compact* subsets.

In the displayed example, the reader familiar with graph theory should note that this point of intersection at $\frac{1}{2}(x+y)$ marks the only distinction between the join $\{0, x\} \cdot \{y, x+y\}$ and the complete bipartite graph $\mathcal{K}(\{0, x\}, \{y, x+y\})$ with vertex sets $\{0, x\}$ and $\{y, x+y\}$.

A first application of joins to geometry is the following.

Proposition 2.8.15. *Let $x_1, \dots, x_k \in \mathbb{R}^n$. Then the convex hull of x_1, \dots, x_k is the linear join of $\text{Conv}(x_1, \dots, x_{k-1})$ with $\{x_k\}$:*

$$(2.8.7) \quad \text{Conv}(x_1, \dots, x_k) = \text{Conv}(x_1, \dots, x_{k-1}) \cdot \{x_k\}.$$

More generally, if $X = \{x_1, \dots, x_k\}$ and if $X = S \cup T$, then

$$(2.8.8) \quad \text{Conv}(X) = \text{Conv}(S) \cdot \text{Conv}(T).$$

Note we are not assuming that $S \cap T = \emptyset$.

Proof. (2.8.7) follows from (2.8.8), so we prove the latter. Note that we may make the following identification:

$$(2.8.9) \quad \text{Conv}(S) = \{a_1x_1 + \dots + a_kx_k : a_i = 0 \text{ if } x_i \notin S\} \subset \text{Conv}(X),$$

with an analogous statement for $\text{Conv}(T)$. (Implicitly, here and below, we assume this is a convex combination, i.e., $\sum_{i=1}^k s_i = 1$ and $x_i \geq 0$ for all i .) By Proposition 2.8.12, $\text{Conv}(X)$ is convex, so $\text{Conv}(S) \cdot \text{Conv}(T) \subset \text{Conv}(X)$.

To show the opposite inclusion, let $y = a_1x_1 + \dots + a_kx_k \in \text{Conv}(X)$. Let

$$b_i = \begin{cases} a_i & \text{if } x_i \in S \\ 0 & \text{otherwise,} \end{cases} \quad c_i = \begin{cases} a_i & \text{if } x_i \in X \setminus S \\ 0 & \text{otherwise.} \end{cases}$$

Let $t = \sum_{i=1}^k c_i$. If $t = 1$, then $y \in \text{Conv}(T) \subset \text{Conv}(S) \cdot \text{Conv}(T)$, and we're done. If $t = 0$, then $y \in \text{Conv}(S)$, and we're done. Otherwise $t \in (0, 1)$, and we set

$$z = \frac{b_1}{1-t}x_1 + \dots + \frac{b_k}{1-t}x_k \in \text{Conv}(S),$$

$$w = \frac{c_1}{t}x_1 + \dots + \frac{c_k}{t}x_k \in \text{Conv}(T),$$

and we have $y = (1-t)z + tw \in \text{Conv}(S) \cdot \text{Conv}(T)$. \square

Corollary 2.8.16. *Let C and D be nonempty convex subsets of \mathbb{R}^n . Then their linear join $C \cdot D$ is convex.*

Proof. Let $a, b \in C \cdot D$. Then $a \in \overline{xy} = \text{Conv}(x, y)$ for $x \in C$ and $y \in D$, and $b \in \text{Conv}(z, w)$ for $z \in C$ and $w \in D$.

Then any convex combination of a and b lies in

$$\begin{aligned} \text{Conv}(x, y) \cdot \text{Conv}(z, w) &= \text{Conv}(x, y, z, w) \\ &= \text{Conv}(x, z) \cdot \text{Conv}(y, w) \subset C \cdot D. \end{aligned} \quad \square$$

2.8.3. Affine maps. Up until now, affine maps have been considered as taking value in \mathbb{R}^m . It is valuable to allow their codomain to be an affine subspace of \mathbb{R}^m .²

Definition 2.8.17. Let $C \subset \mathbb{R}^n$ be convex and let $K \subset \mathbb{R}^m$ be an affine subspace. A map $f : C \rightarrow K$ is affine if the composite $C \xrightarrow{f} K \subset \mathbb{R}^m$ is affine. This simply means $f((1-t)x + ty) = (1-t)f(x) + tf(y)$ for $x, y \in C$ and $t \in [0, 1]$.

If H is an affine subspace of \mathbb{R}^n , then an affine isomorphism $f : H \rightarrow K$ is a bijective affine map. An affine isomorphism $f : H \rightarrow H$ is called an affine automorphism of H .

The following makes it easy to study affine maps.

Lemma 2.8.18. Let $H \subset \mathbb{R}^n$ and $K \subset \mathbb{R}^m$ be affine subspaces and let $x \in H$. Let V and W be the linear bases of H and K , respectively. Then a map $f : H \rightarrow K$ is affine if and only if the composite

$$V \xrightarrow{\tau_x} H \xrightarrow{f} K \xrightarrow{\tau_{-f(x)}} W$$

is linear. In particular, we get a commutative diagram

$$(2.8.10) \quad \begin{array}{ccc} H & \xrightarrow{f} & K \\ \tau_x \uparrow \cong & & \tau_{f(x)} \uparrow \cong \\ V & \xrightarrow{g} & W, \end{array}$$

where $g = \tau_{-f(x)} \circ f \circ \tau_x$. Here, the vertical maps are bijective translations, and the map g is linear if and only if f is affine.

Proof. Translations are isometries and therefore are affine (a direct proof that translations are affine is easy). Composites of affine maps are affine. Linear maps are obviously affine. By Proposition 2.5.1, an affine map between vector spaces that takes 0 to 0 is linear. So the result follows. \square

Thus, we can deduce the properties of affine functions from those of translations and linear maps. Since the vertical maps in (2.8.10) are bijective, we obtain the following.

Corollary 2.8.19. Let $f : H \rightarrow K$ be an affine isomorphism. Then the linear map $g : V \rightarrow W$ between their linear bases in (2.8.10) is a linear isomorphism. Thus, H and K have the same dimension and the inverse map $f^{-1} : K \rightarrow H$ is affine.

The following very simple consequence of Lemma 2.8.18 bears mention.

²If we write $f : X \rightarrow Y$, then Y is the codomain of f . If $Z \subset Y$ and if the image of f is contained in Z , we may regard f as being a map from X to Z . Writing $f : X \rightarrow Z$ amounts to restricting the codomain of f to Z .

Corollary 2.8.20. *Let $H \subset \mathbb{R}^n$ and $K \subset \mathbb{R}^m$ be affine subspaces and let $f : H \rightarrow K$ be affine. Then $f(H)$ is an affine subspace of \mathbb{R}^m (and hence of K).*

Proof. For $x \in H$, $f(H) = \tau_{f(x)}(g(V))$. □

Affine functions are very useful for studying convex and affine hulls:

Proposition 2.8.21. *Let $f : H \rightarrow K$ be an affine function between affine subspaces of \mathbb{R}^n and \mathbb{R}^m , respectively. Then f respects affine combinations of points in H , i.e., if $x_1, \dots, x_k \in H$ and $\sum_{i=1}^k a_i = 1$, then*

$$(2.8.11) \quad f(a_1x_1 + \dots + a_kx_k) = a_1f(x_1) + \dots + a_kf(x_k).$$

Thus,

$$\begin{aligned} f(\text{Aff}(x_1, \dots, x_k)) &= \text{Aff}(f(x_1), \dots, f(x_k)), \\ f(\text{Conv}(x_1, \dots, x_k)) &= \text{Conv}(f(x_1), \dots, f(x_k)). \end{aligned}$$

Proof. (2.8.11) is certainly true if f is linear, so by Lemma 2.8.18, it suffices to assume f is a translation, say $f = \tau_x$. Since $\sum_{i=1}^k a_i = 1$,

$$\begin{aligned} \tau_x(a_1x_1 + \dots + a_kx_k) &= (a_1x_1 + \dots + a_kx_k) + (a_1x + \dots + a_kx) \\ &= a_1(x_1 + x) + \dots + a_k(x_k + x) \\ &= a_1\tau_x(x_1) + \dots + a_k\tau_x(x_k). \end{aligned} \quad \square$$

We immediately obtain the following useful tool for analyzing convex and affine hulls of finite sets.

Corollary 2.8.22. *Let $f : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be the linear map with $f(e_i) = x_i$ for $i = 1, \dots, k$. Then*

$$\begin{aligned} f(\text{Aff}(e_1, \dots, e_k)) &= \text{Aff}(x_1, \dots, x_k), \\ f(\Delta^{k-1}) &= \text{Conv}(x_1, \dots, x_k). \end{aligned}$$

In particular, the convex hull of k points in \mathbb{R}^n is the image of the standard $(k-1)$ -simplex under a linear (hence affine) map from \mathbb{R}^k to \mathbb{R}^n .

We also obtain the following.

Corollary 2.8.23. *Let $X = \{x_1, \dots, x_k\} \subset \mathbb{R}^n$ and let $H = \text{Aff}(X)$. Let K be an affine subspace of \mathbb{R}^m . Then the affine maps from H to K are determined by their restriction to X , i.e., if f and g are affine maps from H to K that agree on X , then $f = g$.*

Proof. The elements of H all have the form $\sum_{i=1}^k a_i x_i$ with $\sum_{i=1}^k a_i = 1$. By (2.8.11), if $f : H \rightarrow K$ is affine, then $f(\sum_{i=1}^k a_i x_i)$ is determined by $f(x_1), \dots, f(x_k)$. □

Proposition 2.8.21 allows us to analyze products of convex sets.

Definition 2.8.24. We identify \mathbb{R}^{n+k} with $\mathbb{R}^n \times \mathbb{R}^k$, writing $\begin{bmatrix} x \\ y \end{bmatrix}$ for the generic element of \mathbb{R}^{n+k} with $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^k$. With this convention, given $C \subset \mathbb{R}^n$ and $D \subset \mathbb{R}^k$, we write

$$(2.8.12) \quad C \times D = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} : x \in C, y \in D \right\} \subset \mathbb{R}^{n+k}.$$

The reader may easily check that if C and D are convex, so is $C \times D$.

The following is useful.

Proposition 2.8.25. *Let $X = \{x_1, \dots, x_\ell\} \subset \mathbb{R}^n$ and $Y = \{y_1, \dots, y_m\} \subset \mathbb{R}^k$. Then*

$$(2.8.13) \quad \begin{aligned} \text{Conv}(X) \times \text{Conv}(Y) &= \text{Conv}(X \times Y) \\ &= \text{Conv} \left(\begin{bmatrix} x_i \\ y_j \end{bmatrix} : 1 \leq i \leq \ell, 1 \leq j \leq m \right). \end{aligned}$$

Proof. We claim that given convex combinations $\sum_{i=1}^{\ell} a_i x_i \in \text{Conv}(X)$ and $\sum_{j=1}^m b_j y_j \in \text{Conv}(Y)$,

$$(2.8.14) \quad \begin{bmatrix} \sum_{i=1}^{\ell} a_i x_i \\ \sum_{j=1}^m b_j y_j \end{bmatrix} = \sum_{\substack{i=1, \dots, \ell \\ j=1, \dots, m}} a_i b_j \begin{bmatrix} x_i \\ y_j \end{bmatrix}.$$

Note that this claim is sufficient to prove the proposition, as

$$\sum_{i=1}^{\ell} \sum_{j=1}^m a_i b_j = \sum_{i=1}^{\ell} a_i \left(\sum_{j=1}^m b_j \right) = \sum_{i=1}^{\ell} a_i = 1,$$

so the claim implies $\text{Conv}(X) \times \text{Conv}(Y) \subset \text{Conv}(X \times Y)$. The opposite inclusion is immediate from Proposition 2.8.12, as $\text{Conv}(X) \times \text{Conv}(Y)$ is convex.

To prove the claim, note that for $z \in \mathbb{R}^k$,

$$\begin{bmatrix} \sum_{i=1}^{\ell} a_i x_i \\ z \end{bmatrix} = \tau_{\begin{bmatrix} 0 \\ z \end{bmatrix}} \left(\sum_{i=1}^{\ell} a_i \begin{bmatrix} x_i \\ 0 \end{bmatrix} \right) = \sum_{i=1}^{\ell} a_i \tau_{\begin{bmatrix} 0 \\ z \end{bmatrix}} \left(\begin{bmatrix} x_i \\ 0 \end{bmatrix} \right) = \sum_{i=1}^{\ell} a_i \begin{bmatrix} x_i \\ z \end{bmatrix}$$

by Proposition 2.8.21. Similarly, for $w \in \mathbb{R}^n$,

$$\begin{bmatrix} w \\ \sum_{j=1}^m b_j y_j \end{bmatrix} = \sum_{j=1}^m b_j \begin{bmatrix} w \\ y_j \end{bmatrix}.$$

The result follows. \square

A useful special case of this is for $Y = \{a, b\} \subset \mathbb{R}$ with $a < b$ so that $\text{Conv}(Y) = [a, b]$. E.g., if $Y = \{0, 1\}$, $\text{Conv}(Y)$ is the unit interval $I = [0, 1]$.

Corollary 2.8.26. *Let $X = \{x_1, \dots, x_k\} \subset \mathbb{R}^n$ and let $a < b \in \mathbb{R}$. Then*

$$(2.8.15) \quad \text{Conv}(X) \times [a, b] = \text{Conv} \left(\begin{bmatrix} x_1 \\ a \end{bmatrix}, \dots, \begin{bmatrix} x_k \\ a \end{bmatrix}, \begin{bmatrix} x_1 \\ b \end{bmatrix}, \dots, \begin{bmatrix} x_k \\ b \end{bmatrix} \right).$$

We may iterate this to study $\text{Conv}(X) \times I^2 = \text{Conv}(X) \times I \times I \subset \mathbb{R}^{n+2}$, or more generally $\text{Conv}(X) \times I^k \subset \mathbb{R}^{n+k}$, where I^k is the product of k copies of the unit interval $I = [0, 1]$. Or, starting with $X = \{0, 1\}$, we may study the convex subset $I^n \subset \mathbb{R}^n$. These polytopes are important enough to merit a name.

Definition 2.8.27. The standard n -cube $I^n \subset \mathbb{R}^n$ is

$$[0, 1]^n = \{a_1 e_1 + \cdots + a_n e_n : a_1, \dots, a_n \in [0, 1]\}.$$

We obtain the following from Corollary 2.8.26 by induction on n .

Corollary 2.8.28. Let $S \subset \mathbb{R}^n$ be the set of vectors whose coordinates are all either 0 or 1:

$$(2.8.16) \quad S = \{\epsilon_1 e_1 + \cdots + \epsilon_n e_n : \epsilon_1, \dots, \epsilon_n \in \{0, 1\}\}.$$

Then $I^n = \text{Conv}(S)$.

Moreover, every element of S is a sum of canonical basis vectors: for $v \in S$, the coordinates of v are all either 0 or 1. Let i_1, \dots, i_k be the coordinates of v that are nonzero, with $i_1 < \cdots < i_k$. Then $v = e_{i_1} + \cdots + e_{i_k}$ (if none of the coordinates of v are nonzero, then $k = 0$ and v is the origin). Thus,

$$(2.8.17) \quad S = \{e_{i_1} + \cdots + e_{i_k} : 1 \leq i_1 < \cdots < i_k \leq n \text{ and } 0 \leq k \leq n\}.$$

An induction argument like that given in Proposition 2.8.12 allows us to strengthen Proposition 2.8.21.

Proposition 2.8.29. Let $C \subset \mathbb{R}^k$ be convex and let $f : C \rightarrow \mathbb{R}^n$ be affine. Then f respects convex combinations of points in C : if $x_1, \dots, x_k \in C$ and $\sum_{i=1}^k a_i = 1$ with $a_i \geq 0$ for all i , then

$$(2.8.18) \quad f(a_1 x_1 + \cdots + a_k x_k) = a_1 f(x_1) + \cdots + a_k f(x_k).$$

Thus, $f(\text{Conv}(x_1, \dots, x_k)) = \text{Conv}(f(x_1), \dots, f(x_k))$.

Proof. We argue by induction on k , with the case $k = 2$ being immediate from the definition of convex functions. For the inductive step we apply f to (2.8.4). \square

It is easy to extend affine mappings on Δ^{k-1} to affine mappings on $\text{Aff}(e_1, \dots, e_k)$.

Lemma 2.8.30. Let $f : \Delta^{k-1} \rightarrow \mathbb{R}^n$ be an affine map. Then f extends to a unique affine map

$$\hat{f} : \text{Aff}(e_1, \dots, e_k) \rightarrow \text{Aff}(f(e_1), \dots, f(e_k))$$

and to a unique linear map $\hat{f} : \mathbb{R}^k \rightarrow \mathbb{R}^n$. Indeed, these extensions are specified by

$$(2.8.19) \quad \hat{f}(a_1 e_1 + \cdots + a_k e_k) = a_1 f(e_1) + \cdots + a_k f(e_k).$$

Proof. (2.8.19) specifies the unique linear map from \mathbb{R}^k to \mathbb{R}^m that agrees with f on the vertices of Δ^{k-1} . It extends f by Proposition 2.8.29. Its restriction to $\text{Aff}(e_1, \dots, e_k)$ is affine because the inclusion of $\text{Aff}(e_1, \dots, e_k)$ is affine and the composite of affine maps is affine. This is the unique affine mapping on $\text{Aff}(e_1, \dots, e_k)$ extending f by Proposition 2.8.21. \square

2.8.4. Affine and convex hulls of infinite sets. Not every convex set is the convex hull of a finite set of points. For instance, the closed unit disk

$$\mathbb{D}^n = \{x \in \mathbb{R}^n : \|x\| \leq 1\}$$

is convex, but cannot be expressed as the convex hull of a finite set because of the curvature of its boundary. The convex hull of a finite set always has vertices, edges, faces, etc., and there are no such phenomena for the disk.

For this and similar examples, it is useful to generalize the notion of convex hull. We shall see that \mathbb{D}^n is the convex hull of the infinite set \mathbb{S}^{n-1} , the unit sphere in \mathbb{R}^n :

$$\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}.$$

Definition 2.8.31. Let $\emptyset \neq X \subset \mathbb{R}^n$. The convex hull $\text{Conv}(X)$ is the union of the convex hulls of the finite subsets of X . The affine hull $\text{Aff}(X)$ is the union of the affine hulls of the finite subsets of X .³

Proposition 2.8.32.

- (1) $\text{Conv}(X)$ is the smallest convex set containing X .
- (2) For any $x \in X$,

$$\text{Aff}(X) = \tau_x(\text{span}(\{y - x : y \in X\}))$$

is the smallest affine subspace containing X .

Proof. By Proposition 2.8.12, $\text{Conv}(X)$ is contained in any convex subset containing X , so (1) follows if we show $\text{Conv}(X)$ to be convex. Let $x, y \in \text{Conv}(X)$. Then there are finite subsets S and T of X such that $x \in \text{Conv}(S)$ and $y \in \text{Conv}(T)$. But then $\overline{xy} \subset \text{Conv}(S \cup T) \subset \text{Conv}(X)$ by Proposition 2.8.12.

(2) follows from Proposition 2.8.11 and the behavior of spans. Note that since \mathbb{R}^n is finite-dimensional there is a finite set $\{y_1, \dots, y_k\} \subset X$ such that

$$\text{span}(\{y - x : y \in X\}) = \text{span}(y_1 - x, \dots, y_k - x). \quad \square$$

The minimality conditions in Proposition 2.8.32 give us the following. Note that $\text{Conv}(X) \subset \text{Aff}(X)$ by (1), as affine subspaces are convex.

Corollary 2.8.33. If $\emptyset \neq X \subset Y \subset \text{Conv}(X)$, then $\text{Conv}(X) = \text{Conv}(Y)$. In particular, $\text{Conv}(\text{Conv}(X)) = \text{Conv}(X)$.

If $X \subset Y \subset \text{Aff}(X)$, then $\text{Aff}(X) = \text{Aff}(Y)$. In particular,

$$\text{Aff}(\text{Aff}(X)) = \text{Aff}(\text{Conv}(X)) = \text{Aff}(X).$$

³While it makes sense to define $\text{Conv}(\emptyset)$ to be \emptyset , $\text{Aff}(\emptyset)$ makes no sense, and \emptyset does not have a well-defined dimension.

These notions allow us to define the dimension of an arbitrary convex set.

Definition 2.8.34. Let $\emptyset \neq C \subset \mathbb{R}^n$ be convex. We define the dimension of C to be the dimension (as an affine subspace) of its affine hull.

Remark 2.8.35. Note that if the convex set C contains at least two points, say x and y , then $\text{Aff}(C)$ contains the affine line $\overleftrightarrow{xy} = \text{Aff}(x, y)$, and hence $\dim C \geq 1$. So a convex set is zero-dimensional if and only if it consists of a single point.

2.8.5. Convex subsets of lines.

Examples 2.8.36. There are some obvious examples of convex subsets of \mathbb{R} :

- (1) closed intervals $[a, b]$ for $a < b \in \mathbb{R}$;
- (2) half-open intervals $(a, b]$ or $[a, b)$ with $a, b \in \mathbb{R}$;
- (3) open intervals (a, b) with $a, b \in \mathbb{R}$;
- (4) open intervals $(-\infty, a)$ or (a, ∞) for $a \in \mathbb{R}$;
- (5) half-open intervals $(-\infty, a]$ or $[a, \infty)$ for $a \in \mathbb{R}$;
- (6) \mathbb{R} ;
- (7) single points $\{a\}$;
- (8) \emptyset .

The reader should check that these types, as listed, are preserved by the affine automorphisms of \mathbb{R} .

Proposition 2.8.37. *The convex subsets of \mathbb{R} are exactly the ones listed in Examples 2.8.36.*

Proof. Disposing of the trivial cases, we assume $C \subset \mathbb{R}$ is a convex subset with at least two points and unequal to all of \mathbb{R} .

Suppose C bounded above and let $b = \sup(C)$.

Case 1. $b \in C$. Then for any $a \in C$, the line segment $[a, b] \subset C$. If C is not bounded below, C must equal $(-\infty, b]$. If C is bounded below, let $a = \inf(C)$. If $a \in C$, then $C = [a, b]$. If $a \notin C$, then if $0 < \epsilon < b - a$, there is a $c \in C$ with $a < c < a + \epsilon$. Since $[c, b] \subset C$ and ϵ may be arbitrarily small, $C = (a, b]$.

Case 2. $b \notin C$. For $\epsilon > 0$, there exists $c \in C$ with $b - \epsilon < c < b$. By the argument just given, $C \cap (-\infty, c]$ must have the form $(a, c]$ for $a \in (-\infty, c)$ or the form $[a, c]$ for $a \in (-\infty, c]$. Since this is true for any ϵ , C must have the form (a, b) for $a \in (-\infty, b)$ or the form $[a, b)$ for $a \in (-\infty, b)$.

If C is not bounded above, then for any $c \in C$, $[c, \infty) \subset C$. Since $C \neq \mathbb{R}$, it is bounded below, and is covered by a reversal of the above arguments. \square

Since affine maps carry convex sets to convex sets and since every line in \mathbb{R}^n is affinely isomorphic to \mathbb{R} this allows us to classify all the convex subsets of an arbitrary line in \mathbb{R}^n . We start with some notation.

Notation 2.8.38. Let $x, y \in \mathbb{R}^n$. We shall use $[x, y]$ as an alternate notation for the line segment \overline{xy} . When $x \neq y$ we shall also write

$$(2.8.20) \quad [x, y] = \{(1-t)x + ty : t \in [0, 1]\} = \overline{xy} \setminus \{y\}.$$

We call this a half-open segment. We shall also write

$$(2.8.21) \quad (x, y) = \{(1-t)x + ty : t \in (0, 1)\} = \overline{xy} \setminus \{x, y\},$$

and call it an open segment.

Corollary 2.8.39. Let ℓ be a line in \mathbb{R}^n . Then any convex subset of ℓ has one of the following forms:

- (1) a line segment $\overline{xy} = [x, y]$ for $x \neq y \in \ell$;
- (2) a half-open segment $[x, y)$ for $x \neq y \in \ell$;
- (3) an open segment (x, y) for $x \neq y \in \ell$;
- (4) an “open ray” $\overrightarrow{xy} \setminus \{x\}$ for $x \neq y \in \ell$;
- (5) a ray \overrightarrow{xy} for $x \neq y \in \ell$;
- (6) ℓ ;
- (7) a single point $x \in \ell$;
- (8) \emptyset .

2.9. Affine independence, interiors and faces.

2.9.1. Affine independence.

Definition 2.9.1. $x_1, \dots, x_k \in \mathbb{R}^n$ are affinely independent if

$$a_1x_1 + \dots + a_kx_k = 0 \quad \text{with} \quad \sum_{i=1}^k a_i = 0 \quad \Rightarrow \quad a_i = 0 \quad \text{for all } i.$$

The following is immediate from the definitions.

Lemma 2.9.2. If x_1, \dots, x_k are linearly independent, then they are affinely independent. In particular, e_1, \dots, e_n are affinely independent.

The converse to Lemma 2.9.2 is false. Indeed, $0, 1 \in \mathbb{R}$ are affinely independent in \mathbb{R} , but are not linearly independent.

Affine independence is important in understanding convex hulls of finite sets.

Proposition 2.9.3. Let $x_1, \dots, x_k \in \mathbb{R}^n$ and let $f : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be the linear map with $f(e_i) = x_i$ for $i = 1, \dots, k$. Then the diagram (2.8.10) becomes

$$(2.9.1) \quad \begin{array}{ccc} \text{Aff}(e_1, \dots, e_k) & \xrightarrow{f} & \text{Aff}(x_1, \dots, x_k) \\ \tau_{e_1} \uparrow \cong & & \tau_{x_1} \uparrow \cong \\ \text{span}(e_2 - e_1, \dots, e_k - e_1) & \xrightarrow{f} & \text{span}(x_2 - x_1, \dots, x_k - x_1), \end{array}$$

We deduce that the following conditions are equivalent:

- (1) x_1, \dots, x_k are affinely independent.

- (2) f restricts to a bijection from $\text{Aff}(e_1, \dots, e_k)$ onto $\text{Aff}(x_1, \dots, x_k)$.
 (3) $x_2 - x_1, \dots, x_k - x_1$ are linearly independent.
 (4) $\text{Aff}(x_1, \dots, x_k)$ has dimension $k - 1$ as an affine subspace.

Proof. The horizontal maps in (2.8.10) are both f because $f : \mathbb{R}^k \rightarrow \mathbb{R}^n$ is linear: $f \circ \tau_{e_1} = \tau_{f(e_1)} \circ f$.

(1) \Rightarrow (2). If $\sum_{i=1}^k a_i x_i = \sum_{i=1}^k b_i x_i$ with $\sum_{i=1}^k a_i = \sum_{i=1}^k b_i = 1$, then $\sum_{i=1}^k (a_i - b_i) x_i = 0$, so $a_i - b_i = 0$ for all i by (1).

(2) \Rightarrow (3). The assumption (2) is that the upper horizontal map in (2.9.1) is a bijection. Since the vertical maps are also bijective, the lower horizontal map is not only bijective but also linear, and hence a linear isomorphism. Since $e_2 - e_1, \dots, e_k - e_1$ are linearly independent and $f(e_i - e_1) = x_i - x_1$, (3) follows.

(3) \Leftrightarrow (4). $\text{span}(x_2 - x_1, \dots, x_k - x_1)$ has dimension $k - 1$ if and only if the $k - 1$ generators $x_2 - x_1, \dots, x_k - x_1$ are linearly independent.

(3) \Rightarrow (1). If $a_1 x_1 + \dots + a_k x_k = 0$ with $\sum_{i=1}^k a_i = 0$, then $\sum_{i=2}^k a_i = -a_1$, so

$$\sum_{i=2}^k a_i (x_i - x_1) = \sum_{i=2}^k a_i x_i - \left(\sum_{i=2}^k a_i \right) x_1 = \sum_{i=1}^k a_i x_i = 0,$$

so $a_i = 0$ for $i \geq 2$ by linear independence, which then forces $a_1 = 0$. \square

Example 2.9.4. Any two distinct points $x, y \in \mathbb{R}^n$ are affinely independent as $\text{Aff}(x, y)$ is a line (Example 2.8.4), and hence is 1-dimensional.

Definition 2.9.5. A set of points in \mathbb{R}^n is collinear if there is a line containing all of them.

Proposition 2.9.6. Three distinct points $x, y, z \in \mathbb{R}^n$ are affinely independent if and only if they are not collinear.

Proof. If x, y and z are contained in the line ℓ , then $\text{Aff}(x, y, z) \subset \ell$, so $\dim \text{Aff}(x, y, z) \leq 1$. So they cannot be affinely independent.

Conversely, if they are affinely dependent, we can find $a, b, c \in \mathbb{R}$, not all 0, with $ax + by + cz = 0$ and $a + b + c = 0$. Suppose $c \neq 0$. then $-c = a + b$, and

$$ax + by = -cz = (a + b)z,$$

so

$$z = \frac{a}{a+b}x + \frac{b}{a+b}y \in \overleftrightarrow{xy}. \quad \square$$

Example 2.9.7. $0, e_1, \dots, e_n \in \mathbb{R}^n$ are affinely independent: if

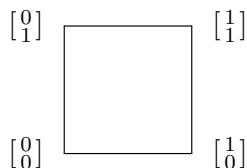
$$a_0 \cdot 0 + a_1 e_1 + \dots + a_n e_n = 0$$

with $\sum_{i=0}^n a_i = 0$, then $a_1 = \dots = a_n = 0$ by the linear independence of e_1, \dots, e_n , and hence $a_0 = -\sum_{i=1}^n a_i$ is 0 as well. Note that

$$\text{Aff}(0, e_1, \dots, e_n) = \mathbb{R}^n.$$

Example 2.9.8. Let I^2 be the unit square in \mathbb{R}^2 . By Corollary 2.8.28,

$$I^2 = \text{Conv}(0, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}) = \text{Conv}(0, e_1, e_2, e_1 + e_2) :$$



As shown in Example 2.9.7,

$$\text{Aff}(0, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}) = \mathbb{R}^2$$

is 2-dimensional (and hence I^2 is 2-dimensional), so these vertices cannot be affinely independent. By Proposition 2.9.6, any three of the vertices are affinely independent.

The same reasoning shows the following.

Example 2.9.9. By Corollary 2.8.28, the n -cube $I^n \subset \mathbb{R}^n$ is the convex hull of the 2^n points whose coordinates are all equal to either 0 or 1. As these points include the affinely independent set $0, e_1, \dots, e_n$, their affine span is \mathbb{R}^n , and hence the n -cube is n -dimensional. But for $n > 1$, $2^n > n + 1$, so the 2^n points in question cannot be affinely independent.

Proposition 2.9.3 gives us a useful calculating tool. Recall that the dimension of a convex set C is set equal to the dimension of its affine hull (as an affine subspace).

Corollary 2.9.10. *Let $\emptyset \neq C \subset \mathbb{R}^n$ be convex. Then the dimension of C is the largest integer k such that C contains an affinely independent set with $k + 1$ points. Moreover, if $C = \text{Conv}(x_1, \dots, x_m)$, we may take those $k + 1$ points to be in $\{x_1, \dots, x_m\}$.*

Proof. Let $H = \text{Aff}(C)$ and let V be its linear base. Let $x \in C$. By Proposition 2.8.32, $V = \text{span}(\{y - x : y \in C\})$, so if $k = \dim V$, then a maximal linear independent subset of V has the form $y_1 - x, \dots, y_k - x$ for $y_1, \dots, y_k \in C$. By Proposition 2.9.3, x, y_1, \dots, y_k are affinely independent.

Conversely, if $x, y_1, \dots, y_k \in C$ are affinely independent, then

$$\text{Aff}(x, y_1, \dots, y_k) \subset \text{Aff}(C)$$

has dimension k , so $\dim C \geq k$.

Finally, if $C = \text{Conv}(x_1, \dots, x_m)$, then take $x = x_1$. We have

$$V = \text{span}(x_2 - x_1, \dots, x_m - x_1).$$

Apply the previous argument. □

Indeed, since affine subspaces are convex, we obtain the following.

Corollary 2.9.11. *Let H be an affine subspace of \mathbb{R}^n . Then the dimension of H is the largest k for which there exists an affinely independent subset of H with $k + 1$ elements. Moreover, if $x_1, \dots, x_{k+1} \in H$ are affinely independent, then $H = \text{Aff}(x_1, \dots, x_{k+1})$.*

Proof. For the last statement, note that if $x_1, \dots, x_{k+1} \in H$ are affinely independent, then $\text{Aff}(x_1, \dots, x_{k+1})$ is a k -dimensional affine subspace of H , and therefore must be all of H by Corollary 2.8.9. \square

The results above assemble to something useful:

Proposition 2.9.12. *Let H be an affine subspace of \mathbb{R}^n with $\dim H = k$. Let $x_1, \dots, x_{k+1} \in H$ be affinely independent. Then:*

- (1) $H = \text{Aff}(x_1, \dots, x_{k+1})$.
- (2) If K is an affine subspace of \mathbb{R}^m , then the affine maps from H to K are determined by their restriction to $\{x_1, \dots, x_{k+1}\}$.
- (3) Any function $f : \{x_1, \dots, x_{k+1}\} \rightarrow K$ extends to an affine map $\bar{f} : H \rightarrow K$.
- (4) If K also has dimension k , then \bar{f} is an affine isomorphism if and only if $f(x_1), \dots, f(x_{k+1})$ are affinely independent.

Thus, there is a one-to-one correspondence between the affine maps from H to K and the functions from $\{x_1, \dots, x_{k+1}\}$ to K .⁴

Proof. (1) is just Corollary 2.9.11, and (2) follows from (1) by Corollary 2.8.23.

To prove (3), let $g : \mathbb{R}^{k+1} \rightarrow \mathbb{R}^n$ with $g(e_i) = x_i$ for $i = 1, \dots, k + 1$. By Proposition 2.9.3, g restricts to an affine isomorphism

$$g_1 : \text{Aff}(e_1, \dots, e_{k+1}) \xrightarrow{\cong} \text{Aff}(x_1, \dots, x_{k+1}).$$

Let $h : \mathbb{R}^{k+1} \rightarrow \mathbb{R}^m$ be the linear map with $h(e_i) = f(x_i)$ for $i = 1, \dots, k + 1$. Since $\text{Aff}(f(x_1), \dots, f(x_{k+1})) \subset K$, h restricts to an affine map

$$h_1 : \text{Aff}(e_1, \dots, e_{k+1}) \rightarrow K.$$

We have a diagram

$$\text{Aff}(x_1, \dots, x_{k+1}) \xleftarrow[\cong]{g_1} \text{Aff}(e_1, \dots, e_{k+1}) \xrightarrow{h_1} K$$

Since g_1 is an affine isomorphism, its inverse function g_1^{-1} is affine. We obtain an affine map

$$\bar{f} = h_1 \circ g_1^{-1} : \text{Aff}(x_1, \dots, x_{k+1}) \rightarrow K$$

that agrees with f on each x_i .

For (4), note that \bar{f} is an affine isomorphism if and only if h_1 is. But if $\dim K = k$, this is equivalent to $f(x_1), \dots, f(x_{k+1})$ being affinely independent. \square

⁴In the language of category theory this says H is the free affine space on $\{x_1, \dots, x_{k+1}\}$.

Since isometries are affine maps, we obtain the following.

Corollary 2.9.13. *Let x, y and z be noncollinear points in \mathbb{R}^2 then any two isometries of \mathbb{R}^2 that agree on x, y and z must be equal.*

Proof. x, y, z are affinely independent by Proposition 2.9.6. Now apply Proposition 2.9.12 with $H = K = \mathbb{R}^2$. \square

Of course, not every affine isomorphism from \mathbb{R}^2 to \mathbb{R}^2 is an isometry, so not every function $f : \{x, y, z\} \rightarrow \mathbb{R}^2$ with $f(x), f(y), f(z)$ noncollinear extends to an isometry. Indeed, at minimum, $f : \{x, y, z\} \rightarrow \{f(x), f(y), f(z)\}$ must be distance-preserving.

2.9.2. Interiors. We can now use Proposition 2.9.3 to show that $\dim C$ gives a reasonable notion for the topological dimension of C . We first need to develop a little elementary topology.

Definition 2.9.14. Let $\emptyset \neq C$ be a convex subset of \mathbb{R}^n . Let $H = \text{Aff}(C)$ and let V be its linear base. For $\epsilon > 0$, let $B_\epsilon(0, V)$ be the open ball of radius epsilon in V :

$$B_\epsilon(0, V) = \{v \in V : \|v\| < \epsilon\}.$$

We say that x is an interior point of C , written $x \in \text{Int}(C)$, if there exists $\epsilon > 0$ such that

$$(2.9.2) \quad x + B_\epsilon(0, V) \subset C.$$

Here,

$$(2.9.3) \quad x + B_\epsilon(0, V) = \tau_x(B_\epsilon(0, V)) = \{x + v : v \in B_\epsilon(0, V)\}.$$

We define the boundary of C by $\partial C = C \setminus \text{Int}(C)$. This is mainly of interest when C is what's known as a closed subspace of $\text{Aff}(C)$.

Example 2.9.15. If $C = \{x\}$, then $\text{Aff}(C) = \{x\}$, and its linear base is the trivial subspace 0 . $B_\epsilon(0, 0) = \{0\}$ for all $\epsilon > 0$, so $\text{Int}(\{x\}) = \{x\}$.

The following elementary observation shows that $\text{Int}(C)$ coincides with what's known as the topological interior of C when viewed as a subspace of $\text{Aff}(C)$.

Lemma 2.9.16. *Let $\emptyset \neq C$ be a convex subset of \mathbb{R}^n . Let $H = \text{Aff}(C)$ and let V be its linear base. Then for $x \in H$,*

$$(2.9.4) \quad \tau_x(B_\epsilon(0, V)) = B_\epsilon(x, H) = \{y \in H : d(x, y) < \epsilon\}.$$

Thus, $x \in \text{Int}(C)$ if and only if $B_\epsilon(x, H) \subset C$ for some $\epsilon > 0$.

Proof. $\tau_x : V \rightarrow \text{Aff}(C)$ is a bijective isometry. \square

In some cases, every point of C is an interior point. For instance, $B_\epsilon(0, V)$ is convex and is equal to its own interior (by the triangle inequality). Here, $\text{Aff}(B_\epsilon(0, V)) = V$. Similarly, a linear or affine subspace is its own interior.

Another important example is the standard simplex Δ^{n-1} .

Lemma 2.9.17. *Let $n \geq 1$. The linear base of $\text{Aff}(e_1, \dots, e_n)$ is*

$$(2.9.5) \quad V = \left\{ a_1 e_1 + \dots + a_n e_n : \sum_{i=1}^n a_i = 0 \right\} = \{x \in \mathbb{R}^n : \langle x, \xi \rangle = 0\},$$

where $\xi = e_1 + \dots + e_n$. As V is the nullspace of the $1 \times n$ matrix ξ^T , it has dimension $n - 1$, as does Δ^{n-1} .

The interior of Δ^{n-1} is the set of points in Δ^{n-1} whose coordinates are all positive:

$$(2.9.6) \quad \text{Int}(\Delta^{n-1}) = \left\{ a_1 e_1 + \dots + a_n e_n : \sum_{i=1}^n a_i = 1 \text{ and } a_i > 0 \text{ for all } i \right\}.$$

Proof. $\text{Aff}(e_1, \dots, e_n) = \{x \in \mathbb{R}^n : \langle x, \xi \rangle = 1\}$. By the bilinearity of the inner product, if $\langle x, \xi \rangle = 1$, then $\langle y, \xi \rangle = 1$ if and only if $y = x + v$ with $v \in V$, i.e., $\text{Aff}(e_1, \dots, e_n) = \tau_x(V)$. So V is the linear base, as claimed.

Let

$$U = \left\{ a_1 e_1 + \dots + a_n e_n : \sum_{i=1}^n a_i = 1 \text{ and } a_i > 0 \text{ for all } i \right\}$$

and let $x = x_1 e_1 + \dots + x_n e_n \in U$. Let $\epsilon = \min(x_1, \dots, x_n)$ and let $v = a_1 e_1 + \dots + a_n e_n \in B_\epsilon(0, V)$. Then $|a_i| < \epsilon$ for all i , so the coordinates of $x + v$ are all positive. Since $x + v \in \text{Aff}(e_1, \dots, e_n)$, it must lie in $U \subset \Delta^{n-1}$. Thus, $U \subset \text{Int}(\Delta^{n-1})$.

Conversely, if $x = x_1 e_1 + \dots + x_n e_n \in \Delta^{n-1}$ with $x_i = 0$, let $j \neq i$. Let $v_\delta = -\delta e_i + \delta e_j \in \text{Aff}(e_1, \dots, e_n)$. Then $x + v_\delta \in \mathbb{R}^n \setminus \Delta^{n-1}$ for all $\delta > 0$. Since

$$\|v_\delta\| = \delta \|e_j - e_i\| = \delta \sqrt{2}$$

may be made arbitrarily small, $x \in \partial \Delta^{n-1}$. \square

The same reasoning gives us the following.

Lemma 2.9.18. *The interior of the unit n -cube $I^n = [0, 1]^n \subset \mathbb{R}^n$ is $(0, 1)^n$, i.e., the set of points in \mathbb{R}^n whose coordinates are all positive and less than 1.*

As shown in Example 2.9.9, $\text{Aff}(I^n) = \mathbb{R}^n$, so its linear base is \mathbb{R}^n as well.

Proof. Let $x = a_1 e_1 + \dots + a_n e_n \in (0, 1)^n$. Let

$$\epsilon = \min(a_1, \dots, a_n, 1 - a_1, \dots, 1 - a_n).$$

Then as in Lemma 2.9.17 we see $x + B_\epsilon(0, \mathbb{R}^n) \subset I^n$. The remainder of the argument is analogous to that in Lemma 2.9.17. \square

Let $f : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be the linear map with $f(e_i) = x_i$ for $i = 1, \dots, k$. Proposition 2.9.3 shows that if x_1, \dots, x_k are affinely independent, then $f : \Delta^{k-1} \rightarrow \text{Conv}(x_1, \dots, x_n)$ is bijective. Lemma 2.9.17 allows us to prove the converse.

Proposition 2.9.19. *Let $f : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be the linear map with $f(e_i) = x_i$ for $i = 1, \dots, k$. Then x_1, \dots, x_k are affinely independent if and only if $f : \Delta^{k-1} \rightarrow \text{Conv}(x_1, \dots, x_n)$ is bijective.*

Proof. The maps

$$(2.9.7) \quad f : \text{Aff}(e_1, \dots, e_k) \rightarrow \text{Aff}(x_1, \dots, x_k)$$

$$(2.9.8) \quad f : \text{Conv}(e_1, \dots, e_k) \rightarrow \text{Conv}(x_1, \dots, x_k)$$

are always onto, essentially by definition of the affine and convex hulls. Thus, Proposition 2.9.3 shows that x_1, \dots, x_k are affinely independent if and only if (2.9.7) is injective. Since $\Delta^{k-1} \subset \text{Aff}(e_1, \dots, e_k)$, injectivity of (2.9.7) implies injectivity of (2.9.8), so it suffices to show the converse.

So suppose (2.9.7) is not injective. The diagram (2.9.1) in the proof of Proposition 2.9.3 then shows that the linear map

$$(2.9.9) \quad f : \text{span}(e_2 - e_1, \dots, e_k - e_1) \rightarrow \text{span}(x_2 - x_1, \dots, x_k - x_1)$$

of linear bases is not injective. In particular, we can find a nonzero vector v in the linear base, V , of $\text{Aff}(e_1, \dots, e_k)$ with $f(v) = 0$. By Lemma 2.9.17, $x = \frac{e_1 + \dots + e_k}{k} \in \text{Int}(\Delta)^{k-1}$, and hence $x + tv \in \Delta^{k-1}$ for t sufficiently small. But $f(x + tv) = f(x) + tf(v) = f(x)$, so (2.9.8) is not injective. \square

We can use Lemma 2.9.17 to characterize the interior of an arbitrary polytope. The following lemma is equivalent to the statement that linear maps are continuous. We include it for simplicity.

Lemma 2.9.20. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be linear. Then for each $\epsilon > 0$, there exists $\delta > 0$ such that $\|f(x)\| < \epsilon$ whenever $\|x\| < \delta$.*

Proof. Let $c = \min(\|f(e_1)\|, \dots, \|f(e_n)\|)$ and let $x = x_1e_1 + \dots + x_n e_n$. Then

$$\begin{aligned} \|f(x)\| &= \|x_1f(e_1) + \dots + x_nf(e_n)\| \leq |x_1|\|f(e_1)\| + \dots + |x_n|\|f(e_n)\| \\ &\leq c(|x_1| + \dots + |x_n|) \leq cn\|x\|, \end{aligned}$$

as $\|x\| = \sqrt{x_1^2 + \dots + x_n^2} \geq \sqrt{x_i^2} = |x_i|$. Take $\delta = \frac{\epsilon}{cn}$. \square

A key step in the above is that $\|x_1e_1 + \dots + x_n e_n\| = \sqrt{x_1^2 + \dots + x_n^2}$. We shall apply the above lemma with \mathbb{R}^n replaced by a linear subspace of \mathbb{R}^k for some k . We can do so by borrowing ideas from Chapter 4, which introduces the idea of an orthonormal basis. Here, v_1, \dots, v_n is an orthonormal basis of $V \subset \mathbb{R}^k$ if

$$(2.9.10) \quad \langle v_i, v_j \rangle = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

and $\text{span}(v_1, \dots, v_n) = V$. By Corollary 4.2.3, every subspace of \mathbb{R}^k has such a basis. By Corollary 4.1.6, if v_1, \dots, v_n is orthonormal, then

$$\|x_1v_1 + \dots + x_nv_n\| = \sqrt{x_1^2 + \dots + x_n^2}.$$

We obtain the following.

Corollary 2.9.21. *Let V be a linear subspace of \mathbb{R}^n and let $f : V \rightarrow \mathbb{R}^m$ be linear. Then for each $\epsilon > 0$, there exists $\delta > 0$ such that $\|f(x)\| < \epsilon$ whenever $\|x\| < \delta$.*

And this, in turn, implies that affine maps satisfy a property called uniform continuity:

Corollary 2.9.22. *Let $f : H \rightarrow K$ be an affine map between affine subspaces of \mathbb{R}^n and \mathbb{R}^m , respectively. Then f is uniformly continuous, meaning that for each $\epsilon > 0$, there exists $\delta > 0$ such that $f(B_\delta(x, H)) \subset B_\epsilon(f(x), K)$ for all $x \in H$ (i.e., δ does not depend on x).*

Proof. This is a direct consequence of Corollary 2.9.21 and (2.8.10). \square

As an easy corollary, we have the following.

Corollary 2.9.23. *Let $C = \text{Conv}(x_1, \dots, x_k)$ be a polytope in \mathbb{R}^n and let $x \in \text{Int}(C)$. Let $f : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be the linear map with $f(e_i) = x_i$ for $i = 1, \dots, k$. Then for each $i = 1, \dots, k$, there exists*

$$y_i = a_1 e_1 + \dots + a_k e_k \in \Delta^{k-1}$$

such that $f(y_i) = x$ and $a_i > 0$.

Proof. Let $H = \text{Aff}(x_1, \dots, x_k)$. Choose $\epsilon > 0$ such that $B_\epsilon(x, H) \subset C$. Define $\gamma : \mathbb{R} \rightarrow \mathbb{R}^n$ by $\gamma(t) = (1-t)x + tx_i$. Then γ is affine. Since $x, x_i \in H$, the image of γ lies in H . By uniform continuity, there exists δ such that $\gamma(B_\delta(0, \mathbb{R})) \subset B_\epsilon(x, H)$. So for $0 < \delta' < \delta$, $\gamma(-\delta')$ lies in the convex set C . Moreover, x is easily seen to lie in the interior of the line segment from $\gamma(-\delta')$ to x_i , i.e., $x = (1-s)\gamma(-\delta') + sx_i$ for $s \in (0, 1)$. Write $\gamma(-\delta')$ as a convex combination

$$\gamma(-\delta') = b_1 x_1 + \dots + b_k x_k,$$

i.e., $\gamma(-\delta') = f(z)$ for $z = b_1 e_1 + \dots + b_k e_k$. But then x is the image under f of $(1-s)z + se_i$, whose i -th coordinate, $(1-s)b_i + s$, is positive. \square

We can now prove the following.

Proposition 2.9.24. *Let $C = \text{Conv}(x_1, \dots, x_k)$ be a polytope in \mathbb{R}^n . Let $f : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be the linear map with $f(e_i) = x_i$ for $i = 1, \dots, k$. Then $f(\text{Int}(\Delta^{k-1})) = \text{Int}(C)$.*

Proof. We first show $f(\text{Int}(\Delta^{k-1})) \subset \text{Int}(C)$. Let $H = \text{Aff}(x_1, \dots, x_k)$ and let W be its linear base.

$$W = \text{span}(x_2 - x_1, \dots, x_k - x_1).$$

$V = \text{span}(e_2 - e_1, \dots, e_k - e_1)$ is the linear base for $\text{Aff}(e_1, \dots, e_k) \subset \mathbb{R}^k$. Let $x \in \text{Int}(\Delta^{k-1})$. Since f is linear, we obtain the following as in (2.9.1)

$$\begin{array}{ccc} \text{Aff}(e_1, \dots, e_k) & \xrightarrow{f} & \text{Aff}(x_1, \dots, x_k) \\ \tau_x \uparrow \cong & & \tau_{f(x)} \uparrow \cong \\ V & \xrightarrow{f} & W. \end{array}$$

Since $x \in \text{Int}(\Delta^{k-1})$, there exists $\epsilon > 0$ such that $\tau_x(B_\epsilon(0, V)) \subset \Delta^{k-1}$, and hence $\tau_{f(x)}(f(B_\epsilon(0, V))) \subset C$. It suffices to show there exists $\delta > 0$ such that $B_\delta(0, W) \subset f(B_\epsilon(0, V))$.

By Corollary 1.5.8, since $x_2 - x_1, \dots, x_k - x_1$ generate W , we can find a subset, $x_{i_1} - x_1, \dots, x_{i_r} - x_1$ that is a basis of W . Let

$$V_1 = \text{span}(e_{i_1} - e_1, \dots, e_{i_r} - e_1) \subset V.$$

Then $f|_{V_1} : V_1 \rightarrow W$ is an isomorphism. Write $g : W \rightarrow V$ for the composite

$$W \xrightarrow{(f|_{V_1})^{-1}} V_1 \subset V.$$

Then $f \circ g$ is the identity map on W . Since g is linear, Corollary 2.9.21 provides a $\delta > 0$ with $g(B_\delta(0, W)) \subset B_\epsilon(0, V)$. But then

$$B_\delta(0, W) = f \circ g(B_\delta(0, W)) \subset f(B_\epsilon(0, V)).$$

We now show $f : \text{Int}(\Delta^{k-1}) \rightarrow \text{Int}(C)$ is onto. Let $x \in \text{Int}(C)$. By Corollary 2.9.23, we can choose $y_i \in \Delta^{k-1}$, $i = 1, \dots, k$, such that $f(y_i) = x$ and the i -th coordinate of y_i is positive. But then $x = f(y)$ for $y = \frac{y_1 + \dots + y_k}{k}$, a convex combination of y_1, \dots, y_k , and hence an element of Δ^{k-1} . Since the coordinates of y are all positive, the result follows. \square

Corollary 2.9.25. *Let $C = \text{Conv}(x_1, \dots, x_k) \subset \mathbb{R}^n$. Then $x \in \text{Int}(C)$ if and only if it can be written as a convex combination $x = a_1x_1 + \dots + a_kx_k$ with $a_i > 0$ for all i .*

Corollary 2.9.26. *Let $C = \text{Conv}(x_1, \dots, x_k)$ be a polytope in \mathbb{R}^n . Let K be an affine subspace of \mathbb{R}^m and let $f : C \rightarrow K$ be affine. Then*

$$f(\text{Int}(C)) = \text{Int}(f(C)).$$

Proof. Let $g : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be the linear map taking e_i to x_i for all i . Let $h : \mathbb{R}^k \rightarrow \mathbb{R}^m$ be the linear map taking e_i to $f(x_i)$ for all i . Then $f \circ g|_{\Delta^{k-1}}$ agrees with $h|_{\Delta^{k-1}}$ on vertices, so the two are equal. The result follows from Proposition 2.9.24. \square

Corollary 2.9.27. *Let $\emptyset \neq C \subset \mathbb{R}^n$ be convex. Then $\text{Int}(C)$ is nonempty. Indeed, if $\dim C = k$ and if $x_1, \dots, x_{k+1} \in C$ are algebraically independent, then $\text{Int}(\text{Conv}(x_1, \dots, x_{k+1})) \subset \text{Int}(C)$.*

Proof. First note that by Corollary 2.9.10, we can find an affinely independent subset $x_1, \dots, x_{k+1} \in C$. By construction, $\text{Aff}(x_1, \dots, x_{k+1}) = \text{Aff}(C)$, and the two have the same linear base, W . We have shown that $\text{Int}(\text{Conv}(x_1, \dots, x_{k+1}))$ is nonempty, and for $x \in \text{Int}(\text{Conv}(x_1, \dots, x_{k+1}))$, we have

$$\tau_x(B_\epsilon(0, W)) \subset \text{Conv}(x_1, \dots, x_{k+1}) \subset C$$

for some $\epsilon > 0$, so $x \in \text{Int}(C)$. \square

We now have the tools we need to prove the following.

Theorem 2.9.28. *Let $\emptyset \neq C \subset \mathbb{R}^n$ be convex and let K be an affine subspace of \mathbb{R}^m . Let $f : C \rightarrow K$ be affine. Then f extends to a unique affine map $\bar{f} : \text{Aff}(C) \rightarrow K$.*

Proof. Let $k = \dim C$ and let $x_1, \dots, x_{k+1} \in C$ be affinely independent. Then $\text{Aff}(C) = \text{Aff}(x_1, \dots, x_{k+1})$, so any affine map $\bar{f} : \text{Aff}(C) \rightarrow K$ is determined by its restriction to $\{x_1, \dots, x_{k+1}\}$, and hence by its restriction to C . Thus, it suffices to show that f extends to an affine map on $\text{Aff}(C)$.

Let $x \in \text{Int}(C)$ and let V be the linear base of $\text{Aff}(C)$. Then $D = \tau_x^{-1}(C)$ is a convex subset of V containing $0 = \tau_x^{-1}(x)$ in its interior. Let g be the composite

$$D \xrightarrow{\tau_x} C \xrightarrow{f} K \xrightarrow{\tau_{-f(x)}} W,$$

where W is the linear base of K . Then g is affine. Since $g(0) = 0$, any extension of g to an affine map $\bar{g} : V \rightarrow W$ is linear. If \bar{g} is such an extension, then $\bar{f} = \tau_{f(x)} \circ \bar{g} \circ \tau_x$ is the desired extension of f to $\text{Aff}(C)$.

Thus, we may assume that $\text{Aff}(C) = V$, $K = W$, $0 \in \text{Int}(C)$ and $f(0) = 0$. We shall show that f then extends to a linear map $\bar{f} : V \rightarrow W$. By assumption, there exists $\epsilon > 0$ such that $B_\epsilon(0, V) \subset C$. Making ϵ slightly smaller, we may assume that the closed ball $\bar{B}_\epsilon(0, V) \subset C$, where

$$\bar{B}_\epsilon(0, V) = \{x \in V : \|x\| \leq \epsilon\}.$$

We shall again use the fact that any linear subspace of \mathbb{R}^n admits an orthonormal basis. Let v_1, \dots, v_k be an orthonormal basis of V , so that

$$\langle v_i, v_j \rangle = \delta_{ij} = \begin{cases} 0 & \text{for } i \neq j, \\ 1 & \text{for } i = j. \end{cases}$$

Then

$$\|a_1 v_1 + \dots + a_k v_k\| = \sqrt{a_1^2 + \dots + a_k^2}.$$

In particular, if $|a_i| \leq \delta$ for all i , then $\|a_1 v_1 + \dots + a_k v_k\| \leq \delta\sqrt{k}$. Let $w_i = \frac{\epsilon}{\sqrt{k}} v_i$. Then $\mathcal{B} = w_1, \dots, w_k$ is a basis of V . Let

$$I_{\mathcal{B}} = \{a_1 w_1 + \dots + a_k w_k : |a_i| \leq 1 \text{ for all } i.\}$$

Note that we have inclusions

$$(2.9.11) \quad \bar{B}_{\frac{\epsilon}{\sqrt{k}}}(0, V) \subset I_{\mathcal{B}} \subset \bar{B}_\epsilon(0, V) \subset C.$$

In particular, f is defined and affine on the convex set $I_{\mathcal{B}}$, with $f(0) = 0$. By Proposition 2.5.1 and induction on k ,

$$(2.9.12) \quad f(a_1 w_1 + \cdots + a_k w_k) = a_1 f(w_1) + \cdots + a_k f(w_k) \\ \text{whenever } |a_i| \leq 1 \text{ for all } i.$$

In other words, f agrees on $I_{\mathcal{B}}$ with the unique linear map $\bar{f} : V \rightarrow W$ agreeing with f on the basis \mathcal{B} . It suffices to show that \bar{f} agrees with f on an arbitrary point $y \in C$.

Let $y_0 = sy$ for $s = \frac{\epsilon}{\|y\|\sqrt{k}}$. Then $y_0 \in \bar{B}_{\frac{\epsilon}{\sqrt{k}}}(0, V) \subset I_{\mathcal{B}}$, so $\bar{f}(y_0) = f(y_0)$. By Proposition 2.5.1,

$$f(y) = f\left(\frac{1}{s}y_0\right) = \frac{1}{s}f(y_0) = \frac{1}{s}\bar{f}(y_0) = \bar{f}\left(\frac{1}{s}y_0\right) = \bar{f}(y). \quad \square$$

Remark 2.9.29. In discussing affine maps whose domain is an affine subspace, Equation (2.4.4) would actually be the definition of choice for an affine map. It is the property one uses.

One would be tempted to use that definition for an affine map on an affine subspace, and then give a different name to maps whose domain is a convex set that satisfy $f((1-t)x + ty) = (1-t)f(x) + tf(y)$ for $t \in [0, 1]$. The name ‘‘convex map’’ has a certain appeal, but that is commonly used for a different concept.

Thanks to Theorem 2.9.28, there is no need for a different name, though we had to work to develop the tools to show it.

2.9.3. Faces. Points not in the interior of $\text{Conv}(x_1, \dots, x_k)$ lie in faces. The following notion will help use develop their theory.

Definition 2.9.30. Let $X = \{x_1, \dots, x_k\} \subset \mathbb{R}^n$ and let $x \in \text{Conv}(X)$. Define the support $S(x) \subset X$ of x in X to be the set of all $x_i \in X$ for which there exists a convex combination $x = a_1 x_1 + \cdots + a_k x_k$ with $a_i > 0$.

The following lemma is useful.

Lemma 2.9.31. *Let $X = \{x_1, \dots, x_k\} \subset \mathbb{R}^n$ and let $x \in \text{Conv}(X)$. Let $y \neq z \in \text{Conv}(X)$ and suppose x lies in the interior, (y, z) of the line segment \overline{yz} . Then the supports of these elements behave as follows:*

$$(2.9.13) \quad S(y) \cup S(z) \subset S(x).$$

Proof. $x = (1-t)y + tz$ for some $t \in (0, 1)$, so if

$$y = a_1 x_1 + \cdots + a_k x_k, \\ z = b_1 x_1 + \cdots + b_k x_k,$$

Then

$$x = ((1-t)a_1 + tb_1)x_1 + \cdots + ((1-t)a_k + tb_k)x_k.$$

If at least one of a_i and b_i is positive, so is $((1-t)a_i + tb_i)$, and the result follows. \square

The following generalizes some of the arguments above and is useful in understanding the points in

$$\partial \operatorname{Conv}(X) = \operatorname{Conv}(X) \setminus \operatorname{Int}(\operatorname{Conv}(X)).$$

Lemma 2.9.32.

(1) Let $X = \{x_1, \dots, x_k\} \subset \mathbb{R}^n$ and let $x \in \operatorname{Conv}(X)$. Then x lies in the interior of the convex hull of its support in X :

$$(2.9.14) \quad x \in \operatorname{Int}(\operatorname{Conv}(S(x))).$$

(2) Let $y \in \operatorname{Conv}(X) \cap \operatorname{Aff}(S(x))$. Then $S(y) \subset S(x)$, so by (1),

$$(2.9.15) \quad y \in \operatorname{Int}(\operatorname{Conv}(S(y))) \subset \operatorname{Conv}(S(y)) \subset \operatorname{Conv}(S(x)).$$

Thus,

$$(2.9.16) \quad \operatorname{Conv}(S(x)) = \operatorname{Conv}(X) \cap \operatorname{Aff}(S(x)).$$

Proof. (1) Let $f : \mathbb{R}^k \rightarrow \mathbb{R}^n$ be the linear map taking e_i to x_i for all i . Let $S(x) = \{x_{i_1}, \dots, x_{i_r}\}$. Note that for $z \in \Delta^{k-1}$ with $f(z) = x$, z must lie in $\operatorname{Conv}(e_{i_1}, \dots, e_{i_r})$. By hypothesis, we may choose $z_j \in \operatorname{Conv}(e_{i_1}, \dots, e_{i_r})$, for $j = 1, \dots, r$, such that

$$z_j = c_{j1}e_{i_1} + \dots + c_{jr}e_{i_r}$$

with $f(z_j) = x$ and $c_{jj} > 0$. then $z = \frac{z_1 + \dots + z_r}{r} \in \operatorname{Int}(\operatorname{Conv}(e_{i_1}, \dots, e_{i_r}))$ with $f(z) = x$, so $x \in \operatorname{Int}(\operatorname{Conv}(S(x)))$, as claimed.

(2) By (1) if $H = \operatorname{Aff}(S(x))$, there exists $\epsilon > 0$ such that

$$B_\epsilon(x, H) \subset \operatorname{Conv}(S(x)).$$

Let $y \in \operatorname{Conv}(X) \cap \operatorname{Aff}(S(x))$. Since $x, y \in H = \operatorname{Aff}(S(x))$, we have an affine map $\gamma : \mathbb{R} \rightarrow H$ via

$$\gamma(t) = (1-t)x + ty.$$

By Corollary 2.9.22, there exists $\delta > 0$ such that

$$\gamma(-\delta) \in B_\epsilon(x, H) \subset \operatorname{Conv}(S(x)).$$

Then x is in the interior of the line segment joining y to $\gamma(-\delta)$, i.e.,

$$x = (1-s)y + s\gamma(-\delta) \quad \text{for } s \in (0, 1).$$

By Lemma 2.9.31, $S(y) \subset S(x)$. □

Corollary 2.9.33. Let $X = \{x_1, \dots, x_k\} \subset \mathbb{R}^n$ and let $x \in \operatorname{Conv}(X)$. Then $x \in \operatorname{Int}(\operatorname{Conv}(X))$ if and only if its support $S(x)$ is equal to X .

Proof. If $S(x) = X$, then $x \in \operatorname{Int}(\operatorname{Conv}(X))$ by Lemma 2.9.32. The converse follows from Corollary 2.9.25. □

We shall see that the subsets $\operatorname{Conv}(S(x))$ comprise the faces of $\operatorname{Conv}(X)$, but it is useful to have a definition of faces that doesn't depend on X :

Definition 2.9.34. Let $X = \{x_1, \dots, x_k\} \subset \mathbb{R}^n$. A face of $\operatorname{Conv}(X)$ is a nonempty subset $F \subset \operatorname{Conv}(X)$ such that

- (1) $F = H \cap \text{Conv}(X)$ for some affine subspace $H \subset \mathbb{R}^n$.
- (2) $\text{Conv}(X) \setminus F$ is convex.

A face of dimension zero is called a vertex. A face of dimension 1 is called an edge. We write $\mathcal{V}(\text{Conv}(X))$ for the set of vertices of X and $\mathcal{E}(\text{Conv}(X))$ for the set of edges.

Remarks 2.9.35. Note that $\text{Conv}(x_1, \dots, x_k)$ is itself a face. The other faces are called proper faces. Since both H and $\text{Conv}(X)$ are convex, F is the intersection of two convex subsets of \mathbb{R}^n , and hence is convex. So it has a dimension.

Note that if F is a face, then (1) is satisfied by taking $H = \text{Aff}(F)$. That could have been built into the definition, but it can be easier to verify that a given subset is a face without having to determine its affine hull in advance.

A vertex consists of a single point $v \in \text{Conv}(X)$, as any convex set containing two distinct points has dimension at least 1. Any point $v \in \text{Conv}(X)$ automatically satisfies (1) with $H = \text{Aff}(v) = \{v\}$, so the sole criterion for v to be a vertex is that $\text{Conv}(X) \setminus \{v\}$ is convex.

Note that for $X = \{x_1, \dots, x_n\}$ it is not necessarily true that every element of X is a vertex. For instance, if $X = \{0, 1, 2\} \subset \mathbb{R}$, then $\text{Conv}(X) \setminus \{1\}$ is not convex, so 1 is not a vertex.

The following is basic, but useful.

Lemma 2.9.36. *Let $F_1 \subset F_2$ be faces of the polytope $C = \text{Conv}(X)$ with $F_1 \neq F_2$. Then $\dim F_1 < \dim F_2$.*

Proof. We have $F_i = \text{Aff}(F_i) \cap C$, and $\text{Aff}(F_1) \subset \text{Aff}(F_2)$. If these two affine subspaces are equal, then $F_1 = F_2$. Otherwise,

$$\dim \text{Aff}(F_1) < \dim \text{Aff}(F_2). \quad \square$$

Lemma 2.9.37. *Let $X = \{x_1, \dots, x_k\} \subset \mathbb{R}^n$ and let F_1 and F_2 be faces of $\text{Conv}(X)$ with $F_1 \cap F_2 \neq \emptyset$. Then $F_1 \cap F_2$ is a face as well.*

Proof. If H_1 and H_2 are affine subspaces and $x \in H_1 \cap H_2$, then

$$\tau_{-x}(H_1 \cap H_2) = \tau_{-x}(H_1) \cap \tau_{-x}(H_2)$$

is a linear subspace, so $H_1 \cap H_2$ is affine. It is easy to see that

$$(\text{Aff}(F_1) \cap \text{Aff}(F_2)) \cap \text{Conv}(X) = F_1 \cap F_2.$$

Thus, it suffices to show $\text{Conv}(X) \setminus (F_1 \cap F_2)$ is convex. The most complicated verification is showing that if $x \in F_1 \setminus F_2$ and $y \in F_2 \setminus F_1$, then $\overline{xy} = [x, y]$ is disjoint from $F_1 \cap F_2$. We argue by contradiction. Suppose $z \in (x, y)$ lies in $F_1 \cap F_2$. Since x, y, z are collinear and distinct, $\text{Aff}(x, z) = \text{Aff}(x, y)$. But $\text{Aff}(x, z)$ lies in $\text{Aff}(F_1)$ and hence $y \in \text{Aff}(F_1) \cap \text{Conv}(X) = F_1$, giving the desired contradiction. \square

Lemma 2.9.38. *$X = \{x_1, \dots, x_k\} \subset \mathbb{R}^n$ and let $x \in \text{Conv}(X)$. Then there is at most one face of $\text{Conv}(X)$ containing x in its interior.*

Proof. Suppose x lies in the interior of both F_1 and F_2 . Let $\epsilon > 0$ be small enough that both $B_\epsilon(x, \text{Aff}(F_1))$ and $B_\epsilon(x, \text{Aff}(F_2))$ are contained in $\text{Conv}(X)$. But then

$$B_\epsilon(x, \text{Aff}(F_1 \cap F_2)) \subset B_\epsilon(x, \text{Aff}(F_1)) \cap B_\epsilon(x, \text{Aff}(F_2)) \subset \text{Conv}(X),$$

so x lies in the interior of $F_1 \cap F_2$, i.e., we may as well assume F_1 is contained in F_2 and that $B_\epsilon(x, \text{Aff}(F_2)) \subset \text{Conv}(X)$.

Let V_i be the linear base for $\text{Aff}(F_i)$ for $i = 1, 2$. Then V_1 is a proper subspace of V_2 and there exists a vector $v \in V_2 \setminus V_1$ of norm less than ϵ . but then $x - v$ and $x + v$ lie in $\text{Conv}(X) \setminus F_1$, but the line segment between them does not. \square

We obtain the following.

Proposition 2.9.39. *Let $X = \{x_1, \dots, x_k\} \subset \mathbb{R}^n$ and let $x \in \text{Conv}(X)$. Let $S(x)$ be its support in X . Then $\text{Conv}(S(x))$ is the unique face of $\text{Conv}(X)$ containing x in its interior. Since every face of $\text{Conv}(X)$ has nonempty interior, the subsets $\text{Conv}(S(x))$ with $x \in X$ are the only faces of $\text{Conv}(X)$.*

Proof. Lemma 2.9.32 shows that

$$\text{Aff}(S(x)) \cap \text{Conv}(X) = \text{Conv}(S(x)).$$

To show $\text{Conv}(S(x))$ is a face, it suffices to show that $\text{Conv}(X) \setminus \text{Conv}(S(x))$ is convex. So let $y, z \in \text{Conv}(X) \setminus \text{Conv}(S(x))$, and suppose $w \in \overline{yz}$ lies in $\text{Conv}(S(x))$. By Lemma 2.9.31, $S(y) \cup S(z) \subset S(w)$. But $S(w) \subset S(x)$ by Lemma 2.9.32, so $y, z \in \text{Conv}(S(x))$, contradicting the existence of such a w .

By Lemma 2.9.32, x is in the interior of $\text{Conv}(S(x))$, and the result follows from Lemma 2.9.38. \square

Corollary 2.9.40. *Let $X = \{x_1, \dots, x_k\} \subset \mathbb{R}^n$. Then any vertex of $\text{Conv}(X)$ lies in X . Moreover, $v \in X$ is a vertex if and only if v does not lie in $\text{Conv}(X \setminus \{v\})$, i.e., there is no subset $S \subset X$ not containing v such that $v \in \text{Conv}(S)$.*

Proof. A vertex is a 0-dimensional face. By Proposition 2.9.39. every face of $\text{Conv}(X)$ has the form $\text{Conv}(S)$ for some $S \subset X$. If S has more than one element, then $\text{Conv}(S)$ has dimension at least one. Moreover, $\text{Conv}(x) = \{x\}$, so v must lie in X , and $S(v) = \{v\}$. But for $v \in X$, $S(v) \neq \{v\}$ if and only if v is a convex combination of the points in $X \setminus \{v\}$. \square

We obtain a unique and unambiguous way to describe a polytope.

Proposition 2.9.41. *Let $X = \{x_1, \dots, x_k\} \subset \mathbb{R}^n$ and let $C = \text{Conv}(X)$. Write $\mathcal{V} = \mathcal{V}(C)$, the set of vertices of C . (Recall from Corollary 2.9.40 that $\mathcal{V} \subset X$.) Then*

$$(2.9.17) \quad C = \text{Conv}(\mathcal{V}).$$

Since vertices depend only on C and not on the choice of a convex generating set X for C , this gives a unique description of a polytope.

Proof. If $x \in X \setminus \mathcal{V}$, then $x \in \text{Conv}(X \setminus \{x\})$ by Corollary 2.9.40, so

$$X \subset \text{Conv}(X \setminus \{x\}).$$

Since $\text{Conv}(X)$ is the smallest convex set containing X ,

$$\text{Conv}(X) \subset \text{Conv}(X \setminus \{x\}),$$

hence $\text{Conv}(X) = \text{Conv}(X \setminus \{x\})$. The result follows by induction on $|X \setminus \mathcal{V}|$. \square

Notation 2.9.42. We shall feel free to discuss a polytope \mathbf{P} without prior reference to a set of convex generators X (i.e., a finite set X with $\mathbf{P} = \text{Conv}(X)$). In particular, our preferred set of convex generators will be its set of vertices $\mathcal{V} = \mathcal{V}(\mathbf{P})$.

Let $x \in \mathbf{P}$. The support of x , $S(x)$, is taken to be its support in \mathcal{V} : if $\mathcal{V} = \{v_1, \dots, v_k\}$, then $S(x) \subset \mathcal{V}$ is the set of all v_i such that x may be written as a convex combination $x = a_1v_1 + \dots + a_kv_k$ with $a_i > 0$.

By Proposition 2.9.39, $\text{Conv}(S(x))$ is the unique face of \mathbf{P} containing x in its interior. We call $\text{Conv}(S(x))$ the carrier of x (in \mathbf{P}).

Recall the boundary of \mathbf{P} is $\partial\mathbf{P} = \mathbf{P} \setminus \text{Int}(\mathbf{P})$. Proposition 2.9.39 gives us the following:

Corollary 2.9.43. *Let \mathbf{P} be a polytope. Then the boundary of \mathbf{P} is the union of its proper faces. In particular, every element of $\partial\mathbf{P}$ lies in a proper face of \mathbf{P} .*

Proof. A point x is in the interior of \mathbf{P} if and only if its carrier is \mathbf{P} . Otherwise, it is in the boundary and its carrier is a proper face of \mathbf{P} . \square

Another consequence of Proposition 2.9.39 is the following.

Corollary 2.9.44. *Let \mathbf{P} be a polytope with vertex set \mathcal{V} . Then any face F of \mathbf{P} is the convex hull of the vertices that lie in it:*

$$(2.9.18) \quad F = \text{Conv}(F \cap \mathcal{V}).$$

Proof. F is the convex hull of some subset $S \subset \mathcal{V}$, all of whose elements must lie in F , i.e., $S \subset F \cap \mathcal{V}$. But we claim S must be equal to $F \cap \mathcal{V}$.

Suppose to the contrary that there exists $v \in F \cap \mathcal{V} \setminus S$. Then

$$v \in F = \text{Conv}(S) \subset \text{Conv}(\mathcal{V} \setminus \{v\}),$$

contradicting that v is a vertex. \square

The following is valuable in identifying carriers.

Proposition 2.9.45. *Let \mathbf{P} be a polytope and let $v, w \in \mathbf{P}$. Let x and y be interior points of the segment $[v, w]$. Then x and y have the same carrier. The carriers of v and w are contained in it.*

Proof. Carriers are determined by supports. We may as well assume that $x \in (v, y)$ and $y \in (x, w)$. By Lemma 2.9.31, $S(y) \subset S(x)$ and vice versa, and both supports contain $S(v)$ and $S(w)$. \square

The following addresses the situation where one polytope is contained in another as a subset of \mathbb{R}^n , with no known relationship between their structure. One application we will find for it is in showing which subsets $T \subset \mathcal{V}(\mathbf{P})$ are vertex sets for faces of \mathbf{P} . We shall use it in our analysis of the Platonic solids.

Corollary 2.9.46. *Let T be a finite subset of the polytope \mathbf{P} and let $\mathbf{Q} = \text{Conv}(T)$. Then every element of $\text{Int}(\mathbf{Q})$ has the same carrier, $F \subset \mathbf{P}$ in \mathbf{P} (i.e., F is a face of \mathbf{P} and $\text{Int}(\mathbf{Q}) \subset \text{Int}(F)$). Moreover, $\mathbf{Q} \subset F$. (We shall hereby refer to F as the carrier of \mathbf{Q} in \mathbf{P} .)*

Proof. Let $x \neq y \in \text{Int}(\mathbf{Q})$ and define $\gamma : \mathbb{R} \rightarrow \text{Aff}(\mathbf{Q}) \subset \text{Aff}(\mathbf{P})$ by $\gamma(t) = (1-t)x + ty$. Since x and y are in the interior of \mathbf{Q} , we can find $\delta > 0$, such that $\gamma([- \delta, 1 + \delta]) \subset \mathbf{Q} \subset \mathbf{P}$. Since x and y are in the interior of $[\gamma(-\delta), \gamma(1 + \delta)]$, they have the same carrier, F , in \mathbf{P} .

It suffices to show that $\partial\mathbf{Q} \subset F$. But if $x \in \partial\mathbf{Q}$ and $y \in \text{Int}(\mathbf{Q})$, let $\gamma(t) = (1-t)x + ty$ for $t \in \mathbb{R}$. Because $y \in \text{Int}(\mathbf{Q})$, there exists $\delta > 0$ such that $\gamma([0, 1 + \delta]) \subset \mathbf{Q} \subset \mathbf{P}$. Since y is in the interior of $[x, \gamma(1 + \delta)]$, the support of x is contained in $S(y)$. So $x \in \text{Conv}(S(y)) = F$. \square

A convenient source of faces is the following. In fact, one can show that all faces arise in this manner. We go back to considering convex generating sets that could include nonvertices, as this result is commonly used to identify which of the convex generators are actually vertices.

Proposition 2.9.47. *Let $X = \{x_1, \dots, x_k\} \subset \mathbb{R}^n$ and let $C = \text{Conv}(X)$. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be affine. Let $f(C) = [a, b]$. Then:*

- (1) $f^{-1}(b) \cap C$ is a face of C .
- (2) Let $S = f^{-1}(b) \cap X$. Then $f^{-1}(b) \cap C = \text{Conv}(S)$. Moreover, b is the maximum value of the restriction, $f|_X$, of f to X , while a is the minimum value.

In particular, every affine function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ that is not constant on X may be used to identify two proper faces of C : $f^{-1}(b)$ and $f^{-1}(a)$.

Proof. (1) $f^{-1}(b)$ is an affine subspace, so it suffices to show $C \setminus f^{-1}(b)$ is convex. Let $x, y \in C \setminus f^{-1}(b)$. Then

$$f((1-t)x + ty) = (1-t)f(x) + tf(y) \leq \max(f(x), f(y)) < b.$$

- (2) First note that $f(\text{Conv}(S)) = \text{Conv}(f(S)) = \text{Conv}(\{b\}) = \{b\}$.

Now let $T = X \setminus S$. Then $f(\text{Conv}(T)) = \text{Conv}(f(T)) \subset [a, b)$, as $f(T) \subset [a, b)$.

Now C is the linear join $\text{Conv}(T) \cdot \text{Conv}(S)$, so any element of C has the form $z = (1-t)x + ty$ with $x \in \text{Conv}(T)$ and $y \in \text{Conv}(S)$, $t \in I$. But then

$$f(z) = (1-t)f(x) + tf(y).$$

By the observations above, $f(x) < b$ and $f(y) = b$. so if $t < 1$, then $f(z) < b$. \square

2.9.4. Examples. We first describe the faces of the standard simplex.

Example 2.9.48. The faces of Δ^{n-1} correspond to the nonempty subsets of $\{1, \dots, n\}$. Let $1 \leq i_1 < \dots < i_k \leq n$, and set

$$(2.9.19) \quad \begin{aligned} F_{i_1, \dots, i_k} &= \text{Conv}(e_{i_1}, \dots, e_{i_k}) \\ &= \{a_1 e_1 + \dots + a_n e_n \in \Delta^{n-1} : a_j = 0 \text{ for } j \notin \{i_1, \dots, i_k\}\}. \end{aligned}$$

Since this is the standard simplex in the Euclidean space $\text{span}(e_{i_1}, \dots, e_{i_k})$, it has dimension $k - 1$ by Lemma 2.9.17. Moreover,

$$F_{i_1, \dots, i_k} = \Delta^{n-1} \cap \text{span}(e_{i_1}, \dots, e_{i_k}),$$

so F_{i_1, \dots, i_k} satisfies (1) in the definition of face. Moreover,

$$(2.9.20) \quad \text{Int}(F_{i_1, \dots, i_k}) = \{a_1 e_{i_1} + \dots + a_k e_{i_k} : a_j > 0 \text{ for } j = 1, \dots, k\}.$$

Finally,

$$\Delta^{n-1} \setminus F_{i_1, \dots, i_k} = \left\{ a_1 e_1 + \dots + a_n e_n \in \Delta^{n-1} : \sum_{j=1}^k a_{i_j} < 1 \right\}.$$

This condition demonstrates the complement is convex. Therefore, F_{i_1, \dots, i_k} is a face of Δ^{n-1} . In particular, we have constructed $\binom{n}{k}$ different $(k - 1)$ -dimensional faces of Δ^{n-1} . Since every nonempty subset of $\{1, \dots, n\}$ is the vertex set of a face, we have constructed all possible faces of Δ^{n-1} .

Remark 2.9.49. The 0-dimensional faces (vertices) are given by $F_i = \{e_i\}$, the i -th basis element. We of course just write e_i for this set.

The 1-dimensional faces (edges) are the line segments $F_{i,j} = \overline{e_i e_j}$ for $i < j$. There are $\binom{n}{2}$ of them.

There are n different $(n - 2)$ -dimensional faces, each of which is obtained by omitting one of the vertices. We write $\partial_i(\Delta^{n-1})$ for the $(n - 2)$ -dimensional face opposite the vertex e_i :

$$(2.9.21) \quad \begin{aligned} \partial_i(\Delta^{n-1}) &= F_{1, \dots, i-1, i+1, \dots, n} \\ &= \text{Conv}(e_j : j \neq i) \\ &= \{a_1 e_1 + \dots + a_n e_n \in \Delta^{n-1} : a_i = 0\}, \end{aligned}$$

We have

$$(2.9.22) \quad \partial \Delta^{n-1} = \bigcup_{i=1}^n \partial_i \Delta^{n-1},$$

and every element of $x \in \partial \Delta^{n-1}$ lies in the interior of exactly one face, specified by the nonzero coordinates of x : we can write x uniquely as

$$(2.9.23) \quad x = a_{i_1} e_{i_1} + \dots + a_{i_k} e_{i_k} \quad \text{with } a_{i_j} \neq 0 \text{ for } j = 1, \dots, k.$$

So F_{i_1, \dots, i_k} is the unique face containing x in its interior. Here, the interior of a vertex is the vertex itself.

Example 2.9.50. We now consider the faces of the n -cube I^n . Let $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ be the projection onto the i -th coordinate:

$$f_i(a_1e_1 + \cdots + a_n e_n) = a_i.$$

Then f_i is linear, so that $f_i^{-1}(0) = \ker f_i$ is a linear subspace, in this case $\text{span}(e_j : j \neq i)$, and $f_i^{-1}(1)$ is an affine subspace, in this case $\tau_{e_i}(f_i^{-1}(0))$. These linear and affine subspaces have dimension $n - 1$.

We write $\partial_i^0(I^n) = I^n \cap f_i^{-1}(0)$ and $\partial_i^1(I^n) = I^n \cap f_i^{-1}(1)$. Then

$$(2.9.24) \quad \begin{aligned} \partial_i^0(I^n) &= \{a_1e_1 + \cdots + a_n e_n \in I^n : a_i = 0\}, \\ \partial_i^1(I^n) &= \{a_1e_1 + \cdots + a_n e_n \in I^n : a_i = 1\}. \end{aligned}$$

This gives

$$(2.9.25) \quad \begin{aligned} I^n \setminus \partial_i^0(I^n) &= \{a_1e_1 + \cdots + a_n e_n \in I^n : a_i > 0\}, \\ I^n \setminus \partial_i^1(I^n) &= \{a_1e_1 + \cdots + a_n e_n \in I^n : a_i < 1\}. \end{aligned}$$

These complements are convex, so $\partial_i^0(I^n)$ and $\partial_i^1(I^n)$ are faces of I^n . Note that $\partial_i^0(I^n)$ is the unit cube in the $(n - 1)$ -dimensional Euclidean space $\text{span}(e_j : j \neq i)$, and hence has dimension $n - 1$ as a convex set. Moreover, $\partial_i^1(I^n) = \tau_{e_i}(\partial_i^0(I^n))$, and hence is $(n - 1)$ -dimensional also. We shall refer to $\partial_i^0(I^n)$ and $\partial_i^1(I^n)$ as *opposite faces* of I^n .

Recall from Corollary 2.8.28 that $I^n = \text{Conv}(S)$ for

$$(2.9.26) \quad S = \{\epsilon_1e_1 + \cdots + \epsilon_n e_n : \epsilon_1, \dots, \epsilon_n \in \{0, 1\}\}.$$

By Corollary 2.9.40, the vertices of I^n must lie in S , and hence its vertices are the elements of S whose complement is convex. In fact, every element of S is a vertex, as it is an intersection of faces: if $\epsilon_i \in \{0, 1\}$ for $i = 1, \dots, n$, then

$$(2.9.27) \quad \epsilon_1e_1 + \cdots + \epsilon_n e_n = \partial_1^{\epsilon_1}(I^n) \cap \cdots \cap \partial_n^{\epsilon_n}(I^n),$$

an intersection of faces, and hence a face.

Note that every element of I^n not in $\text{Int}(I^n)$ lies in at least one $(n - 1)$ -dimensional face, so that

$$(2.9.28) \quad \partial I^n = \bigcup_{i=1}^n (\partial_i^0(I^n) \cup \partial_i^1(I^n)).$$

Remark 2.9.51. Recall that an edge of a polytope is a 1-dimensional face, and hence is the convex hull of two vertices. We call it the edge determined by these two vertices. Note that each vertex of I^2 lies on exactly two edges. For instance, 0 lies on $\text{Conv}(0, e_1)$ and $\text{Conv}(0, e_2)$, but $\text{Conv}(0, e_1 + e_2)$ is not an edge, as it intersects the interior of I^2 .

2.10. Exercises.

1. Prove that the line $\begin{bmatrix} a \\ b \end{bmatrix} + \text{span}(\begin{bmatrix} c \\ d \end{bmatrix})$ coincides with the set of points $\begin{bmatrix} x \\ y \end{bmatrix}$ with $y = \frac{d}{c}x + \frac{bc-ad}{c}$.
2. Put the line $y = mx + b$ in the form $v + \text{span}(w)$.
3. Show that any two distinct affine planes in \mathbb{R}^3 are parallel (i.e., are translates of one another) if and only if they do not intersect.
4. Generalize the preceding problem to appropriate affine subspaces of \mathbb{R}^n .
5. Show that any distance-preserving function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is affine and one-to-one.
6. Show that if $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is affine and one-to-one, then $n \leq m$. Show that the range of f is an affine subspace of dimension m .
7. Show that not every linear isomorphism $T_A : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is an isometry.
8. We say $X \subset \mathbb{R}^n$ is an affine set if $(1-t)x + ty \in X$ for all $x, y \in X$ and $t \in \mathbb{R}$.
 - (a) Show that if $X \subset \mathbb{R}^n$ is an affine set containing 0, then X is a linear subspace.
 - (b) Deduce that every affine set in \mathbb{R}^n is an affine subspace.
9. Let $C \subset \mathbb{R}^n$ be convex and let $f : C \rightarrow \mathbb{R}^m$ be affine. Show that $f(C)$ is a convex subset of \mathbb{R}^m .
10. Let $C \subset \mathbb{R}^n$ be convex. Show that $\text{Int}(C)$ is also convex.
11. Let $C \subset \mathbb{R}^n$ be convex. Let $x \in \text{Int}(C)$ and $y \in \partial C$. Show that $[x, y) \subset \text{Int}(C)$.
12. Show that the closed unit disk

$$\mathbb{D}^n = \{x \in \mathbb{R}^n : \|x\| \leq 1\}$$
 is convex.
13. Show that the open unit disk

$$\text{Int}(\mathbb{D}^n) = \{x \in \mathbb{R}^n : \|x\| < 1\}$$
 is convex.
14. Show that the unit sphere

$$\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$$
 is not convex.
15. Show that \mathbb{D}^n is the convex hull of \mathbb{S}^{n-1} .
16. Show that every face of I^n is the intersection of some collection of $(n-1)$ -dimensional faces.
17. For a given $x \in \partial I^n$, describe the face containing x in its interior.
18. What are the edges of I^n ? How many of them are there?
19. What are the edges of I^n containing 0? Does each of the other vertices lie in the same number of edges?
20. How many $(n-1)$ -dimensional faces of I^n contain 0? Do you get the same number for any other vertex?

3. Groups

Group theory is valuable in understanding isometry, as \mathcal{I}_n , the set of isometries of \mathbb{R}^n , forms a group under composition. Other geometric questions may be solved by group theory as well.

3.1. Definition and examples.

Definition 3.1.1. A group, G , is a set together with a binary operation, which we think of as multiplication. I.e., for $g, h \in G$ the product $gh \in G$. The operation has the following properties.

- (1) Multiplication is associative: $(gh)k = g(hk)$ for all $g, h, k \in G$.
- (2) There is an identity element e for the multiplication: $ge = eg = g$ for all $g \in G$.
- (3) Each element $g \in G$ is invertible: there is an element $h \in G$ with $gh = hg = e$.

If G is finite (i.e., if it has finitely many elements), we refer to the number of elements in it, $|G|$, as its order.

Note that the multiplication is not assumed to be commutative, and it will not be commutative in most of the groups of interest here. Note that the definition of invertibility above is exactly the definition used for invertibility of matrices.

Examples 3.1.2.

- (1) Let X be a set. The permutation group $\Sigma(X)$ of X is the set of all bijections, $\sigma : X \rightarrow X$. It is a group under composition, as the inverse function of a bijection is a bijection. The n -symmetric group Σ_n is defined to be $\Sigma(\{1, \dots, n\})$.
- (2) The set \mathcal{I}_n of isometries of \mathbb{R}^n is a group under composition. It is immediate that the composition of isometries is an isometry. The identity element is the identity map, id , of \mathbb{R}^n . Inverses exist by Corollary 2.5.6.
- (3) The set \mathcal{A}_n of affine automorphisms of \mathbb{R}^n is also a group under composition by Proposition 2.6.4. It contains \mathcal{I}_n as a subgroup (see below).
- (4) The set \mathcal{S}_n of all similarities of \mathbb{R}^n is a group under composition by Lemma 2.7.3 and Corollary 2.7.4.
- (5) The set of $n \times n$ invertible matrices forms a group, denoted $\text{GL}_n(\mathbb{R})$, under matrix multiplication. The identity element is the identity matrix I_n . That the product of invertible matrices is invertible is shown by the proof of Lemma 3.1.3(2) below.
- (6) Note that $\text{GL}_1(\mathbb{R})$ is the set of 1×1 matrices $[s]$ with s a nonzero real number. The product $[s] \cdot [t] = [st]$. Thus, we may identify $\text{GL}_1(\mathbb{R})$ with the group \mathbb{R}^\times of nonzero real numbers under multiplication.
- (7) \mathbb{R}^n is a group under addition. So is any other vector space. Vector spaces are groups with additional structure, with the additional

structure given by the scalar multiplication. We call the group structure here the “additive group” of the vector space.

- (8) The integers, \mathbb{Z} , form a group under addition.

From now on we will suppress the composition symbol and simply write $\alpha\beta$ for the composition of $\alpha, \beta \in \mathcal{I}_n$.

Lemma 3.1.3. *Let G be a group.*

- (1) *Let $g, h, k \in G$ with $hg = gk = e$. Then $h = k$. Thus, inverses are unique, and we may write g^{-1} for the unique element with $gg^{-1} = g^{-1}g = e$.*
 (2) *For $h, k \in G$, $(hk)^{-1} = k^{-1}h^{-1}$.*

Proof. For (1), if $hg = gk = e$, then

$$k = ek = (hg)k = h(gk) = he = h.$$

For (2), we simply multiply:

$$\begin{aligned} k^{-1}h^{-1} \cdot h k &= k^{-1}(h^{-1}h)k = k^{-1}ek = k^{-1}k = e \\ h k \cdot k^{-1}h^{-1} &= h(kk^{-1})h^{-1} = heh^{-1} = hh^{-1} = e \end{aligned} \quad \square$$

We can now define higher powers of elements of G inductively: $g^n = g \cdot g^{n-1}$ for $n > 1$, and define negative powers by $g^{-n} = (g^n)^{-1}$ for $n > 0$. Of course, $g^0 = e$ and $g^1 = g$. The following is tedious to prove, but true (see [17]).

Lemma 3.1.4. *Let G be a group and let $g \in G$. Then for all $m, n \in \mathbb{Z}$,*

- (1) $g^m \cdot g^n = g^{m+n}$,
 (2) $(g^m)^n = g^{mn}$.

Remark 3.1.5. In a group such as \mathbb{R} or \mathbb{Z} whose group operation is written additively, the k -th “power” of an element g is

$$kg = \underbrace{g + \cdots + g}_{k \text{ times}}$$

when $k > 0$. The 0th “power” is, of course, 0, while, if $k < 0$, the k th “power” of g is the (additive) inverse of $(-k)g$, as defined above. Note that if $G = \mathbb{R}$ or \mathbb{Z} this notion of kg coincides with multiplication by the integer k by the definition of multiplication in \mathbb{R} .

The power laws of Lemma 3.1.4 then look like distributivity and associativity, respectively.

We will be very interested in studying subgroups of $\text{GL}_n(\mathbb{R})$ and \mathcal{I}_n .

Definition 3.1.6. Let G be a group. A subgroup $H \subset G$ is a nonempty subset such that:

- (1) H is closed under multiplication: for $h, k \in H$, the product hk is in H .
 (2) H is closed under inverses: for $h \in H$, $h^{-1} \in H$.

A subgroup $H \subset G$ is easily seen to be a group under the operation of G .

- Examples 3.1.7.** (1) \mathcal{I}_n is a subgroup of the group \mathcal{S}_n of similarities of \mathbb{R}^n , which is in turn a subgroup of \mathcal{A}_n .
 (2) The set \mathbb{R}^{pos} of positive real numbers under multiplication, is a subgroup of the group \mathbb{R}^\times of nonzero real numbers under multiplication.
 (3) Write \mathcal{LI}_n for the linear isometries from \mathbb{R}^n to \mathbb{R}^n and write \mathcal{L}_n for the linear automorphisms of \mathbb{R}^n (i.e., the isomorphisms from \mathbb{R}^n to \mathbb{R}^n). Then we have inclusions of subgroups

$$(3.1.1) \quad \begin{array}{ccc} \mathcal{LI}_n & \subset & \mathcal{I}_n \\ \cap & & \cap \\ \mathcal{L}_n & \subset & \mathcal{A}_n. \end{array}$$

- (4) The translations of \mathbb{R}^n form a subgroup, $\mathcal{T}_n \subset \mathcal{I}_n$:

$$\mathcal{T}_n = \{\tau_x : x \in \mathbb{R}^n\}.$$

- (5) The set of $n \times n$ invertible matrices with integer coefficients is *not* a subgroup of $\text{GL}_n(\mathbb{R})$. For instance, if $A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$, then A has integer coefficients and is invertible, but the reader may easily verify that

$$A^{-1} = \begin{bmatrix} 1 & -\frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix},$$

which does not have integer coefficients.

Thus, to obtain a subgroup, we define $\text{GL}_n(\mathbb{Z})$ to be the set of matrices $A \in \text{GL}_n(\mathbb{R})$ such that both A and A^{-1} have integer coefficients.

- (6) Let G be a group and let $g \in G$. By Lemma 3.1.4, the set of all powers of g form a subgroup $\langle g \rangle$ of G , called the cyclic subgroup generated by g :

$$\langle g \rangle = \{g^n : n \in \mathbb{Z}\}.$$

Clearly, any subgroup H containing g must contain $\langle g \rangle$, so $\langle g \rangle$ is the smallest subgroup containing g .

- (7) For $g_1, \dots, g_k \in G$, we write $\langle g_1, \dots, g_k \rangle$ for the smallest subgroup of G containing g_1, \dots, g_k . This subgroup can be difficult to describe, and we must be careful not to confuse this notation for that of the inner product. We shall use a similar notation for “generators and relations” below.

- (8) For $n > 0$ in \mathbb{Z} , $\langle n \rangle = \{kn : k \in \mathbb{Z}\}$ is the set of all multiples of n . In particular, when $n = 2$, $\langle 2 \rangle$ is the set of all even integers.

Regarding Example (5) above, the following is an easy consequence of [17, Corollary 10.3.6].

Proposition 3.1.8. *A matrix $A \in \text{GL}_n(\mathbb{R})$ with integer coefficients lies in $\text{GL}_n(\mathbb{Z})$ if and only if $\det A = \pm 1$.*

Since this is important in 2-dimensional geometry, we include a proof for $n = 2$.

Proof of Proposition 3.1.8 for $n = 2$. Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \text{GL}_2(\mathbb{R})$. Then

$$(3.1.2) \quad A^{-1} = \frac{1}{\det A} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix},$$

e.g., by Exercise 1 in Chapter 1. Thus, if $\det A = \pm 1$, then A^{-1} has integer coefficients.

Conversely, if A^{-1} has integer coefficients, then $\det A^{-1}$ is an integer (as is $\det A$). We have

$$1 = \det(I_2) = \det(AA^{-1}) = \det A \det A^{-1}.$$

Thus, $\det A$ has a multiplicative inverse in \mathbb{Z} , so $\det A = \pm 1$. \square

The case of $n > 2$ just replaces (3.1.2) with the formula $A^{-1} = \frac{1}{\det A} A^{\text{adj}}$, where A^{adj} , the adjoint matrix of A , is an integer matrix by determinant theory.

Definition 3.1.9. A group G is cyclic if $g = \langle g \rangle$ for some $g \in G$.

3.2. Orders of elements. Cyclic subgroups give important information about the elements in a group. Some of that information is encoded in the concept of order.

Definition 3.2.1. An integer k is called an exponent for an element $g \in G$ if $g^k = e$. Of course, 0 is an exponent for every group element. We say g has finite order if it has a positive exponent. In this case, we define the order of g , written $|g|$, to be the smallest positive exponent of g , i.e., the smallest positive integer g for which $g^k = e$.

If g does not have a positive exponent, we write $|g| = \infty$.

Knowing the order of an element allows us to determine the structure of the cyclic subgroup it generates.

Proposition 3.2.2. Let $g \in G$ have finite order, say $|g| = n$. Then:

- (1) $g^k = g^\ell$ if and only if $k - \ell = nq$ for some $q \in \mathbb{Z}$.⁵
- (2) $\langle g \rangle$ has exactly n elements: $e = g^0, g^1, \dots, g^{n-1}$. Thus, $|g| = |\langle g \rangle|$.

If $|g| = \infty$, then the elements $\{g^k : k \in \mathbb{Z}\}$ are all distinct.

Proof. Let $|g| = n$. If $k - \ell = nq$, then $k = \ell + nq$, so

$$g^k = g^{\ell+nq} = g^\ell (g^n)^q = g^\ell e^q = g^\ell,$$

as $g^n = e$. Conversely, if $g^k = g^\ell$ with $k > \ell$, then $e = g^k g^{-\ell} = g^{k-\ell}$. By the standard division properties of integers, we may write $k - \ell = nq + r$

⁵In the language of number theory, this says k is congruent to ℓ mod n , written $k \equiv \ell \pmod{n}$.

with $0 \leq r < n$ (i.e., r is the remainder if we divide $k - \ell$ by n). Then $e = g^{k-\ell} = g^{nq+r} = (g^n)^q g^r = e^q g^r = g^r$. So r is an exponent of g . But n is the smallest positive exponent of g and $r < n$. So r is not positive. Thus $r = 0$, and hence $k - \ell = nq$.

(2) now follows from (1): if $m \in \mathbb{Z}$ and if r is the remainder when we divide m by n , then $g^m = g^r$ by the argument given above (and yes, we can do divisions with remainder even when m is negative), so

$$g^m \in \{g^k : 0 \leq k < n\}.$$

Moreover, these n elements are all distinct, as if $0 \leq \ell < k \leq n$ and if $g^k = g^\ell$, this would force n to divide $k - \ell$, which is impossible, as $0 < k - \ell < n$.

Finally, if $|g| = \infty$ and $g^k = g^\ell$ with $k \geq \ell$, then $e = g^{k-\ell}$. Since g has no positive exponents, $k - \ell$ is not positive, so $k = \ell$. \square

Examples 3.2.3.

- (1) Let $0 \neq x \in \mathbb{R}^n$. Then $\tau_x^k = \tau_{kx} \neq \text{id}$, as $kx \neq 0$, so $|\tau_x| = \infty$ in \mathcal{I}_n .
- (2) Let $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \in \text{GL}_2(\mathbb{R})$. Then $A^2 = -I_2$, $A^3 = -A$ and $A^4 = I_2$, so A has order 4.

3.3. Conjugation and normality. The subgroup of translations introduces a useful concept called normality. Translations do not commute with general isometries, but we can compute their deviation from commuting in an important sense. The key observation is the following.

Lemma 3.3.1. *Let A be an invertible $n \times n$ matrix and let $x \in \mathbb{R}^n$. Then*

$$(3.3.1) \quad T_A \tau_x T_A^{-1} = \tau_{Ax}.$$

Proof. For $y \in \mathbb{R}^n$,

$$\begin{aligned} T_A \tau_x T_A^{-1}(y) &= T_A \tau_x (A^{-1}y) = T_A (A^{-1}y + x) \\ &= AA^{-1}y + Ax = y + Ax = \tau_{Ax}(y). \end{aligned} \quad \square$$

Thus, if β is a linear isometry of \mathbb{R}^n and $x \in \mathbb{R}^n$, then

$$\beta \tau_x \beta^{-1} = \tau_{\beta(x)}.$$

But this extends immediately to the following:

Corollary 3.3.2. *Let $\alpha \in \mathcal{I}_n$ and $x \in \mathbb{R}^n$. Write $\alpha = \tau_z \beta$ with β a linear isometry (and hence $z = \alpha(0)$) Then*

$$(3.3.2) \quad \alpha \tau_x \alpha^{-1} = \tau_{\beta(x)}$$

Phrased entirely in terms of α , this says

$$\alpha \tau_x \alpha^{-1} = \tau_w \quad \text{for } w = \alpha(x) - \alpha(0).$$

Proof. Since $\alpha^{-1} = \beta^{-1} \tau_{-z}$, we have

$$\begin{aligned} \alpha \tau_x \alpha^{-1} &= \tau_z \beta \tau_x \beta^{-1} \tau_{-z} \\ &= \tau_z \tau_{\beta(x)} \tau_{-z} \end{aligned}$$

$$\begin{aligned}
&= \tau_{z+\beta(x)-z} \\
&= \tau_{\beta(x)}. \quad \square
\end{aligned}$$

The same argument shows the following.

Corollary 3.3.3. *Let $f \in \mathcal{A}_n$ and $x \in \mathbb{R}^n$. Write $f = \tau_z \circ g$ with g a linear isomorphism from \mathbb{R}^n to itself (and hence $z = f(0)$). Then*

$$(3.3.3) \quad f\tau_x f^{-1} = \tau_{g(x)}.$$

The operation we are looking at is called conjugation and detects deviation from commuting.

Definition 3.3.4. Let $x, g \in G$, where G is a group. The conjugate of g by x is xgx^{-1} .

Conjugation computes the deviation from commuting in the following way. If $xgx^{-1} = h$, then $xg = hx$, so “pulling x past g ” replaces g by h . Corollary 3.3.2 says every conjugate of a translation is a translation. So if we pull an isometry past a translation, our translation is replaced by a different, and computable translation. This is a very nice property and can be used to describe the multiplication in \mathcal{I}_n in full generality. Recall from Theorem 2.5.3 that every isometry is a composite $\tau_x\beta$ with β a linear isometry.

Proposition 3.3.5. *Let $x, y \in \mathbb{R}^n$ and let β, γ be linear isometries of \mathbb{R}^n . Then*

$$(3.3.4) \quad \tau_x\beta \cdot \tau_y\gamma = \tau_{x+\beta(y)}\beta\gamma.$$

Here, $\beta\gamma$, as the composite of two linear isometries, is a linear isometry.

Proof. $\tau_x\beta \cdot \tau_y\gamma = \tau_x(\beta\tau_y\beta^{-1})\beta\gamma = \tau_x\tau_{\beta(y)}\beta\gamma = \tau_{x+\beta(y)}\beta\gamma. \quad \square$

Of course, exactly the same calculation holds for affine automorphisms, and by the same proof. It is convenient here to use the fact that a linear automorphism of \mathbb{R}^n can be written uniquely in the form T_A for $A \in \text{GL}_n(\mathbb{R})$.

Proposition 3.3.6. *Let $A, B \in \text{GL}_n(\mathbb{R})$ and $x, y \in \mathbb{R}^n$. Then*

$$(3.3.5) \quad \tau_x T_A \cdot \tau_y T_B = \tau_{x+Ay} T_{AB}.$$

Definition 3.3.7. Let H be a subgroup of G .

- (1) The conjugate of H by $x \in G$ is

$$xHx^{-1} = \{xhx^{-1} : h \in H.\}$$

- (2) We say H is normal in G , written $H \triangleleft G$, if $xHx^{-1} = H$ for all $x \in G$.

Lemma 3.3.8.

- (1) For H a subgroup of G and $x \in G$, xHx^{-1} is a subgroup of G .
(2) $H \triangleleft G$ if and only if $xhx^{-1} \in H$ for all $x \in G$ and $h \in H$.

Proof. For (1), for $h, k \in H$, we have

$$xhx^{-1} \cdot xkx^{-1} = xh(x^{-1}x)kx^{-1} = xhkkx^{-1},$$

so xHx^{-1} is closed under multiplication. And

$$(xkx^{-1})^{-1} = (x^{-1})^{-1}h^{-1}x^{-1} = xh^{-1}x^{-1},$$

so xHx^{-1} is closed under inverses.

For (2), suppose $xhx^{-1} \in H$ for all $x \in G$ and $h \in H$. Then

$$(3.3.6) \quad xHx^{-1} \subset H \quad \text{for all } x \in G.$$

But then

$$H = x(x^{-1}Hx)x^{-1} \subset xHx^{-1}$$

since $x^{-1}Hx \subset H$ by substituting x^{-1} for x in (3.3.6). So $xHx^{-1} = H$. \square

By Lemma 3.3.8(2) and Lemma 3.3.2, we obtain:

Proposition 3.3.9. $\mathcal{T}_n \triangleleft \mathcal{I}_n$.

But in fact, Lemma 3.3.2 is stronger than this as it says exactly how \mathcal{T}_n is normal.

3.4. Homomorphisms. In this book, our primary interest in groups is in studying groups of symmetries of geometric figures. This fits into a framework called “group actions,” meaning that the way the group acts on the geometry and the figure is the primary focus of interest. In group theory proper, one is often more interested in the relationships between different groups. These relationships are often captured by the functions between groups that preserve the group structure:

Definition 3.4.1. Let G and H be groups. A homomorphism $f : G \rightarrow H$ is a function with the property that

$$f(xy) = f(x)f(y) \quad \text{for all } x, y \in G.$$

The kernel, $\ker f$, of f is

$$\ker f = \{x \in G : f(x) = e\},$$

the set of elements carried by f to the identity element of H .

Example 3.4.2. The determinant function gives rise to a homomorphism

$$\det : \mathrm{GL}_n(\mathbb{R}) \rightarrow \mathbb{R}^\times,$$

$\det(A) = \det A$. The point here is that a matrix A is invertible if and only if its determinant is nonzero, i.e., $\det A$ is an element of the group \mathbb{R}^\times of nonzero real numbers under multiplication. This map is a homomorphism because $\det(AB) = \det A \cdot \det B$ for any two $n \times n$ matrices A and B .

The kernel of the determinant map is the group

$$(3.4.1) \quad \mathrm{SL}_n(\mathbb{R}) = \{A \in \mathrm{GL}_n(\mathbb{R}) : \det A = 1\}.$$

It is called the n -th special linear group of \mathbb{R} . We shall see that $\mathrm{SL}_2(\mathbb{R})$ is important in understanding the isometries of hyperbolic space.

Recall that $\Sigma_n = \Sigma(\{1, \dots, n\})$, the permutation group of $\{1, \dots, n\}$. The following gives a very important example of a homomorphism.

Lemma 3.4.3. *There is a homomorphism $\iota_n : \Sigma_n \rightarrow \text{GL}(\mathbb{R})$ obtained by setting*

$$(3.4.2) \quad \iota_n(\sigma) = [e_{\sigma(1)} | \dots | e_{\sigma(n)}],$$

the matrix whose i -th column is $e_{\sigma(i)}$, where e_1, \dots, e_n are the canonical basis vectors of \mathbb{R}^n .

This homomorphism is one-to-one.

Proof. This is certainly one-to-one, as if $\iota_n(\sigma) = \iota_n(\tau)$, then $e_{\sigma(i)} = e_{\tau(i)}$ for all i , hence $\sigma(i) = \tau(i)$ for all i and $\sigma = \tau$.

To see it is a homomorphism, we have

$$\begin{aligned} \iota_n(\sigma)\iota_n(\tau) &= \iota_n(\sigma)[e_{\tau(1)} | \dots | e_{\tau(n)}] \\ &= [\iota_n(\sigma) \cdot e_{\tau(1)} | \dots | \iota_n(\sigma) \cdot e_{\tau(n)}] \\ &= [e_{\sigma(\tau(1))} | \dots | e_{\sigma(\tau(n))}], \end{aligned}$$

where the last equality follows as, for any matrix A , Ae_j is the j -th column of A . The result follows as $\sigma\tau$ is the composition of σ and τ . \square

We shall study ι_n in greater detail later. But now is a good time to make the following definition.

Definition 3.4.4. The n -th alternating group A_n is the kernel of the composite

$$\Sigma_n \xrightarrow{\iota_n} \text{GL}_n(\mathbb{R}) \xrightarrow{\det} \mathbb{R}^\times,$$

i.e., A_n is the group of permutations σ with the property that $\iota_n(\sigma)$ has determinant 1.⁶

We now develop some elementary theory about homomorphisms.

Lemma 3.4.5. *Let $f : G \rightarrow H$ be a homomorphism of groups and let $x \in G$. Then $f(x^k) = (f(x))^k$ for all $k \in \mathbb{Z}$.*

Proof. For $k = 0$ this follows from $x^0 = e$ and $e \cdot e = e$. so that

$$f(e) = f(e \cdot e) = f(e) \cdot f(e).$$

Multiplying both sides by $f(e)^{-1}$ we get $f(e) = e$.

⁶There is a different definition of A_n intrinsic to studying permutation groups (see, e.g., [17, 11]). One shows that every permutation is a composite of transpositions: permutations that interchange two indices and fix the rest. One then shows there is a homomorphism $\text{sgn} : \Sigma_n \rightarrow \{\pm 1\}$ whose value on every transposition is -1 and defines A_n to be the kernel of sgn . That sgn is well-defined is an essential component in the development of the determinant function, though there are approaches that work in the other direction.

We are not giving rigorous treatments of either determinant theory or permutation groups in this text. We strongly encourage the reader to consult an upper level algebra text such as [17] or [11] to learn the details.

For $k > 0$, this follows by induction on k from the inductive definition of powers and the homomorphism property. For $k < 0$ it follows from the fact that $x^k x^{-k} = e$. and hence

$$e = f(x^k x^{-k}) = f(x^k) f(x^{-k}) = f(x^k) (f(x))^{-k},$$

as $-k > 0$. Now multiply both sides by $(f(x))^k = ((f(x))^{-k})^{-1}$ and the result follows. \square

Corollary 3.4.6. *Let $f : G \rightarrow H$ be a homomorphism of groups. Then $\ker f \triangleleft G$.*

Proof. $\ker f$ is obviously closed under multiplication, and is closed under inverses by Lemma 3.4.5, so $\ker f$ is a subgroup of G . For $h \in \ker f$ and $x \in G$,

$$f(xhx^{-1}) = f(x)f(h)f(x^{-1}) = f(x)f(h)f(x)^{-1} = f(x)ef(x)^{-1} = e.$$

So $xhx^{-1} \in \ker f$. \square

Example 3.4.7. A linear function $f : V \rightarrow W$ between vector spaces gives a homomorphism between their additive groups.

A more interesting example comes from Theorem 2.5.3, which shows that each $\alpha \in \mathcal{I}_n$ may be written uniquely as a composite $\alpha = \tau_x \beta$ with $\beta \in \mathcal{LI}_n$ (i.e., $\beta : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a linear isometry).

Proposition 3.4.8. *Define*

$$\pi : \mathcal{I}_n \rightarrow \mathcal{LI}_n$$

as follows: if $\alpha = \tau_x \beta$ with $\beta \in \mathcal{LI}_n$, set $\pi(\alpha) = \beta$. Then π is a homomorphism whose kernel is \mathcal{T}_n . Moreover, if $j : \mathcal{LI}_n \subset \mathcal{I}_n$ is the inclusion, then $\pi \circ j = \text{id}_{\mathcal{LI}_n}$, the identity map of \mathcal{LI}_n (i.e., for $\beta \in \mathcal{LI}_n$, $\pi(\beta) = \beta$). Thus, π is onto.

In the language of group theory, this says that

$$1 \longrightarrow \mathcal{T}_n \xrightarrow{i} \mathcal{I}_n \xrightarrow{\pi} \mathcal{LI}_n \longrightarrow 1$$

is a split extension, where i is the natural inclusion.

Proof. Let $\alpha_1, \alpha_2 \in \mathcal{I}_n$ with

$$\alpha_1 = \tau_{x_1} \beta_1, \quad \alpha_2 = \tau_{x_2} \beta_2,$$

with $\beta_1, \beta_2 \in \mathcal{LI}_n$. Then

$$(3.4.3) \quad \begin{aligned} \alpha_1 \alpha_2 &= \tau_{x_1} \beta_1 \tau_{x_2} \beta_2 \\ &= \tau_{x_1 + \beta_1(x_2)} \beta_1 \beta_2 \end{aligned}$$

by Proposition 3.3.5. Thus

$$\pi(\alpha_1 \alpha_2) = \beta_1 \beta_2 = \pi(\alpha_1) \pi(\alpha_2),$$

so π is a homomorphism.

Now, $\ker \pi = \{\tau_x \beta : \beta = \text{id}\} = \mathcal{T}_n$. Finally, if β is a linear isometry, then $\beta = \tau_0 \beta$, and hence $\pi(\beta) = \beta$. \square

Another useful example is the following. Recall that \mathcal{S}_n , the group of similarities of \mathbb{R}^n is defined to be the set of functions $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that there exists $0 < s \in \mathbb{R}$, such that

$$d(f(x), f(y)) = s \cdot d(x, y) \quad \text{for all } x, y \in \mathbb{R}^n.$$

The real number s is called the scaling factor, $\text{scale}(f)$, of f . Recall also that the positive real numbers \mathbb{R}^{pos} form a group under multiplication (with identity element 1). Finally, for $s \in \mathbb{R}^{\text{pos}}$, we have a linear similarity

$$\mu_s : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

given by $\mu_s(x) = sx$ for all $x \in \mathbb{R}^n$, by (2.7.2).

Proposition 3.4.9. *The scaling factor*

$$\text{scale} : \mathcal{S}_n \rightarrow \mathbb{R}^{\text{pos}}$$

is a surjective homomorphism with kernel \mathcal{I}^n . Thus, $\mathcal{I}_n \triangleleft \mathcal{S}_n$. Moreover, there is a homomorphism $\sigma : \mathbb{R}^{\text{pos}} \rightarrow \mathcal{S}_n$ given by $\sigma(s) = \mu_s$. Since $\text{scale}(\mu_s) = s$, $\text{scale} \circ \sigma = \text{id}_{\mathbb{R}^{\text{pos}}}$. In the language of group theory this says that

$$1 \rightarrow \mathcal{I}_n \xrightarrow{i} \mathcal{S}_n \xrightarrow{\text{scale}} \mathbb{R}^{\text{pos}} \rightarrow 1$$

is a split extension, where i is the inclusion of \mathcal{I}_n in \mathcal{S}_n .

Proof. That scale is a homomorphism is immediate from Lemma 2.7.3. Since isometries are similarities of scaling factor 1, \mathcal{I}_n is the kernel of scale . Moreover, scale is surjective because $\text{scale} \circ \sigma = \text{id}_{\mathbb{R}^{\text{pos}}}$. \square

We have already studied the kernel of a linear function. Kernels of general group homomorphisms are useful for the same reasons.

Lemma 3.4.10. *Let $f : G \rightarrow H$ be a homomorphism and let $x, y \in G$. Then $f(x) = f(y)$ if and only if $x^{-1}y \in \ker f$. Thus, f is one-to-one if and only if $\ker f = \{e\}$.*

Proof.

$$\begin{aligned} f(x) = f(y) &\Leftrightarrow e = f(x)^{-1}f(y) = f(x^{-1}y) \\ &\Leftrightarrow x^{-1}y \in \ker f. \end{aligned}$$

Thus $\ker f = \{e\}$ implies f is one-to-one. But $f(e) = e$, so if f is one-to-one, then no other element may be carried by f to e , and hence $\ker f = \{e\}$. \square

Two groups may be identified with one another if they are isomorphic:

Definition 3.4.11. An isomorphism $f : G \rightarrow H$ of groups is a homomorphism that is one-to-one and onto. The inverse function $f^{-1} : H \rightarrow G$ is then easily seen to be a group homomorphism, and hence an isomorphism.

We say that G and H are isomorphic if there is an isomorphism between them. We write

$$f : G \xrightarrow{\cong} H$$

to indicate that f is an isomorphism from G to H .

A one-to-one homomorphism $\iota : G \rightarrow K$ is called an embedding of groups.

The homomorphism $\iota_n : \Sigma_n \rightarrow \text{GL}_n(\mathbb{R})$ is an embedding. We also have an interesting example of an isomorphism of groups.

Proposition 3.4.12. *There is an isomorphism $\nu : \mathbb{R}^n \rightarrow \mathcal{T}_n$ from the additive group \mathbb{R}^n to the translation subgroup of \mathcal{I}_n , given by $\nu(x) = \tau_x$.*

Proof. $\tau_x \tau_y = \tau_{x+y}$. □

We also have the following:

Example 3.4.13. There is an isomorphism $T : \text{GL}_n(\mathbb{R}) \rightarrow \mathcal{L}_n$ given by $T(A) = T_A$, the transformation induced by A . We will see below that this restricts to an isomorphism $T : \text{O}_n \rightarrow \mathcal{LI}_n$, where O_n is the n th orthogonal group, defined below.

The group \mathbb{R}^{pos} , of positive real numbers under multiplication, is an old friend.

Lemma 3.4.14. *There is an isomorphism $\exp : \mathbb{R} \rightarrow \mathbb{R}^{\text{pos}}$ given by*

$$\exp(x) = e^x.$$

Here \mathbb{R} is the group of real numbers under addition.

Proof. \exp is a homomorphism as $e^{x+y} = e^x \cdot e^y$. It is a bijection whose inverse function is the natural logarithm. □

Isomorphic groups have the same group structure. Every group theoretic property is preserved by an isomorphism. An important example of isomorphism is the following.

Proposition 3.4.15. *Let G be a group and let $x \in G$. Define the conjugation by x , $c_x : G \rightarrow G$, by*

$$c_x(y) = xyx^{-1}.$$

Then c_x is an isomorphism from G to itself. Moreover, if H is a subgroup of G , then $c_x(H) = xHx^{-1}$, the conjugate subgroup of H by x . In fact,

$$c_x : H \xrightarrow{\cong} xHx^{-1}.$$

Thus, conjugate subgroups are isomorphic.

Proof. $c_x(y)c_x(z) = xyx^{-1}xzx^{-1} = xyzx^{-1} = c_x(yz)$, so c_x is a homomorphism. But $c_{x^{-1}}$ is easily seen to provide an inverse function for c_x , so c_x is bijective.

Now $xHx^{-1} = c_x(H)$ by definition, and the rest follows by restriction of the properties above. □

In fact, two subgroups being conjugate is stronger than being isomorphic. A very important example of homomorphisms is the following.

Proposition 3.4.16. *Let G be a group and let $g \in G$. Then there is a unique homomorphism*

$$f_g : \mathbb{Z} \rightarrow G$$

with $f_g(1) = g$. Explicitly, $f_g(k) = g^k$ for all $k \in \mathbb{Z}$.

The image of f_g is $\langle g \rangle$. If $|g| = n < \infty$, then $\ker f_g = \langle n \rangle$, the cyclic subgroup of \mathbb{Z} generated by $|g| = n$. Otherwise $\ker f_g = \{0\}$, the identity subgroup of \mathbb{Z} , and hence f_g is one-to-one. In consequence, if $|g| = \infty$, f_g induces an isomorphism

$$f_g : \mathbb{Z} \xrightarrow{\cong} \langle g \rangle.$$

Proof. $k \in \mathbb{Z}$ is the k th power of 1 in the (additive) group structure of \mathbb{Z} , so if $f_g : \mathbb{Z} \rightarrow G$ is a homomorphism with $f_g(1) = g$, then $f_g(k) = g^k$ by Lemma 3.4.5. Conversely, if we define $f_g : \mathbb{Z} \rightarrow G$ by $f_g(k) = g^k$, then f_g is a homomorphism by the rules of exponents. By construction, the image of f_g is $\langle g \rangle$.

The kernel of f_g is the set of exponents of g , which, by Proposition 3.2.2(1) is the set of multiples of $|g|$ if $|g|$ is finite. But if $|g|$ is infinite, g has no exponents other than 0, hence $\ker f_g = \{0\}$, then. \square

3.5. A matrix model for isometries and affine maps. Much of this material could have been presented in Chapter 2, but is easier to understand with a little group theory.

Definition 3.5.1. Let $A \in \text{GL}_n(\mathbb{R})$ (i.e., A is an invertible $n \times n$ matrix with coefficients in \mathbb{R}) and let $x \in \mathbb{R}^n$. We write $M(A, x)$ for the $(n+1) \times (n+1)$ block matrix

$$(3.5.1) \quad M(A, x) = \left[\begin{array}{c|c} A & x \\ \hline 0 & 1 \end{array} \right].$$

These assemble into an useful collection of matrices

$$(3.5.2) \quad \mathfrak{A}_n = \left\{ \left[\begin{array}{c|c} A & x \\ \hline 0 & 1 \end{array} \right] : A \in \text{GL}_n(\mathbb{R}), x \in \mathbb{R}^n \right\}.$$

Indeed, \mathfrak{A}_n is a subgroup of $\text{GL}_{n+1}(\mathbb{R})$:

Lemma 3.5.2. *The product of two elements of \mathfrak{A}_n is given as follows:*

$$(3.5.3) \quad \left[\begin{array}{c|c} A & x \\ \hline 0 & 1 \end{array} \right] \left[\begin{array}{c|c} B & y \\ \hline 0 & 1 \end{array} \right] = \left[\begin{array}{c|c} AB & x + Ay \\ \hline 0 & 1 \end{array} \right].$$

The inverse of $M(A, x)$ is given by

$$(3.5.4) \quad \left[\begin{array}{c|c} A & x \\ \hline 0 & 1 \end{array} \right]^{-1} = \left[\begin{array}{c|c} A^{-1} & -A^{-1}x \\ \hline 0 & 1 \end{array} \right],$$

and hence \mathfrak{A}_n is closed under inverses.

Proof. (3.5.3) is immediate from (1.4.6). (3.5.4) then follows by an easy calculation. \square

This now allows us to establish an isomorphism between \mathfrak{A}_n and the group \mathcal{A}_n of affine automorphisms of \mathbb{R}^n .

Proposition 3.5.3. *There is an isomorphism $\varphi : \mathfrak{A}_n \rightarrow \mathcal{A}_n$ given by*

$$(3.5.5) \quad \varphi(M(A, x)) = \tau_x T_A.$$

Proof. By Proposition 3.3.6 and (3.5.3), φ is a homomorphism. Since every affine automorphism of \mathbb{R}^n can be written uniquely in the form $\tau_x T_A$, φ is bijective. \square

Block multiplication also gives a nice model for the way an affine automorphism acts on a vector in \mathbb{R}^n .

Lemma 3.5.4. *Let $A \in M_n(\mathbb{R})$ and $x \in \mathbb{R}^n$. Then for $y \in \mathbb{R}^n$, we have*

$$(3.5.6) \quad \left[\begin{array}{c|c} A & x \\ \hline 0 & 1 \end{array} \right] \left[\begin{array}{c} y \\ 1 \end{array} \right] = \left[\begin{array}{c} \tau_x T_A(y) \\ 1 \end{array} \right].$$

Thus, the action of the affine automorphism $\tau_x T_A$ on \mathbb{R}^n can be read off from the action of the block matrix $M(A, x)$ on the affine subspace

$$\tau_{e_{n+1}}(\mathbb{R}^n) \subset \mathbb{R}^{n+1}.$$

Here, we are identifying \mathbb{R}^n with $\text{span}(e_1, \dots, e_n) \subset \mathbb{R}^{n+1}$.

Proof. This follows from Proposition 1.4.4. \square

Let us consider this correspondence in the context of isometries. The subgroup $\mathcal{I}_n \subset \mathcal{A}_n$ consists of the composites $\tau_x T_A$ such that T_A is a linear isometry. The $n \times n$ matrices A such that T_A is an isometry are called the orthogonal matrices and form a subgroup $O_n \subset GL_n(\mathbb{R})$. We will study O_n in Chapter 4.

Definition 3.5.5. Define $\mathfrak{I}_n \subset \mathfrak{A}_n$ by

$$(3.5.7) \quad \mathfrak{I}_n = \left\{ \left[\begin{array}{c|c} A & x \\ \hline 0 & 1 \end{array} \right] : A \in O_n, x \in \mathbb{R}^n \right\}.$$

Since O_n is a subgroup of $GL_n(\mathbb{R})$, Lemma 3.5.2 shows \mathfrak{I}_n to be a subgroup of \mathfrak{A}_n . Indeed, since $\varphi(M(A, x))$ is an isometry if and only if T_A is an isometry, we have:

Proposition 3.5.6. $\mathfrak{I}_n = \varphi^{-1}(\mathcal{I}_n)$, and hence $\varphi : \mathfrak{I}_n \rightarrow \mathcal{I}_n$ is an isomorphism of groups. As before,

$$(3.5.8) \quad \varphi \left(\left[\begin{array}{c|c} A & x \\ \hline 0 & 1 \end{array} \right] \right) = \tau_x T_A,$$

where $A \in O_n$ and $x \in \mathbb{R}^n$.

3.6. G -sets. We are concerned here on the ways that groups of symmetries act on geometric objects. For instance, we wish to study how \mathcal{I}_n acts on \mathbb{R}^n and how subgroups of \mathcal{I}_n act on subsets of \mathbb{R}^n . It is useful to develop some language and basic results that cover a variety of cases.

Definition 3.6.1. Let G be a group. A G -set X is a set together with a function

$$(3.6.1) \quad G \times X \rightarrow X$$

$$(3.6.2) \quad (g, x) \mapsto gx$$

called the action of G on X , that satisfies:

- (1) $g(hx) = (gh)x$ for all $g, h \in G$ and $x \in X$.
- (2) $ex = x$ for all $x \in X$.

We will write $g \cdot x$ at times for gx to avoid ambiguity. We will write “ G acts on X via (3.6.2)” to say that X is a G -set with the specified action.

If X and Y are G -sets, a G -map (or G -equivariant map) $f : X \rightarrow Y$ is a function f such that $f(gx) = gf(x)$ for all $g \in G$ and $x \in X$.

In most cases of interest, we will want X to have a topology and want the action to preserve the topology. This will be implicit in most of our examples, but is discussed in detail in Section A.4. For the most part we will work implicitly with regard to the topology.

Examples 3.6.2.

- (1) \mathcal{I}_n acts on \mathbb{R}^n via

$$(\alpha, x) \mapsto \alpha(x).$$

- (2) If G acts on X and H is a subgroup of G , then the restriction of the action map to $H \times X$ specifies an action of H on X : $h \cdot x = hx$. In particular, every subgroup of \mathcal{I}_n acts in this way on \mathbb{R}^n .
- (3) If G acts on X and $f : K \rightarrow G$ is a group homomorphism, then K acts on X via $(k, x) \mapsto f(k)x$.

There are two important concepts regarding G -actions: orbit and isotropy.

Definition 3.6.3. Let X be a G -set and let $x \in X$. The isotropy subgroup G_x of X is

$$(3.6.3) \quad G_x = \{g \in G : gx = x\},$$

the set of all elements of G that fix x . It is easily seen to be a subgroup of G by properties (1) and (2) of Definition 3.6.1.

The orbit $G \cdot x$ is

$$(3.6.4) \quad G \cdot x = \{gx : g \in G\}.$$

The isotropy subgroup expresses the redundancy in expressing the elements in the orbit:

$$(3.6.5) \quad g_1x = g_2x \iff g_2^{-1}g_1x = x \iff g_2^{-1}g_1 \in G_x.$$

Different elements in an orbit can have different isotropy subgroups, but they are conjugate.

Lemma 3.6.4. *Let X be a G -set and let $x \in X$ and $g \in G$. Then*

$$(3.6.6) \quad G_{gx} = gG_xg^{-1}.$$

Proof. As in (3.6.5), $g_1gx = gx$ if and only if $g^{-1}g_1g \in G_x$. Now conjugate by g . \square

A major theme here is the following:

Definition 3.6.5. Let X be a G -set and let $Y \subset X$. We say that $g \in G$ preserves Y if $g(Y) = Y$, where $g(Y)$ is the image of Y under g :

$$g(Y) = \{g(y) : y \in Y\}.$$

We define the symmetries of Y under this action to be the subgroup

$$(3.6.7) \quad \mathcal{S}_G(Y) = \{g \in G : g(Y) = Y\},$$

consisting of the elements of G that preserve Y .

Indeed, $\mathcal{S}_G(Y)$ is the isotropy subgroup of Y under the obvious action of G on the set of all subsets of X : the action in which $g \cdot Y = g(Y)$. Given that observation, the following is an immediate consequence of Lemma 3.6.4.

Corollary 3.6.6. *Let X be a G -set and let $Y \subset X$. Then for any $g \in G$, we have*

$$(3.6.8) \quad \mathcal{S}_G(g(Y)) = g\mathcal{S}_G(Y)g^{-1}.$$

Another important concept regarding G -sets is that of fixed-points.

Definition 3.6.7. Let X be a G -set and let $g \in G$. The fixed-point set of X under g is

$$(3.6.9) \quad X^g = \{x \in X : gx = x\}.$$

For a subgroup $H \subset G$, the H -fixed-point set is

$$(3.6.10) \quad X^H = \{x \in X : hx = x \text{ for all } h \in H\}.$$

Since the elements fixing a given x form a subgroup, we obtain:

$$(3.6.11) \quad X^g = X^{\langle g \rangle}.$$

A useful observation is the following.

Lemma 3.6.8. *Let X be a G -set, H a subgroup of G , and $g \in G$. Then*

$$(3.6.12) \quad g(X^H) = X^{gHg^{-1}}.$$

Similarly, $g(X^h) = X^{ghg^{-1}}$.

Proof. The inclusion $g(X^H) \subset X^{gHg^{-1}}$ is obvious. Conversely, $ghg^{-1}y = y$ if and only if $hg^{-1}y = g^{-1}y$. \square

3.7. Direct products. Let G_1, \dots, G_k be groups. Then the Cartesian product $G_1 \times \cdots \times G_k$ has a group structure given by coordinatewise multiplication:

$$(x_1, \dots, x_k)(y_1, \dots, y_k) = (x_1y_1, \dots, x_ky_k).$$

We call this group structure the direct product of G_1, \dots, G_k and denote it simply by $G_1 \times \cdots \times G_k$. As the reader may verify, it has the property that a function $f : H \rightarrow G_1 \times \cdots \times G_k$ is a group homomorphism if and only if each of its coordinate functions $f_i : H \rightarrow G_i$ is a group homomorphism.

4. Linear isometries

We study the linear isometries from \mathbb{R}^n to itself (i.e., the group of linear automorphisms of \mathbb{R}^n).

4.1. Orthonormal bases and orthogonal matrices.

Definition 4.1.1.

- (1) A unit vector $u \in \mathbb{R}^n$ is a vector of norm 1: $\|u\| = 1$.
- (2) The set of all unit vectors in \mathbb{R}^n is the $(n - 1)$ -sphere \mathbb{S}^{n-1} :

$$\mathbb{S}^{n-1} = \{u \in \mathbb{R}^n : \|u\| = 1\}.$$

- (3) The vectors $x, y \in \mathbb{R}^n$ are orthogonal if $\langle x, y \rangle = 0$. In this case we write $x \perp y$.
- (4) The vectors $x_1, \dots, x_k \in \mathbb{R}^n$ form an orthogonal set if $\langle x_i, x_j \rangle = 0$ for all $i \neq j$ and if $x_i \neq 0$ for all i .
- (5) The vectors $x_1, \dots, x_k \in \mathbb{R}^n$ form an orthonormal set if $\langle x_i, x_j \rangle = 0$ for all $i \neq j$ and each x_i is a unit vector.
- (6) The Kronecker delta, δ_{ij} is defined by the formula

$$\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases}$$

Thus, $x_1, \dots, x_k \in \mathbb{R}^n$ is an orthonormal set if and only if

$$\langle x_i, x_j \rangle = \delta_{ij}$$

for all i, j .

The Kronecker delta is a very useful notation. For instance, the $n \times n$ identity matrix I_n has ij th coordinate δ_{ij} , hence

$$I_n = (\delta_{ij}).$$

Lemma 4.1.2. *Let $x_1, \dots, x_k \in \mathbb{R}^n$ form an orthogonal set. Then x_1, \dots, x_k are linearly independent.*

Proof. Suppose $a_1x_1 + \dots + a_kx_k = 0$. Then

$$\begin{aligned} 0 &= \langle x_i, a_1x_1 + \dots + a_kx_k \rangle = a_1\langle x_i, x_1 \rangle + \dots + a_k\langle x_i, x_k \rangle \\ &= a_i\langle x_i, x_i \rangle, \end{aligned}$$

as $\langle x_i, x_j \rangle = 0$ for $i \neq j$. Since $x_i \neq 0$, $\langle x_i, x_i \rangle \neq 0$, hence $a_i = 0$ for all i . \square

Thus, if x_1, \dots, x_n is an orthonormal set in \mathbb{R}^n , it is a basis of \mathbb{R}^n . In general, if $V \subset \mathbb{R}^n$ is a subspace, an orthonormal basis of V is an orthonormal set that spans V , and hence forms a basis for V . Orthonormal bases have a number of applications. They are very easy to work with if we can calculate the inner products of the vectors involved:

Lemma 4.1.3. *Let $\mathcal{B} = v_1, \dots, v_k$ be an orthonormal basis for a subspace $V \subset \mathbb{R}^n$ and let $v \in V$. Then*

$$(4.1.1) \quad v = \langle v, v_1 \rangle v_1 + \cdots + \langle v, v_k \rangle v_k.$$

In particular,

$$(4.1.2) \quad [v]_{\mathcal{B}} = \begin{bmatrix} \langle v, v_1 \rangle \\ \vdots \\ \langle v, v_k \rangle \end{bmatrix}.$$

Proof. Since v_1, \dots, v_k is a basis for V and $v \in V$, we can find coefficients a_1, \dots, a_k with $v = a_1 v_1 + \cdots + a_k v_k$. It suffices to show $a_i = \langle v, v_i \rangle$ for all i . But

$$\begin{aligned} \langle v, v_i \rangle &= \langle a_1 v_1 + \cdots + a_k v_k, v_i \rangle = a_1 \langle v_1, v_i \rangle + \cdots + a_k \langle v_k, v_i \rangle \\ &= a_i \langle v_i, v_i \rangle = a_i, \end{aligned}$$

as $\langle v_j, v_i \rangle = \delta_{ji}$ for all j . \square

We wish to find the matrices A that induce linear isometries. Orthogonality is the key idea. We first investigate the relationship between orthogonality and the Pythagorean formula. The following is a special case of the cosine law to be proven below.

Lemma 4.1.4. *Let $x, y \in \mathbb{R}^n$. Then*

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 \quad \Leftrightarrow \quad x \perp y.$$

Proof.

$$\|x + y\|^2 = \langle x + y, x + y \rangle = \langle x, x \rangle + 2\langle x, y \rangle + \langle y, y \rangle.$$

The right-hand side is $\|x\|^2 + \|y\|^2$ if and only if $\langle x, y \rangle = 0$. \square

We may now characterize not only the linear isometries of \mathbb{R}^n , but more generally the linear isometric embeddings of \mathbb{R}^n into \mathbb{R}^m , i.e., the linear functions $T_A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ with $d(T_A(x), T_A(y)) = d(x, y)$ for all $x, y \in \mathbb{R}^n$. Here, A is an $m \times n$ matrix.

Theorem 4.1.5. *Let $A = [v_1 | \dots | v_n]$ be $m \times n$ and let T_A be the linear function induced by A . Then the following conditions are equivalent.*

- (1) T_A is an isometric embedding.
- (2) $\|Ax\| = \|x\|$ for all $x \in \mathbb{R}^n$.
- (3) The columns, v_1, \dots, v_n , of A are an orthonormal set.
- (4) $\langle Ax, Ay \rangle = \langle x, y \rangle$ for all $x, y \in \mathbb{R}^n$.

Proof. (1) \Rightarrow (2):

$$\|Ax\| = d(Ax, 0) = d(T_A(x), T_A(0)) = d(x, 0) = \|x\|.$$

(2) \Rightarrow (3): $\|v_i\| = \|Ae_i\| = 1$, so each v_i is a unit vector. For $i \neq j$,

$$\|v_i + v_j\|^2 = \|A(e_i + e_j)\|^2 = \|e_i + e_j\|^2 = \|e_i\|^2 + \|e_j\|^2 = \|v_i\|^2 + \|v_j\|^2.$$

The third equality is by Lemma 4.1.4, as e_i and e_j are orthogonal. (The others are by (2) and linearity.) So v_i and v_j are orthogonal by Lemma 4.1.4.

(3) \Rightarrow (4): Let $x = x_1e_1 + \cdots + x_ne_n$ and $y = y_1e_1 + \cdots + y_ne_n$. Then

$$\begin{aligned} \langle Ax, Ay \rangle &= \langle A(x_1e_1 + \cdots + x_ne_n), A(y_1e_1 + \cdots + y_ne_n) \rangle \\ &= \sum_{i,j=1}^n x_i y_j \langle Ae_i, Ae_j \rangle \\ &= \sum_{i,j=1}^n x_i y_j \delta_{ij} \\ &= \sum_{i=1}^n x_i y_i = \langle x, y \rangle. \end{aligned}$$

(4) \Rightarrow (1):

$$\begin{aligned} d(T_A(x), T_A(y)) &= \|Ay - Ax\| = \|A(y - x)\| \\ &= \sqrt{\langle A(y - x), A(y - x) \rangle} \\ &= \sqrt{\langle y - x, y - x \rangle} = d(x, y). \quad \square \end{aligned}$$

Corollary 4.1.6. Let v_1, \dots, v_n be an orthonormal set in \mathbb{R}^m and $a_1, \dots, a_n \in \mathbb{R}$. Then,

$$(4.1.3) \quad \|a_1v_1 + \cdots + a_nv_n\| = \sqrt{a_1^2 + \cdots + a_n^2}.$$

Proof. Let $A = [v_1 | \dots | v_n]$. Then $a_1v_1 + \cdots + a_nv_n = A \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$. Now apply

the equivalence of (2) and (3) above. \square

Of course, the columns being orthonormal, and hence linearly independent, implies $n \leq m$. We obtain a nice characterization of linear isometric embeddings via transposes.

Definition 4.1.7. Let $A = (a_{ij})$ be an $m \times n$ matrix. We write A^T for the transpose of A : the $n \times m$ matrix whose ij th entry is a_{ji} . Thus, if v_i is the i -th column of A , then the i th row of A^T is v_i^T .

The following elementary result is useful.

Lemma 4.1.8. Let A be $m \times n$ and let B be $n \times k$. Then $(AB)^T = B^T A^T$.

Proof. The ij th coordinate of $(AB)^T$ is the j th coordinate of AB :

$$\sum_{k=1}^n a_{jk} b_{ki} = \sum_{k=1}^n b_{ki} a_{jk}.$$

This is dot product of the i th column of B with the transpose of the j th row of A , and that's exactly what we get from the ij th coordinate of $B^T A^T$.

(As the displayed equation hints, this lemma holds for matrices over any commutative ring.) \square

Transposes are useful in uncovering the relationship between the matrix product and the dot product.

Lemma 4.1.9. *Let A be $m \times n$ and let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Then*

$$(4.1.4) \quad \langle Ax, y \rangle = \langle x, A^T y \rangle = y^T Ax = x^T A^T y.$$

In particular, if $x, y \in \mathbb{R}^n$ and $A = I_n$ we obtain the dot product of x and y as the matrix product of the row matrix x^T with y :

$$(4.1.5) \quad \langle x, y \rangle = x^T y.$$

Proof. Each term in (4.1.4) expands to

$$\sum_{\substack{i=1, \dots, m \\ j=1, \dots, n}} a_{ij} x_j y_i. \quad \square$$

Corollary 4.1.10. *Let A be the $m \times n$ matrix whose i th row is the row matrix w_i for $i = 1, \dots, m$ and let B be the $n \times k$ whose j th column is the column matrix v_j for $j = 1, \dots, k$. Then the ij th coordinate of AB is given by*

$$(AB)_{ij} = \langle w_i^T, v_j \rangle.$$

Proof. By definition, $(AB)_{ij}$ is the matrix product $w_i v_j$. Apply (4.1.5). \square

Corollary 4.1.11. *Let $A = [v_1 | \dots | v_n]$ be $m \times n$. Then the ij th coordinate of $A^T A$ is $\langle v_i, v_j \rangle$. Thus T_A is a linear isometric embedding if and only if $A^T A = I_n$.*

Proof. Since the transpose of the i th row of A^T is just v_i , the ij th entry of $A^T A$ is $\langle v_i, v_j \rangle$ by Corollary 4.1.10. So $A^T A = I_n$ if and only if $\langle v_i, v_j \rangle = \delta_{ij}$ for all i, j . \square

Of course, if $m = n$, the columns form an orthonormal set if and only if they form an orthonormal basis of \mathbb{R}^n . In this case, A is invertible, so $A^T A = I_n \Rightarrow A^T = A^{-1}$. Of course, a linear isometric embedding from \mathbb{R}^n to \mathbb{R}^n is by definition a linear isometry. Summarizing the above information, we obtain a characterization of the linear isometries of \mathbb{R}^n .

Theorem 4.1.12. *Let A be $n \times n$. Then the following conditions are equivalent.*

- (1) $T_A : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a linear isometry.
- (2) The columns of A form an orthonormal basis of \mathbb{R}^n .
- (3) $\langle Ax, Ay \rangle = \langle x, y \rangle$ for all $x, y \in \mathbb{R}^n$.
- (4) A is invertible with inverse A^T .

Definition 4.1.13. An $n \times n$ matrix A such that T_A is a linear isometric embedding is called an orthogonal matrix. It is characterized by its columns being an orthonormal basis of \mathbb{R}^n . We write $O(n)$ for the set of $n \times n$ orthogonal matrices.

Proposition 4.1.14. $O(n)$ is a subgroup of $GL_n(\mathbb{R})$.

Proof. For $A, B \in O(n)$,

$$(AB)^T = B^T A^T = B^{-1} A^{-1} = (AB)^{-1},$$

so $O(n)$ is closed under multiplication. Also, since $A^T = A^{-1}$, $AA^T = I_n$. But $AA^T = (A^T)^T A^T$, so A^T satisfies Theorem 4.1.12, and hence lies in $O(n)$. So $O(n)$ is closed under inverses. \square

That last observation gives:

Corollary 4.1.15. Let $A \in O(n)$. Then $A^T \in O(n)$, hence the rows of A form an orthonormal basis of \mathbb{R}^n .

Theorem 4.1.12 now gives:

Corollary 4.1.16. There is a group isomorphism

$$T : O(n) \rightarrow \mathcal{LI}_n$$

from the orthogonal group to the group of linear isometries of \mathbb{R}^n , given by $T(A) = T_A$.

For any $n \times n$ matrix A , $\det(A) = \det(A^T)$. We obtain the following.

Corollary 4.1.17. Let $A \in O(n)$ then $\det A = \pm 1$. We obtain a group homomorphism

$$\det : O(n) \rightarrow \{\pm 1\}.$$

Its kernel is

$$SO(n) = \{A \in O(n) : \det A = 1\},$$

a subgroup of $O(n)$ called the n -th special orthogonal group.

Proof. Since $I_n = A^T A$,

$$1 = \det I_n = \det(A^T) \det(A) = \det(A)^2,$$

so $\det A = \pm 1$. Kernels of homomorphisms are always subgroups. \square

Recall the embedding $\iota_n : \Sigma_n \rightarrow GL_n(\mathbb{R})$ of Lemma 3.4.3 given by

$$(4.1.6) \quad \iota_n(\sigma) = [e_{\sigma(1)} | \dots | e_{\sigma(n)}],$$

the matrix whose i -th column is $e_{\sigma(i)}$. In particular, the columns of $\iota_n(\sigma)$ are obtained by permuting the columns of I_n and hence form an orthonormal basis of \mathbb{R}^n . Thus, $\iota_n(\sigma)$ is an orthogonal matrix, and we obtain the following.

Corollary 4.1.18. *The embedding ι_n takes value in O_n , i.e., (4.1.6) gives an embedding*

$$\iota_n : \Sigma_n \rightarrow O_n.$$

In particular, $\det(\iota_n(\sigma)) = \pm 1$ for all $\sigma \in \Sigma_n$.

Definition 4.1.19. We write $\text{sgn} : \Sigma_n \rightarrow \{\pm 1\}$ for the composite

$$\Sigma_n \xrightarrow{\iota_n} O(n) \xrightarrow{\det} \{\pm 1\}.$$

Recall that a transposition is a permutation that interchanges two indices and leaves all the others fixed.

Lemma 4.1.20. *Let $\tau \in \Sigma_n$ be a transposition. Then $\text{sgn}(\tau) = -1$. Thus $\text{sgn} : \Sigma_n \rightarrow \{\pm 1\}$ is onto. Its kernel is the alternating group A_n defined in Definition 3.4.4.*

Proof. $\iota_n(\tau)$ is obtained from the identity matrix by exchanging two columns. So its determinant is $-\det I_n = -1$. \square

Remark 4.1.21. Using group theoretic analysis, one can show there is a unique homomorphism $s : \Sigma_n \rightarrow \{\pm 1\}$ taking each transposition to -1 . (See, e.g., [17, Proposition 3.5.8] for existence. Uniqueness follows because the transpositions generate Σ_n [17, Corollary 3.5.6]). By Lemma 4.1.20, that unique homomorphism coincides with the definition we've given of sgn .

However, the existence and uniqueness of s are often used in the proof that $\det(AB) = \det A \det B$. So care is needed to avoid circular arguments. The reader is encouraged to read a careful development of the group theory and determinant theory used here. The treatment of determinants, in particular, is generally nonrigorous in elementary linear algebra courses.

Recall:

Definition 4.1.22. $c \in \mathbb{R}$ is an eigenvalue for the $n \times n$ matrix A if there is a nonzero vector v with $Av = cv$. If c is an eigenvalue, the eigenspace of (A, c) is the set of all vectors v with $Av = cv$. The eigenspace is a subspace of \mathbb{R}^n . Its elements are called eigenvectors of (A, c) .

Lemma 4.1.23. *Let c be a real eigenvalue for the orthogonal matrix A . Then $c = \pm 1$.*

Proof. Let v be a nonzero eigenvector for A, c . Then

$$\langle v, v \rangle = \langle Av, Av \rangle = \langle cv, cv \rangle = c^2 \langle v, v \rangle.$$

Since $v \neq 0$, $\langle v, v \rangle \neq 0$, so $c^2 = 1$. \square

Another characterization of eigenvalues is that they are the roots of the characteristic polynomial $\text{ch}_A(x)$ of A . We shall discuss this further below. Here, we note that orthogonal matrices may have nonreal eigenvalues, i.e., nonreal roots of $\text{ch}_A(x)$. For instance $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ has characteristic polynomial $x^2 + 1$ and hence has eigenvalues $\pm i$.

4.2. Gramm–Schmidt. Orthogonal matrices play a very important role in both Euclidean and spherical geometry. Since the columns of an orthogonal matrix form an orthonormal basis of \mathbb{R}^n it will be worth our while to find ways to produce orthonormal bases. One such method is called Gramm–Schmidt orthogonalization. It will be very useful in our study of spherical geometry below.

Algorithm 4.2.1 (Gramm–Schmidt orthogonalization). Let v_1, \dots, v_k be linearly independent in \mathbb{R}^n . We first give an inductive procedure to obtain an orthogonal set w_1, \dots, w_k such that $\text{span}(v_1, \dots, v_i) = \text{span}(w_1, \dots, w_i)$ for $i = 1, \dots, k$.

We then replace each w_i with $z_i = \frac{w_i}{\|w_i\|}$ obtaining an orthonormal set z_1, \dots, z_k such that $\text{span}(v_1, \dots, v_i) = \text{span}(z_1, \dots, z_i)$ for $i = 1, \dots, k$. The end result is called the Gramm–Schmidt orthogonalization of v_1, \dots, v_k . Note in particular that z_1, \dots, z_k is an orthonormal basis for $\text{span}(v_1, \dots, v_k)$. In addition it respects the particular nest of subspaces $\text{span}(v_1, \dots, v_i)$ for $i = 1, \dots, k$.

We give the procedure here and show in the proposition below that it behaves as stated.

We define w_1, \dots, w_k inductively. We first set $w_1 = v_1$. Suppose now that we have inductively found w_1, \dots, w_i for $1 \leq i \leq k - 1$ such that:

- (1) w_1, \dots, w_i is an orthogonal set.
- (2) $\text{span}(v_1, \dots, v_j) = \text{span}(w_1, \dots, w_j)$ for $j = 1, \dots, i$.

We now give the procedure for finding w_{i+1} . We show in the proposition below that (1) and (2) remain true with i replaced by $i + 1$, and hence the process may continue.

In particular, we set

$$(4.2.1) \quad w_{i+1} = v_{i+1} - \frac{\langle w_1, v_{i+1} \rangle}{\langle w_1, w_1 \rangle} w_1 - \dots - \frac{\langle w_i, v_{i+1} \rangle}{\langle w_i, w_i \rangle} w_i.$$

Proposition 4.2.2. *The Gramm–Schmidt process works as stated: given w_1, \dots, w_i satisfying (1) and (2), and choosing w_{i+1} by the formula (4.2.1), the resulting set w_1, \dots, w_{i+1} satisfies (1) and (2) with i replaced by $i + 1$.*

Proof. Since w_1, \dots, w_i is an orthogonal set,

$$\langle w_j, w_{i+1} \rangle = \langle w_j, v_{i+1} \rangle - \frac{\langle w_j, v_{i+1} \rangle}{\langle w_j, w_j \rangle} \langle w_j, w_j \rangle = 0$$

for $1 \leq j \leq i$ by bilinearity and the fact that $\langle w_j, w_m \rangle = 0$ for $j \neq m \in \{1, \dots, i\}$. It suffices to show that $\text{span}(v_1, \dots, v_{i+1}) = \text{span}(w_1, \dots, w_{i+1})$: since v_1, \dots, v_{i+1} are linearly independent, this forces w_1, \dots, w_{i+1} to be linearly independent, and hence $w_{i+1} \neq 0$, making w_1, \dots, w_{i+1} an orthogonal set.

By the inductive assumption of (2), $\text{span}(w_1, \dots, w_i) = \text{span}(v_1, \dots, v_i)$. By Lemma 1.5.5, $\text{span}(w_1, \dots, w_i, x) = \text{span}(v_1, \dots, v_i, x)$ for any vector x .

In particular,

$$(4.2.2) \quad \text{span}(w_1, \dots, w_i, v_{i+1}) = \text{span}(v_1, \dots, v_{i+1}).$$

By (4.2.1), $w_{i+1} \in \text{span}(w_1, \dots, w_i, v_{i+1})$, so Lemma 1.5.5 gives

$$\text{span}(w_1, \dots, w_{i+1}) \subset \text{span}(v_1, \dots, v_{i+1}).$$

But (4.2.1) also gives $v_{i+1} \in \text{span}(w_1, \dots, w_{i+1})$, which also includes v_1, \dots, v_i by induction. So

$$\text{span}(v_1, \dots, v_{i+1}) \subset \text{span}(w_1, \dots, w_{i+1}). \quad \square$$

Corollary 4.2.3. *Every subspace $W \subset \mathbb{R}^n$ has an orthonormal basis.*

Proof. Start with a basis v_1, \dots, v_k of W and apply the Gram–Schmidt process. The resulting set z_1, \dots, z_k is orthonormal, and hence linearly independent. Its span is $\text{span}(v_1, \dots, v_k) = W$. \square

Remark 4.2.4. There is nothing sacred about restricting attention to \mathbb{R}^n in the above. The Gram–Schmidt process works precisely as stated for linearly independent vectors v_1, \dots, v_k in any inner product space V (Definition 2.3.9). In particular, every finite-dimensional inner product space admits an orthonormal basis.

4.3. Orthogonal complements. Lemma 4.1.3 has important applications to both theoretical and practical questions. We shall use it repeatedly. We shall now use it to study orthogonal complements, which are also important for both theoretical and practical questions.

Definition 4.3.1. Let $S \subset \mathbb{R}^n$ be any subset. We write S^\perp for the set of vectors orthogonal to every element of S :

$$S^\perp = \{v \in \mathbb{R}^n : \langle v, s \rangle = 0 \text{ for all } s \in S\}.$$

By the bilinearity of the inner product, S^\perp is a subspace of \mathbb{R}^n . We will be particularly interested in V^\perp , where V is a subspace of \mathbb{R}^n .

Definition 4.3.2. Let V be a subspace of \mathbb{R}^n . Then V^\perp is called the orthogonal complement of V .

Lemma 4.3.3. *Let $S = \{v_1, \dots, v_k\}$. Then S^\perp is the orthogonal complement of $\text{span}(v_1, \dots, v_k)$. In particular, if v_1, \dots, v_k is a basis for V , then $V^\perp = \{v_1, \dots, v_k\}^\perp$, the set of vectors orthogonal to each of v_1, \dots, v_k .*

Proof. Since $\{v_1, \dots, v_k\} \subset \text{span}(v_1, \dots, v_k)$,

$$\text{span}(v_1, \dots, v_k)^\perp \subset \{v_1, \dots, v_k\}^\perp.$$

The reverse inclusion follows from the bilinearity of the inner product: if a vector is orthogonal to each of v_1, \dots, v_k , then it must be orthogonal to any linear combination of v_1, \dots, v_k . \square

The following algorithm is useful both theoretically and practically.

Lemma 4.3.4. *Let V be a subspace of \mathbb{R}^n and let v_1, \dots, v_k be an orthonormal basis for V (obtained, for instance, by applying the Gram–Schmidt process to an arbitrary basis of V). Extend it to a basis $v_1, \dots, v_k, w_{k+1}, \dots, w_n$ of \mathbb{R}^n (e.g., by inductive application of Lemma 1.5.6(3)). Now apply the Gram–Schmidt process to $v_1, \dots, v_k, w_{k+1}, \dots, w_n$, respecting the stated order of the basis. Note this does not change the first k vectors, v_1, \dots, v_k , as they are already orthonormal. We obtain an orthonormal basis v_1, \dots, v_n of \mathbb{R}^n , where v_1, \dots, v_k is the original orthonormal basis of V .*

Then under this procedure, v_{k+1}, \dots, v_n is a basis for V^\perp .

Proof. Let $v \in \mathbb{R}^n$. By Lemma 4.1.3, we may write

$$v = \langle v, v_1 \rangle v_1 + \cdots + \langle v, v_n \rangle v_n.$$

By Lemma 4.3.3, $v \in V^\perp$ if and only if $\langle v, v_i \rangle = 0$ for $i = 1, \dots, k$. Thus, $v \in V^\perp$ if and only if $v \in \text{span}(v_{k+1}, \dots, v_n)$. \square

Corollary 4.3.5. *Let V be a subspace of \mathbb{R}^n . Then $\dim V + \dim V^\perp = n$. Moreover $(V^\perp)^\perp = V$.*

Proof. For the last statement,

$$\begin{aligned} (V^\perp)^\perp &= \{v_{k+1}, \dots, v_n\}^\perp \\ &= \text{span}(v_1, \dots, v_k) = V \end{aligned}$$

by the proof given above. \square

Corollary 4.3.6. *Let V be a subspace of \mathbb{R}^n . Let v_1, \dots, v_k be an orthonormal basis of V and w_1, \dots, w_l an orthonormal basis of V^\perp . Then*

$$v_1, \dots, v_k, w_1, \dots, w_l$$

is an orthonormal basis of \mathbb{R}^n .

Proof. By the orthogonality of V and V^\perp , $v_1, \dots, v_k, w_1, \dots, w_l$ is an orthonormal set. By Corollary 4.3.5, it has n elements. \square

Corollary 4.3.7. *Let V be a subspace of \mathbb{R}^n . Then each $y \in \mathbb{R}^n$ may be written uniquely in the form $y = v + w$ with $v \in V$ and $w \in V^\perp$. Indeed, if v_1, \dots, v_k is an orthonormal basis for V and w_1, \dots, w_{n-k} is an orthonormal basis for V^\perp there are linear functions $\pi_V : \mathbb{R}^n \rightarrow V$ and $\pi_{V^\perp} : \mathbb{R}^n \rightarrow V^\perp$ given by*

$$(4.3.1) \quad \begin{aligned} \pi_V(y) &= \langle y, v_1 \rangle v_1 + \cdots + \langle y, v_k \rangle v_k, \\ \pi_{V^\perp}(y) &= \langle y, w_1 \rangle w_1 + \cdots + \langle y, w_{n-k} \rangle w_{n-k}, \end{aligned}$$

called the orthogonal projections of \mathbb{R}^n onto V and V^\perp , respectively, and

$$(4.3.2) \quad y = \pi_V(y) + \pi_{V^\perp}(y) \quad \text{for all } y \in \mathbb{R}^n.$$

These projections are independent of the choices of orthonormal bases for V and V^\perp and satisfy

$$(4.3.3) \quad \pi_V|_V = \text{id}_V \quad \pi_{V^\perp}|_{V^\perp} = \text{id}_{V^\perp}$$

$$(4.3.4) \quad \pi_V|_{V^\perp} = 0 \qquad \pi_{V^\perp}|_V = 0.$$

In consequence, we have

$$(4.3.5) \quad \operatorname{Im} \pi_V = V \qquad \operatorname{Im} \pi_{V^\perp} = V^\perp$$

$$(4.3.6) \quad \ker \pi_V = V^\perp \qquad \ker \pi_{V^\perp} = V.$$

Finally,

$$(4.3.7) \quad \pi_V \circ \pi_V = \pi_V \qquad \pi_{V^\perp} \circ \pi_{V^\perp} = \pi_{V^\perp}.$$

Proof. The maps in (4.3.1) are linear by the bilinearity of the inner product. (4.3.3) is immediate from Lemma 4.1.3, and Corollary 4.3.6 gives (4.3.2). This gives the desired decomposition $y = v + w$ with $v \in V$ and $w \in V^\perp$.

(4.3.4) follows since $\langle v_i, w_j \rangle = 0$ for $i = 1, \dots, k$ and $j = 1, \dots, n - k$. This in turn gives the uniqueness of the decomposition $y = v + w$ as follows: let $y = v + w$ with $v \in V$ and $w \in V^\perp$. Then

$$(4.3.8) \quad \pi_V(y) = \pi_V(v) + \pi_V(w) = v + 0 = v,$$

as π_V restricts to the identity on V and the zero map on V^\perp . Similarly, $\pi_{V^\perp}(y) = w$. This shows both the uniqueness of the decomposition $y = v + w$ and the independence of the projection maps from the choices of bases. The remaining results follow from this, also. \square

Corollary 4.3.7 can be partially restated in terms of direct sum decomposition.

Corollary 4.3.8. *Let V be a subspace of \mathbb{R}^n . Then there are inverse isomorphisms*

$$\begin{aligned} \iota : V \oplus V^\perp &\xrightarrow{\cong} \mathbb{R}^n, & \iota(v, w) &= v + w, \\ \pi : \mathbb{R}^n &\xrightarrow{\cong} V \oplus V^\perp, & \pi(y) &= (\pi_V(y), \pi_{V^\perp}(y)). \end{aligned}$$

Proof. Corollary 4.3.7 shows $\iota \circ \pi = \operatorname{id}_{\mathbb{R}^n}$. And (4.3.8) and its analogue for π_{V^\perp} show that $\pi \circ \iota = \operatorname{id}_{V \oplus V^\perp}$. \square

In studying linear isometries, orthogonal complements are extremely important. A key idea in linear algebra is invariant subspaces.

Definition 4.3.9. Let A be an $n \times n$ matrix. An invariant subspace of A (or of T_A) is a subspace $V \subset \mathbb{R}^n$ such that $Av \in V$ for all $v \in V$, i.e., $T_A(V) \subset V$. To save words, we will simply say V is A -invariant (or T_A -invariant).

The simplest example of invariant subspaces comes from eigenvectors.

Lemma 4.3.10. *A one-dimensional subspace $\operatorname{span}(v)$ is A -invariant if and only if v is an eigenvector for (A, c) for some c .*

Proof. $\operatorname{span}(v)$ is one-dimensional if and only if $v \neq 0$. Since $A \cdot av = aAv$, $\operatorname{span}(v)$ is A -invariant if and only if $Av \in \operatorname{span}(v)$, i.e., if and only if $Av = cv$ for some $c \in \mathbb{R}$. \square

More generally, of course, the eigenspace of (A, c) is A -invariant for any eigenvalue c of A . Lemma 4.3.10 can be used to find additional invariant subspaces if A is orthogonal. We shall make use of this in studying $O(3)$, below.

Lemma 4.3.11. *Let V be an invariant subspace for an orthogonal matrix A . Then V^\perp is A -invariant as well.*

Proof. We first show that $A(V) = V$, i.e., that the restriction of T_A to V carries V onto V . To see this, let v_1, \dots, v_k be an orthonormal basis of V . Then Av_1, \dots, Av_k is an orthonormal set contained in V , and hence a basis for V , as $\dim V = k$. So the range of the restriction of T_A to V is V .

This implies that for $v \in V$, $A^{-1}v \in V$, so V is A^{-1} -invariant as well.

Let $w \in V^\perp$. We wish to show $Aw \in V^\perp$ as well. Recall that A^{-1} is orthogonal. Thus, for $v \in V$,

$$\langle v, Aw \rangle = \langle A^{-1}v, A^{-1}Aw \rangle = \langle A^{-1}v, w \rangle = 0,$$

as $w \in V^\perp$ and $A^{-1}v \in V$. □

4.4. Applications to rank. The notion of orthogonal complement is useful in studying nonorthogonal matrices as well. Recall the following.

Definition 4.4.1. Let A be $m \times n$. Then $\ker T_A = \{x \in \mathbb{R}^n : Ax = 0\}$ is called the null space of A , and may be denoted $N(A)$.

Note that by Corollary 4.1.10, $x \in N(A)$ if and only if x is orthogonal to every column of A^T . We obtain the following.

Lemma 4.4.2. *Let A be $m \times n$, then $N(A)$ is the orthogonal complement of the span of the columns of A^T . Since the columns span the range of A^T we get*

$$(4.4.1) \quad N(A) = A^T(\mathbb{R}^m)^\perp.$$

Thus, $\dim N(A) = n - \text{rank}(A^T)$, and hence $\text{rank } A = \text{rank}(A^T)$.

In particular, this gives us an algorithm for finding a basis for $\{v_1, \dots, v_k\}^\perp$ for $v_1, \dots, v_k \in \mathbb{R}^n$: set $A = [v_1 | \dots | v_k]$ and then use Gauss elimination to find a basis for $N(A^T) = A(\mathbb{R}^k)^\perp$, since $(A^T)^T = A$.

Since the rank of A is the dimension of the span of its columns, $\text{rank } A$ is the maximal number of linearly independent columns of A . Let $A = [v_1, \dots, v_n]$ and $r = \text{rank } A$. Suppose that v_{i_1}, \dots, v_{i_r} are linearly independent and let $B = [v_{i_1} | \dots | v_{i_r}]$. Then $r = \text{rank } B = \text{rank}(B^T)$ so B^T has r linearly independent columns, hence B has r linearly independent rows. Let C be the $r \times r$ matrix obtained from B by restricting to those r linearly independent rows. Then C^T has rank r , hence so does C , hence C is invertible. In particular $\det C \neq 0$. We obtain:

Corollary 4.4.3. *Let A be $m \times n$ with $r = \text{rank } A$. Then we can throw away some of the rows and columns of A to obtain an invertible $r \times r$ matrix. Conversely, if A has such an $r \times r$ invertible submatrix, then $\text{rank } A \geq r$. Thus, $\text{rank } A$ is the largest number r such that A has an invertible $r \times r$ submatrix obtained by throwing away some of the rows and columns of A .*

Proof. We prove the converse. If A has such a submatrix, C , and if B is obtained from A by deleting the columns but not the rows, then B is an $m \times r$ matrix r of whose rows are linearly independent. So $\text{rank } B = r$, and the columns of B are linearly independent. So A has at least r linearly independent columns and $\text{rank } A \geq r$. \square

4.5. Invariant subspaces for linear isometries. For simplicity, we work for the moment in \mathbb{R}^n , though the result generalizes to an arbitrary finite-dimensional inner product space.

Proposition 4.5.1. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a linear isometry and let V be an f -invariant subspace. Then V^\perp is f -invariant by Lemma 4.3.11. Let $\mathcal{B}' = v_1, \dots, v_k$ be an orthonormal basis for V and let $\mathcal{B}'' = w_1, \dots, w_{n-k}$ be an orthonormal basis for V^\perp . Let $\mathcal{B} = v_1, \dots, v_k, w_1, \dots, w_{n-k}$, a basis of \mathbb{R}^n by Corollary 4.3.6. Then*

$$[f]_{\mathcal{B}} = \left[\begin{array}{c|c} [f|_V]_{\mathcal{B}'} & 0 \\ \hline 0 & [f|_{V^\perp}]_{\mathcal{B}''} \end{array} \right],$$

hence $\det[f]_{\mathcal{B}} = \det[f|_V]_{\mathcal{B}'} \det[f|_{V^\perp}]_{\mathcal{B}''}$.

Proof. Simply apply the proof of Lemma 1.8.18, noting in this case that $X = 0$, as V^\perp is f -invariant. \square

Proposition 4.5.2. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be linear and let $\mathcal{B} = v_1, \dots, v_n$ be an orthonormal basis of \mathbb{R}^n . Then f is a linear isometry if and only if $[f]_{\mathcal{B}}$ is orthogonal.*

Proof. We already know f is an isometry if and only if $[f] = [f]_{\mathcal{E}}$ is orthogonal. But if \mathcal{B} is orthonormal, then $[I]_{\mathcal{E}\mathcal{B}} = [v_1 | \dots | v_n]$ is orthogonal. $[f]_{\mathcal{B}} = [I]_{\mathcal{E}\mathcal{B}}^{-1} [f]_{\mathcal{E}} [I]_{\mathcal{E}\mathcal{B}}$. Since $O(n)$ is a subgroup of $GL(n, \mathbb{R})$, $[f]_{\mathcal{E}}$ is orthogonal if and only if $[f]_{\mathcal{B}}$ is orthogonal. \square

We can generalize this to inner product spaces, and in process give a different proof of Proposition 4.5.2. The point is that an inner product space V has a norm and distance coming from the inner product in exactly the same way the norm and distance come from the inner product in \mathbb{R}^n : $\|v\| = \sqrt{\langle v, v \rangle}$ and $d(v, w) = \|v - w\|$. The properties of the Euclidean norm and distance immediately generalize to this context, and Theorem 4.1.5 generalizes to the following.

Theorem 4.5.3. *Let V and W be inner product spaces. Let v_1, \dots, v_n be an orthonormal basis of V and let w_1, \dots, w_n be arbitrary. Let $f : V \rightarrow W$*

be the unique linear function with $f(v_i) = w_i$ for $i = 1, \dots, n$. Then the following conditions are equivalent:

- (1) f is an isometry, i.e., $d(f(v), f(v')) = d(v, v')$ for all $v, v' \in V$.
- (2) $\|f(v)\| = \|v\|$ for all $v \in V$.
- (3) w_1, \dots, w_n is an orthonormal set.
- (4) $\langle f(v), f(v') \rangle = \langle v, v' \rangle$ for all $v, v' \in V$.

Proof. The proof of Theorem 4.1.5 generalizes word for word. □

Corollary 4.5.4. Let $\mathcal{B} = v_1, \dots, v_n$ be an orthonormal basis for the inner product space V and let $f : V \rightarrow V$ be linear. Then f is an isometry if and only if $[f]_{\mathcal{B}}$ is orthogonal.

Proof. Let $\Phi_{\mathcal{B}} : \mathbb{R}^n \rightarrow V$ be the isomorphism induced by \mathcal{B} . Then $\Phi_{\mathcal{B}}$ and $\Phi_{\mathcal{B}}^{-1}$ are isometries by Theorem 4.5.3. Thus, if f is an isometry, so is $T = \Phi_{\mathcal{B}}^{-1} \circ f \circ \Phi_{\mathcal{B}}$, hence the matrix of T is orthogonal. But the matrix of T is $[f]_{\mathcal{B}}$.

The converse follows similarly, as if T is an isometry, so is

$$f = \Phi_{\mathcal{B}} \circ T \circ \Phi_{\mathcal{B}}^{-1}. \quad \square$$

5. Isometries of \mathbb{R}^2

5.1. Reflections. Orthonormal basis is the key idea needed to understand reflections. We may reflect across a line in \mathbb{R}^2 , a plane in \mathbb{R}^3 , etc. In this section, we study reflections across lines in \mathbb{R}^2 .

Let $\ell = x + \text{span}(y)$ be a line in \mathbb{R}^n . Then $\text{span}(y)$ is a line through the origin. As such, it contains exactly two unit vectors, $v = \frac{y}{\|y\|}$ and $-v$. These may be thought of as giving orientations to $\text{span}(y) = \text{span}(v)$, as

$$\text{span}(v) = \{tv : t \in \mathbb{R}\}$$

gives a parametrization of $\text{span}(v)$ in which v appears on the positive side of 0 and $-v$ on the negative. These positions reverse if we parametrize $\text{span}(v)$ as $\{t(-v) : t \in \mathbb{R}\}$. Formally:

Definition 5.1.1. An orientation for a line $\ell = x + \text{span}(y)$ in \mathbb{R}^n is a choice of unit vector $v \in \text{span}(y)$, i.e., a unit vector parallel to ℓ .

So every line has exactly two orientations. An orientation of a line may be thought of as a choice of orthonormal basis for $\text{span}(y)$. The same idea may be used to orient a plane in \mathbb{R}^n , except we must then use an equivalence class of orthonormal bases. We will discuss this later. The following trick is useful for extending an orthonormal basis of $\text{span}(y)$ to an orthonormal basis of the plane.

Definition 5.1.2. For $y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \in \mathbb{R}^2$, write $y^\perp = \begin{bmatrix} -y_2 \\ y_1 \end{bmatrix}$.

This does not conflict with the notation $\{y\}^\perp = \text{span}(y)^\perp$, and they are certainly related:

Lemma 5.1.3. For $y \neq 0$, y and y^\perp are orthogonal, and

$$\text{span}(y^\perp) = \text{span}(y)^\perp,$$

the orthogonal complement of $\text{span}(y)$. Also, $\|y\| = \|y^\perp\|$ and the slope of $\text{span}(y^\perp)$ is $\frac{y_1}{-y_2}$, the negative reciprocal of the slope of $\text{span}(y)$.

Proof. y^\perp is nonzero and orthogonal to y . so $\text{span}(y^\perp)$ is a 1-dimensional subspace of $\{y\}^\perp = \text{span}(y)^\perp$, which, by Corollary 4.3.5 is 1-dimensional. The rest is straightforward. \square

Note that if $v = \frac{y}{\|y\|}$, then $v^\perp = \frac{y^\perp}{\|y^\perp\|}$ gives one of the two orientations for $\text{span}(y^\perp)$. In particular, we have the following.

Corollary 5.1.4. For any unit vector $v \in \mathbb{R}^2$ there are exactly two unit vectors orthogonal to it: $\pm v^\perp$. Thus, there are exactly two orthonormal bases whose first vector is v : v, v^\perp and $v, -v^\perp$.

Back to lines, we can make the following definition.

Definition 5.1.5. Let $\ell = x + \text{span}(y)$ be a line in \mathbb{R}^2 . A unit normal N for ℓ is a unit vector N with $\text{span}(N) = \text{span}(y)^\perp$.

Corollary 5.1.4 gives the following.

Corollary 5.1.6. *A line $\ell = x + \text{span}(y)$ in \mathbb{R}^2 has exactly two unit normals: $\pm v^\perp$, with $v = \frac{y}{\|y\|}$. Moreover, for either choice of unit normal N , $\mathcal{B} = v, N$ is an orthonormal basis of \mathbb{R}^2 .*

We can use these orthonormal bases to obtain a grid based on the line ℓ .

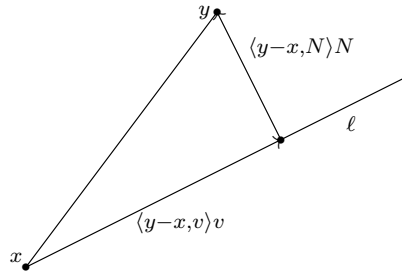
Theorem 5.1.7. *Let $\ell = x + \text{span}(v)$ be a line in \mathbb{R}^2 with v a unit vector. Let N be a unit normal for ℓ . Then every $y \in \mathbb{R}^2$ may be written uniquely as the sum*

$$y = (x + sv) + tN$$

of a point $x + sv \in \ell$ and a point $tN \in \text{span}(N) = \text{span}(v)^\perp$. The coordinates s and t are affine functions $\mathbb{R}^2 \rightarrow \mathbb{R}$:

$$(5.1.1) \quad \begin{aligned} s(y) &= \langle y - x, v \rangle \\ t(y) &= \langle y - x, N \rangle. \end{aligned}$$

Moreover, the value $t(y)$ is independent of the choice of $x \in \ell$ (but $s(y)$ is not).



Proof. If $x = 0$ this is just a restatement of Lemma 4.1.3. If $x \neq 0$, then $0 \in \tau_{-x}(\ell) = \text{span}(v)$, and we get

$$\tau_{-x}(y) = \langle \tau_{-x}(y), v \rangle v + \langle \tau_{-x}(y), N \rangle N.$$

So

$$\begin{aligned} y &= \tau_x \tau_{-x}(y) = x + \langle \tau_{-x}(y), v \rangle v + \langle \tau_{-x}(y), N \rangle N \\ &= x + sv + tN, \end{aligned}$$

with $s = s(y)$ and $t = t(y)$ as in (5.1.1), as $\tau_{-x}(y) = y - x$. Uniqueness of s and t follow from the uniqueness of coefficients after applying τ_{-x} (Lemma 4.1.3). s and t are affine as each is the composite of a linear function and a translation.

To see that t is independent of the choice of x , note that any other point $z \in \ell$ has the form $z = x + cv$ for some $c \in \mathbb{R}$, and hence

$$\langle y - z, N \rangle = \langle y - x - cv, N \rangle = \langle y - x, N \rangle - c \langle v, N \rangle = \langle y - x, N \rangle,$$

as $\langle v, N \rangle = 0$. □

This allows us to define the reflection of \mathbb{R}^2 across ℓ .

Definition 5.1.8. Let $\ell = x + \text{span}(v)$ be a line in \mathbb{R}^2 with v a unit vector. Let N be a unit normal for ℓ . The reflection, σ_ℓ , of \mathbb{R}^2 across ℓ takes $y = (x + sv) + tN$ to $(x + sv) - tN$. Thus,

$$(5.1.2) \quad \sigma_\ell(y) = y - 2t(y)N = y - 2\langle y - x, N \rangle N,$$

with $t(y) = \langle y - x, N \rangle$ as studied in Theorem 5.1.7.

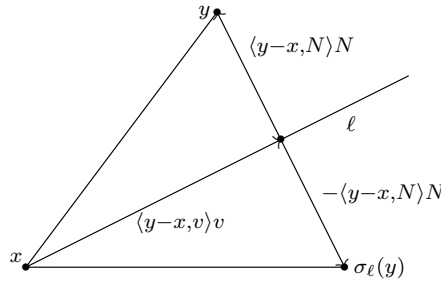
The formula (5.1.2) (and therefore the function σ_ℓ) is independent of the choice of unit normal: the only other choice would be $-N$, and

$$\langle y - x, -N \rangle (-N) = \langle y - x, N \rangle N$$

by the bilinearity of the inner product.

We call ℓ the axis of σ_ℓ and call σ_ℓ the reflection with axis ℓ .

(5.1.3)



The following is immediate from the definition. A function whose square is the identity is called an involution.

Lemma 5.1.9. $\sigma_\ell \circ \sigma_\ell = \text{id}$, so reflections are involutions.

Reflections give us a new family of isometries.

Proposition 5.1.10. Let ℓ be a line in \mathbb{R}^2 . Then σ_ℓ is an isometry.

Proof. Let $\ell = w + \text{span}(v)$ with v a unit vector and N a unit normal. Let $x, y \in \mathbb{R}^2$. Then,

$$\begin{aligned} x &= w + \langle x - w, v \rangle v + \langle x - w, N \rangle N, \\ y &= w + \langle y - w, v \rangle v + \langle y - w, N \rangle N. \end{aligned}$$

So

$$\begin{aligned} (5.1.4) \quad d(\sigma_\ell(x), \sigma_\ell(y)) &= \|(w + \langle y - w, v \rangle v - \langle y - w, N \rangle N) \\ &\quad - (w + \langle x - w, v \rangle v - \langle x - w, N \rangle N)\| \\ &= \|(\langle y - w, v \rangle - \langle x - w, v \rangle)v \\ &\quad - (\langle y - w, N \rangle - \langle x - w, N \rangle)N\| \\ &= \|\langle y - x, v \rangle v - \langle y - x, N \rangle N\|. \end{aligned}$$

Because v, N is an orthonormal basis, $\|av + bN\| = \sqrt{a^2 + b^2}$ for any $a, b \in \mathbb{R}$ (Corollary 4.1.6). In particular, the last line of (5.1.4) is equal to

$$\|\langle y - x, v \rangle v + \langle y - x, N \rangle N\|,$$

which is just $\|y - x\|$ by Lemma 4.1.3. \square

We now have two infinite families of isometries of \mathbb{R}^2 : translations and reflections. These two families differ in two important ways. First, translations preserve orientation of the plane (to be discussed below) and reflections reverse it. Second, translations have no fixed-points and reflections do have fixed-points.

Definition 5.1.11. Let $\alpha : X \rightarrow X$ be a function. The fixed-point set, X^α , of α is

$$X^\alpha = \{x \in X : \alpha(x) = x\}.$$

We leave it to the reader to show $(\mathbb{R}^n)^{\tau_x} = \emptyset$ for all $0 \neq x \in \mathbb{R}^n$.

Lemma 5.1.12. Let ℓ be a line in \mathbb{R}^2 . Then

$$(\mathbb{R}^2)^{\sigma_\ell} = \ell,$$

i.e., the fixed-point set of σ_ℓ is precisely ℓ .

Proof. Let $\ell = x + \text{span}(v)$ with v a unit vector. Let N be a unit normal for ℓ . From the definition of σ_ℓ we see $\sigma_\ell(y) = y$ if and only if $\langle y - x, N \rangle = 0$, and this holds if and only if $y - x \in \text{span}(v) = \{N\}^\perp$. But that is equivalent to saying $y \in \ell$. \square

The output of Theorem 5.1.7 is useful as it shows that the complement of ℓ in \mathbb{R}^2 is the union of two convex pieces.

Corollary 5.1.13. Let $\ell = x + \text{span}(v)$ be a line in \mathbb{R}^2 with v a unit vector and N a unit normal. Let

$$c = \langle x, N \rangle.$$

Then

$$(5.1.5) \quad \ell = \{y \in \mathbb{R}^2 : \langle y, N \rangle = c\} = \{y \in \mathbb{R}^2 : \langle y - x, N \rangle = 0\}.$$

Define the positive and negative parts with respect to N of the complement of ℓ to be

$$\begin{aligned} (\mathbb{R}^2 - \ell)^+ &= \{y \in \mathbb{R}^2 : \langle y, N \rangle > c\} = \{y \in \mathbb{R}^2 : \langle y - x, N \rangle > 0\} \\ (\mathbb{R}^2 - \ell)^- &= \{y \in \mathbb{R}^2 : \langle y, N \rangle < c\} = \{y \in \mathbb{R}^2 : \langle y - x, N \rangle < 0\}. \end{aligned}$$

In particular, these depend on the orientation of $\text{span}(v)^\perp$ given by N .

Then each of these parts is convex, and their union is $\mathbb{R}^2 - \ell$, the complement of ℓ in \mathbb{R}^2 . If y and z are in different parts, the line segment from y to z intersects ℓ .

Finally, the reflection σ_ℓ interchanges these two pieces.

Proof. To establish (5.1.5), note that

$$\langle x + tv, N \rangle = \langle x, N \rangle + t\langle v, N \rangle = \langle x, N \rangle = c,$$

as v is orthogonal to N . Moreover, if $\langle y - x, N \rangle = 0$, then $y - x$ is orthogonal to N , and hence lies in $\text{span}(v)$.

The rest is immediate from Theorem 5.1.7, as $t : \mathbb{R}^2 \rightarrow \mathbb{R}$ is affine. The last sentence follows as σ_ℓ changes the sign of t . \square

The following will prove useful.

Lemma 5.1.14. *Let ℓ be a line in \mathbb{R}^2 and let $m = \tau_x(\ell)$ for $x \in \mathbb{R}^2$. Then*

$$\sigma_m = \tau_x \sigma_\ell \tau_{-x}.$$

Proof. This is equivalent to showing $\sigma_m \tau_x = \tau_x \sigma_\ell$. Let $\ell = z + \text{span}(v)$ with v a unit vector. Then $m = \tau_x(z) + \text{span}(v)$, so the two lines have the same unit normal, $N = v^\perp$. Let $y \in \mathbb{R}^2$ and write

$$y = z + sv + tN.$$

Then

$$\begin{aligned} \sigma_m \tau_x(y) &= \sigma_m(\tau_x(z) + sv + tN) \\ &= \tau_x(z) + sv - tN \\ &= x + (z + sv - tN) \\ &= \tau_x(\sigma_\ell(y)). \end{aligned} \quad \square$$

We now introduce the classical Euclidean notion of dropping a perpendicular.

Definition 5.1.15. The lines $\ell = x + \text{span}(y)$ and $m = z + \text{span}(w)$ are perpendicular (written $\ell \perp m$) if y and w are orthogonal, i.e., if $\text{span}(w) = \text{span}(y^\perp)$. We say a nonzero vector z is perpendicular to ℓ ($z \perp \ell$) if z is orthogonal to y .

Corollary 5.1.16. *For any line $\ell = x + \text{span}(y)$ and any $z \in \mathbb{R}^2$ there is a unique line through z perpendicular to ℓ : the line $z + \text{span}(y^\perp)$.*

This allows us to define the perpendicular bisector of a line segment.

Definition 5.1.17. Let $x \neq y \in \mathbb{R}^2$. The perpendicular bisector of the segment \overline{xy} is the line $z + \text{span}((y - x)^\perp)$ with $z = \frac{x+y}{2}$, the midpoint of the line segment \overline{xy} .

This is perpendicular to \overleftrightarrow{xy} by Corollary 2.1.8 and bisects \overline{xy} as it passes through the midpoint.

Proposition 5.1.18. *Let ℓ be a line in \mathbb{R}^2 and let $y \notin \ell$. Then ℓ is the perpendicular bisector of $\overline{y\sigma_\ell(y)}$.*

Conversely if $x \neq y \in \mathbb{R}^2$ and if ℓ is the perpendicular bisector of \overline{xy} , then σ_ℓ exchanges x and y .

Proof. Let $\ell = w + \text{span}(v)$ with v a unit vector. Let N be a unit normal for ℓ and let $y \notin \ell$. Then

$$y = w + \langle y - w, v \rangle v + \langle y - w, N \rangle N,$$

so the midpoint, z , of $\overline{y\sigma_\ell(y)}$ is given by

$$z = \frac{y + \sigma_\ell(y)}{2} = w + \langle y - w, v \rangle v,$$

which lies in ℓ . In particular, $\ell = z + \text{span}(v)$. Moreover,

$$y - \sigma_\ell(y) = 2\langle y - w, N \rangle N,$$

so $(y - \sigma_\ell(y))^\perp$ is a multiple of v , and hence ℓ is the perpendicular bisector of $\overline{y\sigma_\ell(y)}$, as claimed.

Conversely, let $x \neq y \in \mathbb{R}^2$, let $z = \frac{x+y}{2}$ and let v be a unit vector orthogonal to $y - x$, so that $\ell = z + \text{span}(v)$ is the perpendicular bisector of \overline{xy} . Say $N = \frac{y-x}{\|y-x\|}$. Then

$$\begin{aligned} x &= z - tN \\ y &= z + tN \end{aligned}$$

where $t = \frac{\|y-x\|}{2}$. So σ_ℓ does exchange x and y . \square

There is an important relationship between reflections and translations: the product of two reflections through parallel lines is a translation, and every translation can be obtained that way. Let us first define the directed distance between two parallel lines. Recall that two lines ℓ and m are parallel if they are translations of the same line through the origin.

Definition 5.1.19. Let $\ell = x + \text{span}(v)$ and $m = y + \text{span}(v)$ be parallel lines in \mathbb{R}^2 . The directed distance from ℓ to m is the vector obtained as follows: let n be any line perpendicular to ℓ (and hence also to m). Then the directed distance from ℓ to m is $n \cap m - n \cap \ell$, i.e., the vector whose initial point is $n \cap \ell$ and whose endpoint is $n \cap m$.

Proposition 5.1.20. *Let $\ell \parallel m$. Then the directed distance from ℓ to m is the unique vector $w \perp \ell$ such that $m = \tau_w(\ell)$. In particular, the directed distance is independent of the choice of the line n perpendicular to ℓ .*

Proof. Let $n \perp \ell$ and let $x = n \cap \ell$. Since $x \in \ell$, $\ell = x + \text{span}(v)$ for a unit vector $v \parallel \ell$. Since $x \in n$ and $n \perp \ell$, $n = x + \text{span}(N)$, where N is a unit normal for ℓ . In particular, $n \cap m = x + tN$ for some t . So

$$n \cap m - n \cap \ell = (x + tN) - x = tN$$

is perpendicular to ℓ , and

$$\tau_{tN}(\ell) = \tau_{tN}(x + \text{span}(v)) = (x + tN) + \text{span}(v) = m,$$

as $x + tN \in m$.

For the uniqueness, if $w \perp \ell$, then $w = uN$ for some u . With n and x as above, $\tau_w(\ell) = (x + uN) + \text{span}(v)$. If this is equal to $m = (x + tN) + \text{span}(v)$,

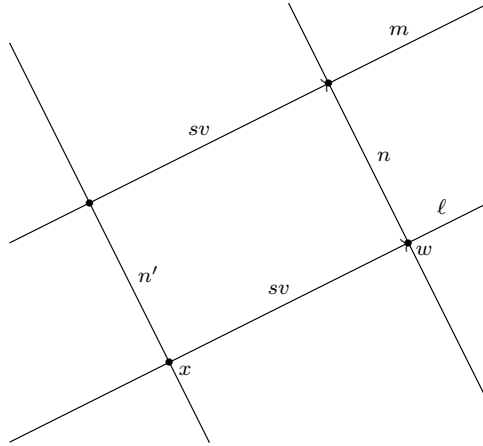
then $(x+tN)-(x+uN) \in \text{span}(v)$, i.e., $(t-u)N \in \text{span}(v)$. Since $\langle v, N \rangle = 0$, $t-u=0$, and uniqueness is achieved. \square

A perhaps simpler geometric argument for the independence part of the above is as follows:

Second proof that the directed distance between parallel lines is independent of the choice of the perpendicular n . Write

$$\ell = x + \text{span}(v) \quad \text{and} \quad m = y + \text{span}(v)$$

with v a unit vector. Let N be a unit normal for ℓ (and hence for m). Let $n \perp \ell$ and let $w = n \cap \ell$. Then $n = w + \text{span}(N)$.



We have an alternative perpendicular to ℓ given by $n' = x + \text{span}(N)$. By the prescription above, it suffices to show that

$$(5.1.6) \quad n \cap m - w = n' \cap m - x.$$

Since $w \in \ell$, we may write $w = x + sv$ for some $s \in \mathbb{R}$. We may and shall assume $s \neq 0$. In particular, $w = \tau_{sv}(x)$. By (5.1.6), it suffices to show $n \cap m = n' \cap m + sv = \tau_{sv}(n' \cap m)$.

Since $sv \parallel m$, $\tau_{sv}(m) = m$ by Proposition 2.1.14, so $\tau_{sv}(n' \cap m) \in m$. So $\tau_{sv}(n' \cap m) = \tau_{sv}(n') \cap m$. But

$$\tau_{sv}(n') = (x + sv) + \text{span}(N) = w + \text{span}(N) = n,$$

and the result follows. \square

We can now compute the composition of two reflections in parallel lines.

Proposition 5.1.21. *Let ℓ and m be parallel lines in \mathbb{R}^2 . Then $\sigma_m \sigma_\ell$ is the translation by twice the directed distance from ℓ to m . In particular, if $v \perp \ell$ and $m = \tau_v(\ell)$, then $\sigma_m \sigma_\ell = \tau_{2v}$.*

Proof. We may assume $\ell \neq m$, as we already know reflections are involutions. Write $\ell = x + \text{span}(v)$ with v a unit vector. Let N be a unit normal for ℓ . Let $n = x + \text{span}(N)$ and let $y = n \cap m$. Then $m = y + \text{span}(v)$ and it suffices to show

$$(5.1.7) \quad \sigma_m \sigma_\ell = \tau_{2(y-x)}.$$

Let $z \in \mathbb{R}^2$. Then

$$\begin{aligned} \sigma_m \sigma_\ell(z) &= \sigma_m(z - 2\langle z - x, N \rangle N) \\ &= z - 2\langle z - x, N \rangle N - 2\langle z - 2\langle z - x, N \rangle N - y, N \rangle N \\ &= z - 2\langle z - x, N \rangle N - 2\langle z - y, N \rangle N + 4\langle z - x, N \rangle \langle N, N \rangle N \\ &= z + 2\langle (z - x) - (z - y), N \rangle N \\ &= z + 2\langle y - x, N \rangle N. \end{aligned}$$

Since $y \in n = x + \text{span}(N)$, $y - x = cN$ for some c , hence

$$\langle y - x, N \rangle N = \langle cN, N \rangle N = c\langle N, N \rangle N = cN = y - x,$$

so (5.1.7) follows. \square

Corollary 5.1.22. *Let $0 \neq w \in \mathbb{R}^2$. Let ℓ be any line perpendicular to w , say $\ell = x + \text{span}(w^\perp)$. Let $m = \tau_{\frac{w}{2}}(\ell) = (x + \frac{w}{2}) + \text{span}(w^\perp)$ and let $n = \tau_{-\frac{w}{2}}(\ell) = (x - \frac{w}{2}) + \text{span}(w^\perp)$. Then*

$$\tau_w = \sigma_m \sigma_\ell = \sigma_\ell \sigma_n.$$

The characterization of line segments in terms of distance given in Proposition 2.3.8 has been of considerable help in understanding isometries. We have a similarly useful characterization of perpendicular bisectors in terms of distance.

Proposition 5.1.23. *Let $x \neq y \in \mathbb{R}^2$ and let ℓ be the perpendicular bisector of \overline{xy} . Then*

$$(5.1.8) \quad \ell = \{z \in \mathbb{R}^2 : d(x, z) = d(y, z)\}.$$

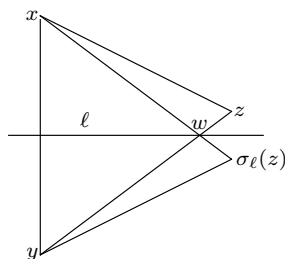
Proof. One direction is easy: if $z \in \ell$, then

$$\begin{aligned} d(x, z) &= d(\sigma_\ell(x), \sigma_\ell(z)) && (\sigma_\ell \text{ is an isometry}) \\ &= d(y, z), \end{aligned}$$

by Proposition 5.1.18 and Lemma 5.1.12.

Thus, suppose $z \in \mathbb{R}^2$ with $d(x, z) = d(y, z)$. Suppose, by contradiction that $z \notin \ell$. By construction, x and y are on opposite sides of ℓ under the decomposition of Corollary 5.1.13. Say x and z are in $(\mathbb{R}^2 - \ell)^+$ and $y \in (\mathbb{R}^2 - \ell)^-$. By Corollary 5.1.13, the line segment \overline{xz} is contained in $(\mathbb{R}^2 - \ell)^+$, while \overline{yz} crosses ℓ . Say $\overline{yz} \cap \ell = w$.

$$\begin{aligned} d(x, z) &= d(y, z) && (\text{by assumption}) \\ &= d(y, w) + d(w, z) && (\text{Proposition 2.3.8}) \end{aligned}$$



$$\begin{aligned}
 &= d(\sigma_\ell(y), \sigma_\ell(w)) + d(w, z) && \text{(reflections are isometries)} \\
 &= d(x, w) + d(w, z),
 \end{aligned}$$

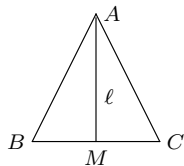
as σ_ℓ exchanges x and y , and w lies on ℓ , the fixed-point set of σ_ℓ . But now Proposition 2.3.8 forces w to be on \overline{xz} , contradicting that \overline{xz} doesn't intersect ℓ . So $z \in \ell$ as claimed. \square

The following is now useful for congruence proofs such as the side-side-side theorem.

Corollary 5.1.24. *Let $x \neq y \in \mathbb{R}^2$ and let $z \neq w \in \mathbb{R}^2$ with $d(x, z) = d(y, z)$ and $d(x, w) = d(y, w)$. Let ℓ be the unique line containing z and w . Then ℓ is the perpendicular bisector of \overline{xy} and hence σ_ℓ interchanges x and y .*

And here is an illustration of its use in Euclidean geometry. We shall not discuss angle measure until Section 5.4. All we need for the discussion here is that unsigned angle measure is preserved by isometries (Proposition 5.4.9).

Proposition 5.1.25 (Pons asinorum). *Let $\triangle ABC$ be a triangle with two sides of equal length. Say $d(A, B) = d(A, C)$.*



Then the angles opposite these two sides have equal (unsigned) measure.

Proof. Note that by Lemma 2.4.3, isometries preserve line segments, so if α is an isometry, then $\alpha(\triangle ABC) = \triangle \alpha(A)\alpha(B)\alpha(C)$. So our notion of congruence via isometries is compatible with Euclidean geometry.

Let M be the midpoint of \overline{BC} . Then

$$d(M, B) = d(M, C) = \frac{1}{2}d(B, C).$$

So M lies on the perpendicular bisector of \overline{BC} (in fact, it lies there by definition of the perpendicular bisector). Since $d(A, B) = d(A, C)$, A also

lies on the perpendicular bisector. So the unique line ℓ containing A and M is the perpendicular bisector. So σ_ℓ interchanges B and C .

Since $A \in \ell$, σ_ℓ fixes A . Since σ_ℓ is an isometry, it gives a congruence from $\triangle ABC$ to itself that fixes A and exchanges B and C . So it exchanges the angles $\angle ABC$ and $\angle ACB$. We shall show in Proposition 5.4.9, that isometries preserve angle measure. So $\angle ABC$ and $\angle ACB$ have the same measure. \square

5.2. Trigonometric functions. We develop the basic properties of trig functions here as they are essential to studying the linear isometries of \mathbb{R}^2 .

A key property of the trigonometric functions is the following, which is often presented as some form of revealed truth. We shall derive it here, along with the other important trig identities. Unless otherwise stated, all angle measures in this book will be in radians.

Theorem 5.2.1. *Let $\theta, \phi \in \mathbb{R}$. Then*

$$(5.2.1) \quad \cos(\theta + \phi) = \cos \theta \cos \phi - \sin \theta \sin \phi$$

$$(5.2.2) \quad \sin(\theta + \phi) = \cos \theta \sin \phi + \sin \theta \cos \phi.$$

To prove this we will exploit the relationship between the trig functions and complex exponentials. So we shall assume some basic material on real and complex power series. The reader is welcome to skip this section and simply apply the above theorem at will.

First, we define the sine and cosine functions by their Taylor series. We will then derive their other properties from this definition.

Definition 5.2.2. For $x \in \mathbb{R}$,

$$\begin{aligned} \cos(x) &= \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} \\ \sin(x) &= \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}. \end{aligned}$$

By the ratio test, the radius of convergence for these series is ∞ , and therefore we can differentiate them term by term on all of \mathbb{R} :

Lemma 5.2.3. $\frac{d}{dx} \cos x = -\sin x$ and $\frac{d}{dx} \sin x = \cos x$.

We now make the connection to the complex exponential. Complex numbers have the form $z = x + iy$ with $x, y \in \mathbb{R}$ and may be identified with the points (x, y) of the plane. There is an important relationship between complex numbers and polar coordinates we will describe below. Write \mathbb{C} for the complex numbers. We define functions $\operatorname{Re} : \mathbb{C} \rightarrow \mathbb{R}$ and $\operatorname{Im} : \mathbb{C} \rightarrow \mathbb{R}$ by $\operatorname{Re}(x + iy) = x$ and $\operatorname{Im}(x + iy) = y$ for $x, y \in \mathbb{R}$. Thus, Re and Im correspond to the coordinate projections of \mathbb{R}^2 onto \mathbb{R} . We define addition

and multiplication in the complex numbers as follows. For $z = a + bi$ and $w = c + di$ with $a, b, c, d \in \mathbb{R}$, we set

$$(5.2.3) \quad z + w = (a + c) + (b + d)i$$

$$(5.2.4) \quad zw = (ac - bd) + (ad + bc)i.$$

We identify \mathbb{R} with the subring $\{a + 0i : a \in \mathbb{R}\}$. Note that addition in \mathbb{C} corresponds to vector addition in \mathbb{R}^2 and that multiplication of a complex number by a real number corresponds to scalar multiplication in \mathbb{R}^2 . The multiplication rule (5.2.4) is then forced by distributivity and the property that $i^2 = -1$, where $i = 0 + 1i$.

The following is straightforward.

Proposition 5.2.4. \mathbb{C} is a commutative ring:

- (1) Addition is commutative and associative with identity element 0. Every element has an additive inverse.
- (2) Multiplication is commutative and associative with identity element 1.
- (3) The distributive law holds: $z(w_1 + w_2) = zw_1 + zw_2$ for all $z, w_1, w_2 \in \mathbb{C}$.

In fact, \mathbb{C} is a field, meaning that in addition to being a commutative ring, every nonzero element has a multiplicative inverse. To show this, define the complex conjugate \bar{z} of $z = a + bi$, $a, b \in \mathbb{R}$, to be $\bar{z} = a - bi$. The following is easily verified.

Lemma 5.2.5. For $z, w \in \mathbb{C}$,

$$(5.2.5) \quad \overline{z + w} = \bar{z} + \bar{w}$$

$$(5.2.6) \quad \overline{zw} = \bar{z}\bar{w}.$$

For $z = a + bi$, $a, b \in \mathbb{R}$,

$$(5.2.7) \quad z\bar{z} = a^2 + b^2,$$

so identifying z with the vector $(a, b) \in \mathbb{R}^2$, $\|z\| = \sqrt{z\bar{z}}$. We write $|z| = \sqrt{z\bar{z}}$ and call it the modulus of z . Since both complex conjugation and the square root are product-preserving, $|zw| = |z||w|$. Finally,

$$\mathbb{R} = \{z \in \mathbb{C} : z = \bar{z}\}.$$

Corollary 5.2.6. \mathbb{C} is a field. For $0 \neq z \in \mathbb{C}$, $z^{-1} = \frac{\bar{z}}{z\bar{z}}$.

Proof. The key point is that $\frac{\bar{z}}{z\bar{z}}$ makes sense: $z\bar{z} = a^2 + b^2$ is a positive real number for $z \neq 0$, so it has a multiplicative inverse $\frac{1}{z\bar{z}}$ in \mathbb{R} , and hence also in \mathbb{C} . So

$$z \cdot \left(\bar{z} \cdot \frac{1}{z\bar{z}} \right) = 1. \quad \square$$

Power series in \mathbb{C} follow the same rules as power series in \mathbb{R} . They have a radius of convergence, calculated in terms of the distance in \mathbb{R}^2 , that can be found by the ratio test.

Definition 5.2.7. The complex exponential function is given by

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}$$

for $z \in \mathbb{C}$.

By the ratio test, the radius of convergence for e^z is ∞ so e^z is defined everywhere. The trig functions are obtained by restricting the complex exponential to the pure imaginary axis.

Theorem 5.2.8. Let $\theta \in \mathbb{R}$. Then

$$(5.2.8) \quad e^{i\theta} = \cos \theta + i \sin \theta.$$

Proof.

$$e^{i\theta} = \sum_{n=0}^{\infty} i^n \frac{\theta^n}{n!}.$$

We need to keep track of the powers of i . $i^2 = -1$, $i^3 = -i$, $i^4 = 1$, $i^5 = i$, and the pattern repeats: for each $k \geq 0$, $i^{4k} = 1$, $i^{4k+1} = i$, $i^{4k+2} = -1$, $i^{4k+3} = -i$. The even powers are ± 1 and the odd powers are $\pm i$. Specifically, $i^{2\ell} = (i^2)^\ell = (-1)^\ell$ and therefore $i^{2\ell+1} = (-1)^\ell i$.

Collecting terms, we have

$$\begin{aligned} e^{i\theta} &= \sum_{\ell=0}^{\infty} i^{2\ell} \frac{\theta^{2\ell}}{(2\ell)!} + \sum_{\ell=0}^{\infty} i^{2\ell+1} \frac{\theta^{2\ell+1}}{(2\ell+1)!} \\ &= \sum_{\ell=0}^{\infty} (-1)^\ell \frac{\theta^{2\ell}}{(2\ell)!} + i \sum_{\ell=0}^{\infty} (-1)^\ell \frac{\theta^{2\ell+1}}{(2\ell+1)!} \\ &= \cos \theta + i \sin \theta. \end{aligned} \quad \square$$

We shall make use of the following basic result from algebra. See [5] or [17] for a proof.

Proposition 5.2.9 (Binomial theorem). Let $z, w \in \mathbb{C}$ and $n \geq 1$. Then

$$(5.2.9) \quad (z + w)^n = \sum_{k=0}^n \binom{n}{k} z^{n-k} w^k,$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is an integer for $0 \leq k \leq n$.

In fact, the theorem holds for z, w in any commutative ring. We deduce a key property of the complex exponential.

Theorem 5.2.10. For $z, w \in \mathbb{C}$,

$$e^z e^w = e^{z+w}.$$

Proof.

$$\begin{aligned}
 e^z e^w &= \sum_{n=0}^{\infty} \frac{z^n}{n!} \sum_{n=0}^{\infty} \frac{w^n}{n!} \\
 &= \sum_{n=0}^{\infty} \sum_{j=0}^n \frac{z^j}{j!} \frac{w^{n-j}}{(n-j)!} \\
 &= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{j=0}^n \frac{n!}{j!(n-j)!} z^j w^{n-j} \\
 &= \sum_{n=0}^{\infty} \frac{(z+w)^n}{n!} \\
 &= e^{z+w}
 \end{aligned}$$

□

We obtain Theorem 5.2.1.

Proof of Theorem 5.2.1. By Theorem 5.2.10,

$$e^{i(\theta+\phi)} = e^{i\theta} e^{i\phi},$$

so

$$\begin{aligned}
 \cos(\theta + \phi) + i \sin(\theta + \phi) &= (\cos \theta + i \sin \theta)(\cos \phi + i \sin \phi) \\
 &= (\cos \theta \cos \phi - \sin \theta \sin \phi) + i(\cos \theta \sin \phi + \sin \theta \cos \phi).
 \end{aligned}$$

Equating the real parts of both sides gives (5.2.1), and equating the pure imaginary parts of both sides gives (5.2.2). □

We now wish to show that $\cos^2 \theta + \sin^2 \theta = 1$. This is equivalent to saying the vector $(\cos \theta, \sin \theta)$ has norm 1, or that the complex number $e^{i\theta}$ has modulus 1.

The complex numbers of modulus 1 comprise the unit circle:

$$\mathbb{S}^1 = \{z \in \mathbb{C} : |z| = 1\} = \{z \in \mathbb{C} : z\bar{z} = 1\},$$

and for our trig identity it suffices to show $e^{i\theta} \overline{e^{i\theta}} = 1$. By (5.2.5) and (5.2.6),

$$\overline{e^z} = \sum_{n=0}^{\infty} \frac{\overline{z^n}}{n!} = \sum_{n=0}^{\infty} \frac{\bar{z}^n}{n!} = e^{\bar{z}},$$

so

$$e^{i\theta} \overline{e^{i\theta}} = e^{i\theta} e^{-i\theta} = e^{i\theta-i\theta} = e^0 = 1.$$

We obtain:

Proposition 5.2.11. For $\theta \in \mathbb{R}$, $\cos^2 \theta + \sin^2 \theta = 1$, i.e., $e^{i\theta} \in \mathbb{S}^1$.

The remaining properties of the trig functions may be best seen using a combination of group theory and calculus. Note that the nonzero elements of \mathbb{C} form a group under multiplication, as each one has a multiplicative inverse. We denote this group

$$\mathbb{C}^\times = \{z \in \mathbb{C} : z \neq 0\}.$$

Since the modulus function is multiplicative, $\mathbb{S}^1 \subset \mathbb{C}^\times$ is closed under multiplication. Moreover, the inverse of $z \in \mathbb{S}^1$, $\frac{\bar{z}}{z\bar{z}}$, is just \bar{z} as $z\bar{z} = 1$. But $|\bar{z}| = |z|$, so \mathbb{S}^1 is also closed under inverses, and hence is a subgroup of \mathbb{C}^\times .

By Proposition 5.2.11, we may define a function $\exp : \mathbb{R} \rightarrow \mathbb{S}^1$ by

$$\exp(\theta) = e^{i\theta}.$$

Theorem 5.2.10 then gives:

Proposition 5.2.12. $\exp : \mathbb{R} \rightarrow \mathbb{S}^1$ is a group homomorphism. Here \mathbb{R} is the group of real numbers under addition.

In particular, we can use calculus to study its kernel. Identifying \mathbb{C} with \mathbb{R}^2 we see that $\frac{d}{dx} \exp(x)$ corresponds to the ordered pair $(-\sin x, \cos x) \in \mathbb{S}^1$. In particular, if $x \in \ker \exp$, then $\cos x = 1$ and $\sin x = 0$. So the second coordinate of $\frac{d}{dx} \exp(x)$ is nonzero. So $\text{Im} \circ \exp$ is one-to-one in a neighborhood of x by the inverse function theorem. In particular, this holds for $x = 0$ so \exp is one-to-one on $(-\epsilon, \epsilon)$ for some $\epsilon > 0$, hence $(-\epsilon, \epsilon) \cap \ker \exp = \{0\}$. By Lemma 3.4.10, any two distinct elements of $\ker \exp$ must be at least ϵ apart. Thus, if $\{x_n\}$ is a sequence of elements of $\ker \exp$ with $\lim_{n \rightarrow \infty} x_n = x$, then there exists N such that $x = x_n$ for all $n \geq N$. Thus, the greatest lower bound of the set of positive elements of $\ker \exp$ must lie in $\ker \exp$: there is a smallest positive element in $\ker \exp$.

Definition 5.2.13. Define 2π to be the smallest positive element of $\ker \exp$.

Proposition 5.2.14. $\ker \exp = \langle 2\pi \rangle = \{2\pi k : k \in \mathbb{Z}\}$, the subgroup of \mathbb{R} generated by 2π .

Proof. Let $x \in \ker \exp$. Then there is an integer k such that x lies in the half-open interval $[2\pi k, 2\pi(k+1))$. But then $x - 2\pi k$ is an element of $\ker \exp$ lying in $[0, 2\pi)$, so $x - 2\pi k = 0$. \square

The following is now immediate from Lemma 3.4.10.

Corollary 5.2.15. $\exp x = \exp y$ if and only if $y = x + 2k\pi$ for some $k \in \mathbb{Z}$. In particular, \exp is one-to-one on $[\theta, \theta + 2\pi)$ for all $\theta \in \mathbb{R}$.

The most important remaining verification is that $\exp : \mathbb{R} \rightarrow \mathbb{S}^1$ is onto. That and the remaining properties of the trig functions can be obtained from calculus and the theorems above.

A priori, there is no connection between the geometric intuition we've built about trig functions and the analytic definitions given here. We must remedy this.

First, note that if $\sin \theta = 0$, then $\cos \theta = \pm 1$ since $\cos^2 \theta + \sin^2 \theta = 1$. So $e^{i\theta} = \pm 1$. In particular, consider $\theta = \pi$. We have $(e^{i\pi})^2 = e^{2\pi i} = 1$, so $e^{i\pi} = \pm 1$. But π is not a multiple of 2π , so $e^{i\pi} \neq 1$. We obtain de Moivre's theorem:

$$(5.2.10) \quad e^{i\pi} = -1.$$

By Corollary 5.2.15, we obtain:

Lemma 5.2.16. $\sin \theta = 0$ if and only if $\theta = n\pi$ for some $n \in \mathbb{Z}$.

Now $\frac{d}{d\theta} \sin \theta = \cos \theta$ is continuous and is positive at $\theta = 0$, so the sine function is strictly increasing on an interval $(-\epsilon, \epsilon)$ for some $\epsilon > 0$. By Lemma 5.2.16 and the continuity of the sine we obtain:

Corollary 5.2.17. $\sin \theta$ is positive for $\theta \in (0, \pi)$ and is negative for $\theta \in (\pi, 2\pi)$. Thus, the cosine is strictly decreasing on $[0, \pi]$ and is strictly increasing on $[\pi, 2\pi]$.

We now consider the zeros of the cosine. If $\cos \theta = 0$, then $\sin \theta = \pm 1$, so $e^{i\theta} = \pm i$. We generalize the preceding argument. The key property about $\pm i$ is that they are the two square roots of -1 . We know there are only two square roots because \mathbb{C} is a field:

Lemma 5.2.18. Let $0 \neq z \in \mathbb{C}$ have a square root. Say $z = w^2$. Then z has exactly two square roots in \mathbb{C} : $\pm w$.

Proof. Any square root of z is a root of the polynomial $x^2 - z$. A polynomial f of degree n with coefficients in a field \mathbb{F} has at most n roots. Since $x^2 - z$ has the two distinct roots $\pm w$, there are no others. \square

But \exp can produce two square roots of -1 : $(e^{i\frac{\pi}{2}})^2 = e^{2i\frac{\pi}{2}} = e^{i\pi} = -1$, and $(e^{i\frac{3\pi}{2}})^2 = e^{i3\pi} = -1$. By the preceding lemma, $\{e^{i\frac{\pi}{2}}, e^{i\frac{3\pi}{2}}\} = \{\pm i\}$. By Corollary 5.2.17, we obtain the following.

Corollary 5.2.19. $e^{i\frac{\pi}{2}} = i$ and $e^{i\frac{3\pi}{2}} = -i$. In particular, $\frac{\pi}{2}$ and $\frac{3\pi}{2}$ are the only values of $\theta \in [0, 2\pi)$ for which $\cos \theta = 0$. Since \cos is continuous, $\cos 0 = 1$ and $\cos \pi = -1$, the cosine is positive on $(-\frac{\pi}{2}, \frac{\pi}{2})$ and is negative on $(\frac{\pi}{2}, \frac{3\pi}{2})$. Thus, the sine is strictly increasing on $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and is strictly decreasing on $[\frac{\pi}{2}, \frac{3\pi}{2}]$.

In particular, the sine restricts to a strictly increasing function

$$f = \sin |_{[-\frac{\pi}{2}, \frac{\pi}{2}]} : \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \rightarrow [-1, 1],$$

as $\sin(-\frac{\pi}{2}) = -1$ and $\sin(\frac{\pi}{2}) = 1$. As f is increasing, it is one-to-one. By the intermediate value theorem, f is onto. Thus, there is a one-to-one and onto inverse function

$$\arcsin = f^{-1} : [-1, 1] \rightarrow \left[-\frac{\pi}{2}, \frac{\pi}{2}\right].$$

Now let $z = x + yi \in \mathbb{S}^1$ with $x \geq 0$. Since $x^2 + y^2 = 1$, $x = \sqrt{1 - y^2}$. Let $\theta = \arcsin y$. So $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, hence $\cos \theta > 0$, giving $\cos \theta = \sqrt{1 - \sin^2 \theta} = x$. Thus $z = e^{i\theta}$, and every point in \mathbb{S}^1 with nonnegative x -coordinate is in the image of \exp .

If $z = x + iy$ with $x < 0$, then $-z$ has positive real part, so $-z = e^{-\theta}$ for $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ by the argument just given, but then $e^{i(\theta+\pi)} = e^{i\theta}e^{i\pi} = -e^{i\theta} = z$. Thus:

Corollary 5.2.20. $\exp : \mathbb{R} \rightarrow \mathbb{S}^1$ is onto. By restriction, $\exp : [\theta, \theta + 2\pi) \rightarrow \mathbb{S}^1$ is bijective for all $\theta \in \mathbb{R}$.

Let $0 \neq z \in \mathbb{C}$. Then $\frac{z}{|z|}$ has modulus 1, i.e., $\frac{z}{|z|} \in \mathbb{S}^1$. By Corollary 5.2.20 there is a unique $\theta \in [0, 2\pi)$, called the argument, $\arg z$, of z with $e^{i\theta} = \frac{z}{|z|}$. We obtain the complex version of polar coordinates.

Corollary 5.2.21. Let $0 \neq z \in \mathbb{C}$. Then there are unique real numbers r, θ with $r > 0$ and $\theta \in [0, 2\pi)$ such that $z = re^{i\theta}$. $r = |z|$, the modulus of z , and θ is called the argument of z .

Of course $re^{i\theta}se^{i\phi} = rse^{i(\theta+\phi)}$. In particular:

Corollary 5.2.22. Every complex number has a square root. If $z = re^{i\theta}$ then the square roots of z are $\pm\sqrt{r}e^{i\frac{\theta}{2}}$.

Corollary 5.2.21 translates directly into the usual form of polar coordinates.

Corollary 5.2.23. Let $0 \neq v = (x, y) \in \mathbb{R}^2$. Then there are unique real numbers r, θ with $r > 0$ and $\theta \in [0, 2\pi)$ with $v = (r \cos \theta, r \sin \theta)$. $r = \|v\|$, the norm of v .

Remark 5.2.24. Let $0 < \theta < \frac{\pi}{2}$. Then the usual derivation of $\sin \theta$ as the opposite over the hypotenuse for a right triangle based at the origin that makes the angle θ with respect to the positive x -axis follows from this, as the quotient of opposite over hypotenuse is just $\frac{r \sin \theta}{r}$. Similarly for the cosine.

Corollary 5.2.22 is a special case of a much deeper theorem called the Fundamental theorem of algebra. Its proof is beyond the scope of this book. See [17] for a proof using algebraic topology or [5] for a purely algebraic proof. The statement is as follows:

Theorem 5.2.25 (Fundamental theorem of algebra). *Every polynomial of positive degree with complex coefficients has a root in \mathbb{C} .*

This applies to Corollary 5.2.22 as any square root of the complex number z is a root of the polynomial $x^2 - z$. However, the proofs of the Fundamental theorem of algebra are not constructive, while the proof Corollary 5.2.22 is constructive. Indeed, it shows that the square roots of $z \in \mathbb{S}^1$ also lie in \mathbb{S}^1 , with nicely specified angles.

Corollary 5.2.26. $e^{i\frac{\pi}{4}} = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}i$, i.e., $\cos \frac{\pi}{4} = \frac{1}{\sqrt{2}}$ and $\sin \frac{\pi}{4} = \frac{1}{\sqrt{2}}$.

Proof. Each side of the equality coincides with the unique square root of i lying in the first quadrant. \square

We can get similar benefit out of complex cube roots. Note that $e^{i\frac{2\pi}{3}}$ and $e^{i\frac{4\pi}{3}}$ are both cube roots of 1, one lying in the second quadrant and one in the third. If $z = re^{i\theta}$ with $r > 0$, then $w = \sqrt[3]{r}e^{i\frac{\theta}{3}}$ is a cube root of z , as are $we^{i\frac{2\pi}{3}} = \sqrt[3]{r}e^{i(\frac{\theta}{3} + \frac{2\pi}{3})}$ and $we^{i\frac{4\pi}{3}} = \sqrt[3]{r}e^{i(\frac{\theta}{3} + \frac{4\pi}{3})}$. In particular, these give three distinct roots of $x^3 - z$, and hence they are the only roots of $x^3 - z$. We have:

Lemma 5.2.27. $z = re^{i\theta}$ with $r > 0$. Then there are exactly three cube roots of z in \mathbb{C} : $\sqrt[3]{r}e^{i\frac{\theta}{3}}$, $\sqrt[3]{r}e^{i(\frac{\theta}{3} + \frac{2\pi}{3})}$ and $\sqrt[3]{r}e^{i(\frac{\theta}{3} + \frac{4\pi}{3})}$. At most one of them is in the first quadrant.

We can use this to recover the trigonometric functions of familiar angles:

Corollary 5.2.28. $e^{i\frac{\pi}{3}} = \frac{1}{2} + \frac{\sqrt{3}}{2}i$, i.e., $\cos \frac{\pi}{3} = \frac{1}{2}$ and $\sin \frac{\pi}{3} = \frac{\sqrt{3}}{2}$. $e^{i\frac{\pi}{6}} = \frac{\sqrt{3}}{2} + \frac{1}{2}i$, i.e., $\cos \frac{\pi}{6} = \frac{\sqrt{3}}{2}$ and $\sin \frac{\pi}{6} = \frac{1}{2}$.

Proof. The binomial expansion for $(\frac{1}{2} + \frac{\sqrt{3}}{2}i)^3$ simplifies to -1 , so $\frac{1}{2} + \frac{\sqrt{3}}{2}i$ is the unique cube root of -1 lying in the first quadrant. But so is $e^{i\frac{\pi}{3}}$, so they must be equal. Similarly, both $e^{i\frac{\pi}{6}}$ and $\frac{\sqrt{3}}{2} + \frac{1}{2}i$ are first quadrant cube roots of i . \square

Note we did not need to use the fact that the sum of the angles in a Euclidean triangle is π , which would be used in the most familiar proofs of these calculations.

Similar results to Lemma 5.2.27 are available for n th roots.

The only remaining issue for the trig functions is to approximate the value of π . But standard calculus techniques deduce from only the results here that the area of the unit disk

$$\mathbb{D}^2 = \{v \in \mathbb{R}^2 : \|v\| \leq 1\}$$

is π . This area can now be approximated by the areas of inscribed polygons.

5.3. Linear isometries of \mathbb{R}^2 : calculation of $\mathbf{O}(2)$. In this section, as we use matrices, we will write vectors as column matrices.

By Theorem 4.1.12, the linear isometries of \mathbb{R}^n are the linear mappings induced by the matrices whose columns form an orthonormal basis of \mathbb{R}^n . For $n = 2$, Corollary 5.1.4 shows us that every orthonormal basis either has the form v, v^\perp or $v, -v^\perp$ for some unit vector v . But the unit vectors in \mathbb{R}^2 are precisely $\left\{ \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} : \theta \in [0, 2\pi) \right\}$. Recalling that $\begin{bmatrix} a \\ b \end{bmatrix}^\perp = \begin{bmatrix} -b \\ a \end{bmatrix}$ we obtain:

Corollary 5.3.1. *The matrices in $O(2)$ are precisely*

$$\left\{ R_\theta = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \middle| \theta \in [0, 2\pi) \right\} \\ \cup \left\{ S_\theta = \begin{bmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{bmatrix} \middle| \theta \in [0, 2\pi) \right\}.$$

The determinant of R_θ is 1 and the determinant of S_θ is -1 . Thus:

Corollary 5.3.2.

$$SO(2) = \left\{ R_\theta = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \middle| \theta \in [0, 2\pi) \right\}$$

We shall see that the matrices R_θ induce rotations and the matrices S_θ induce reflections.

Proposition 5.3.3. *The matrix R_θ rotates the plane about the origin by the angle θ in the counterclockwise direction. We write $\rho_{(0,\theta)} = T_{R_\theta}$.*

Proof. We use polar coordinates.

$$\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} r \cos \phi \\ r \sin \phi \end{bmatrix} = \begin{bmatrix} r(\cos \theta \cos \phi - \sin \theta \sin \phi) \\ r(\sin \theta \cos \phi + \cos \theta \sin \phi) \end{bmatrix} \\ = \begin{bmatrix} r \cos(\theta + \phi) \\ r \sin(\theta + \phi) \end{bmatrix}$$

by Theorem 5.2.1. So the plane is indeed being rotated by θ about 0. \square

So the elements of $SO(2)$ are rotations of the plane. Their composition law is as follows:

Proposition 5.3.4. *The matrix product $R_\theta R_\phi = R_{\theta+\phi}$. Thus, the matrices in $SO(2)$ commute with each other. The induced linear functions satisfy $\rho_{(0,\theta)} \circ \rho_{(0,\phi)} = \rho_{(0,\theta+\phi)}$.*

Proof.

$$\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{bmatrix} \\ = \begin{bmatrix} \cos \theta \cos \phi - \sin \theta \sin \phi & -\cos \theta \sin \phi - \sin \theta \cos \phi \\ \sin \theta \cos \phi + \cos \theta \sin \phi & -\sin \theta \sin \phi + \cos \theta \cos \phi \end{bmatrix} \\ = \begin{bmatrix} \cos(\theta + \phi) & -\sin(\theta + \phi) \\ \sin(\theta + \phi) & \cos(\theta + \phi) \end{bmatrix} \quad \square$$

Rotations help explain the perp operation for vectors:

Lemma 5.3.5. *Let $v \in \mathbb{R}^2$. Then $v^\perp = \rho_{(0,\frac{\pi}{2})}(v)$.*

Proof. $R_{\frac{\pi}{2}} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. Multiplication by this matrix has the desired effect on coordinates. \square

There are a number of approaches available now for studying the linear reflections. We shall use that fact that every isometry preserving the origin is linear. Let

$$\ell_\theta = \text{span} \left(\begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} \right),$$

the line through the origin meeting \mathbb{S}^1 in $\pm \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}$. (Of course, $\ell_\theta = \ell_{\theta+\pi}$.)

The reflection σ_{ℓ_θ} fixes the origin (Lemma 5.1.12), and hence is linear. So σ_{ℓ_θ} is the linear isometry represented by the matrix $[\sigma_{\ell_\theta}(e_1) | \sigma_{\ell_\theta}(e_2)]$, the matrix whose columns are the images under σ_{ℓ_θ} of the canonical basis vectors of \mathbb{R}^2 . We have

$$\begin{aligned} \sigma_{\ell_\theta}(e_1) &= e_1 - 2 \left\langle e_1, \begin{bmatrix} -\sin \theta \\ \cos \theta \end{bmatrix} \right\rangle \begin{bmatrix} -\sin \theta \\ \cos \theta \end{bmatrix} = \begin{bmatrix} 1 - 2 \sin^2 \theta \\ 2 \sin \theta \cos \theta \end{bmatrix} = \begin{bmatrix} \cos(2\theta) \\ \sin(2\theta) \end{bmatrix}, \\ \sigma_{\ell_\theta}(e_2) &= e_2 - 2 \left\langle e_2, \begin{bmatrix} -\sin \theta \\ \cos \theta \end{bmatrix} \right\rangle \begin{bmatrix} -\sin \theta \\ \cos \theta \end{bmatrix} = \begin{bmatrix} 2 \cos \theta \sin \theta \\ 1 - 2 \cos^2 \theta \end{bmatrix} = \begin{bmatrix} \sin(2\theta) \\ -\cos(2\theta) \end{bmatrix}. \end{aligned}$$

Thus, we have proven:

Proposition 5.3.6. σ_{ℓ_θ} is the linear transformation induced by the matrix $S_{2\theta}$.

Reversing it, we see that $T_{S_\theta} = \sigma_{\ell_{\frac{\theta}{2}}}$. So the matrices S_θ all represent reflections.

We now show how to express the effect of a linear reflection on a vector in polar coordinates.

Lemma 5.3.7. $\sigma_{\ell_\theta} \left(\begin{bmatrix} r \cos \phi \\ r \sin \phi \end{bmatrix} \right) = \begin{bmatrix} r \cos(2\theta - \phi) \\ r \sin(2\theta - \phi) \end{bmatrix}$.

Proof. We multiply matrices:

$$\begin{aligned} S_{2\theta} \begin{bmatrix} r \cos \phi \\ r \sin \phi \end{bmatrix} &= \begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix} \begin{bmatrix} r \cos \phi \\ r \sin \phi \end{bmatrix} \\ &= \begin{bmatrix} r(\cos 2\theta \cos \phi + \sin 2\theta \sin \phi) \\ r(\sin 2\theta \cos \phi - \cos 2\theta \sin \phi) \end{bmatrix} = \begin{bmatrix} r \cos(2\theta - \phi) \\ r \sin(2\theta - \phi) \end{bmatrix}. \quad \square \end{aligned}$$

We now show how to compose linear reflections. Since reflection matrices have determinant -1 , the product of two reflection matrices will have determinant 1, and hence will be a rotation matrix by our analysis above.

Proposition 5.3.8. The composite of two linear reflections is a linear rotation. Specifically,

$$\sigma_{\ell_\theta} \sigma_{\ell_\phi} = \rho_{(0, 2(\theta - \phi))},$$

the rotation about $0 = \ell_\theta \cap \ell_\phi$ by twice the directed angle from ℓ_ϕ to ℓ_θ .

Proof. The argument is very similar to that of Lemma 5.3.7.

$$S_{2\theta} S_{2\phi} = \begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix} \begin{bmatrix} \cos 2\phi & \sin 2\phi \\ \sin 2\phi & -\cos 2\phi \end{bmatrix}$$

$$\begin{aligned}
&= \begin{bmatrix} \cos 2\theta \cos 2\phi + \sin 2\theta \sin 2\phi & \cos 2\theta \sin 2\phi - \sin 2\theta \cos 2\phi \\ \sin 2\theta \cos 2\phi - \cos 2\theta \sin 2\phi & \sin 2\theta \sin 2\phi + \cos 2\theta \cos 2\phi \end{bmatrix} \\
&= \begin{bmatrix} \cos 2(\theta - \phi) & -\sin 2(\theta - \phi) \\ \sin 2(\theta - \phi) & \cos 2(\theta - \phi) \end{bmatrix}. \quad \square
\end{aligned}$$

Note that the directed angle from ℓ_ϕ to ℓ_θ does not make sense, as $\ell_\phi = \ell_{\phi+\pi}$ and similarly for ℓ_θ . But when you double the difference, the added π ceases to matter.

We now show how to compose rotations with reflections. We could do this by simply multiplying matrices again, but there is a more conceptual way to do both this and the calculation above. Note that ℓ_0 is the x -axis and that σ_{ℓ_0} is induced by the matrix

$$(5.3.1) \quad S_0 = \begin{bmatrix} \cos 0 & \sin 0 \\ \sin 0 & -\cos 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Thus, the effect of σ_{ℓ_0} on the plane is the same as that of complex conjugation if we identify the plane with \mathbb{C} in the usual way.

Lemma 5.3.9. $\sigma_{\ell_\theta} = \rho_{(0,2\theta)}\sigma_{\ell_0}$ for all $\theta \in \mathbb{R}$.

Proof. This follows from a special case of Proposition 5.3.8: $\sigma_{\ell_\theta}\sigma_{\ell_0} = \rho_{(0,2\theta)}$. Just multiply both sides on the right by σ_{ℓ_0} and use the fact that σ_m is an involution (i.e., $\sigma_m^2 = \text{id}$) for every line m .

However, we can prove it much more simply by direct calculation of matrix products:

$$R_{2\theta}S_0 = \begin{bmatrix} \cos 2\theta & -\sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = S_{2\theta}. \quad \square$$

Now we can show what happens when we conjugate a linear rotation by a linear reflection.

Proposition 5.3.10. Let ℓ be a line through the origin and $\theta \in \mathbb{R}$. Then

$$\sigma_\ell \rho_{(0,\theta)} \sigma_\ell^{-1} = \rho_{(0,-\theta)}.$$

Proof. When $\ell = \ell_0$ it is immediate from the matrix multiplication

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

For $\ell = \ell_\phi$, we can deduce it from this case and Lemma 5.3.9:

$$\begin{aligned}
\sigma_{\ell_\phi} \rho_{(0,\theta)} \sigma_{\ell_\phi}^{-1} &= (\rho_{(0,2\phi)} \sigma_{\ell_0}) \rho_{(0,\theta)} (\rho_{(0,2\phi)} \sigma_{\ell_0})^{-1} \\
&= \rho_{(0,2\phi)} (\sigma_{\ell_0} \rho_{(0,\theta)} \sigma_{\ell_0}^{-1}) \rho_{(0,2\phi)}^{-1} \\
&= \rho_{(0,2\phi)} \rho_{(0,-\theta)} \rho_{(0,2\phi)}^{-1} = \rho_{(0,-\theta)},
\end{aligned}$$

as any two rotations about 0 commute. \square

As usual, we obtain the following.

Corollary 5.3.11. *If $0 \in \ell$, then*

$$\sigma_\ell \rho_{(0,\theta)} = \rho_{(0,-\theta)} \sigma_\ell.$$

We obtain a simpler proof of Proposition 5.3.8.

Alternative proof of Proposition 5.3.8.

$$\begin{aligned} \sigma_{\ell_\theta} \sigma_{\ell_\phi} &= \rho_{(0,2\theta)} \sigma_{\ell_0} \rho_{(0,2\phi)} \sigma_{\ell_0} \\ &= \rho_{(0,2\theta)} \rho_{(0,-2\phi)} \sigma_{\ell_0} \sigma_{\ell_0} \\ &= \rho_{(0,2(\theta-\phi))}. \end{aligned} \quad \square$$

Similarly, we can use Lemma 5.3.9, Corollary 5.3.11 and the composition rule for rotations about 0 to compute an arbitrary composition of rotations about 0 and reflections in lines through 0. Group theoretically, we have the following.

Corollary 5.3.12. *O(2) may be decomposed as*

$$O(2) = \{R_\theta, R_\theta S_0 : \theta \in \mathbb{R}\}$$

with the multiplication law given by $S_0^2 = \text{id}$, $R_\theta R_\phi = R_{\theta+\phi}$ and $S_0 R_\theta = R_{-\theta} S_0$. Here, of course $R_\theta = R_\phi$ if and only if $\theta - \phi$ is a multiple of 2π .

5.4. Angles in \mathbb{R}^2 and \mathbb{R}^n ; the cosine law; orientation in \mathbb{R}^2 .

5.4.1. Angles in \mathbb{R}^2 . The angle between a pair of lines doesn't make complete sense as, for instance, $\text{span}(\begin{bmatrix} 1 \\ 0 \end{bmatrix})$ and $\text{span}(\begin{bmatrix} 1 \\ 1 \end{bmatrix})$ have angles of both $\frac{\pi}{4}$ and $\frac{3\pi}{4}$ between them in the counterclockwise direction. To be more specific, we should work with the angle between two rays. Recall that for $x \neq y \in \mathbb{R}^n$, The ray from x through y is

$$\overrightarrow{xy} = \{(1-t)y + ty : t \geq 0\}.$$

The following is immediate:

Lemma 5.4.1.

$$\begin{aligned} \overrightarrow{xy} &= \{x + t(y-x) : t \geq 0\} \\ &= \tau_x(\overrightarrow{0v}), \end{aligned}$$

for $v = \frac{y-x}{\|y-x\|}$. We say \overrightarrow{xy} has initial point x or that it emanates from x .

The angle between two rays in \mathbb{R}^2 with the same initial point is now easily defined.

Definition 5.4.2. Let \overrightarrow{xy} and \overrightarrow{xz} be rays in \mathbb{R}^2 with initial point x . Let $y-x = \begin{bmatrix} r \cos \theta \\ r \sin \theta \end{bmatrix}$ and $z-x = \begin{bmatrix} s \cos \phi \\ s \sin \phi \end{bmatrix}$ with $r, s > 0$. Then the (directed) angle from \overrightarrow{xz} to \overrightarrow{xy} is $\theta - \phi$.

Note that the angle is defined by first translating the rays to emanate from the origin and then taking the angle of the translated rays. The following is the main ingredient in the proof of the cosine law.

Lemma 5.4.3. *Let $0 \neq x, y \in \mathbb{R}^2$. Then*

$$(5.4.1) \quad \langle x, y \rangle = \|x\| \|y\| \cos \theta,$$

where θ is the angle from $\vec{0x}$ to $\vec{0y}$.

Proof. Let $x = \begin{bmatrix} r \cos \phi \\ r \sin \phi \end{bmatrix}$ and $y = \begin{bmatrix} s \cos \psi \\ s \sin \psi \end{bmatrix}$ with $r, s > 0$. Then, $r = \|x\|$, $s = \|y\|$ and $\theta = \psi - \phi$.

$$\begin{aligned} \langle x, y \rangle &= \left\langle \begin{bmatrix} r \cos \phi \\ r \sin \phi \end{bmatrix}, \begin{bmatrix} s \cos \psi \\ s \sin \psi \end{bmatrix} \right\rangle \\ &= rs(\cos \phi \cos \psi + \sin \phi \sin \psi) \\ &= rs \cos(\psi - \phi). \end{aligned} \quad \square$$

5.4.2. Angles in \mathbb{R}^n . Note that the definition of directed angle in \mathbb{R}^2 depends on the parametrization of the unit circle by the complex exponential function. We don't have this tool in higher dimensions. There, it's easiest to define the unsigned angle between two rays. To do better, we will need to discuss oriented planes in \mathbb{R}^n . In the meantime we have the following.

Definition 5.4.4. The unsigned angle, θ , between two rays \vec{xy} and \vec{xz} in \mathbb{R}^n is given by

$$(5.4.2) \quad \theta = \cos^{-1} \left(\frac{\langle y-x, z-x \rangle}{\|y-x\| \|z-x\|} \right) = \cos^{-1} \left(\frac{\langle \tau_{-x}(y), \tau_{-x}(z) \rangle}{\|\tau_{-x}(y)\| \|\tau_{-x}(z)\|} \right).$$

Here \cos^{-1} is the inverse function of $\cos : [0, \pi] \rightarrow [-1, 1]$. Note that $\frac{\langle y-x, z-x \rangle}{\|y-x\| \|z-x\|}$ lies in the domain of \cos^{-1} by the Cauchy-Schwarz inequality.

Despite the fact that τ_{-x} is an isometry, we cannot simplify the far right-hand side of (5.4.2), as τ_{-x} is not linear unless $x = 0$. We must show the following:

Lemma 5.4.5. (5.4.2) is independent of the choices of $y \in \vec{xy}$ and $z \in \vec{xz}$.

Proof. If $v = \frac{y-x}{\|y-x\|}$ and $w = \frac{z-x}{\|z-x\|}$ we may replace y by $x+tv$ and replace z by $x+sw$ for any $s, t > 0$ and have the same two rays. Since (5.4.2) incorporates the translation by $-x$ it suffices to note that

$$\frac{\langle tv, sw \rangle}{\|tv\| \|sw\|} = \frac{st \langle v, w \rangle}{|s| |t|} = \langle v, w \rangle$$

for any $s, t > 0$, as v and w are unit vectors. \square

Of course, unsigned angles are exactly what is used in Euclidean geometry and hence are appropriate in the following, which we may as well state in \mathbb{R}^n . The cosine law generalizes the Pythagorean theorem to nonright triangles.

Theorem 5.4.6 (Cosine law). *Let $\triangle ABC$ be a triangle in \mathbb{R}^n and let θ be the unsigned angle corresponding to the vertex C . Write a, b, c for the lengths of the sides opposite the vertices A, B, C , respectively. Then*

$$(5.4.3) \quad c^2 = a^2 + b^2 - 2ab \cos \theta.$$

Proof. θ is the unsigned angle between \overrightarrow{CA} and \overrightarrow{CB} , so

$$ab \cos \theta = \langle B - C, A - C \rangle$$

by (5.4.2). We have

$$\begin{aligned} c^2 &= \langle B - A, B - A \rangle \\ &= \langle (B - C) - (A - C), (B - C) - (A - C) \rangle \\ &= \langle B - C, B - C \rangle + \langle A - C, A - C \rangle - 2\langle B - C, A - C \rangle \\ &= a^2 + b^2 - 2ab \cos \theta. \end{aligned} \quad \square$$

Remark 5.4.7. Note that the rays emanating from $x \in \mathbb{R}^n$ are in one-to-one correspondence with the oriented lines containing x . If ℓ is such a line the orientation specifies a unique unit vector v such that $\ell = x + \text{span}(v)$. This corresponds to a unique ray $\tau_x(\overrightarrow{0v})$, i.e., the ray from x through $x + v$. Thus, we can define the directed angle between nonparallel oriented lines in \mathbb{R}^2 by taking x to be their point of intersection. Similarly, we may define the unsigned angle between intersecting lines in \mathbb{R}^n .

Isometries of \mathbb{R}^n are affine maps and therefore preserve rays: for $\alpha \in \mathcal{I}_n$ and $t \in \mathbb{R}$,

$$\alpha((1-t)x + ty) = (1-t)\alpha(x) + t\alpha(y).$$

Thus:

Lemma 5.4.8. For $\alpha \in \mathcal{I}_n$ and $x \neq y \in \mathbb{R}^n$, $\alpha(\overrightarrow{xy}) = \overline{\alpha(x)\alpha(y)}$.

Thus, we can ask whether isometries preserve signed and/or unsigned angles.

Proposition 5.4.9. *Isometries of \mathbb{R}^n preserve unsigned angles.*

Proof. By Theorem 2.5.3, every isometry is a composite $\tau_x\beta$ where β is a linear isometry. Thus, it suffices to assume our isometry is either a translation or is linear. Formula (5.4.2) is clearly invariant under translation, as

$$\tau_z(y) - \tau_z(x) = y - x$$

for any x, y and z , so we may assume our isometry, β , is linear. But linear isometries preserve inner product, differences and norms, so (5.4.2) is invariant under linear isometry. \square

Similarities of \mathbb{R}^n are also affine functions, so we can ask if they, also, preserve unsigned angles. Indeed, this is an important aspect of the theory of similar triangles.

Proposition 5.4.10. *Similarities of \mathbb{R}^n preserve unsigned angles.*

Proof. By Corollary 2.7.5, it suffices to show that $\mu_s : \mathbb{R}^n \rightarrow \mathbb{R}^n$ preserves unsigned angles, where $\mu_s(x) = sx$. (Here, $0 < s \in \mathbb{R}$.) But this is obvious from (5.4.2). \square

Indeed, we can prove a converse for this. This matches conventional wisdom about similar triangles.

Theorem 5.4.11. *An affine automorphism of \mathbb{R}^n is a similarity if and only if it preserves unsigned angles.*

Proof. Let f be an affine automorphism of \mathbb{R}^n that preserves unsigned angles. Write $f = \tau_x g$ with g linear. Then g preserves unsigned angles, and it suffices to show g is a linear similarity, i.e., that $g = \mu_s \beta$ for β a linear isometry of \mathbb{R}^n and $s > 0$.

Write $g = T_A$, where $A = [v_1 | \dots | v_n]$ is the $n \times n$ matrix whose columns are $v_1, \dots, v_n \in \mathbb{R}^n$. Since g is an automorphism, v_1, \dots, v_n are linearly independent.

Of course $v_i = g(e_i)$; since g preserves angles, $\langle v_i, v_j \rangle = 0$ for $i \neq j$. It suffices to show that $\|v_i\| = \|v_j\|$ for all i, j , as then if $s = \|v_i\|$, $g = \mu_s T_B$, for $B = [\frac{v_1}{s} | \dots | \frac{v_n}{s}]$. Since $\frac{v_1}{s}, \dots, \frac{v_n}{s}$ is an orthonormal basis of \mathbb{R}^n , B is an orthogonal matrix, and the result follows.

By (5.4.2),

$$(5.4.4) \quad \frac{\langle Ax, Ay \rangle}{\|Ax\| \|Ay\|} = \frac{\langle x, y \rangle}{\|x\| \|y\|}$$

for all $x, y \in \mathbb{R}^n$.

Now,

$$(5.4.5) \quad \frac{\langle e_i, e_i + e_j \rangle}{\|e_i\| \|e_i + e_j\|} = \frac{1}{\sqrt{2}} = \frac{\langle e_j, e_i + e_j \rangle}{\|e_j\| \|e_i + e_j\|}$$

for all $i \neq j$, so

$$(5.4.6) \quad \frac{\langle v_i, v_i + v_j \rangle}{\|v_i\| \|v_i + v_j\|} = \frac{\langle v_j, v_i + v_j \rangle}{\|v_j\| \|v_i + v_j\|}$$

for $i \neq j$. Since $\langle v_i, v_j \rangle = 0$, this evaluates to

$$(5.4.7) \quad \frac{\langle v_i, v_i \rangle}{\|v_i\| \|v_i + v_j\|} = \frac{\langle v_j, v_j \rangle}{\|v_j\| \|v_i + v_j\|}.$$

Multiplying through by $\|v_i + v_j\|$, this gives $\|v_i\| = \|v_j\|$. \square

5.4.3. Orientation in \mathbb{R}^2 . For isometries of the plane, we can ask if signed angles are preserved, or perhaps reversed.

Definition 5.4.12. Let α be an isometry of \mathbb{R}^2 . We say α is orientation-preserving if it preserves signed (directed) angles. We say α orientation-reversing if it reverses the sign of every directed angle.

A priori, there could be isometries of \mathbb{R}^2 that preserve the signs of some angles and reverse the signs of other angles, but that is not the case. Recall that every linear isometry of \mathbb{R}^2 is either a reflection in a line through the origin or a rotation about the origin.

Proposition 5.4.13. *Let $\alpha = \tau_x \beta$ be an isometry of \mathbb{R}^2 with β linear. Then α is orientation-preserving if β is a rotation and is orientation-reversing if β is a reflection.*

Proof. Translations obviously preserve signed angles, so it suffices to assume $\alpha = \beta$ is linear. Two rays emanating from y have the form $\tau_y(\vec{0v})$ and $\tau_y(\vec{0w})$ for unit vectors v, w . For β linear, $\beta\tau_y = \tau_{\beta(y)}\beta$. So the angle from $\beta\tau_y(\vec{0v})$ to $\beta\tau_y(\vec{0w})$ is the angle from $\tau_{\beta(y)}\beta(\vec{0v})$ to $\tau_{\beta(y)}\beta(\vec{0w})$, which in turn is the angle from $\beta(\vec{0v}) = \overline{0\beta(v)}$ to $\beta(\vec{0w}) = \overline{0\beta(w)}$.

Write $v = \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}$ and $w = \begin{bmatrix} \cos \phi \\ \sin \phi \end{bmatrix}$. Then the angle from $\tau_y(\vec{0v})$ to $\tau_y(\vec{0w})$ is $\phi - \theta$. We have two cases to consider. If β is a rotation, $\beta = \rho_{(0,\psi)}$ for some ψ , so

$$\begin{aligned}\beta(v) &= \begin{bmatrix} \cos(\theta+\psi) \\ \sin(\theta+\psi) \end{bmatrix}, \\ \beta(w) &= \begin{bmatrix} \cos(\phi+\psi) \\ \sin(\phi+\psi) \end{bmatrix},\end{aligned}$$

by Proposition 5.3.3. Clearly, the angle is preserved.

In the remaining case, $\beta = \sigma_{\ell_\psi}$ for some ψ . Here $\ell_\psi = \text{span} \left(\begin{bmatrix} \cos \psi \\ \sin \psi \end{bmatrix} \right)$, and by Lemma 5.3.7,

$$\begin{aligned}\beta(v) &= \begin{bmatrix} \cos(2\psi-\theta) \\ \sin(2\psi-\theta) \end{bmatrix}, \\ \beta(w) &= \begin{bmatrix} \cos(2\psi-\phi) \\ \sin(2\psi-\phi) \end{bmatrix}.\end{aligned}$$

This clearly reverses the sign of the angle. □

We extend this easily to similarities of \mathbb{R}^2 .

Proposition 5.4.14. *A similarity $\tau_x \mu_s \beta$ of \mathbb{R}^2 , with β a linear isometry, is orientation-preserving if β is a rotation and is orientation-reversing if β is a reflection.*

Proof. It suffices to show that μ_s is orientation-preserving. This is obvious from the proof of Proposition 5.4.13. □

By Proposition 5.4.13, an isometry of \mathbb{R}^2 either preserves all angles or reverses all angles. The following now makes sense.

Definition 5.4.15. Let $\alpha \in \mathcal{I}_2$. We define the sign, $\text{sgn } \alpha$, of α by

$$\text{sgn } \alpha = \begin{cases} 1 & \text{if } \alpha \text{ is orientation-preserving} \\ -1 & \text{if } \alpha \text{ is orientation-reversing.} \end{cases}$$

Now consider the composite $\alpha_1 \alpha_2$ of $\alpha_1, \alpha_2 \in \mathcal{I}_2$. If both α_1 and α_2 reverse all angles, then the composite preserves all angles. We may analyze the other cases similarly and obtain:

Corollary 5.4.16. *The composite $\alpha_1\alpha_2$ of $\alpha_1, \alpha_2 \in \mathcal{I}_2$ is orientation-reversing if exactly one of α_1 and α_2 orientation-reversing. In all other cases, $\alpha_1\alpha_2$ is orientation-preserving. Thus, $\text{sgn}(\alpha_1\alpha_2) = \text{sgn}(\alpha_1)\text{sgn}(\alpha_2)$, so*

$$\text{sgn} : \mathcal{I}_2 \rightarrow \{\pm 1\}$$

is a group homomorphism. Here $\{\pm 1\}$ is a group under multiplication in the standard way.

In particular, the product of two orientation-preserving isometries is orientation-preserving. But if α preserves all angles, α^{-1} must also. Thus:

Corollary 5.4.17. *The orientation-preserving isometries form a subgroup,*

$$\mathcal{O}_2 \subset \mathcal{I}_2.$$

Another proof of this comes from the fact that $\mathcal{O}_2 = \ker \text{sgn}$. This also shows that $\mathcal{O}_2 \triangleleft \mathcal{I}_2$, though that is also implicit in Corollary 5.4.16: any conjugate of an orientation-preserving isometry is orientation-preserving.

5.5. Calculus of isometries of \mathbb{R}^2 . In principle, Chapter 2 and Section 5.3 tell us everything we want to know about \mathcal{I}_2 and its composition law. But in fact, there is more geometry to be uncovered. There are two families of isometries we have not discussed yet.

Definition 5.5.1. Let $x \in \mathbb{R}^2$. The rotation, $\rho_{(x,\theta)}$, about x by the angle θ is $\tau_x \rho_{(0,\theta)} \tau_{-x}$. It rotates the rays emanating from x radially by the angle θ .

Rotations about a fixed x compose with one another as expected.

Lemma 5.5.2. $\rho_{(x,\theta)}\rho_{(x,\phi)} = \rho_{(x,\theta+\phi)}$. For $0 < \theta < 2\pi$ the fixed-point set $(\mathbb{R}^2)^{\rho_{(x,\theta)}} = \{x\}$.

Proof.

$$\begin{aligned} \rho_{(x,\theta)}\rho_{(x,\phi)} &= \tau_x \rho_{(0,\theta)} \tau_{-x} \tau_x \rho_{(0,\phi)} \tau_{-x} \\ &= \tau_x \rho_{(0,\theta)} \rho_{(0,\phi)} \tau_{-x} \\ &= \tau_x \rho_{(0,\theta+\phi)} \tau_{-x} \\ &= \rho_{(x,\theta+\phi)}. \end{aligned}$$

For the second statement we first consider the case $x = 0$. Here $\rho_{(0,\theta)}$ is the linear transformation T_{R_θ} induced by the rotation matrix R_θ . A fixed-point y of T_{R_θ} is a vector y such that $R_\theta y = y$, i.e., $(I - R_\theta)y = 0$, with I the identity matrix. But

$$I - R_\theta = \begin{bmatrix} 1 - \cos \theta & \sin \theta \\ -\sin \theta & 1 - \cos \theta \end{bmatrix}$$

has determinant $2(1 - \cos \theta)$, which is nonzero for $0 < \theta < 2\pi$. Therefore, $I - R_\theta$ is invertible for $0 < \theta < 2\pi$, and hence $(I - R_\theta)y = 0$ implies that

$y = 0$. Thus, the fixed-point set of $\rho_{(0,\theta)}$ is $\{0\}$.⁷ For general x , we apply the following lemma, which actually applies to arbitrary group actions on sets. \square

Lemma 5.5.3. *Let $\alpha, \beta \in \mathcal{I}_n$. Then $(\mathbb{R}^n)^{\alpha\beta\alpha^{-1}} = \alpha((\mathbb{R}^n)^\beta)$.*

Proof.

$$\begin{aligned} x \in (\mathbb{R}^n)^{\alpha\beta\alpha^{-1}} &\Leftrightarrow \alpha\beta\alpha^{-1}(x) = x \\ &\Leftrightarrow \beta\alpha^{-1}(x) = \alpha^{-1}(x) \\ &\Leftrightarrow \alpha^{-1}(x) \in (\mathbb{R}^n)^\beta \\ &\Leftrightarrow x \in \alpha((\mathbb{R}^n)^\beta). \end{aligned} \quad \square$$

Thus, every nonidentity rotation has exactly one fixed-point. When we complete the classification of the isometries of \mathbb{R}^2 we will obtain the converse: every isometry of \mathbb{R}^2 with exactly one fixed-point is a rotation about that point.

5.5.1. Glide reflections. We have one more infinite family of isometries to define.

Definition 5.5.4. Let ℓ be a line in \mathbb{R}^2 . A glide reflection with axis ℓ is a composite $\tau_x\sigma_\ell$ with $x \parallel \ell$. (By our definitions, this requires that $x \neq 0$, so a glide reflection is not a reflection.) We call $\tau_x\sigma_\ell$ with $x \parallel \ell$ the standard form of this glide reflection and show below it is unique.

Lemma 5.5.5. *Let ℓ be a line in \mathbb{R}^2 and let $x \parallel \ell$. Then τ_x commutes with σ_ℓ , and hence $(\tau_x\sigma_\ell)^2 = \tau_x^2 = \tau_{2x}$. The fixed-point set $(\mathbb{R}^2)^{\tau_x\sigma_\ell} = \emptyset$, i.e., a glide reflection has no fixed-points.*

Proof. Let $\ell = y + \text{span}(v)$ with v a unit vector. Then

$$\begin{aligned} \sigma_\ell\tau_x(z) &= z + x - 2\langle z + x - y, v^\perp \rangle v^\perp \\ &= z - 2\langle z - y, v^\perp \rangle v^\perp + x - 2\langle x, v^\perp \rangle v^\perp. \end{aligned}$$

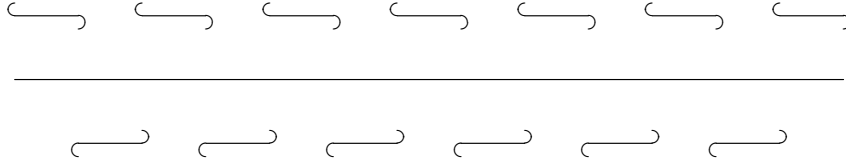
But $\langle x, v^\perp \rangle = 0$ as $x \parallel \ell$, so the latter is just $\tau_x\sigma_\ell(z)$. Thus τ_x and σ_ℓ commute.

Thus, $(\tau_x\sigma_\ell)^2 = \tau_x\sigma_\ell\tau_x\sigma_\ell = \tau_x^2\sigma_\ell^2 = \tau_x^2 = \tau_{2x}$, as $\sigma_\ell^2 = \text{id}$. Now, τ_{2x} is a nonidentity translation, and has no fixed-points. But any fixed-point of α is a fixed-point of α^2 , so $\tau_x\sigma_\ell$ has no fixed-points. \square

Pictorially, a glide reflection with axis ℓ looks like footsteps walking along the line ℓ . The glide reflection flips each left-footstep to a right-footstep, advanced along ℓ by the amount of the glide τ_x . Similarly each right-footstep flips and glides to a left-footstep:

We can detect the axis of a glide reflection in the following way.

⁷The argument here is a direct proof that 1 is not an eigenvalue for R_θ for $0 < \theta < 2\pi$. In fact, for these θ , there are no real eigenvalues for R_θ , as the characteristic polynomial of R_θ has no real roots.



Lemma 5.5.6. *Let $\alpha = \tau_x\sigma_\ell$ be a glide reflection in standard form (i.e., $x \parallel \ell$). Then ℓ is the only line preserved by α (i.e., the only line m with $\alpha(m) = m$).*

Proof. First note that ℓ is preserved by α as σ_ℓ fixes ℓ pointwise, while τ_x preserves ℓ by Proposition 2.1.14. (In particular, the effect of α on ℓ is translation by x along ℓ .)

Let m be a line not parallel to ℓ and let $z = m \cap \ell$. Then

$$\alpha(z) = \alpha(m) \cap \alpha(\ell) = \alpha(m) \cap \ell.$$

Since α has no fixed-points, $\alpha(z) \neq z$. Moreover, $\alpha(z) \in \ell$, but $\alpha(z) \neq z = \ell \cap m$, so $\alpha(z) \notin m$. So $\alpha(m) \neq m$.

Now let $m \parallel \ell$ with $m \neq \ell$. Since $m \cap \ell = \emptyset$, m must lie in either the positive or negative part of $\mathbb{R}^2 - \ell$ as described in Corollary 5.1.13. Since these two parts are interchanged by σ_ℓ and preserved by τ_x , $\alpha(m) \neq m$.

More explicitly, if $m \parallel \ell$ and $m \neq \ell$, write $\ell = y + \text{span}(v)$ for a unit vector v . Since $x \parallel \ell$, $x = sv$ for some $0 \neq s \in \mathbb{R}$. Since $m \parallel \ell$, $m = z + \text{span}(v)$ for some z . Write $z = y + uv + tv^\perp$. Then $z - uv \in m$, so $m = z' + \text{span}(v)$ with $z' = y + tv^\perp$. Since $m \neq \ell$, $tv^\perp = z' - y \notin \text{span}(v)$, so $t \neq 0$.

The generic element of m then has the form $y + tv^\perp + av$, and

$$\sigma_\ell(y + tv^\perp + av) = y - tv^\perp + av.$$

So $\tau_x\sigma_\ell(y + tv^\perp + av) = y - tv^\perp + (a + s)v$. In particular,

$$\alpha(m) = (y - tv^\perp) + \text{span}(v) = \sigma_\ell(z') + \text{span}(v).$$

But $\sigma_\ell(z') - z' = -2tv^\perp \notin \text{span}(v)$ So $\alpha(m) \neq m$. □

Corollary 5.5.7. *The standard form of a glide reflection is unique: if*

$$\tau_x\sigma_\ell = \tau_y\sigma_m$$

with $x \parallel \ell$ and $y \parallel m$, then $x = y$ and $\ell = m$.

Proof. By Lemma 5.5.6, ℓ and m must coincide with the unique line preserved by the glide reflection in question, so $\ell = m$. But then

$$\tau_x = \tau_x\sigma_\ell\sigma_\ell = \tau_y\sigma_m\sigma_\ell = \tau_y,$$

so $x = y$. □

5.5.2. Calculating composites of isometries. We have seen that every isometry of \mathbb{R}^2 either has the form $\tau_x\rho_{(0,\theta)}$ or $\tau_x\sigma_{\ell_\theta}$ for some $x \in \mathbb{R}^2$ and $\theta \in \mathbb{R}$. We shall show that if $0 < \theta < 2\pi$ then $\tau_x\rho_{(0,\theta)}$ is a rotation about some point by θ and that $\tau_x\sigma_{\ell_\theta}$ is either a reflection or a glide reflection depending on whether x is perpendicular to ℓ_θ or not. Thus, we have constructed all the isometries of \mathbb{R}^2 already and understand their geometric properties. We first treat the case of rotations.

We first put $\rho_{(y,\theta)}$ as the composite of a translation and a linear rotation.

Lemma 5.5.8. *Let $y \in \mathbb{R}^2$ and $\theta \in \mathbb{R}$. Then*

$$(5.5.1) \quad \rho_{(y,\theta)} = \tau_{(I-R_\theta)y}\rho_{(0,\theta)},$$

where R_θ is the standard 2×2 rotation matrix and I is the identity matrix.

Proof. By definition, $\rho_{(y,\theta)} = \tau_y\rho_{(0,\theta)}\tau_{-y}$. Since $\rho_{(0,\theta)}$ is linear, we may apply Proposition 3.3.5 (with γ the identity transformation), obtaining

$$\rho_{(y,\theta)} = \tau_{y+\rho_{(0,\theta)}(-y)}\rho_{(0,\theta)}.$$

But $\rho_{(0,\theta)}$ is multiplication by the rotation matrix R_θ , so

$$y + \rho_{(0,\theta)}(-y) = y - R_\theta y = (I - R_\theta)y. \quad \square$$

Proposition 5.5.9. *Let $0 < \theta < 2\pi$ and let $x \in \mathbb{R}^2$. Then $\tau_x\rho_{(0,\theta)} = \rho_{(y,\theta)}$ for some $y \in \mathbb{R}^2$.*

Proof. By (5.5.1) it suffices to solve $(I - R_\theta)y = x$. Since $0 < \theta < 2\pi$, $(I - R_\theta)$ is invertible as shown in the proof of Lemma 5.5.1. hence $y = (I - R_\theta)^{-1}x$ is the unique solution. \square

Remark 5.5.10. We have seen that the orientation-preserving isometries of \mathbb{R}^2 are precisely those of the form $\tau_x\rho_{(0,\theta)}$ for some $x \in \mathbb{R}^2$ and $\theta \in \mathbb{R}$. By the above analysis, these are either translations (θ a multiple of 2π) or rotations. The composite of orientation-preserving isometries is orientation-preserving. We shall compute these composites more precisely.

Corollary 5.5.11. *Let $x, y \in \mathbb{R}^2$ and $\theta, \phi \in \mathbb{R}$. If $\theta + \phi$ is not a multiple of 2π then*

$$\rho_{(x,\theta)}\rho_{(y,\phi)} = \rho_{(z,\theta+\phi)}$$

for some $z \in \mathbb{R}^2$. Otherwise, $\rho_{(x,\theta)}\rho_{(y,\phi)}$ is a translation.

Proof. By (5.5.1) there are vectors v and w with

$$\begin{aligned} \rho_{(x,\theta)} &= \tau_v\rho_{(0,\theta)}, \\ \rho_{(y,\phi)} &= \tau_w\rho_{(0,\phi)}. \end{aligned}$$

So

$$\begin{aligned} \rho_{(x,\theta)}\rho_{(y,\phi)} &= \tau_v\rho_{(0,\theta)}\tau_w\rho_{(0,\phi)} \\ &= \tau_{v+R_\theta w}\rho_{(0,\theta+\phi)} \end{aligned}$$

by Proposition 3.3.5. If $\theta + \phi$ is not a multiple of 2π the result follows from Proposition 5.5.9. Otherwise, $\rho_{(0,\theta+\phi)}$ is the identity and we are done. \square

We can also give a nice theoretical calculation of compositions of translations with rotations.

Corollary 5.5.12. *Let $x, y \in \mathbb{R}^2$ and $\theta \in \mathbb{R}$. Then there exist $z, w \in \mathbb{R}^2$ with*

$$\begin{aligned}\tau_x \rho_{(y,\theta)} &= \rho_{(z,\theta)}, \\ \rho_{(y,\theta)} \tau_x &= \rho_{(w,\theta)}.\end{aligned}$$

Proof. These are immediate from Proposition 3.3.5, (5.5.1) and Proposition 5.5.9. \square

All the above calculations can be carried out explicitly in full using the formula for inverting a 2×2 matrix. The results are numerically ugly as trig functions are numerically ugly. We will introduce a geometric calculus for carrying out these calculations, below, but the results are numerically appealing only when the trig functions are nicely computable.

We shall now analyze the orientation-reversing isometries of \mathbb{R}^2 and their compositions with each other and with the orientation-preserving isometries. One immediate result is the following. Recall from Proposition 5.1.21 that if ℓ and m are parallel lines in \mathbb{R}^2 , then $\sigma_m \sigma_\ell$ is the translation by twice the directed distance from ℓ to m . The other composites of two reflections are given as follows.

Lemma 5.5.13. *Let ℓ and m be nonparallel lines in \mathbb{R}^2 and let $x = \ell \cap m$. Let v and w be unit vectors parallel to ℓ and m , respectively, and let $v = \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}$ and $w = \begin{bmatrix} \cos \phi \\ \sin \phi \end{bmatrix}$. Then*

$$(5.5.2) \quad \sigma_m \sigma_\ell = \rho_{(x, 2(\phi - \theta))},$$

the rotation about $x = \ell \cap m$ by twice the directed angle from ℓ to m .

Proof. Recall that for $\psi \in \mathbb{R}$, $\ell_\psi = \text{span} \left(\begin{bmatrix} \cos \psi \\ \sin \psi \end{bmatrix} \right)$. In particular, $\text{span}(v) = \ell_\theta$ and $\text{span}(w) = \ell_\phi$, and hence $\ell = \tau_x(\ell_\theta)$ and $m = \tau_x(\ell_\phi)$. By Lemma 5.1.14,

$$\begin{aligned}\sigma_\ell &= \tau_x \sigma_{\ell_\theta} \tau_{-x}, \\ \sigma_m &= \tau_x \sigma_{\ell_\phi} \tau_{-x}.\end{aligned}$$

Thus,

$$\begin{aligned}\sigma_m \sigma_\ell &= \tau_x \sigma_{\ell_\phi} \tau_{-x} \tau_x \sigma_{\ell_\theta} \tau_{-x} \\ &= \tau_x \sigma_{\ell_\phi} \sigma_{\ell_\theta} \tau_{-x} \\ &= \tau_x \rho_{(0, 2(\phi - \theta))} \tau_{-x} \\ &= \rho_{(x, 2(\phi - \theta))}.\end{aligned}$$

Here, we used Proposition 5.3.8 to evaluate $\sigma_{\ell_\phi} \sigma_{\ell_\theta}$. \square

Remark 5.5.14. As in Proposition 5.3.8, while the directed angle from ℓ to m is not well-defined, twice the directed angle is well-defined. Choosing v and w above amounts to orienting ℓ and m , and with those orientations, the directed angle makes sense. But a choice of the opposite orientation in either case would add or subtract π from that directed angle, and when the angle is doubled, the extra π goes away.

In other words, if ψ is the directed angle from v to w , then

$$\sigma_m \sigma_\ell = \rho_{(x, 2\psi)} = \rho_{(x, 2(\psi \pm \pi))}.$$

Moreover,

$$m = \rho_{(x, \psi)}(\ell) = \rho_{(x, \psi \pm \pi)}(\ell).$$

We obtain the following corollary.

Corollary 5.5.15. *Let $x \in \mathbb{R}^2$ and $0 < \theta < 2\pi$. Let ℓ be any line through x . Let $m = \rho_{(x, \frac{\theta}{2})}(\ell)$ and let $n = \rho_{(x, -\frac{\theta}{2})}(\ell)$. Then*

$$(5.5.3) \quad \rho_{(x, \theta)} = \sigma_m \sigma_\ell = \sigma_\ell \sigma_n.$$

Moreover, m and n are the unique lines satisfying (5.5.3).

Proof. Uniqueness follows from the argument in the remark above, but it also follows from group theory: if $\sigma_m \sigma_\ell = \sigma_{m'} \sigma_\ell$, then

$$\begin{aligned} \sigma_m \sigma_\ell \sigma_\ell^{-1} &= \sigma_{m'} \sigma_\ell \sigma_\ell^{-1} \\ \sigma_m &= \sigma_{m'}. \end{aligned}$$

Similarly multiplication on the left by σ_ℓ^{-1} proves the uniqueness of n . Of course, since reflections are involutions, $\sigma_\ell^{-1} = \sigma_\ell$. \square

A key now is to evaluate the composite of a translation τ_x and a reflection σ_ℓ . We know that if $x \parallel \ell$ then the result is a glide reflection and cannot be meaningfully simplified. When $x \perp \ell$ we get a nice, clean result.

Lemma 5.5.16. *Let $x \perp \ell$ in \mathbb{R}^2 . Then*

$$\begin{aligned} \tau_x \sigma_\ell &= \sigma_{\tau_{\frac{x}{2}}(\ell)}, \\ \sigma_\ell \tau_x &= \sigma_{\tau_{-\frac{x}{2}}(\ell)}, \end{aligned}$$

reflections in lines parallel to ℓ .

Proof. We apply Corollary 5.1.22. Let $m = \tau_{\frac{x}{2}}(\ell)$ and let $n = \tau_{-\frac{x}{2}}(\ell)$. Then $\tau_x = \sigma_m \sigma_\ell$, so $\tau_x \sigma_\ell = \sigma_m \sigma_\ell \sigma_\ell = \sigma_m$, as reflections are involutions. Also, $\tau_x = \sigma_\ell \sigma_n$, so $\sigma_\ell \tau_x = \sigma_\ell \sigma_\ell \sigma_n = \sigma_n$. \square

Finally, we address the general case of composition of reflections and translations.

Proposition 5.5.17. *Let $\ell = x + \text{span}(v)$ be a line in \mathbb{R}^2 with v a unit vector. Let $y \in \mathbb{R}^2$. Write $y = z + w$ with $z \in \text{span}(v)$ and $w \in \text{span}(v^\perp)$ (hence $z = \langle y, v \rangle v$ and $w = \langle y, v^\perp \rangle v^\perp$, as v, v^\perp is an orthonormal basis of \mathbb{R}^2). Then*

$$\begin{aligned}\tau_y \sigma_\ell &= \tau_z \sigma_{\tau_{\frac{w}{2}}(\ell)}, \\ \sigma_\ell \tau_y &= \tau_z \sigma_{\tau_{-\frac{w}{2}}(\ell)}.\end{aligned}$$

If $y \perp \ell$, then $z = 0$ and these are reflections in lines parallel to ℓ . Otherwise, $z \neq 0$ and these are glide reflections in standard form, with axes are parallel to ℓ .

Regardless, if α is either a reflection or a glide reflection with axis ℓ , then $\tau_y \alpha$ and $\alpha \tau_y$ are either reflections or glide reflections with axis parallel to ℓ .

Proof. We have $\tau_y = \tau_z \tau_w$, and τ_z commutes with τ_w and σ_ℓ as either $z = 0$ or $z \parallel \ell$. Now apply Lemma 5.5.16 to the appropriate composite of τ_w and σ_ℓ . \square

Since every orientation-reversing isometry has the form $\tau_x \sigma_{\ell_\theta}$ for some θ , we obtain the following.

Corollary 5.5.18. *Every orientation-reversing isometry of \mathbb{R}^2 is either a reflection or a glide reflection.*

In summation, we have:

Theorem 5.5.19. *Every isometry of \mathbb{R}^2 is either a translation, a rotation, a reflection or a glide reflection. The former two are orientation-preserving and the latter two are orientation-reversing. Both translations and glide reflections are without fixed-points. The fixed-point set of a nonidentity rotation consists of only the point about which it rotates. The fixed-point set of a reflection is the line of reflection.*

We can now analyze conjugation in \mathcal{I}_2 in some detail:

Theorem 5.5.20. *Let $\alpha \in \mathcal{I}_2$. Then:*

- (1) *If $\alpha = \tau_z \beta$ with $\beta : \mathbb{R}^n \rightarrow \mathbb{R}^n$ a linear isomorphism, then*

$$\alpha \tau_x \alpha^{-1} = \tau_{\beta(x)}.$$

Phrased entirely in terms of α this says $\alpha \tau_x \alpha^{-1} = \tau_w$, where $w = \alpha(x) - \alpha(0)$.

- (2) $\alpha \sigma_\ell \alpha^{-1} = \sigma_{\alpha(\ell)}$.
 (3) $\alpha \rho_{(x,\theta)} \alpha^{-1} = \rho_{(\alpha(x),\psi)}$, where $\psi = \theta$ if α is orientation-preserving and $\psi = -\theta$ if α is orientation-reversing.
 (4) *If γ is a glide reflection with axis ℓ , then $\alpha \gamma \alpha^{-1}$ is a glide reflection with axis $\alpha(\ell)$.*

Proof. (1) is just Corollary 3.3.2 for $n = 2$. For (2), we have

$$(\mathbb{R}^2)^{\alpha\sigma_\ell\alpha^{-1}} = \alpha((\mathbb{R}^2)^{\sigma_\ell})$$

by Lemma 5.5.3, but this is just $\alpha(\ell)$. The only isometry with fixed-point set $\alpha(\ell)$ is $\sigma_{\alpha(\ell)}$.

In case (3), a similar argument shows the fixed set of $\alpha\rho_{(x,\theta)}\alpha^{-1}$ is $\alpha(x)$, so $\alpha\rho_{(x,\theta)}\alpha^{-1}$ is the rotation about $\alpha(x)$ by some angle ψ . But determining ψ requires further work. So we take a different approach. Let ℓ be any line through x and let $m = \rho_{(x, \frac{\theta}{2})}(\ell)$. Then $\rho_{(x,\theta)} = \sigma_m\sigma_\ell$ by Corollary 5.5.15. Thus,

$$\begin{aligned} \alpha\rho_{(x,\theta)}\alpha^{-1} &= (\alpha\sigma_m\alpha^{-1})(\alpha\sigma_\ell\alpha^{-1}) \\ &= \sigma_{\alpha(m)}\sigma_{\alpha(\ell)}. \end{aligned}$$

This last is the rotation about $\alpha(m \cap \ell)$ by twice the directed angle from $\alpha(\ell)$ to $\alpha(m)$. Since orientation-preserving isometries preserve angles and orientation-reversing isometries reverse them, (3) follows.

For (4), write $\gamma = \tau_x\sigma_\ell$, with $x \parallel \ell$. Thus, we may write $\ell = y + \text{span}(x)$ for some $y \in \mathbb{R}^2$. We have

$$\begin{aligned} \alpha\gamma\alpha^{-1} &= (\alpha\tau_x\alpha^{-1})(\alpha\sigma_\ell\alpha^{-1}) \\ &= \tau_w\sigma_{\alpha(\ell)} \end{aligned}$$

with $w = \alpha(x) - \alpha(0)$, by (1) and (2). By Corollary 2.5.4,

$$\alpha(\ell) = \alpha(y) + \text{span}(\alpha(x) - \alpha(0)),$$

so $w \parallel \alpha(\ell)$ and this is a glide reflection in standard form with axis $\alpha(\ell)$. \square

We also have all the ingredients to analyze compositions of isometries. Let us consider the compositions of rotations with reflections.

Proposition 5.5.21. *Let ℓ be a line in \mathbb{R}^2 and let $x \in \ell$. Then $\rho_{(x,\theta)}\sigma_\ell$ and $\sigma_\ell\rho_{(x,\theta)}$ are both reflections in lines through x for every $\theta \in \mathbb{R}$.*

Proof. Let m be the line through x such that the directed angle from some orientation of ℓ to some orientation of m is $\frac{\theta}{2}$. Then $\rho_{(x,\theta)} = \sigma_m\sigma_\ell$ so $\rho_{(x,\theta)}\sigma_\ell = \sigma_m\sigma_\ell\sigma_\ell = \sigma_m$.

The other case is similar. \square

Proposition 5.5.22. *Let ℓ be a line in \mathbb{R}^2 and let $x \notin \ell$. Let $0 < \theta < 2\pi$. Then $\rho_{(x,\theta)}\sigma_\ell$ and $\sigma_\ell\rho_{(x,\theta)}$ are both glide reflections.*

Proof. Let m be the unique line through x parallel to ℓ and let n be the line through x such that the directed angle from some orientation of m to some orientation of n is $\frac{\theta}{2}$. Then $\rho_{(x,\theta)} = \sigma_n\sigma_m$ so $\rho_{(x,\theta)}\sigma_\ell = \sigma_n\sigma_m\sigma_\ell$. By Proposition 5.1.21, $\sigma_m\sigma_\ell = \tau_v$ for a nonzero vector $v \perp m$ (nonzero because $m \neq \ell$). Thus, $\rho_{(x,\theta)}\sigma_\ell = \sigma_n\tau_v$. Because $0 < \theta < 2\pi$, $0 < \frac{\theta}{2} < \pi$, and hence v is not perpendicular to n , so the result is a glide reflection by Proposition 5.5.17. Note the axis is parallel to n and not ℓ .

The other case is similar. \square

5.5.3. Calculus of reflections. Corollary 5.1.22 shows that every translation is the product of two reflections. Corollary 5.5.15 shows that every rotation is the product of two reflections. Thus, every orientation-preserving isometry of \mathbb{R}^2 is a product of two reflections.

The orientation-reversing isometries of \mathbb{R}^2 are either reflections or glide reflections, and the latter are composites of reflections and translations. So every orientation-reversing isometry of \mathbb{R}^2 is either a reflection or the product of three reflections.

We can use Corollaries 5.1.22 and 5.5.15 to develop a calculus for composing isometries useful for both practical and theoretical results. The following example is representative.

The following example is useful in studying wallpaper groups.

Example 5.5.23. We calculate the composite $\alpha = \rho_{(0, \frac{\pi}{3})} \rho_{([\frac{2}{0}], \frac{2\pi}{3})}$. We do so by writing $\rho_{(0, \frac{\pi}{3})} = \sigma_\ell \sigma_m$ and writing $\rho_{([\frac{2}{0}], \frac{2\pi}{3})} = \sigma_m \sigma_n$. This then gives

$$\alpha = \sigma_\ell \sigma_m \sigma_m \sigma_n = \sigma_\ell \sigma_n,$$

as σ_m is an involution.

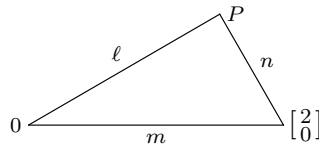
We can do this because of the flexibility of Corollary 5.5.15. The equation $\rho_{(0, \frac{\pi}{3})} = \sigma_\ell \sigma_m$ is equivalent to saying that $\ell \cap m = 0$ and the directed angle from m to ℓ is $\frac{\pi}{6}$, while $\rho_{([\frac{2}{0}], \frac{2\pi}{3})} = \sigma_m \sigma_n$ says $m \cap n = [\frac{2}{0}]$ and the directed angle from n to m is $\frac{\pi}{3}$. In particular, m must go through both 0 and $[\frac{2}{0}]$, and hence must be the x -axis.

The directed angles now allow us to precisely calculate the lines ℓ and n . ℓ is the line through 0 such that the directed angle from the positive x -axis to ℓ is $\frac{\pi}{6}$. Thus, ℓ has slope $\tan \frac{\pi}{6} = \frac{1}{\sqrt{3}}$. Since $0 \in \ell$, ℓ is the line $y = \frac{1}{\sqrt{3}}x$.

On the other hand, n is the line through $[\frac{2}{0}]$ such that the angle from n to the x -axis is $\frac{\pi}{3}$, so the angle from the x -axis to n is $-\frac{\pi}{3}$. But this says n has slope $-\sqrt{3}$. Since n goes through $[\frac{2}{0}]$, the point-slope formula gives

$$\frac{y - 0}{x - 2} = -\sqrt{3},$$

so n is the line $y = -\sqrt{3}x + 2\sqrt{3}$.



Now $\alpha = \sigma_\ell \sigma_n = \rho_{(P, \theta)}$, where $P = \ell \cap n$ and θ is twice the directed angle from n to ℓ . By Corollary 5.5.11, $\theta = \frac{\pi}{3} + \frac{2\pi}{3} = \pi$, so it suffices to calculate P . We do this by setting $\frac{1}{\sqrt{3}}x = -\sqrt{3}x + 2\sqrt{3}$. This gives $P = \left[\begin{array}{c} \frac{3}{2} \\ \sqrt{3} \end{array} \right]$.

We can use the same argument to prove the following.

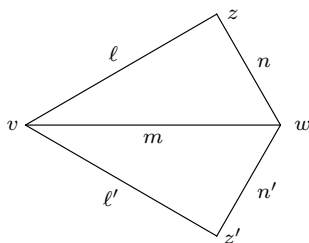
Lemma 5.5.24. *Let $v \neq w \in \mathbb{R}^2$ and suppose $\theta + \phi$ is not a multiple of 2π . Write $\rho_{(v,\theta)}\rho_{(w,\phi)} = \rho_{(z,\theta+\phi)}$ as in Corollary 5.5.11 and let m be the line containing v and w . Then $z \notin m$ and $\rho_{(w,\phi)}\rho_{(v,\theta)} = \rho_{(\sigma_m(z),\theta+\phi)}$. In particular, $\rho_{(v,\theta)}$ and $\rho_{(w,\phi)}$ do not commute.*

Proof. As in Example 5.5.23, we write $\rho_{(v,\theta)} = \sigma_\ell\sigma_m$ and $\rho_{(w,\phi)} = \sigma_m\sigma_n$, giving $\rho_{(v,\theta)}\rho_{(w,\phi)} = \sigma_\ell\sigma_n$.

But we could just as easily have used Corollary 5.5.15 to write $\rho_{(v,\theta)} = \sigma_m\sigma_{\ell'}$ and $\rho_{(w,\phi)} = \sigma_{n'}\sigma_m$, and this allows us to write

$$\rho_{(w,\phi)}\rho_{(v,\theta)} = \sigma_{n'}\sigma_m\sigma_m\sigma_{\ell'} = \sigma_{n'}\sigma_{\ell'}.$$

In this procedure, the directed angle from ℓ' to m is opposite to the directed angle from m to ℓ , while the directed angle from m to n' is opposite to the directed angle from m to n (Corollary 5.5.15):



Since orientation-reversing isometries reverse directed angles and since σ_m fixes m (and hence also fixes v and w), the angle reversal defining ℓ' and n' shows that $\ell' = \sigma_m(\ell)$ and $n' = \sigma_m(n)$. Thus, the intersection, z' , of ℓ' and n' is $\sigma_m(\ell) \cap \sigma_m(n) = \sigma_m(z)$.

By Corollary 5.5.11, $\rho_{(w,\phi)}\rho_{(v,\theta)} = \rho_{(z',\theta+\phi)}$, so the result follows. \square

This covers the most general case of the following proposition, which is useful in characterizing the finite groups of symmetries in \mathbb{R}^2 .

Proposition 5.5.25. *Let $v \neq w \in \mathbb{R}^2$ and let θ, ϕ be arbitrary elements of $(0, 2\pi)$. Then $\rho_{(v,\theta)}$ and $\rho_{(w,\phi)}$ do not commute.*

Proof. By Lemma 5.5.24 we need only consider the case where $\phi + \theta$ is a multiple of 2π . In this case, both $\rho_{(v,\theta)}\rho_{(w,\phi)}$ and $\rho_{(w,\phi)}\rho_{(v,\theta)}$ are translations. Again we can write

$$\rho_{(v,\theta)} = \sigma_\ell\sigma_m = \sigma_m\sigma_{\ell'}$$

$$\rho_{(w,\phi)} = \sigma_m\sigma_n = \sigma_{n'}\sigma_m$$

with m the line containing v and w , and again

$$\rho_{(v,\theta)}\rho_{(w,\phi)} = \sigma_\ell\sigma_n$$

$$\rho_{(w,\phi)}\rho_{(v,\theta)} = \sigma_{n'}\sigma_{\ell'}.$$

Since these composites are translations, we have $\ell \parallel n$ and $n' \parallel \ell'$, and the argument in Lemma 5.5.24 again shows that $\sigma_m(\ell) = \ell'$ and $\sigma_m(n) = n'$.

We first consider the case $\theta = \pi$ (and hence $\phi = \pi$). In this case ℓ and n are perpendicular to m . By the angle reversal, this gives $\ell' = \ell$ and $n' = n$, so $\rho_{(w,\phi)}\rho_{(v,\theta)} = \sigma_n\sigma_\ell$. This just reverses the direction in the directed distance used to calculate the translation vector, so $\rho_{(w,\phi)}\rho_{(v,\theta)}$ is the translation inverse to $\rho_{(v,\theta)}\rho_{(w,\phi)}$.

In the remaining case, ℓ and ℓ' have different slopes. Since the translation vectors for the two composites are perpendicular to the lines ℓ and ℓ' , respectively, these translation vectors also have different slopes. \square

The calculus of reflections permits a significant strengthening of Proposition 5.5.22. It will play an important role in our study of wallpaper groups.

Proposition 5.5.26. *Let ℓ be a line in \mathbb{R}^2 and let $y \notin \ell$. Let A be the directed distance from y to ℓ , i.e., if p is the line through y perpendicular to ℓ , then $p \cap \ell = y + A$. Let $B = \rho_{(0, \frac{\pi}{2})}(A)$. Let $0 \neq \theta \in \mathbb{R}$ and let $\phi = \frac{\theta}{2}$. Let q be the line through $y + A$ such that the directed angle from ℓ to q is ϕ and let*

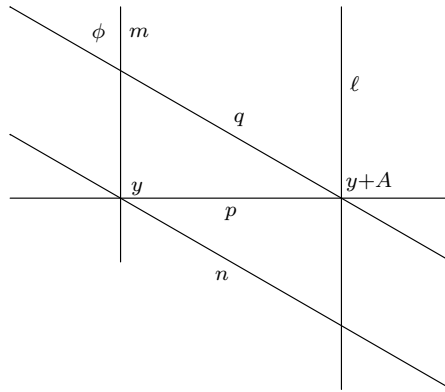
$$w = -\sin \phi A + \cos \phi B.$$

Then $w \parallel q$, and

$$(5.5.4) \quad \rho_{(y,\theta)}\sigma_\ell = \tau_{(2 \sin \phi w)}\sigma_q,$$

a glide reflection in standard form. Pictorially, we have:

(5.5.5)



Here, m is the line through y parallel to ℓ and n is the line through y parallel to q .

Since $\|w\| = \|A\|$, the length of the glide is $2|\sin \phi| \|A\|$.

Proof. Since the directed angle from m to n is ϕ , our calculus of reflections gives

$$\rho_{(y,\theta)}\sigma_\ell = \sigma_n\sigma_m\sigma_\ell.$$

Since A is the directed distance from m to ℓ , $\sigma_m\sigma_\ell = \tau_{-2A}$, and hence

$$\rho_{(y,\theta)}\sigma_\ell = \sigma_n\tau_{(-2A)}.$$

The directed angle from p to n is $\phi + \frac{\pi}{2}$. Since w is visibly equal to $\rho_{(0, \phi + \frac{\pi}{2})}(A)$, w is parallel to n , and hence to q .

Let

$$v = \rho_{(0, \frac{\pi}{2})}(w) = \rho_{(0, \phi + \pi)}(A) = -\cos \phi A - \sin \phi B.$$

Then $v \perp w$, hence $v \perp n$, and

$$(5.5.6) \quad \cos \phi v + \sin \phi w = -\cos^2 \phi A - \sin^2 \phi A = -A.$$

Thus, $\tau_{-2A} = \tau_{(2 \cos \phi v)} \tau_{(2 \sin \phi w)}$, and hence

$$\rho_{(y, \theta)} \sigma_\ell = \sigma_n \tau_{(2 \cos \phi v)} \tau_{(2 \sin \phi w)}.$$

Since $v \perp n$, $\sigma_n \tau_{(2 \cos \phi v)}$ is the reflection across $\tau_{(-\frac{1}{2}(2 \cos \phi v))}(n)$. It suffices to show that $\tau_{(-\cos \phi v)}(n) = q$.

Now $n = y + \text{span}(w)$, so $\tau_{(-\cos \phi v)}(n) = y - \cos \phi v + \text{span}(w)$. But by (5.5.6), $y + A \in y - \cos \phi v + \text{span}(w)$. Since q is the line through $y + A$ parallel to n , the result follows. \square

The reverse composition behaves similarly. Here is the result. We leave the proof to the reader.

Proposition 5.5.27. *Let ℓ be a line in \mathbb{R}^2 and let $y \notin \ell$. Let A be the directed distance from y to ℓ , i.e., if p is the line through y perpendicular to ℓ , then $p \cap \ell = y + A$. Let $B = \rho_{(0, \frac{\pi}{2})}(A)$. Let $0 \neq \theta \in \mathbb{R}$ and let $\phi = \frac{\theta}{2}$. Let q be the line through $y + A$ such that the directed angle from q to ℓ is ϕ and let*

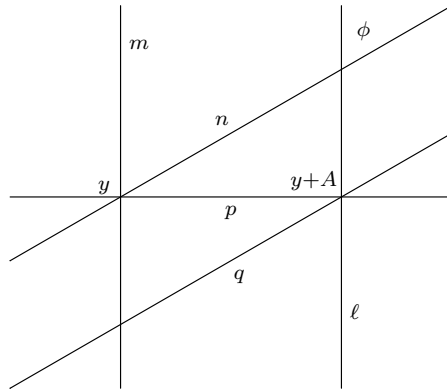
$$w = \sin \phi A + \cos \phi B.$$

Then $w \parallel q$, and

$$(5.5.7) \quad \sigma_\ell \rho_{(y, \theta)} = \tau_{(2 \sin \phi w)} \sigma_q,$$

a glide reflection in standard form. Pictorially, we have:

(5.5.8)

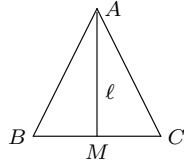


Here, m is the line through y parallel to ℓ , n is the line through y parallel to q . Again, the glide has length $2|\sin \phi| \|A\|$.

5.6. Classical results from Euclidean geometry. We now give derivations from our analytic geometry of some standard results in Euclidean geometry. We first give a converse to the Pons asinorum.

Proposition 5.6.1. *Let $\triangle ABC$ be a triangle such that $\angle ABC$ and $\angle ACB$ have the same unsigned measure. Then $d(A, B) = d(A, C)$.*

Proof. Let ℓ bisect the angle $\angle BAC$ and let $M = \ell \cap \overline{BC}$.



Since the measures of the three interior angles of a triangle must add up to π (Corollary 2.1.17), $\angle AMC$ and $\angle AMB$ have the same measure. But since these two measures add up to π , both must be right angles. But then, since ℓ bisects $\angle BAC$, $\frac{d(M, B)}{d(A, M)}$ and $\frac{d(M, C)}{d(A, M)}$ are equal to the tangents of equal angles. So $d(M, B) = d(M, C)$. But then $d(A, B) = d(A, C)$ by the Pythagorean theorem (a special case of the cosine law). \square

5.7. Exercises.

1. Show that a similarity of \mathbb{R}^2 with two fixed-points is an isometry.
2. Show that an orientation-preserving isometry of \mathbb{R}^2 with two fixed-points is the identity.
3. Show that an isometry of \mathbb{R}^2 fixing three noncollinear points is the identity. Here, the three points are noncollinear if there is no line containing all three of them.
4. What can you say about an orientation-reversing isometry with two fixed-points? Can you identify the isometry by knowing the two points?
5. Here, the standard form of an isometry is one of τ_x , $\rho_{(x, \theta)}$, σ_ℓ or $\tau_x \sigma_\ell$ with $x \parallel \ell$. Please specify the explicit values of x , θ , ℓ .
 - (a) Let ℓ be the line $y = \frac{1}{\sqrt{3}}x - \frac{2}{\sqrt{3}}$. Write $\alpha = \rho_{(0, \frac{\pi}{3})} \sigma_\ell$ in standard form.
 - (b) Let ℓ be the line $y = -\sqrt{3}x + 2$. Write $\alpha = \sigma_\ell \rho_{(0, -\frac{\pi}{3})}$ in standard form.
 - (c) Write $\alpha = \rho_{(0, \frac{\pi}{3})} \rho\left(\begin{bmatrix} \sqrt{3} \\ 1 \end{bmatrix}, \frac{\pi}{3}\right)$ in standard form.
 - (d) Write $\alpha = \rho\left(\begin{bmatrix} 2 \\ 0 \end{bmatrix}, \frac{\pi}{3}\right) \rho_{(0, \frac{2\pi}{3})}$ in standard form.
 - (e) Write $\alpha = \rho\left(\begin{bmatrix} 2 \\ 0 \end{bmatrix}, \frac{\pi}{3}\right) \rho_{(0, -\frac{\pi}{3})}$ in standard form.

- (f) Write $\alpha = \tau \begin{bmatrix} 2 \\ -2 \end{bmatrix} \rho_{(0, \frac{\pi}{2})}$ in standard form. (Hint: write the translation as $\sigma_\ell \sigma_m$ and the rotation as $\sigma_m \sigma_n$. Then ℓ and m are perpendicular to the translation vector and $m \cap n = 0$. This specifies m and that determines the other lines.)
- (g) Show that if $x \notin \ell$ and θ is not a multiple of 2π , then both $\sigma_\ell \rho_{(x, \theta)}$ and $\rho_{(x, \theta)} \sigma_\ell$ are glide reflections.

6. Groups of symmetries: planar figures

We introduce the general theory of symmetry groups by studying symmetry groups in the plane. In the planar case, we have developed enough background to study the symmetries in depth. We first explain the notion of the symmetry group of a subset of \mathbb{R}^n .

6.1. Symmetry in \mathbb{R}^n ; congruence and similarity.

6.1.1. The group of symmetries of $X \subset \mathbb{R}^n$.

Definition 6.1.1. Let $X \subset \mathbb{R}^n$. We write

$$\mathcal{S}(X) = \{\alpha \in \mathcal{I}_n : \alpha(X) = X\}.$$

We call this the group of symmetries of X , though in fact they are the symmetries of \mathbb{R}^n that carry X onto itself. (Thus, in some contexts, $\mathcal{S}(\mathbb{R}^n, X)$ would be a better notation: the symmetries of the pair (\mathbb{R}^n, X) .)

$\mathcal{S}(X)$ is easily seen to be a subgroup of \mathcal{I}_n : it is certainly closed under composition, and, because isometries are bijections, if $\alpha(X) = X$, the inverse function must carry X onto X as well.

Alternatively, we can ask when a subgroup of \mathcal{I}_n lies in $\mathcal{S}(X)$.

Definition 6.1.2. Let $H \subset \mathcal{I}_n$ and let $X \subset \mathbb{R}^n$. We say X is H -invariant if $\alpha(X) \subset X$ for all $\alpha \in H$.

These concepts fit together as follows.

Lemma 6.1.3. *Let $H \subset \mathcal{I}_n$ and let $X \subset \mathbb{R}^n$. Then X is H -invariant if and only if $H \subset \mathcal{S}(X)$.*

Proof. It suffices to show that if X is H -invariant then $\alpha(X) = X$ for all $\alpha \in H$. But this follows from the fact that H is a subgroup. If $\alpha \in H$, so is α^{-1} , and hence $\alpha^{-1}(X) \subset X$. But applying α to both sides now gives $X \subset \alpha(X)$. As $\alpha(X) \subset X$, the result follows. \square

Our first calculation has already been done for us by Proposition 2.5.1:

Example 6.1.4. The symmetry group of the origin in \mathbb{R}^n consists of the linear isometries of \mathbb{R}^n :

$$(6.1.1) \quad \mathcal{S}(\{0\}) = \mathcal{LI}_n,$$

which in turn is isomorphic to the orthogonal group $O(n)$.

Currently, we have a good understanding of $O(2)$, and have used it to classify the isometries of \mathbb{R}^2 . We will use that understanding to get a good handle on the symmetry groups of subsets of \mathbb{R}^2 .

We will study $O(3)$ in Chapter 7. It will be the basis of our understanding of the isometries of the sphere \mathbb{S}^2 (and hence our understanding of the geometry of the earth's surface). It is also the starting point for studying the isometries of \mathbb{R}^3 . The study of $O(n)$ for $n \geq 4$ is more difficult.

6.1.2. The subgroups $\mathcal{T}(X)$ and $\mathcal{O}(X)$ of $\mathcal{S}(X)$.

Definition 6.1.5. Let $X \subset \mathbb{R}^n$. We write

$$\mathcal{T}(X) = \{\tau_x : \tau_x \in \mathcal{S}(X)\}.$$

More generally, if H is a subgroup of \mathcal{I}_n , we write

$$\mathcal{T}(H) = H \cap \mathcal{T}_n,$$

the set of translations in H .

Note these are subgroups of $\mathcal{S}(X)$ and H , respectively, as the intersection of two subgroups of a group is always a subgroup.

It is also of value to study the subgroup of orientation-preserving symmetries of X . We have so far only developed the requisite theory for $n = 2$, where we are able to use signed angles to define orientation-preservation for nonlinear maps (Definition 5.4.12), and then show in Corollary 5.4.17 that the collection, \mathcal{O}_2 , of orientation-preserving isometries of \mathbb{R}^2 is a subgroup of \mathcal{I}^2 .

For $n > 2$ we do not have well-defined signed angles between pairs of vectors, and we need additional theory. We give such a treatment in Section 8.2. Corollary 8.2.2 identifies the orientation-preserving isometries of \mathbb{R}^n and shows that they form a subgroup $\mathcal{O}_n \subset \mathcal{I}_n$. Given that, we can make the following definitions.

Definition 6.1.6. Let $X \subset \mathbb{R}^n$, The orientation-preserving symmetries of X are

$$\mathcal{O}(X) = \{\alpha \in \mathcal{S}(X) : \alpha \text{ is orientation-preserving}\} = \mathcal{S}(X) \cap \mathcal{O}_n.$$

For an arbitrary subgroup $H \subset \mathcal{I}_n$ we write $\mathcal{O}(H)$ for the orientation-preserving elements of H :

$$\mathcal{O}(H) = H \cap \mathcal{O}_n.$$

There are inclusions of subgroups

$$\begin{aligned} \mathcal{T}(X) &\subset \mathcal{O}(X) \subset \mathcal{S}(X), \\ \mathcal{T}(H) &\subset \mathcal{O}(H) \subset H. \end{aligned}$$

6.1.3. Congruence and similarity.

Definition 6.1.7. The subsets X and Y of \mathbb{R}^n are congruent if there is an isometry α of \mathbb{R}^n with $\alpha(X) = Y$. The isometry α is said to be a congruence from X to Y . Thus, the symmetry group $\mathcal{S}(X)$ is the set of all congruences from X to itself.

We can now use the idea of conjugacy to relate the symmetry groups of congruent figures. Recall that if H is a subgroup of G and if $g \in G$ then the conjugate of H by g is the subgroup

$$gHg^{-1} = \{ghg^{-1} : h \in H\}.$$

By Proposition 3.4.15, conjugate subgroups of G are isomorphic.

Lemma 6.1.8. *Let $X \subset \mathbb{R}^n$ and let $\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be an isometry. Then the symmetry groups of X and of $\alpha(X)$ are conjugate by α , and hence isomorphic:*

$$\mathcal{S}(\alpha(X)) = \alpha\mathcal{S}(X)\alpha^{-1}.$$

Proof.

$$\begin{aligned} \beta\alpha(X) = \alpha(X) &\Leftrightarrow \alpha^{-1}\beta\alpha(X) = X \\ &\Leftrightarrow \alpha^{-1}\beta\alpha \in \mathcal{S}(X) \\ &\Leftrightarrow \beta \in \alpha\mathcal{S}(X)\alpha^{-1}. \quad \square \end{aligned}$$

Since $\{x\} = \tau_x(\{0\})$ we obtain the following:

Corollary 6.1.9. $\mathcal{S}(\{x\}) = \tau_x\mathcal{L}\mathcal{I}_n\tau_x^{-1}$.

Similar subsets of \mathbb{R}^n also have isomorphic symmetry groups. We have seen this concept before in the study of similar triangles in Euclidean geometry.

Definition 6.1.10. The subsets X and Y of \mathbb{R}^n are similar if there is a similarity $f \in \mathcal{S}_n$ of \mathbb{R}^n with $f(X) = Y$. We say f is a similarity from X to Y .

In this case, the fact that \mathcal{I}_n is normal in \mathcal{S}_n produces an interesting result. The point is that if H is a subgroup of \mathcal{I}_n and $f \in \mathcal{S}_n$, then

$$fHf^{-1} \subset f\mathcal{I}_nf^{-1} = \mathcal{I}_n.$$

In particular, if we conjugate an element of H by f , we get an isometry and not just a similarity. So conjugation by f produces an isomorphism

$$(6.1.2) \quad c_f : H \xrightarrow{\cong} fHf^{-1}$$

between two subgroups of \mathcal{I}_n .

Lemma 6.1.11. *Let $X \subset \mathbb{R}^n$ and let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a similarity. Then the symmetry groups of X and of $f(X)$ are conjugate by f , and hence isomorphic:*

$$\mathcal{S}(f(X)) = f\mathcal{S}(X)f^{-1}.$$

Proof. Let $\alpha \in \mathcal{I}_n$. Then

$$\begin{aligned} \alpha f(X) = f(X) &\Leftrightarrow f^{-1}\alpha f(X) = X \\ &\Leftrightarrow f^{-1}\alpha f \in \mathcal{S}(X) \\ &\Leftrightarrow \alpha \in f\mathcal{S}(X)f^{-1}. \end{aligned}$$

Here, the second equivalence uses that $f^{-1}\alpha f$ is an isometry. □

We will see that the analogous result where f is replaced by an affine isomorphism of \mathbb{R}^n is false, as then $f^{-1}\alpha f$ need not be an isometry.

Lemma 6.1.11 is useful for comparing the symmetry groups of different models of the n -cube. The standard n -cube is $I^n = [0, 1]^n$. But it is easier

to calculate the symmetry group of $[-1, 1]^n$, as the latter symmetry group is given by linear isometries. The two models are obviously similar.

Another such example is given by equilateral triangles. Any two equilateral triangles can be shown to be similar. One particular standard model for an equilateral triangle is shown in Section 6.5 to have symmetry group D_6 , the dihedral group of order 6. Therefore every equilateral triangle has symmetry group isomorphic to D_6 , and we can identify the generators in terms of the geometry of the triangle.

Indeed, it can be shown that any two triangles are affinely isomorphic. But not every triangle has symmetry group D_6 . Some have symmetry group D_2 , and some have no nonidentity symmetries.

6.2. Symmetries of polytopes.

6.2.1. Generalities. We have developed the theory of polytopes in Sections 2.8 and 2.9. We defined them to be the convex hulls of finite sets. Here, if $S = \{x_1, \dots, x_k\} \subset \mathbb{R}^n$, the convex hull, $\text{Conv}(S)$ is the set of all convex combinations of the points in S . Here, a convex combination of x_1, \dots, x_k is a sum

$$a_1x_1 + \cdots + a_kx_k$$

with $a_i \geq 0$ for all i and $\sum_{i=1}^k a_i = 1$. $\text{Conv}(S)$ is the smallest convex subset of \mathbb{R}^n containing S . In particular, if $x \in \text{Conv}(S)$, then $\text{Conv}(S) = \text{Conv}(S \cup \{x\})$ so the polytope does not determine the set S . So we will write \mathbf{P} for the polytope and refer to S as a convex generating set for \mathbf{P} .

Isometries of \mathbb{R}^n are affine. By Proposition 2.8.21, if $x = a_1x_1 + \cdots + a_kx_k$, then

$$(6.2.1) \quad \alpha(x) = a_1\alpha(x_1) + \cdots + a_k\alpha(x_k) \quad \text{for } \alpha \in \mathcal{I}_n.$$

In particular, the effect of α on $\mathbf{P} = \text{Conv}(S)$ is determined by its effect on S . Moreover, $\alpha(\mathbf{P}) = \text{Conv}(\alpha(S))$. So if $\alpha(S) = S$, then $\alpha(\mathbf{P}) = \mathbf{P}$. We obtain the following.

Lemma 6.2.1. *For a finite set $S \subset \mathbb{R}^n$ and for $\mathbf{P} = \text{Conv}(S)$, $\mathcal{S}(S)$ is a subgroup of $\mathcal{S}(\mathbf{P})$.*

In fact, the same thing holds for infinite sets S , by the same argument.

However, for some sets S , there are isometries of $\mathcal{S}(\text{Conv}(S))$ that do not preserve S . This can happen when there are elements in S which are not *vertices* of $\text{Conv}(S)$ (see below). We give an example and then discuss the issue in greater detail.

Example 6.2.2. Let $S = \{-1, \frac{3}{4}, 1\} \subset \mathbb{R}$. Then

$$\text{Conv}(S) = \text{Conv}(\{-1, 1\}) = [-1, 1].$$

Let $\alpha \in \mathcal{I}_1$ be multiplication by -1 . Then $\alpha \in \mathcal{S}([-1, 1])$. But $\alpha(S) \neq S$ as $\alpha(\frac{3}{4}) \notin S$.

A key in understanding the symmetries of a polytope is the notion of face. Let $\mathbf{P} = \text{Conv}(S)$. A face of \mathbf{P} is a nonempty subset of the form $F = \mathbf{P} \cap H$, where H is an affine subspace of \mathbb{R}^n and $\mathbf{P} \setminus F$ is convex. It is shown in Proposition 2.9.39 that $F = \text{Conv}(T)$ for some $T \subset S$. Thus, it is a polytope, and has a dimension, given as the dimension of the affine hull $\text{Aff}(T)$ of T , an affine subspace of \mathbb{R}^n . Moreover, we may take the affine subspace H in the definition of face to be $\text{Aff}(T)$.

A vertex is a face of dimension 0. By Corollary 2.9.40, every vertex of $\mathbf{P} = \text{Conv}(S)$ must lie in S . Write $\mathcal{V} = \mathcal{V}(\mathbf{P})$ for the set of vertices of \mathbf{P} . By Proposition 2.9.41, $\mathbf{P} = \text{Conv}(\mathcal{V})$. Thus, \mathcal{V} is the unique smallest convex generating set for \mathbf{P} .

Proposition 6.2.3. *Let \mathbf{P} be a polytope and let $\alpha \in \mathcal{S}(\mathbf{P})$. Then for each face F of \mathbf{P} , $\alpha(F)$ is a face of the same dimension as F .*

Proof. Let $\mathcal{V} = \mathcal{V}(\mathbf{P})$ be the vertices of \mathbf{P} . Let F be a face of \mathbf{P} , and write $F = \text{Conv}(T)$ for $T \subset \mathcal{V}$. By Proposition 2.8.21,

$$\alpha(F) = \text{Conv}(\alpha(T)) = \mathbf{P} \cap \text{Aff}(\alpha(T)),$$

and its complement in \mathbf{P} is convex, as α is one-to-one and preserves convexity. \square

Since a vertex is simply a face of dimension 0, the following is immediate.

Corollary 6.2.4. *Let \mathbf{P} be a polytope and let $\alpha \in \mathcal{S}(\mathbf{P})$. Let $v \in S$ be a vertex of \mathbf{P} . Then $\alpha(v)$ is also a vertex of \mathbf{P} .*

Corollary 6.2.5. *Let $\mathbf{P} \subset \mathbb{R}^n$ be a polytope with vertex set \mathcal{V} . Let $\alpha \in \mathcal{S}(\mathbf{P})$. Then α restricts to a bijection*

$$(6.2.2) \quad \alpha|_{\mathcal{V}} : \mathcal{V} \xrightarrow{\cong} \mathcal{V}.$$

In particular, $\alpha \in \mathcal{S}(\mathcal{V})$, and hence the inclusion $\mathcal{S}(\mathcal{V}) \subset \mathcal{S}(\mathbf{P})$ of Corollary 6.2.1 is the identity:

$$(6.2.3) \quad \mathcal{S}(\mathcal{V}) = \mathcal{S}(\mathbf{P}).$$

Moreover, the passage from α to $\alpha|_{\mathcal{V}}$ induces a group homomorphism from $\mathcal{S}(\mathbf{P})$ to the group of permutations of \mathcal{V} :

$$(6.2.4) \quad \begin{aligned} \rho : \mathcal{S}(\mathbf{P}) &\rightarrow \Sigma(\mathcal{V}) \\ \rho(\alpha) &= \alpha|_{\mathcal{V}}. \end{aligned}$$

This restriction map is injective if $\dim \mathbf{P} = n$.

Proof. α is injective, hence its restriction to \mathcal{V} is also. By Corollary 6.2.4, $\alpha(\mathcal{V}) \subset \mathcal{V}$. Since \mathcal{V} is finite,

$$\alpha|_{\mathcal{V}} : \mathcal{V} \rightarrow \mathcal{V}$$

is bijective. So $\alpha \in \mathcal{S}(\mathcal{V})$, and ρ is well-defined. ρ is a homomorphism because it respects composition.

If $\dim \mathbf{P} = n$, then $\text{Aff}(\mathbf{P})$ is an n -dimensional affine subspace of \mathbb{R}^n , hence $\text{Aff}(\mathbf{P}) = \mathbb{R}^n$. Of course, $\text{Aff}(\mathbf{P}) = \text{Aff}(\mathcal{V})$, and an affine map on $\text{Aff}(\mathcal{V})$ is determined by its effect on \mathcal{V} (Corollary 2.8.23). So ρ is injective. \square

Indeed, using exactly the same argument as that in Proposition 6.2.3, we can compare the faces of congruent, similar, or even affinely isomorphic polytopes.

Proposition 6.2.6. *Let α be an isometry of \mathbb{R}^n (or more generally a similarity or affine automorphism). Let \mathbf{P} be a polytope in \mathbb{R}^n . Then*

$$\alpha(\text{Int}(\mathbf{P})) = \text{Int}(\alpha(\mathbf{P})),$$

and if F is a face of \mathbf{P} , then $\alpha(F)$ is a face of $\alpha(\mathbf{P})$ of the same dimension.

Remark 6.2.7. By (2.8.11), the restriction of α to $\text{Aff}(\mathbf{P})$ is determined by $\rho(\alpha)$. But if $\text{Aff}(\mathbf{P}) \neq \mathbb{R}^n$ (i.e., if $\dim \mathbf{P} < n$), then α will not be determined by its restriction to $\text{Aff}(\mathbf{P})$. For instance, if $\mathbf{P} = \text{Conv}(-e_1, e_1) \subset \mathbb{R}^2$, then the linear isometry of \mathbb{R}^2 induced by $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ (i.e., complex conjugation) restricts to the identity on \mathbb{R} (and hence on \mathbf{P}), but is not the identity isometry. In particular, this transformation is a nontrivial element of $\mathcal{S}(\mathbf{P})$.

Moreover, $\rho : \mathcal{S}(\mathbf{P}) \rightarrow \Sigma(\mathcal{V})$ is rarely onto. Indeed, for $\alpha \in \mathcal{S}(\mathbf{P})$, we must have

$$(6.2.5) \quad d(\alpha(x), \alpha(y)) = d(x, y) \quad \text{for } x, y \in \mathcal{V},$$

and not all pairs of elements of \mathcal{V} will, in general, have the same distance from each other that x has from y . So not every permutation of the vertices can be realized by an isometry. For instance, if \mathbf{P} is the square and if x and y share an edge, then no symmetry of \mathbf{P} can take x and y to points diagonally opposite one another.

An exception is the standard simplex $\Delta^{n-1} \subset \mathbb{R}^n$. Recall that Δ^{n-1} has vertices e_1, \dots, e_n , the standard basis elements of \mathbb{R}^n . Note here that $d(e_i, e_j) = \sqrt{2}$ for all $i \neq j$.

Proposition 6.2.8. *The restriction map*

$$(6.2.6) \quad \rho : \mathcal{S}(\Delta^{n-1}) \xrightarrow{\Sigma} (\{e_1, \dots, e_n\}) \cong \Sigma_n$$

is onto.

Proof. Let $\sigma \in \Sigma_n$. Then σ corresponds to the permutation of $\{e_1, \dots, e_n\}$ taking e_i to $e_{\sigma(i)}$ for all i . But this permutation is induced by the matrix $A_\sigma = [e_{\sigma(1)} | \dots | e_{\sigma(n)}]$ whose i -th column is $e_{\sigma(i)}$ for all i . Since the columns of A_σ form an orthonormal basis of \mathbb{R}^n , A_σ is an orthogonal matrix and hence induces an isometry of \mathbb{R}^n having the desired effect on the canonical basis vectors. Thus σ is in the image of ρ , and ρ is onto. \square

Remark 6.2.9. The map $\rho : \mathcal{S}(\Delta^{n-1}) \xrightarrow{\Sigma} (\{e_1, \dots, e_n\})$ from the last example is not one-to-one for the same reason the analogous map for

$$\text{Conv}(-e_1, e_1) \subset \mathbb{R}^2$$

was not one-to-one. There is a reflection map of \mathbb{R}^n across the hyperplane $\text{Aff}(e_1, \dots, e_n)$, and this provides a nontrivial element of $\mathcal{S}(\Delta^{n-1})$ that restricts to the identity on Δ^{n-1} .

Indeed let $\mathbf{P} \subset \mathbb{R}^n$ be a polytope with vertex set \mathcal{V} . If \mathbf{P} has dimension less than n , then \mathcal{I}_n will always contain nontrivial elements that restrict to the identity on $\text{Aff}(\mathbf{P})$, and hence ρ is not one-to-one.

6.2.2. Centroids. A useful concept in studying polytopes is the centroid.

Definition 6.2.10. Let $\mathbf{P} \subset \mathbb{R}^n$ be a polytope with vertex set \mathcal{V} . The centroid of \mathbf{P} is

$$(6.2.7) \quad c(\mathbf{P}) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} v,$$

the average of the points in \mathcal{V} .⁸

Proposition 6.2.11. Let $\mathbf{P} \subset \mathbb{R}^n$ be a polytope with vertex set

$$\mathcal{V} = \{v_1, \dots, v_k\}.$$

Then every $\alpha \in \mathcal{S}(\mathbf{P})$ preserves the centroid, $c(\mathbf{P})$, of \mathbf{P} . i.e., $\mathcal{S}(\mathbf{P})$ is a subgroup of $\mathcal{S}(c(\mathbf{P}))$. In particular, if $c(\mathbf{P}) = 0$, then $\mathcal{S}(\mathbf{P})$ is a subgroup of the group of linear isometries of \mathbb{R}^n .

Proof. $c(\mathbf{P}) = \frac{1}{k}v_1 + \dots + \frac{1}{k}v_k$, a convex combination of v_1, \dots, v_k . Thus, for $\alpha \in \mathcal{S}(\mathbf{P})$, (2.8.11) gives

$$\alpha(c(\mathbf{P})) = \frac{1}{k}\alpha(v_1) + \dots + \frac{1}{k}\alpha(v_k) = \frac{1}{k} \sum_{i=1}^k \alpha(v_i).$$

Since α restricts to a permutation on \mathcal{V} , $\sum_{i=1}^k \alpha(v_i) = \sum_{i=1}^k v_i$, and hence $\alpha(c(\mathbf{P})) = c(\mathbf{P})$. \square

A similar argument gives the following.

Proposition 6.2.12. Let $\mathbf{P} \subset \mathbb{R}^n$ be a polytope and let α be an isometry of \mathbb{R}^n (or, more generally, a symmetry or an affine isomorphism). Then α carries the centroid of \mathbf{P} to the centroid of $\alpha(\mathbf{P})$.

Corollary 6.2.13. Let \mathbf{P} be a polytope and let F be a face of \mathbf{P} . Let $\alpha \in \mathcal{S}(\mathbf{P})$. Then α carries the centroid of F to the centroid of $\alpha(F)$.

⁸It is essential that we are taking the average of the vertices and not simply the average of some convex generating set. If we were to expand \mathcal{V} to a larger convex generating set, the average of that expanded generating set might be different, and the results to follow would be false.

6.2.3. Symmetries of the n -cube. We can use Proposition 6.2.12 to compute the centroid of the n -cube. The standard n -cube I^n is $[0, 1]^n \subset \mathbb{R}^n$. Its vertices are computed in Example 2.9.50. Here, we shall model the n -cube as $[-1, 1]^n$, which we shall call the standard balanced n -cube. It is related by an obvious similarity to the standard model. In particular, its vertices are

$$(6.2.8) \quad S = \{\epsilon_1 e_1 + \cdots + \epsilon_n e_n : \epsilon_1, \dots, \epsilon_n \in \{\pm 1\}\}.$$

Corollary 6.2.14. *The standard balanced n -cube has centroid 0. Therefore its symmetries are all linear.*

Proof. For each v in the vertex set S of (6.2.8), its negative, $-v$, also lies in S . So the sum of the vertices is 0. \square

Now apply the obvious similarity from the balanced n -cube to the standard one:

Corollary 6.2.15. *The centroid of the standard n -cube is $\frac{1}{2}e_1 + \cdots + \frac{1}{2}e_n$.*

To compute the symmetries of the balanced n -cube, we shall pay attention to their effect on faces.

We now compute the symmetry group of the balanced n -cube. We first display the group itself and then show it gives the desired symmetries.

Definition 6.2.16. The group of $n \times n$ signed permutation matrices, $O(n, \mathbb{Z})$, consists of the matrices of the form $A = [\epsilon_1 e_{\sigma(1)} | \cdots | \epsilon_n e_{\sigma(n)}]$ with $\sigma \in \Sigma_n$ and $\epsilon_i \in \{\pm 1\}$ for all i . These are precisely the invertible matrices taking each canonical basis element to a signed canonical basis element, so they form a subgroup of $GL_n(\mathbb{R})$. Indeed, their columns form an orthonormal basis of \mathbb{R}^n , so $O(n, \mathbb{Z})$ is a subgroup of $O(n)$. For each $\sigma \in \Sigma_n$, there are 2^n ways to sign the permutation matrix A_σ , so $O(n, \mathbb{Z})$ has $2^n \cdot n!$ elements.

Note that for $A \in O(n, \mathbb{Z})$, the transformation T_A preserves the balanced n -cube $[-1, 1]^n$. Thus, we obtain an injective group homomorphism

$$(6.2.9) \quad \begin{aligned} T : O(n, \mathbb{Z}) &\rightarrow \mathcal{S}([-1, 1]^n) \\ A &\rightarrow T_A. \end{aligned}$$

Proposition 6.2.17. *The map T of (6.2.9) is an isomorphism.*

Proof. As shown in Example 2.9.50, the $(n-1)$ -dimensional faces of $[-1, 1]^n$ have the form

$$(6.2.10) \quad \partial_i^\epsilon([-1, 1]^n) = \{a_1 e_1 + \cdots + a_n e_n \in [-1, 1]^n : a_i = \epsilon\}$$

for $i = 1, \dots, n$ and $\epsilon \in \pm 1$. Note that $\partial_i^\epsilon([-1, 1]^n)$ is the image under $\tau_{\epsilon e_i}$ of the standard balanced $(n-1)$ -cube in $\text{span}(\{e_j : j \neq i\})$. By Proposition 6.2.12, the centroid of $\partial_i^\epsilon([-1, 1]^n)$ is ϵe_i . But each $\alpha \in \mathcal{S}([-1, 1]^n)$ takes $(n-1)$ -dimensional faces to $(n-1)$ -dimensional faces, and must take centroids to centroids. So $\alpha(e_i)$ is a signed canonical basis vector for all i . Thus, T is onto. \square

6.2.4. Symmetries of the regular n -gon in \mathbb{R}^2 . We shall also study the symmetries of the standard regular n -gon in \mathbb{R}^2 , $n \geq 3$, which we take to be $P_n = \text{Conv}(v_0, \dots, v_{n-1})$ with $v_j = \begin{bmatrix} \cos \frac{2\pi j}{n} \\ \sin \frac{2\pi j}{n} \end{bmatrix}$ for $j = 0, \dots, n-1$.

We shall verify here that each v_j is a vertex of P_n and that its centroid is the origin. We shall complete the calculation of $\mathcal{S}(P_n)$ in Section 6.5.

For simplicity of notation, let $\rho = \rho_{(0, \frac{2\pi}{n})}$. We shall make use of the fact that

$$(6.2.11) \quad \rho(v_j) = v_{j+1} \quad \text{for } j = 0, \dots, n-2, \quad \text{and} \quad \rho(v_{n-1}) = v_0.$$

Thus, ρ permutes the elements of $\{v_0, \dots, v_{n-1}\}$. The following is now immediate from (2.8.11).

Lemma 6.2.18. $\rho \in \mathcal{S}(P_n)$.

We immediately obtain the following.

Corollary 6.2.19. *The origin is the centroid of P_n . Thus every symmetry of P_n is linear.*

Proof. By Proposition 6.2.11, ρ preserves the centroid of P_n . But the only fixed-point of ρ is the origin. \square

Note that $v_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

Lemma 6.2.20. v_0 is a vertex of P_n .

Proof. Since the cosine function is decreasing on $[0, \pi]$ and $\cos(-t) = \cos t$, the x -coordinate of v_j is in $[-1, \cos \frac{2\pi}{n}]$ for $j = 1, \dots, n-1$. By (2.8.11), the x -coordinate of each element of $\text{Conv}(v_1, \dots, v_{n-1})$ lies in that interval. Thus, v_0 is not in $\text{Conv}(v_1, \dots, v_{n-1})$. So v_0 is a vertex by Corollary 2.9.40. \square

Corollary 6.2.21. v_j is a vertex of P_n for all $j = 0, \dots, n-1$.

Proof. For $j = 1, \dots, n-1$, $v_j = \rho^j(v_0)$. Since $\rho \in \mathcal{S}(P_n)$, so is ρ^j . By Corollary 6.2.4, symmetries of P_n carry vertices to vertices. \square

6.3. Geometry meets number theory: the golden mean. The golden mean is defined to be the number $\Phi = \frac{1+\sqrt{5}}{2}$. It is important in number theory, as it generates what is known as the ring of integers in the number field $\mathbb{Q}(\sqrt{5}) = \{a + b\sqrt{5} : a, b \in \mathbb{Q}\}$. Note that the quadratic formula shows Φ to be a root of the quadratic $x^2 - x - 1$. We obtain the following.

Lemma 6.3.1. *The golden mean satisfies*

$$(6.3.1) \quad \Phi^2 = \Phi + 1.$$

Moreover, the multiplicative inverse of Φ in \mathbb{R} is given by

$$(6.3.2) \quad \frac{1}{\Phi} = \Phi - 1 = \frac{-1 + \sqrt{5}}{2}.$$

Proof. Φ is a root of $x^2 - x - 1$, so $\Phi^2 - \Phi - 1 = 0$, and (6.3.1) follows immediately. Similarly,

$$\Phi(\Phi - 1) = \Phi^2 - \Phi = 1,$$

and we obtain (6.3.2). \square

Notation 6.3.2. We write $\phi = \frac{1}{\Phi} = \Phi - 1 = \frac{-1+\sqrt{5}}{2}$.

Since $2 = \sqrt{4} < \sqrt{5} < 3 = \sqrt{9}$, we obtain:

Lemma 6.3.3. *We have $1.5 < \Phi < 2$, hence $.5 < \phi < .75$.*

This gives rise to the famous geometric observation:

Proposition 6.3.4. *Let R be a rectangle one of whose sides has length 1 and the other has length Φ . Cut it into a square of side 1 and a rectangle S whose sides have lengths 1 and $\Phi - 1$. Then we take the ratio of the smaller side over the larger side in R and S is the same: the ratio is ϕ*

Proof. Since $1 < \Phi < 2$, the smaller side of S has length $\Phi - 1 = \phi$, and the larger side of S has length 1. \square

A second connection with geometry now follows from our work on the regular pentagon P_5 . We make use of the identification of \mathbb{R}^2 with the complex numbers \mathbb{C} . Under this identification, the vertex $v_j = \begin{bmatrix} \cos \frac{2\pi j}{5} \\ \sin \frac{2\pi j}{5} \end{bmatrix}$ corresponds to the complex exponential $e^{\frac{2\pi j}{5} \cdot i} = \cos \frac{2\pi j}{5} + i \sin \frac{2\pi j}{5}$. So, as a complex number, $v_j^5 = 1$ for all j . Indeed, we can say more. A standard notation is to set $\zeta_5 = e^{\frac{2\pi i}{5}}$. We then have

$$(6.3.3) \quad v_j = \zeta_5^j \quad \text{for } j = 0, \dots, 4.$$

Of course $v_0 = e^0 = 1$.

We obtain the following very useful calculations.

Proposition 6.3.5. *We have*

$$(6.3.4) \quad \cos \frac{2\pi}{5} = \frac{\phi}{2},$$

$$(6.3.5) \quad \cos \frac{4\pi}{5} = -\frac{\Phi}{2}.$$

Proof. Write $\xi_5 = \zeta_5 + \zeta_5^4 = v_1 + v_4$. Since $\zeta_5^5 = 1$, $\zeta_5^4 = \zeta_5^{-1}$. Since $\|\zeta_5\| = 1$, $\zeta_5^{-1} = \bar{\zeta}_5$, the complex conjugate of ζ_5 . Thus,

$$(6.3.6) \quad \begin{aligned} \xi_5 &= \zeta_5 + \bar{\zeta}_5 = \left(\cos \frac{2\pi}{5} + i \sin \frac{2\pi}{5} \right) + \left(\cos \frac{2\pi}{5} - i \sin \frac{2\pi}{5} \right) \\ &= 2 \cos \frac{2\pi}{5} \end{aligned}$$

So (6.3.4) is equivalent to showing that $\xi_5 = \phi$.

Let $\omega = \zeta_5^2 + \zeta_5^3 = v_2 + v_3$. Note that ζ_5^3 is the inverse of ζ_5^2 . So an argument similar to that for (6.3.6) shows that

$$(6.3.7) \quad \omega = 2 \cos \frac{4\pi}{5}.$$

So (6.3.5) is equivalent to showing that $\omega = -\Phi$. Now

$$(6.3.8) \quad \xi_5 + \omega + 1 = v_0 + \cdots + v_5 = 0,$$

as the centroid of P_5 is 0, so

$$(6.3.9) \quad \omega = -\xi_5 - 1.$$

Now

$$(6.3.10) \quad \xi_5^2 = (\zeta_5 + \zeta_5^{-1})^2 = \zeta_5^2 + 2\zeta_5\zeta_5^{-1} + \zeta_5^{-2} = \omega + 2 = -\xi_5 + 1,$$

so ξ_5 is a root of $x^2 + x - 1$. Since $\xi_5 = 2 \cos \frac{2\pi}{5} > 0$, the quadratic formula gives

$$(6.3.11) \quad \xi_5 = \frac{-1 + \sqrt{5}}{2} = \phi,$$

as the other root is negative. We obtain (6.3.4). But now

$$(6.3.12) \quad \omega = -\xi_5 - 1 = -\frac{1 + \sqrt{5}}{2} = -\Phi,$$

and (6.3.5) follows. \square

6.4. Symmetries of points and lines in \mathbb{R}^2 . Example 6.1.4 shows that the symmetry group of the origin in \mathbb{R}^n is the group of linear isometries of \mathbb{R}^n . When $n = 2$ we can use our calculation of $O(2)$ to say more:

Example 6.4.1. In \mathbb{R}^2 , we have

$$(6.4.1) \quad \mathcal{S}(\{0\}) = \mathcal{LI}_2 = \{\rho_{(0,\theta)} : \theta \in \mathbb{R}\} \cup \{\sigma_{\ell_\theta} : \theta \in \mathbb{R}\}.$$

Thus, for $n = 2$, $\mathcal{T}(\{0\}) = \{\text{id}\}$ and $\mathcal{O}(\{0\}) = \{\rho_{(0,\theta)} : \theta \in \mathbb{R}\}$, which we may identify with $SO(2)$.

We can now use Corollary 6.1.9 to find $\mathcal{S}(\{x\})$ for $x \neq 0$.

Corollary 6.4.2. *Let $x \in \mathbb{R}^2$. Then*

$$\mathcal{S}(\{x\}) = \{\rho_{(x,\theta)} : \theta \in \mathbb{R}\} \cup \{\sigma_\ell : x \in \ell\}.$$

Proof. $\rho_{(x,\theta)} = \tau_x \rho_{(0,\theta)} \tau_{-x}$ (by definition), and if $x \in \ell$, then $\sigma_\ell = \tau_x \sigma_{\ell_\theta} \tau_{-x}$ for $\ell_\theta = \tau_{-x}(\ell)$ (Lemma 5.1.14). \square

We next calculate the symmetries of a line.

Proposition 6.4.3. *Let $\ell = y + \text{span}(v)$ be a line in \mathbb{R}^2 with v a unit vector. Then*

$$\begin{aligned}\mathcal{T}(\ell) &= \{\tau_{sv} : s \in \mathbb{R}\}, \\ \mathcal{O}(\ell) &= \mathcal{T}(\ell) \cup \{\rho_{(x,\pi)} : x \in \ell\}.\end{aligned}$$

Moreover, the reflections in $\mathcal{S}(\ell)$ are $\{\sigma_\ell\} \cup \{\sigma_m : m \perp \ell\}$ and the glide reflections in $\mathcal{S}(\ell)$ are $\{\sigma_\ell \tau_x : x \parallel \ell\}$.

Proof. The calculation of $\mathcal{T}(\ell)$ is Proposition 2.1.14. The calculation of the glide reflections in $\mathcal{S}(\ell)$ is Lemma 5.5.6.

Let's now consider the rotations in $\mathcal{S}(\ell)$. First note that if $x \in \ell$, then $\ell = x + \text{span}(v)$ and $\rho_{(x,\pi)}(x + sv) = x - sv$, so $\rho_{(x,\pi)} \in \mathcal{S}(\ell)$. Thus, for rotations, it suffices to show:

- (1) If $x \notin \ell$ and θ is not a multiple of 2π , then $\rho_{(x,\theta)} \notin \mathcal{S}(\ell)$.
- (2) If $x \in \ell$ and θ is not a multiple of π , then $\rho_{(x,\theta)} \notin \mathcal{S}(\ell)$.

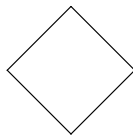
To prove (1), we may as well translate the problem and assume $x = 0$. This simplifies the calculations. We drop a perpendicular from $x = 0$ to ℓ , and that perpendicular is easily seen to be ℓ_ϕ , where $\begin{bmatrix} \cos \phi \\ \sin \phi \end{bmatrix} = v^\perp$. Let $z = \ell_\phi \cap \ell$. Then the Pythagorean theorem shows z to be the closest point on ℓ to the origin. Since rotations are isometries, $\rho_{(0,\theta)}(z)$ is the closest point on $\rho_{(0,\theta)}(\ell)$ to the origin. Since $x \notin \ell$, $z \neq 0$, and since θ is not a multiple of 2π , $\rho_{(0,\theta)}(z) \neq z$. Thus $\rho_{(0,\theta)}(\ell) \neq \ell$.

To prove (2), we again assume $x = 0$. We then have $\ell = \ell_\phi$ where $\begin{bmatrix} \cos \phi \\ \sin \phi \end{bmatrix} = v$. But $\rho_{(0,\theta)}(\ell_\phi) = \ell_{\phi+\theta}$, and this is ℓ_ϕ if and only if θ is a multiple of π .

Finally, we calculate the reflections in $\mathcal{S}(\ell)$. If $m \perp \ell$ and if $x = \ell \cap m$, then $\ell = x + \text{span}(v)$ and $m = x + \text{span}(v^\perp)$. An easy calculation shows $\sigma_m(\ell) = x + \text{span}(-v) = \ell$. Of course, σ_ℓ is the identity on ℓ , so $\sigma_\ell \in \mathcal{S}(\ell)$.

If $m \parallel \ell$ and $m \neq \ell$, then $\sigma_m(\ell) \neq \ell$ by an argument similar to that of Lemma 5.5.6. Thus, it suffices to consider the case where m is neither parallel nor perpendicular to ℓ . Let $x = \ell \cap m$. As above we may translate the problem and assume that $x = 0$. Thus, we may assume $\ell = \ell_\phi$ and $m = \ell_{\phi+\psi}$ where ψ is not a multiple of $\frac{\pi}{2}$. By Lemma 5.3.7, $\sigma_m(\ell) = \ell_{\phi+2\psi}$, and that is not equal to ℓ_ϕ , as 2ψ is not a multiple of π . \square

6.5. Dihedral groups. We study the symmetries of the regular n -gon for $n \geq 3$. Let $v_i = \begin{bmatrix} \cos \frac{2\pi i}{n} \\ \sin \frac{2\pi i}{n} \end{bmatrix}$ and let P_n be the polygon whose vertices are v_0, \dots, v_{n-1} . So P_4 is the following:



Formally, P_n is the convex hull, $\text{Conv}(v_0, \dots, v_{n-1})$, and we have demonstrated in Corollary 6.2.21 that each v_i satisfies the formal definition of a vertex for a polytope. Since $P_n = \text{Conv}(v_0, \dots, v_{n-1})$, these are the only vertices of P_n .

By Corollary 6.2.5, every symmetry of P_n must permute the vertex set $\{v_0, \dots, v_{n-1}\}$, and since the vertex set contains 3 noncollinear points, the symmetries of P_n are determined by their effect on the vertices.

Proposition 6.5.1. *The symmetry group of P_n is finite, with $2n$ elements:*

$$\mathcal{S}(P_n) = \left\{ \rho_{\left(0, \frac{2\pi}{n}\right)}^i : 0 \leq i < n \right\} \cup \left\{ \rho_{\left(0, \frac{2\pi}{n}\right)}^i \sigma_{\ell_0} : 0 \leq i < n \right\},$$

with ℓ_0 the x -axis as usual. Thus,

$$\mathcal{O}(P_n) = \left\{ \rho_{\left(0, \frac{2\pi}{n}\right)}^i : 0 \leq i < n \right\}.$$

In particular, these isometries are all linear.⁹ Finally, note that

$$(6.5.1) \quad \rho_{\left(0, \frac{2\pi}{n}\right)}^i \sigma_{\ell_0} = \sigma_{\ell_{\frac{\pi i}{n}}},$$

where $\ell_{\frac{\pi i}{n}} = \text{span} \left(\begin{pmatrix} \cos \frac{\pi i}{n} \\ \sin \frac{\pi i}{n} \end{pmatrix} \right)$. This line through the origin contains either a vertex of P_n or the midpoint of an edge, or both. Every line through the origin and one of these vertices or midpoints is included in this list.

Note that there are exactly n rotations (including the identity) and n reflections in $\mathcal{S}(P_n)$.

Proof. As discussed above, a symmetry of P_n is determined by its effect on the vertices. But it cannot permute the vertices arbitrarily, as it must preserve distance.

The cosine law tells us the distance between two points, v and w , on the unit circle if we know the unsigned angle between the rays $\overrightarrow{0v}$ and $\overrightarrow{0w}$: if that angle is θ , then the distance, d , is determined by

$$(6.5.2) \quad d^2 = \|v\|^2 + \|w\|^2 - 2\|v\|\|w\|\cos\theta = 2(1 - \cos\theta).$$

In particular, this distance increases with θ , so the two closest vertices to v_i are v_{i-1} and v_{i+1} , where we identify v_0 as v_n to make sense of this. We call these the adjacent vertices to v_i . Any $\alpha \in \mathcal{S}(P_n)$ must take adjacent vertices to adjacent vertices.

Thus, if we know $\alpha(v_0)$ there are exactly two choices for $\alpha(v_1)$. And since $\alpha(v_0)$ occupies one of the vertices adjacent to $\alpha(v_1)$, v_2 must go to the other vertex adjacent to $\alpha(v_1)$. Continuing in this manner, we see that the targets of the rest of the vertices are determined, once we know $\alpha(v_0)$ and $\alpha(v_1)$.

Now consider the displayed isometries. $\rho_{\left(0, \frac{2\pi}{n}\right)}^i = \rho_{\left(0, \frac{2\pi i}{n}\right)}$ takes v_0 to v_i and takes v_1 to the next vertex in the counterclockwise direction, while

⁹We already knew this from Corollary 6.2.19, but the argument here is more elementary.

$\rho_{(0, \frac{2\pi}{n})}^i \sigma_{\ell_0}$ takes v_0 to v_i and takes v_1 to the adjacent vertex in the clockwise direction. Thus, no other isometries of P_n are possible.

Now, (6.5.1) is just Lemma 5.3.9. When i is even, this line goes through $v_{\frac{i}{2}}$. When $i = 2k + 1$, the line goes through the the midpoint of $\overline{v_k v_{k+1}}$. The rest follows since $\ell_\theta = \ell_{\theta+\pi}$ for all θ . \square

Definition 6.5.2. The group $\mathcal{S}(P_n)$ displayed above is the standard model for the dihedral group, D_{2n} of order $2n$. A standard notation would be to set $\rho_{(0, \frac{2\pi}{n})} = b$, or b_n if n varies, and set $\sigma_{\ell_0} = a$. So

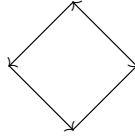
$$D_{2n} = \{b^i : 0 \leq i < n\} \cup \{b^i a : 0 \leq i < n\}.$$

We have $b^n = \rho_{(0, 2\pi)} = \text{id}$ and this is the smallest power of b that gives the identity. As discussed in Chapter 3, this says the order of b is n . We also have $a^2 = \sigma_{\ell_0}^2 = \text{id}$, so a has order 2. Note that $ab^i a^{-1} = b^{-i}$ by Proposition 5.3.10. Thus, $ab^i = b^{-i}a$, allowing us to compute the product of any two of the listed elements.

The group $\mathcal{O}(D_{2n}) = \{b^i : 0 \leq i < n\}$ is the cyclic group of order n , written C_n . In particular, $C_n = \langle b \rangle$.

We can easily find a subset X with symmetry group C_n .

Example 6.5.3. Let Q_n be obtained from P_n by replacing the edge $\overline{v_i v_{i+1}}$ by an arrow pointing from v_i to v_{i+1} . Q_4 :



Obviously, any symmetry of Q_n must preserve P_n , so $\mathcal{S}(Q_n)$ is a subgroup of $\mathcal{S}(P_n)$. Since all the arrows point counterclockwise, there are no reflections in $\mathcal{S}(Q_n)$, so $\mathcal{S}(Q_n) \subset \mathcal{O}(P_n)$. Moreover, each b^i does preserve Q_n , so $\mathcal{S}(Q_n) = \mathcal{O}(Q_n) = C_n$.

Let $\alpha : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be an isometry. By Lemma 6.1.8, $\mathcal{S}(\alpha(P_n)) = \alpha D_{2n} \alpha^{-1}$, a conjugate subgroup to $\mathcal{S}(P_n)$. In general, if H is a subgroup of the group G and if $g \in G$, there is an isomorphism

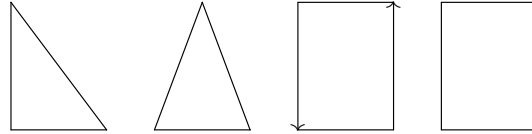
$$\begin{aligned} c_g : H &\rightarrow gHg^{-1} \\ h &\mapsto ghg^{-1}. \end{aligned}$$

This is easily seen to be a bijective group homomorphism. Its inverse is given by conjugating by g^{-1} . Thus there are lots of subgroups of \mathcal{I}_2 isomorphic to D_{2n} . In fact, we will show that every finite subgroup of \mathcal{I}_2 is conjugate either to the standard copy of D_{2n} or the standard copy of C_n .

Before we continue, there are some additional cases to consider for small values of n .

Definition 6.5.4. We set $C_1 = \text{id}$, $D_2 = \{\sigma_{\ell_0}, \text{id}\}$, $C_2 = \{\rho_{(0,\pi)}, \text{id}\}$, and $D_4 = \{\rho_{(0,\pi)}, \sigma_{\ell_0}, \rho_{(0,\pi)}\sigma_{\ell_0} = \sigma_{\ell_{\frac{\pi}{2}}}, \text{id}\}$. Note that in every case, the orientation-preserving subgroup of D_{2n} is C_n and that C_n is cyclic of order n .

As the reader may verify, the following figures have symmetry groups conjugate to C_1 , D_2 , C_2 and D_4 , respectively:



Note that C_2 and D_2 are isomorphic as groups, but are not conjugate in \mathcal{I}_2 as $\rho_{(0,\pi)}$ is orientation-preserving and σ_{ℓ_0} is orientation-reversing.

6.6. Index 2 subgroups. We show how to derive the structure of $\mathcal{S}(X)$ from the structure of $\mathcal{O}(X)$. The techniques used apply to much more general situations, so we state them in that context.

Definition 6.6.1. Let H be a subgroup of G and let $x \in G$. The right coset, Hx , of H by x is

$$Hx = \{hx : h \in H\}.$$

Note that $x = ex \in Hx$.

The right cosets partition G into disjoint sets:

Lemma 6.6.2. *Let H be a subgroup of G then the following statements are equivalent:*

- (1) $Hx \cap Hy \neq \emptyset$.
- (2) $Hx = Hy$.
- (3) $xy^{-1} \in H$.

Proof. (2) \Rightarrow (1) is immediate. To see that (1) \Rightarrow (3), let $z \in Hx \cap Hy$. Then $z = hx = ky$ with $h, k \in H$. Then $xy^{-1} = h^{-1}k \in H$.

Finally, suppose $xy^{-1} = h \in H$. Then $x = hy$, so $kx = (kh)y \in Hy$ for all $k \in H$, so $Hx \subset Hy$. But $y = h^{-1}x$, so $ky = (kh^{-1})x \in Hx$ for all $k \in H$, so $Hy \subset Hx$. Thus $Hx = Hy$. We've shown that (3) \Rightarrow (2). \square

$He = \{he : h \in H\}$ is precisely H , and (3) shows that $Hx = He$ if and only if $x = xe^{-1} \in H$.

Definition 6.6.3. We say H has finite index in G if the number of distinct right cosets of H in G is finite. We then write $[G : H]$ for the number of these right cosets, and call it the index of H in G . If there are infinitely many right cosets we write $[G : H] = \infty$.

We obtain the following as a bonus.

Theorem 6.6.4 (Lagrange's theorem). *Let G be a finite group and let H be a subgroup. Then $|G| = [G : H] \cdot |H|$, so the order of H divides the order of G .*

Proof. Each element of G lies in exactly one right coset of H . There are $[G : H]$ such cosets. For a given coset Hx there is a function

$$\begin{aligned} H &\rightarrow Hx \\ h &\mapsto hx. \end{aligned}$$

This is a bijection as its inverse function is given by multiplying on the right by x^{-1} . In particular, each right coset has $|H|$ elements in it, and the result follows. \square

Corollary 6.6.5. *Let G be a finite group and let $g \in G$. Then $|g|$ divides $|G|$.*

Proof. $|g| = |\langle g \rangle|$. \square

Corollary 6.6.6. *Let G be a group whose order is a prime number, p . Then G is cyclic. Indeed, any nonidentity element of G is a generator.*

Proof. Let $e \neq g \in G$. Then $|\langle g \rangle|$ divides p and is greater than one. So $|\langle g \rangle| = |G|$, hence $\langle g \rangle = G$. \square

Our focus in this section is on index 2 subgroups. These arise in many important contexts. The following helps us recognize index 2 subgroups even when the groups are infinite. Recall that if $Y \subset X$, the set-theoretic difference $X \setminus Y$ is $\{x \in X : x \notin Y\}$.

Proposition 6.6.7. *Let H be a subgroup of G with $H \neq G$. Then the following conditions are equivalent.*

- (1) H has index 2 in G (i.e., $[G : H] = 2$).
- (2) For all $x \in G \setminus H$, $Hx = G \setminus H$.
- (3) For all $x, y \in G \setminus H$, $xy \in H$.

Proof. (1) \Rightarrow (2): Here, there are exactly 2 right cosets of H in G . One is $He = H$. Since the right cosets partition G , this says every element not in the coset H must be in the other coset. In particular, if $x \notin H$, then Hx is the other coset, which must consist of all elements of $G \setminus H$.

(2) \Rightarrow (3): Let $x, y \in G \setminus H$. Since H is a subgroup, $y^{-1} \in G \setminus H$. So

$$G \setminus H = Hx = Hy^{-1}$$

by (2). Thus, $xy \in H$ by Lemma 6.6.2(3).

(3) \Rightarrow (1): $H = He$ is one coset of H in G . If $x \notin H$, then $Hx \neq He$, as $xe^{-1} \notin H$. So $H \neq G$ implies there are at least two cosets. It suffices to show that if $x, y \in G \setminus H$, then $Hx = Hy$, i.e., $xy^{-1} \in H$. But this last follows from (3), as $y^{-1} \notin H$. \square

Note that (2) and (3) hold vacuously if $H = G$, so the requirement that $H \neq G$ is necessary for the equivalence above. (2) or (3) alone is equivalent to saying that $[G : H] = 1$ or 2.

We can now give some examples of index 2 subgroups, using (3), above.

Examples 6.6.8.

- (1) The orientation-preserving isometries \mathcal{O}_2 have index 2 in \mathcal{I}_2 as the product of any two orientation-reversing isometries is orientation-preserving.
- (2) The even integers $\langle 2 \rangle$ have index 2 in \mathbb{Z} as the sum of any two odd integers is even.
- (3) The special orthogonal group $\text{SO}(n)$ has index 2 in $\text{O}(n)$, as if $A, B \in \text{O}(n) \setminus \text{SO}(n)$, then $\det(AB) = \det A \det B = (-1)^2 = 1$.
- (4) The cyclic group C_n has index 2 in the dihedral group D_{2n} : every element of $D_{2n} \setminus C_n$ lies in the right coset of σ_{ℓ_0} .

There is a nice group-theoretic characterization of index 2 subgroups:

Corollary 6.6.9. *Index 2 subgroups are always normal. In fact, a subgroup $H \subset G$ has index 2 if and only if H is the kernel of a surjective group homomorphism $f : G \rightarrow \{\pm 1\}$.*

Proof. Let $[G : H] = 2$. Define $f : G \rightarrow \{\pm 1\}$ by

$$f(x) = \begin{cases} 1 & \text{if } x \in H \\ -1 & \text{otherwise.} \end{cases}$$

If exactly one of x and y is in H , then $xy \notin H$, so $-1 = f(xy) = f(x)f(y)$. Otherwise, $xy \in H$, and $1 = f(xy) = f(x)f(y)$. Thus, f is a homomorphism. $\ker f = f^{-1}(1) = H$.

Conversely, suppose $f : G \rightarrow \{\pm 1\}$ is a homomorphism with $\ker f = H$. Then if $x, y \in G \setminus H$, $f(xy) = f(x)f(y) = (-1)^2 = 1$, so $xy \in \ker f = H$. \square

By Lemma 4.1.20, the alternating group A_n is the kernel of the sign homomorphism $\text{sgn} : \Sigma_n \rightarrow \{\pm 1\}$. We obtain:

Corollary 6.6.10. *The alternating group A_n has index 2 in Σ_n .*

The following is a key in understanding symmetry groups.

Proposition 6.6.11. *Let H be an index 2 subgroup of G and let K be an arbitrary subgroup of G . If K is not contained in H , then $H \cap K$ has index 2 in K .*

Proof. Let $x, y \in K \setminus K \cap H$. Then $x, y \in G \setminus H$, so $xy \in H$. But $x, y \in K$, and K is a subgroup, so $xy \in K$. Thus, $xy \in K \cap H$, and the result follows. \square

We obtain the following:

Corollary 6.6.12. *Let H be a subgroup of \mathcal{I}_2 with $\mathcal{O}(H) \neq H$. Then $\mathcal{O}(H)$ has index 2 in H . Thus, if $\beta \in H \setminus \mathcal{O}(H)$, then $H \setminus \mathcal{O}(H) = \mathcal{O}(H)\beta$.*

Another useful consequence of Proposition 6.6.11 requires some group theory.

Corollary 6.6.13. *Let $n \geq 5$. Then A_n is the only index 2 subgroup of Σ_n .*

Proof. We argue by contradiction. Suppose H is an index 2 subgroup of Σ_n with $H \neq A_n$. Then $K \cap A_n$ has index 2 in A_n , and hence $K \cap A_n$ is a nontrivial proper normal subgroup of A_n . But for $n \geq 5$, A_n is what's known as a simple group: a group with no normal subgroups other than the trivial group and itself (see, e.g., [17, Theorem 4.2.7]).¹⁰ We obtain the desired contradiction. \square

Remark 6.6.14. By studying the groups in question, one can show that A_n is also the only index 2 subgroup of Σ_n for $n < 5$.

6.7. Left cosets; orbit counting; the first Noether theorem. This material is not needed in this chapter, but it is important. We include it here because the idea of cosets has been introduced.

Let H be a subgroup of G . A left coset of H in G is a subset of the form

$$(6.7.1) \quad xH = \{xh : h \in H\}$$

for $x \in G$. Of course, $x = xe \in xH$. The proof of the following is analogous to that of Lemma 6.6.2.

Lemma 6.7.1. *Let H be a subgroup of G then the following statements are equivalent:*

- (1) $xH \cap yH \neq \emptyset$.
- (2) $xH = yH$.
- (3) $y^{-1}x \in H$.

Definition 6.7.2. We write G/H for the set of left cosets of H in G :

$$G/H = \{xH : x \in G\}.$$

Thus, G/H is a collection of subsets of G . Note that since each $x \in G$ lies in the left coset xH and since any two left cosets are either disjoint or identical, every element of G lies in exactly one left coset of H in G . G/H is sometimes called the homogeneous space of G with respect to H .

Similarly, we write $H \backslash G$ for the set of right cosets of H in G . If H has finite index in G , then by definition, $[G : H] = |H \backslash G|$.

Lemma 6.7.3. *There is a bijection $\chi : G/H \rightarrow H \backslash G$ given by*

$$\chi(xH) = Hx^{-1}.$$

¹⁰The simplicity of A_n for $n \geq 5$ is also an essential ingredient in Galois' famous proof that there is no formula involving iterated roots for solving polynomials of degree $n \geq 5$.

Proof. We have that $xH = yH$ if and only if $y^{-1}x \in H$. But $y^{-1}x = y^{-1}(x^{-1})^{-1}$, so this is equivalent to saying $Hy^{-1} = Hx^{-1}$. \square

Definition 6.7.4. Let H be a subgroup of G . Then the canonical map $\pi : G \rightarrow G/H$ is defined by $\pi(x) = xH$ for all $x \in G$.

There is an important connection between G -sets and homogeneous spaces. Recall that if X is a G -set (i.e., G acts on X) and if $x \in X$, then the isotropy subgroup G_x is the set of elements of G that fix x , while the orbit Gx (or $G \cdot x$) is the set of images of x under the action of G :

$$G_x = \{g \in G : gx = x\},$$

$$Gx = \{gx : g \in G\}.$$

Two orbits that intersect nontrivially must be equal, so every element of X lies in exactly one orbit.

Lemma 6.7.5. Let H be a subgroup of G . Then G/H is a G -set via

$$g \cdot xH = (gx)H$$

for $g \in G$ and $xH \in G/H$. The isotropy subgroup of eH under this action is H . The canonical map $\pi : G \rightarrow G/H$ is a G -map, where the action of G on G is given by left multiplication.

Proof. Setting $g \cdot xH = (gx)H$ is well-defined as if $xH = yH$, then $x = yh$ for some $h \in H$, hence $gx = gyh$. This determines an action of G on G/H by associativity. The isotropy subgroup of eH is $\{g \in G : gH = H\}$. But this is precisely H .

Regarding π , we have $\pi(gx) = g\pi(x)$. \square

Note that G/H consists of a single orbit, as $G/H = G \cdot eH$.

Definition 6.7.6. The action of G on a set X is transitive if X consists of a single G -orbit.

In fact, every transitive G -set is G -isomorphic to a homogeneous space:

Proposition 6.7.7. Let X be a G -set and let $x \in X$. Then there is a G -map $f_x : G/H \rightarrow X$ with $f_x(eH) = x$ if and only if H is contained in the isotropy subgroup G_x . There is a unique such f_x in this case: $f_x(gH) = gx$ for all $g \in G$. The image of f_x is the orbit Gx , and f_x is one-to-one if and only if $H = G_x$. Thus, taking $H = G_x$, we obtain a G -isomorphism

$$(6.7.2) \quad f_x : G/G_x \xrightarrow{\cong} Gx.$$

Proof. If there is a G -map $f_x : G/H \rightarrow X$ with $f_x(eH) = x$, then the composite $f = f_x \circ \pi : G \rightarrow X$ is a G -map, so if $g \in G$, $f(g) = gx$. But $f(g) = f_x(gH)$, so $f_x(gH) = gx$ for all $g \in G$. But if $h \in H$, $hH = eH$, so we must have $hx = x$. Thus $H \subset G_x$.

Conversely, if $H \subset G_x$ and if $g_1H = g_2H$, then $g_2^{-1}g_1 \in H \subset G_x$. So $g_2^{-1}g_1x = x$, and hence $g_1x = g_2x$. So setting $f_x(gH) = gx$ gives a well-defined function $f_x : G/H \rightarrow X$. It is obviously a G -map, and its image is Gx .

Now $f_x(g_1H) = f_x(g_2H)$ if and only if $g_1x = g_2x$, and this in turn holds if and only if $g_2^{-1}g_1 \in G_x$. So f_x is one-to-one precisely when $H = G_x$. \square

Corollary 6.7.8. *A G -set is transitive if and only if it is G -isomorphic to a homogeneous space.*

The following is immediate from Lagrange's theorem (Theorem 6.6.4).

Corollary 6.7.9. *Let X be a G -set and $x \in X$. Then the orbit Gx is finite if and only if G_x has finite index in G . In this case $|Gx| = [G : G_x]$. Thus, if G itself is finite, $|G| = |G_x| \cdot |Gx|$ for any $x \in X$.*

The left and right cosets of H are generally different subsets of G . They coincide precisely when $H \triangleleft G$.

Proposition 6.7.10. *Let H be a subgroup of G . Then the following conditions are equivalent.*

- (1) $H \triangleleft G$.
- (2) $xH = Hx$ for all $x \in G$.
- (3) There is a well-defined operation on G/H given by

$$(xH)(yH) = xyH$$

for all $x, y \in G$.

- (4) There is a group structure on G/H making $\pi : G \rightarrow G/H$ a homomorphism.

Proof. (1) \Leftrightarrow (2) Multiplying by x on the right shows $xHx^{-1} = H$ if and only if $xH = Hx$.

(1) \Leftrightarrow (3) Assuming (1), it suffices to show that $(xy)^{-1}xhyk \in H$ for all $h, k \in H$. By (2), $hy = yh'$ for some $h' \in H$, and the result follows.

(3) \Rightarrow (4) The operation specified in (3) inherits associativity from the group operation in G . It has eH as an identity element, and $x^{-1}H$ is an inverse for xH .

(4) \Rightarrow (1) $H = \ker \pi$. Kernels are always normal. \square

Remark 6.7.11. Note that (4) forces the operation on G/H to be the one specified in (3). When $H \triangleleft G$, when we write G/H , we shall always intend it to have this group structure.

Example 6.7.12. Consider the case $G = \mathbb{Z}$ and $H = \langle n \rangle$. Here, the group operation is additive, so the cosets are written additively: $a + \langle n \rangle$ is the coset containing a . Since \mathbb{Z} is abelian, there is no distinction between left and right cosets. It is customary to write \mathbb{Z}_n for $\mathbb{Z}/\langle n \rangle$. It is the standard

cyclic group of order n . We write \bar{a} for the coset $a + \langle n \rangle$. In this notation, we have

$$\begin{aligned}\bar{a} + \bar{b} &= \overline{a + b}, \\ \bar{a} \cdot \bar{b} &= \overline{ab}.\end{aligned}$$

For $a > 0$,

$$\bar{a} = \underbrace{\bar{1} + \cdots + \bar{1}}_{a \text{ times}} = a \cdot \bar{1},$$

so $\mathbb{Z}_n = \langle \bar{1} \rangle$, and $\bar{0} = \bar{n} = n \cdot \bar{1}$, while, if $0 < a < n$, then $a \notin \langle n \rangle$, so $a \cdot \bar{1} \neq \bar{0}$. Thus \mathbb{Z}_n is indeed cyclic of order n .

Theorem 6.7.13 (First Noether isomorphism theorem). *Let $H \triangleleft G$ and let $f : G \rightarrow K$ be a group homomorphism. Then f factors through a homomorphism $\bar{f} : G/H \rightarrow K$ making the following diagram commute if and only if $H \subset \ker f$.*

$$(6.7.3) \quad \begin{array}{ccc} G & \xrightarrow{f} & K \\ & \searrow \pi & \nearrow \bar{f} \\ & G/H & \end{array}$$

Moreover, \bar{f} is an injective if and only if $H = \ker f$, and is surjective if and only if f is surjective.

Proof. If the diagram commutes, then

$$H = \ker \pi \subset \ker(\bar{f} \circ \pi) = \ker f.$$

Conversely, if $H \subset \ker f$, then $f(xh) = f(x)$ for all $x \in G$ and $h \in H$, so there is a function $\bar{f} : G/H \rightarrow K$ making the diagram commute. It is a group homomorphism by (3), above.

By construction, the image of \bar{f} is the image of f , hence the surjectivity statement. If \bar{f} is injective, then $\ker f = \ker(\bar{f} \circ \pi) = \ker \pi = H$. Conversely if $\ker f = H$ and if $xH \in \ker \bar{f}$, then $e = \bar{f}(xH) = f(x)$, and hence $x \in H$. But then $xH = eH$, and $\ker \bar{f}$ is trivial, making \bar{f} injective. \square

As a first application, recall from Proposition 3.4.16 that for any group G and any $g \in G$, there is a unique homomorphism $f_g : \mathbb{Z} \rightarrow G$ with $f_g(1) = g$. Explicitly, $f_g(m) = g^m$ for all $m \in \mathbb{Z}$, so that the image of f_g is $\langle g \rangle$ and the kernel of f_g is

$$\{m \in \mathbb{Z} : g^m = e\},$$

the set of all exponents of g . In particular, if $|g|$ is finite, $\ker f_g = \langle |g| \rangle$, and if $|g|$ is infinite, $\ker f_g = 0$. We ask now when f_g factors as

$$\begin{array}{ccc} \mathbb{Z} & \xrightarrow{f_g} & G \\ & \searrow \pi & \nearrow \bar{f}_g \\ & \mathbb{Z}_n & \end{array}$$

Note that since $\pi(1) = \bar{1}$ such a factorization exists if and only if there is a homomorphism $\bar{f}_g : \mathbb{Z}_n \rightarrow G$ with $\bar{f}_g(\bar{1}) = g$.

Corollary 6.7.14. *There is a homomorphism $\bar{f}_g : \mathbb{Z}_n \rightarrow G$ with $\bar{f}_g(\bar{1}) = g$ if and only if n is an exponent for g . Such a homomorphism is unique. Its image is $\langle g \rangle$. It is injective if and only if $|g| = n$. In particular, every cyclic group $\langle g \rangle$ of order n is isomorphic to \mathbb{Z}_n by the isomorphism $\bar{f}_g : \mathbb{Z}_n \rightarrow \langle g \rangle$ given by $\bar{f}_g(\bar{a}) = g^a$ for all $a \in \mathbb{Z}$.*

Corollary 6.6.6 now gives:

Corollary 6.7.15. *Any group of prime order p is isomorphic to \mathbb{Z}_p .*

6.8. Leonardo's theorem. The finite subgroups of \mathcal{I}_2 are known as rosette groups, as they occur as groups of symmetries of finite polygons. The following theorem of Leonardo da Vinci will verify that fact, once we know that C_n is the symmetry group of a polygon.

Theorem 6.8.1 (Leonardo's theorem). *Every finite subgroup of \mathcal{I}_2 is conjugate to either the standard copy of D_{2n} or the standard copy of C_n .*

The proof will take up this whole section. Let $H \subset \mathcal{I}_2$ be finite. First note that H contains no nonidentity translations, as τ_x has infinite order if $x \neq 0$, so $\langle \tau_x \rangle$ is infinite.

Next let $x \parallel \ell$, and let $\gamma = \tau_x \sigma_\ell$. Then $\gamma^2 = \tau_{2x}$. If $\gamma^k = \text{id}$, then $\gamma^{2k} = \tau_{2kx} = \text{id}$ as well. This forces $k = 0$, so glide reflections have infinite order as well, and hence cannot lie in H .

Thus only rotations and reflections can lie in H . We wish next to show that all rotations in H must have a common center. A key idea is the notion of commutators.

Definition 6.8.2. Let G be a group and let $x, y \in G$. The commutator of x and y is

$$[x, y] = xyx^{-1}y^{-1}.$$

These measure the deviation from commutativity.

Lemma 6.8.3. *The elements $x, y \in G$ commute if and only if $[x, y] = e$.*

Proof.

$$xyx^{-1}y^{-1} = e \quad \Leftrightarrow \quad xyx^{-1}y^{-1}y = y$$

$$\Leftrightarrow xyx^{-1}x = yx$$

$$\Leftrightarrow xy = yx. \quad \square$$

Corollary 6.8.4. *Let H be a finite subgroup of \mathcal{I}_2 and let $\rho_{(x,\theta)}, \rho_{(y,\phi)} \in H$, with $\theta, \phi \in (0, 2\pi)$. Then $x = y$. Thus, all the rotations in H must have a common center of rotation.*

Proof. Suppose $x \neq y$. By Proposition 5.5.25, $\rho_{(x,\theta)}$ and $\rho_{(y,\phi)}$ do not commute, so the commutator $[\rho_{(x,\theta)}, \rho_{(y,\phi)}] \neq \text{id}$. Now,

$$[\rho_{(x,\theta)}, \rho_{(y,\phi)}] = \rho_{(x,\theta)}\rho_{(y,\phi)}\rho_{(x,-\theta)}\rho_{(y,-\phi)}.$$

Since the rotational angles add up to a multiple of 2π , this is a translation by Corollary 5.5.11. But H contains no nonidentity translations, so $x = y$. \square

We now study the elements of finite order in $\mathcal{O}(\{x\})$.

Proposition 6.8.5. *The rotation $\rho_{(x,\theta)}$ has finite order if and only if θ is a rational multiple of 2π . If $\theta = \frac{2\pi k}{n}$ with $\frac{k}{n}$ in lowest terms, then $\rho_{(x,\theta)}$ has order n .*

Proof. Suppose $\text{id} = \rho_{(x,\theta)}^n = \rho_{(x,n\theta)}$. Then $n\theta$ is a multiple of 2π , say $n\theta = 2\pi k$, so $\theta = 2\pi \frac{k}{n}$, a rational multiple of 2π . This must be the case for any element of finite order.

Let $\theta = \frac{2\pi k}{n}$ with $n > 0$, where $\frac{k}{n}$ is in lowest terms. This means k and n are relatively prime, and hence have no common prime divisor. If $\text{id} = \rho_{(x,\theta)}^m = \rho_{(x,m\theta)}$, then $m\theta$ is a multiple of 2π , say $m\theta = 2\pi \ell$ with $\ell \in \mathbb{Z}$. Then $\frac{mk}{n} = \ell$, or $mk = n\ell$, so n divides mk . Since n and k are relatively prime, standard elementary number theory shows n must divide m . In particular, the lowest positive exponent for $\rho_{(x,\theta)}$ is n , so $\rho_{(x,\theta)}$ has order n . \square

$$\text{Write } C_n(x) = \left\langle \rho_{\left(x, \frac{2\pi}{n}\right)} \right\rangle = \tau_x C_n \tau_{-x}.$$

Proposition 6.8.6. *Let H be a finite subgroup of \mathcal{I}_2 with $\mathcal{O}(H) \neq \{\text{id}\}$. Then $\mathcal{O}(H) = C_n(x)$ for some $n > 1$ and $x \in \mathbb{R}^2$.*

Proof. By Corollary 6.8.4, there exists $x \in \mathbb{R}^2$ such that each element of $\mathcal{O}(H)$ has the form $\rho_{(x,\theta)}$ for some θ . Each such θ which occurs must be a rational multiple of 2π . Let $\phi = \frac{2\pi k}{n}$ be the smallest positive such θ which occurs, with $\frac{k}{n}$ in lowest terms and $n > 0$.

We first show that $\mathcal{O}(H) = \langle \rho_{(x,\phi)} \rangle$. To see this, suppose $\rho_{(x,\theta)} \in \mathcal{O}(H)$. Then there is a unique integer m such that θ is in the half-open interval $[m\phi, (m+1)\phi)$. Let $\psi = \theta - m\phi$. Then $\psi \in [0, \phi)$ and

$$\rho_{(x,\psi)} = \rho_{(x,\theta)}\rho_{(x,-m\phi)} = \rho_{(x,\theta)}\rho_{(x,\phi)}^{-m}.$$

The right-hand side is in $\mathcal{O}(H)$, so $\rho_{(x,\psi)} \in \mathcal{O}(H)$. But $0 \leq \psi < \phi$, and ϕ is the smallest positive angle such that the rotation about x by that angle is

in $\mathcal{O}(H)$, so ψ is not positive. Thus, $\psi = 0$, and hence $\theta = m\phi$, and $\rho_{(x,\theta)}$ is a power of $\rho_{(x,\phi)}$.

We now show that this, together with our assumption on the minimality of $\phi = \frac{2\pi k}{n}$ forces $k = 1$. Thus, let $\phi_0 = \frac{2\pi}{n}$. By Proposition 6.8.5, both $\rho_{(x,\phi)}$ and $\rho_{(x,\phi_0)}$ have order n , and hence $\langle \rho_{(x,\phi)} \rangle$ and $\langle \rho_{(x,\phi_0)} \rangle$ have n elements. But $\rho_{(x,\phi)} = \rho_{(x,\phi_0)}^k \in \langle \rho_{(x,\phi_0)} \rangle$. Since $\langle \rho_{(x,\phi)} \rangle$ is the smallest subgroup of \mathcal{I}_2 containing $\rho_{(x,\phi)}$, this forces $\langle \rho_{(x,\phi)} \rangle \subset \langle \rho_{(x,\phi_0)} \rangle$. But each of these groups has n elements, so $\langle \rho_{(x,\phi)} \rangle = \langle \rho_{(x,\phi_0)} \rangle$. Thus, $\rho_{(x,\phi_0)} \in \langle \rho_{(x,\phi)} \rangle \subset \mathcal{O}(H)$. But $\phi_0 \leq \phi$, so the minimality of ϕ forces $\phi_0 = \phi$, and hence $k = 1$. \square

Proof of Leonardo's theorem. If $\mathcal{O}(H) = H$, Proposition 6.8.6 completes the proof. Otherwise, by Corollary 6.6.12, $\mathcal{O}(H)$ has index 2 in H , and if $\beta \in H \setminus \mathcal{O}(H)$, $H \setminus \mathcal{O}(H) = \mathcal{O}(H)\beta$. Such a β must be orientation-reversing. Since H contains no glide reflections, $\beta = \sigma_\ell$ for some ℓ .

If $\mathcal{O}(H) = \{\text{id}\}$, let x be any element of ℓ in the argument below. Otherwise, let x be the unique point in \mathbb{R}^2 such that $\mathcal{O}(H) = C_n(x)$ (uniqueness follows because x is the unique fixed-point for every nonidentity element in $C_n(x)$). Note, then, that x must lie in ℓ , as otherwise, if $\rho_{(x,\theta)} \in \mathcal{O}(H)$, then $\rho_{(x,\theta)}\sigma_\ell$ is a glide reflection in H by Proposition 5.5.22.

In particular, $0 \in \tau_{-x}(\ell)$, so $\tau_{-x}(\ell) = \ell_\phi = \rho_{(0,\phi)}(\ell_0)$ for some ϕ , with ℓ_0 the x -axis. Let $\alpha = \tau_x\rho_{(0,\phi)}$. Then $\alpha(0) = x$ and $\alpha(\ell_0) = \ell$. By Theorem 5.5.20, $\alpha C_n \alpha^{-1} = C_n(x)$ and $\alpha \sigma_{\ell_0} \alpha^{-1} = \sigma_\ell$. Since conjugation by α is a homomorphism, it must take the coset $C_n \sigma_{\ell_0}$ in D_{2n} onto the coset $C_n(x) \sigma_\ell$ in H . We obtain that $H = \alpha D_{2n} \alpha^{-1}$. \square

6.9. Orbits and isotropy in the plane. Fix a subgroup $H \subset \mathcal{I}_2$. We study the way H acts on points in the plane.

Definition 6.9.1. For $x \in \mathbb{R}^2$, the orbit Hx (or $H \cdot x$ if emphasis is needed) of x under the action of H is

$$Hx = \{\alpha(x) : \alpha \in H\}.$$

The isotropy subgroup H_x of x under this action is

$$H_x = \{\alpha \in H : \alpha(x) = x\}.$$

Thus, the orbit is the set of images of x under the transformations in H and the isotropy subgroup is the set of group elements fixing x . It is easily seen to be a subgroup of H .

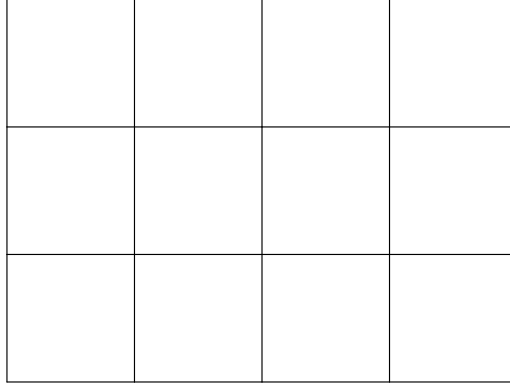
Of course if $H = \mathcal{I}_2$, then $Hx = \mathbb{R}^2$ while $H_x = \mathcal{S}(\{x\})$, studied above. We are more interested in what are called discrete subgroups of \mathcal{I}_2 , in which H is much smaller and the isotropy subgroups H_x are finite, and are therefore either cyclic or dihedral.

Definition 6.9.2. Let $n > 1$. We say that x is an n -center for H if H_x is isomorphic to either C_n or D_{2n} , i.e., the orientation preserving transformations in H_x are precisely $C_n(x)$. A point of symmetry for H is an n -center

for some $n > 1$. We shall also refer to an n -center as a point of symmetry of period n .

In the same spirit, we say ℓ is a line of symmetry for H if $\sigma_\ell \in H$.

Example 6.9.3. We illustrate these concepts with a simple rectangular grid, which we call X .



Assume the grid continues infinitely in both the vertical and horizontal directions, tiling the whole plane. Let $H = \mathcal{S}(X)$, the symmetry group of this pattern. Then the center of each square in the grid can be seen to be a 4-center: the shortest rotation about the center of a square that preserves the pattern is by $90^\circ = \frac{\pi}{2}$. The vertices of the squares are also 4-centers. One can also see that the midpoints of the edges of the squares are 2-centers.

As for lines of symmetry, the grid lines themselves are reflection lines. So are the perpendicular bisectors of the edges of the squares, and so are the lines extending the diagonals of the squares. In particular, each 4-center lies on 4 lines of symmetry, so its isotropy subgroup H_x is D_8 . Each 2-center lies on two lines of symmetry, so its isotropy subgroup H_x is D_4 .

The symmetry group $H = \mathcal{S}(X)$ is the wallpaper group \mathcal{W}_4^1 . We shall study it further in Section 6.14

The study of n -centers will be important in understanding the discrete subgroups of \mathcal{I}_2 . Note the emphasis on orientation-preserving symmetries in the definition of n -center.

We will in general be interested in studying the restriction of the action to $\mathcal{O}(H)$ and $\mathcal{T}(H)$. Note that $\mathcal{O}(H)_x = \mathcal{O}(H_x)$ is exactly what contributes the n in n -center. $\mathcal{T}(H)_x = \{\text{id}\}$ for all x , as nontrivial translations have no fixed-points. But the orbit $\mathcal{T}(H)x$ is important:

Definition 6.9.4. The \mathcal{T} -orbit of x under H is

$$\mathcal{T}(H)x = \{\alpha(x) : \alpha \in \mathcal{T}(H)\}.$$

There is a nice relationship between isotropy and conjugacy.

Lemma 6.9.5. *Let $\alpha \in H$ and $x \in \mathbb{R}^2$. Then $H_{\alpha(x)} = \alpha H_x \alpha^{-1}$. Thus, if x is an n -center for H , so is $\alpha(x)$. In particular, the set of all n -centers for H is H -invariant.*

Proof.

$$\begin{aligned} \beta(\alpha(x)) = \alpha(x) &\Leftrightarrow (\alpha^{-1}\beta\alpha)(x) = x \\ &\Leftrightarrow \alpha^{-1}\beta\alpha \in H_x \\ &\Leftrightarrow \beta \in \alpha H_x \alpha^{-1}. \quad \square \end{aligned}$$

We will see lots of examples of these ideas in the upcoming sections.

6.10. Frieze groups. A frieze group \mathcal{F} is a subgroup of \mathcal{I}_2 such that

$$\mathcal{T}(\mathcal{F}) = \langle \tau_v \rangle = \{ \tau_v^k : k \in \mathbb{Z} \} = \{ \tau_{kv} : k \in \mathbb{Z} \}$$

for some $v \neq 0$. We shall see that frieze groups are symmetry groups of repeating patterns called friezes, and that each is isomorphic to one of seven specific groups.

As a first example, we consider the case where $\mathcal{F} = \mathcal{T}(\mathcal{F}) = \langle \tau_v \rangle$. We call this group $\mathcal{F}_1(v)$, or simply \mathcal{F}_1 . Note that Theorem 5.5.20 shows that the conjugacy class of $\mathcal{F}_1(v)$ depends on $\|v\|$, as τ_v conjugate to τ_w implies $\|v\| = \|w\|$. Moreover, the conjugacy class depends only on $\|v\|$, as if $\|v\| = \|w\|$ there is a rotation $\rho = \rho_{(0,\theta)}$ with $\rho(v) = w$, and hence $\rho\tau_v\rho^{-1} = \tau_w$, and hence $\rho\mathcal{F}_1(v)\rho^{-1} = \mathcal{F}_1(w)$.

It is easy to find a pattern whose symmetry group is \mathcal{F}_1 . For instance, let X be the pattern that repeats infinitely in both directions:

$$\dots \text{ F F F F F F F F F F F F } \dots$$

Since the letter F detects orientation and there are no reverse F's in the pattern, there are no orientation-reversing isometries in $\mathcal{S}(X)$. Also, there are no rotated F's, so $\mathcal{O}(X) = \mathcal{T}(X)$, and the translation subgroup is obviously generated by a single translation. So $\mathcal{S}(X) = \mathcal{T}(X) = \langle \tau_v \rangle$ where τ_v translates each F to the next F to the right.

This pattern X is a typical frieze pattern. It repeats infinitely in both directions and occurs in a narrow strip. Friezes are used in architecture as decorative borders.

Let us now analyze how we can add rotations to \mathcal{F}_1 without adding any additional translations.

Lemma 6.10.1. *Let \mathcal{F} be a subgroup of \mathcal{I}_2 with $\mathcal{T}(\mathcal{F}) = \langle \tau_v \rangle$ for $v \neq 0$. Suppose that $\rho_{(x,\theta)} \in \mathcal{F}$ with $\theta \in (0, 2\pi)$. Then $\theta = \pi$. Thus, every point of symmetry for \mathcal{F} is a 2-center.*

Proof. Write $\rho_{(x,\theta)} = \tau_y\rho_{(0,\theta)}$. Then $\rho_{(x,\theta)}\tau_v\rho_{(x,\theta)}^{-1} = \tau_{\rho_{(0,\theta)}(v)}$ by Theorem 5.5.20. Since \mathcal{F} is closed under conjugation, $\rho_{(0,\theta)}(v) = kv$ for some $k \in \mathbb{Z}$, but since $\rho_{(0,\theta)}(v)$ has the same norm as v , $k = \pm 1$. But $k = 1$ only occurs when θ is a multiple of 2π , which has been ruled out, and $k = -1$ implies $\theta = \pi$ in the range given. \square

π -rotations behave nicely with respect to translations.

Lemma 6.10.2.

- (1) Let $x, w \in \mathbb{R}^2$. Then $\tau_w \rho_{(x, \pi)} = \rho_{(x + \frac{1}{2}w, \pi)}$.
 (2) For $x, y \in \mathbb{R}^2$, $\rho_{(y, \pi)} \rho_{(x, \pi)} = \tau_{2(y-x)}$.

Proof. First note that if ℓ and m are perpendicular, then the directed angle from ℓ to m and the directed angle from m to ℓ are both $\frac{\pi}{2}$, so $\sigma_\ell \sigma_m$ and $\sigma_m \sigma_\ell$ are both equal to the rotation about $\ell \cap m$ by π . This will occur frequently in our analysis of the symmetry groups of patterns.

In this case, we prove (2) by setting m equal to the the line through x and y , ℓ the perpendicular to m through y , and n the perpendicular to m through x . Then

$$\rho_{(y, \pi)} \rho_{(x, \pi)} = \sigma_\ell \sigma_m \sigma_m \sigma_n = \sigma_\ell \sigma_n.$$

Now $\ell \parallel n$, and the directed distance from n to ℓ can be calculated along the perpendicular m to these two lines. We get

$$\sigma_\ell \sigma_n = \tau_{2(m \cap \ell - m \cap n)} = \tau_{2(y-x)},$$

giving (2). (1) now follows from (2) by taking $y = x + \frac{1}{2}w$ and then multiplying both sides of the resulting equation (2) on the right by $\rho_{(x, \pi)}$. Since a rotation by π is an involution (is its own inverse), we obtain (1). \square

We can now characterize the orientation-preserving subgroups of frieze groups.

Proposition 6.10.3. Let \mathcal{F} be a subgroup of \mathcal{I}_2 with $\mathcal{T}(\mathcal{F}) = \langle \tau_v \rangle$ for $v \neq 0$ and with $\mathcal{T}(\mathcal{F}) \neq \mathcal{O}(\mathcal{F})$. Then $\mathcal{T}(\mathcal{F})$ has index 2 in $\mathcal{O}(\mathcal{F})$.

If $\beta \in \mathcal{O}(\mathcal{F}) \setminus \mathcal{T}(\mathcal{F})$, $\beta = \rho_{(x, \pi)}$ for some $x \in \mathbb{R}^2$, and the other elements of $\mathcal{O}(\mathcal{F}) \setminus \mathcal{T}(\mathcal{F})$ are precisely the elements

$$(6.10.1) \quad \tau_{kv} \rho_{(x, \pi)} = \rho_{(x + \frac{k}{2}v, \pi)}, \quad k \in \mathbb{Z}.$$

In particular, there are exactly two \mathcal{T} -orbits of 2-centers for \mathcal{F} : $\mathcal{T}(\mathcal{F})x$ and $\mathcal{T}(\mathcal{F})(x + \frac{1}{2}v)$.

The multiplication in $\mathcal{O}(\mathcal{F})$ is then determined by

$$(6.10.2) \quad \rho_{(x, \pi)} \tau_{kv} \rho_{(x, \pi)}^{-1} = \tau_{-kv},$$

so that $\rho_{(x, \pi)} \tau_{kv} = \tau_{-kv} \rho_{(x, \pi)}$.

Finally, there is a unique line preserved by $\mathcal{O}(\mathcal{F})$: $\ell = x + \text{span}(v)$.

Proof. Lemma 6.10.1 shows that any element of $\mathcal{O}(\mathcal{F}) \setminus \mathcal{T}(\mathcal{F})$ has the form $\rho_{(x, \pi)}$ for some x . Lemma 6.10.2(1) verifies the equation in (6.10.1). These elements must of course lie in $\mathcal{O}(\mathcal{F})$. But Lemma 6.10.2(2) shows that if $\rho_{(y, \pi)} \in \mathcal{O}(\mathcal{F})$, then $\tau_{2(y-x)} \in \mathcal{T}(\mathcal{F}) = \langle \tau_v \rangle$, so $2(y-x) = kv$ for some k , and hence $y = x + \frac{k}{2}v$, as stated.

Therefore, the elements $y = x + \frac{k}{2}v$ are all the 2-centers for \mathcal{F} , and lie in $\mathcal{T}(\mathcal{F})x$ when k is even and in $\mathcal{T}(\mathcal{F})(x + \frac{1}{2}v)$ when k is odd. These orbits are distinct because x and $x + \frac{1}{2}v$ are less than $\|v\|$ apart.

Formula (6.10.2) is easy, using Lemma 6.10.2(2). Let $y = x + \frac{k}{2}v$. Then $\tau_{kv} = \rho_{(y,\pi)}\rho_{(x,\pi)}$, so

$$\rho_{(x,\pi)}\tau_{kv}\rho_{(x,\pi)}^{-1} = \rho_{(x,\pi)}\rho_{(y,\pi)}\rho_{(x,\pi)}\rho_{(x,\pi)}^{-1} = \rho_{(x,\pi)}\rho_{(y,\pi)},$$

and this last is equal to τ_{-kv} by another application of Lemma 6.10.2(2).

By Proposition 6.4.3, a line ℓ is preserved by $\rho_{(y,\pi)}$ if and only if $y \in \ell$. But there is only one line containing all the 2-centers of \mathcal{F} : $\ell = x + \text{span}(v)$. \square

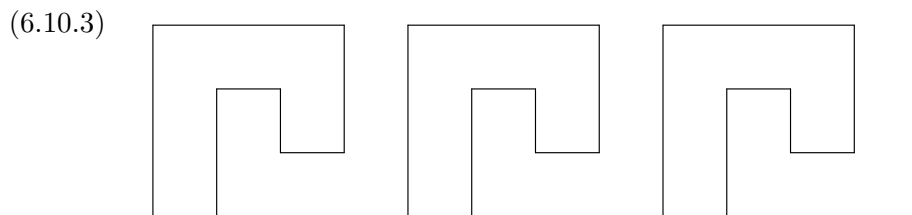
Remark 6.10.4. Proposition 6.10.3 actually gives a construction, depending only on the choice of x and v , for a frieze group \mathcal{F}_2 consisting of orientation-preserving isometries, and shows that any orientation-preserving frieze group containing rotations is isomorphic to it, for appropriate x and v .

We may write this group as $\mathcal{F}_2(v, x)$ if we wish to avoid ambiguity. The reader may check that the conjugacy class of $\mathcal{F}_2(v, x)$ in \mathcal{I}_2 depends only on $\|v\|$.

These groups are isomorphic as x and v vary. Abstractly, they are isomorphic to what's called the infinite dihedral group, D_∞ , which is generated by elements a and b where b has infinite order, a has order 2, and $aba^{-1} = b^{-1}$. Note then that $\langle b \rangle$ has index 2 in D_∞ , and is normal. One may construct a homomorphism of D_∞ onto D_{2n} taking b to b_n and a to a .

This will be the paradigm from here on for the frieze groups studied. A given class of frieze groups consists of groups that are abstractly isomorphic and whose conjugacy class in \mathcal{I}_2 depends only on $\|v\|$.

The following, if continued infinitely in both directions, provides a pattern X whose symmetry group is \mathcal{F}_2 . Find its 2-centers and its shortest translation.



The result of this process of continuing a visual pattern infinitely in both directions will be referred to as the frieze pattern generated by the piece displayed.

We now consider how we can add orientation-reversing isometries to \mathcal{F}_1 or \mathcal{F}_2 to obtain a frieze group leaving $\mathcal{O}(\mathcal{F})$ as stated. We need a lemma to determine which orientation-reversing transformations may occur.

Lemma 6.10.5. *Let \mathcal{F} be a frieze group with $\mathcal{T}(\mathcal{F}) = \mathcal{F}_1 = \langle \tau_v \rangle$. If $\sigma_\ell \in \mathcal{F}$, then ℓ is either parallel or perpendicular to v . If a glide reflection $\tau_x \sigma_\ell \in \mathcal{F}$, then ℓ is parallel to v and $x = \frac{k}{2}v$ for some integer k .*

Proof. The generic orientation-reversing isometry has the form $\alpha = \tau_x \sigma_{\ell_\phi}$ with $\ell_\phi = \text{span}\left(\begin{bmatrix} \cos \phi \\ \sin \phi \end{bmatrix}\right)$. If $\alpha \in \mathcal{F}$, then $\alpha \tau_v \alpha^{-1} \in \mathcal{F}$. By Theorem 5.5.20, $\alpha \tau_v \alpha^{-1} = \tau_{\sigma_{\ell_\phi}(v)}$. Since σ_{ℓ_ϕ} is an isometry, w has the same norm as v , and $\tau_w \in \langle \tau_v \rangle$ if and only if $w = \pm v$. This occurs if and only if ℓ_ϕ is either parallel or perpendicular to v .

By Proposition 5.5.17, α is a reflection in a line parallel to ℓ_ϕ if $x \perp \ell_\phi$, and a glide reflection with axis parallel to ℓ_ϕ otherwise. In particular, we are done if α is a reflection.

We next rule out glide reflections whose axis is perpendicular to v . But that is easy. If ℓ is perpendicular to v and x is parallel to ℓ , then $(\tau_x \sigma_\ell)^2 = \tau_{2x}$, a translation in a direction perpendicular to v , and hence cannot lie in $\mathcal{T}(\mathcal{F}) = \langle \tau_v \rangle$.

The remaining case is a glide reflection $\tau_x \sigma_\ell$ with $\ell \parallel v$, and we apply the same argument. $(\tau_x \sigma_\ell)^2 = \tau_{2x} \in \langle \tau_v \rangle$, so $2x = kv$ for some $k \in \mathbb{Z}$. \square

We may now determine the frieze groups whose translation subgroup is $\langle \tau_v \rangle$. By Corollary 6.6.12, the result will be determined by finding a single orientation-reversing transformation in \mathcal{F} . We first consider the case where $\mathcal{O}(\mathcal{F}) = \mathcal{F}_1$.

Proposition 6.10.6. *Let \mathcal{F} be a frieze group with $\mathcal{O}(\mathcal{F}) = \mathcal{F}_1 = \langle \tau_v \rangle$. Suppose \mathcal{F} contains a reflection σ_ℓ with $\ell \parallel v$. Then*

$$(6.10.4) \quad \mathcal{F} = \langle \tau_v \rangle \cup \{\tau_{kv} \sigma_\ell : k \in \mathbb{Z}\}.$$

Thus $\mathcal{F} \setminus \mathcal{O}(\mathcal{F})$ consists of the reflection σ_ℓ and the glide reflections $\tau_{kv} \sigma_\ell$ with $k \neq 0$.

We call this group \mathcal{F}_1^1 . Its multiplication is determined by the fact that σ_ℓ commutes with all the elements in $\mathcal{O}(\mathcal{F}) = \langle \tau_v \rangle$.

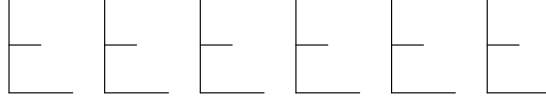
Finally, ℓ is the unique line preserved by \mathcal{F} .

Proof. (6.10.4) follows immediately from Corollary 6.6.12. The transformations $\tau_{kv} \sigma_\ell$ with $k \neq 0$ are glide reflections since $kv \parallel \ell$. The fact that σ_ℓ commutes with τ_v shows that the elements displayed in (6.10.4) do, in fact, form a subgroup of \mathcal{I}_2 .

Finally, Proposition 6.4.3 shows that the only line preserved by a glide reflection is its axis. \square

Remark 6.10.7. Since σ_ℓ commutes with τ_{kv} and since $\langle \tau_v \rangle$ is isomorphic to \mathbb{Z} , \mathcal{F}_1^1 is isomorphic to $\mathbb{Z} \times D_2$. The frieze pattern generated by the

following image has symmetry group \mathcal{F}_1^1 .



Note that the glide reflections $\tau_{kv}\sigma_\ell$ square to τ_v^{2k} , so that every nonidentity even power of τ_v occurs as the square of a glide reflection in \mathcal{F}_1^1 , but no odd power does.

The next case to consider is where \mathcal{F} contains a reflection in a line perpendicular to v .

Proposition 6.10.8. *Let \mathcal{F} be a frieze group with $\mathcal{O}(\mathcal{F}) = \mathcal{F}_1 = \langle \tau_v \rangle$. Suppose \mathcal{F} contains a reflection σ_m with $m \perp v$. Then*

$$(6.10.5) \quad \mathcal{F} = \langle \tau_v \rangle \cup \{ \tau_{kv} \sigma_m : k \in \mathbb{Z} \} = \langle \tau_v \rangle \cup \{ \sigma_{\tau_{\frac{k}{2}v}(m)} : k \in \mathbb{Z} \}.$$

Thus $\mathcal{F} \setminus \mathcal{O}(\mathcal{F})$ consists of the reflections in the translates of m by multiples of $\frac{1}{2}v$. The translations and reflections in \mathcal{F} relate as follows. For $n = \tau_{\frac{k}{2}v}(m)$ we have

$$(6.10.6) \quad \sigma_n \tau_{kv} \sigma_n^{-1} = \tau_{-kv},$$

so $\sigma_n \tau_{kv} = \tau_{-kv} \sigma_n$.

We call this group \mathcal{F}_1^2 . The lines preserved by \mathcal{F}_1^2 consist of all lines parallel to v .

Proof. The first equality in (6.10.5) is just Corollary 6.6.12, and the second equality is Lemma 5.5.16. Equation (6.10.6) also follows from Lemma 5.5.16, which shows both $\sigma_n \tau_{kv}$ and $\tau_{-kv} \sigma_n$ to be equal to the reflection in $\tau_{-\frac{k}{2}v}(n)$.

Note that (6.10.6) shows that the transformations listed in (6.10.5) are closed under multiplication and inverses, and therefore form a subgroup of \mathcal{I}_2 , so this group does exist. The statement about preserved lines is immediate from Proposition 6.4.3. \square

Remark 6.10.9. By (6.10.6), \mathcal{F}_1^2 is abstractly isomorphic to the infinite dihedral group D_∞ . The frieze pattern generated by the following image has symmetry group \mathcal{F}_1^2 .



By Lemma 6.10.5, the only other possibility for a frieze group \mathcal{F} with $\mathcal{O}(\mathcal{F}) = \mathcal{T}(\mathcal{F})$ is that \mathcal{F} contains glide reflections but not reflections.

Proposition 6.10.10. *Let \mathcal{F} be a frieze group with $\mathcal{O}(\mathcal{F}) = \mathcal{F}_1 = \langle \tau_v \rangle$, and suppose \mathcal{F} contains a glide reflection but no reflections. Then the axis*

ℓ of the glide reflection is parallel to v , and \mathcal{F} contains the glide reflection $\gamma = \tau_{\frac{1}{2}v}\sigma_\ell$ that squares to τ_v . Moreover,

$$(6.10.7) \quad \mathcal{F} = \langle \gamma \rangle = \{ \gamma^k : k \in \mathbb{Z} \}.$$

We call this group \mathcal{F}_1^3 . The orientation-preserving elements of \mathcal{F}_1^3 are $\{ \gamma^{2k} = \tau_{kv} : k \in \mathbb{Z} \}$ and the orientation-reversing elements are the glide reflections

$$\{ \gamma^{2k+1} = \tau_{\frac{2k+1}{2}v}\sigma_\ell : k \in \mathbb{Z} \}.$$

Thus, the squares of the glide reflections in \mathcal{F}_1^3 are precisely the odd powers of τ_v . Finally, ℓ is the only line preserved by \mathcal{F}_1^3 .

Proof. By Lemma 6.10.5, \mathcal{F} contains a glide reflection of the form $\tau_{\frac{n}{2}v}\sigma_\ell$, with $\ell \parallel v$, for some $n \in \mathbb{Z}$. If n were even, say $n = 2k$, this would be $\tau_{kv}\sigma_\ell$. But then $\sigma_\ell = \tau_{-kv}\tau_{kv}\sigma_\ell$ is in \mathcal{F} , and \mathcal{F} has been assumed to contain no reflections. Thus, $n = 2k + 1$ is odd. But then

$$\tau_{-kv}\tau_{\frac{2k+1}{2}v}\sigma_\ell = \tau_{\frac{1}{2}v}\sigma_\ell = \gamma$$

is in \mathcal{F} as claimed.

By Corollary 6.6.12, $\mathcal{F} = \mathcal{O}(\mathcal{F}) \cup \mathcal{O}(\mathcal{F})\gamma$. But $\mathcal{O}(\mathcal{F}) = \langle \tau_v \rangle = \langle \gamma^2 \rangle$ is the set of even powers of γ , while $\mathcal{O}(\mathcal{F})\gamma = \langle \gamma^2 \rangle\gamma$ is the set of odd powers of γ , so $\mathcal{F} = \langle \gamma \rangle$. The rest follows easily as above. \square

Remark 6.10.11. \mathcal{F}_1^3 is the symmetry group of the frieze pattern generated by the following.



By Lemma 6.10.5 and the statements of the propositions above, the groups \mathcal{F}_1 , \mathcal{F}_1^1 , \mathcal{F}_1^2 and \mathcal{F}_1^3 exhaust all possibilities for frieze groups with $\mathcal{O}(\mathcal{F}) = F_1$. Each of these has been realized as the symmetry group of a pattern.

We shall now find the frieze groups with $\mathcal{O}(\mathcal{F}) = \mathcal{F}_2$. Lemma 6.10.5 will again come into play.

Proposition 6.10.12. *Let \mathcal{F} be a frieze group with $\mathcal{O}(\mathcal{F}) = \mathcal{F}_2$. Suppose \mathcal{F} contains a reflection σ_ℓ with $\ell \parallel v$. Then ℓ is the line containing the 2-centers for $\mathcal{O}(\mathcal{F})$ (and hence for \mathcal{F}). Moreover, σ_ℓ commutes with all the elements of $\mathcal{O}(\mathcal{F})$, and if y is a 2-center for \mathcal{F} , then $\rho_{(y,\pi)}\sigma_\ell$ is the reflection in the line, $m(y)$, through y perpendicular to ℓ . In particular, the isotropy subgroup \mathcal{F}_y for every 2-center y of \mathcal{F} is D_4 .*

We call this group \mathcal{F}_2^1 . Its orientation-reversing elements consist of the reflections σ_ℓ and $\sigma_{m(y)}$ for y a 2-center of \mathcal{F} , and the glide reflections $\tau_{kv}\sigma_\ell$ for $0 \neq k \in \mathbb{Z}$. Since σ_ℓ commutes with all the elements of $\mathcal{O}(\mathcal{F})$, we have isomorphisms

$$\mathcal{F}_2^1 \cong \mathcal{F}_2 \times D_2 \cong D_\infty \times D_2.$$

Finally, ℓ is the only line preserved by \mathcal{F} .

Proof. Let X_2 be the set of 2-centers. They all lie on a line m parallel to v . By Lemma 6.9.5, if $\sigma_\ell \in \mathcal{F}$, then $\sigma_\ell(X_2) = X_2$. This forces $\sigma_\ell(m) = m$. But since $m \parallel \ell$, that forces $\ell = m$ by Proposition 6.4.3.

In particular, every 2-center y lies on ℓ . But since $m(y) \perp \ell$, the composites $\sigma_{m(y)}\sigma_\ell$ and $\sigma_\ell\sigma_{m(y)}$ are both equal to $\rho_{(y,\pi)}$. So $\sigma_\ell\rho_{(y,\pi)}$ and $\rho_{(y,\pi)}\sigma_\ell$ are both equal to $\sigma_{m(y)}$.

Since ℓ is parallel to v , σ_ℓ commutes with the translations in $\mathcal{O}(\mathcal{F})$, and its composites with them give the stated glide reflections. That σ_ℓ commutes with all the elements in $\mathcal{O}(\mathcal{F})$ shows that adjoining σ_ℓ to $\mathcal{O}(\mathcal{F})$ does in fact provide a group with the stated elements.

Finally, ℓ was the only line preserved by $\mathcal{O}(\mathcal{F})$, and is also preserved by σ_ℓ , so it is the only line preserved by $\mathcal{F} = \mathcal{F}_2^1$. \square

Remark 6.10.13. Note that the squares of the glide reflections in \mathcal{F}_2^1 are precisely the nonidentity even powers of τ_v . \mathcal{F}_2^1 is the symmetry group of the frieze pattern generated by the following.

(6.10.8)



The last frieze group is the most complicated and interesting.

Proposition 6.10.14. Let \mathcal{F} be a frieze group with $\mathcal{O}(\mathcal{F}) = \mathcal{F}_2$ and suppose \mathcal{F} contains a reflection σ_m with $m \perp v$, but does not contain a reflection in any line parallel to v . Let ℓ be the line containing the 2-centers of \mathcal{F} and let $y = m \cap \ell$. Then the closest 2-centers to y are $y \pm \frac{1}{4}v$. Write $x = y + \frac{1}{4}v$. Then $\rho_{(x,\pi)}\sigma_m = \tau_{\frac{1}{2}v}\sigma_\ell = \gamma$, the same γ as in \mathcal{F}_1^3 .

We call this group \mathcal{F}_2^2 . The orientation-reversing isometries in \mathcal{F}_2^2 consist of the reflections

$$\left\{ \tau_{kv}\sigma_m = \sigma_{\tau_{\frac{k}{2}v}(m)} : k \in \mathbb{Z} \right\},$$

and the glide reflections

$$\left\{ \tau_{kv}\rho_{(x,\pi)}\sigma_m = \tau_{kv}\gamma = \gamma^{2k+1} : k \in \mathbb{Z} \right\}.$$

The multiplication in \mathcal{F}_2^2 may be obtained from the equations

$$(6.10.9) \quad \sigma_m\tau_{kv}\sigma_m^{-1} = \tau_{-kv}$$

$$(6.10.10) \quad \sigma_m\rho_{(x,\pi)}\sigma_m^{-1} = \rho_{(x-\frac{1}{2}v,\pi)}.$$

The isotropy subgroups of the 2-centers are all C_2 , and ℓ is the unique line preserved by \mathcal{F}_2^2 .

Proof. Let X_2 be the set of 2-centers for \mathcal{F} . None of the points in X_2 can lie on m , as if $y = m \cap \ell$, then $\rho_{(y,\pi)}\sigma_m = \sigma_\ell$. Since $\sigma_\ell \notin \mathcal{F}$, y cannot be a 2-center for \mathcal{F} .

By Lemma 6.9.5, $\sigma_m(X_2) = X_2$. But closest 2-centers are $\frac{1}{2}\|v\|$ apart. If x is a closest 2-center to y , $\sigma_m(x)$ will have the same distance from y as x does, and will be a closest 2-center to x (the distances are additive, as y lies on the line segment between x and $\sigma_m(x)$). Therefore, both x and $\sigma_m(x)$ must be $\frac{1}{4}\|v\|$ away from y . We may choose $x = y + \frac{1}{4}v$ as stated.

Now $\rho_{(x,\pi)}$ is the product of σ_ℓ with the reflection in the line through x perpendicular to ℓ . That line is $\tau_{\frac{1}{4}v}(m)$. So

$$\rho_{(x,\pi)}\sigma_m = \sigma_\ell\sigma_{\tau_{\frac{1}{4}v}(m)}\sigma_m = \sigma_\ell\tau_{\frac{1}{2}v},$$

as the directed distance from m to $\tau_{\frac{1}{4}v}(m)$ is $\frac{1}{4}v$. But this is γ , as claimed. The listing of orientation-reversing elements is now immediate from Corollary 6.6.12.

Equation (6.10.9) was shown in the analysis of \mathcal{F}_1^2 . For Equation (6.10.10), we have

$$\sigma_m\rho_{(x,\pi)}\sigma_m^{-1} = \sigma_m\sigma_\ell\tau_{\frac{1}{2}v} = \sigma_\ell\sigma_m\tau_{\frac{1}{2}v} = \sigma_\ell\sigma_{\tau_{-\frac{1}{4}v}(m)},$$

with the last equality from Lemma 5.5.16. But $\tau_{-\frac{1}{4}v}(y) = \tau_{-\frac{1}{2}v}(x)$, so the result is $\rho_{(x-\frac{1}{2}v,\pi)}$, as claimed.

Finally, ℓ is the only line preserved by \mathcal{F}_2 , and is still preserved by \mathcal{F}_2^2 . \square

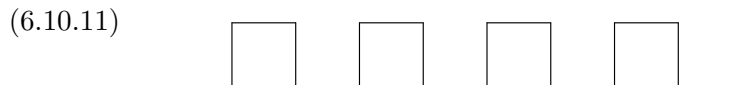
Remark 6.10.15. Note that both \mathcal{F}_1^2 and \mathcal{F}_1^3 are subgroups of \mathcal{F}_2^2 . In fact, \mathcal{F}_2^2 is generated by σ_m and γ , as $\gamma^2 = \tau_v$ and $\gamma = \rho_{(x,\pi)}\sigma_m$. Moreover,

$$\sigma_m\gamma\sigma_m^{-1} = \sigma_m\tau_{\frac{1}{2}v}\sigma_\ell\sigma_m^{-1} = \tau_{-\frac{1}{2}v}\sigma_m\sigma_\ell\sigma_m^{-1} = \tau_{-\frac{1}{2}v}\sigma_\ell = \gamma^{-1},$$

so \mathcal{F}_2^2 is abstractly isomorphic to D_∞ , but perhaps in a surprising way.

Note that the glide reflections in \mathcal{F}_2^2 square to the odd powers of τ_v . In fact these glide reflections all lie in the subgroup \mathcal{F}_1^3 , where they already had this property.

\mathcal{F}_2^2 is the symmetry group of the frieze pattern generated by the following:



Here, the vertical lines of symmetry pass through the midpoints of the horizontal lines of the pattern, while the 2-centers occur at the midpoints of the vertical lines of the pattern.

Finally, we make the now expected observation that there are no other frieze groups.

Proposition 6.10.16. *There are no frieze groups \mathcal{F} with $\mathcal{O}(\mathcal{F}) = \mathcal{F}_2$ that contain glide reflections by not reflections. Thus, the only frieze groups with $\mathcal{O}(\mathcal{F}) = \mathcal{F}_2$ are \mathcal{F}_2 , \mathcal{F}_2^1 and \mathcal{F}_2^2 .*

Proof. Let \mathcal{F} be a frieze group with $\mathcal{O}(\mathcal{F}) = \mathcal{F}_2$ and let α be a glide reflection in \mathcal{F} . Then α must preserve the line ℓ containing the 2-centers, and hence $\alpha = \tau_{\frac{k}{2}v}\sigma_\ell$ for some k (as it squares to a power of τ_v). If k is even, $\alpha = \tau_{nv}\sigma_\ell$ for $n = \frac{k}{2}$, and hence $\sigma_\ell \in \mathcal{F}$. But this implies $\mathcal{F} = \mathcal{F}_2^1$. If k is odd, then $\alpha = \tau_{nv}\gamma$ for $n = \frac{k-1}{2}$, so $\gamma \in \mathcal{F}$. But then $\rho_{(x,\pi)}\gamma = \sigma_m \in \mathcal{F}$, and hence $\mathcal{F} = \mathcal{F}_2^2$. \square

6.11. Fundamental regions and orbit spaces. Let's start by framing an example to analyze. Let \mathcal{F} be a frieze group, and assume the translation subgroup is generated by a horizontal translation, say from left to right. In all cases other than \mathcal{F}_1 and \mathcal{F}_1^2 there is a unique line ℓ preserved by \mathcal{F} . Let us identify ℓ with the x -axis. In the cases of \mathcal{F}_1 and \mathcal{F}_1^2 , where every horizontal line is now preserved, we may also focus attention on the x -axis and call it ℓ . Let

$$(6.11.1) \quad Y = \left\{ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2 : x_2 \in [-c, c] \right\}$$

for some fixed $c > 0$. Thus, Y consists of the points in the plane of distance less than or equal to c from ℓ .¹¹ An analysis of the frieze groups now shows that Y is \mathcal{F} -invariant, i.e., $\mathcal{F} \subset \mathcal{S}(Y)$.

We now wish to focus on how the transformations \mathcal{F} affect Y . We give a more intuitive treatment of some material addressed more rigorously in Appendix A.

Definition 6.11.1. An action of a group G on a set Y gives, for each $g \in G$ and $y \in Y$ an element $gy \in Y$ (sometimes written $g \cdot y$) such that

$$(6.11.2) \quad g_1 \cdot (g_2 \cdot y) = (g_1g_2) \cdot y$$

$$(6.11.3) \quad e \cdot y = y$$

for all $g_1, g_2 \in G$ and $y \in Y$. Here g_1g_2 is the product in G and e is the identity element of G .

Example 6.11.2. If G is a subgroup of $\mathcal{S}(Y)$, then G acts on Y via $g \cdot y = g(y)$.

As the example illustrates, for an action of G on Y and a $g \in G$ there is an induced transformation $\mu_g : Y \rightarrow Y$ given by $\mu_g(y) = g \cdot y$. (6.11.2) is then equivalent to saying that $\mu_{g_1} \circ \mu_{g_2} = \mu_{g_1g_2}$ for all $g_1, g_2 \in G$, while (6.11.3) is equivalent to $\mu_e = \text{id}_Y$, the identity function of Y .

In particular, then $\mu_g \circ \mu_{g^{-1}} = \mu_{g^{-1}} \circ \mu_g = \text{id}_Y$ so each μ_g is bijective with inverse function $\mu_{g^{-1}}$.

Our motivating examples, (6.11.1) and the more general Example 6.11.2, are more than just actions on sets. The geometry of the action is important and is our basic object of study here.

¹¹We take the distance from a point to a line in the plane to be the length of the perpendicular line segment “dropped” from the point to the line.

Definition 6.11.3. Let $Y \subset \mathbb{R}^n$. An action of G on Y is isometric if each $\mu_g \in \mathcal{S}(Y)$ as in Example 6.11.2. More generally, an action is smooth if each μ_g is smooth as described in Chapter 8.1. More generally still, the action is continuous if each μ_g is continuous, as described in Appendix A. In any of these cases, we call Y a G -space, as the structure of Y as a topological space becomes part of the structure of the action.

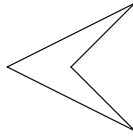
If G acts on both X and Y , a G -map $f : X \rightarrow Y$ is a function such that

$$f(gx) = gf(x)$$

for all $g \in G$ and $x \in X$. If X and Y are G -spaces, we can talk about f being continuous, smooth or isometric, as appropriate.

If f is bijective, its inverse function is a G -map, and we call f an isomorphism of G -sets. If X and Y are G -spaces, we call f a continuous, smooth or metric G -isomorphism if both f and f^{-1} are continuous, smooth or isometric, respectively.

We will find the above concepts very useful in analyzing wallpaper groups. A very important concept is that of fundamental region. We should discuss polygons a bit first. The boundary of a polygon consists of a union of line segments that form a circuit with the property that no two edges meet in an interior point and each vertex lies on exactly two edges. Here is an example of a nonconvex polygon:



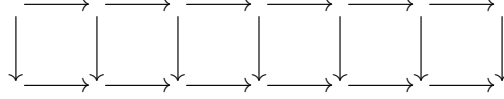
The complement of such a circuit in the plane has exactly two connected pieces: a bounded piece that we call the interior of the polygon, and an unbounded piece, the exterior. The polygon itself consists of the interior plus the boundary. We write ∂P for the boundary of the polygon P and write $\text{Int}(P)$ for its interior.

Definition 6.11.4. Let G act isometrically on $Y \subset \mathbb{R}^2$. A fundamental region for the action of G on Y is a polygon $P \subset Y$ such that:

- (1) $Y = \bigcup_{g \in G} g(P)$.
- (2) If $P \cap g(P) \neq \emptyset$ for $g \neq e$, then $P \cap g(P) \subset \partial P$.

One may define fundamental regions for smooth or continuous actions, but one would not expect them to be polygons. Polygons work in the isometric case, because isometries take polygons to polygons. There are also higher dimensional analogues of fundamental regions. For an isometric action in \mathbb{R}^n , the fundamental region is an n -dimensional polytope.

Example 6.11.5. A pattern with symmetry group \mathcal{F}_1 is generated by:



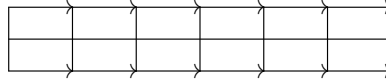
Here, we can take Y to be a slight widening of X (to include the arrow heads). A fundamental region for the action of \mathcal{F}_1 on Y is given by a single rectangle T obtained by vertically thickening the following chamber.



Note the splitting of the arrow heads on the vertical lines, as the remainder of the heads lies in the interiors of adjacent translates of T . In fact, even the slightest horizontal translate of T is also a fundamental region for \mathcal{F}_1 , as it still satisfies the two properties. It doesn't matter how exactly you chop the pattern as long as those two properties are satisfied.


The situation of \mathcal{F}_1^1 is more complicated, as \mathcal{F}_1^1 consists of more than just translations.

Example 6.11.6. A pattern with symmetry group \mathcal{F}_1^1 is generated by



Here, we might first ask for a fundamental region T for the translation subgroup $\mathcal{T}(\mathcal{F}_1^1) = \mathcal{F}_1$, getting a vertical thickening of the following:



The only element of \mathcal{F}_1^1 that carries interior points of T to interior points of T is σ_ℓ , which exchanges the top and bottom chambers of T . So a fundamental region R for the action of \mathcal{F}_1^1 on Y is given by either one of these chambers, say .

Remark 6.11.7. Note that if G acts on Y and if P is a fundamental region for this action, each G -orbit meets P in at least one point by Definition 6.11.4(1). But Definition 6.11.4(2) shows that if a G -orbit Gx meets P in more than one point, then $Gx \cap P$ is contained in ∂P . Thus, if Gx meets P in an interior point, it meets P in no other point.

Definition 6.11.8. Let Y be a G -space. The orbit space, Y/G of the action of G on Y is the set of all orbits of this action. Thus, the points in Y/G are the G -orbits in Y . The canonical map $\pi : Y \rightarrow Y/G$ takes each $y \in Y$ to the orbit containing it.

It is customary to give Y/G the quotient topology induced by the canonical map π . This is discussed in detail in Appendix A. Let P be a fundamental region for the action of G on Y . By Remark 6.11.7, the composite

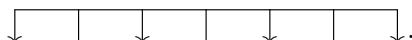
$$P \subset Y \xrightarrow{\pi} Y/G$$

is surjective, and as a set we may regard Y/G as obtained from P by making some identifications on ∂P . In fact, in our situation we may identify the topological space Y/G with the quotient topology obtained from P with these identifications. A rigorous proof is given in Appendix A for the special case of the Klein bottle (Corollary A.5.12). The general case for frieze and wallpaper groups follows from a standard theorem about continuous functions from compact spaces to Hausdorff spaces. We shall not treat it here, but shall content ourselves with describing the orbit spaces in terms of identifications on the fundamental region.

In particular, we can describe the orbit spaces for the actions of \mathcal{F}_1 and \mathcal{F}_1^1 on the horizontal strip Y discussed above. In the case of \mathcal{F}_1 , each point on the left vertical boundary is identified with its image under τ_v , which occurs on the right vertical boundary. There are no other identifications on the fundamental region, so the orbit space is the result of making this identification of the left boundary with the right, resulting in wrapping up the fundamental region to form a cylinder.

In the case of \mathcal{F}_1^1 , we get exactly the same result, only restricting to the smaller fundamental region $\square \rightrightarrows$. The result is again a cylinder by the same reasoning.

Example 6.11.9. A pattern with symmetry group \mathcal{F}_1^2 is generated by



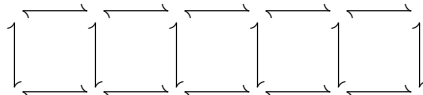
A fundamental region T for the translation subgroup $\mathcal{T}(\mathcal{F}_1^2) = \mathcal{F}_1$, is given by $\square \square$. The only element of \mathcal{F}_1^2 that carries interior points of T to interior points of T is the reflection across the nonbarbed vertical line, which exchanges left and right halves of T . So a fundamental region for the action of \mathcal{F}_1^2 on Y may be given by $\square \rightrightarrows$. There are no identifications on the boundary of this region from the action of \mathcal{F}_1^2 , so the fundamental region and the orbit space coincide in this case.

For \mathcal{F}_1^3 , we use the pattern



A fundamental region, T , for $\mathcal{T}(\mathcal{F}_1^3)$ is given by $\square \uparrow \square$. Of this, the left side, $\square \uparrow$, forms a fundamental region S for \mathcal{F}_1^3 . The glide reflection generating \mathcal{F}_1^3 identifies the left edge of S with the right edge, but twisted so the arrow heads line up together. The result is a Möbius band.

For \mathcal{F}_2 we use the following pattern:




There are 2-centers at the center of each square and at the midpoints of the vertical edges of the squares. A fundamental region T for $\mathcal{T}(\mathcal{F}_2)$ is given by a single square with the outer arrowheads chopped:



The only element of \mathcal{F}_2 taking interior points of T to interior points of T is the rotation by π about the center of T . For a fundamental region S for \mathcal{F}_2 we can divide T in half by any line through the center, and then take either of the halves as S . For instance, divide T in half with a vertical line. Then on each vertical edge of S , the upper half of the edge is identified with the lower half by the rotation about the center point of the edge. In particular, the orbit space looks like a pillow case.

For \mathcal{F}_2^1 we can use (6.10.8) for a pattern, getting a single square as a fundamental region T for $\mathcal{T}(\mathcal{F}_2^1)$. The elements of \mathcal{F}_2^1 that carry interior points of T to interior points of T consist of σ_ℓ , the reflection in the line perpendicular to σ_ℓ going through the center of the square, and the rotation about the center by π . A fundamental region S for \mathcal{F}_2^1 is given by any of the four squares whose boundaries are given by these lines of reflection, and the orbit space coincides with S .

For \mathcal{F}_2^2 we use (6.10.11) for a pattern. The 2-centers are at the midpoints of the vertical lines of the pattern and lines of symmetry are the vertical lines through the midpoints of the horizontal lines in the pattern. A fundamental region T for $\mathcal{T}(\mathcal{F})$ is given by . The elements of \mathcal{F}_2^2 taking interior points of T to interior points of T are the two reflections and the rotation about the center point. This allows us to choose the left quarter of T , from the left edge to the first line of symmetry, as a fundamental region, S , for \mathcal{F}_2^2 . In this case, the upper half of the left edge of S is identified to the lower half of the left edge by the rotation about the center point of that edge. The resulting orbit space is a cone with two corners on its edge.

6.12. Translation lattices in \mathbb{R}^n . Frieze groups are the isometry groups in \mathcal{I}_2 whose translation subgroups are cyclic of infinite order: $\mathcal{T}(\mathcal{F}) = \langle \tau_v \rangle$ for some $v \neq 0$. Of course $v \neq 0$ is equivalent to saying $\text{span}(v) \neq 0$, and hence the singleton v is linearly independent.

Wallpaper groups, a.k.a. 2-dimensional crystallographic groups, are subgroups $\mathcal{W} \subset \mathcal{I}_2$ whose translation subgroup is given by

$$\mathcal{T}(\mathcal{W}) = \langle \tau_v, \tau_w \rangle = \{ \tau_v^k \tau_w^\ell : k, \ell \in \mathbb{Z} \} = \{ \tau_{kv+\ell w} : k, \ell \in \mathbb{Z} \},$$

where v, w are linearly independent in \mathbb{R}^2 . Our goal in this section is to understand these translation subgroups. They are what's known as 2-dimensional translation lattices. Since translation lattices are also important in studying higher dimensional crystallographic groups (with important physical applications when $n = 3$), we shall study translation lattices in \mathbb{R}^n .

Recall from Proposition 3.4.12 that there is an isomorphism of groups

$$\nu : \mathbb{R}^n \xrightarrow{\cong} \mathcal{T}_n$$

given by $\nu(x) = \tau_x$, where \mathbb{R}^n is an additive group in the usual way. The translation subgroup of a wallpaper group is the image under ν of a lattice in \mathbb{R}^2 :

Definition 6.12.1. A lattice in \mathbb{R}^n is the additive subgroup generated by a basis v_1, \dots, v_n of \mathbb{R}^n as a vector space over \mathbb{R} . Thus, v_1, \dots, v_n are linearly independent and their associated lattice is

$$\Lambda = \langle v_1, \dots, v_n \rangle = \{ a_1 v_1 + \dots + a_n v_n : a_1, \dots, a_n \in \mathbb{Z} \}.$$

We call $\mathcal{B} = v_1, \dots, v_n$ a \mathbb{Z} -basis for Λ , and write $\Lambda_{\mathcal{B}} = \Lambda(v_1, \dots, v_n)$ for the lattice with \mathbb{Z} -basis \mathcal{B} .

We write \mathcal{T}_{Λ} for the image of Λ under $\nu : \mathbb{R}^n \cong \mathcal{T}_n$ and call it the translation lattice induced by Λ . In particular, for $\Lambda = \Lambda(v_1, \dots, v_n)$,

$$\begin{aligned} \mathcal{T}_{\Lambda} &= \langle \tau_{v_1}, \dots, \tau_{v_n} \rangle = \{ \tau_{v_1}^{a_1} \dots \tau_{v_n}^{a_n} : a_1, \dots, a_n \in \mathbb{Z} \} \\ &= \{ \tau_{a_1 v_1 + \dots + a_n v_n} : a_1, \dots, a_n \in \mathbb{Z} \}. \end{aligned}$$

In particular, a wallpaper group is a subgroup $\mathcal{W} \subset \mathcal{I}_2$ such that $\mathcal{T}(\mathcal{W})$ is a translation lattice.

Example 6.12.2. The standard lattice $\Lambda_{\mathcal{E}}$ in \mathbb{R}^n is $\Lambda(e_1, \dots, e_n)$: the lattice whose \mathbb{Z} -basis is the canonical basis \mathcal{E} of \mathbb{R}^n . Note that

$$\Lambda_{\mathcal{E}} = \left\{ \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} : a_1, \dots, a_n \in \mathbb{Z} \right\} = \mathbb{Z}^n,$$

the Cartesian product of n copies of \mathbb{Z} . The group structure on \mathbb{Z}^n here is the direct product of n copies of \mathbb{Z} , as the addition in \mathbb{R}^n is coordinatewise.

The following helps us get a handle on lattices.

Lemma 6.12.3. Let $\mathcal{B} = v_1, \dots, v_n$ be a \mathbb{Z} -basis for a lattice in \mathbb{R}^n and let $A_{\mathcal{B}} = [v_1 | \dots | v_n]$, the $n \times n$ matrix whose i th column is v_i for all i . Then the linear transformation $T_{A_{\mathcal{B}}}$ induced by $A_{\mathcal{B}}$ is a linear isomorphism

$$T_{A_{\mathcal{B}}} : \mathbb{R}^n \xrightarrow{\cong} \mathbb{R}^n$$

and restricts to an isomorphism of groups:

$$f_{\mathcal{B}} : \mathbb{Z}^n = \Lambda_{\mathcal{E}} \xrightarrow{\cong} \Lambda_{\mathcal{B}}.$$

Specifically,

$$(6.12.1) \quad f_{\mathcal{B}} \left(\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \right) = a_1 v_1 + \cdots + a_n v_n$$

for $a_1, \dots, a_n \in \mathbb{Z}$.

Proof. $T_{A_{\mathcal{B}}}$ is an isomorphism because the columns of $A_{\mathcal{B}}$ form a basis of \mathbb{R}^n . In fact, (6.12.1) holds for $T_{A_{\mathcal{B}}}$ for any $a_1, \dots, a_n \in \mathbb{R}$ by matrix multiplication, so it holds for $f_{\mathcal{B}} : \mathbb{Z}^n \rightarrow \Lambda_{\mathcal{B}}$ as well. But that shows $f_{\mathcal{B}}$ maps \mathbb{Z}^n onto $\Lambda_{\mathcal{B}}$, while $f_{\mathcal{B}}$ is one-to-one because $T_{A_{\mathcal{B}}}$ is. \square

Corollary 6.12.4. *Any two n -dimensional translation lattices are conjugate in the group \mathcal{A}_n of affine automorphisms of \mathbb{R}^n .*

Proof. By Lemma 3.3.1, $T_{A_{\mathcal{B}}} \tau_v T_{A_{\mathcal{B}}}^{-1} = \tau_{A_{\mathcal{B}}v} = \tau_{f_{\mathcal{B}}(v)}$ for all $v \in \Lambda_{\mathcal{E}}$. \square

For simplicity of notation, write elements of \mathbb{Z}^n as row vectors. Note that we obtain a composite isomorphism $\nu \circ f_{\mathcal{B}} : \mathbb{Z}^n \rightarrow \mathcal{T}_{\Lambda_{\mathcal{B}}}$ via

$$\nu \circ f_{\mathcal{B}}(a_1, \dots, a_n) = \tau_{a_1 v_1 + \cdots + a_n v_n} = \tau_{v_1}^{a_1} \cdots \tau_{v_n}^{a_n}.$$

Now, $\mathcal{T}_{\Lambda_{\mathcal{B}}}$ acts isometrically on \mathbb{R}^n , as translations are isometries. So we get an isometric action of \mathbb{Z}^n on \mathbb{R}^n via

$$(a_1, \dots, a_n) \cdot w = \tau_{a_1 v_1 + \cdots + a_n v_n}(w).$$

We write $\mathbb{R}_{\mathcal{B}}^n$ for \mathbb{R}^n with this action of \mathbb{Z}^n .

But \mathbb{Z}^n also acts on \mathbb{R}^n via the isometric action induced by $\Lambda_{\mathcal{E}}$:

$$(a_1, \dots, a_n) \cdot w = \tau_{a_1 e_1 + \cdots + a_n e_n}(w).$$

We write $\mathbb{R}_{\mathcal{E}}^n$ for \mathbb{R}^n with this action of \mathbb{Z}^n . Of course, if a is the column vector with ordered coordinates a_1, \dots, a_n , this is just $\tau_a(w)$.

Proposition 6.12.5. *Let $\mathcal{B} = v_1, \dots, v_n$ be a \mathbb{Z} -basis for a lattice in \mathbb{R}^n . Then*

$$T_{A_{\mathcal{B}}} : \mathbb{R}_{\mathcal{E}}^n \xrightarrow{\cong} \mathbb{R}_{\mathcal{B}}^n$$

is a linear (but not necessarily isometric) isomorphism of \mathbb{Z}^n -spaces.

Proof. $T_{A_{\mathcal{B}}}$ is a linear isomorphism, and is isometric if and only if the columns of $A_{\mathcal{B}}$, v_1, \dots, v_n , form an orthonormal basis of \mathbb{R}^n . It suffices to show $T_{A_{\mathcal{B}}}$ is a \mathbb{Z}^n -map.

$$\begin{aligned} T_{A_{\mathcal{B}}}((a_1, \dots, a_n) \cdot w) &= A_{\mathcal{B}} \cdot (w + a_1 e_1 + \cdots + a_n e_n) \\ &= A_{\mathcal{B}} w + a_1 A_{\mathcal{B}} e_1 + \cdots + a_n A_{\mathcal{B}} e_n \\ &= T_{A_{\mathcal{B}}}(w) + a_1 v_1 + \cdots + a_n v_n \\ &= (a_1, \dots, a_n) \cdot T_{A_{\mathcal{B}}}(w), \end{aligned}$$

where the final action is evaluated in $\mathbb{R}_{\mathcal{E}}^n$. □

This actually gives us all we need to define fundamental regions for translation lattices. Recall that the n -cube I^n is the set of all vectors $x \in \mathbb{R}^n$ whose coordinates all lie in the unit interval $I = [0, 1]$. Thus,

$$I^n = \{x_1e_1 + \cdots + x_ne_n : x_i \in I \text{ for all } i\}.$$

The boundary ∂I^n consists of those elements of I^n with the property that at least one of their coordinates lies in $\{0, 1\}$. ∂I^n is the union of $(n - 1)$ -dimensional faces, $\partial_i^0 I^n$ and $\partial_i^1 I^n$, for $i = 1, \dots, n$. Here, for $\epsilon = 0$ or 1 ,

$$\partial_i^\epsilon I^n = \{x_1e_1 + \cdots + x_ne_n \in I^n : x_i = \epsilon\}.$$

There is an obvious affine bijection $\iota_i^\epsilon : I^{n-1} \rightarrow \partial_i^\epsilon I^n$,

$$\iota_i^\epsilon \left(\begin{bmatrix} x_1 \\ \vdots \\ x_{n-1} \end{bmatrix} \right) = \begin{bmatrix} x_1 \\ \vdots \\ x_{i-1} \\ \epsilon \\ x_i \\ \vdots \\ x_{n-1} \end{bmatrix},$$

so the faces of I^n are $(n - 1)$ -dimensional polytopes. We have

$$\partial I^n = \bigcup_{i=1}^n (\partial_i^0 I^n \cup \partial_i^1 I^n).$$

The interior of I^n is

$$\text{Int}(I^n) = \{x_1e_1 + \cdots + x_ne_n : x_i \in (0, 1) \text{ for all } i\}.$$

Lemma 6.12.6. I^n is a fundamental region for the action of \mathbb{Z}^n (or $\Lambda_{\mathcal{E}}$) on $\mathbb{R}_{\mathcal{E}}^n$ in the sense that:

- (1) $\mathbb{R}^n = \bigcup_{a \in \mathbb{Z}^n} \tau_a(I^n)$.
- (2) If $I^n \cap \tau_a(I^n) \neq \emptyset$ for $a \neq 0$, then $I^n \cap \tau_a(I^n) \subset \partial I^n$.

Moreover, each \mathbb{Z}^n -orbit

$$\mathbb{Z}^n \cdot x = \Lambda_{\mathcal{E}} \cdot x$$

intersects I^n . If $\Lambda_{\mathcal{E}} \cdot x$ intersects the interior of I^n , then $(\Lambda_{\mathcal{E}} \cdot x) \cap I^n$ consists of just this one interior point. If $\Lambda_{\mathcal{E}} \cdot x$ intersects the boundary of I^n , say $y \in (\Lambda_{\mathcal{E}} \cdot x) \cap \partial I^n$, then $(\Lambda_{\mathcal{E}} \cdot x) \cap I^n$ is contained entirely in ∂I^n , and consists of 2^k points, where k is the number of coordinates of y that lie in $\{0, 1\}$.

Proof. Let $x = x_1e_1 + \cdots + x_ne_n \in \mathbb{R}^n$ and let $a = [x_1]e_1 + \cdots + [x_n]e_n$ where $[t]$ denotes the greatest integer less than or equal to the real number t . Then $x \in \tau_a(I^n)$, so (1) holds.

For (2), let $I^n \cap \tau_a(I^n) \neq \emptyset$ with $0 \neq a = a_1e_1 + \cdots + a_ne_n \in \mathbb{Z}^n$. Then τ_a translates the i th coordinate by a_i for all i . This says $[0, 1] \cap [a_i, a_i + 1] \neq \emptyset$ for $i = 1, \dots, n$, and hence $a_i \in \{-1, 0, 1\}$ for all i . Since $a \neq 0$, at least

one $a_i \in \{-1, 1\}$, and hence any point in $I^n \cap \tau_a(I^n)$ must have at least one coordinate in $\{0, 1\}$.

Regarding orbits, each orbit intersects I^n by (1). More specifically, let x and a be as above and let $z_i = (x_i - \lfloor x_i \rfloor)$ for $i = 1, \dots, n$. Then

$$z = z_1 e_1 + \dots + z_n e_n \in (\Lambda_{\mathcal{E}} \cdot x) \cap I^n.$$

Note that $z_i \in [0, 1)$ by construction. If $z_i \in (0, 1)$ and $\tau_a(z) \in I^n$, then $a_i = 0$. If $z_i = 0$ and $\tau_a(z) \in I^n$, then $a_i \in \{0, 1\}$. Considering those i such that $z_i = 0$, there are 2^k choices of a for which $\tau_a(z) \in I^n$. Compare the resulting elements $\tau_a(z)$ to the y in the statement, and the result follows. \square

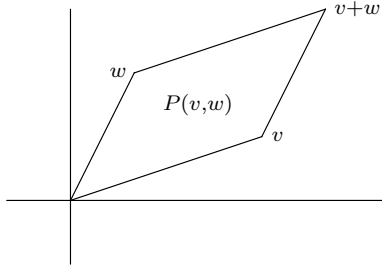
Corollary 6.12.7. *Let $\mathcal{B} = v_1, \dots, v_n$ be a \mathbb{Z} -basis for a lattice in \mathbb{R}^n and let $T_{A_{\mathcal{B}}} : \mathbb{R}_{\mathcal{E}}^n \xrightarrow{\cong} \mathbb{R}_{\mathcal{B}}^n$ be the induced linear \mathbb{Z}^n -isomorphism. Then $T_{A_{\mathcal{B}}}(I^n)$ is a fundamental region for the action of \mathbb{Z}^n (and hence $\Lambda_{\mathcal{B}}$) on $\mathbb{R}_{\mathcal{B}}^n$. Moreover,*

$$T_{A_{\mathcal{B}}}(I^n) = \{x_1 v_1 + \dots + x_n v_n : x_i \in I \text{ for all } i\}.$$

We call it the parallelepiped $P(v_1, \dots, v_n)$ generated by v_1, \dots, v_n .

Proof. By Proposition 6.2.6, $T_{A_{\mathcal{B}}}(I^n)$ is an n -dimensional polytope with boundary $T_{A_{\mathcal{B}}}(\partial I^n)$. \square

Of special interest here is the case $n = 2$. Here, I^2 is the usual unit square and if $\mathcal{B} = v, w$ is a \mathbb{Z} -basis for a lattice in \mathbb{R}^2 , then the parallelepiped $P(v, w)$ generated by v, w is the parallelogram in \mathbb{R}^2 with vertices $0, v, w$ and $v + w$:



The edges of this parallelogram are the images of the edges of I^2 under $T_{A_{\mathcal{B}}}$, and its interior points are the images of the interior points of I^2 under $T_{A_{\mathcal{B}}}$: $\{sv + tw : s, t \in (0, 1)\}$.

By Corollary 6.12.7, $P(v, w)$ is a fundamental region for the action of $\Lambda_{\mathcal{B}}$ (or $\mathcal{T}_{\Lambda_{\mathcal{B}}}$) on \mathbb{R}^n . This is obvious from the linear algebra and the proof of Corollary 6.12.7, but not obvious from the fact that $\Lambda_{\mathcal{B}}$ is generated by v, w as an abelian group. In fact, we can use fundamental regions to find alternative \mathbb{Z} -bases for $\Lambda_{\mathcal{B}}$.

In fact, a given lattice will have infinitely many \mathbb{Z} -bases, and infinitely many noncongruent parallelograms as fundamental regions. But any two fundamental regions have the same area. The area is calculated as follows.

Proposition 6.12.8. *Let $\mathcal{B} = v, w$ be a \mathbb{Z} -basis for a lattice in \mathbb{R}^2 . Then the area of $P(v, w)$ is $|\det A_{\mathcal{B}}|$.*

Proof. Let $B = A_{\mathcal{B}}^T A_{\mathcal{B}}$, where T represents the transpose. Then

$$\det B = \det A_{\mathcal{B}}^T \det A_{\mathcal{B}} = (\det A_{\mathcal{B}})^2,$$

as the determinant of a square matrix is equal to the determinant of its transpose. Thus, it suffices to show that the square of the area of $P(v, w)$ is equal to $\det B$.

By Corollary 4.1.11,

$$B = \begin{bmatrix} \langle v, v \rangle & \langle v, w \rangle \\ \langle v, w \rangle & \langle w, w \rangle \end{bmatrix},$$

so $\det B = \langle v, v \rangle \langle w, w \rangle - \langle v, w \rangle^2$. We calculate the area of the parallelogram $P(v, w)$ as the product of its base and its height. We leave it as an exercise to the reader that this formula works. We take v as the base vector, so the base length is $\|v\|$. For the height, we use the formula from the Gram–Schmidt orthogonalization: the height vector is $w - \frac{\langle v, w \rangle}{\langle v, v \rangle} v$. So the square of the length of the height vector is

$$\begin{aligned} \left\langle w - \frac{\langle v, w \rangle}{\langle v, v \rangle} v, w - \frac{\langle v, w \rangle}{\langle v, v \rangle} v \right\rangle &= \langle w, w \rangle - 2 \frac{\langle v, w \rangle}{\langle v, v \rangle} \langle v, w \rangle + \frac{\langle v, w \rangle^2}{\langle v, v \rangle^2} \langle v, v \rangle \\ &= \langle w, w \rangle - \frac{\langle v, w \rangle^2}{\langle v, v \rangle}. \end{aligned}$$

So the square of the area of $P(v, w)$ is this times $\|v\|^2$, which is precisely $\det B$. \square

We now go back to n -dimensional lattices. Recall that $\mathrm{GL}_n(\mathbb{Z})$ is the subgroup of $\mathrm{GL}_n(\mathbb{R})$ consisting of the matrices with integer coefficients whose inverse matrix also has integer coefficients. We saw in Proposition 3.1.8 that a matrix $A \in \mathrm{GL}_n(\mathbb{R})$ with integer coefficients lies in $\mathrm{GL}_n(\mathbb{Z})$ if and only if $\det A = \pm 1$.

Proposition 6.12.9. *Let Λ be a lattice in \mathbb{R}^n with \mathbb{Z} -basis $\mathcal{B} = v_1, \dots, v_n$. Then $\mathcal{B}' = w_1, \dots, w_n$ is also a \mathbb{Z} -basis of Λ if and only if there is a matrix $B \in \mathrm{GL}_n(\mathbb{Z})$ such that w_i is the i th column of $A_{\mathcal{B}} B$ for all i (i.e., $A_{\mathcal{B}} B = A_{\mathcal{B}'}$).*

Proof. Let $B \in \mathrm{GL}_n(\mathbb{Z})$ and let w_i be the i th column of $A_{\mathcal{B}} B$ for $i = 1, \dots, n$. We show $\mathcal{B}' = w_1, \dots, w_n$ is a \mathbb{Z} -basis for $\Lambda_{\mathcal{B}}$, i.e., that $\Lambda_{\mathcal{B}} = \Lambda_{\mathcal{B}'}$. First note that since B is invertible, $A_{\mathcal{B}} B$ is invertible, and hence its columns, w_1, \dots, w_n form a basis \mathcal{B}' of \mathbb{R}^n over \mathbb{R} , and hence do form a \mathbb{Z} -basis for a lattice. Since $\Lambda_{\mathcal{B}}$ is generated as an abelian group by v_1, \dots, v_n , it is the smallest subgroup of \mathbb{R}^n containing v_1, \dots, v_n , so if $v_1, \dots, v_n \in \Lambda_{\mathcal{B}'}$, then $\Lambda_{\mathcal{B}} \subset \Lambda_{\mathcal{B}'}$. Similarly, if $w_1, \dots, w_n \subset \Lambda_{\mathcal{B}}$, then $\Lambda_{\mathcal{B}'} \subset \Lambda_{\mathcal{B}}$. Thus, it suffices to show that

$$(6.12.2) \quad v_1, \dots, v_n \in \Lambda_{\mathcal{B}'} = \langle w_1, \dots, w_n \rangle,$$

$$(6.12.3) \quad w_1, \dots, w_n \in \Lambda_{\mathcal{B}} = \langle v_1, \dots, v_n \rangle.$$

Let $B = (b_{ij})$, i.e., the ij th entry of B is b_{ij} . Since w_i is the i th column of $A_{\mathcal{B}}B$ and since $A_{\mathcal{B}} = [v_1, \dots, v_n]$, we have

$$w_i = b_{1i}v_1 + \dots + b_{ni}v_n \in \Lambda_{\mathcal{B}},$$

so (6.12.3) holds. By the definition of \mathcal{B}' , $A_{\mathcal{B}}B = A_{\mathcal{B}'}$. So $A_{\mathcal{B}'}B^{-1} = A_{\mathcal{B}}$. Since B^{-1} has integer coefficients, the same argument gives (6.12.2).

For the converse, suppose $\Lambda_{\mathcal{B}} = \Lambda_{\mathcal{B}'}$. Then both (6.12.2) and (6.12.3) hold. By (6.12.3), there are integers b_{ij} , $1 \leq i, j \leq n$ with $w_i = b_{1i}v_1 + \dots + b_{ni}v_n$ for all i , so if $B = (b_{ij})$, then $A_{\mathcal{B}}B = A_{\mathcal{B}'}$. By (6.12.2), there are integers b'_{ij} , $1 \leq i, j \leq n$ with $v_i = b'_{1i}w_1 + \dots + b'_{ni}w_n$ for all i . Setting $B' = (b'_{ij})$ we get $A_{\mathcal{B}'}B' = A_{\mathcal{B}}$. So $A_{\mathcal{B}}BB' = A_{\mathcal{B}}$. Since $A_{\mathcal{B}}$ is invertible $BB' = I_n$ so B and B' are inverse integer matrices and hence $B \in \text{GL}_n(\mathbb{Z})$. \square

Corollary 6.12.10. *Let $\mathcal{B} = v_1, v_2$ and $\mathcal{B}' = w_1, w_2$ be two different \mathbb{Z} -bases for a lattice Λ in \mathbb{R}^2 . Then $P(v_1, v_2)$ and $P(w_1, w_2)$ have the same area.*

Proof. By Proposition 6.12.9, there is a matrix $B \in \text{GL}_2(\mathbb{Z})$ with $A_{\mathcal{B}}B = A_{\mathcal{B}'}$. So

$$|\det A_{\mathcal{B}'}| = |\det A_{\mathcal{B}} \det B| = |\det A_{\mathcal{B}}|,$$

as matrices in $\text{GL}_n(\mathbb{Z})$ have determinant ± 1 . The result now follows from Proposition 6.12.8. \square

Proposition 6.12.9 provides many examples of alternative \mathbb{Z} -bases for a lattice $\Lambda(v, w)$ in \mathbb{R}^2 . For instance $\mathcal{B}' = 2v + w, 3v + 2w$ is one such, as $\det \begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix} = 1$. The following gives an infinite family of \mathbb{Z} -bases that may be of some interest.

Corollary 6.12.11. *Let $\mathcal{B} = v, w$ be a \mathbb{Z} -basis for a lattice Λ in \mathbb{R}^2 . Then so are $\mathcal{B}' = v, v+w$, $\mathcal{B}'' = v, 2v+w$, etc. The fundamental region $P(v, v+w)$ is the parallelogram generated by v and the main diagonal of $P(v, w)$. We obtain a sequence, $P(v, w)$, $P(v, v+w)$, \dots , $P(v, kv+w)$, \dots of fundamental regions for Λ , infinitely many of which are noncongruent.*

Proof. Let $B = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. Then $A_{\mathcal{B}}B = A_{\mathcal{B}'}$, $A_{\mathcal{B}'}B = A_{\mathcal{B}''}$, etc., so Proposition 6.12.9 shows that $v, kv+w$ is a \mathbb{Z} -basis for $\Lambda_{\mathcal{B}}$ for all $k \geq 0$.

It is not impossible for $P(v, w)$ and $P(v, v+w)$ to be congruent. E.g., if $v = e_1$ and $w = \begin{bmatrix} \cos \frac{2\pi}{3} \\ \cos \frac{2\pi}{3} \end{bmatrix}$, then $v+w = \begin{bmatrix} \cos \frac{\pi}{3} \\ \cos \frac{\pi}{3} \end{bmatrix}$. The two parallelograms $P(v, w)$ and $P(v, v+w)$ are congruent, but not by a congruence preserving the origin. In $P(v, w)$, the origin is at one of the angles of measure $\frac{2\pi}{3}$ occurring in the parallelogram, and in $P(v, v+w)$, the origin is at one of the angles of measure $\frac{\pi}{3}$.

Regardless, as we move forward in the sequence of fundamental regions, the angle at 0 gets smaller and smaller, and we get infinitely many noncongruent fundamental regions for whatever our choice of Λ was. To see this, note that all the parallelograms in the sequence have the same area.

Because the all have the same base vector v , they also all have the same height h . Let θ_k be the angle at 0 in the parallelogram $P(v, kv + w)$. For simplicity of calculation, rotate so that $v = te_1$ for $t > 0$, and reflect, if necessary, so that $w = \begin{bmatrix} x \\ h \end{bmatrix}$. Then $kv + w = \begin{bmatrix} kt+x \\ h \end{bmatrix}$, so $\tan \theta_k = \frac{h}{kt+x}$, and hence $\lim_{k \rightarrow \infty} \theta_k = 0$. \square

In classifying wallpaper groups, we'll start with an abstract subgroup of \mathcal{I}_2 whose translation subgroup is a translation lattice and then deduce information about the group. In particular, we will not be given a preferred \mathbb{Z} -basis for the lattice and we'll need to be able to find one.

The following definition is nonstandard but is nicely adapted to studying groups of isometries. It will eliminate some of the point set topology that might come into play here.

Definition 6.12.12. A subset $X \subset \mathbb{R}^n$ is uniformly discrete if there exists $\epsilon > 0$ such that $d(x, y) > \epsilon$ for all $x \neq y \in X$.

The following suggests this will be useful.

Lemma 6.12.13. Let Λ be a lattice in \mathbb{R}^n and let $x \in \mathbb{R}^n$. Then the orbit $\{x + v : v \in \Lambda\}$ of x under the translation action by Λ is uniformly discrete.

Proof. Since $d(x + v, x + w) = d(v, w)$, the result is independent of the choice of x (and the same ϵ can be used for any x). Moreover, for a fixed x and for $w, v \in \Lambda$, $d(x + v, x + w) = d(x, x + w - v)$. As $w - v \in \Lambda$, it suffices to find an ϵ such that the distance from x to any of its translates by nonzero elements of Λ must be greater than ϵ .

Let $\mathcal{B} = v_1, \dots, v_n$ be a \mathbb{Z} -basis for Λ , and consider the linear isomorphism $T_{A_{\mathcal{B}}} : \mathbb{R}_{\mathcal{E}}^n \xrightarrow{\cong} \mathbb{R}_{\mathcal{B}}^n$. Since the result is independent of x , we can take $x = \frac{1}{2}v_1 + \dots + \frac{1}{2}v_n$, the center of mass of $P(v_1, \dots, v_n)$. Let $y = \frac{1}{2}e_1 + \dots + \frac{1}{2}e_n = T_{A_{\mathcal{B}}}^{-1}(x)$. Then if $z \in \mathbb{R}^n$ with $d(y, z) < \frac{1}{2}$, it is easy to see that z lies in the interior of $I^n = P(e_1, \dots, e_n)$. By the continuity of $T_{A_{\mathcal{B}}}^{-1}$ there exists $\epsilon > 0$ such that $d(T_{A_{\mathcal{B}}}^{-1}(w), y) < \frac{1}{2}$ whenever $d(w, x) < \epsilon$. Thus, for $d(w, x) < \epsilon$, w lies in the interior of $P(v_1, \dots, v_n)$. But since $P(v_1, \dots, v_n)$ is a fundamental region for $\Lambda_{\mathcal{B}}$, the interior of $P(v_1, \dots, v_n)$ is disjoint from its image under any element of $\Lambda_{\mathcal{B}}$. So any translate of x by a nonzero element of $\Lambda_{\mathcal{B}}$ has distance more than ϵ from x . \square

Remark 6.12.14. This underlines the difference between saying that v, w is a \mathbb{Z} -basis for a lattice in \mathbb{R}^2 and the weaker condition that $\langle v, w \rangle$ is isomorphic to $\mathbb{Z} \times \mathbb{Z}$. The latter condition holds for $v = e_1$ and $w = re_1$ for some irrational number r , as no integral multiple of r can be an integer. Of course, for this v, w there can be no fundamental region. But also, there are elements in $\langle v, w \rangle$ of arbitrarily small nonzero norm.

Our motivation in using this concept is the following.

Lemma 6.12.15. *Let $X \subset \mathbb{R}^n$ be uniformly discrete. Let $\{x_i : i \geq 1\}$ be a sequence in X with $\lim_{i \rightarrow \infty} x_i = y \in \mathbb{R}^n$. Then there exists $N > 0$ such that $x_i = y$ for all $i \geq N$. Thus, $\{x_i : i \geq 1\}$ is eventually constant and the limit element y must lie in X .*

Proof. By hypothesis, there is an $\epsilon > 0$ such that $d(x, x') > \epsilon$ for all $x \neq x'$ in X . Since $\lim_{i \rightarrow \infty} x_i = y$, there exists $N > 0$ such that $d(x_i, y) < \frac{\epsilon}{2}$ for all $i \geq N$. But then $d(x_i, x_j) < \epsilon$ for all $i, j > N$. But this forces $x_i = x_j$ for $i, j \geq N$. This, in turn forces $x_i = y$ for $i \geq N$ \square

We deduce the following “intuitively obvious” fact.

Corollary 6.12.16. *Let $X \subset \mathbb{R}^n$ be uniformly discrete and let $x \in X$. Then there exists a closest element, y , to x in $X \setminus \{x\}$, i.e., there exists $y \neq x$ in X such that $d(x, z) \geq d(x, y)$ for all $z \neq x$ in X . Moreover, there are only finitely many such y . By abuse of language, we shall call them closest elements to x in X .*

Proof. There exists $\epsilon > 0$ such that $d(y, z) > \epsilon$ for all $y \neq z$ in X . So $S = \{d(z, x) : z \in X \setminus \{x\}\}$ is bounded below by a positive number. Let s be the greatest lower bound of the elements in S . Then $s > 0$ and for each $\delta > 0$, there is an element $z \in X \setminus \{x\}$ with $d(z, x) < s + \delta$. Thus, we can choose a sequence $\{z_i \in X \setminus \{x\} : i \geq 1\}$ such that $d(z_i, x) < s + \frac{1}{i}$.

We shall now make use of some basic results in point-set topology. (See, e.g., [8].) Note that $\{z_i : i \geq 1\}$ lies in the closed ball of radius $s + 1$ about x (i.e., $\{z \in \mathbb{R}^n : d(x, z) \leq s + 1\}$), which is *compact* by the Heine–Borel theorem. By a basic result on compactness, there is a subsequence of $\{z_i : i \geq 1\}$ converging to some $y \in \mathbb{R}^n$. In other words, after passing to a subsequence, if necessary, we can assume $\lim_{i \rightarrow \infty} z_i = y$. Note that any subset of a uniformly discrete set is uniformly discrete. Thus, by Lemma 6.12.15, $y \in X \setminus \{x\}$ and there exists $N > 0$ such that $x_i = y$ for all $i \geq N$. So $d(x, y) < s + \frac{1}{i}$ for all $i \geq N$, and hence $d(x, y) = s$.

The same argument shows there are only finitely many such y . Otherwise, we could form a sequence $\{y_i : i \geq 1\}$ of distinct such y , which would all lie in the closed ball of radius $d(x, y)$ about x . As above, we may assume the sequence converges. But uniform discreteness forces the sequence to be eventually constant, violating the assumption the y_i are distinct. \square

Definition 6.12.17. In a lattice Λ we refer to a closest element to 0 as a minimal length element (i.e., v is a minimal length element if $\|v\|$ is minimal among the norms of the nonzero elements of Λ). A shortest translation for \mathcal{T}_Λ is a translation τ_v where v is a minimal length element of Λ .

Corollary 6.12.16 shows that every lattice has minimal length elements. The same argument actually shows more.

Corollary 6.12.18. *Let v be a minimal length element of the lattice $\Lambda \subset \mathbb{R}^n$. Then there are finitely many minimal length vectors in*

$$\Lambda \setminus \{kv : 0 \neq k \in \mathbb{Z}\},$$

i.e., there are finitely many vectors $w \in \Lambda \setminus \langle v \rangle$ such that $\|w\| \leq \|z\|$ for all $z \in \Lambda \setminus \langle v \rangle$.

Also, if S is the set of minimal length vectors in Λ , then there are finitely many nonzero vectors of minimal length in $\Lambda \setminus (S \cup \{0\})$. We call them vectors of subminimal length in Λ .

Proof. $\Lambda \setminus \{kv : 0 \neq k \in \mathbb{Z}\}$ and $\Lambda \setminus S$ are uniformly discrete. Apply Corollary 6.12.16 to find the elements closest to 0. \square

Lemma 6.12.19. *Let v be a minimal length element of the lattice Λ in \mathbb{R}^n . Then $\langle v \rangle = \Lambda \cap \text{span}(v)$.*

Proof. $\text{span}(v)$ is the union over $k \in \mathbb{Z}$ of the line segment from kv to $(k+1)v$. Each of these segments has length $\|v\|$. If w lies in the interior of one of these segments, then $\|w - kv\| < \|v\|$, so w cannot lie in Λ by the minimality of $\|v\|$. \square

We can apply these ideas to find a \mathbb{Z} -basis for a lattice in \mathbb{R}^2 .

Proposition 6.12.20. *Let Λ be a lattice in \mathbb{R}^2 . Let v be a minimal length vector in Λ and let w be a vector of minimal length in $\Lambda \setminus \langle v \rangle$. Then v, w is a \mathbb{Z} -basis for Λ . We call such a basis a minimal length \mathbb{Z} -basis for Λ .*

Proof. By Lemma 6.12.19, v, w are linearly independent over \mathbb{R} , and hence form a \mathbb{Z} -basis, \mathcal{B} , for a lattice $\Lambda_{\mathcal{B}} \subset \Lambda$. We wish to show $\Lambda_{\mathcal{B}} = \Lambda$.

Since $P(v, w)$ is a fundamental region for $\Lambda_{\mathcal{B}}$, for any point $z \in \mathbb{R}^2$ there is a $y \in \Lambda_{\mathcal{B}}$ with $z - y \in P(v, w)$. Applying this to $z \in \Lambda$, it suffices to show that

$$(6.12.4) \quad P(v, w) \cap \Lambda = \{0, v, w, v + w\}.$$

So let $z \in P(v, w) \cap \Lambda$. Then $z = sv + tw$ with $s, t \in [0, 1]$. If $t = 0$, $z \in \text{span}(v) \cap \Lambda = \langle v \rangle$, and hence $z = 0$ or $z = v$. If $t = 1$, then $z - w \in \text{span}(v) \cap \Lambda = \langle v \rangle$, so $z = w$ or $z = v + w$. Thus, it suffices to assume $t \in (0, 1)$ and derive a contradiction.

For $t \in (0, 1)$, $z \notin \text{span}(v)$. By our minimality assumption on $\|w\|$, $\|z\| \geq \|w\|$. Note that the line segment \overline{vw} cuts $P(v, w)$ into two triangles, one, which we'll call T_1 with vertices $0, v, w$, and the other, T_2 , with vertices $v, w, v + w$. Each vertex of T_1 lies in the convex set $\{y \in \mathbb{R}^2 : \|y\| \leq \|w\|\}$. Since a triangle is the convex hull of its vertices, $T_1 \subset \{y \in \mathbb{R}^2 : \|y\| \leq \|w\|\}$, and hence every element in T_1 has norm less than or equal to $\|w\|$. In fact, we can be even more specific. We can identify T_1 as

$$T_1 = \{uy : y \in \overline{vw}, u \in [0, 1]\},$$

and $\|uy\| = u\|y\|$ for u and y as above, so $\|uy\| < \|w\|$ if $u < 1$. So the only elements of norm $\|w\|$ in T_1 lie in \overline{vw} . But on \overline{vw} , $\|(1-t)v + tw\|$ is a quadratic in t with positive leading coefficient, and hence attains its maximum on $[0, 1]$ precisely on one or both endpoints, so $\|y\| < \|w\|$ for all

y in the interior of \overline{vw} . In consequence, the only points of $T_1 \cap \Lambda$ are 0 , v and w .

Now, for T_2 , note that the rotation about $\frac{1}{2}(v+w)$ by π takes T_2 onto T_1 and preserves Λ , as

$$\rho_{(y,\pi)}(x) = 2y - x$$

for $x, y \in \mathbb{R}^2$. So the result follows here, also. \square

In analyzing wallpaper groups, it is useful to study the minimal length \mathbb{Z} -bases for the lattice of translations. Note that if v is a minimal length vector in Λ so is $-v$.

Proposition 6.12.21. *Let Λ be a lattice in \mathbb{R}^2 and let S be the set of minimal length vectors in Λ . Then S has one of the following forms:*

- (1) $S = \{\pm v\}$ for some v .
- (2) $S = \{\pm v, \pm w\}$, where the smallest unsigned angle θ between $\text{span}(v)$ and $\text{span}(w)$ satisfies $\frac{\pi}{3} < \theta \leq \frac{\pi}{2}$.
- (3) $S = \{\pm v, \pm w, \pm(v-w)\}$ where the smallest unsigned angle θ between $\text{span}(v)$ and $\text{span}(w)$ is $\frac{\pi}{3}$.

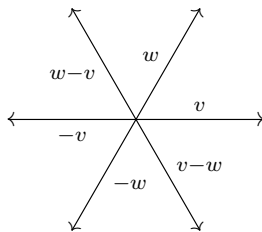
Proof. (1) is certainly a possibility, e.g., if $\mathcal{B} = e_1, 2e_2$, then $S = \{\pm e_1\}$.

Suppose there are vectors $v, w \in S$ such that the unsigned angle θ between $\overrightarrow{0v}$ and $\overrightarrow{0w}$ is less than $\frac{\pi}{3}$. Then as in the proof of the cosine law,

$$\begin{aligned} \|v - w\|^2 &= \|v\|^2 + \|w\|^2 - 2\|v\|\|w\|\cos\theta \\ &= 2\|v\|^2(1 - \cos\theta), \end{aligned}$$

as $\|v\| = \|w\|$. But $0 < \theta < \frac{\pi}{3}$, so $\frac{1}{2} < \cos\theta < 1$, and hence $\|v - w\|^2 < \|v\|^2$, contradicting the minimality of $\|v\|$. Thus, any two vectors in S not negatives of each other must determine an unsigned angle $\geq \frac{\pi}{3}$.

If $\theta = \frac{\pi}{3}$, then $\cos\theta = \frac{1}{2}$ and the same argument shows $\|v - w\| = \|v\|$, and hence $\pm(v-w) \in S$. Pictorially, we get an array of equal-length vectors in S forming the vertices of a regular hexagon.



That the angle from $\overrightarrow{0(w-v)}$ to $\overrightarrow{0(-v)}$ is $\frac{\pi}{3}$ can be seen by translating the equilateral triangle $\triangle 0vw$ by $-v$.

There cannot be any additional elements of S as that would introduce angles between rays of S that are less than $\frac{\pi}{3}$, contradicting our earlier calculation. So we are precisely in case (3). This situation is realized geometrically by $\Lambda_{\mathcal{B}}$ for $\mathcal{B} = v, w$ for this choice of v and w .

If $\theta > \frac{\pi}{2}$, i.e., if θ is obtuse, we may exchange it for the angle between the rays determined by v and $-w$, which is then acute. Thus, by renaming w if necessary, we may assume θ is acute. By our calculations above, if we rule out (3), we may assume $\frac{\pi}{3} < \theta \leq \frac{\pi}{2}$. But the angle $\pi - \theta$ between v and $-w$ is then less than $\frac{2\pi}{3}$. But then additional elements in S would produce angles between rays of S less than $\frac{\pi}{3}$, which is impossible. So if $\frac{\pi}{3} < \theta \leq \frac{\pi}{4}$, then (2) holds. \square

6.13. Orientation-preserving wallpaper groups. Recall that a wallpaper group is a two-dimensional crystallographic group, i.e., a subgroup $\mathcal{W} \subset \mathcal{I}_2$ such that the translation subgroup $\mathcal{T}(\mathcal{W})$ of \mathcal{W} is a translation lattice in \mathbb{R}^2 , i.e.,

$$\mathcal{T}(\mathcal{W}) = \langle \tau_v, \tau_w \rangle = \{ \tau_v^i \tau_w^j : i, j \in \mathbb{Z} \},$$

where v, w is a basis of \mathbb{R}^2 as a vector space over \mathbb{R} . In this section we study wallpaper groups not containing any orientation-reversing isometries.

Recall that a point of symmetry for a subgroup $H \subset \mathcal{I}_2$ is a point x whose isotropy subgroup

$$H_x = \{ \alpha \in H : \alpha(x) = x \}$$

is isomorphic to either C_n or D_{2n} for some $n > 1$. The value n is then called the period of x . A point of symmetry of period n is called an n -center for H .

By Leonardo's theorem, x is a point of symmetry whose period is divisible by n if and only if H_x is finite and $\rho(x, \frac{2\pi}{n}) \in H$.

Recall that a subset $Y \subset \mathbb{R}^n$ is uniformly discrete if there exists $\epsilon > 0$ such that $d(x, y) > \epsilon$ for all $x, y \in Y$ with $x \neq y$. The following is key.

Lemma 6.13.1. *Let \mathcal{W} be a wallpaper group and let \tilde{X}_n be the set of points of symmetry of \mathcal{W} whose period is divisible by $n > 1$. Then \tilde{X}_n is uniformly discrete. In fact, if ℓ is the shortest length of a nonidentity translation in \mathcal{W} (which exists by Lemma 6.12.13), then $d(x, y) \geq \frac{1}{2}\ell$ for all $x, y \in \tilde{X}_n$ with $x \neq y$.*

Proof. By hypothesis, $\rho(x, \frac{2\pi}{n})$ and $\rho(y, -\frac{2\pi}{n})$ lie in \mathcal{W} , hence so does their product. By Corollary 5.5.11,

$$\rho(x, \frac{2\pi}{n}) \rho(y, -\frac{2\pi}{n}) = \tau_v$$

for some $\tau_v \in \mathcal{T}(\mathcal{W})$, and hence $\|v\| \geq \ell$ by our definition of ℓ . Now,

$$\rho(x, \frac{2\pi}{n}) = \tau_v \rho(y, \frac{2\pi}{n}),$$

so

$$\rho(x, \frac{2\pi}{n})(y) = \tau_v \rho(y, \frac{2\pi}{n})(y) = \tau_v(y).$$

Since $d(x, y) = d\left(x, \rho_{\left(x, \frac{2\pi}{n}\right)}(y)\right) = d(x, \tau_v(y))$, we have an isosceles triangle with vertices x , y and $\tau_v(y)$. By the triangle inequality,

$$d(y, \tau_v(y)) \leq d(x, y) + d(x, \tau_v(y)) = 2d(x, y).$$

But $d(y, \tau_v(y)) = \|v\| \geq \ell$, and the result follows. \square

We now wish to show the period of a point of symmetry in a wallpaper group cannot be too large. In fact, the period is quite constrained, and this will make it easy to classify the orientation-preserving wallpaper groups.

Proposition 6.13.2. *Let \mathcal{W} be a wallpaper group that admits n -centers. Then $n = 2, 3, 4$ or 6 .*

Proof. Let x be an n -center for \mathcal{W} and let y be a closest n -center to x (Corollary 6.12.16), and write

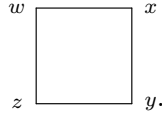
$$d = d(x, y).$$

Let $z = \rho_{\left(y, \frac{2\pi}{n}\right)}(x)$, and let $w = \rho_{\left(z, \frac{2\pi}{n}\right)}(y)$. These are again n -centers, as the set of n -centers is \mathcal{W} -invariant (Lemma 6.9.5). Moreover, we have

$$(6.13.1) \quad d = d(x, y) = d(y, z) = d(z, w),$$

as rotations are isometries.

If $n = 4$, the angles $\angle xyz$ and $\angle yzw$ are right angles. Since the distances in (6.13.1) are equal we obtain a square with vertices x, y, z, w :



If $n = 6$, the angles $\angle xyz$ and $\angle yzw$ have measure $\frac{\pi}{3}$. Since the distances in (6.13.1) are equal, we must have $x = w$, and $\triangle xyz$ is equilateral. Since $w = x$, it is not an n -center distinct from x .

Finally, we claim that if $n = 5$ or if $n \geq 7$, then $d(x, w) < d(x, y) = d$, which contradicts that y is a closest n -center to x , and hence shows there cannot be n -centers in a wallpaper group \mathcal{W} if $n = 5$ or $n \geq 7$.

Thus, assume that $n = 5$ or if $n \geq 7$. Since isometries preserve distance, we may as well rotate and translate our points so that $z = 0$ and $y = \begin{bmatrix} d \\ 0 \end{bmatrix}$.

Then by our choice of angles, $w = \begin{bmatrix} d \cos \frac{2\pi}{n} \\ d \sin \frac{2\pi}{n} \end{bmatrix}$ and $x = \begin{bmatrix} d - d \cos \frac{2\pi}{n} \\ d \sin \frac{2\pi}{n} \end{bmatrix}$, and hence

$$(6.13.2) \quad d(x, w) = \left| d \left(1 - 2 \cos \frac{2\pi}{n} \right) \right|.$$

Since $n > 4$, these angles are acute, so $0 < \cos \frac{2\pi}{n} < 1$, and hence

$$-1 < 1 - 2 \cos \frac{2\pi}{n} < 1.$$

By (6.13.2), $d(x, w) < d$. Since w is not equal to x in this case, we obtain our desired contradiction. \square

Corollary 6.13.3. *Let \mathcal{W} be a wallpaper group that admits 4-centers. Then \mathcal{W} does not admit either 3-centers or 6-centers. Thus, every point of symmetry in \mathcal{W} has period 2 or 4.*

Proof. Let x be a 4-center for \mathcal{W} and let y be either a 3-center or a 6-center. Then $\rho(x, \frac{2\pi}{4})$ and $\rho(y, \frac{2\pi}{3})$ lie in \mathcal{W} , hence so does

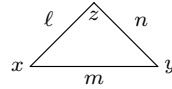
$$\rho(y, \frac{2\pi}{3})\rho(x, -\frac{2\pi}{4}) = \rho(z, \frac{2\pi}{12})$$

for some z . But this has period divisible by 12, which is ruled out by Proposition 6.13.2. \square

6.13.1. Groups admitting 4-centers. We shall make use of the following.

Lemma 6.13.4. *Let $x \neq y \in \mathbb{R}^2$. Then $\rho(x, \frac{\pi}{2})\rho(y, \frac{\pi}{2}) = \rho(z, \pi)$ where x, y, z form an isosceles right triangle with right angle at z .*

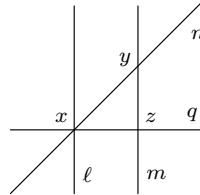
Proof. Let m be the line containing x and y . Let ℓ be the line through x such that the directed angle from m to ℓ is $\frac{\pi}{4}$ and let n be the line through y such that the directed angle from n to m is $\frac{\pi}{4}$. Then $\rho(x, \frac{\pi}{2})\rho(y, \frac{\pi}{2}) = \sigma_\ell\sigma_m\sigma_m\sigma_n = \sigma_z$,



with $z = \ell \cap n$. \square

We now show there is a unique orientation-preserving wallpaper group with a given 4-center x and a given shortest nonzero translation τ_v . First consider the composite $\tau_v\rho(x, \frac{\pi}{2})$. Using our calculus of isometries, we can compute it as $\sigma_m\sigma_\ell\sigma_\ell\sigma_n$ with ℓ the line through x perpendicular to v , $m = \ell + \frac{1}{2}v$ and n the line through x bisecting the directed angle from $q = x + \text{span}(v)$ to ℓ :

(6.13.3)

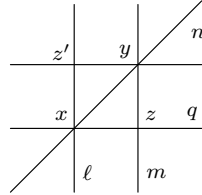


The diagram provides a 4-center, y . We also obtain a point of symmetry $z = m \cap q$, as $\tau_v\rho(x, \pi) = \sigma_m\sigma_q = \rho(z, \pi)$.

Note the distance from x to z is $\frac{1}{2}\|v\|$. Moreover, no point of symmetry can be closer to x than z is, as the period of each point of symmetry for \mathcal{W} is divisible by 2, and if $u \in \mathbb{R}^2$, then $\rho(u, \pi)\rho(x, \pi)$ is the translation by $2(u - x)$, a vector whose norm is twice the distance from x to u .

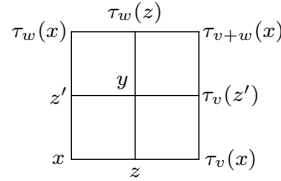
By Lemma 6.13.4, this forces z to be a 2-center: otherwise $\rho(x, \frac{\pi}{2})\rho(z, \frac{\pi}{2})$ is a rotation at a point closer to x than z is. Lemma 6.13.4 also establishes that $\rho(y, \frac{\pi}{2})\rho(x, \frac{\pi}{2}) = \rho(z, \pi)$. Similarly, $\rho(x, \frac{\pi}{2})\rho(y, \frac{\pi}{2}) = \rho(z', \pi)$, with z' in the following diagram.

(6.13.4)



Again, z' must be a 2-center, as otherwise, there is a point of symmetry closer to x than z or z' is. Additionally, $\rho(z', \pi)\rho(x, \pi)$ is the translation by $w = 2(z' - x)$, a vector orthogonal to v and having the same length as v . By Proposition 6.12.20, v, w is a \mathbb{Z} -basis for $\mathcal{T}(\mathcal{W})$, and a fundamental region, R , for $\mathcal{T}(\mathcal{W})$ is given as follows.

(6.13.5)



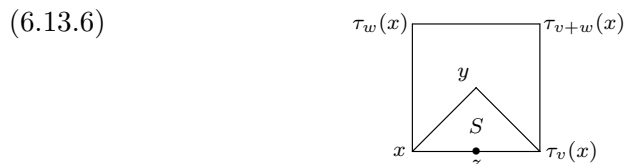
Moreover, the labelled points are the only points of symmetry in R , as any other point would be closer to one of these than is permissible, as that would violate the minimality assumption on the length of v . Since every \mathcal{T} -orbit in \mathbb{R}^2 meets the fundamental region for $\mathcal{T}(\mathcal{W})$, there are exactly two \mathcal{T} -orbits of 4-centers for \mathcal{W} : one of them represented here by y and the other by x , $\tau_v(x)$, $\tau_w(x)$ and $\tau_{v+w}(x)$, all of which lie in the same \mathcal{T} -orbit. Similarly, there are exactly two \mathcal{T} -orbits of 2-centers for \mathcal{W} , one of them represented by z and $\tau_w(z)$ and the other by z' and $\tau_v(z')$.

Thus, we have identified all the translations and rotations in \mathcal{W} . Since \mathcal{W} is orientation-preserving, we know all its elements. Note that the argument above shows that both τ_w and the rotations at all the points of symmetry in R are generated by τ_v and $\rho(x, \frac{\pi}{2})$. The same argument, applied to the translates of R by the elements of $\mathcal{T}(\mathcal{W})$, will show that τ_v and $\rho(x, \frac{\pi}{2})$ generate all of \mathcal{W} , provided we show they generate the rotations by $\frac{\pi}{2}$ at all elements in the \mathcal{T} -orbit of x . But the same argument as that given for the diagram (6.13.3) shows that

$$\begin{aligned} \tau_{k(v+w)}\rho(x, \frac{\pi}{2}) &= \rho(\tau_{kw}(x), \frac{\pi}{2}), \\ \tau_{k(v-w)}\rho(x, \frac{\pi}{2}) &= \rho(\tau_{kv}(x), \frac{\pi}{2}), \end{aligned}$$

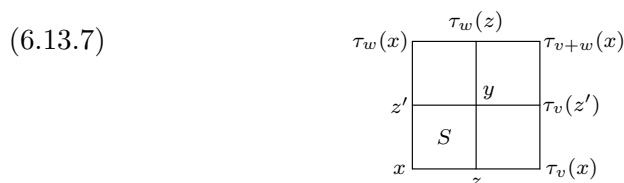
for all $k \in \mathbb{Z}$. We can then iterate this argument at all of these translates of x to get the others.

Note that the only elements of \mathcal{W} that carry interior points of \mathbb{R} to interior points of R are the rotations about y . Thus, a fundamental region for \mathcal{W} is given by the triangle, S with vertices x , y and $\tau_v(x)$:



There are two sets of identifications on S induced by elements of \mathcal{W} . The edges \overline{xy} and $\overline{y\tau_v(x)}$ are folded together via $\rho_{(y, \frac{\pi}{2})}$, joining x to $\tau_v(x)$; the edge $\overline{x\tau_v(x)}$ is folded in half via $\rho_{(z, \pi)}$ with the crease point at z . The orbit space is topologically a sphere \mathbb{S}^2 .

An alternate choice for a fundamental region S for \mathcal{W} is given by the square with vertices x , z' , y and z .



The orbit space is, of course, the same, and is instructive to see.

Note that by Theorem 5.5.20, we can conjugate \mathcal{W} via a translation and a rotation so that $x = 0$ and v lies on the positive x -axis. But conjugation will not change the length of v . Write $\mathcal{W}(r)$ for the case where $x = 0$ and $v = re_1$ with $r > 0$. Note that conjugation by the linear map induced by rI_2 takes τ_{e_1} to τ_{re_1} and preserves $\rho_{(0, \frac{\pi}{2})}$. While rI_2 does not induce an isometry, it does induce an affine isomorphism from \mathbb{R}^2 to itself, so the groups $\mathcal{W}(r)$ are all conjugate in the group of affine isomorphisms. In particular, they are isomorphic. We have shown:

Theorem 6.13.5. *For a given x and v in \mathbb{R}^2 with $v \neq 0$ there is a unique orientation-preserving wallpaper group containing $\rho_{(x, \frac{\pi}{2})}$ in which τ_v is a shortest translation. We call it \mathcal{W}_4 . It is generated by $\rho_{(x, \frac{\pi}{2})}$ and τ_v , and a fundamental region R for $\mathcal{T}(\mathcal{W}_4)$ is given by (6.13.5), with the labelled points being the only points of symmetry in R . We see there are two \mathcal{T} -orbits of 4-centers and two \mathcal{T} -orbits of 2-centers.*

A fundamental region S for \mathcal{W}_4 is given in (6.13.6). The orbit space is a sphere, \mathbb{S}^2 .

The conjugacy class of this group in \mathcal{O}_2 depends only on $\|v\|$, and as $\|v\|$ varies, these groups are isomorphic.

Example 6.13.6. A pattern whose symmetry group is \mathcal{W}_4 is given by propagating the pattern in Figure 6.13.1 across the plane. The arrow-bordered

squares are fundamental regions for $\mathcal{T}(\mathcal{W})$, with 4-centers at their vertices and centers and with 2-centers at the midpoints of the double-headed arrows. The orientations of the barbs implies there are no reflections or glide reflections that preserve this pattern, so its symmetry group must be orientation-preserving and hence be \mathcal{W}_4 .

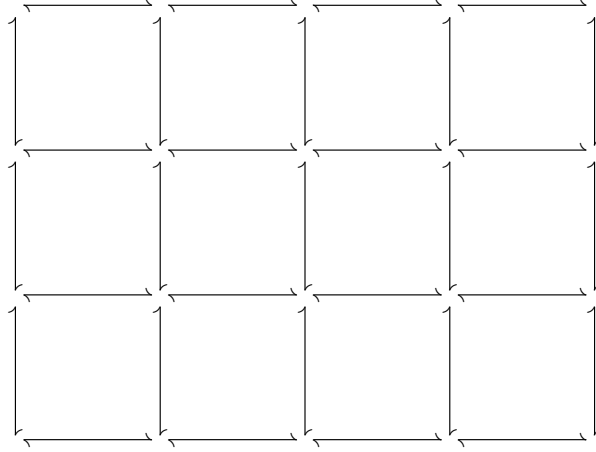


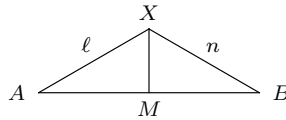
FIGURE 6.13.1. A pattern with symmetry group \mathcal{W}_4 .

6.13.2. Groups admitting 6-centers. We now give a similar uniqueness theorem for wallpaper groups that admit 6-centers. By Proposition 6.13.2 and Corollary 6.13.3, all points of symmetry in such a group have period 2, 3, or 6.

Lemma 6.13.7. *Let A be a 6-center for a wallpaper group \mathcal{W} and let B be a 6-center closest to A . Then $\rho_{(A, \frac{\pi}{3})}\rho_{(B, \frac{\pi}{3})} = \rho_{(X, \frac{2\pi}{3})}$, where X is a 3-center closer to A than B is. In fact, X is a 3-center closest to A , and the midpoint M of \overline{AB} is a 2-center. Moreover, M is a point of symmetry closest to A , and τ_{B-A} is a shortest translation in \mathcal{W} . Thus, closest 6-centers differ by a shortest translation.*

Proof. Let $m = \overleftrightarrow{AB}$. Let ℓ be the line through A such that the directed angle from m to ℓ is $\frac{\pi}{6}$ and let n be the line through B such that the directed angle from n to m is $\frac{\pi}{6}$. Then $\rho_{(A, \frac{\pi}{3})}\rho_{(B, \frac{\pi}{3})} = \sigma_\ell\sigma_m\sigma_m\sigma_n = \rho_{(X, \frac{2\pi}{3})}$ with X in the diagram below.

(6.13.8)



The period of X is divisible by 3. Since X is closer to A than B is, X cannot be a 6-center, so it must be a 3-center. If X' were a 3-center closer to A than

X is, then a similar diagram displays $\rho_{(A, \frac{\pi}{3})}\rho_{(B', \frac{\pi}{3})} = \rho_{(X', \frac{2\pi}{3})}$, where B' is a 6-center closer to A than B is. Since this is impossible, X is a 3-center closest to A .

(6.13.8) also shows that $\rho_{(X, \frac{2\pi}{3})}\rho_{(A, \frac{\pi}{3})} = \rho_{(M, \pi)}$, so M is a point of symmetry of period divisible by 2. Since it is closer to A than B is, M is a 2-center. We have $\rho_{(M, \pi)}\rho_{(A, \pi)} = \tau_{2(M-A)} = \tau_{B-A}$. Since the 6-centers are \mathcal{W} -invariant and since $\tau_{B-A}(A) = B$, this is a shortest translation in \mathcal{W} . If N were a 2-center closer to A than M is, we'd have $\rho_{(N, \pi)}\rho_{(A, \pi)} = \tau_{2(N-A)}$, a translation shorter than τ_{B-A} . So M is a point of symmetry closest to A . \square

We assume now that \mathcal{W} is a wallpaper group that admits 6-centers and that A and B are closest 6-centers in \mathcal{W} . Rotating (6.13.8) by increments of $\frac{2\pi}{3}$ about X , we obtain an equilateral triangle whose vertices are 6-centers.



Here, X is the marked point at the centroid of the triangle, and the translations $\tau_v = \tau_{B-A}$, $\tau_w = \tau_{C-A}$ and $\tau_{w-v} = \tau_{C-B}$ are shortest translations in \mathcal{W} by Lemma 6.13.7. By Proposition 6.12.20, v, w is a \mathbb{Z} -basis for $\mathcal{T}(\mathcal{W})$, and a fundamental region, R , for $\mathcal{T}(\mathcal{W})$ is given as follows.



Here, the vertices of R are all translates of A : $B = \tau_v(A)$, $C = \tau_w(A)$ and $D = \tau_{v+w}(A)$. Since $\|v\| = \|w\|$, R is a rhombus. It is the union of two equilateral triangles along the common edge \overline{BC} . The points marked \bullet are 3-centers, and occur at the centroids of the two equilateral triangles. The points marked \circ are 2-centers. The ones on the edges of R occur at the midpoints of those edges. The other is at the center point of R and occurs at the midpoint of \overline{BC} , which is also the midpoint of \overline{AD} .

Since v and w are shortest translations in \mathcal{W} , the argument of Lemma 6.13.7 shows there are no other points of symmetry in R than those marked. The 2-centers on opposite edges of R are translates of one another, so there are three \mathcal{T} -orbits of 2-centers. No interior point of R is a translate of any other interior point of R , so there are two \mathcal{T} -orbits of 3-centers. There is one \mathcal{T} -orbit of 6-centers, given by the vertices of R .

The only elements of \mathcal{W} that carry interior points of R to interior points of R are the rotations about the marked centers in the interior of R . As a result, a fundamental region S for \mathcal{W} is given by the region shown in (6.13.8):

the triangle whose vertices are A , B and the centroid of the triangle $\triangle ABC$. The orbit space is again a sphere for the same reasons as for \mathcal{W}_4 .

The argument given in Lemma 6.13.7 shows both τ_w and all the rotations at the marked points in R are generated by $\rho_{(A, \frac{\pi}{3})}$ and τ_v (or, in fact, by $\rho_{(A, \frac{\pi}{3})}$ and any of the rotations of maximal period for one of the marked points). An argument similar to that used in Theorem 6.13.5 shows that \mathcal{W} is generated by $\rho_{(A, \frac{\pi}{3})}$ and τ_v , and is the unique wallpaper group containing $\rho_{(A, \frac{\pi}{3})}$ for which τ_v is a shortest translation. Again as in Theorem 6.13.5, we obtain the following.

Theorem 6.13.8. *For a given x and v in \mathbb{R}^2 with $v \neq 0$ there is a unique orientation-preserving wallpaper group containing $\rho_{(x, \frac{\pi}{3})}$ in which τ_v is a shortest translation. We call it \mathcal{W}_6 . It is generated by $\rho_{(x, \frac{\pi}{3})}$ and τ_v , and a fundamental region R for $\mathcal{T}(\mathcal{W}_6)$ is given by (6.13.10), with the vertices and the marked points being the only points of symmetry in R . We see there are three \mathcal{T} -orbits of 2-centers (marked \circ), two \mathcal{T} -orbits of 3-centers (marked \bullet) and one \mathcal{T} -orbit of 6-centers (the vertices of R).*

A fundamental region S for \mathcal{W}_6 is given in (6.13.8). The orbit space is a sphere \mathbb{S}^2 .

The conjugacy class of this group in \mathcal{O}_2 depends only on $\|v\|$, and as $\|v\|$ varies, these groups are isomorphic.

Example 6.13.9. A pattern whose symmetry group is \mathcal{W}_6 is given in Figure 6.13.2. There are 6-centers at the center of each honeycomb cell. The double-headed arrows prevent orientation-reversing symmetries, so the group must be \mathcal{W}_6 .

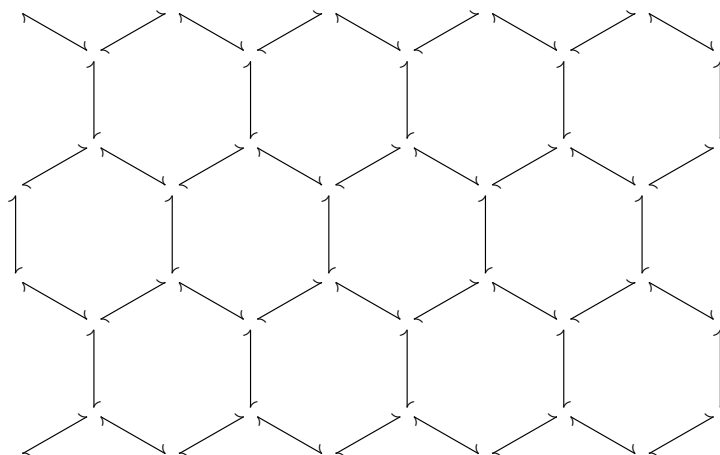


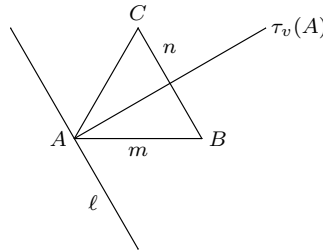
FIGURE 6.13.2. A pattern whose symmetry group is \mathcal{W}_6

6.13.3. Groups admitting 3-centers but not 6-centers. Thus, here, \mathcal{W} is a wallpaper group admitting 3-centers but not 6-centers. Since $\rho_{(x, \frac{2\pi}{3})}\rho_{(y, \pi)}$ has order 6, any wallpaper group containing both a 3-center and a 2-center must contain a 6-center. Thus, if \mathcal{W} contains 3-centers but not 6-centers, then every point of symmetry for \mathcal{W} must be a 3-center.

Lemma 6.13.10. *Let \mathcal{W} be a wallpaper whose only points of symmetry are 3-centers. Let A be a 3-center and let B be a 3-center closest to A . Then $\rho_{(A, \frac{2\pi}{3})}\rho_{(B, \frac{4\pi}{3})} = \rho_{(C, \frac{4\pi}{3})}$ where $\triangle ABC$ is equilateral, and hence C is also a closest 3-center to A .*

Moreover, $\rho_{(B, \frac{4\pi}{3})}\rho_{(A, \frac{2\pi}{3})} = \tau_v$, where $v = 2(M - A)$, with M the midpoint of \overline{BC} .

(6.13.11)



Conversely, if $v' \in \mathcal{T}(\mathcal{W})$, then an analogous diagram holds with 3-centers B' and C' in place of B and C , respectively, and with v' in place of v . If $\|v'\| < \|v\|$, then $d(A, B') < d(A, B)$. Since B is a 3-center closest to A , that forces τ_v to be a shortest translation in $\mathcal{T}(\mathcal{W})$.

Proof. That $\rho_{(A, \frac{2\pi}{3})}\rho_{(B, \frac{4\pi}{3})} = \rho_{(C, \frac{4\pi}{3})}$ with $\triangle ABC$ equilateral follows as in Lemma 6.13.7.

The composite $\rho_{(B, \frac{4\pi}{3})}\rho_{(A, \frac{2\pi}{3})}$ is a translation because the angle sum is a multiple of 2π . To calculate it, write $\overleftarrow{AB} = m$ and $\overleftarrow{BC} = n$. Then

$$\rho_{(B, \frac{4\pi}{3})}\rho_{(A, \frac{2\pi}{3})} = \sigma_n \sigma_m \sigma_m \sigma_\ell = \sigma_n \sigma_\ell,$$

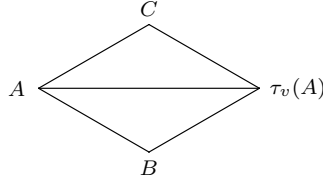
with the directed angle from ℓ to n being $\frac{\pi}{3}$, so that ℓ is as pictured in (6.13.11). Since $\ell \parallel n$, this composite is the translation by $2(q \cap n - q \cap \ell)$, where q is the perpendicular to ℓ and n containing A , i.e., with $q \cap \ell = A$. Since \overline{AB} and \overline{AC} have equal length, the proof of the Pons asinorum shows $q \cap n$ is M , the midpoint of \overline{BC} .

So $\rho_{(B, \frac{4\pi}{3})}\rho_{(A, \frac{2\pi}{3})} = \tau_v$, with v as claimed, giving $\rho_{(B, \frac{4\pi}{3})} = \tau_v \rho_{(A, \frac{4\pi}{3})}$. Given v' as in the converse, we can reverse engineer the entire diagram with B' and C' in the analogous positions, and we get $\rho_{(B', \frac{4\pi}{3})} = \tau_{v'} \rho_{(A, \frac{4\pi}{3})}$, so if $\tau_{v'} \in \mathcal{W}$, B' is a 3-center for \mathcal{W} . The result follows. \square

Note that since B and C are closer to A than $\tau_v(A)$ and since v is a shortest translation for \mathcal{W} , B and C cannot be in the \mathcal{T} -orbit of A . But

the argument for the converse shows that if A is a 3-center for \mathcal{W} and v is a shortest translation, then the points B and C obtained from the diagram

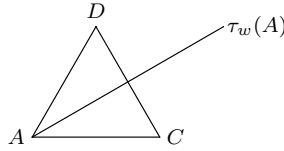
(6.13.12)



are 3-centers for \mathcal{W} . Here $\angle CA\tau_v(A)$, $\angle BA\tau_v(A)$, $\angle C\tau_v(A)A$ and $\angle B\tau_v(A)A$ are all $\frac{\pi}{6}$ as in (6.13.8).

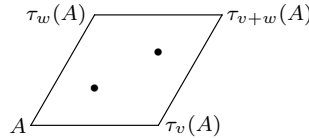
We can now repeat the argument for Lemma 6.13.10 with C in place of B , obtaining a diagram

(6.13.13)



with D a 3-center for \mathcal{W} and w a shortest translation in \mathcal{W} . Note that $w = \rho_{(0, \frac{\pi}{3})}(v)$, so v, w are linearly independent. Since both are shortest translations in \mathcal{W} , v, w is a \mathbb{Z} -basis for $\mathcal{T}(\mathcal{W})$, and generate a rhombic fundamental region R for $\mathcal{T}(\mathcal{W})$.

(6.13.14)



A more complete picture of the rotations and translations for \mathcal{W} is given in Figure 6.13.3. The 3-centers are indicated by \bullet and the solid line segments between 3-centers all represent shortest translations. Note that $D = \tau_{w-v}(B)$ so that $\tau_w(B) = \tau_v(D)$.

The fundamental region R for $\mathcal{T}(\mathcal{W})$ is the union of two equilateral triangles with 3-centers at their centroids and vertices. There are no other 3-centers in R , as that would violate the minimality of $\|v\|$. Thus, there are exactly three \mathcal{T} -orbits of 3-centers for \mathcal{W} , represented by A , B and C , respectively.

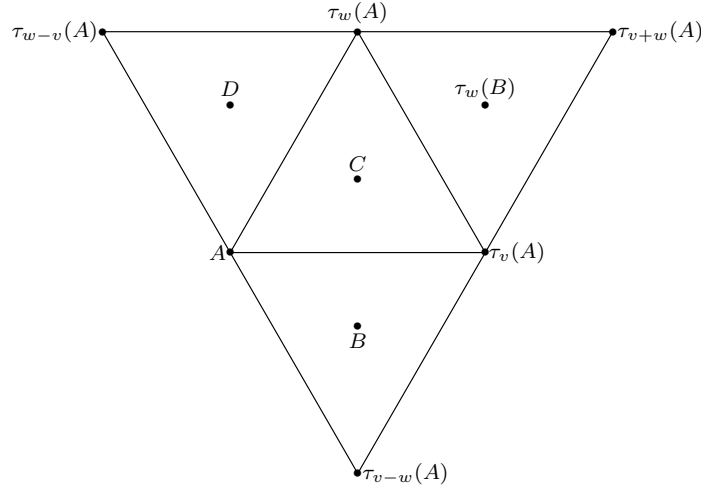
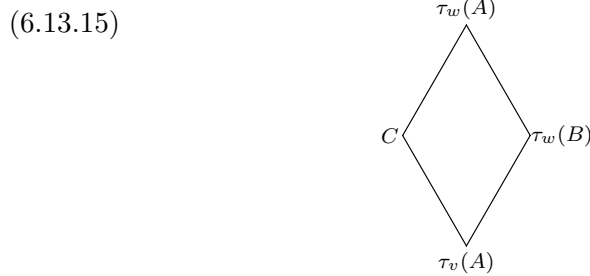


FIGURE 6.13.3. Array of 3-centers and translations for \mathcal{W}_3 .

A fundamental region S for \mathcal{W} is not obvious. But the following blowup from Figure 6.13.3 suffices.



Note we have rotated the appropriate section of Figure 6.13.3 about $\tau_w(A)$ by $-\frac{\pi}{6}$ in order to fit the grid for our graphics generator.

If we rotate this region about C by multiples of $\frac{2\pi}{3}$ it sweeps out the full circle about C and covers the lower half of R . If we rotate by multiples of $\frac{2\pi}{3}$ about $\tau_w(B)$ it covers the upper half of R . Thus, $\mathbb{R}^2 = \bigcup_{\alpha \in \mathcal{W}} \alpha(S)$. But it's easy to see that no element of \mathcal{W} carries interior points of S to interior points of S . So S is indeed a fundamental region for \mathcal{W} .

The identifications on S induced by \mathcal{W} are given by gluing the segment $\overline{C\tau_w(A)}$ to $\overline{C\tau_v(A)}$ by the rotation about C and gluing the segment $\overline{\tau_w(B)\tau_w(A)}$ to $\overline{\tau_w(B)\tau_v(A)}$ by the rotation about $\tau_w(B)$. Once again, the orbit space is topologically a sphere, \mathbb{S}^2 .

As above, we get:

Theorem 6.13.11. *For a given x and v in \mathbb{R}^2 with $v \neq 0$ there is a unique orientation-preserving wallpaper group with no 6-centers containing $\rho_{(x, \frac{2\pi}{3})}$ in which τ_v is a shortest translation. We call it \mathcal{W}_3 . It is generated by $\rho_{(x, \frac{2\pi}{3})}$*

and τ_v , and a fundamental region R for $\mathcal{T}(\mathcal{W}_3)$ is given by (6.13.14), with the vertices and marked points being the only points of symmetry in R . The only points of symmetry are 3-centers, of which there are three \mathcal{T} -orbits: one given by the vertices of R and the other two by the marked points.

A fundamental region S for \mathcal{W}_3 is given in (6.13.15). The orbit space is a sphere.

The conjugacy class of this group in \mathcal{O}_2 depends only on $\|v\|$, and as $\|v\|$ varies, these groups are isomorphic.

Example 6.13.12. A pattern with symmetry group \mathcal{W}_3 is given in Figure 6.13.4.

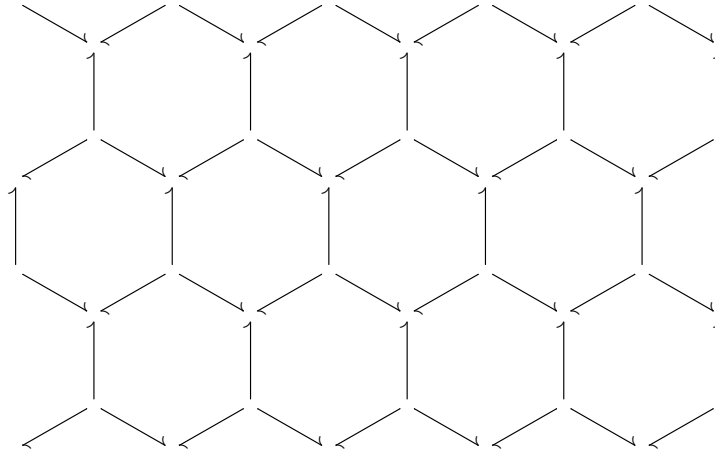


FIGURE 6.13.4. A pattern with symmetry group \mathcal{W}_3 .

6.13.4. The remaining cases. In the cases we've studied so far, the presence of rotations of certain periods determines the structure of the translation lattice of \mathcal{W} . If there are 4-centers, there is a \mathbb{Z} -basis consisting of orthogonal vectors of equal length. If there are 3-centers, the translation lattice has a \mathbb{Z} -basis consisting of vectors of equal length and forming an angle of $\frac{\pi}{3}$ with one another.

In the remaining cases, the translation lattice $\mathcal{T}(\mathcal{W})$ can be arbitrary. This introduces additional complication. There are exactly two cases remaining:

- All points of symmetry have period 2.
- There are no points of symmetry and $\mathcal{W} = \mathcal{T}(\mathcal{W})$ is a translation lattice.

Let us deal with the latter case first, as that will shed light on the former. Thus, we assume $\mathcal{W} = \mathcal{T}(\mathcal{W})$. By Theorem 5.5.20 and our calculation of $\text{SO}(2)$, the conjugacy class of $\mathcal{T}(\mathcal{W})$ in \mathcal{O}_2 determines and is determined by the following data.

- (1) The length of a shortest translation τ_v .
- (2) The length of a shortest translation τ_w with w not in $\text{span}(v)$.
- (3) The shortest possible directed angle from v to w among pairs v and w as above.

For conjugacy in \mathcal{I}_2 , replace the directed angle by the unsigned angle in (3). Thus, many of these translation lattices are geometrically distinct, even if you fix the length of the shortest translation.

On the other hand, by Corollary 6.12.4, any two translation lattices are conjugate in \mathcal{A}_2 , so they are linearly equivalent, just not geometrically. We shall refer to a wallpaper group for which $\mathcal{W} = \mathcal{T}(\mathcal{W})$ as \mathcal{W}_1 .

A fundamental region for \mathcal{W}_1 is simply the fundamental region R for $\mathcal{T}(\mathcal{W})$ in any of these examples: the parallelogram with vertices x , $\tau_v(x)$, $\tau_w(x)$ and $\tau_{v+w}(x)$ for any $x \in \mathbb{R}^2$ and for v and w satisfying (1) and (2) above:

$$(6.13.16) \quad \begin{array}{c} \tau_v(x) \qquad \qquad \tau_{v+w}(x) \\ \diagdown \qquad \qquad \diagup \\ x \qquad \qquad \qquad \tau_w(x) \end{array}$$

The orbit space is obtained by identifying the left edge with the right edge via τ_w and identifying the bottom edge with the top edge via τ_v . The result looks a bit twisted when v and w are not perpendicular, but as shown in Corollary 6.12.7, the linear map T_{A_B} induces a linear isomorphism from the fundamental region for the standard lattice onto the fundamental region here (when we take $x = 0$) respecting the boundary identifications used to construct the orbit space. For the standard lattice, the fundamental region is the unit square. The identifications of the left and right edges can be seen to form a cylinder. Identifying the top and bottom then creates a figure called a torus (denoted T^2) which is topologically equivalent to the surface of a doughnut. We have shown:

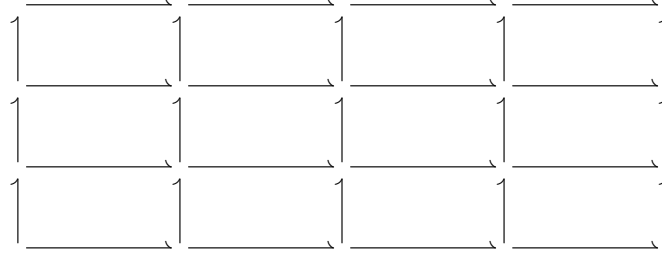
Theorem 6.13.13. *An orientation-preserving wallpaper group with no points of symmetry is a translation lattice. Its fundamental region is a parallelogram as in (6.13.16) and its orbit space is a torus. Any two such groups are linearly conjugate, but not necessarily equivalent geometrically. We call them \mathcal{W}_1 .*

Example 6.13.14. Figure 6.13.5 displays a pattern whose symmetry group is \mathcal{W}_1 .

There is one remaining case for orientation-preserving wallpaper groups. We let \mathcal{W} be an orientation-preserving wallpaper group whose only points of symmetry are 2-centers. Note that if x and y are 2-centers, then

$$\rho_{(y,\pi)}\rho_{(x,\pi)} = \tau_{2(y-x)} \in \mathcal{T}(\mathcal{W}).$$

But then $\rho_{(y,\pi)} = \tau_{2(y-x)}\rho_{(x,\pi)}$ lies in the right coset $\mathcal{T}(\mathcal{W})\rho_{(x,\pi)}$, so $\mathcal{T}(\mathcal{W})$ has index 2 in \mathcal{W} . Thus, if v, w is a \mathbb{Z} -basis for the lattice inducing $\mathcal{T}(\mathcal{W})$, then τ_v, τ_w and $\rho_{(x,\pi)}$ generate \mathcal{W} .

FIGURE 6.13.5. A pattern whose symmetry group is \mathcal{W}_1 .

Note that for any $\tau_z \in \mathcal{T}(\mathcal{W})$, τ_z and $\rho_{(x,\pi)}$ generate the frieze group $\mathcal{F}_2(z, x)$ defined in Remark 6.10.4. Its elements are

$$\mathcal{F}_2(z, x) = \{\tau_{kz}, \tau_{kz}\rho_{(x,\pi)} : k \in \mathbb{Z}\} = \{\tau_{kz}, \rho_{(x+\frac{k}{2}z,\pi)} : k \in \mathbb{Z}\}.$$

The multiplication is given by the conjugation formula

$$(6.13.17) \quad \rho_{(x,\pi)}\tau_u\rho_{(x,\pi)}^{-1} = \tau_{-u}$$

for any $x, u \in \mathbb{R}^2$.

Thus, a rotation $\rho_{(x,\pi)}$ together with a translation do not generate a wallpaper group: the subgroup they generate in \mathcal{I}_2 has a translation subgroup isomorphic to \mathbb{Z} , not $\mathbb{Z} \times \mathbb{Z}$. So three generators are necessary for \mathcal{W} : two for $\mathcal{T}(\mathcal{W})$ and one rotation.

As above, if τ_z and $\rho_{(x,\pi)}$ are in \mathcal{W} , so is $\rho_{(x+\frac{1}{2}z,\pi)}$, the rotation by π about the midpoint of the segment $\overline{x\tau_z(x)}$. Thus, if v, w form a \mathbb{Z} -basis \mathcal{B} for the lattice inducing $\mathcal{T}(\mathcal{W})$ and if x is a 2-center for \mathcal{W} , then the marked points are 2-centers in the following diagram for a fundamental region R for $\mathcal{T}(\mathcal{W})$.

$$(6.13.18) \quad \begin{array}{c} \tau_v(x) \quad \bullet \quad \bullet \quad \tau_{v+w}(x) \\ \diagup \quad \bullet \quad \bullet \quad \diagdown \\ x \quad \bullet \quad \bullet \quad \tau_w(x) \end{array}$$

Here, the marked point in the center is $x + \frac{1}{2}(v + w)$, the midpoint of the diagonal $\overline{x\tau_{v+w}(x)}$. It coincides with the midpoint of the other diagonal $\overline{\tau_v(x)\tau_w(x)}$.

Note that there cannot be any other 2-centers in R , as that would induce a new translation not in $\Lambda_{\mathcal{B}} = \mathcal{T}(\mathcal{W})$.

The only element of \mathcal{W} that carries interior points of R to interior points of R is the rotation by π about the center point of R . So a fundamental region S for \mathcal{W} is given by the left half of R :

$$(6.13.19) \quad \begin{array}{c} \tau_v(x) \quad \bullet \quad \bullet \quad \tau_{v+w}(x) \\ \diagup \quad S \quad \bullet \quad \diagdown \\ x \quad \bullet \quad \bullet \quad \tau_w(x) \end{array}$$

The identifications on the boundary of S given by \mathcal{W} are obtained by folding each of the right and left edges in half along the midpoint and identifying the top and bottom edges by τ_v . The result is sewn up completely around the boundary, topologically forming a sphere, \mathbb{S}^2 .

Conversely, if we start with a translation lattice \mathcal{T}_Λ and an $x \in \mathbb{R}^2$, then

$$(6.13.20) \quad \mathcal{W}_2(\Lambda, x) = \left\{ \tau_z, \tau_z \rho_{(x, \pi)} : z \in \Lambda \right\} = \left\{ \tau_z, \rho_{(x + \frac{1}{2}z, \pi)} : z \in \Lambda \right\}$$

forms a orientation-preserving subgroup of \mathcal{I}_2 with $\mathcal{T}(\mathcal{W}) = \mathcal{T}_\Lambda$ and with all points of symmetry of period 2. (That $\mathcal{W}_2(\Lambda, x)$ is closed under multiplication in \mathcal{I}_2 follows from (6.13.17). It is closed under inverses follows from the last expression in (6.13.20).) We have shown:

Theorem 6.13.15. *For any translation lattice \mathcal{T}_Λ and any $x \in \mathbb{R}^2$, there is a unique orientation-preserving wallpaper group \mathcal{W} whose points of symmetry all have period 2 such that $\mathcal{T}(\mathcal{W}) = \mathcal{T}_\Lambda$ and x is a 2-center. The points of symmetry for \mathcal{W} are precisely the points $\tau_{\frac{1}{2}z}(x)$ with $\tau_z \in \mathcal{T}(\mathcal{W})$.*

\mathcal{W} is generated by τ_v, τ_w and $\rho_{(x, \pi)}$ for any \mathbb{Z} -basis v, w of Λ . A fundamental region R for $\mathcal{T}(\mathcal{W})$ is given in (6.13.18) with the marked points being the only points of symmetry in R . A fundamental region S for \mathcal{W} is given by the left half of R as in (6.13.19). The orbit space is topologically a sphere, \mathbb{S}^2 .

Two such groups are conjugate in \mathcal{O}_2 or \mathcal{I}_2 if and only if their translation lattices are, but any two such groups are conjugate in \mathcal{A}_2 . We call them \mathcal{W}_2 .

Example 6.13.16. A pattern whose symmetry group is \mathcal{W}_2 is given in Figure 6.13.6.

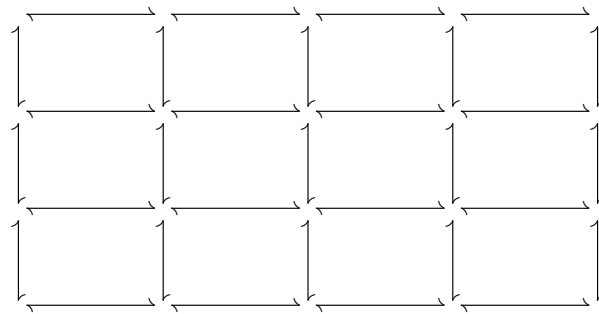


FIGURE 6.13.6. A pattern whose symmetry group is \mathcal{W}_2 .

6.14. General wallpaper groups. Given an orientation-preserving wallpaper group \mathcal{W}_0 we study how to add orientation-reversing isometries to \mathcal{W}_0 to obtain a wallpaper group \mathcal{W} whose orientation-preserving subgroup $\mathcal{O}(\mathcal{W})$ is equal to \mathcal{W}_0 .

Note that by Corollary 6.6.12 this forces \mathcal{W}_0 to have index 2 in \mathcal{W} , so that if $\alpha \in \mathcal{W} \setminus \mathcal{W}_0$, then $\mathcal{W} \setminus \mathcal{W}_0$ is precisely equal to the right coset

$$\mathcal{W}_0\alpha = \{\beta\alpha : \beta \in \mathcal{W}_0\}.$$

Index 2 subgroups are always normal, so $\alpha\mathcal{W}_0\alpha^{-1} = \mathcal{W}_0$.

The following is elementary group theory.

Lemma 6.14.1. *Let \mathcal{W}_0 be an orientation-preserving wallpaper group and let α be an orientation-reversing isometry. Then $\mathcal{W} = \mathcal{W}_0 \cup \mathcal{W}_0\alpha$ is a wallpaper group if and only if the following hold:*

- (1) $\alpha\mathcal{W}_0\alpha^{-1} = \mathcal{W}_0$.
- (2) $\alpha^2 \in \mathcal{W}_0$.

Proof. If $\mathcal{W}_0 \cup \mathcal{W}_0\alpha$ is a group, then (1) is a consequence of the normality of index 2 subgroups, while (2) is a consequence of the index 2 property, as if $\alpha^2 = \beta\alpha$ with $\beta \in \mathcal{W}_0$, then $\alpha = \beta \in \mathcal{W}_0$, which we have ruled out. Thus α^2 must lie in \mathcal{W}_0 .

Conversely, suppose (1) and (2) hold. Then (1) implies $\mathcal{W}_0 \cup \mathcal{W}_0\alpha$ is closed under multiplication. The key case is that for $\beta_1, \beta_2 \in \mathcal{W}_0$,

$$\beta_1\alpha\beta_2 = \beta_1(\alpha\beta_2\alpha^{-1})\alpha = \beta_1\beta_3\alpha$$

for some $\beta_3 \in \mathcal{W}_0$ by (1). To see that $\mathcal{W}_0 \cup \mathcal{W}_0\alpha$ is closed under inverses, the key case is that $(\beta\alpha)^{-1} = \alpha^{-1}\beta^{-1}$ and $\alpha^{-1} = \alpha^{-2}\alpha$ is in $\mathcal{W}_0\alpha$ by (1). Thus, $\mathcal{W} = \mathcal{W}_0 \cup \mathcal{W}_0\alpha$ is a subgroup of \mathcal{I}_2 . \square

We now see how to apply this to the situation at hand. First, we are assuming α is orientation-reversing. So $\alpha = \tau_x\sigma_{\ell_\phi}$ for some line ℓ_ϕ through the origin and some $x \in \mathbb{R}^2$. If $x \perp \ell_\phi$, α is then a reflection with axis parallel to ℓ_ϕ ; otherwise, α is a glide reflection with axis parallel to ℓ_ϕ . Since \mathcal{W}_0 is a wallpaper group, $\mathcal{T}(\mathcal{W}_0) = \mathcal{T}_\Lambda$ for some lattice $\Lambda \subset \mathbb{R}^2$.

Lemma 6.14.2. *Let \mathcal{W} be a wallpaper group with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_0$. Let α be an orientation-reversing isometry in \mathcal{W} , and write $\alpha = \tau_x\sigma_{\ell_\phi}$ with ℓ_ϕ a line through the origin. Let $\mathcal{T}(\mathcal{W}_0) = \mathcal{T}_\Lambda$ for the lattice $\Lambda \subset \mathbb{R}^2$. Then $\alpha\mathcal{T}_\Lambda\alpha^{-1} = \mathcal{T}_\Lambda$. Moreover,*

$$(6.14.1) \quad \alpha\tau_v\alpha^{-1} = \tau_{\sigma_{\ell_\phi}(v)},$$

so $\sigma_{\ell_\phi} : \Lambda \rightarrow \Lambda$ is a bijective group homomorphism, and, as an isometry, preserves the norm.

In particular, if S is the set of nonzero vectors of minimal length in Λ , then $\sigma_{\ell_\phi} : S \rightarrow S$ is bijective. The analogous result holds for the set of vectors T in Λ whose norm is the next size up.

Proof. Theorem 5.5.20 gives (6.14.1). Since $\mathcal{W}_0 \triangleleft \mathcal{W}$, $\tau_{\sigma_{\ell_\phi}(v)} \in \mathcal{W}_0$. But it is also in \mathcal{T}_2 , and $\mathcal{W}_0 \cap \mathcal{T}_2 = \mathcal{T}(\mathcal{W}_0)$ by definition, and this in turn is equal to \mathcal{T}_Λ . So $\alpha\mathcal{T}_\Lambda\alpha^{-1} \subset \mathcal{T}_\Lambda$. But α^{-1} is also orientation-reversing, so the same argument shows $\alpha^{-1}\mathcal{T}_\Lambda\alpha \subset \mathcal{T}_\Lambda$. Conjugating this by α gives $\mathcal{T}_\Lambda \subset \alpha\mathcal{T}_\Lambda\alpha^{-1}$.

Thus, the map $c_\alpha : \mathcal{T}_\Lambda \rightarrow \mathcal{T}_\Lambda$ given by conjugating by α is a bijection. The map $\nu : \Lambda \rightarrow \mathcal{T}_\Lambda$ given by $\nu(v) = \tau_v$ is an isomorphism. (6.14.1) says the following diagram commutes:

$$\begin{array}{ccc} \Lambda & \xrightarrow[\cong]{\nu} & \mathcal{T}_\Lambda \\ \sigma_{\ell_\phi} \downarrow & & \downarrow c_\alpha \cong \\ \Lambda & \xrightarrow[\cong]{\nu} & \mathcal{T}_\Lambda, \end{array}$$

so $\sigma_{\ell_\phi} : \Lambda \rightarrow \Lambda$ is a bijective. □

Of course, we can use S and T to find a \mathbb{Z} -basis for Λ . And knowing S and T will tell us which possible α can arise in this context.

We will now sharpen Lemma 6.14.1.

Proposition 6.14.3. *Let \mathcal{W}_0 be an orientation-preserving wallpaper group with $\mathcal{T}(\mathcal{W}_0) = \mathcal{T}_\Lambda$. Let α be an orientation-reversing isometry. Write $\alpha = \tau_x \sigma_{\ell_\phi}$ with ℓ_ϕ a line through the origin. Then $\mathcal{W} = \mathcal{W}_0 \cup \mathcal{W}_0 \alpha$ is a wallpaper group if and only if the following hold:*

- (1) *There is a \mathbb{Z} -basis v, w for Λ such that $\sigma_{\ell_\phi}(v), \sigma_{\ell_\phi}(w)$ is also a \mathbb{Z} -basis for Λ .*
- (2) *For all $y \in \mathbb{R}^2$, $\alpha(y)$ and y have the same isotropy group under \mathcal{W}_0 , i.e., if y is an n -center for \mathcal{W}_0 , then so are $\alpha(y)$ and $\alpha^{-1}(y)$.*
- (3) *If α is a glide reflection, then $\alpha^2 \in \mathcal{T}_\Lambda$.*

Proof. Suppose $\mathcal{W} = \mathcal{W}_0 \cup \mathcal{W}_0 \alpha$ is a wallpaper group. Then (1) follows from Lemma 6.14.2, while (2) follows from the fact that each y and $\alpha(y)$ must have the same isotropy subgroup under the action of \mathcal{W} . But isotropy subgroups for wallpaper groups are either cyclic or dihedral, so their orientation-preserving subgroups must be isomorphic. Finally, if α is a glide reflection, then α^2 is a translation, which must lie in \mathcal{W}_0 by Lemma 6.14.1.

Conversely, suppose (1)–(3) hold. Then conjugation by α induces an isomorphism from \mathcal{T}_Λ to itself by (6.14.1). But every element of \mathcal{W}_0 not in \mathcal{T}_Λ has the form $\rho_{(y,\theta)}$ for some θ , and $\alpha \rho_{(y,\theta)} \alpha^{-1} = \rho_{(\alpha(y), \pm\theta)}$ by Theorem 5.5.20. By (2), conjugation by α induces a bijection from \mathcal{W}_0 to itself. The result now follows from Lemma 6.14.1. □

Some other useful general results are as follows.

Lemma 6.14.4. *Let ℓ be a line of symmetry for a wallpaper group \mathcal{W} . Then there are infinitely many lines of symmetry for \mathcal{W} parallel to ℓ . The directed distance between a closest pair of such lines is $\pm \frac{1}{2}v$, where τ_v is the shortest translation perpendicular to ℓ .*

Proof. By Theorem 5.5.20, $\alpha \sigma_\ell \alpha^{-1} = \sigma_{\alpha(\ell)}$ for all $\alpha \in \mathcal{I}_2$. In particular, if ℓ is a line of symmetry for \mathcal{W} and if $\alpha \in \mathcal{W}$, then $\alpha(\ell)$ is a line of symmetry for \mathcal{W} .

Since $\mathcal{T}(\mathcal{W})$ is a lattice, there exists a translation $\tau_z \in \mathcal{T}(\mathcal{W})$ with z not parallel to ℓ . So $\tau_z(\ell) \neq \ell$, and is a line of symmetry for \mathcal{W} . But any translation of ℓ is parallel to ℓ so there exist lines of symmetry for \mathcal{W} other than ℓ that are parallel to ℓ . In fact, infinitely many of them, as the translates $\tau_{kz}(\ell)$ with $k \in \mathbb{Z}$ are all distinct.

Let m and n be distinct lines of symmetry for \mathcal{W} parallel to ℓ . Then $\sigma_m\sigma_n$ is the translation by twice the directed distance from n to m . The directed distance is perpendicular to n and hence to ℓ . Conversely, if w is perpendicular to ℓ , then $\tau_w\sigma_n$ is the reflection in $\tau_{\frac{1}{2}w}(n)$.

Since any subset of $\mathcal{T}(\mathcal{W})$ is uniformly discrete, there is a shortest translation τ_v perpendicular to ℓ , and for each line of symmetry n parallel to ℓ , $\tau_{\frac{1}{2}v}(n)$ must be the line of symmetry closest to n in the direction of v . \square

We can now investigate the relationship between these lines of symmetry and the points of symmetry of even period.

Proposition 6.14.5. *Let ℓ be a line of symmetry for a wallpaper group \mathcal{W} . Let x be a point of symmetry for \mathcal{W} with even period such that x does not lie on a line of symmetry parallel to ℓ . Then x lies exactly half way between two closest lines of symmetry parallel to ℓ , i.e., if τ_v is a shortest translation in $\mathcal{T}(\mathcal{W})$ perpendicular to ℓ , then x lies on a line $\tau_{\frac{1}{4}v}(m)$ where m is a line of symmetry parallel to ℓ .*

Proof. By Lemma 6.14.4 there is a line of symmetry m parallel to ℓ such that x lies on $\tau_{tv}(m)$ for $t \in (0, \frac{1}{2})$. Let n be the line through x perpendicular to ℓ . Then

$$\rho_{(x,\pi)}\sigma_m = \sigma_n\sigma_{(\tau_{tv}(m))}\sigma_m = \sigma_n\tau_{2tv},$$

a glide reflection with axis n . Thus, $(\rho_{(x,\pi)}\sigma_m)^2 = \tau_{4tv}$ is a translation perpendicular to ℓ . But this forces $t = \frac{1}{4}$. \square

We wish to extend these results to glide reflections.

Definition 6.14.6. Let $\gamma = \tau_z\sigma_\ell$ be a glide reflection in standard form (i.e., $z \parallel \ell$). We say γ is an essential glide reflection for a wallpaper group \mathcal{W} if $\gamma \in \mathcal{W}$, but τ_z is not in \mathcal{W} ; equivalently, σ_ℓ is not in \mathcal{W} . In this case, we say ℓ is an essential glide axis for \mathcal{W} , i.e., ℓ is the axis for a glide reflection in \mathcal{W} , but is not a line of symmetry for \mathcal{W} .

For $\gamma = \tau_z\sigma_\ell$ as above, $\gamma^2 = \tau_{2z}$ is in \mathcal{W} , so τ_z is a square root of a translation in \mathcal{W} .

We say a glide reflection γ is inessential for \mathcal{W} if it is not essential. In this case, it is the composite of two elements of \mathcal{W} , τ_z and σ_ℓ .

We say γ is primitive for \mathcal{W} if it is essential and τ_{2z} is a shortest translation in \mathcal{W} parallel to ℓ .

The following is elementary but useful.

Lemma 6.14.7. *Let $\gamma_1 = \tau_z\sigma_\ell$ be a glide reflection in standard form, and suppose γ_1 is an essential glide reflection for a wallpaper group \mathcal{W} . Then there is a glide reflection γ , primitive for \mathcal{W} , with the property that the set of all glide reflections in \mathcal{W} with axis ℓ is $\{\gamma^{2k+1} = \tau_{(k+\frac{1}{2})v}\sigma_\ell : k \in \mathbb{Z}\}$, where τ_v is a shortest translation in \mathcal{W} parallel to ℓ .*

Proof. Since any subset of $\mathcal{T}(\mathcal{W})$ is uniformly discrete, there is a shortest translation τ_v parallel to ℓ . But this implies that the set of translations in \mathcal{W} parallel to ℓ is precisely $\langle \tau_v \rangle = \{\tau_v^k : k \in \mathbb{Z}\} = \{\tau_{kv} : k \in \mathbb{Z}\}$. This is because the translations parallel to ℓ in \mathcal{I}_2 are $\{\tau_w : w \in \text{span}(v)\}$, and this forces w to lie in the line segment from kv to $(k+1)v$ for some $k \in \mathbb{Z}$. But if $\tau_w \in \mathcal{W}$ and if τ_v is a shortest translation in \mathcal{W} lying in $\text{span}(v)$, this forces w to be an endpoint of this segment. Otherwise, $w - kv$ gives rise to a translation in \mathcal{W} parallel to ℓ and shorter than τ_v .

For any $z \parallel \ell$, τ_z and σ_ℓ commute. So $(\tau_z\sigma_\ell)^2 = \tau_z^2\sigma_\ell^2 = \tau_{2z}$, as $\sigma_\ell^2 = \text{id}$. So $(\tau_z\sigma_\ell)^{2k} = \tau_{2kz}$ and $(\tau_z\sigma_\ell)^{2k+1} = \tau_{(2k+1)z}\sigma_\ell$.

For $\gamma_1 = \tau_z\sigma_\ell$ as in the statement, $\gamma_1^2 = \tau_{2z}$ is in \mathcal{W} , so $2z = kv$ for some k , and k must be odd, as otherwise $\tau_z \in \mathcal{W}$. Let $k = 2r + 1$. then $\gamma = \gamma_1\tau_{-rv} = \tau_{\frac{1}{2}v}\sigma_\ell$ is the desired primitive glide reflection for \mathcal{W} . \square

Lemma 6.14.8. *Let γ be an essential glide reflection for a wallpaper group \mathcal{W} , and let ℓ be the axis of γ . Then there are infinitely many essential glide axes parallel to ℓ . The directed distance between a closest pair of such axes is $\pm\frac{1}{2}v$, where τ_v is the shortest translation perpendicular to ℓ .*

Proof. The proof is similar to that of Lemma 6.14.4. Let w be a shortest translation in \mathcal{W} parallel to ℓ . By the proof of Lemma 6.14.7, we may assume $\gamma = \tau_{\frac{1}{2}w}\sigma_\ell$ is primitive for \mathcal{W} . Let $\tau_z \in \mathcal{W}$ with z not parallel to ℓ . Then

$$\begin{aligned}\tau_z\gamma\tau_z^{-1} &= (\tau_z\tau_{\frac{1}{2}w}\tau_z^{-1})(\tau_z\sigma_\ell\tau_z^{-1}) \\ &= \tau_{\frac{1}{2}w}\sigma_{\tau_z(\ell)},\end{aligned}$$

a primitive glide reflection whose axis is parallel to ℓ . Let $\gamma_1 = \tau_{\frac{1}{2}w}\sigma_m$ and $\gamma_2 = \tau_{\frac{1}{2}w}\sigma_n$ be primitive glide reflections with axes parallel to ℓ . Since, $\tau_{\frac{1}{2}w}$ commutes with the reflection in any line parallel to w ,

$$\gamma_2\gamma_1 = \tau_w\sigma_n\sigma_m = \tau_w\tau_u,$$

where u is twice the directed distance from m to n . In particular, $\tau_u \in \mathcal{W}$, and $u \perp \ell$. The result now follows precisely as in Lemma 6.14.4. \square

As we shall see, it is possible that there are essential glide axes parallel to lines of symmetry. In this case, the following holds.

Proposition 6.14.9. *Suppose \mathcal{W} has an essential glide axis ℓ parallel to a line of symmetry n . Then the line of symmetry is half way between two closest essential glide axes parallel to ℓ .*

Proof. Let w be a shortest translation parallel to ℓ and let v be a shortest translation perpendicular to ℓ . Since no essential glide axis is a line of symmetry, there is an essential glide axis m parallel to ℓ such that $n = \tau_{tv}(m)$ for $t \in (0, \frac{1}{2})$.

Now, $\gamma = \tau_{\frac{1}{2}w}\sigma_m$ is a primitive glide reflection with axis m , and

$$\sigma_n\gamma = \tau_{\frac{1}{2}w}\sigma_n\sigma_m = \tau_{\frac{1}{2}w}\tau_{2tv} \in \mathcal{W}.$$

Since $\tau_{\frac{1}{2}w} \notin \mathcal{W}$, $\tau_{2tv} \notin \mathcal{W}$. But $\tau_{\frac{1}{2}w}^2 \in \mathcal{W}$, so $\tau_{4tv} = \tau_{2tv}^2 \in \mathcal{W}$, so $t = \frac{1}{4}$. \square

The next result will follow as in Proposition 6.14.5.

Proposition 6.14.10. *Let ℓ be an axis for an essential glide reflection for a wallpaper group \mathcal{W} . Let x be a point of symmetry for \mathcal{W} with even period such that x does not lie on an essential glide axis parallel to ℓ . Then x lies exactly half way between two closest such axes, i.e., if τ_v is a shortest translation in $\mathcal{T}(\mathcal{W})$ perpendicular to ℓ , then x lies on a line $\tau_{\frac{1}{4}v}(m)$ where $m \parallel \ell$ is the axis of an essential glide reflection in \mathcal{W} .*

Moreover, there is a line p orthogonal to ℓ that is the axis for an essential glide reflection with glide $\tau_{\frac{1}{2}v}$, and x lies on $\tau_{\frac{1}{4}w}(p)$, where w is the shortest translation in \mathcal{W} parallel to ℓ .

Proof. As in the proof of Proposition 6.14.5, we may assume $x \in \tau_{tv}(m)$ where $t \in (0, \frac{1}{2})$ and $m \parallel \ell$ is the axis of a glide reflection in \mathcal{W} . In particular, if w is a shortest translation parallel to ℓ , then $\gamma = \tau_{\frac{1}{2}w}\sigma_m$ is a glide reflection in \mathcal{W} . Let $n = x + \text{span}(v)$, the line through x perpendicular to ℓ . Then

$$\rho(x,\pi)\gamma = \sigma_n\sigma_{\tau_{tv}(m)}\tau_{\frac{1}{2}w}\sigma_m = \sigma_n\tau_{\frac{1}{2}w}\sigma_{\tau_{tv}(m)}\sigma_m = \sigma_n\tau_{\frac{1}{2}w}\tau_{2tv} = \sigma_p\tau_{2tv},$$

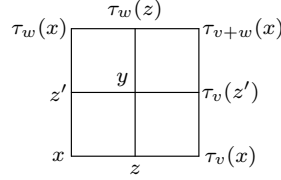
where $p = n - \frac{1}{4}v = \tau_{-\frac{1}{4}v}(n)$. This is a glide reflection with axis p , so its square is τ_{4tv} , a translation in \mathcal{W} perpendicular to ℓ . Thus, $t = \frac{1}{4}$ and x is half way between m and $\tau_{\frac{1}{2}v}(m)$ as claimed. Finally, we have the glide reflection

$$(6.14.2) \quad \gamma' = \rho(x,\pi)\gamma = \sigma_p\tau_{\frac{1}{2}v}$$

whose axis, p , is orthogonal to ℓ , and $n = \tau_{\frac{1}{4}w}(p)$, as claimed. \square

6.14.1. Wallpaper groups with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_4$. Let \mathcal{W} be a wallpaper group with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_4$ and let $\alpha \in \mathcal{W} \setminus \mathcal{W}_4$. The nonzero translation vectors in \mathcal{W}_4 of minimal length form a set $S = \{\pm v, \pm w\}$, where $v \perp w$ and $\|v\| = \|w\|$. So S is the vertex set for a square X centered at the origin. Let $\alpha = \tau_x\sigma_{\ell_\phi}$. By Lemma 6.14.2, σ_{ℓ_ϕ} is one of the reflections in the dihedral group $\mathcal{S}(X) \cong D_8$, so $\ell_\phi = \text{span}(v)$, $\ell_\phi = \text{span}(w)$, $\ell_\phi = \text{span}(v+w)$ or $\ell_\phi = \text{span}(v-w)$. Thus, the axis of α must be parallel to one of v , w , $v+w$ or $v-w$.

As shown in (6.13.5), a fundamental region R for $\mathcal{T}(\mathcal{W}_4)$ is given by



for any $x \in \mathbb{R}^2$, and if x is a 4-center, the points of symmetry in \mathbb{R}^2 are the 4-centers x, y and their translates, and the 2-centers z, z' and their translates.

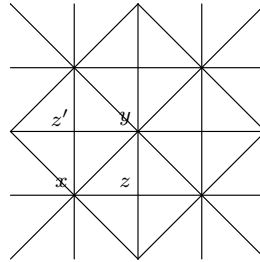
Let us now consider the possibilities for α above to have the form $\alpha = \sigma_\ell$ where ℓ meets R . By our analysis above, then in regard to the picture, ℓ must be vertical, horizontal or diagonal.

Since $\alpha \in \mathcal{W}$, the isotropy subgroup $\mathcal{W}_{\alpha(a)}$ must equal \mathcal{W}_a for all $a \in \mathbb{R}^2$. So σ_ℓ must take 4-centers to 4-centers and take 2-centers to 2-centers. Note that the centers lie at the vertices of a grid made by vertical and horizontal lines, where the distance between closest lines in either direction is $\frac{1}{2}\|v\|$.

Suppose first that ℓ is vertical. The reflection across ℓ must carry vertical lines in this grid to vertical lines in the grid. Therefore, either ℓ must be one of the grid lines, or it must be half way between adjacent grid lines. But the latter case is impossible as σ_ℓ would then take 2-centers to 4-centers and vice versa.

Thus, if ℓ is vertical, it must coincide with one of the grid lines, and therefore must go through a 4-center, say y . But then \mathcal{W}_y is dihedral, and hence is D_8 and there are four lines of symmetry through y : a vertical line, a horizontal line, and two diagonal lines, as in the British flag.

(6.14.3)



Tracing these out in our fundamental region R , we see that every point of symmetry has dihedral isotropy. The diagram (6.14.3) displays all the lines of symmetry meeting R that result from this fact. Note that this exhausts the possibilities for lines of symmetry meeting R , as additional lines would intersect the ones displayed, producing new rotations. Note that the presence of a reflection across any one of the lines displayed in (6.14.3) would produce the entire array of lines of symmetry displayed here, using the patterns associated with the lines of symmetry in D_8 , since each one of these lines contains a 4-center.

Indeed, \mathcal{W}_4 together with the reflection in any vertical, horizontal or diagonal line meeting a 4-center generates exactly this group. And it is indeed a group, as the assumptions of Proposition 6.14.3 are satisfied. We call it \mathcal{W}_4^1 . A pattern whose symmetry group is \mathcal{W}_4^1 is shown in Figure 6.14.1.

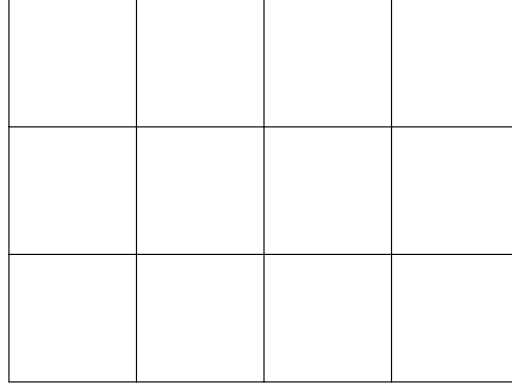
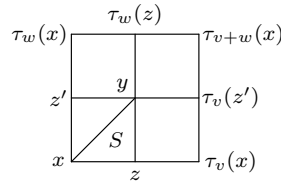


FIGURE 6.14.1. A pattern with symmetry group \mathcal{W}_4^1

The only elements of \mathcal{W}_4^1 that carry interior points of R to interior points of R are the elements in the isotropy subgroup of y . Thus, a fundamental region for \mathcal{W}_4^1 is given by any of the small isosceles right triangles in (6.14.3), e.g., the one with vertices x , y and z .

(6.14.4)



Note that there are no identifications on S coming from elements of \mathcal{W}_4^1 and the orbit space is just S . We can think of the orbit space $\mathbb{R}^2/\mathcal{W}_4^1$ as obtained from the orbit space $\mathbb{R}^2/\mathcal{W}_4$, a sphere, by making the identifications induced by the reflections. These additional identifications amount to folding the sphere over onto itself and flattening it out.

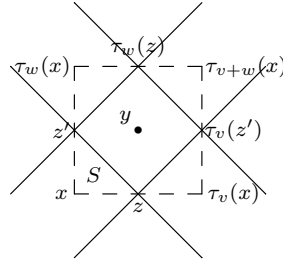
We've seen that \mathcal{W}_4 together with the reflection in any of the lines displayed in (6.14.3) produces \mathcal{W}_4^1 . This covers all possible reflections in lines parallel to v or w that preserve the arrays of centers. But it only covers the diagonal lines going through a 4-center. Let

$$\ell = \overset{\leftrightarrow}{zz'}.$$

ℓ is diagonal, but is not a line of symmetry for \mathcal{W}_4^1 . The points of symmetry on ℓ are all 2-centers. Note that σ_ℓ does preserve the set of 4-centers and also preserves the set of 2-centers. Thus, Proposition 6.14.3 shows that

$\mathcal{W} = \mathcal{W}_4 \cup \mathcal{W}_4\sigma_\ell$ is a wallpaper group. We call it \mathcal{W}_4^2 . Since z and z' lie on ℓ , their isotropy subgroups under \mathcal{W}_4^2 are dihedral. So the diagonal lines through z and z' perpendicular to ℓ must also be lines of symmetry for \mathcal{W}_4^2 . In the picture below, the dotted lines surround a fundamental region R for $\mathcal{T}(\mathcal{W}_4^2)$ and the solid lines represent lines of symmetry for \mathcal{W}_4^2 .

(6.14.5)

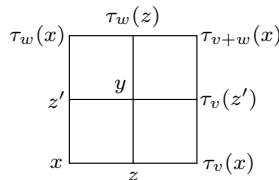


There cannot be any additional lines of symmetry, as that would introduce new rotations. Thus, there are two \mathcal{T} -orbits of 2-centers for \mathcal{W}_4^2 , represented by z and z' , each with isotropy D_4 , and two \mathcal{T} -orbits of 4-centers for \mathcal{W}_4^2 , represented by x and y , each with isotropy C_4 . A fundamental region S for \mathcal{W}_4^2 is given by the triangle with vertices x , z and z' , as indicated in the picture. The identifications on S induced by \mathcal{W}_4^2 are simply to identify \overline{xz} and $\overline{xz'}$ via $\rho(x, \frac{\pi}{2})$. The orbit space is a cone.

A pattern with symmetry group \mathcal{W}_4^2 is given in Figure 6.14.2. A fundamental region R for $\mathcal{T}(\mathcal{W}_4^2)$ is overlaid in dotted lines. Its vertices and center point are 4-centers. The 2-centers are at the midpoints of its edges. Its lines of symmetry are the vertical and horizontal lines through its 2-centers.

Our analysis above shows that \mathcal{W}_4^1 and \mathcal{W}_4^2 are the only wallpaper groups obtained from \mathcal{W}_4 by adjoining a reflection. One can still ask if one can obtain a different wallpaper group by adjoining a glide reflection. “Different” here is key, as Proposition 5.5.22 shows that if $\rho_{(x,\theta)}$ is a nontrivial rotation and $x \notin \ell$, then both $\rho_{(x,\theta)}\sigma_\ell$ and $\sigma_\ell\rho_{(x,\theta)}$ are glide reflections. In particular, both \mathcal{W}_4^1 and \mathcal{W}_4^2 have numerous glide reflections.

Thus, we consider wallpaper groups of the form $\mathcal{W} = \mathcal{W}_4 \cup \mathcal{W}_4\alpha$ with α a glide reflection. So $\alpha = \tau_u\sigma_\ell$ with $u \parallel \ell$, and $\alpha^2 = \tau_{2u}$ must lie in $\mathcal{T}_\Lambda = \mathcal{T}(\mathcal{W}_4)$. The case where $\tau_u \in \mathcal{T}_\Lambda$ is uninteresting, as then $\sigma_\ell = \tau_u^{-1}\alpha$ is in \mathcal{W} , so \mathcal{W} is already known to be one of \mathcal{W}_4^1 and \mathcal{W}_4^2 . So the interesting cases are where τ_u is not in \mathcal{T}_Λ but τ_{2u} is. As shown above, ℓ (and hence u) must be vertical, horizontal or diagonal with respect to region R in the layout



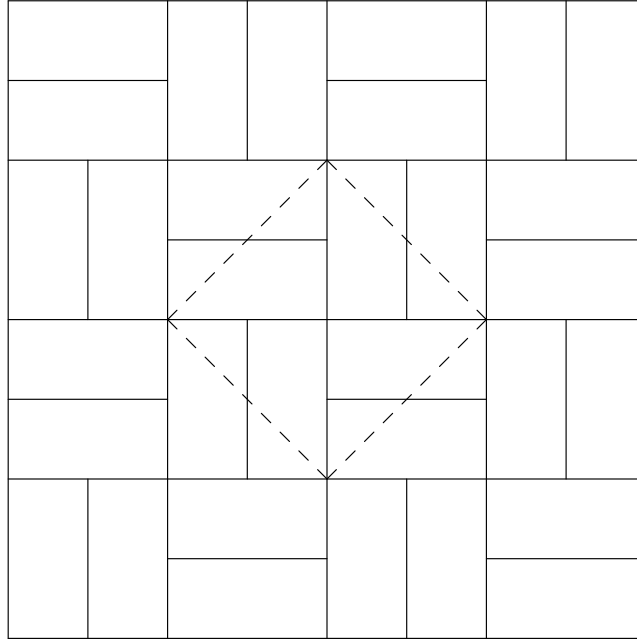


FIGURE 6.14.2. A pattern with symmetry group \mathcal{W}_4^2 .

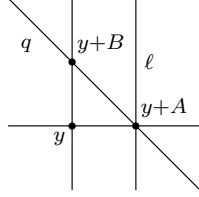
where x and y are 4-centers and z and z' are 2-centers. We shall assume that ℓ intersects R in more than one point.

First, consider the case where ℓ is vertical. If ℓ contains one of the points of symmetry in R , then σ_ℓ is already known to preserve the periods of all points of symmetry, and hence we have to translate by a multiple of τ_w to again preserve the periods. As above, this implies $\sigma_\ell \in \mathcal{W}$ and hence $\mathcal{W} = \mathcal{W}_4^1$. Thus, the only possibility for something new with ℓ vertical is if ℓ passes half way between the vertical columns of centers, i.e., if ℓ passes through either $x + \frac{1}{4}v$ or $x + \frac{3}{4}v$. In either case, u must be an odd multiple of $\frac{1}{2}w$ in order to preserve the periods of the points of symmetry for \mathcal{W}_4 . But then $\tau_u = \tau_{\frac{2k+1}{2}w} = \tau_{kw}\tau_{\frac{1}{2}w}$. Since $\tau_{kw} \in \mathcal{W}$, this implies $\tau_{\frac{1}{2}w}\sigma_\ell \in \mathcal{W}$. Thus, we may assume $u = \frac{1}{2}w$.

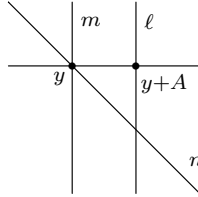
We shall show that these glide reflections lie in \mathcal{W}_4^2 . To do so, we make use of the following lemma, which gives a precise calculation of the composite of a rotation by $\frac{\pi}{2}$ with a reflection not containing the rotation point. The lemma is actually a special case of Proposition 5.5.26, but it is simpler and easier to picture than the general case, so we present it here in full detail.

Lemma 6.14.11. *Let $0 \neq A \in \mathbb{R}^2$ and let $B = \rho_{(0, \frac{\pi}{2})}(A)$. Let $y \in \mathbb{R}^2$ and let ℓ be the line through $y + A$ parallel to B : $\ell = (y + A) + \text{span}(B)$. Then $\rho_{(y, \frac{\pi}{2})}\sigma_\ell$ is the glide reflection which in standard form is given by $\tau_{(B-A)}\sigma_q$,*

where q is the line through $y + A$ and $y + B$, i.e., $q = (y + A) + \text{span}(B - A)$.



Proof. Let m be the line through y parallel to ℓ and let n be the line through y parallel to q .



Then $\rho_{(y, \frac{\pi}{2})} = \sigma_n \sigma_m$, hence

$$\begin{aligned} \rho_{(y, \frac{\pi}{2})} \sigma_\ell &= \sigma_n (\sigma_m \sigma_\ell) \\ &= \sigma_n \tau_{-2A} \\ &= \sigma_n \tau_{-(B+A)} \tau_{(B-A)} \\ &= \sigma_{(n + \frac{1}{2}(B+A))} \tau_{(B-A)} \end{aligned}$$

By Lemma 5.5.16, as $(B + A) \perp n$. In particular, $\rho_{(y, \frac{\pi}{2})} \sigma_\ell = \sigma_q \tau_{(B-A)}$ for $q = n + \frac{1}{2}(B + A)$. Now $n = y + \text{span}(B - A)$, hence

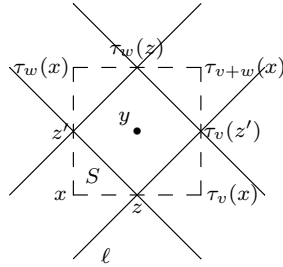
$$q = \left(y + \frac{1}{2}(B + A) \right) + \text{span}(B - A),$$

and it suffices to show that $y + A$ and $y + B$ are on q . But

$$y + \frac{1}{2}(B + A) + \frac{1}{2}(B - A) = y + B,$$

$$y + \frac{1}{2}(B + A) - \frac{1}{2}(B - A) = y + A. \quad \square$$

We now apply Lemma 6.14.11 in \mathcal{W}_4^2 with ℓ the line through z and $\tau_v(z')$ in the fundamental region R :



This gives $A = \frac{1}{4}(v - w)$ and $B = \frac{1}{4}(v + w)$. This gives $\rho_{(y, \frac{\pi}{2})}\sigma_\ell = \sigma_q\tau_{\frac{1}{2}v}$ with $q = x + \frac{3}{4}v + \text{span}(w)$. This is one of the vertical glide reflections we uncovered above. Since $\sigma_\ell \in \mathcal{W}_4^2$, this glide reflection is as well.

A similar argument with ℓ the line through z' and $\tau_w(z)$ shows the glide reflection $\sigma_q\tau_{-\frac{1}{2}v}$ is in \mathcal{W}_4^2 where $q = x + \frac{1}{4}v + \text{span}(w)$. This is the inverse of the vertical glide reflection through $x + \frac{1}{4}v$ discussed above.

If instead we use the other two reflections in \mathcal{W}_4^2 that intersect R , we see that the two nontrivial horizontal glide reflections that generate wallpaper groups when added to \mathcal{W}_4 also produce \mathcal{W}_4^2 .

The only remaining possibility is glide reflections whose axis is diagonal with respect to R . In the diagonal rows of points of symmetry, the isotropy is the same for all points of symmetry in the row, and is different between closest rows. So the axis for a diagonal glide reflection in a wallpaper group containing \mathcal{W}_4 must lie along one of the rows of centers rather than in between.

One case is easy, as it already occurred in the frieze group \mathcal{F}_2^2 : if m is the line containing the points z and z' in R and if ℓ is the line containing x and y , then $\rho_{(y, \pi)}\sigma_m = \tau_{\frac{1}{2}(v+w)}\sigma_\ell$, a glide reflection with axis ℓ . In particular, this glide reflection is in \mathcal{W}_4^2 . So are all other possible glide reflections along diagonal lines of 4-centers.

Finally, we apply Lemma 6.14.11 with ℓ the vertical line through x . Here, $A = -\frac{1}{2}v$, $B = -\frac{1}{2}w$, and $\rho_{(y, \frac{\pi}{2})}\sigma_\ell = \sigma_q\tau_{\frac{1}{2}(v-w)}$, a glide reflection whose axis q is the line through z and z' . So this glide reflection, along with all other candidates for glide reflections through diagonal lines of 2-centers, is in \mathcal{W}_4^1 . We have proven the following.

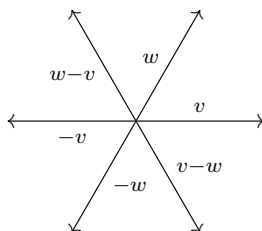
Theorem 6.14.12. *There are exactly two wallpaper groups obtained by adding orientation-reversing symmetries to \mathcal{W}_4 : \mathcal{W}_4^1 and \mathcal{W}_4^2 . In \mathcal{W}_4^1 , the two \mathcal{T} -orbits of 4-centers have isotropy D_8 and the two \mathcal{T} -orbits of 2-centers have isotropy D_4 , so every point of symmetry lies on a line of symmetry. The pattern of lines of symmetry in the fundamental region R for $\mathcal{T}(\mathcal{W}_4^1)$ is given in (6.14.3), and a fundamental region for \mathcal{W}_4^1 is indicated in (6.14.4). The orbit space is just this fundamental region S .*

In \mathcal{W}_4^2 , the two \mathcal{T} -orbits of 2-centers have isotropy D_4 , but the two \mathcal{T} -orbits of 4-centers have isotropy C_4 . So every 2-center is on a line of symmetry, but none of the 4-centers are. The pattern of lines of symmetry in a fundamental region R for $\mathcal{T}(\mathcal{W}_4^2)$ is given in (6.14.5), and a fundamental region for \mathcal{W}_4^2 is indicated by S there. The orbit space is a cone.

6.14.2. Wallpaper groups with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_6$. Let \mathcal{W} be a wallpaper group with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_6$ and let $\alpha \in \mathcal{W} \setminus \mathcal{W}_6$. The nonzero translation vectors in \mathcal{W}_6 of minimal length form the vertex set for a regular hexagon

X centered at the origin.

(6.14.6)

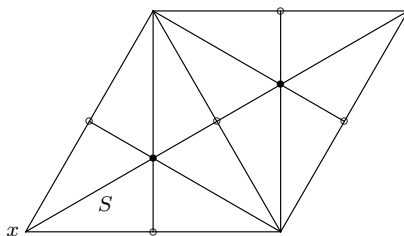


The identification of the upper left arrow as $w - v$ may be obtained either via algebra in \mathbb{C} or by translating the equilateral triangle with vertices 0 , v and w by $-v$.

Let $\alpha = \tau_x \sigma_{\ell_\phi}$. By Lemma 6.14.2, σ_{ℓ_ϕ} is one of the reflections in the dihedral group $\mathcal{S}(X) \cong D_{12}$, so ℓ_ϕ either coincides with the span of one of the displayed vectors, or bisects the angle between an adjacent pair.

Suppose \mathcal{W} admits a reflection through a 6-center x . Then the isotropy subgroup \mathcal{W}_x is dihedral, and since the six reflections in \mathcal{W}_x must be parallel to those preserving the regular hexagon X , the three lines emanating from x in the following diagram of the fundamental region R for $\mathcal{T}(\mathcal{W}_6)$ must be lines of symmetry.

(6.14.7)



Here, the figure gives a fundamental region R for $\mathcal{T}(\mathcal{W})$, with the 2-centers indicated as \circ , the 3-centers as \bullet and the vertices of the rhombus are x , $\tau_v(x)$, $\tau_w(x)$ and $\tau_{v+w}(x)$, the 6-centers in R . Every center that meets one of the three lines of symmetry emanating from x must have dihedral isotropy, which forces all the other indicated lines to be lines of symmetry.

There can be no other lines of symmetry meeting R , as that would introduce new rotations. So any one of the little triangles in (6.14.7) is a fundamental region for \mathcal{W} , as indicated by S in the figure. A complete diagram with these symmetries is given in Figure 6.14.3, showing that we have in fact constructed a wallpaper group. We call it \mathcal{W}_6^1 , and (6.14.7) gives a fundamental region for $\mathcal{T}(\mathcal{W}_6^1)$ with all its symmetries.

We can now ask if there are alternative ways we could add reflections to \mathcal{W}_6 to obtain a wallpaper group. (6.14.7) makes a good guide for answering this question, because the axis for any such reflection would have to be parallel to one of the lines drawn in that figure. If a line ℓ meets R and is parallel to the bottom edge of (6.14.7), then $\sigma_\ell(x)$ is directly above x , and cannot be a 6-center unless ℓ coincides with the upper edge of R .

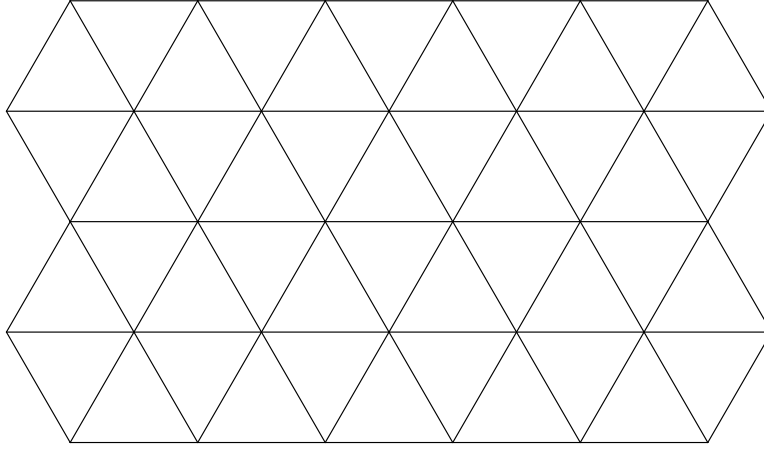
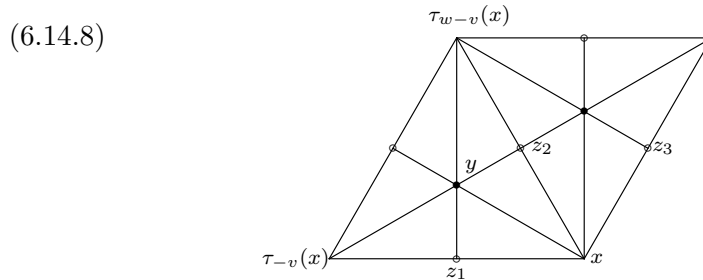


FIGURE 6.14.3. A figure with symmetry group \mathcal{W}_6^1 . The vertices of the small equilateral triangles are all 6-centers.

A similar analysis applies to all the other possible directions for lines of symmetry. The only permissible reflections that carry 6-centers to 6-centers are the ones displayed in (6.14.7). Thus, \mathcal{W}_6^1 is the only wallpaper group obtained from \mathcal{W}_6 by adding reflections. We must yet consider the possibility that glide reflections could be added to \mathcal{W}_6 to produce a wallpaper group with no reflections. We repeat the diagram for the fundamental region R for $\mathcal{T}(\mathcal{W}_6^1)$ and label some points.



Up to symmetry, the glides we need to consider are as follows:

- (1) The axis is the line containing z_2 and z_3 . The glide takes z_3 to z_2 .
- (2) The axis is the line containing z_1 and z_3 . The glide takes z_3 to z_1 .
- (3) The axis is the vertical line through z_2 . The glide takes z_2 to $\tau_{w-v}(z_3)$.

By Propositions 5.5.26 and 5.5.27, these glide reflections all lie in \mathcal{W}_6^1 . The one in (1) is $\rho(y, \frac{2\pi}{3})\sigma_\ell$ where ℓ is the line through x and z_2 . The one in (2) is $\rho(x, \frac{2\pi}{3})\sigma_m$ where m is the line through z_3 and $\tau_{w-v}(x)$. The one in (3) is $\sigma_n\rho(y, \frac{2\pi}{3})$ where n is the line through z_3 and $\tau_{w-v}(x)$.

We have obtained the following.

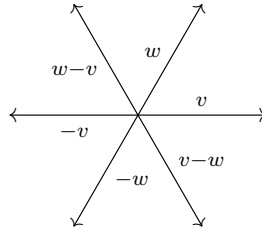
Theorem 6.14.13. *There is only one way to extend \mathcal{W}_6 to a wallpaper group containing orientation-reversing isometries. The result is \mathcal{W}_6^1 . A fundamental region for $\mathcal{T}(\mathcal{W}_6^1)$ showing its points and lines of symmetry is shown in (6.14.7). There are three \mathcal{T} -orbits of 2-centers, each with isotropy D_4 ; two \mathcal{T} -orbits of 3-centers, each with isotropy D_6 ; one \mathcal{T} -orbit of 6-centers with isotropy D_{12} . A fundamental region for \mathcal{W}_6^1 is the small triangle marked S in (6.14.7). There are no identifications on S induced by \mathcal{W}_6^1 , so the orbit space is just S .*

6.14.3. Wallpaper groups with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_3$. As was the case for \mathcal{W}_4 we will construct two different wallpaper groups \mathcal{W} with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_3$ (and with $\mathcal{W} \setminus \mathcal{O}(\mathcal{W}) \neq \emptyset$) and then show there are no others.

First note that the translation subgroups of \mathcal{W}_3 and \mathcal{W}_6 coincide. Indeed, \mathcal{W}_3 is a subgroup of \mathcal{W}_6 , and $\mathcal{W}_6 \setminus \mathcal{W}_3$ consists of rotations only. For instance, if you remove the arrowheads from the diagram in Figure 6.13.4 (a diagram whose symmetry group is \mathcal{W}_3), you get a diagram whose symmetries are \mathcal{W}_6^1 , providing an embedding of \mathcal{W}_3 into $\mathcal{O}(\mathcal{W}_6^1) = \mathcal{W}_6$. The image of this embedding consists of the obvious elements.

Thus, the nonzero translation vectors in \mathcal{W}_3 of minimal length form the vertex set for a regular hexagon X centered at the origin.

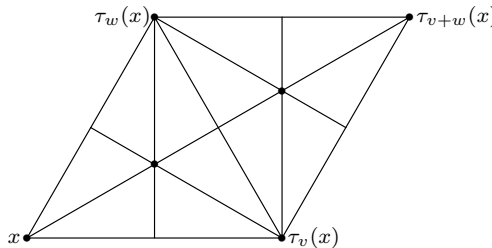
(6.14.9)



Let \mathcal{W} be a wallpaper group with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_3$ and let $\alpha = \tau_x \sigma_{\ell_\phi}$ be in $\mathcal{W} \setminus \mathcal{W}_3$ with ℓ_ϕ a line through the origin. Then by Lemma 6.14.2, σ_{ℓ_ϕ} is one of the reflections in the dihedral group $\mathcal{S}(X) \cong D_{12}$, so ℓ_ϕ either coincides with the span of one of the displayed vectors, or bisects the angle between an adjacent pair.

Precisely as in the case of \mathcal{W}_6 , the axis of α (either as a reflection or a glide reflection) must be parallel to one of the lines in the following diagram for a fundamental region R for $\mathcal{T}(\mathcal{W}_3)$.

(6.14.10)



Here, x can be any 3-center.

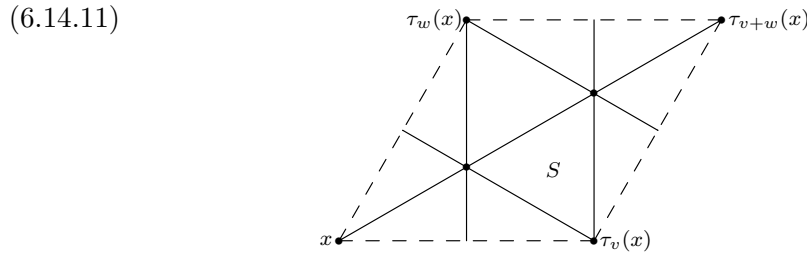
The complication here is that, just as the 4-centers in a \mathcal{W}_4 -group needn't lie on lines of symmetry, a particular \mathcal{T} -orbit of 3-centers in a \mathcal{W}_3 -group needn't lie on lines of symmetry.

Let us first assume that the 3-center x does lie on a line of symmetry. Thus, the isotropy subgroup of x with respect to \mathcal{W} is D_6 . Thus, there are exactly three lines of symmetry containing x , and they make unsigned angles of $\frac{2\pi}{3}$ with one another.

Thus, there are exactly two possibilities:

- (1) The long diagonal $\overline{x\tau_{v+w}(x)}$ lies on a line of symmetry.
- (2) The edges $\overline{x\tau_v(x)}$ and $\overline{x\tau_w(x)}$ lie on lines of symmetry.

In Case (1), representatives of all three \mathcal{T} -orbits of 3-centers lie on $\overline{x\tau_{v+w}(x)}$ so the isotropy group for all three orbits is D_6 . That gives us the following lines of symmetry intersecting R in more than one point.



Here, the dotted lines represent the edges of R . Note that there can be no further reflections meeting R in more than one point as that would introduce additional rotations. We call this group \mathcal{W}_3^1 . Note that the equilateral triangle labelled S gives a fundamental region for \mathcal{W}_3^1 , as its image under iterated rotations and reflections can be seen to cover all of \mathbb{R}^2 , and no element of \mathcal{W}_3^1 carries interior points of S to interior points of S . In fact, no two points of S are identified by elements of \mathcal{W}_3^1 and the orbit space of \mathcal{W}_3^1 is just S .

Another way of seeing this is that the union of S with its reflection across $\overline{x\tau_{v+w}(x)}$ is the fundamental region for \mathcal{W}_3 given in (6.13.15). So S can be seen as the result of folding a fundamental region for \mathcal{W}_3 in half along the reflection line, and the orbit space of \mathcal{W}_3^1 can be seen as the result of flattening out the spherical orbit space for \mathcal{W}_3 via this fold.

A pattern with symmetry group \mathcal{W}_3^1 is given in Figure 6.14.4. Note that the “local pattern” around the 3-centers is different for 3-centers in different \mathcal{T} -orbits. One \mathcal{T} -orbit of 3-centers looks like a target for the arrows. Another is a source. The third is the center of an empty hexagon. This underlines the fact that 3-centers from different \mathcal{T} -orbits for \mathcal{W}_3^1 are not mapped to one another by either rotations or reflections. That gives another verification that the orbit space of \mathcal{W}_3^1 is just S .

In Case (2) the vertices of R all have isotropy D_6 . Since the two specified edges lie on lines of symmetry, we obtain the following lines of symmetry

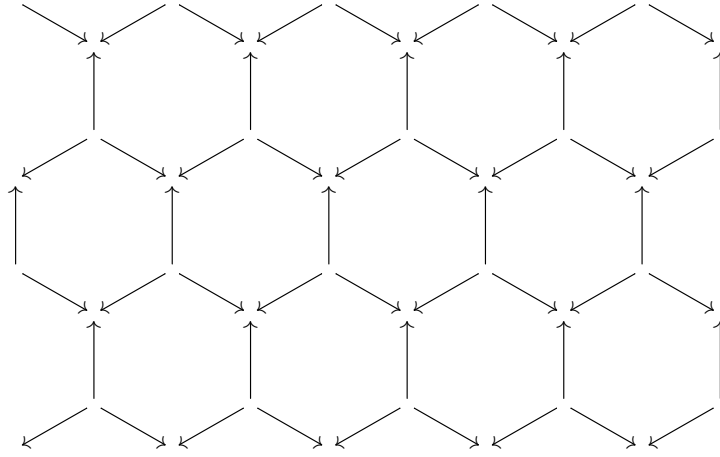
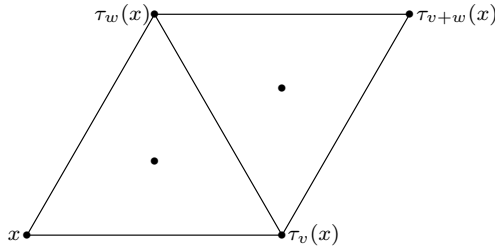


FIGURE 6.14.4. A figure with symmetry group \mathcal{W}_3^1 .

meeting R in more than one point.

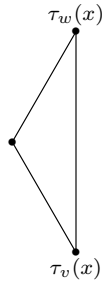
(6.14.12)



There cannot be any additional lines of symmetry meeting R in more than one point, as that would introduce new rotations on the boundary of R . Thus the two \mathcal{T} -orbits of 3-centers represented by the 3-centers in the interior of R have isotropy C_3 while the \mathcal{T} -orbit given by the vertices of R has isotropy D_6 . Thus not all 3-centers lie on lines of symmetry, and it makes a difference to our diagram that we chose x so it did. We call the resulting group \mathcal{W}_3^2 .

A fundamental region for S is given by the following triangle, where the unlabelled vertex coincides with the unlabelled centroid in the lower (or left-hand) equilateral triangle in (6.14.12).

(6.14.13)



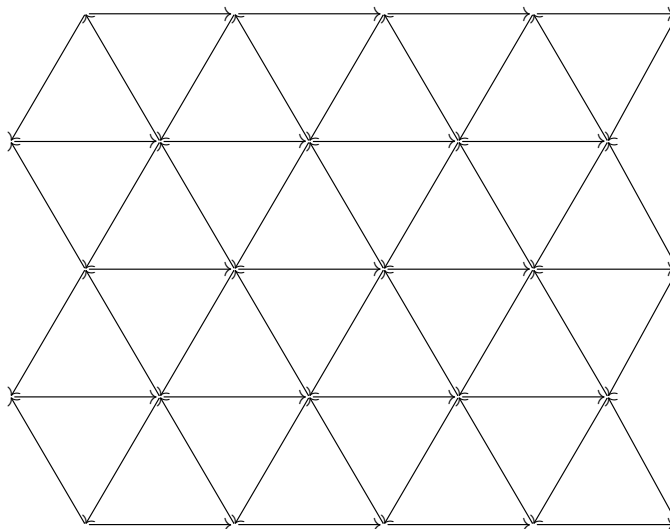


FIGURE 6.14.5. A pattern with symmetry group \mathcal{W}_3^2 .

We have rotated the appropriate section of (6.14.12) by $-\frac{\pi}{6}$ about $\tau_w(x)$ to suit our graphics generator. Note this is precisely the left half of the fundamental region for \mathcal{W}_3 given in (6.13.15). The orbit space is obtained by gluing together the two left-hand edges of S via the rotation about their common vertex, and is topologically a cone.

A pattern whose symmetry group is \mathcal{W}_3^2 is given in Figure 6.14.5. Again the three \mathcal{T} -orbits of 3-centers have different local pattern, but two of them are mirror images of one another and are identified via the reflection that takes one center to the other.

All that remains now is to show that any wallpaper group \mathcal{W} containing orientation-reversing isometries with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_3$ must be one of \mathcal{W}_3^1 and \mathcal{W}_3^2 . As shown above, this amounts to showing that there is at least one 3-center for \mathcal{W} that lies on a line of symmetry.

Thus, suppose first that \mathcal{W} admits a line ℓ of symmetry. We shall show ℓ must contain a 3-center. First note from Figure 6.13.3 that there are three different possible orientations for a rhombic fundamental region R for $\mathcal{T}(\mathcal{W}_3)$. In the framework of that picture, the long diagonal in one of them is vertical, while the long diagonals of the other two have positive and negative slope, respectfully.

This implies that any line of symmetry for \mathcal{W} is parallel to either the long diagonal or the lower edge of some fundamental region R for $\mathcal{T}(\mathcal{W}_3)$. Suppose first that \mathcal{W} admits a line of symmetry ℓ parallel to the lower edge of R , and meeting R . So consider Figure 6.14.6, and suppose ℓ is parallel to the line m through v and $\tau_v(x)$. If $\ell = m$, then the 3-center A lies on ℓ

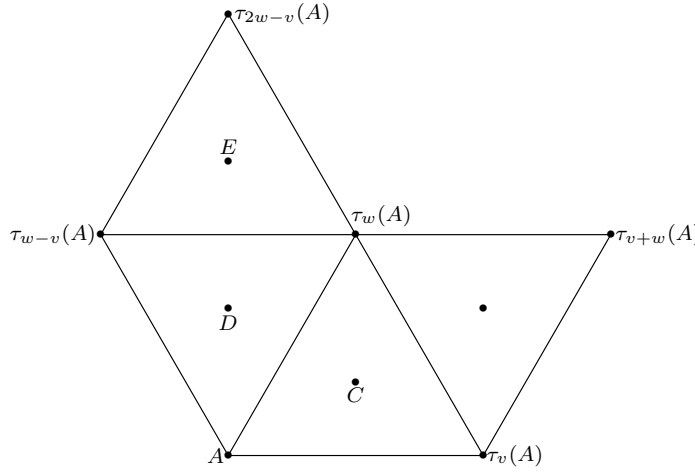


FIGURE 6.14.6. Array of 3-centers and translations for \mathcal{W}_3 .

and we're done. Otherwise, $\sigma_\ell(A)$ lies on the line through A perpendicular to ℓ . Since $\sigma_\ell(A)$ is a 3-center, it must equal either D , E or $\tau_{2w-v}(A)$. In the first case, since C lies on the perpendicular bisector of \overline{AD} , the 3-center C lies on ℓ . In the second, since $d(A, D) = d(D, E)$ is the shortest distance between two 3-centers, D lies on ℓ . In the last case, $\tau_w(A)$ lies on ℓ . In all of these cases, $\mathcal{W} = \mathcal{W}_3^2$.

Now assume ℓ meets R and is parallel to the long diagonal m of R , i.e., to the line containing A and $\tau_{v+w}(A)$. The lines parallel to m either contain no 3-centers or contain a whole line of 3-centers, so if ℓ is parallel to m , then either ℓ contains a vertex of R , in which case we are done, or ℓ is half way between two such lines of 3-centers. But in this case, σ_ℓ takes either $\tau_v(A)$ or $\tau_w(A)$ to the midpoint of the segment between them, which is impossible as that point is not a point of symmetry for \mathcal{W}_3 .

Thus, it remains to show that if \mathcal{W} contains a glide reflection, it must contain a reflection. The argument here is identical to the one given for \mathcal{W}_6 . Thus, we have shown the following.

Theorem 6.14.14. *There are exactly two wallpaper groups obtained by adding orientation-reversing isometries to \mathcal{W}_3 . In the first, \mathcal{W}_3^1 , every \mathcal{T} -orbit of 3-centers has isotropy D_6 . A fundamental region R for $\mathcal{T}(\mathcal{W}_3^1)$ showing all points and lines of symmetry is given in (6.14.11). The equilateral triangle S in that diagram is a fundamental region for \mathcal{W}_3^1 , and also gives its orbit space.*

In the second, \mathcal{W}_3^2 , two of the \mathcal{T} -orbits of 3-centers have isotropy C_3 and the third has isotropy D_6 . A fundamental R region for $\mathcal{T}(\mathcal{W}_3^2)$ showing all points and lines of symmetry is given in (6.14.12). A fundamental S region for \mathcal{W}_3^2 is given in (6.14.13). The orbit space is a cone.

6.14.4. Wallpaper groups with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_2$. Here, since the translation lattice can be arbitrary, there are more complications possible than in the previous cases. In fact, there are four different wallpaper groups \mathcal{W} containing orientation-reversing isometries such that $\mathcal{O}(\mathcal{W}) = \mathcal{W}_2$. One of them contains glide reflections but no reflections. But all of them are forecast by the discussions above.

Recall that if Λ is a lattice in \mathbb{R}^2 and if $x \in \mathbb{R}^2$, there is a unique orientation-preserving wallpaper group $\mathcal{W}_2 = \mathcal{W}_2(\Lambda, x)$ whose points of symmetry all have period 2 such that $\mathcal{T}(\mathcal{W}_2) = \mathcal{T}_\Lambda$ and x is a 2-center (see (6.13.20)). If $\mathcal{B} = v, w$ is a \mathbb{Z} -basis for Λ then a fundamental region for $\mathcal{T}(\mathcal{W}_2)$ is given in the following diagram. The marked points are its 2-centers.

$$(6.14.14) \quad \begin{array}{c} \tau_v(x) \bullet \text{---} \bullet \text{---} \tau_{v+w}(x) \\ \diagup \quad \quad \quad \diagdown \\ x \bullet \text{---} \bullet \text{---} \tau_w(x) \end{array}$$

Here, the marked point in the center is $x + \frac{1}{2}(v + w)$, the midpoint of the diagonal $\overline{x\tau_{v+w}(x)}$. It coincides with the midpoint of the other diagonal $\overline{\tau_v(x)\tau_w(x)}$.

In the previous cases, the rotational symmetries determined the shape of a preferred fundamental region for \mathcal{T}_Λ . In this case and for \mathcal{W}_1 -groups, the shape will be influenced by the orientation-reversing isometries present.

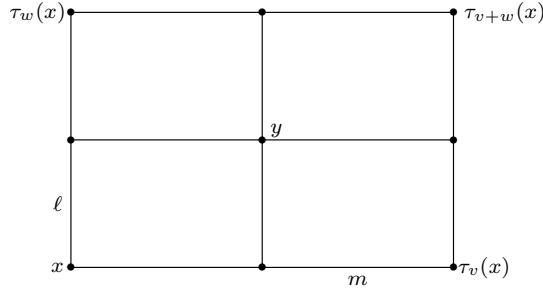
Recall from Lemma 6.14.4 that if ℓ is a line of symmetry for a wallpaper group \mathcal{W} , then there are infinitely many lines of symmetry parallel to ℓ . There are three cases:

- (1) There are reflections in more than one direction, i.e., not all lines of symmetry are parallel.
- (2) There are reflections, and all lines of symmetry are parallel.
- (3) There are no reflections, but there are glide reflections.

Let us first consider Case (1). Of course if ℓ and m are nonparallel lines of symmetry for \mathcal{W} , then $\sigma_m\sigma_\ell = \rho_{(x,2\theta)} \in \mathcal{W}$, where $x = \ell \cap m$ and θ is the directed angle from ℓ to m . In a \mathcal{W}_2 -group, 2θ must equal π , and hence ℓ and m must be perpendicular. By Lemma 6.14.4, we obtain a rectangular grid of lines of symmetry. The lines parallel to ℓ are $\{\tau_{\frac{k}{2}v}(\ell) : k \in \mathbb{Z}\}$, where τ_v is a shortest translation in \mathcal{W} perpendicular to ℓ , and the lines parallel to m are $\{\tau_{\frac{k}{2}w}(m) : k \in \mathbb{Z}\}$, where τ_w is a shortest translation in \mathcal{W} perpendicular to m (hence parallel to ℓ). Note there can be no other lines of symmetry for ℓ as we have ruled out additional lines parallel to either ℓ or m , and any other line would introduce a rotation by an angle other than π , which cannot exist in a \mathcal{W}_2 -group.

Each intersection point between lines parallel to ℓ and lines parallel to m is a 2-center. We obtain a region T as follows.

(6.14.15)



By construction, $\bigcup_{z \in \mathcal{T}(\mathcal{W})} \tau_z(T) = \mathbb{R}^2$, so T is a fundamental region for $\mathcal{T}(\mathcal{W})$ if and only if there is no translation $\tau_z \in \mathcal{T}(\mathcal{W})$ taking an interior point of T to an interior point of T . However, each translation $\tau_z \in \mathcal{T}(\mathcal{W})$ preserves parallel lines, carries 2-centers of \mathcal{W} to 2-centers of \mathcal{W} , and carries lines of symmetry of \mathcal{W} to lines of symmetry of \mathcal{W} . Since $\tau_{\frac{1}{2}v}$ and $\tau_{\frac{1}{2}w}$ are not in \mathcal{W} , the only way τ_z can carry an interior point of T to an interior point of T is if it takes some vertex of T to the center point y of T . In other words, $\pm z$ must equal either z_1 or z_2 where

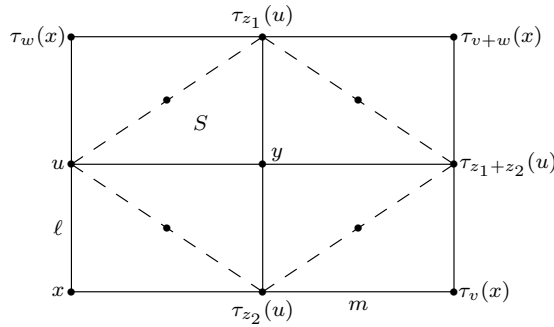
$$z_1 = \frac{1}{2}(v + w)$$

$$z_2 = \frac{1}{2}(v - w).$$

Note that $z_1 + z_2 = v$, so if either of τ_{z_1} or τ_{z_2} is in $\mathcal{T}(\mathcal{W})$, so is the other.

Let $u = \tau_{\frac{1}{2}w}(x)$. Then u is a 2-center, and if $\tau_z \in \mathcal{W}$, then $\tau_{\frac{1}{2}z}(u)$ is a 2-center for \mathcal{W} . In particular, if τ_{z_1} and τ_{z_2} are in \mathcal{W} , we obtain the displayed 2-centers in the centers of the four rectangles of (6.14.15), precisely as in Proposition 6.14.5. Moreover, by Proposition 6.14.5, there are no other 2-centers in T , so the region bounded by the dotted lines below must be the fundamental region for $\mathcal{T}(\mathcal{W}) = \mathcal{T}_\Lambda$. In particular z_1, z_2 must be a \mathbb{Z} -basis for Λ .

(6.14.16)



In particular, if τ_{z_1} and τ_{z_2} are in $\mathcal{T}(\mathcal{W})$, then the fundamental region R for $\mathcal{T}(\mathcal{W})$ is rhombic (its sides have equal length), and the two diagonals

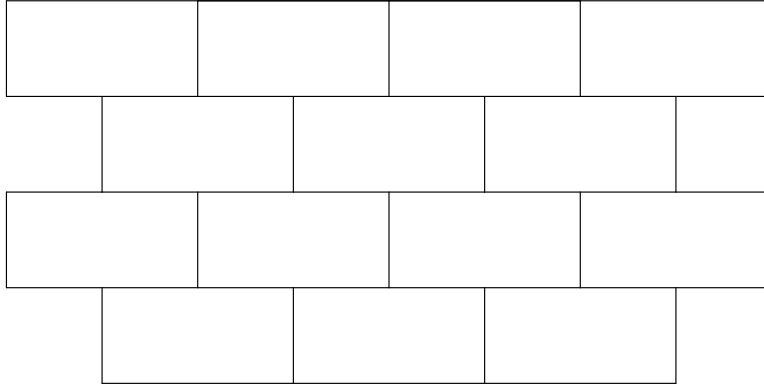
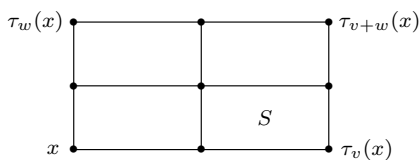


FIGURE 6.14.7. A pattern with symmetry group \mathcal{W}_2^1 .

of this rhombus are lines of symmetry for \mathcal{W} . Note that any rhombus can be obtained in this manner. We call this group \mathcal{W}_2^1 . It has two \mathcal{T} -orbits of 2-centers with isotropy D_4 and two \mathcal{T} -orbits of 2-centers with isotropy C_2 .

The triangle labeled S in (6.14.16) gives a fundamental region for \mathcal{W}_2^1 . The only identification on S given by elements of \mathcal{W}_2^1 is induced by the rotation about the midpoint of $\overline{u\tau_{z_1}(u)}$ by π . It folds the edge $\overline{u\tau_{z_1}(u)}$ in half. So the orbit space is a cone. A pattern with symmetry group \mathcal{W}_2^1 is given in Figure 6.14.7. A fundamental region for $\mathcal{T}(\mathcal{W}_2^1)$ can be taken to have vertices at the center points of the bricks. Its edges are diagonal in the picture.

The other possibility when there are reflections in two different directions is that the region T in (6.14.15) is a fundamental region for $\mathcal{T}(\mathcal{W}) = \mathcal{T}_\Lambda$. In this case Λ has a \mathbb{Z} -basis v, w . As shown above, there can be no further lines or points of symmetry in T .

(6.14.17) 

We call this group \mathcal{W}_2^2 . Its four \mathcal{T} -orbits of 2-centers all have isotropy D_4 .

The small rectangle labelled S is a fundamental region for \mathcal{W}_2^2 . There are no identifications on S induced by \mathcal{W}_2^2 , so S is the orbit space. A pattern with symmetry group \mathcal{W}_2^2 is given in Figure 6.14.8. A single brick represents the fundamental region R .

We now consider Case (2), where all lines of symmetry are parallel. There cannot be any points of symmetry on a line of symmetry, as if x were such a point, the isotropy group \mathcal{W}_x would be D_4 , requiring a line of symmetry through x perpendicular to the original line of symmetry.

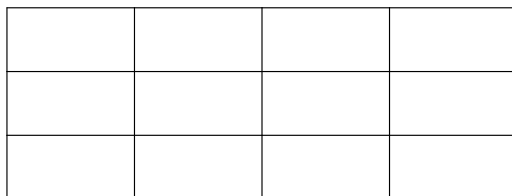


FIGURE 6.14.8. A pattern with symmetry group \mathcal{W}_2^2

Let ℓ be a line of symmetry for \mathcal{W} . By Lemma 6.14.4 the lines of symmetry for \mathcal{W} are precisely $\{\tau_{\frac{k}{2}v}(\ell) : k \in \mathbb{Z}\}$ where v is a shortest translation in \mathcal{W} perpendicular to ℓ .

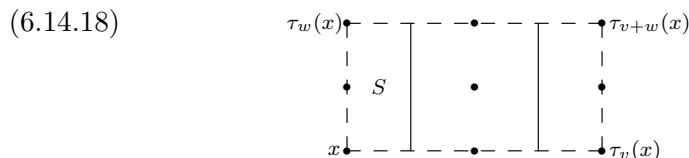
By Proposition 6.14.5 any point x of symmetry must be exactly half way between any two such lines. Thus, there is a line $m = \tau_{\frac{k}{2}v}(\ell)$ such that $x \in \tau_{\frac{1}{4}v}(m)$. We now have $\tau_{kv}\rho(x,\pi) = \rho(\tau_{\frac{k}{2}v}(x),\pi)$, providing a 2-center in each permissible location along $x + \text{span}(v)$.

Since \mathcal{W} is a wallpaper group, there is a translation $\tau_z \in \mathcal{W}$ with $z \notin \text{span}(v)$. So there is a 2-center $\tau_z(x)$ not on $x + \text{span}(v)$. As above, this produces a 2-center along $\tau_z(x) + \text{span}(v)$ half way between each adjacent pair of lines of symmetry for \mathcal{W} , including a 2-center, u , between m and $\tau_{\frac{1}{2}v}(m)$, i.e., in the same chamber as x . But then $\rho(u,\pi)\rho(x,\pi) = \tau_{2(u-x)}$ is a translation parallel to ℓ .

Let τ_w be a shortest translation in \mathcal{W} parallel to ℓ . Then for each 2-center $z = \tau_{\frac{k}{2}v}(x)$ along $x + \text{span}(v)$ there is an infinite family $\{\tau_{\frac{r}{2}w}(z) : r \in \mathbb{Z}\}$ of 2-centers along $z + \text{span}(w)$. By the minimality of w , these are the only 2-centers between the lines of symmetry $z \pm \tau_{\frac{1}{4}v}(\text{span}(w))$. Therefore, the only 2-centers for \mathcal{W} are

$$\{\tau_{\frac{1}{2}(kv+rw)}(x) : k, r \in \mathbb{Z}\}.$$

These are precisely the 2-centers in $\mathcal{W}_2(\Lambda_{\mathcal{B}}, x)$ where $\mathcal{B} = v, w$, so $\mathcal{T}(\mathcal{W}) = \mathcal{T}_{\Lambda_{\mathcal{B}}}$, and the fundamental region R for $\mathcal{T}(\mathcal{W})$ is as follows.



The solid vertical lines are the only lines of symmetry that meet R and the marked points are its 2-centers, all of which have isotropy C_2 . We call this group \mathcal{W}_2^3 . A pattern with symmetry group \mathcal{W}_2^3 is given in Figure 6.14.9. The marked points in it give all its 2-centers. The lines of symmetry are the vertical lines between the columns of 2-centers.

The left-hand quarter panel in (6.14.18), labelled S , is a fundamental region for \mathcal{W}_2^3 . The identifications on it glue its lower boundary to its upper boundary via τ_w and fold it's left hand boundary in half via $\rho_{(x+\frac{1}{2}w,\pi)}$, producing a shape like a pillow case.

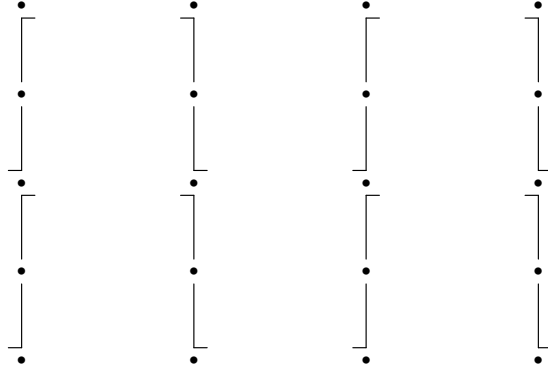


FIGURE 6.14.9. A pattern with symmetry group \mathcal{W}_2^3 .

Finally, we address Case (3), where \mathcal{W} contains glide reflections but not reflections. \mathcal{W}_2 -groups are the first case in which that can occur.

First let $\gamma = \tau_v \sigma_\ell$ be a glide reflection in standard form. Suppose there is a 2-center x on ℓ . Then $\rho_{(x,\pi)} = \sigma_\ell \sigma_m$, where m is the line through x perpendicular to ℓ , so

$$\gamma \rho_{(x,\pi)} = \tau_v \sigma_\ell \sigma_\ell \sigma_m = \tau_v \sigma_m = \sigma_{(\tau_{\frac{1}{2}v}(m))},$$

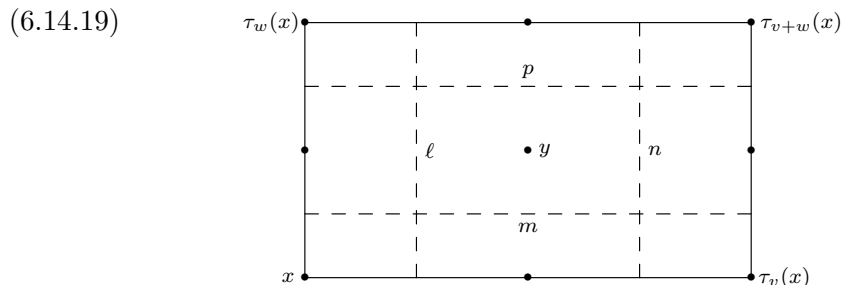
as we saw in the frieze group \mathcal{F}_2^2 . In particular, this cannot happen in a wallpaper group without reflections.

So we now assume \mathcal{W} has no reflections, but does have glide reflections. Thus, there is no 2-center on a glide axis, and we can apply both Lemma 6.14.8 and Proposition 6.14.10. We obtain that there is a rectangular grid of glide axes and that each rectangular box contains exactly one 2-center that occurs at its center. Moreover, if ℓ is a glide axis for \mathcal{W} and if m is a glide axis orthogonal to ℓ , then the glide axes parallel to ℓ are $\{\tau_{\frac{k}{2}v}(\ell) : k \in \mathbb{Z}\}$ and the glide axes parallel to m are $\{\tau_{\frac{k}{2}w}(m) : k \in \mathbb{Z}\}$, where τ_v is a shortest translation perpendicular to ℓ and τ_w is a shortest translation parallel to ℓ . Thus, precisely as was the case for \mathcal{W}_2^3 , the set of 2-centers for \mathcal{W} is

$$\{\tau_{\frac{1}{2}(kv+rw)}(x) : k, r \in \mathbb{Z}\}.$$

These are precisely the 2-centers in $\mathcal{W}_2(\Lambda_{\mathcal{B}}, x)$ where $\mathcal{B} = v, w$, so $\mathcal{T}(\mathcal{W}) = \mathcal{T}_{\Lambda_{\mathcal{B}}}$, and the fundamental region R for $\mathcal{T}(\mathcal{W})$ is as follows, where the dotted

lines denote glide axes.

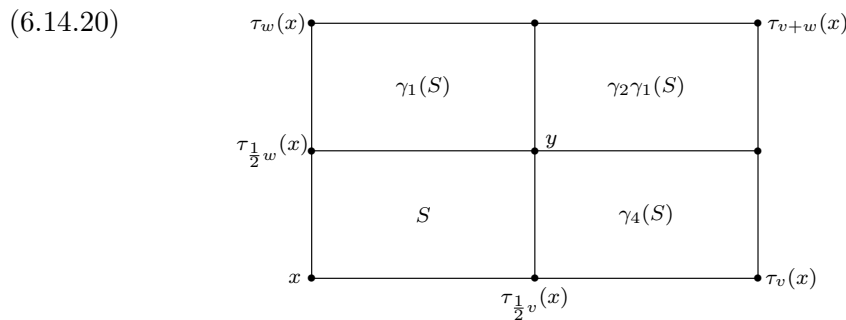


There can be no glide axes parallel to neither ℓ nor m , as if $\ell \cap q \neq \emptyset$, then the product of a glide reflection with axis ℓ and a glide reflection with axis q is the composite of $\sigma_\ell \sigma_q$ with translations on either side, and gives a rotation about some point by twice the directed angle from q to ℓ . So (6.14.5) gives the entire picture of the symmetries of \mathcal{W} in R . We call this group \mathcal{W}_2^4 .

Let us analyze the primitive glide reflections with the axes indicated in (6.14.19). Specifically, let $\gamma_1 = \tau_{\frac{1}{2}w} \sigma_\ell$, $\gamma_2 = \tau_{\frac{1}{2}v} \sigma_p$, $\gamma_3 = \tau_{\frac{1}{2}w} \sigma_n$ and $\gamma_4 = \tau_{\frac{1}{2}v} \sigma_m$. Then

$$\gamma_2 \gamma_1 = \tau_{\frac{1}{2}v} \sigma_p \tau_{\frac{1}{2}w} \sigma_\ell = \sigma_p \tau_{\frac{1}{2}w} \tau_{\frac{1}{2}v} \sigma_\ell = \sigma(\tau_{-\frac{1}{4}w}(p)) \sigma(\tau_{\frac{1}{4}v}(\ell)) = \rho_{(y,\pi)}.$$

Let S be the lower left rectangular block in (6.14.5). Then the following illustrates that S is a fundamental region for \mathcal{W}_2^4 .



The orbit space of \mathcal{W}_2^4 is perhaps the most interesting of all the orbit spaces of wallpaper groups. (Some might prefer the Klein bottle, which is the orbit space of \mathcal{W}_1^3 , below.) The lower edge of S is identified to upper edge by γ_1 , which takes x to y and takes $\tau_{\frac{1}{2}v}(x)$ to $\tau_{\frac{1}{2}w}(x)$. Thus, the upper and lower edges are identified via the same twist used in making the Möbius band. And the Möbius band is what we get if we make that identification and nothing else.

Similarly, the left-hand edge is identified with the right-hand edge via γ_4 , which applies the same twist, taking x to y and $\tau_{\frac{1}{2}w}(x)$ to $\tau_{\frac{1}{2}v}(x)$. So again, if we made only this identification, but not the identification of the lower edge with the upper, we would get a Möbius band.

But here, we are making both identifications at once. The result is a rectangle with the identifications indicated by the arrows below.



Thus, the double-headed arrows are identified with each other preserving the direction of the arrowheads, and similarly for the single-headed arrows. The result is what's known as a real projective space, \mathbb{RP}^2 . It is a surface, or 2-dimensional manifold, as can be shown by methods similar to those applied to the Klein bottle in Appendix A.

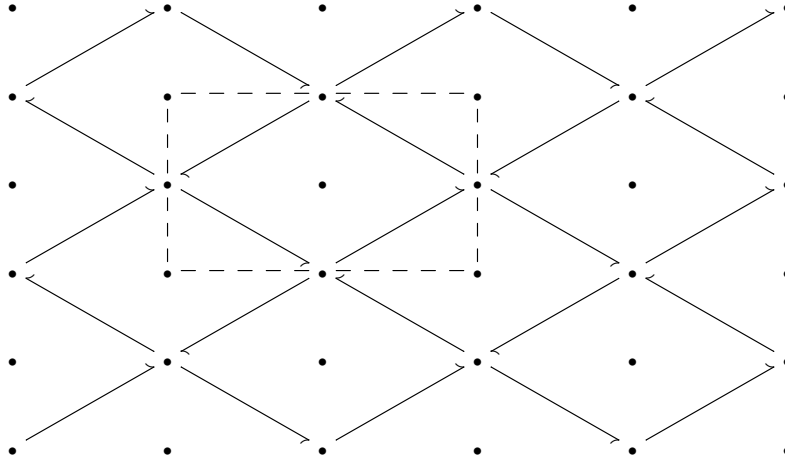


FIGURE 6.14.10. A pattern with symmetry group \mathcal{W}_2^4 .

A pattern with symmetry group \mathcal{W}_2^4 is given in Figure 6.14.10. The marked points are its points of symmetry, and the dotted region superimposed on it is a fundamental region R for $\mathcal{T}(\mathcal{W})$.

From the isotropy data and reflections present we can see that the groups \mathcal{W}_2^1 – \mathcal{W}_2^4 are all distinct. But we have exhausted all possibilities for adding orientation-reversing isometries to \mathcal{W}_2 to obtain a wallpaper group. We have obtained the following.

Theorem 6.14.15. *There are exactly four different wallpaper groups obtained by adding orientation-reversing isometries to \mathcal{W}_2 . In one of them, \mathcal{W}_2^2 , all four \mathcal{T} -orbits of 2-centers have isotropy D_4 . The fundamental regions for $\mathcal{T}(\mathcal{W}_2^2)$ and for \mathcal{W}_2^2 are given in (6.14.17). The orbit space is S . The fundamental region for $\mathcal{T}(\mathcal{W}_2^2)$ is rectangular.*

In a second, \mathcal{W}_2^1 , there are two \mathcal{T} -orbits of 2-centers with isotropy D_4 and two with isotropy C_2 . The fundamental regions for $\mathcal{T}(\mathcal{W}_2^1)$ and for \mathcal{W}_2^1 are given in (6.14.16). The orbit space is a cone. The fundamental region for $\mathcal{T}(\mathcal{W}_2^1)$ is rhombic.

A third, \mathcal{W}_2^3 has four \mathcal{T} -orbits of 2-centers with isotropy C_2 , but does have lines of symmetry. The fundamental regions for $\mathcal{T}(\mathcal{W}_2^3)$ and for \mathcal{W}_2^3 are given in (6.14.18). The orbit space looks like a pillow case. The fundamental region for $\mathcal{T}(\mathcal{W}_2^3)$ is rectangular.

The fourth, \mathcal{W}_2^4 , has no lines of symmetry, but does admit glide reflections. Each \mathcal{T} -orbit of 2-centers has isotropy C_2 . The fundamental region for $\mathcal{T}(\mathcal{W}_2^4)$ is given in (6.14.19) and that for \mathcal{W}_2^4 is given in (6.14.20). The orbit space is a real projective space $\mathbb{R}P^2$. The fundamental region for $\mathcal{T}(\mathcal{W}_2^4)$ is rectangular.

6.14.5. Wallpaper groups with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_1$. \mathcal{W}_1 is just a translation lattice:

$$\mathcal{W}_1 = \mathcal{T}(\mathcal{W}_1) = \mathcal{T}_\Lambda = \{\tau_z : z \in \Lambda\}$$

for some arbitrary lattice $\Lambda \subset \mathbb{R}^2$. But the presence of orientation-reversing isometries imposes stronger conditions on what lattices may occur.

Let us first suppose that \mathcal{W} contains a reflection σ_ℓ . Since \mathcal{W}_1 has index 2 in \mathcal{W} , this says

$$(6.14.21) \quad \mathcal{W} \setminus \mathcal{W}_1 = \{\tau_z \sigma_\ell : \tau_z \in \mathcal{W}_1\}.$$

Moreover, the isometries $\tau_z \sigma_\ell$ are reflections if $z \perp \ell$ and are glide reflections otherwise. Whether some of these glide reflections are essential will depend on the relationship between ℓ and the translations that occur in \mathcal{W}_1 .

Let ℓ_ϕ be the line through the origin parallel to ℓ . By Lemma 6.14.2,

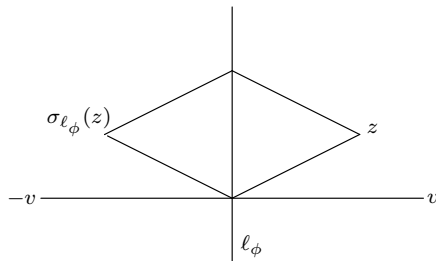
$$(6.14.22) \quad \sigma_\ell \tau_z \sigma_\ell^{-1} = \tau_{\sigma_{\ell_\phi}(z)}$$

for any $z \in \mathbb{R}^2$. In particular, if $\tau_z \in \mathcal{W}_1 = \mathcal{T}_\Lambda$, so is $\tau_{\sigma_{\ell_\phi}(z)}$.

By Lemma 6.14.4, there is an infinite family of lines of symmetry parallel to ℓ , $\{\tau_{\frac{k}{2}v}(\ell) : k \in \mathbb{Z}\}$, where τ_v is a shortest translation perpendicular to ℓ . These are the only lines of symmetry parallel to ℓ , and there can be no other lines of symmetry for \mathcal{W} , as if two lines of symmetry intersect, there is a rotation about their point of intersection.

So far, we know there are translations $\{\tau_v^k : k \in \mathbb{Z}\}$ in \mathcal{T}_Λ perpendicular to ℓ (and no other translations perpendicular to ℓ , as τ_v is a shortest such translation). We next claim there are translations parallel to ℓ . To see this, suppose $\tau_z \in \mathcal{T}_\Lambda$ is neither parallel nor perpendicular to ℓ . Then we obtain the following situation.

$$(6.14.23)$$



By construction, ℓ_ϕ is the perpendicular bisector of $\overline{z\sigma_{\ell_\phi}(z)}$, so the midpoint $\frac{1}{2}(z + \sigma_{\ell_\phi}(z))$ is on ℓ_ϕ , as is the origin. So $z + \sigma_{\ell_\phi}(z)$ is on ℓ_ϕ . But $z \in \Lambda$, as is $\sigma_{\ell_\phi}(z)$ by (6.14.22), so $\tau_{(z+\sigma_{\ell_\phi}(z))}$ is a translation parallel to ℓ in \mathcal{T}_Λ .

Let τ_w be a shortest translation parallel to ℓ in \mathcal{T}_Λ . Then as shown earlier, $\{\tau_w^k : k \in \mathbb{Z}\}$ is the set of all translations parallel to ℓ in \mathcal{T}_Λ .

Lemma 6.14.16. *Let \mathcal{W} be wallpaper group with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_1 = \mathcal{T}_\Lambda$ and let $\sigma_\ell \in \mathcal{W}$. Let τ_v be a shortest translation perpendicular to ℓ and τ_w a shortest translation parallel to ℓ . Let $\tau_z \in \mathcal{W}$. Then*

$$(6.14.24) \quad z = \frac{k}{2}w + \frac{r}{2}v \quad \text{for } k, r \in \mathbb{Z}.$$

Moreover, k and r are either both even or both odd.

Proof. Since v and w are perpendicular, they form a basis of \mathbb{R}^2 as a vector space over \mathbb{R} . So $z = sw + tv$ for $s, t \in \mathbb{R}$. Now $\tau_z(\ell)$ is the axis for the reflection $\tau_z\sigma_\ell\tau_z^{-1}$ by Theorem 5.5.20, so $\tau_z(\ell) \in \{\tau_{\frac{k}{2}v}(\ell) : k \in \mathbb{Z}\}$. On the other hand $\tau_z(\ell) = \tau_{tv}(\tau_{sw}(\ell)) = \tau_{tv}(\ell)$, as sw is parallel to ℓ . Thus, t is an integral multiple of $\frac{1}{2}$.

Again by Theorem 5.5.20, $\sigma_\ell\tau_z\sigma_\ell^{-1} = \tau_{\sigma_{\ell_\phi}(z)}$, so $\sigma_{\ell_\phi}(z) \in \Lambda$. Since $w \in \ell_\phi$ and $v \perp \ell_\phi$, we have

$$\sigma_{\ell_\phi}(sw + tv) = sw - tv,$$

so $z + \sigma_{\ell_\phi}(z) = 2sw \in \Lambda$. So s is an integral multiple of $\frac{1}{2}$ also.

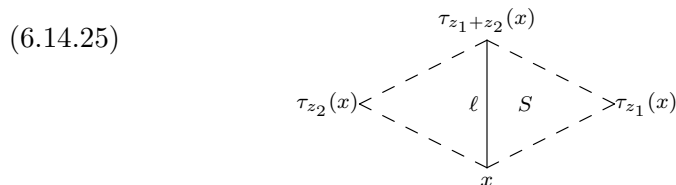
We know that neither $\frac{1}{2}v$ nor $\frac{1}{2}w$ is in Λ . We claim this implies that if $z = \frac{k}{2}w + \frac{r}{2}v \in \Lambda$, then either k and r are both even or k and r are both odd. To see this, note that if $z = \frac{2m+1}{2}w + \frac{2n}{2}v$, then $z - mw - nv = \frac{1}{2}w$. Similarly, if $z = \frac{2m}{2}w + \frac{2n+1}{2}v$, then $z - mw - nv = \frac{1}{2}v$. \square

It is possible that v, w is a \mathbb{Z} -basis for Λ , in which case k and r must always be even. Otherwise, there exists a pair of odd integers k, r with $\frac{k}{2}w + \frac{r}{2}v \in \Lambda$. But then, subtracting multiples of v and w as above, we see that $z_1 = \frac{1}{2}w + \frac{1}{2}v$ and $z_2 = \frac{1}{2}w - \frac{1}{2}v$ are in Λ . Since $w = z_1 + z_2$ and $v = z_1 - z_2$, the lattice $\Lambda_{\mathcal{B}}$ with $\mathcal{B} = z_1, z_2$ must contain all sums $\frac{k}{2}w + \frac{r}{2}v$ with k and r either both even or both odd. Since such sums do form an additive subgroup of \mathbb{R}^2 , the presence of a single such sum in which both k and r are odd forces $\Lambda = \Lambda_{\mathcal{B}}$.

Of course, $z_2 = \sigma_{\ell_\phi}(z_1)$. We have shown that exactly one of the following must hold:

- (1) $z_1 = \frac{1}{2}w + \frac{1}{2}v$ and $\sigma_{\ell_\phi}(z_1) = z_2 = \frac{1}{2}w - \frac{1}{2}v$ form a \mathbb{Z} -basis for Λ .
- (2) v and w form a \mathbb{Z} -basis for Λ .

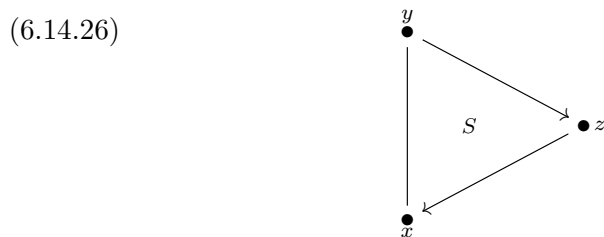
In Case (1), for $x \in \ell$, we obtain a fundamental region R for $\mathcal{T}(\mathcal{W})$ as follows.



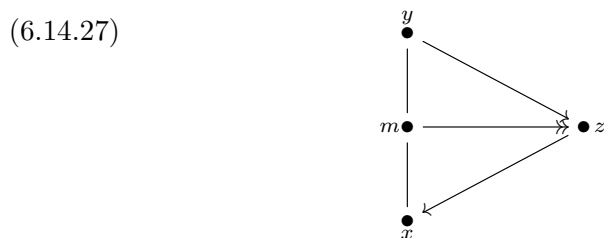
Here, the dotted lines are the edges of R and the solid line ℓ is the only line of symmetry meeting R in more than one point. Since $z_2 = \sigma_{\ell_\phi}(z_1)$, R is rhombic. We can obtain any rhombus we like in this way by varying the lengths of v and w . We call this group \mathcal{W}_1^1 , and the triangle labelled S forms a fundamental region for \mathcal{W}_1^1 . Its orbit space is rather complicated, as we identify the edge $\overline{\tau_{z_1}(x)\tau_{z_1+z_2}(x)}$ with the edge $\overline{x\tau_{z_1}(x)}$ via the composite $\sigma_\ell\tau_{-z_1}$. This identification applies a twist, identifying $\tau_{z_1}(x)$ to x and identifying $\tau_{z_1+z_2}(x)$ to $\tau_{z_1}(x)$, wrapping these two edges up into a single circle.

Lemma 6.14.17. *The orbit space of \mathcal{W}_1^1 is a Möbius band.*

Proof. Write $y = \tau_{z_1+z_2}(x)$ and $z = \tau_{z_1}(x)$. Then the orbit space is the result of making the indicated identifications on S :



Let m be the midpoint of \overline{xy} . We shall cut S along the segment \overline{mz} :



This separates S into two right triangles, which we then glue to each other along their hypotenuses via that stated identification between \overline{yz} and \overline{zx} . We obtain the following rectangle, where the opposite edges are to be identified with one another according to the orientations specified by the double

arrows obtained from the cut we made:



The result of this identification is a Möbius band. □

Another feature of \mathcal{W}_1^1 is the presence of essential glide reflections.

Proposition 6.14.18. \mathcal{W}_1^1 has essential glide axes half way between each closest pair of lines of symmetry. These are the only essential glide axes for \mathcal{W}_1^1 .

Proof. Let v, w, z_1 and z_2 be as above. Let m be a line of symmetry for \mathcal{W}_1^1 . Then Lemma 5.5.16 gives

$$\tau_{z_1}\sigma_m = \tau_{\frac{w}{2}}\tau_{\frac{v}{2}}\sigma_m = \tau_{\frac{w}{2}}\sigma_{\tau_{\frac{v}{4}}(m)}$$

Since $\tau_{\frac{w}{2}} \notin \Lambda$, this is an essential (in fact primitive) glide reflection. Its axis is half way between m and $\tau_{\frac{v}{2}}(m)$, a closest pair of lines of symmetry.

To see this gives all essential glide axes for \mathcal{W}_1^1 , note that the orientation-reversing elements of \mathcal{W}_1^1 have the form $\tau_{kz_1+rz_2}\sigma_\ell$ with $k, r \in \mathbb{Z}$, by (6.14.21), since z_1, z_2 is a \mathbb{Z} -basis for Λ . Now,

$$kz_1 + rz_2 = k\left(\frac{1}{2}w + \frac{1}{2}v\right) + r\left(\frac{1}{2}w - \frac{1}{2}v\right) = \frac{k+r}{2}w + \frac{k-r}{2}v,$$

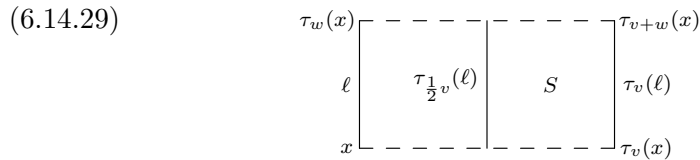
so

$$\tau_{kz_1+rz_2}\sigma_\ell = \tau_{\frac{k+r}{2}w}\sigma_n,$$

where $n = \tau_{\frac{k-r}{4}v}(\ell)$. If $k+r$ is even, so is $k-r$. In this case n is a line of symmetry for \mathcal{W}_1^1 and $\tau_{\frac{k+r}{2}w} \in \Lambda$, so the result is either a reflection or an inessential glide reflection. If $k+r$ is odd, so is $k-r$, and the result is an essential glide reflection whose axis is $\tau_{\frac{v}{4}}(m)$, where m is the line of symmetry $\tau_{\frac{k-r-1}{4}v}(\ell)$. □

A pattern with symmetry group \mathcal{W}_1^1 is given in Figure 6.14.11. The lines of symmetry in it are horizontal and follow the arrows.

In Case (2), v, w form a \mathbb{Z} -basis for Λ , so a fundamental region R for $\mathcal{T}(\mathcal{W})$ is given as follows.



We call this group \mathcal{W}_1^2 . A fundamental region for \mathcal{W}_1^2 is given by the square marked S . The only identification on S is the identification of its lower

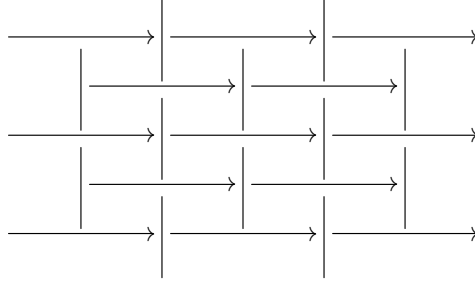


FIGURE 6.14.11. A pattern with symmetry group \mathcal{W}_1^1 .

edge to its upper edge via τ_w . The orbit space is a cylinder. An argument similar to that of Proposition 6.14.18 shows the following.

Proposition 6.14.19. *There are no essential glide reflections in \mathcal{W}_1^2 .*

A pattern with symmetry group \mathcal{W}_1^2 is given in Figure 6.14.12.

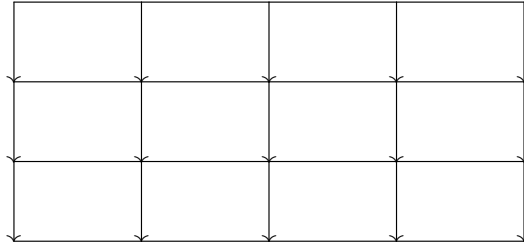


FIGURE 6.14.12. A pattern with symmetry group \mathcal{W}_1^2 .

We now consider the case where \mathcal{W} has glide reflections by no reflections. Let ℓ be a glide axis for \mathcal{W} . Let τ_w be a shortest translation parallel to ℓ and write $\gamma_\ell = \tau_{\frac{1}{2}w}\sigma_\ell$. By Lemma 6.14.7, the glide reflections in \mathcal{W} with axis ℓ are precisely $\{\gamma_\ell^{2k+1} = \tau_{\frac{2k+1}{2}w}\sigma_\ell : k \in \mathbb{Z}\}$. The glide reflections γ_ℓ and γ_ℓ^{-1} are called primitive for ℓ . Note $\gamma_\ell^2 = \tau_w$.

If m is a glide axis parallel to ℓ , then this same τ_w is the shortest translation parallel to m , so the primitive glide reflections with axis m are $\gamma_m = \tau_{\frac{1}{2}w}\sigma_m$ and its inverse. By Lemma 6.14.8, the glide axes parallel to ℓ are precisely $\{\tau_{\frac{k}{2}v}(\ell) : k \in \mathbb{Z}\}$, where v is a shortest translation perpendicular to ℓ .

These are in fact the only glide axes for \mathcal{W} : if q is a glide axis not parallel to ℓ , then $\ell \cap q \neq \emptyset$. If $\tau_z\sigma_q$ is a glide reflection with axis q , then

$$\tau_z\sigma_q\tau_{\frac{w}{2}}\sigma_\ell = \tau_z\sigma_q\sigma_\ell\tau_{\frac{w}{2}}$$

is the product of the rotation $\sigma_q\sigma_\ell$ with translations on either side, and hence is a rotation about some point by twice the directed angle from ℓ to q . Since there are no nonidentity rotations in \mathcal{W} , no such glide axis q can exist.

Thus, having found the glide axes in \mathcal{W} , and having identified the shortest translations, v and w , perpendicular and parallel to these axes, respectively, it suffices to determine the lattice Λ inducing $\mathcal{W}_1 = \mathcal{T}_\Lambda$, to find its relationship to v and w , and to describe its fundamental region R .

Lemma 6.14.20. *Let \mathcal{W} be a wallpaper group with $\mathcal{O}(\mathcal{W}) = \mathcal{W}_1 = \mathcal{T}_\Lambda$. Suppose \mathcal{W} contains no reflections, but does contain a glide reflection with axis ℓ . Let τ_v be a shortest translation perpendicular to ℓ and let τ_w be a shortest translation parallel to ℓ . Then v, w form a \mathbb{Z} -basis for Λ . Thus, for $x \in \ell$ we obtain a fundamental region R for $\mathcal{T}_\Lambda = \mathcal{W}_1$ as follows:*

$$(6.14.30) \quad \begin{array}{c} \tau_w(x) \text{---} \text{---} \text{---} \tau_{v+w}(x) \\ | \qquad \qquad \qquad | \qquad \qquad \qquad | \\ | \qquad \qquad \tau_{\frac{1}{2}v}(\ell) \qquad \qquad | \qquad \qquad \qquad | \\ | \qquad \qquad \qquad | \qquad \qquad \qquad | \qquad \qquad \qquad | \\ | \qquad \qquad \qquad | \qquad \qquad \qquad | \qquad \qquad \qquad | \\ x \text{---} \text{---} \text{---} \tau_v(x) \end{array}$$

Here, the dashed lines are glide axes and the solid lines are the other edges of R .

Proof. As above we write $\gamma_\ell = \tau_{\frac{w}{2}}\sigma_\ell$ for a primitive glide reflection in \mathcal{W} with axis ℓ .

The argument here is similar to that of Lemma 6.14.16. Let $\tau_z \in \mathcal{W}$. Then $z = sw + tv$ for some $s, t \in \mathbb{R}$. By Theorem 5.5.20,

$$\tau_z\gamma_\ell\tau_z^{-1} = \tau_z\tau_{\frac{w}{2}}\sigma_\ell\tau_z^{-1} = \tau_{\frac{w}{2}}\tau_z\sigma_\ell\tau_z^{-1} = \tau_{\frac{w}{2}}\sigma_m,$$

where $m = \tau_z(\ell) = \tau_{tv}(\ell)$. This is a primitive glide reflection in \mathcal{W} , so t must be an integral multiple of $\frac{1}{2}$. Again by Theorem 5.5.20,

$$\gamma\tau_z\gamma^{-1} = \tau_{\sigma_{\ell_\phi}(z)},$$

where σ_{ℓ_ϕ} is the line through the origin parallel to ℓ , as $\sigma_\ell = \tau_y\sigma_{\ell_\phi}$ for some $y \perp \ell$. As in the proof of Lemma 6.14.16, this implies s an integral multiple of $\frac{1}{2}$. So $z = \frac{k}{2}w + \frac{r}{2}v$ for $k, r \in \mathbb{Z}$, and again as in the lemma, k and r are either both even or both odd, as neither $\frac{w}{2}$ nor $\frac{v}{2}$ is in Λ .

But if k and r are both odd, then $z_1 = \frac{w}{2} + \frac{v}{2} \in \Lambda$. However,

$$\tau_{z_1}\gamma = \tau_{\frac{w}{2}}\tau_{\frac{v}{2}}\tau_{\frac{w}{2}}\sigma_\ell = \tau_w\tau_{\frac{v}{2}}\sigma_\ell = \tau_w\sigma_{\tau_{\frac{v}{4}}(\ell)},$$

as $v \perp \ell$. This is a glide reflection with axis $\tau_{\frac{v}{4}}(\ell)$, which is not a glide axis for \mathcal{W} . Thus, k and r must be both even, and v, w is a \mathbb{Z} -basis for Λ . And this implies (6.14.30) is a fundamental region R for $\mathcal{T}_\Lambda = \mathcal{W}_1$. \square

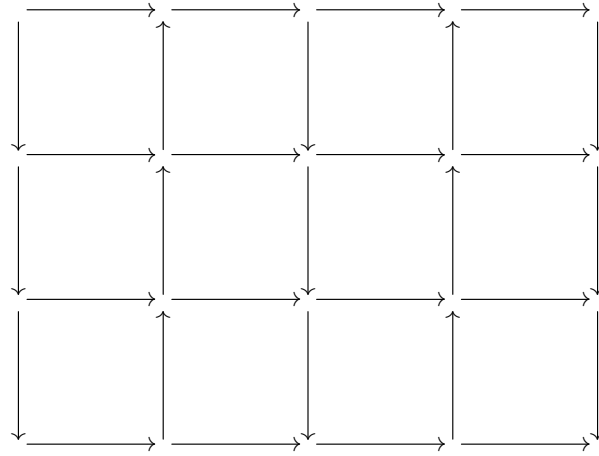
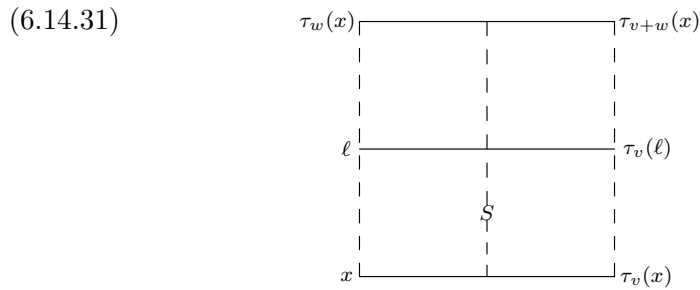


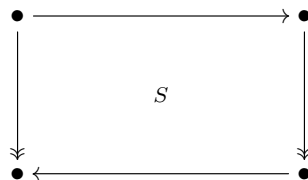
FIGURE 6.14.13. A pattern with symmetry group \mathcal{W}_1^3 .

Lemma 6.14.20 completely characterizes the unique wallpaper group with glide axis ℓ and with shortest translations τ_v and τ_w as stated. We call this group \mathcal{W}_1^3 .

Despite the similarity in appearance between the fundamental regions R for translation in \mathcal{W}_1^2 and \mathcal{W}_1^3 (here, glide axes replace lines of symmetry), the fundamental regions S are quite different. The fundamental region S for \mathcal{W}_1^3 , is the lower half of R . The glide reflection $\tau_{\frac{w}{2}}\sigma_{\tau_{\frac{v}{2}}(\ell)}$ carries the lower half onto the upper half with a twist:



In particular, while the left edge of S is identified with the right edge by translation, its lower edge is identified with its upper edge with a twist. So the orbit space is a rectangle with the following identifications on its edges:



These are precisely the identifications used to construct the Klein bottle in Figure A.3.1. In particular, the Klein bottle is studied extensively in Appendix A. It is shown there to be a 2-dimensional manifold, or surface. We shall confine ourselves here to noting that if we just make the identification between the lower and upper edge, we obtain a Möbius strip, but if we just make the identification between the left and right edges, we obtain a cylinder. So we can either think of the Klein bottle as obtained by gluing together the two edges of a cylinder with an orientation reversal, or can think of it as obtained from an identification on the boundary of a Möbius strip. A pattern with symmetry group \mathcal{W}_1^3 is given in Figure 6.14.13.

We have shown the following.

Theorem 6.14.21. *There are exactly three wallpaper groups \mathcal{W} containing orientation-reversing isometries such that $\mathcal{O}(\mathcal{W}) = \mathcal{W}_1$. In the first of them, \mathcal{W}_1^1 , there is a rhombic fundamental region for $\mathcal{T}(\mathcal{W}_1^1)$ with a line of symmetry as its diagonal. See (6.14.25) for fundamental regions for $\mathcal{T}(\mathcal{W}_1^1)$ and \mathcal{W}_1^1 . The orbit space is a Möbius band.*

In the second, \mathcal{W}_1^2 , the fundamental region for its translation subgroup, shown in (6.14.29), is rectangular, with lines of symmetry parallel to one set of edges. The orbit space is a cylinder.

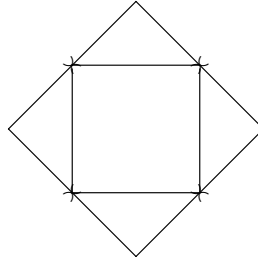
Finally, \mathcal{W}_1^3 has no lines of symmetry, but has glide reflections. The fundamental regions for $\mathcal{T}(\mathcal{W}_1^3)$ and \mathcal{W}_1^3 are shown in (6.14.30) and (6.14.31), respectively. The orbit space is a Klein bottle.

6.15. Exercises.

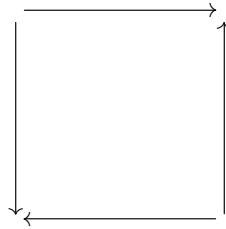
- Let C be a polytope in \mathbb{R}^k with centroid \bar{x} and let D be a polytope in \mathbb{R}^{n-k} with centroid \bar{y} . Show that $\begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}$ is the centroid of $C \times D \subset \mathbb{R}^n$.
- What are the isotropy subgroups of the following points under the action of $\mathcal{S}([-1, 1]^3)$? What points are these on the cube? (What are the faces of which these are interior points? Are they centroids?)
 - $e_1 + e_2 + e_3$.
 - $e_2 + e_3$.
 - $\frac{1}{2}e_1 + e_2 + e_3$.
 - e_3 .
 - $\frac{1}{2}e_3$.
- What is the orbit of $e_1 + e_2 + e_3$ under the action of $\mathcal{S}([-1, 1]^3)$? What is the orbit of e_3 ?
- Consider the action of $\mathcal{S}([-1, 1]^n)$ on $[-1, 1]^n$. Describe the orbit and isotropy subgroup of each of the following points.
 - $e_1 + \cdots + e_n$.
 - e_n .
- Consider the action of $\mathcal{S}(P_n)$ on the regular n -gon P_n .
 - Describe the orbit and isotropy subgroup of the vertex v_i .
 - Describe the orbit and isotropy subgroup of the midpoint of an edge.

- (c) Describe the orbit and isotropy subgroup of a point on an edge that is neither the midpoint of the edge nor a vertex.
6. Show that both \mathcal{F}_1^2 and \mathcal{F}_1^3 are isomorphic to subgroups of \mathcal{F}_2^2 .
 7. Show that \mathcal{W}_1^1 is isomorphic to a subgroup of \mathcal{W}_2^1 .
 8. Show that \mathcal{W}_1^3 is isomorphic to a subgroup of \mathcal{W}_2^4 .
 9. Show that both \mathcal{W}_2^1 and \mathcal{W}_2^4 are isomorphic to subgroups of \mathcal{W}_4^2 .
 10. Show that \mathcal{W}_1^2 is isomorphic to a subgroup of \mathcal{W}_2^2 .
 11. Show that both \mathcal{W}_2^1 and \mathcal{W}_2^2 are isomorphic to subgroups of \mathcal{W}_4^1 .
 12. Is \mathcal{W}_2^3 isomorphic to a subgroup of a \mathcal{W}_4 -group?
 13. Show that both \mathcal{W}_3^1 and \mathcal{W}_3^2 are isomorphic to subgroups of \mathcal{W}_6^1 .
 14. The following are rosette patterns. Indicate the following for each:
 - all lines of symmetry;
 - the shortest rotation preserving the pattern;
 - the name of the rosette group.

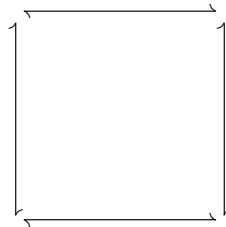
(a)



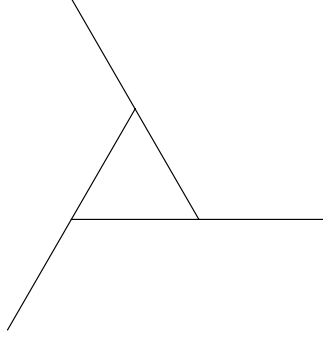
(b)



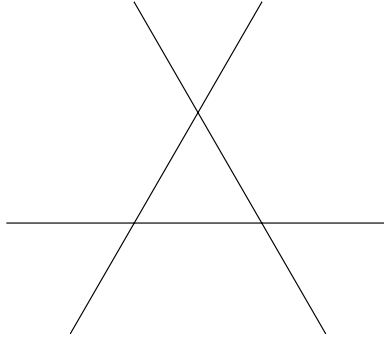
(c)



(d)



(e)



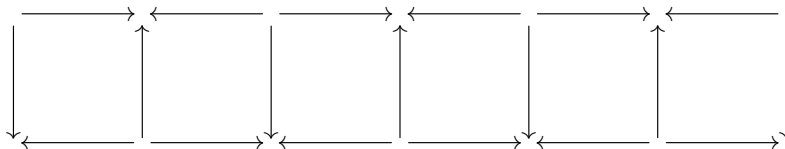
(f)



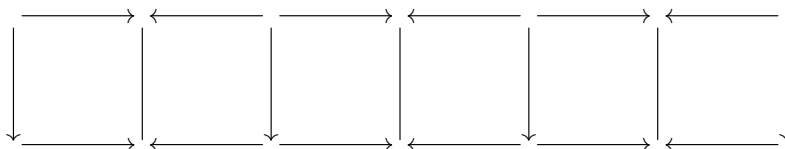
15. The following are frieze patterns. For each one, indicate the following:

- The shortest translation, τ_v , that preserves the pattern, X .
- All points of symmetry.
- All lines of symmetry.
- A fundamental region, R , for $\mathcal{T}(X)$.
- A fundamental region, S for $\mathcal{F} = \mathcal{S}(X)$.
- Which translations are squares of a glide reflection in \mathcal{F} ?
- What are the isotropy subgroups of the points of symmetry, if any?
- Which frieze group is \mathcal{F} ?

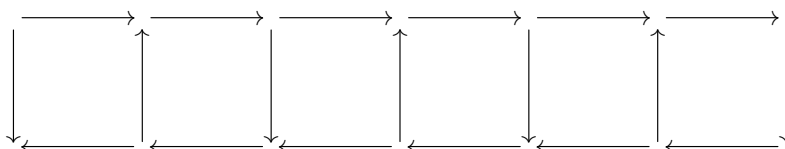
(a)



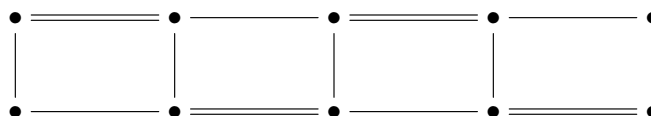
(b)



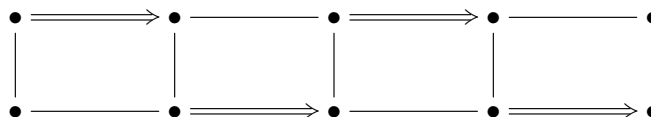
(c)



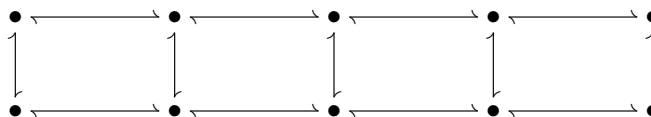
(d)



(e)



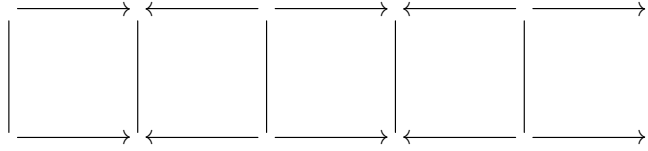
(f)



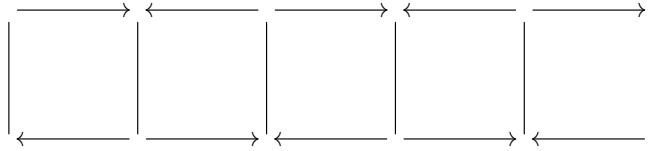
(g)



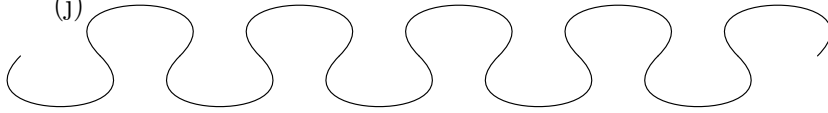
(h)



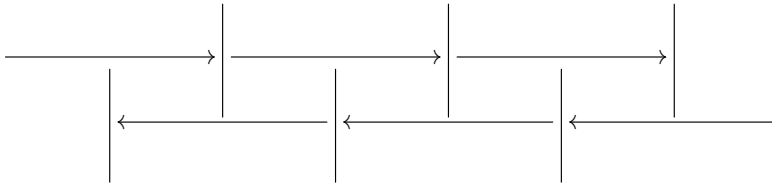
(i)



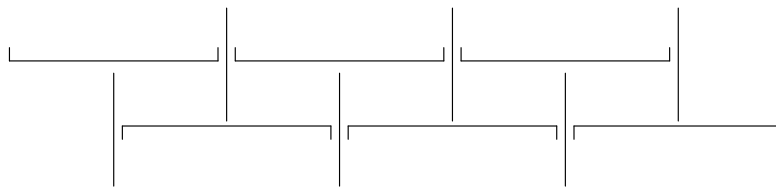
(j)



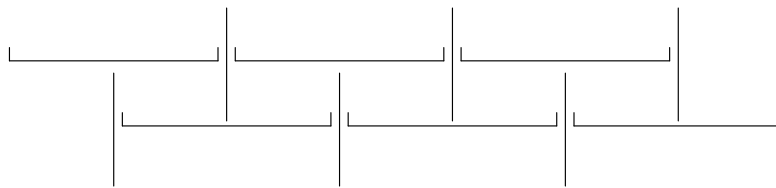
(k)



(l)



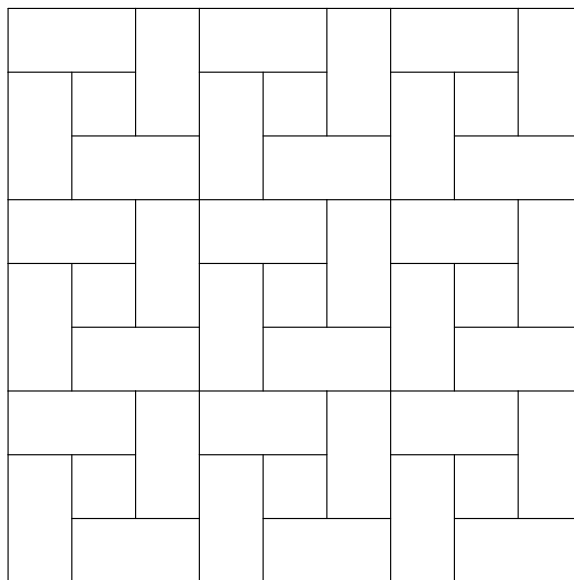
(m)



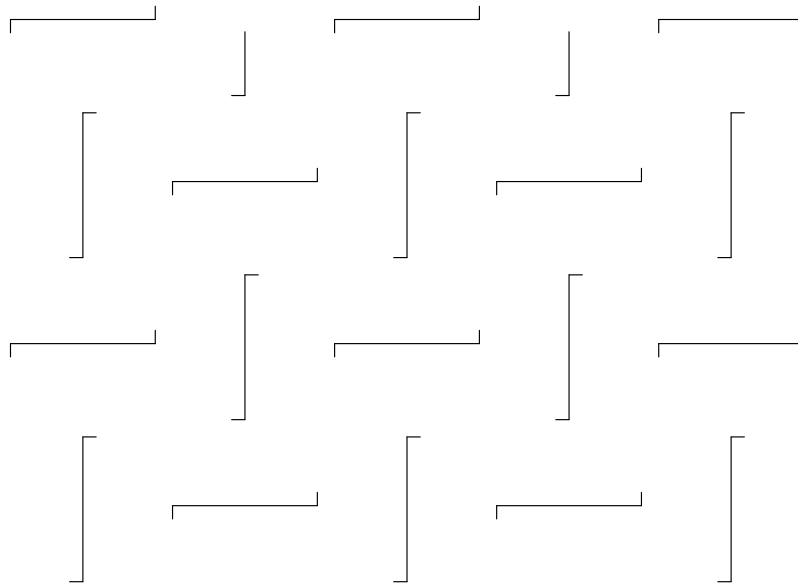
16. The following are wallpaper patterns. For each one, indicate the following:

- Shortest translations, τ_v and τ_w , in two different directions, that preserve the pattern and form the boundary of a fundamental region R for $\mathcal{T}(X)$.
- All n -centers for each possible n .
- All lines of symmetry.
- If there are glide reflections but no reflections, give the axes for the glide reflections.
- A fundamental region, R , for $\mathcal{T}(X)$. If $\mathcal{W} = \mathcal{S}(X)$ is a W_3 -group that contains lines of symmetry, base it at a 3-center on a line of symmetry. Otherwise base it at an n -center for the largest possible n .
- A fundamental region, S for \mathcal{W} .
- What are the \mathcal{T} -orbits of n -centers for each n ? What is their isotropy?
- Which wallpaper group is \mathcal{W} ?

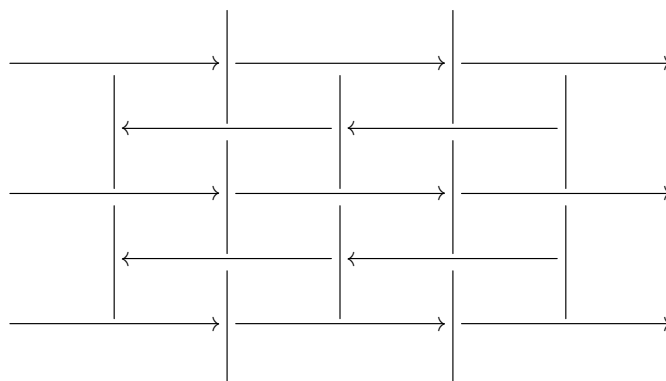
(a)



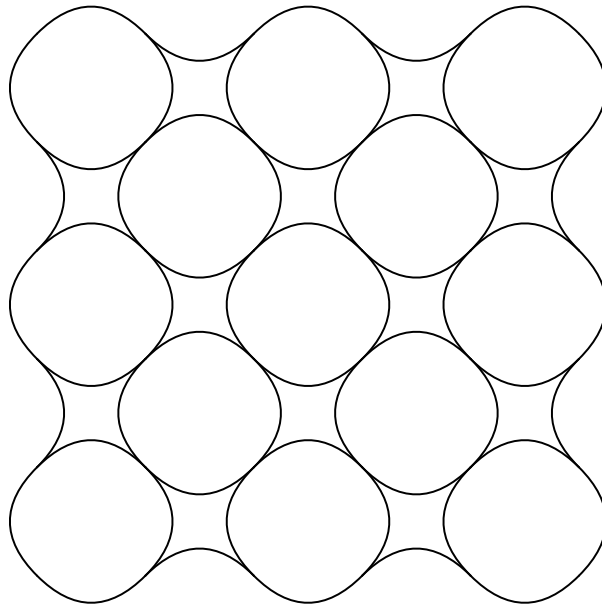
(b)



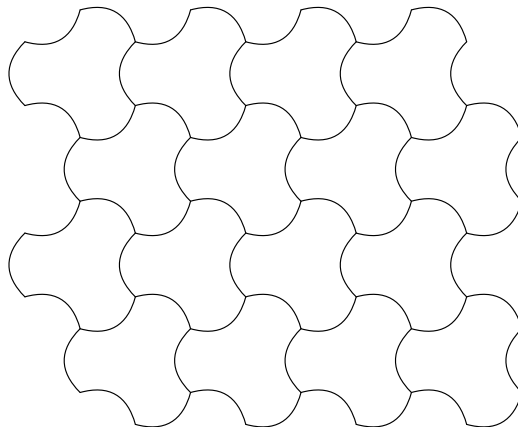
(c)



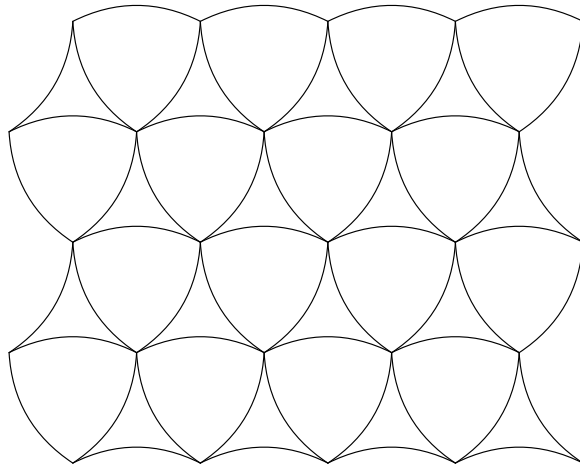
(d)



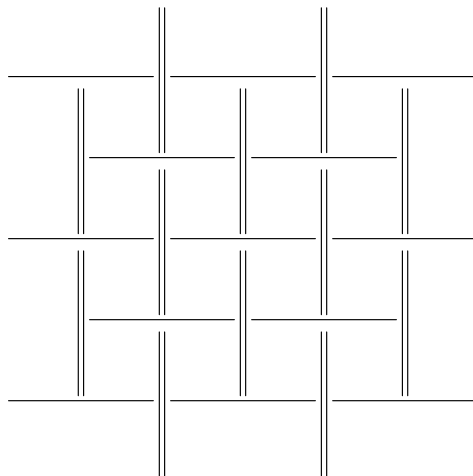
(e)



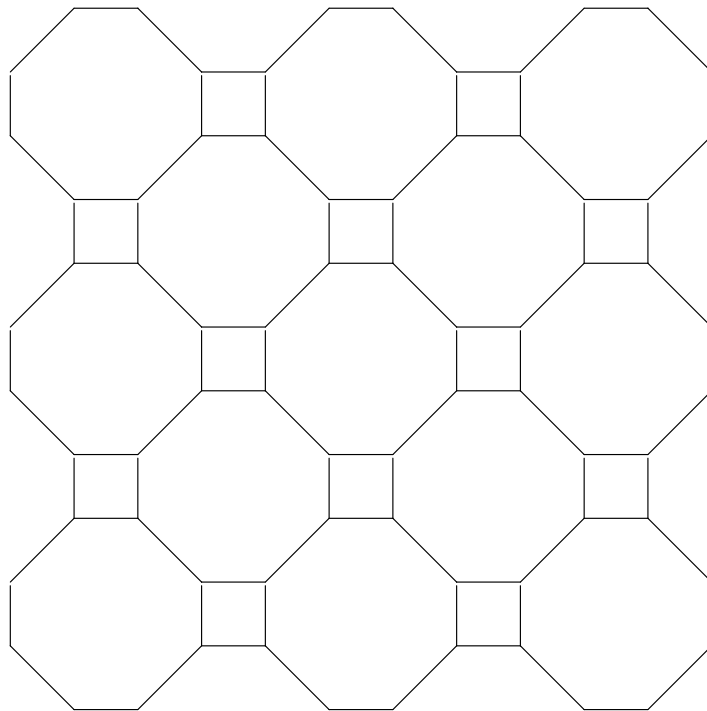
(f)



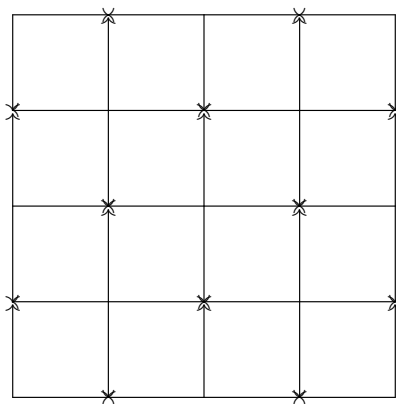
(g)



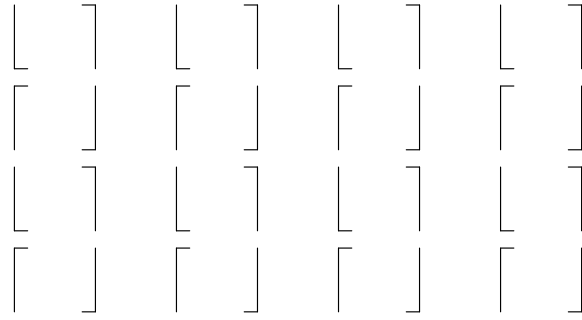
(h)



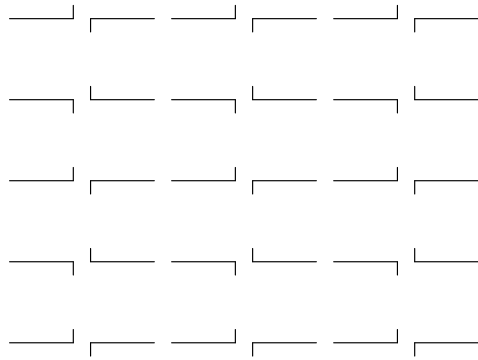
(i)



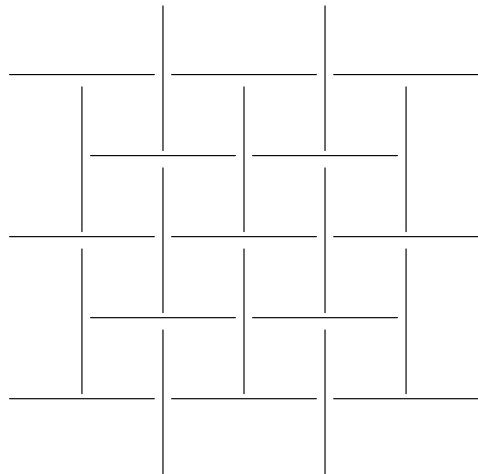
(j)



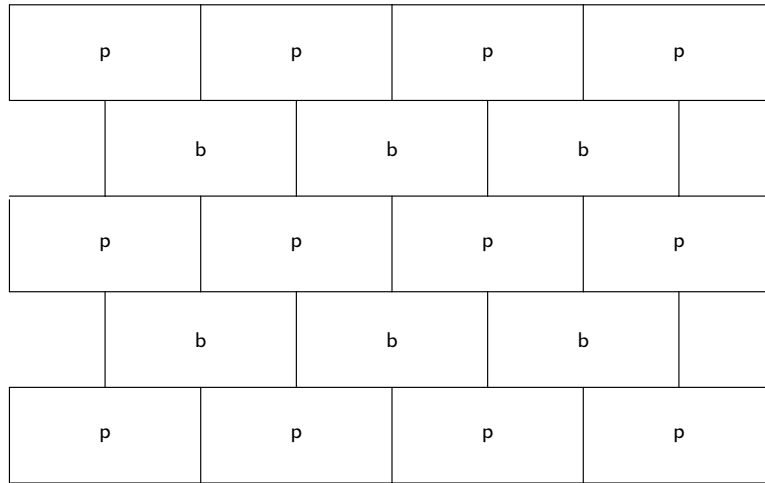
(k)



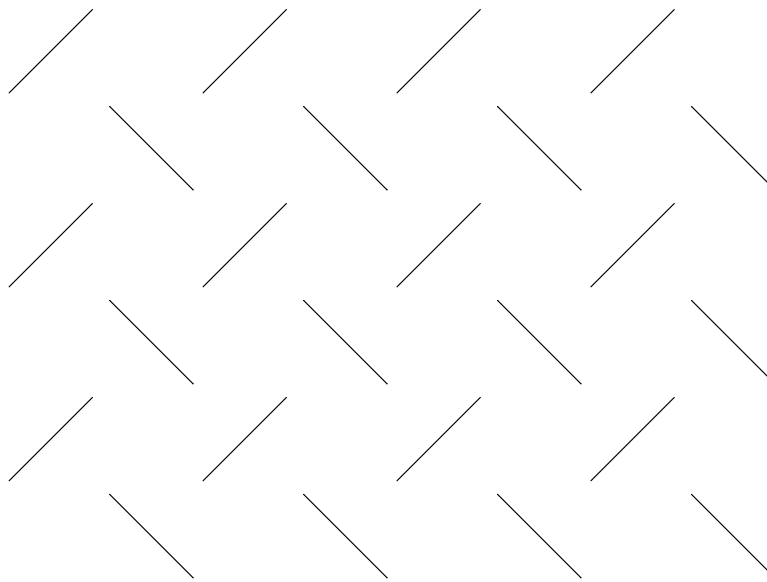
(l)



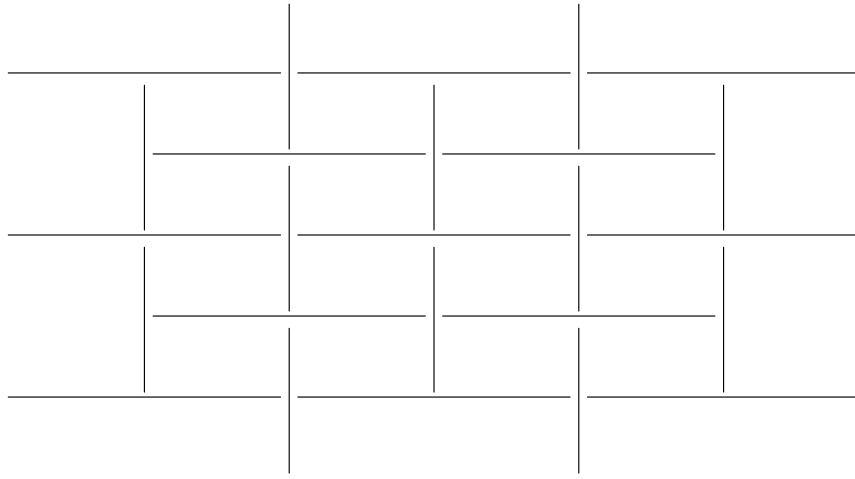
(m)



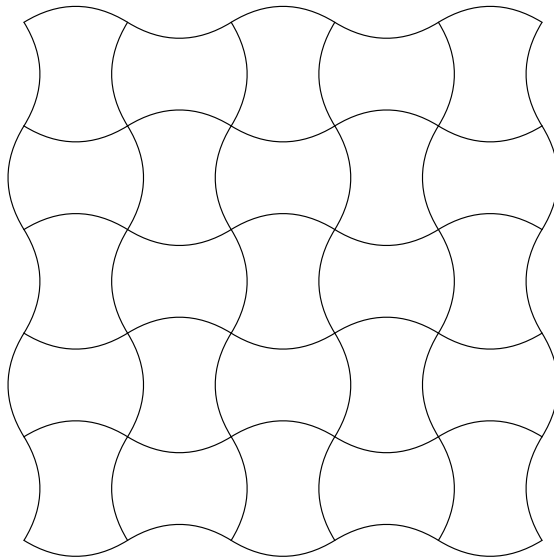
(n)



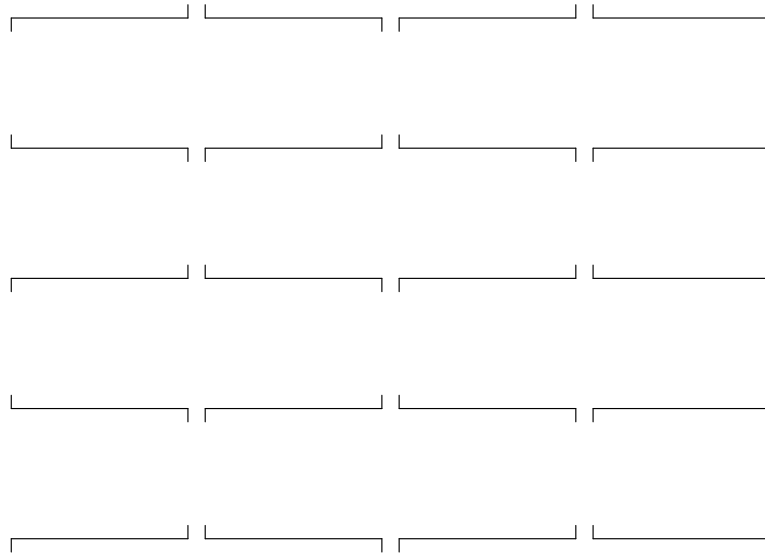
(o)



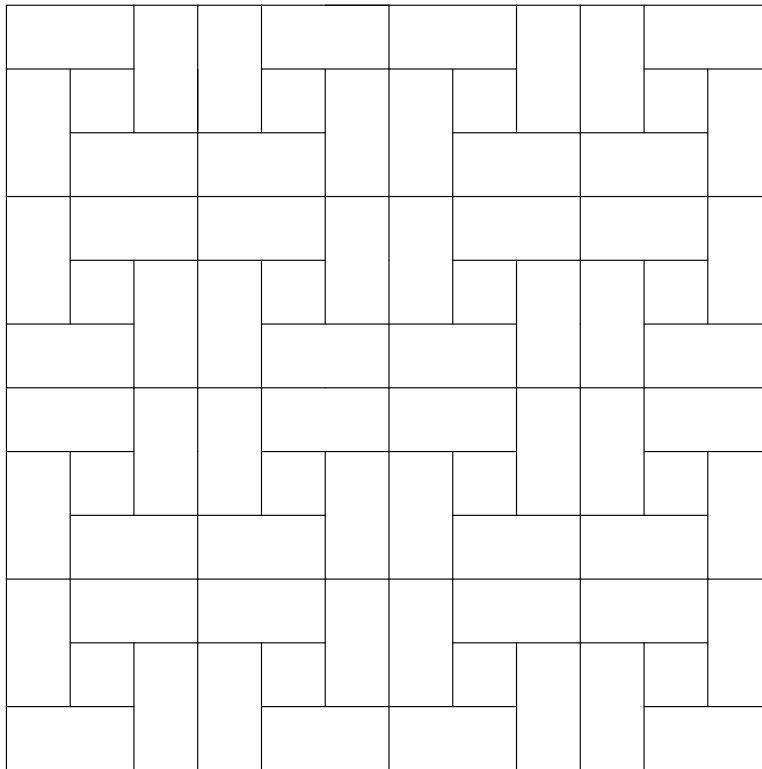
(p)



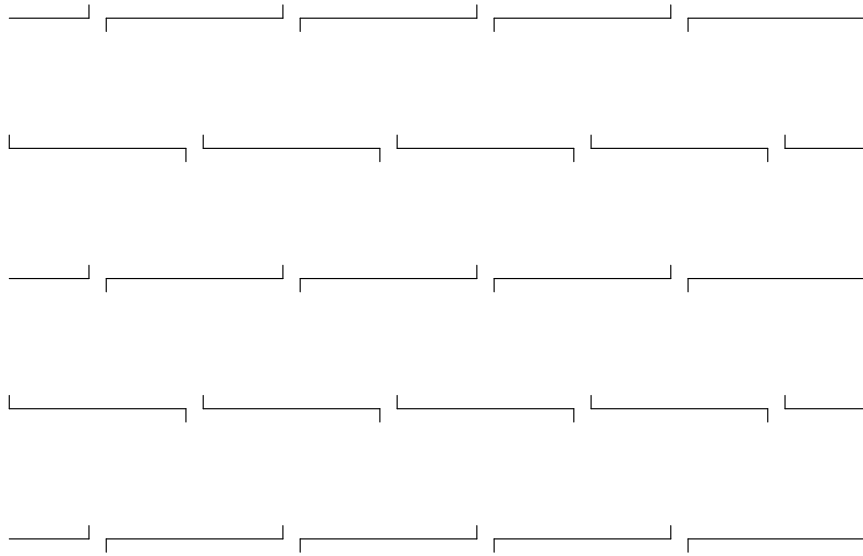
(q)



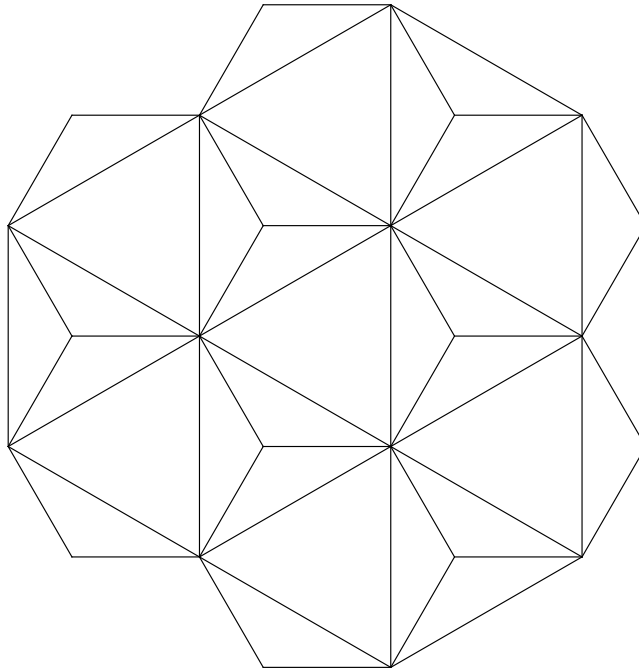
(r)



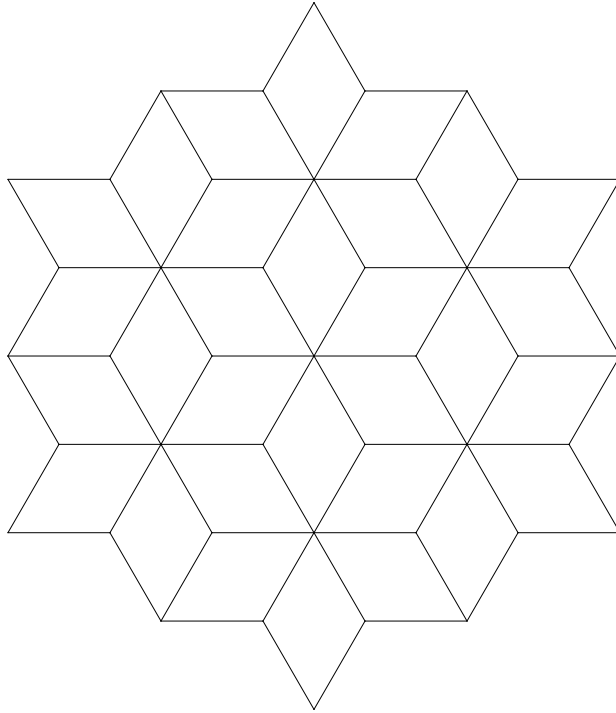
(s)



(t)



(u)



7. Linear isometries of \mathbb{R}^3

7.1. Linear orientations of \mathbb{R}^n . Let's begin by reviewing what we know about orientations in the Euclidean plane. We can think of the standard orientation of \mathbb{R}^2 as being given by the information required to identify the counterclockwise direction for calculating angles. This in turn can be seen as coming from the usual ordering of the canonical basis $\mathcal{E} = e_1, e_2$. This then determines the the sign of the directed angle from the ray $\overrightarrow{0e_1}$ to the ray $\overrightarrow{0x}$ by finding the unique $\theta \in [0, 2\pi)$ with

$$\frac{x}{\|x\|} = \cos \theta e_1 + \sin \theta e_2.$$

In the very same way, an (ordered) orthonormal basis will be seen to provide an orientation for any 2-dimensional inner product space (e.g., any 2-dimensional subspace of \mathbb{R}^n).

Using the directed angle determined by the (standard) orientation, we were then able to detect whether an isometry of \mathbb{R}^2 preserves or reverses orientation by seeing whether it preserves or reverses the signs of directed angles.

Notice that the orientation of the plane does not provide a preferred orientation to lines in the plane: not even for lines through the origin. Each line ℓ in the plane has two orientations, each given by a choice of unit vector parallel to ℓ . In particular, the orientation of a line corresponds to a choice or orthonormal basis for its translation through the origin.

We begin to see a pattern. An orientation of an inner product space V should correspond in some way to a choice of orthonormal basis. And that choice will not automatically orient the subspaces of V .

In fact, an inner product is not necessary for orienting a vector space. Inner products induce lengths and unsigned angles, but are not needed for orientations themselves. Recall the one-to-one correspondence of Corollary 1.2.8 between the bases of an n -dimensional vector space V and the linear isomorphisms from \mathbb{R}^n to V . This correspondence takes the basis \mathcal{B} to the isomorphism $\Phi_{\mathcal{B}} : \mathbb{R}^n \rightarrow V$. The inverse of this correspondence takes the linear isomorphism $f : \mathbb{R}^n \rightarrow V$ to the basis $f(e_1), \dots, f(e_n)$ of V .

Definition 7.1.1. The linear isomorphisms $f, g : \mathbb{R}^n \rightarrow V$ are *orientation equivalent* if the determinant of $g^{-1} \circ f$ is positive. The bases \mathcal{B} and \mathcal{B}' are orientation equivalent if $\Phi_{\mathcal{B}}$ and $\Phi_{\mathcal{B}'}$ are orientation equivalent.

This is easily seen to be an equivalence relation:

Lemma 7.1.2. *Orientation equivalence is an equivalence relation between linear isomorphisms $\mathbb{R}^n \rightarrow V$.*

Proof. Write $f \sim g$ if $\det(g^{-1}f)$ is positive. Then certainly $f \sim f$ as $\det(I) = 1$, so \sim is reflexive. To see it is symmetric, suppose $f \sim g$. We wish to show $g \sim f$, i.e., that $f^{-1}g$ has positive determinant. But

$f^{-1}g = (g^{-1}f)^{-1}$, so

$$\det(f^{-1}g) = \frac{1}{\det(g^{-1}f)}.$$

Since the latter is positive, so is the former.

Finally we show transitivity. Suppose $f \sim g$ and $g \sim h$. We have $h^{-1}f = (h^{-1}g)(g^{-1}f)$. So

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

is positive. \square

Note that $\Phi_{\mathcal{B}'}^{-1}\Phi_{\mathcal{B}}$ is the linear transformation induced by the transition matrix $[I]_{\mathcal{B}'\mathcal{B}}$ and hence the bases \mathcal{B} and \mathcal{B}' are orientation equivalent if and only if this transition matrix has positive determinant.

Definition 7.1.3. A linear orientation of the n -dimensional vector space V consists of a choice of orientation equivalence class of linear isomorphisms $f : \mathbb{R}^n \rightarrow V$ (or equivalently of bases \mathcal{B} of V). A specific linear isomorphism or basis in the given class is said to induce the orientation of V .

A vector space together with a choice of orientation is called an oriented vector space.

The canonical orientation of \mathbb{R}^n is the one given by the canonical basis $\mathcal{E} = e_1, \dots, e_n$. This corresponds to the identity map of \mathbb{R}^n .

Lemma 7.1.4. *An n -dimensional vector space, $n \geq 1$, has exactly two linear orientations. If $\mathcal{B} = v_1, \dots, v_n$ represents one of them, then $\mathcal{B}' = -v_1, v_2, \dots, v_n$ represents the other.*

Proof. For \mathcal{B} and \mathcal{B}' as above, the transition matrix is given by

$$[I]_{\mathcal{B}'\mathcal{B}} = \begin{bmatrix} -1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & 1 \end{bmatrix},$$

which has determinant -1 . So \mathcal{B} and \mathcal{B}' lie in different equivalence classes and it suffices to show there are at most two classes.

Thus, assume neither g nor h is orientation equivalent to f . Then $g^{-1}f$ and $h^{-1}f$ both have negative determinant. Now,

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

is the product of two negative numbers, and hence is positive. \square

We have seen that an orientation of \mathbb{R}^2 does not induce an orientation of a one-dimensional subspace. Nor, of course, does an orientation of a subspace induce an orientation of the whole. What we have is the following. For simplicity, we state the result in \mathbb{R}^n , but any inner product space will do.

Proposition 7.1.5. *Let V be a subspace of \mathbb{R}^n . Then an orientation of V together with an orientation of V^\perp determine an orientation of \mathbb{R}^n as follows: if $\mathcal{B}_1 = v_1, \dots, v_k$ is a basis of V and $\mathcal{B}_2 = w_1, \dots, w_{n-k}$ is a basis of V^\perp , then the basis $\mathcal{B} = v_1, \dots, v_k, w_1, \dots, w_{n-k}$ determines an orientation of \mathbb{R}^n that depends only on the orientation classes of \mathcal{B}_1 and \mathcal{B}_2 .*

If we reverse the orientation on either one of V and V^\perp , the resulting orientation of \mathbb{R}^n is reversed, but if we reverse the orientation on both V and V^\perp , then the orientation of \mathbb{R}^n is unchanged.

Conversely, if we are given orientations on both V and \mathbb{R}^n , there is a unique orientation of V^\perp compatible with these under the above association.

Proof. If $\mathcal{B}'_1 = v_1, \dots, v_k$ and $\mathcal{B}'_2 = w'_1, \dots, w'_{n-k}$ are alternative bases of V and V^\perp , respectively, and if $\mathcal{B}' = v'_1, \dots, v'_k, w'_1, \dots, w'_{n-k}$, then the transition matrix $[I]_{\mathcal{B}'\mathcal{B}}$ is given by

$$[I]_{\mathcal{B}'\mathcal{B}} = \left[\begin{array}{c|c} [I]_{\mathcal{B}'_1\mathcal{B}_1} & 0 \\ \hline 0 & [I]_{\mathcal{B}'_2\mathcal{B}_2} \end{array} \right].$$

So $\det[I]_{\mathcal{B}'\mathcal{B}} = \det[I]_{\mathcal{B}'_1\mathcal{B}_1} \det[I]_{\mathcal{B}'_2\mathcal{B}_2}$, hence reversing the orientation class of exactly one of the two bases will reverse the orientation of the induced basis of \mathbb{R}^n . Reversing both will preserve it.

Since there are exactly two orientations of \mathbb{R}^n , fixing the orientation class of the basis of V and allowing the orientation class on V^\perp to vary, we obtain the two orientations of \mathbb{R}^n via this process, each uniquely associated with an orientation of V^\perp . \square

Definition 7.1.6. Let V be an oriented vector space, with its orientation induced by the linear isomorphism $g : \mathbb{R}^n \rightarrow V$. Let $f : V \rightarrow V$ be a linear isomorphism. We say that f is orientation-preserving if $f \circ g$ induces the same orientation as g , and orientation-reversing otherwise.

The following generalizes the orientation behavior of linear isometries of \mathbb{R}^2 .

Lemma 7.1.7. *Let V be an oriented vector space and let $f : V \rightarrow V$ be a linear isomorphism. Then f is orientation-preserving if and only if $\det f$ is positive. In particular, this is independent of the choice of linear orientation of V .*

Proof. Let $g = \Phi_{\mathcal{B}} : \mathbb{R}^n \rightarrow V$ induce the orientation of V . Then $g^{-1} \circ (f \circ g)$ is the linear transformation induced by the matrix $[f]_{\mathcal{B}}$ and hence has the same determinant as f . \square

7.2. Rotations. We first show every element in $\text{SO}(3)$ is a rotation. But what does that mean? Let's first examine rotating about the north pole e_3 . Rotating about a pole should fix that pole and rotate in the plane orthogonal to that pole. So we are fixing e_3 and will rotate the xy -plane. If we rotate

by the angle θ , the resulting matrix is

$$R_{(e_3, \theta)} = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

This is indeed a special orthogonal matrix (i.e., orthogonal with determinant 1), and induces a linear isometry

$$\rho_{(e_3, \theta)} = T_{R_{(e_3, \theta)}}.$$

Note that we are implicitly orienting the xy -plane by looking down on it. If we look up from the south pole, the displayed transformation would rotate the orthogonal plane counterclockwise in the implicit orientation of the plane given by looking up at it. We should begin by making this precise.

In this section we write elements of \mathbb{R}^3 as column vectors so we can use linear algebra. Let u be a unit vector in \mathbb{R}^3 , i.e., $u \in \mathbb{S}^2$. Then $\{u\}^\perp = \text{span}(u)^\perp$ is a 2-dimensional subspace of \mathbb{R}^3 and can be identified as the nullspace of the row matrix u^T .

Let v be a unit vector in $\{u\}^\perp$. Then u, v is an orthonormal set, so $\text{span}(u, v)$ is 2-dimensional. Thus, $\{u, v\}^\perp$ is 1-dimensional, and contains exactly two unit vectors, say z and $-z$. Now u, v, z and $u, v, -z$ are both orthonormal bases of \mathbb{R}^3 , so $[u|v|z]$ and $[u|v|-z]$ are both orthogonal matrices. Since

$$[u|v|-z] = [u|v|z] \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix},$$

$\det[u|v|-z] = -\det[u|v|z]$. So exactly one of $[u|v|z]$ and $[u|v|-z]$ is special orthogonal. Let $w = \pm z$ such that $\det[u|v|w] = 1$. We have shown:

Lemma 7.2.1. *For any orthonormal set $u, v \in \mathbb{R}^3$ there is a unique vector w such that u, v, w is an orthonormal basis of \mathbb{R}^3 and $\det[u|v|w] = 1$. This choice of w induces the unique linear orientation on $\{u\}^\perp$, coming from the orthonormal basis v, w of $\{u\}^\perp$, such that the basis u, v, w induces the standard orientation of \mathbb{R}^3 .*

Note that the uniqueness of the orientation on $\{u\}^\perp$ was shown in Proposition 7.1.5.

Remark 7.2.2. We shall refer to the orientation on $\{u\}^\perp$ given by the basis v, w as the orientation induced by the pole u . If we replace w by $-w$ we get the opposite orientation on $\{u\}^\perp$. And $-w$ is the unique vector with the property that $-u, v, -w$ is orthonormal and $\det[-u|v|-w] = 1$. So the orientation on $\{u\}^\perp = \{-u\}^\perp$ induced by the pole $-u$ is the opposite of the orientation induced by u . This expresses the difference between “looking down” from u and “looking up” from $-u$.

Definition 7.2.3. Define $\rho_{(u,\theta)}$, the rotation about the pole u by the angle θ , to be the unique linear transformation of \mathbb{R}^3 with

$$(7.2.1) \quad \begin{aligned} \rho_{(u,\theta)}(v) &= (\cos \theta)v + (\sin \theta)w, \\ \rho_{(u,\theta)}(w) &= (-\sin \theta)v + (\cos \theta)w, \\ \rho_{(u,\theta)}(u) &= u, \end{aligned}$$

with v, w as above. In other words, if $\mathcal{B} = v, w, u$, then the matrix of $\rho_{(u,\theta)}$ with respect to \mathcal{B} is precisely the matrix $R_{(e_3,\theta)}$ above. We call $\text{span}(u)$ the axis of $\rho_{(u,\theta)}$.

The proof of the following is immediate from the constructions.

Lemma 7.2.4. *Since \mathcal{B} is orthonormal and $[\rho_{(u,\theta)}]_{\mathcal{B}} = R_{(e_3,\theta)}$, an orthogonal matrix, $\rho_{(u,\theta)}$ is an isometry.*

The matrix of $\rho_{(u,\theta)}$ with respect to the standard basis of \mathbb{R}^3 is

$$R_{(u,\theta)} := [\rho_{(u,\theta)}] = PR_{(e_3,\theta)}P^{-1},$$

where $P = [I]_{\mathcal{E}\mathcal{B}} = [v|w|u]$. Note that

$$\det[v|w|u] = -\det[v|u|w] = \det[u|v|w],$$

so \mathcal{B} induces the standard linear orientation of \mathbb{R}^3 (i.e., $[v|w|u]$ is special orthogonal).

But we also wish to show this transformation is independent of the choice of the unit vector $v \in \{u\}^\perp$.

Lemma 7.2.5. *Let $v, w \in \mathbb{R}^n$ be an orthonormal set and let $V = \text{span}(v, w)$. Then the unit vectors in V are precisely the elements*

$$(7.2.2) \quad x = (\cos \phi)v + (\sin \phi)w \quad \text{for } \phi \in [0, 2\pi).$$

Moreover, given x satisfying (7.2.2), the unique orthonormal basis x, y inducing the same orientation of V as v, w is given by

$$(7.2.3) \quad y = (-\sin \phi)v + (\cos \phi)w = \cos\left(\phi + \frac{\pi}{2}\right)v + \sin\left(\phi + \frac{\pi}{2}\right)w.$$

Finally, if $n = 3$ and $[u|v|w]$ is special orthogonal, so is $[u|x|y]$.

Proof. Let $A = [v|w]$. Then $T_A : \mathbb{R}^2 \rightarrow V$ is a linear isometric isomorphism. The vectors satisfying (7.2.2) are precisely $T_A(\mathbb{S}^1)$, which are the unit vectors of V as T_A preserves the norm. T_A also preserves orthogonality, so the unit vectors orthogonal to x are precisely $\pm y$. The transition matrix from x, y to v, w is the standard rotation matrix R_ϕ , which has determinant 1, so these two bases induce the same orientation. Replacing y by $-y$ reverses the sign of the determinant of the transition matrix, and hence $x, -y$ induces the opposite orientation.

Finally, if $n = 3$ and $[u|v|w]$ is special orthogonal, so is $[v|w|u]$, and the transition matrix from the basis x, y, u to the basis v, w, u is the rotation matrix $R_{(e_3,\phi)}$, which has determinant 1. The result follows. \square

We must show the following.

Proposition 7.2.6. *Let $u \in \mathbb{S}^2$. Then the linear transformation $\rho_{(u,\theta)}$ defined in (7.2.1) is independent of the choice of $v \in \{u\}^\perp$.*

Proof. As stated above, if we use v to define $\rho_{(u,\theta)}$, and if $\mathcal{B} = v, w, u$, then the matrix $[\rho_{(u,\theta)}]_{\mathcal{B}}$ of $\rho_{(u,\theta)}$ with respect to \mathcal{B} is $R_{(e_3,\theta)}$. But if v' and w' are given by (7.2.2) and (7.2.3), respectively, and if $\mathcal{B}' = v', w', u$, then $[I]_{\mathcal{B}\mathcal{B}'} = R_{(e_3,\phi)}$. So

$$\begin{aligned} [\rho_{(u,\theta)}]_{\mathcal{B}'} &= [I]_{\mathcal{B}\mathcal{B}'}^{-1} [\rho_{(u,\theta)}]_{\mathcal{B}} [I]_{\mathcal{B}\mathcal{B}'} \\ &= R_{(e_3,-\phi)} R_{(e_3,\theta)} R_{(e_3,\phi)} \\ &= R_{(e_3,\theta)}. \end{aligned} \quad \square$$

Note that Remark 7.2.2 gives:

Lemma 7.2.7. $\rho_{(u,\theta)} = \rho_{(-u,-\theta)}$.

Proof.

$$\begin{aligned} \rho_{(u,\theta)}(v) &= \cos(-\theta)v + \sin(-\theta) \cdot (-w), \\ \rho_{(u,\theta)}(-w) &= -\sin(-\theta)v + \cos(-\theta) \cdot (-w), \\ \rho_{(u,\theta)}(-u) &= -u. \end{aligned} \quad \square$$

We've seen in the planar case that fixed-point sets are important.

Lemma 7.2.8. *Let $u \in \mathbb{S}^2$ and let $\theta \in (0, 2\pi)$ then the fixed-point set of $\rho_{(u,\theta)}$ is $\text{span}(u)$, the axis of rotation.*

Proof. The fixed-point set of a linear transformation f is the eigenspace of $(f, 1)$. We know that u , and hence $\text{span}(u)$ are fixed by $\rho_{(u,\theta)}$, so it suffices to show the eigenspace is 1-dimensional.

Let $A = [f]_{\mathcal{B}} = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Then the eigenspace of $(f, 1)$ is

the image under $\Phi_{\mathcal{B}}$ of the eigenspace of $(A, 1)$, so it suffices to show this last eigenspace, which is the nullspace of $I_n - A$, is 1-dimensional.

$$I_n - A = \begin{bmatrix} 1 - \cos \theta & \sin \theta & 0 \\ -\sin \theta & 1 - \cos \theta & 0 \\ 0 & 0 & 0 \end{bmatrix}. \text{ Now,}$$

$$\det \begin{bmatrix} 1 - \cos \theta & \sin \theta \\ -\sin \theta & 1 - \cos \theta \end{bmatrix} = 2(1 - \cos \theta),$$

which is nonzero for $\theta \in (0, 2\pi)$, so $\begin{bmatrix} 1 - \cos \theta & \sin \theta \\ -\sin \theta & 1 - \cos \theta \end{bmatrix}$ reduces via

Gauss elimination to the identity matrix. So A reduces to $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$,

which has rank 2, so its nullspace has dimension 1, as desired. \square

We next show that every element of $\text{SO}(3)$ is a rotation. Let V be an n -dimensional vector space and let $f : V \rightarrow V$ be linear. Recall that the characteristic polynomial $\text{ch}_f(x) = \det(xI - f)$ of f is a monic polynomial of degree n , i.e.,

$$\text{ch}_f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$$

with $a_0, \dots, a_{n-1} \in \mathbb{R}$. Its roots are the eigenvalues of f . Recall also that $\text{ch}_f(x) = \text{ch}_{[f]_{\mathcal{B}}}(x)$ for any basis \mathcal{B} of V .

Lemma 7.2.9. *If n is odd then f has at least one real eigenvalue.*

Proof. This is just the standard result that every odd degree polynomial over \mathbb{R} has at least one real root. In this case note that

$$\frac{\text{ch}_f(x)}{x^n} = 1 + \frac{a_{n-1}}{x} + \frac{a_{n-2}}{x^2} + \cdots + \frac{a_0}{x^n},$$

so $\lim_{x \rightarrow \pm\infty} \frac{\text{ch}_f(x)}{x^n} = 1$. But this implies $\lim_{x \rightarrow \pm\infty} \text{ch}_f(x) = \lim_{x \rightarrow \pm\infty} x^n = \pm\infty$ when n is odd. In particular, $\text{ch}_f(x)$ must take on both positive and negative values, and hence must have a root by the intermediate value theorem. \square

Recall from Corollary 4.5.4 that if V is an inner product space with orthonormal basis \mathcal{B} then the linear function $f : V \rightarrow V$ is an isometry if and only if $[f]_{\mathcal{B}}$ is orthogonal.

Proposition 7.2.10. *Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ be a linear isometry of determinant 1. Then 1 is an eigenvalue of f .*

Proof. Write $f = T_A$ for $A \in \text{SO}(3)$. By Lemma 7.2.9, A has at least one eigenvalue, which by Lemma 4.1.23 must be ± 1 . If the eigenvalue is 1 we are done. If not, let u be a unit eigenvector for $A, -1$. By Lemma 4.3.10, $\text{span}(u)$ is an invariant subspace of A . By Lemma 4.3.11, $V = \text{span}(u)^\perp$ is A -invariant as well. Let $\mathcal{B}' = v, w$ be an orthonormal basis of V and let $\mathcal{B} = u, v, w$ be the induced orthonormal basis of \mathbb{R}^3 . Since f is an isometry and V is f -invariant, $f|_V : V \rightarrow V$ is an isometry, hence $[f|_V]_{\mathcal{B}'}$ is an orthogonal matrix. Again, since $\{u\}$ and V are f -invariant, by construction of the basis \mathcal{B} , $[f]_{\mathcal{B}}$ is the block sum

$$[f]_{\mathcal{B}} = \left[\begin{array}{c|c} -1 & 0 \\ \hline 0 & [f|_V]_{\mathcal{B}'} \end{array} \right],$$

and hence $\det f = \det [f]_{\mathcal{B}} = \det[-1] \det [f|_V]_{\mathcal{B}'} = -\det [f|_V]_{\mathcal{B}'}$. So

$$\det [f|_V]_{\mathcal{B}'} = -1.$$

By our analysis of $\text{O}(2)$, this makes $[f|_V]_{\mathcal{B}'}$ the matrix representing a reflection in a line ℓ through the origin. Since ℓ is pointwise fixed by $[f|_V]_{\mathcal{B}'}$, it consists of eigenvectors for the eigenvalue 1. But then 1 is an eigenvalue for $f|_V$ and hence also for f . \square

We obtain:

Theorem 7.2.11. *Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ be a linear isometry of determinant 1. Then f is a rotation of \mathbb{R}^3 about a unit eigenvector u of $(f, 1)$.*

Proof. The proof is very similar to that of Proposition 7.2.10. Start with a unit eigenvector u of $(f, 1)$ and let $V = \{u\}^\perp$. Then $f|_V : V \rightarrow V$ is an isometry, so if $\mathcal{B}' = v, w$ is an orthonormal basis of V , $[f|_V]_{\mathcal{B}'}$ is orthogonal. Now choose \mathcal{B}' so that $\det[v, w, u] = 1$ to give the correct orientation data as above. Now $\mathcal{B} = v, w, u$ is an orthonormal basis of \mathbb{R}^3 , and since $f(u) = u$,

$$[f]_{\mathcal{B}} = \left[\begin{array}{cc|c} [f|_V]_{\mathcal{B}'} & & 0 \\ \hline 0 & & 1 \end{array} \right].$$

We have

$$1 = \det f = \det[f|_V]_{\mathcal{B}'} \cdot 1,$$

so by our analysis of $\text{SO}(2)$, $[f|_V]_{\mathcal{B}'}$ is a 2×2 rotation matrix R_θ for some θ . But then, visibly, $[f]_{\mathcal{B}} = R_{(e_3, \theta)}$ and hence $f = \rho_{(u, \theta)}$. To explicitly solve for θ we solve

$$\cos \theta = \langle f(v), v \rangle, \quad \sin \theta = \langle f(v), w \rangle. \quad \square$$

Remark 7.2.12. The proof of Theorem 7.2.11 can be carried out algorithmically. Starting with $A \in \text{SO}(3)$, the eigenspace of $(T_A, 1)$ is the nullspace of $I - A$, which can be computed by Gauss elimination. Having chosen a unit vector $u \in N(I - A)$, find an orthonormal basis v, w for $N(u^T)$, and replace w by $-w$, if necessary, to get the correct orientations. We can now compute θ by calculating

$$\begin{aligned} T_A(v) &= \langle T_A(v), v \rangle v + \langle T_A(v), w \rangle w \\ T_A(w) &= \langle T_A(w), v \rangle v + \langle T_A(w), w \rangle w \end{aligned}$$

and then using inverse trig functions.

In practice, one must do these calculations in order to compute the composite of two rotations: given unit vectors $u, v \in \mathbb{R}^3$ and angles θ, ϕ , we know that

$$\rho_{(u, \theta)} \circ \rho_{(v, \phi)}$$

is a linear isometry of determinant 1, and hence is equal to $\rho_{(w, \psi)}$ for some w, ψ . The above steps may be used to calculate w and ψ .

7.3. Cross products. Cross products are a useful computational tool for finding orthonormal bases in \mathbb{R}^3 .

Definition 7.3.1. Write $\text{Hom}_{\mathbb{R}}(V, W)$ for the set of linear functions from V to W , and note that $\text{Hom}_{\mathbb{R}}(V, W)$ is a vector space via

$$\begin{aligned} (f + g)(x) &= f(x) + g(x), \\ (cf)(x) &= cf(x). \end{aligned}$$

Lemma 7.3.2. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be linear. Then there is a unique vector $y = \varphi(f)$ such that*

$$f(x) = \langle y, x \rangle$$

for all $x \in \mathbb{R}^n$. This gives a linear isomorphism

$$\varphi : \text{Hom}_{\mathbb{R}}(\mathbb{R}^n, \mathbb{R}) \xrightarrow{\cong} \mathbb{R}^n.$$

Proof. $f = T_A$ for $A = [f] = [f(e_1) \ \dots \ f(e_n)]$, so just take

$$(7.3.1) \quad y = A^T = \begin{bmatrix} f(e_1) \\ \vdots \\ f(e_n) \end{bmatrix},$$

as, for a row matrix A , $Ax = \langle A^T, x \rangle$ for all $x \in \mathbb{R}^n$. φ is linear by (7.3.1) and is an isomorphism as a linear map is uniquely determined by what it does to basis elements. \square

Definition 7.3.3. Let $u, v \in \mathbb{R}^3$ and define $d(u, v) : \mathbb{R}^3 \rightarrow \mathbb{R}$ by

$$d(u, v)(x) = \det[x|u|v].$$

Since the determinant is linear in each column when its other column entries are fixed, $d(u, v) : \mathbb{R}^3 \rightarrow \mathbb{R}$ is linear. Define the cross product $u \times v$ of u and v to be $\varphi(d(u, v))$, i.e., $u \times v$ is the unique vector in \mathbb{R}^3 such that

$$\langle u \times v, x \rangle = \det[x|u|v]$$

for all $x \in \mathbb{R}^3$.

Our main purpose in introducing cross products is (3), below, which has obvious applications to rotations in \mathbb{R}^3 . The other properties then allow valuable calculations.

Proposition 7.3.4. *The cross product gives a bilinear function*

$$\mathbb{R}^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}^3$$

satisfying the following properties:

- (1) $u \times v = -v \times u$.
- (2) $u \times v$ is orthogonal to both u and v .
- (3) If u, v is an orthonormal set, then $[u \times v|u|v]$ is a special orthogonal matrix, i.e., $u \times v, u, v$ is an orthonormal basis of \mathbb{R}^3 inducing the standard orientation.
- (4) $u \times v \neq 0$ if and only if u, v are linearly independent.
- (5) $\langle u \times v, w \rangle = \langle u, v \times w \rangle$.
- (6) $(u \times v) \times w = \langle u, w \rangle v - \langle v, w \rangle u$.
- (7) $\langle u \times v, w \times z \rangle = \langle u, w \rangle \langle v, z \rangle - \langle v, w \rangle \langle u, z \rangle$.

Proof. The cross product is bilinear by the linearity of the map φ in Lemma 7.3.2 together with the fact the determinant is linear in each column when the other column entries are fixed.

(1) follows because the determinant changes sign if you interchange two columns of the matrix.

(2) follows because any matrix with two equal columns has determinant 0.

For (3), let w, u, v be an orthonormal basis with $\det[w|u|v] = 1$. Then

$$\begin{aligned} u \times v &= \langle u \times v, w \rangle w + \langle u \times v, u \rangle u + \langle u \times v, v \rangle v \\ &= \det[w|u|v]w \\ &= w, \end{aligned}$$

as $u \times v$ is orthogonal to u and v .

For (4), $u \times u = 0$ because any matrix with two equal columns has determinant 0. It then follows that $u \times v = 0$ if u, v are linearly dependent. At the other extreme, if u and v are orthogonal and both nonzero, then $\frac{u}{\|u\|} \times \frac{v}{\|v\|}$ has norm 1 by (3), hence $u \times v$ has norm $\|u\|\|v\|$ by bilinearity. Finally, if u, v are linearly independent, apply the first step in the Gram–Schmidt process. By the preceding case,

$$0 \neq u \times \left(v - \frac{\langle u, v \rangle}{\langle u, u \rangle} u \right) = u \times v,$$

where the equality follows from the bilinearity of the cross product and the fact $u \times u = 0$.

(5) simply says $\det[w|u|v] = \det[u|v|w]$.

In (6), both sides are linear in w , keeping u and v fixed. We can now calculate both sides when w is one of the canonical basis vectors, noting the coordinates of $u \times v$ can be calculated from the expansions with respect to the first column of $\det[e_i|u|v]$ for $i = 1, 2, 3$. Since both sides of (6) agree when w is a canonical basis vector, they must agree for arbitrary vectors w .

(7) may now be obtained from (5) and (6). \square

Either a direct calculation or (3) now gives:

Corollary 7.3.5. $e_1 \times e_2 = e_3$, $e_2 \times e_3 = e_1$ and $e_3 \times e_1 = e_2$. The other cross products of canonical basis vectors can now be obtained from Proposition 7.3.4(1).

We now give our application to rotations.

Corollary 7.3.6. Let $u \in \mathbb{S}^2$ and let v be a unit vector in $\{u\}^\perp$ (e.g., $v = \frac{u \times e_i}{\|u \times e_i\|}$ when u, e_i are linearly independent). Then $\mathcal{B} = v, u \times v, u$ is an orthonormal basis inducing the standard orientation of \mathbb{R}^3 , and hence $[\rho(u, \theta)]\mathcal{B} = R_{(e_3, \theta)}$.

7.4. Reflections. Reflections in \mathbb{R}^3 behave very much like reflections in \mathbb{R}^2 . In \mathbb{R}^2 we reflect over a line. In \mathbb{R}^3 we reflect over a plane. What's in common is that we reflect across a set having a fixed normal direction.

A linear reflection reflects across a linear subspace. A plane $V \subset \mathbb{R}^3$ is 2-dimensional, so V^\perp is 1-dimensional. Thus, there are exactly two unit vectors in V^\perp , and we call them unit normals for V .

Definition 7.4.1. Let V be a 2-dimensional linear subspace of \mathbb{R}^3 . Then the reflection across V is given by

$$\sigma_V(x) = x - 2\langle x, N \rangle N$$

for N a unit normal of V .

This is easily seen to be independent of the choice of unit normal, as the only other choice is $-N$. The bilinearity of the inner product shows σ_V to be a linear function.

Proposition 7.4.2. σ_V is a linear isometry of determinant -1 . If v, w is an orthonormal basis of V and if $\mathcal{B} = N, v, w$, then

$$[\sigma_V]_{\mathcal{B}} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus, σ_V is an involution, i.e., $\sigma_V^2 = I$. Moreover, the fixed-point set of σ_V is V .

Proof. If $x \in V$, $\langle x, N \rangle = 0$, so x is fixed by σ_V . Since $\langle N, N \rangle = 1$, $\sigma_V(N) = -N$. Thus, $[\sigma_V]_{\mathcal{B}}$ is the displayed matrix. Since that matrix is orthogonal and \mathcal{B} is orthonormal, σ_V is an isometry. We have

$$[\sigma_V^2]_{\mathcal{B}} = [\sigma_V]_{\mathcal{B}}^2 = I_n,$$

so $\sigma_V^2 = I$. To find the fixed-point set, note that $V = \text{span}(N)^\perp = \{N\}^\perp$, so $V = \{y \in \mathbb{R}^3 : \langle y, N \rangle = 0\}$, so if $y \notin V$, then $\langle y, N \rangle \neq 0$, and hence $\sigma_V(y) \neq y$ by the definition of σ_V . \square

So what happens when we compose two linear reflections? The result will have determinant 1 and hence be a rotation. But which one?

Lemma 7.4.3. Let V and V' be distinct 2-dimensional subspaces of \mathbb{R}^3 with unit normals N and N' , respectively. Then $V \cap V'$ is the 1-dimensional subspace whose unit vectors are $\pm \frac{N \times N'}{\|N \times N'\|}$.

Proof. $V = \{N\}^\perp$ and $V' = \{N'\}^\perp$. So

$$\begin{aligned} V \cap V' &= \{v \in \mathbb{R}^3 : \langle v, N \rangle = \langle v, N' \rangle = 0\} \\ &= \{N, N'\}^\perp. \end{aligned}$$

Since $V \neq V'$, $\text{span}(N) \neq \text{span}(N')$, so N, N' are linearly independent. So $N \times N'$ is nonzero and $\text{span}(N, N')$ is 2-dimensional. But then $\{N, N'\}^\perp$ is 1-dimensional. Since $N \times N'$ is nonzero and lies in $\{N, N'\}^\perp$, it must span $\{N, N'\}^\perp$. \square

We can now calculate the product of two linear reflections. Note how useful the cross product is for keeping track of orientations.

Proposition 7.4.4. *Let V and V' be distinct 2-dimensional subspaces of \mathbb{R}^3 with unit normals N and N' , respectively. Let $u = \frac{N \times N'}{\|N \times N'\|}$. Then*

$$(7.4.1) \quad \sigma_{V'}\sigma_V = \rho_{(u, 2\cos^{-1}\langle N, N' \rangle)}.$$

Indeed, we may interpret $\cos^{-1}\langle N, N' \rangle$ as being the directed angle from V to V' at the pole u . (At $-u$, the direction would be opposite, corresponding to the reversal of the order of the cross product.)

Proof. Let $v = N \times u$ and $v' = N' \times u$, so that v, N and v', N' are both orthonormal bases of $W = \{u\}^\perp$ giving the orientation induced by u .

Since V is fixed by σ_V and V' by $\sigma_{V'}$, $V \cap V' = \text{span}(u)$ is fixed by the rotation $\sigma_{V'}\sigma_V$, so the possible poles for $\sigma_{V'}\sigma_V$ are $\pm u$. We calculate the matrix of $\sigma_{V'}\sigma_V$ with respect to the basis $\mathcal{B} = v, N, u$. Since we know $\sigma_{V'}\sigma_V$ to be a rotation, it suffices to find $[\sigma_{V'}\sigma_V(v)]_{\mathcal{B}}$. Write

$$\begin{aligned} v' &= (\cos \theta)v + (\sin \theta)N, \\ N' &= -(\sin \theta)v + (\cos \theta)N, \end{aligned}$$

so that $\cos \theta = \langle v', v \rangle = \langle N', N \rangle$, $\sin \theta = \langle v', N \rangle = -\langle N', v \rangle$. (That $\langle v', v \rangle = \langle N', N \rangle$ also follows from Proposition 7.3.4(7).) Since σ_V fixes v , we have

$$\begin{aligned} \sigma_{V'}\sigma_V(v) &= \sigma_{V'}(v) \\ &= v - 2\langle v, N' \rangle N' \\ &= v + 2(\sin \theta)N' \\ &= (1 - 2\sin^2 \theta)v + 2(\sin \theta \cos \theta)N \\ &= \cos(2\theta)v + \sin(2\theta)N, \end{aligned}$$

so it suffices to show $\theta = \cos^{-1}\langle N, N' \rangle$. Since $\cos \theta = \langle N, N' \rangle$ it suffices to show $\sin \theta > 0$, i.e., $\langle N, v' \rangle > 0$. We have

$$\begin{aligned} \langle N, v' \rangle &= \langle N, N' \times u \rangle \\ &= -\langle u \times N', N \rangle \\ &= -\langle u, N' \times N \rangle \\ &= \langle u, N \times N' \rangle \\ &= \|N \times N'\| \end{aligned}$$

by the definition of u . □

It is easy to reverse-engineer this process, and we obtain the following.

Corollary 7.4.5. *Every linear rotation of \mathbb{R}^3 is the product of two linear reflections.*

The following is important in spherical geometry.

Proposition 7.4.6. *Let $u \neq v \in \mathbb{S}^2$. Then there is a unique linear reflection of \mathbb{R}^3 interchanging u and v . Specifically, if $N = \frac{v-u}{\|v-u\|}$ and if V is the 2-dimensional subspace with unit normal N , then $\sigma_V(u) = v$.*

Proof. If V is a two-dimensional subspace with unit normal N and if $\sigma_V(u) = v$, then

$$v = u - 2\langle u, N \rangle N,$$

and hence $N = \frac{v-u}{-2\langle u, N \rangle}$. The denominator is nonzero, as $u \neq v$. Since N is a unit vector,

$$1 = \|N\| = \frac{\|v-u\|}{|-2\langle u, N \rangle|},$$

so $-2\langle u, N \rangle = \pm\|v-u\|$. Thus, $N = \pm \frac{v-u}{\|v-u\|}$. This gives uniqueness.

It suffices to show that if $N = \frac{v-u}{\|v-u\|}$ and if $V = \{N\}^\perp$, then $\sigma_V(u) = v$. Now

$$\begin{aligned} \sigma_V(u) &= u - 2\langle u, N \rangle N \\ &= u - 2 \left(\frac{\langle u, v \rangle - \langle u, u \rangle}{\|v-u\|} \right) \frac{v-u}{\|v-u\|} \\ &= u - 2 \frac{\langle u, v \rangle - 1}{\langle v-u, v-u \rangle} (v-u) \\ &= u - 2 \frac{\langle u, v \rangle - 1}{\langle u, u \rangle + \langle v, v \rangle - 2\langle u, v \rangle} (v-u) \\ &= u - 2 \frac{\langle u, v \rangle - 1}{2 - 2\langle u, v \rangle} (v-u) = u + (v-u) = v. \end{aligned}$$

Here, we have used twice that $\langle u, u \rangle = \langle v, v \rangle = 1$, as $u, v \in \mathbb{S}^2$. \square

7.5. Rotation-reflections.

Lemma 7.5.1. *Let V be a 2-dimensional linear subspace of \mathbb{R}^3 with unit normal N . Then σ_V and $\rho_{(N, \theta)}$ commute for all $\theta \in \mathbb{R}$.*

Proof. Let $v \in V$ and let $\mathcal{B} = v, N \times v, N$. Then $[\sigma_V]_{\mathcal{B}}$ and $[\rho_{(N, \theta)}]_{\mathcal{B}}$ commute. \square

Definition 7.5.2. A rotation-reflection of \mathbb{R}^3 is a composite

$$(7.5.1) \quad \sigma_V \rho_{(N, \theta)} = \rho_{(N, \theta)} \sigma_V$$

for V a 2-dimensional linear subspace with unit normal N , and $\theta \in (0, 2\pi)$.

Note that both V and $\text{span}(N)$ are invariant subspaces for the rotation-reflection $\sigma_V \rho_{(N, \theta)}$.

We can now complete our classification of the linear isometries of \mathbb{R}^3 .

Proposition 7.5.3. *Every orientation-reversing linear isometry of \mathbb{R}^3 is either a reflection or a rotation-reflection.*

Proof. Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ be an orientation-reversing linear isometry. Then f has at least one real eigenvalue. Since both reflections and rotation-reflections have an eigenvector for -1 we shall begin by supposing that -1 is an eigenvalue of f . Let N be a unit eigenvector for $(f, -1)$. Then $\text{span}(N)$ is an f -invariant subspace and hence so is $V = \text{span}(N)^\perp$. Let $v \in V$ and let $w = N \times v$. Let $\mathcal{B} = v, w, N$. Since $f|_V : V \rightarrow V$ is a linear isometry, there exists $\theta \in \mathbb{R}$ with $f(v) = (\cos \theta)v + (\sin \theta)w$ and $f(w) = \pm w'$ with $w' = (-\sin \theta)v + (\cos \theta)w$.

If $f(w) = -w'$, then

$$[f]_{\mathcal{B}} = \begin{bmatrix} \cos \theta & \sin \theta & 0 \\ \sin \theta & -\cos \theta & 0 \\ 0 & 0 & -1 \end{bmatrix},$$

which has determinant 1. So $f(w) = w'$, which gives

$$[f]_{\mathcal{B}} = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & -1 \end{bmatrix},$$

hence $f = \sigma_V \rho_{(N, \theta)}$. When θ is a multiple of 2π this is just the reflection σ_V , and otherwise it is a rotation-reflection.

Suppose, then, that 1 is an eigenvalue of f and let u be a unit eigenvector for $(f, 1)$. Again, $\text{span}(u)$ is an invariant subspace as is $V = \text{span}(u)^\perp$. Let $w = u \times v$ and $\mathcal{B} = v, w, u$. Again $f(v) = (\cos \theta)v + (\sin \theta)w$ and $f(w) = \pm w'$ as above. If $f(w) = w'$, then

$$[f]_{\mathcal{B}} = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

which has determinant 1. Thus, $f(w) = -w'$ so if $\mathcal{B}' = v, w$,

$$[f|_V]_{\mathcal{B}'} = \begin{bmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{bmatrix},$$

a 2×2 reflection matrix. Every 2×2 reflection matrix has -1 as an eigenvalue, hence so does $f|_V$. But any eigenvector for $f|_V$ is an eigenvector for f with the same eigenvalue, so we are back in case one and we're done. \square

Remark 7.5.4. The transformation $\alpha = \sigma_V \rho_{(N, \theta)}$ of (7.5.1) does not in general determine the subspace V . Indeed, for any 2-dimensional subspace V with unit normal N , the composite $\sigma_V \rho_{(N, \pi)}$ is the isometry induced by the orthogonal matrix $-I_3$.

However, since a rotation-reflection has determinant -1 , the eigenspace of $(\alpha, -1)$ must have odd dimension. So if $\alpha \neq T_{-I_3}$, the eigenspace of $(\alpha, -1)$ must be 1-dimensional, and hence must equal $\text{span}(N)$. This, in turn determines V .

7.6. Symmetries of the Platonic solids. The Platonic solids are the regular polyhedra:¹² the tetrahedron, the cube, the octahedron, the dodecahedron and the icosahedron.

7.6.1. The cube and the regular tetrahedron. We first give some basics on the cube and its symmetries. Proposition 6.2.17 computes the symmetries of the n -cube: taking $[-1, 1]^n$ as the model for the n -cube, the centroid is 0, so the symmetries are all linear. The symmetries are induced by the signed permutation matrices, which form the group $O(n, \mathbb{Z}) \subset O(n)$. The elements of $O(n, \mathbb{Z})$ are the matrices $[\epsilon_1 e_{\sigma(1)} | \dots | \epsilon_n e_{\sigma(n)}]$, where $\epsilon_i \in \{\pm 1\}$ for $i = 1, \dots, n$, and $\sigma \in \Sigma_n$, the group of permutations of $\{1, \dots, n\}$. Note that $|O(n, \mathbb{Z})| = 2^n n!$, as there are 2^n choices of signs and $n!$ permutations.

By Corollary 8.2.2 below, the topological notion of the orientation-preserving property for a linear isometry of \mathbb{R}^n coincides with the linear one, so the group of orientation-preserving isometries is given by

$$(7.6.1) \quad \mathcal{O}([-1, 1]^n) \cong O(n, \mathbb{Z}) \cap SO(n) = \{A \in O(n, \mathbb{Z}) : \det A = 1\}.$$

We shall denote it by $SO(n, \mathbb{Z})$. Since $O(n, \mathbb{Z})$ contains orientation-reversing isometries, $SO(n, \mathbb{Z})$ has index 2 in $O(n, \mathbb{Z})$. and hence has order $2^{n-1} n!$.

In particular, the symmetry group of the standard 3-dimensional cube $\mathbf{C} = [-1, 1]^3$ is $O(3, \mathbb{Z})$ and has 48 elements, while $\mathcal{O}(\mathbf{C})$ has order 24. We will show that $\mathcal{O}(\mathbf{C})$ is isomorphic to Σ_4 .

For now, let us review the vertices, edges and faces of \mathbf{C} . (For the remainder of Section 7.6, the word “face” means two-dimensional face.) The vertex set of \mathbf{C} is

$$(7.6.2) \quad S = \left\{ \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix} : \epsilon_i \in \{\pm 1\} \text{ for } i = 1, \dots, 3 \right\}.$$

The faces are

$$(7.6.3) \quad \partial_i^\epsilon(\mathbf{C}) = \left\{ \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \in \mathbf{C} : a_i = \epsilon \right\} \quad \text{for } i = 1, \dots, 3 \text{ and } \epsilon = \pm 1.$$

In particular, the vertex

$$(7.6.4) \quad v = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix} = \partial_1^{\epsilon_1}(\mathbf{C}) \cap \partial_2^{\epsilon_2}(\mathbf{C}) \cap \partial_3^{\epsilon_3}(\mathbf{C}),$$

and lies in no other faces of \mathbf{C} .

Each edge is the intersection of two faces, and is therefore given by specifying two coordinates by particular elements of $\{\pm 1\}$. In particular, two vertices share an edge if and only if they have two coordinates in common.

Let $v = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}$ and $w = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix}$ be vertices of \mathbf{C} . Then

$$d(v, w) = \sqrt{(\epsilon_1 - \delta_1)^2 + (\epsilon_2 - \delta_2)^2 + (\epsilon_3 - \delta_3)^2}$$

¹²Here, we take “polyhedron” to mean a 3-dimensional polytope.

But

$$(\epsilon_i - \delta_i)^2 = \begin{cases} 0 & \text{if } \epsilon_i = \delta_i \\ 4 & \text{otherwise.} \end{cases}$$

Thus, $d(v, w) = 2\sqrt{k}$, where k is the number of coordinates in which v and w differ. In particular, if v and w share an edge, the distance is 2. If they agree on exactly one coordinate (and hence are diagonally apart on a face), the distance is $2\sqrt{2}$. If they agree in no coordinate (and hence $v = -w$), the distance is $2\sqrt{3}$.

Since each vertex is contained in exactly three faces, there are exactly three vertices of distance $2\sqrt{2}$ from it.

7.6.2. The regular tetrahedron. We shall show that the regular tetrahedron is the convex hull of some of the coordinates of the cube $\mathbf{C} = [-1, 1]^3$. We'll be able to use this to say more about $\mathcal{S}(\mathbf{C})$ and $\mathcal{O}(\mathbf{C})$.

Recall that the set of vertices of \mathbf{C} is denoted by S . Consider the set

$$(7.6.5) \quad T = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix} \right\} \subset S.$$

Then each pair of distinct vertices in T agrees on exactly one coordinate, so the distance between any pair of distinct elements of T is $2\sqrt{2}$.

So any three distinct elements of T form the vertex set for an equilateral triangle in \mathbb{R}^3 . The triangles are all congruent, and assemble to form a tetrahedron, $\mathbf{T} = \text{Conv}(T)$. Since the faces of \mathbf{T} are congruent to one another, the tetrahedron is regular.

Since each vertex of \mathbf{C} has exactly three vertices of distance $2\sqrt{2}$ from it, T contains every vertex of distance $2\sqrt{2}$ from any of its vertices.

Now let

$$T' = S \setminus T$$

and let $\mathbf{T}' = \text{Conv}(T')$. Then the same argument given above shows that \mathbf{T}' is a regular tetrahedron and that for $w \in T'$, $T' \setminus \{w\}$ is the set of all vertices of \mathbf{C} of distance $2\sqrt{2}$ from w . Moreover, $T' = \{-v : v \in T\}$. For $v \in T$ and $w \in T'$, either $d(v, w) = 2$ or $d(v, w) = 2\sqrt{3}$.

Proposition 7.6.1. $\mathcal{S}(\mathbf{T})$ is an index two subgroup of $\mathcal{S}(\mathbf{C})$ and hence has order 24.

Proof. First note that the sum of the vectors in T is 0, so the centroid of \mathbf{T} is 0. Thus, $\mathcal{S}(\mathbf{T})$ is a subgroup of the group of linear isometries of \mathbb{R}^3 . So if $\alpha \in \mathcal{S}(\mathbf{T})$ and $v \in T$, $\alpha(-v) = -\alpha(v) \in T'$. So α permutes the vertices of \mathbf{C} , and hence $\alpha \in \mathcal{S}(\mathbf{C})$.

Now let $\alpha \in \mathcal{S}(\mathbf{C})$. Then α may not permute the elements of T , as $\alpha\left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\right)$ can be any vertex of S by our calculation of $\mathcal{S}(\mathbf{C})$.

Let $w = \alpha\left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\right)$, and suppose w is not in T . Then α must carry every other vertex in T to a vertex of distance $2\sqrt{2}$ from w , which must lie in T' .

So if $w \notin T$, then $\alpha(T) = T'$, and hence $\alpha(\mathbf{T}) = \mathbf{T}'$. Since α permutes S , we must also have $\alpha(T') = T$.

We see that $\mathcal{S}(\mathbf{C})$ permutes the two-element set $\{T, T'\}$. We obtain a surjective homomorphism

$$f : \mathcal{S}(\mathbf{C}) \rightarrow \Sigma(\{T, T'\}) \cong \Sigma_2,$$

with $f(\alpha)(T) = \alpha(T)$ and $f(\alpha)(T') = \alpha(T')$. Since Σ_2 has $2! = 2$ elements, $\ker f$ has index 2 in $\mathcal{S}(\mathbf{C})$. But $\ker f$ is the set of $\alpha \in \mathcal{S}(\mathbf{C})$ that permute the elements of T , and hence lie in $\mathcal{S}(\mathbf{T})$. \square

Corollary 7.6.2. *The symmetry group $\mathcal{S}(\mathbf{T})$ of the regular tetrahedron \mathbf{T} is isomorphic to the full permutation group of its vertex set, i.e., to Σ_4 .*

Proof. Corollary 6.2.5 gives a homomorphism $\rho : \mathcal{S}(\mathbf{T}) \rightarrow \Sigma(T)$, which is injective as $\text{Aff}(T) = \mathbb{R}^3$ (it is easy to show that the vertices in T are affinely independent), and hence an element of $\mathcal{S}(\mathbf{T})$ is determined by its effect on T .

But both $\mathcal{S}(\mathbf{T})$ and Σ_4 have 24 elements, so ρ is an isomorphism. \square

We can now use this to derive information about $\mathcal{O}(\mathbf{C})$.

7.6.3. Calculation of $\mathcal{O}(\mathbf{C})$. Consider the pairs of points $\{\pm v\}$ with $v \in T$. Then there are four such pairs. Write X for the four-element set they comprise:

$$(7.6.6) \quad X = \{\{\pm v\} : v \in T\}.$$

Then $\mathcal{S}(\mathbf{C})$ permutes the elements of X , i.e., there is a homomorphism

$$(7.6.7) \quad \begin{aligned} \pi : \mathcal{S}(\mathbf{C}) &\rightarrow \Sigma(X) \\ \pi(\alpha)(\{\pm v\}) &= \{\pm \alpha(v)\}, \end{aligned}$$

for $\alpha \in \mathcal{S}(\mathbf{C})$ and $v \in T$. By Corollary 7.6.2, π restricts to an isomorphism

$$(7.6.8) \quad \pi|_{\mathcal{S}(\mathbf{T})} : \mathcal{S}(\mathbf{T}) \xrightarrow{\cong} \Sigma(X).$$

We use this to obtain the following:

Corollary 7.6.3. *The restriction of π to $\mathcal{O}(\mathbf{C})$ is an isomorphism:*

$$(7.6.9) \quad \pi|_{\mathcal{O}(\mathbf{C})} : \mathcal{O}(\mathbf{C}) \xrightarrow{\cong} \Sigma(X).$$

Thus, $\mathcal{O}(\mathbf{C})$ is isomorphic to Σ_4 .

Proof. Since $\mathcal{O}(\mathbf{C})$ and $\Sigma(X)$ both have order 24, it suffices to show (7.6.3) is onto. Since π restricts to an isomorphism from $\mathcal{S}(\mathbf{T})$ onto $\Sigma(X)$, it suffices to show that for $\alpha \in \mathcal{S}(\mathbf{T})$, there exists $\beta \in \mathcal{O}(\mathbf{C})$ with $\pi(\alpha) = \pi(\beta)$.

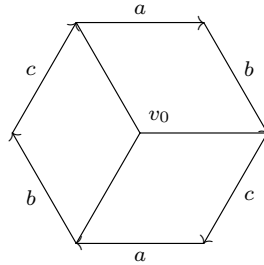
If α is orientation-preserving, take $\beta = \alpha$. Otherwise, take $\beta = -\alpha$, the composite of α with multiplication by -1 . Since $-I_3$ is orientation-reversing, $\beta \in \mathcal{O}(\mathbf{C})$. And multiplication by -1 acts as the identity on X . \square

Remark 7.6.4. The boundary of the cube induces a tiling of the 2-sphere \mathbb{S}^2 by radial projection (i.e., the function that takes a nonzero vector v to $\frac{v}{\|v\|}$). The symmetry group of this tiling coincides precisely with $\mathcal{S}(\mathbf{C})$.

The projective space \mathbb{RP}^2 is the quotient of \mathbb{S}^2 obtained by identifying each $x \in \mathbb{S}^2$ with its antipode $-x$. Thus, we obtain \mathbb{RP}^2 from \mathbb{S}^2 by identifying any two points that differ by multiplication by $-I_3$. Since the action of $-I_3$ carries faces to faces in \mathbf{C} , the tiling of \mathbb{S}^2 induces a tiling of \mathbb{RP}^2 with four vertices, six edges and three faces.

The proof of Corollary 7.6.3 amounts to studying the induced action of $\mathcal{S}(\mathbf{C})$ on \mathbb{RP}^2 , where $-I_3$ acts as the identity. The set X projects to the vertex set for this tiling of \mathbb{RP}^2 , which is invariant under the action of $\mathcal{S}(\mathbf{C})$. The argument studies the action of $\mathcal{S}(\mathbf{C})$ on that four-element set. In fact, both $\mathcal{O}(\mathbf{C})$ and $\mathcal{S}(\mathbf{T})$ map isomorphically onto the symmetry group of this tiling of \mathbb{RP}^2 , despite being different subgroups of $\mathcal{S}(\mathbf{C})$. (The isometry group of \mathbb{RP}^2 is the quotient group $\mathrm{O}(3)/(\pm I_3)$.)

Under this model, we can view \mathbb{RP}^2 as obtained as follows. Let Y be the union of the three faces of \mathbf{C} containing the vertex $v_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ and identify the six edges of Y not meeting v_0 as indicated in the following diagram: specifically, we identify these edges in opposite pairs according to the orientations given in the following diagram:



So a , b and c correspond to the three of the edges of \mathbb{RP}^2 that don't meet v_0 ; the other three edges of \mathbb{RP}^2 come from the edges of the diagram that emanate from v_0 . The identifications of the edges in Y correspond to multiplication by -1 in \mathbf{C} . These identifications also reduce the six vertices of $Y \setminus v_0$ to three vertices of \mathbb{RP}^2 , indicated by the endpoints of the edges emanating from v_0 . Note that each pair of vertices in \mathbb{RP}^2 is connected by a unique edge.

Using this model, you can study the induced action of $\mathcal{O}(\mathbf{C})$ on \mathbb{RP}^2 . Corollary 7.6.3 shows this action is effective, meaning that no nonidentity element of $\mathcal{O}(\mathbf{C})$ acts as the identity on \mathbb{RP}^2 .

7.6.4. The dodecahedron. We shall make use of the golden mean

$$\Phi = \frac{1 + \sqrt{5}}{2}$$

studied in Section 6.3. As shown in (6.3.2) there, the multiplicative inverse $\phi = \frac{1}{\Phi}$ satisfies

$$\phi = \Phi - 1 = \frac{-1 + \sqrt{5}}{2}.$$

Thus,

$$(7.6.10) \quad \Phi + \phi = \frac{1 + \sqrt{5}}{2} + \frac{-1 + \sqrt{5}}{2} = \sqrt{5} < 3.$$

Moreover, Lemma 6.3.3 gives the numerical estimates

$$\Phi \in (1.5, 2), \quad \phi \in (.5, .75).$$

We shall also make use of the following:

$$(7.6.11) \quad \Phi^2 - \phi^2 = \Phi^2 - (\Phi - 1)^2 = 2\Phi - 1 = \sqrt{5}.$$

The last equality follows from $\Phi = \frac{1+\sqrt{5}}{2}$. Also,

$$(7.6.12) \quad \Phi^2 + \phi^2 = \Phi^2 + (\Phi - 1)^2 = 2\Phi^2 - 2\Phi + 1 = 3.$$

Definition 7.6.5. The standard regular dodecahedron \mathbf{D} is defined to be $\text{Conv}(V)$, where

$$(7.6.13) \quad V = \left\{ \begin{bmatrix} 0 \\ \pm\phi \\ \pm\Phi \end{bmatrix}, \begin{bmatrix} \pm\Phi \\ 0 \\ \pm\phi \end{bmatrix}, \begin{bmatrix} \pm\phi \\ \pm\Phi \\ 0 \end{bmatrix}, \begin{bmatrix} \pm 1 \\ \pm 1 \\ \pm 1 \end{bmatrix} \right\},$$

where the signs ± 1 in the various coordinates are independent of each other. Thus, there are 20 elements in V , including the eight elements

$$(7.6.14) \quad S = \left\{ \begin{bmatrix} \pm 1 \\ \pm 1 \\ \pm 1 \end{bmatrix} \right\}.$$

Recall that S is the set of vertices of the standard balanced cube

$$(7.6.15) \quad \mathbf{C} = [-1, 1]^3.$$

Thus, \mathbf{C} is a convex subset of the dodecahedron \mathbf{D} . We single out two elements of V for special consideration:

$$(7.6.16) \quad v_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad v_1 = \begin{bmatrix} 0 \\ \phi \\ \Phi \end{bmatrix}.$$

Recall that $\mathcal{S}(\mathbf{C})$ consists of the linear isometries induced by the signed permutation matrices $O(n, \mathbb{Z})$. In particular, we have very good control over these isometries. The subgroup $\mathcal{S}(\mathbf{C}) \cap \mathcal{S}(\mathbf{D})$ will help us get good control on the geometry of \mathbf{D} and on its full group of isometries.

We do not yet know that each $v \in V$ is a vertex of \mathbf{D} . If we did, then Proposition 6.2.5 would tell us that $\mathcal{S}(V) = \mathcal{S}(\mathbf{D})$. But at this point we only have the result of Lemma 6.2.1 that $\mathcal{S}(V) \subset \mathcal{S}(\mathbf{D})$. We shall use this to show that every $v \in V$ is a vertex of \mathbf{D} .

Lemma 7.6.6. *Let*

$$H = (\mathcal{S}(\mathbf{C}) \cap \mathcal{S}(V)) \subset (\mathcal{S}(\mathbf{C}) \cap \mathcal{S}(\mathbf{D})).$$

Then H consists of the isometries induced by the following specific signed permutation matrices:

$$\left\{ \begin{bmatrix} \pm 1 & 0 & 0 \\ 0 & \pm 1 & 0 \\ 0 & 0 & \pm 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & \pm 1 \\ \pm 1 & 0 & 0 \\ 0 & \pm 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & \pm 1 & 0 \\ 0 & 0 & \pm 1 \\ \pm 1 & 0 & 0 \end{bmatrix} \right\}.$$

Here, as above, the signs in a given matrix are independent of each other, so the order of H is 24.

The action of H on the convex generating set V has two orbits: $S = Hv_0$ and $V \setminus S = Hv_1$.

Proof. The permutations in these matrices are either the identity or what is called the cyclic permutations $\sigma = (1\ 2\ 3)$ and $\sigma^{-1} = (1\ 3\ 2)$:

$$\begin{array}{ccc} \sigma(1) = 2 & \sigma(2) = 3 & \sigma(3) = 1 \\ \sigma^{-1}(1) = 3 & \sigma^{-1}(3) = 2 & \sigma^{-1}(2) = 1. \end{array}$$

These are precisely the permutations that preserve the “cyclic” ordering of 0 , ϕ and Φ encoded into the elements in V . So these are precisely the matrices for elements of $\mathcal{S}(\mathbf{C})$ that preserve V .

Regarding the orbits, S is H -invariant and every element of S lies in the orbit of v_0 . Similarly, $V \setminus S$ is H -invariant and each of its elements is in the orbit of v_1 . \square

The following calculations will be useful. They also provide evidence suggesting that $\mathcal{S}(\mathbf{D})$ acts transitively on V . (Recall that a G -action on a set V is transitive if V consists of a single G -orbit.)

Lemma 7.6.7. *The inner products of elements of V satisfy the following.*

(7.6.17)

$$\langle v, v_0 \rangle = \begin{cases} 3 & \text{for } v = v_0 \\ \sqrt{5} & \text{for } v = v_1, \begin{bmatrix} \Phi \\ 0 \\ \phi \end{bmatrix}, \begin{bmatrix} \phi \\ \Phi \\ 0 \end{bmatrix} \\ 1 & \text{for } v = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ -\phi \\ \Phi \end{bmatrix}, \begin{bmatrix} -\phi \\ \Phi \\ 0 \end{bmatrix}, \begin{bmatrix} \Phi \\ 0 \\ -\phi \end{bmatrix} \\ -1 & \text{for } v = \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} \phi \\ \Phi \\ -\phi \end{bmatrix}, \begin{bmatrix} -\phi \\ \Phi \\ 0 \end{bmatrix}, \begin{bmatrix} \Phi \\ 0 \\ \phi \end{bmatrix} \\ -\sqrt{5} & \text{for } v = \begin{bmatrix} 0 \\ -\phi \\ -\Phi \end{bmatrix}, \begin{bmatrix} -\Phi \\ 0 \\ -\phi \end{bmatrix}, \begin{bmatrix} -\phi \\ -\Phi \\ 0 \end{bmatrix} \\ -3 & \text{for } v = -v_0. \end{cases}$$

(7.6.18)

$$\langle v, v_1 \rangle = \begin{cases} 3 & \text{for } v = v_1 \\ \sqrt{5} & \text{for } v = v_0, \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ -\phi \\ \Phi \end{bmatrix} \\ 1 & \text{for } v = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} \Phi \\ \phi \\ 0 \end{bmatrix}, \begin{bmatrix} -\Phi \\ \phi \\ 0 \end{bmatrix}, \begin{bmatrix} \phi \\ \Phi \\ 0 \end{bmatrix}, \begin{bmatrix} -\phi \\ -\Phi \\ 0 \end{bmatrix} \\ -1 & \text{for } v = \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -\Phi \\ 0 \\ -\phi \end{bmatrix}, \begin{bmatrix} \Phi \\ 0 \\ -\phi \end{bmatrix}, \begin{bmatrix} -\phi \\ -\Phi \\ 0 \end{bmatrix}, \begin{bmatrix} \phi \\ -\Phi \\ 0 \end{bmatrix} \\ -\sqrt{5} & \text{for } v = -v_0, \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ \phi \\ -\Phi \end{bmatrix} \\ -3 & \text{for } v = -v_1. \end{cases}$$

Proof. These are straightforward from (7.6.10), (7.6.11) and (7.6.12), along with the calculations that $\Phi - \phi = \Phi - (\Phi - 1) = 1$ and $\Phi\phi = \Phi^2 - \Phi = 1$. \square

Corollary 7.6.8. *The elements of V are all vertices of \mathbf{D} . Thus, $\mathcal{S}(V) = \mathcal{S}(\mathbf{D})$, and hence*

$$(7.6.19) \quad H = \mathcal{S}(\mathbf{C}) \cap \mathcal{S}(\mathbf{D}).$$

Proof. Since $H \subset \mathcal{S}(\mathbf{D})$ and since symmetries of a polytope carry vertices to vertices, it suffices to show v_0 and v_1 are vertices.

We use Proposition 2.9.47. Let $f_0, f_1 : \mathbb{R}^3 \rightarrow \mathbb{R}$ be the linear maps given by $f_i(x) = \langle x, v_i \rangle$ for $i = 1, 2$. By Lemma 7.6.7, $f_i(\mathbf{D}) = [-3, 3]$ and $f_i^{-1}(3) = \{v_i\}$. So v_i is a 0-dimensional face of \mathbf{D} for $i = 1, 2$. \square

For each $v \in V$, the negative $-v$ is also in V , so the vectors in V add up to 0. We obtain:

Corollary 7.6.9. *The centroid of \mathbf{D} is the origin, so $\mathcal{S}(\mathbf{D})$ is a subgroup of the group of linear isometries of \mathbb{R}^3 .*

Note that the vertices of \mathbf{D} all have norm $\sqrt{3}$. Moreover, if v and w are vertices of \mathbf{D} , we have

$$(7.6.20) \quad \|v - w\|^2 = \langle v, v \rangle + \langle w, w \rangle - 2\langle v, w \rangle,$$

so Lemma 7.6.7 gives a calculation of the distances between vertices of \mathbf{D} . The larger the inner product $\langle v, w \rangle$, the smaller the distance between v and w . Since there are at most two orbits of V under the action of $\mathcal{S}(\mathbf{D})$, we obtain the following.

Corollary 7.6.10. *The shortest distance between distinct vertices of \mathbf{D} is*

$$(7.6.21) \quad \sqrt{3 + 3 - 2\sqrt{5}} = 2\phi.$$

Each vertex v has distance 2ϕ from exactly three other vertices. Moreover, there are six vertices of distance 2 from v , six of distance $2\sqrt{2}$, three of distance 2Φ and one of distance $2\sqrt{3}$.

Proof. It suffices to verify the equality in (7.6.21) and to show that

$$\sqrt{6 + 2\sqrt{5}} = 2\Phi.$$

For the former, we have

$$6 - 2\sqrt{5} = 4 \left(\frac{3 - \sqrt{5}}{2} \right) = 4(2 - \Phi).$$

But $\phi^2 = 2 - \Phi$. For the latter, we add $\sqrt{5}$ rather than subtract it and note that $\frac{3+\sqrt{5}}{2} = \Phi + 1 = \Phi^2$. \square

We now determine the faces of \mathbf{D} . We start by constructing one of the three faces containing v_0 . Let $v_1 = \begin{bmatrix} 0 \\ \phi \end{bmatrix}$, $v_2 = \begin{bmatrix} 0 \\ -\phi \end{bmatrix}$, $v_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $v_4 = \begin{bmatrix} \Phi \\ 0 \\ \phi \end{bmatrix}$, and set

$$U = \{v_0, v_1, v_2, v_3, v_4\}.$$

We shall show that $F = \text{Conv}(U)$ is a face of \mathbf{D} .

Let $N = v_0 + \cdots + v_4 = \begin{bmatrix} \Phi+2 \\ 0 \\ 3\Phi+1 \end{bmatrix}$. Then the centroid of F will be $\frac{1}{5}N$ when we show F is in fact a face. We use the linear function $f(x) = \langle x, N \rangle$ to show this. As the reader may calculate,

$$(7.6.22) \quad f(v_i) = 4\Phi + 3 \quad \text{for } v_i \in U.$$

Now let $x_0 = \begin{bmatrix} \Phi \\ 0 \\ -\phi \end{bmatrix}$, $x_1 = \begin{bmatrix} \phi \\ \phi \\ 0 \end{bmatrix}$, $x_2 = \begin{bmatrix} \phi \\ -\phi \\ 0 \end{bmatrix}$, $x_3 = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}$ and $x_4 = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}$. Then an easy calculation gives

$$(7.6.23) \quad f(x_i) = 2\Phi - 1 = \sqrt{5} < 4\Phi + 3 \quad \text{for } i = 0, \dots, 4.$$

The other vertices are the negatives of these, so their values under f are negatives of these. So $f(\mathbf{D}) = [-4\Phi - 3, 4\Phi + 3]$ and Proposition 2.9.47 gives the following.

Proposition 7.6.11. $F = \mathbf{D} \cap f^{-1}(4\Phi + 3)$ is a face of \mathbf{D} .

The following is useful in understanding what is going on.

Lemma 7.6.12. *We have*

$$(7.6.24) \quad N = (\Phi + 2) \begin{bmatrix} 1 \\ 0 \\ \Phi \end{bmatrix}.$$

Proof. $(\Phi + 2)\Phi = \Phi^2 + 2\Phi = \Phi + 1 + 2\Phi = 3\Phi + 1$. \square

Write

$$(7.6.25) \quad \rho_F = \rho \left(\frac{N}{\|N\|}, \frac{2\pi}{5} \right),$$

the rotation about $\frac{N}{\|N\|}$ by $\frac{2\pi}{5}$. We shall show that ρ_F lies in $\mathcal{S}(\mathbf{D})$ and permutes the vertices of F . A key in doing this is that $\cos \frac{2\pi}{5} = \frac{\phi}{2}$ as shown in (6.3.4). Using this, the following can be obtained by a good computer

algebra program (we used Maple) and can be verified by a tedious, but direct calculation. But we can also use theory to simplify that verification a bit:

Proposition 7.6.13. *The matrix inducing ρ_F is given by*

$$(7.6.26) \quad R_F = \begin{bmatrix} \frac{1}{2} & -\frac{\Phi}{2} & \frac{\phi}{2} \\ \frac{\Phi}{2} & \frac{\phi}{2} & -\frac{1}{2} \\ \frac{\phi}{2} & \frac{1}{2} & \frac{\Phi}{2} \end{bmatrix}.$$

The rotation ρ_F acts on the vertices v_i and x_i in V as follows:

$$(7.6.27) \quad v_0 \mapsto v_1 \mapsto v_2 \mapsto v_3 \mapsto v_4 \mapsto v_0$$

$$(7.6.28) \quad x_0 \mapsto x_1 \mapsto x_2 \mapsto x_3 \mapsto x_4 \mapsto x_0.$$

Since every vertex or its negative is one of these, and since ρ_F is linear, ρ_F permutes the elements of V , and hence lies in $\mathcal{S}(\mathbf{D})$.

Proof. It is easy to verify (7.6.27) and (7.6.28) by hand and that $R_F \cdot N = N$. Since any three distinct elements of U are linearly independent, this verifies that R_F has order 5 and fixes N . (The order is also verified in what follows.)

It is also easy to verify that the columns of R_F are orthonormal, so that R_F is an orthogonal matrix. To see it gives the desired rotation, we argue as follows. Let

$$u = \frac{N}{|N|} = \frac{1}{\sqrt{\Phi^2 + 1}} \begin{bmatrix} 1 \\ 0 \\ \Phi \end{bmatrix}.$$

Then e_2 is orthogonal to u and an easy calculation shows that

$$u \times e_2 = \frac{1}{\sqrt{\Phi^2 + 1}} \begin{bmatrix} -\Phi \\ 0 \\ 1 \end{bmatrix}.$$

We obtain an orthonormal basis $\mathcal{B} = e_2, w, u$, with $w = u \times e_2$, and it suffices to show that the matrix $|T_{R_F}|_{\mathcal{B}}$ of T_{R_F} with respect to \mathcal{B} is $R_{(e_3, \frac{2\pi}{5})}$. Since \mathcal{B} is orthonormal and R_F is orthogonal and fixes u ,

$$(7.6.29) \quad |T_{R_F}|_{\mathcal{B}} = \begin{bmatrix} \langle R_F \cdot e_2, e_2 \rangle & \langle R_F \cdot w, e_2 \rangle & 0 \\ \langle R_F \cdot e_2, w \rangle & \langle R_F \cdot w, w \rangle & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

If we accept the Maple calculation that $\det R_F = 1$, then it suffices to notice that $\langle R_F \cdot e_2, e_2 \rangle$ is the (2, 2) coordinate of R_F , which is $\frac{\phi}{2} = \cos \frac{2\pi}{5}$ by (6.3.4), and that $\langle R_F \cdot e_2, w \rangle$ is positive. Without the determinant calculation, one must additionally show that $\langle R_F \cdot w, e_2 \rangle$ is negative. By (4.1.4),

$$\langle R_F \cdot w, e_2 \rangle = \langle w, R_F^T e_2 \rangle,$$

which is again easy to calculate. \square

Corollary 7.6.14. $\mathcal{S}(\mathbf{D})$ acts transitively on the vertices of \mathbf{D} .

Proof. Every element of V lies in exactly one $\mathcal{S}(\mathbf{D})$ -orbit. Since $\rho_F(v_0) = v_1$, v_0 and v_1 lie in the same $\mathcal{S}(\mathbf{D})$ -orbit. Since $H \subset \mathcal{S}(\mathbf{D})$, $H \cdot v_0 \subset \mathcal{S}(\mathbf{D}) \cdot v_0$, i.e., $H \cdot v_0 = S$ is contained in this $\mathcal{S}(\mathbf{D})$ -orbit, as is $H \cdot v_1 = V \setminus S$. So this $\mathcal{S}(\mathbf{D})$ -orbit consists of all of V . \square

Since v_0 and v_1 are 2ϕ apart and ρ_F is an isometry, the following is immediate from Proposition 7.6.13.

Corollary 7.6.15. *F is a regular pentagon with edge length 2ϕ .*

We now calculate the order of $\mathcal{S}(\mathbf{D})$.

Corollary 7.6.16. *The isotropy subgroup of any vertex under the action of $\mathcal{S}(\mathbf{D})$ has order six and is isomorphic to Σ_3 . Thus, $\mathcal{S}(\mathbf{D})$ has order 120.*

Proof. Since $\mathcal{S}(\mathbf{D})$ acts transitively on the vertices, our isotropy claim will follow if we prove it for v_0 . Since v_0, v_1, v_4 are linearly independent, any element in the isotropy subgroup of v_0 is determined by its effect on v_1 and v_4 (it fixes v_0). The only vertices of distance 2ϕ from v_0 are v_1 , v_4 and $z = \begin{bmatrix} \phi \\ \phi \\ 0 \end{bmatrix}$. So $\mathcal{S}(\mathbf{D})_{v_0}$ embeds in the permutation group $\Sigma(\{v_1, v_4, z\}) \cong \Sigma_3$.

Σ_3 has six elements. The cyclic permutation induced by $\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ fixes v_0 , so $\mathcal{S}(\mathbf{D})_{v_0}$ has order divisible by three.

Let

$$(7.6.30) \quad \sigma = T_A \quad \text{for} \quad A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Then σ permutes the vertices of F and fixes v_4 . σ is the reflection across the xz -plane and has order 2.

Write

$$(7.6.31) \quad \tau = \rho_F \sigma \rho_F^{-1}.$$

Then τ permutes the vertices of F , fixes v_0 and has order 2. So the order of $\mathcal{S}(\mathbf{D})_{v_0}$ is divisible by 2. Thus, $\mathcal{S}(\mathbf{D})_{v_0}$ induces the full permutation group $\Sigma(\{v_1, v_4, z\})$ and has order six.

But $\mathcal{S}(\mathbf{D})$ acts transitively on the vertices of \mathbf{D} so the orbit $\mathcal{S}(\mathbf{D})v_0$ has 20 elements. By Corollary 6.7.9, $\mathcal{S}(\mathbf{D})$ has order $|\mathcal{S}(\mathbf{D})_{v_0}| |\mathcal{S}(\mathbf{D})v_0| = 120$. \square

We next determine the symmetries of \mathbf{D} that take F to F , i.e., the group $\mathcal{S}(\mathbf{D}) \cap \mathcal{S}(F)$.

Lemma 7.6.17. *$\mathcal{S}(\mathbf{D}) \cap \mathcal{S}(F)$ is a dihedral group of order 10. Its elements are*

$$\{\rho_F^k, \rho_F^k \tau : 0 \leq k \leq 4\},$$

where τ is given in (7.6.31).

Proof. Since both σ and ρ_F are in $\mathcal{S}(\mathbf{D}) \cap \mathcal{S}(F)$, so is τ . τ fixes v_0 . It also exchanges v_1 and v_4 and exchanges v_2 and v_3 . Since v_0, v_1 and v_4 are linearly independent, a linear isometry of \mathbb{R}^3 is determined by its effect on v_0, v_1 and v_4 , and hence by its effect on F . So τ has order 2 and satisfies $\tau\rho_F\tau^{-1} = \rho_F^{-1}$. In particular, τ and ρ_F generate a group isomorphic to the symmetry group of the standard regular pentagon P_5 , which is the dihedral group D_{10} . Note that adjacent vertices of F are 2ϕ apart, while nonadjacent vertices have distance 2 from each other. So the argument given for the calculation of $\mathcal{S}(P_n)$ (Proposition 6.5.1) applies here to show there are no other elements in $\mathcal{S}(\mathbf{D}) \cap \mathcal{S}(F)$. \square

By Proposition 6.2.12, an isometry $\alpha \in \mathcal{S}(\mathbf{D})$ carries the centroid of F to the centroid of $\alpha(F)$. Thus, $\alpha(F) = F$ if and only if $\alpha(\frac{1}{5}N) = \frac{1}{5}N$. We obtain the following:

Lemma 7.6.18. *Let K be a subgroup of $\mathcal{S}(\mathbf{D})$. Then set of faces*

$$K \cdot F = \{\alpha(F) : \alpha \in K\}$$

is in one-to-one correspondence with the orbit $K \cdot \frac{1}{5}N$. So

$$|K \cdot F| = [K : K_N] = \frac{|K|}{|K_N|}$$

is the index of the isotropy subgroup, K_N , of N under the action of K .

Proof. For the last statement, we note that since the isometries in K are linear the isotropy subgroups of N and $\frac{1}{5}N$ coincide. \square

By Lemma 7.6.17, $\mathcal{S}(\mathbf{D})_N$ has order 10, so $\mathcal{S}(\mathbf{D}) \cdot F$ consists of all 12 visible faces of the dodecahedron. This begs the question of whether there are any other faces of \mathbf{D} . In other words, if the convex hull of a subset of V is a face of \mathbf{D} , is it one of the faces $\alpha(F)$ for $\alpha \in \mathcal{S}(\mathbf{D})$? (I.e., does $\mathcal{S}(\mathbf{D})$ act transitively on the faces of \mathbf{D} ?)

We shall not give a detailed argument for this. One could use some more advanced topology, but one can also use Proposition 2.9.39 to show the following.

Proposition 7.6.19. *Let $v \neq w$ be vertices of \mathbf{D} . Then the segment*

$$[v, w] = \overline{vw} = \text{Conv}(v, w)$$

is an edge of \mathbf{D} if and only if the distance $d(v, w)$ from v to w is 2ϕ . Moreover:

- (1) *If $d(v, w) = 2$, then $(v, w) = \text{Int}([v, w])$ is contained in the interior of a face $\alpha(F)$ with $\alpha \in \mathcal{S}(\mathbf{D})$.*
- (2) *If $v, w \in V$ have distance greater than 2 from one another, then (v, w) lies in the interior of \mathbf{D} .*

Proof. Since $\mathcal{S}(\mathbf{D})$ acts transitively on the vertices, we may assume $v = v_0$. There are exactly three vertices of distance 2ϕ from v_0 : v_1, v_4 and z . They are cyclically permuted by the symmetry $\rho = T_A$, $A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$, which lies in $H \subset \mathcal{S}(\mathbf{D})$.¹³

The segment $[v_0, v_1]$ is contained in $F \cap \rho^2(F)$, the intersection of two 2-dimensional faces. Since these two faces are distinct, their intersection is a face of dimension less than 2. As it contains a segment, $F \cap \rho^2(F)$ must be an edge. Since v_0 and v_1 are vertices, they cannot lie in the interior of an edge. Thus $F \cap \rho^2(F) = [v_0, v_1]$ is an edge. But $[v_0, v_4]$ and $[v_0, z]$ are images of $[v_0, v_1]$ under powers of ρ , and hence are edges also.

The vertices of distance 2 from v_0 are all images under powers of ρ of either v_2 or v_3 . So (1) will follow if we show (v_0, v_2) and (v_0, v_3) are contained in $\text{Int}(F)$. But this follows from Corollary 2.9.46, as both (v_0, v_2) and (v_0, v_3) intersect (v_1, v_4) , so all three segments have the same carrier, which must perforce contain the vertices v_0, v_1, v_2, v_3, v_4 : the full vertex set of F .

Note that the map τ of (7.6.31) fixes v_0 and exchanges v_2 and v_3 . It is then easy to see that the isotropy subgroup $\mathcal{S}(\mathbf{D})_{v_0}$ acts transitively on the vertices of distance 2 from v_0 . The vertices of distance $2\sqrt{2}$ from v_0 are the negatives of those of distance 2. Since the isometries are linear, $\mathcal{S}(\mathbf{D})_{v_0}$ acts transitively on those of distance $2\sqrt{2}$ as well. So in analyzing this case, we may simply study $[v_0, w]$ for $w = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}$. The midpoint of $[v_0, w]$ is the canonical basis vector e_3 . It suffices to show the carrier of e_3 is \mathbf{D} .

Note that $\begin{bmatrix} 0 \\ 0 \\ \Phi \end{bmatrix}$ is the midpoint of the vertices $\begin{bmatrix} 0 \\ \phi \\ \phi \end{bmatrix}$ and $\begin{bmatrix} 0 \\ -\phi \\ \phi \end{bmatrix}$, while $\begin{bmatrix} 0 \\ 0 \\ -\Phi \end{bmatrix}$ is the midpoint of their negatives, so both these points are in \mathbf{D} . And both e_3 and 0 lie in the interior of the segment from $\begin{bmatrix} 0 \\ 0 \\ \Phi \end{bmatrix}$ to $\begin{bmatrix} 0 \\ 0 \\ -\Phi \end{bmatrix}$. So e_3 and 0 have the same carrier. Since 0 is the centroid of \mathbf{D} , that carrier is \mathbf{D} .

The segments of length 2Φ are easy, as their midpoints are easily seen to lie in $\text{Int}(\mathbf{C})$, which perforce must be contained in $\text{Int}(\mathbf{D})$. The remaining segment, of length $2\sqrt{3}$, has 0 as its midpoint. \square

Corollary 7.6.20. *The faces $\alpha(F)$ with $\alpha \in \mathcal{S}(\mathbf{D})$ are the only faces of \mathbf{D} .*

Proof. Let G be a face of \mathbf{D} with vertex set $R \subset V$. Then R cannot contain a pair of vertices of distance greater than 2 from each other, as then G would intersect $\text{Int}(\mathbf{D})$.

Suppose there is a pair $v, w \in R$ of distance two from each other. Then there is a face $\alpha(F)$ containing $[v, w]$, so that $G \cap \alpha(F)$ has dimension at least 1. The intersection of two faces, if nonempty, is either a vertex, and edge or a face. But $[v, w]$ meets the interior of $\alpha(F)$, so it is not an edge. And $\alpha(F)$ is the unique face containing any point in its interior. So $G = \alpha(F)$.

¹³One can show that $\rho = \rho_{(v_0, \frac{2\pi}{3})}$ by the methods used in the proof of Proposition 7.6.13.

Finally, it is impossible that each pair in R has distance 2ϕ from each other: there are at least three vertices in R . If $u, v, w \in R$ with

$$d(u, v) = d(v, w) = 2\phi,$$

then, applying an isometry taking v to v_0 we see that $d(u, w) = 2$. \square

The following is now useful.

Proposition 7.6.21. *The subgroup $H = \mathcal{S}(\mathbf{D}) \cap \mathcal{S}(\mathbf{C})$ acts transitively on the 12 faces of the dodecahedron. So does, $\mathcal{O}(H) \subset H$, the subgroup of orientation-preserving transformations in H . As $|\mathcal{O}(H)| = 12$, we obtain a bijection from $\mathcal{O}(H)$ to the set of faces of \mathbf{D} given by*

$$\alpha \mapsto \alpha(F).$$

Proof. The only signed cyclic permutation matrices that preserve N are I_3 and the matrix A in (7.6.30). So the isotropy subgroup H_N has order 2, and hence index 12 in H . Since $\det A = -1$, $\mathcal{O}(H)_N$ is the identity subgroup, and the result follows. \square

The following is a nonstandard definition we find useful here.

Definition 7.6.22. A chord in a face F' of \mathbf{D} is a segment \overline{vw} between nonadjacent vertices v, w of F' (i.e., \overline{vw} is not an edge of F').

Note that each face of \mathbf{D} has exactly five chords. Each has length 2. (Each is the image under an isometry of the chord $\overline{v_0v_3}$ of F .)

Corollary 7.6.23. *Let F' be a face of \mathbf{D} . Then $F' \cap \mathbf{C}$ is both an edge of \mathbf{C} and a chord of F' .*

Proof. Each face F' of \mathbf{D} has the form $\alpha(F)$ for some $\alpha \in H$. The face F satisfies

$$F = \mathbf{D} \cap f^{-1}(4\Phi + 3)$$

where f is the linear function $f(x) = \langle x, N \rangle$ with N the sum of the vertices of F . The image of f on \mathbf{D} is $[-4\Phi - 3, 4\Phi + 3]$, and the same is true if we restrict f to \mathbf{C} . Indeed,

$$(7.6.32) \quad f^{-1}(4\Phi + 3) \cap \mathbf{C} = \text{Conv}(v_0, v_3)$$

by our calculation of f on the vertices of \mathbf{C} . Thus, our assertion is true for $F' = F$. Since the elements of H are symmetries of \mathbf{C} as well as \mathbf{D} , the same is true for F' . \square

Corollary 7.6.24. *Let $\alpha \in \mathcal{S}(\mathbf{D})$ and let F' be a face of \mathbf{D} . Then the intersection of the cube $\alpha(\mathbf{C})$ with F' is both an edge of $\alpha(\mathbf{C})$ and a chord of F' .*

Proof. This is simply the image under α of $\mathbf{C} \cap \alpha^{-1}(F')$. \square

We shall make use of the following.

Lemma 7.6.25. *A cube with centroid at the origin is determined by any one of its edges.*

Proof. Here, we define a cube as being a polytope similar to the standard cube \mathbf{C} . Let e be an edge of \mathbf{C} and let u be a unit vector parallel to e . Then the vertices of \mathbf{C} are obtained by rotating the vertices of e by increments of $\frac{\pi}{2}$ about u . The same relationship will hold in any similar polytope centered at 0. \square

Indeed, a cube in \mathbb{R}^3 is determined by its centroid and any one of its edges.

Proposition 7.6.26. *Let X be the set of cubes*

$$X = \{\alpha(\mathbf{C}) : \alpha \in \mathcal{S}(\mathbf{D})\}$$

and let Y be the set of chords of F . Then there is a one-to-one correspondence $g : X \rightarrow Y$ given by setting $g(\alpha(\mathbf{C})) = \alpha(\mathbf{C}) \cap F$. Thus, there are five elements in X .

Proof. The map g is well-defined by Corollary 7.6.24 and is one-to-one by Lemma 7.6.25. It is onto because each chord in F is the image of $\overline{v_0v_3}$ under a power of ρ_F . \square

We obtain the following.

Theorem 7.6.27. *There is a group homomorphism*

$$(7.6.33) \quad \begin{aligned} \varepsilon : \mathcal{S}(\mathbf{D}) &\rightarrow \Sigma(X) \cong \Sigma_5 \\ \varepsilon(\alpha) &= \alpha(\mathbf{C}). \end{aligned}$$

The kernel of ε is $\{\pm I_3\}$. The restriction

$$\varepsilon : \mathcal{O}(\mathbf{D}) \rightarrow \Sigma_5$$

is injective and induces an isomorphism

$$(7.6.34) \quad \varepsilon : \mathcal{O}(\mathbf{D}) \xrightarrow{\cong} A_5$$

of $\mathcal{O}(\mathbf{D})$ onto the alternating group A_5 . The image of $\varepsilon : \mathcal{S}(\mathbf{D}) \rightarrow \Sigma_5$ is also A_5 .

Proof. The kernel of ε is

$$\ker \varepsilon = \mathcal{S}(\mathbf{D}) \cap \bigcap_{\alpha \in \mathcal{S}(\mathbf{D})} \mathcal{S}(\alpha(\mathbf{C})),$$

the set of isometries of \mathbf{D} that also preserve each of the cubes $\alpha(\mathbf{C})$. Since each $\alpha \in \mathcal{S}(\mathbf{D})$ is linear, $\{\pm I_3\} \subset \ker \varepsilon$. Moreover, $\ker \varepsilon \subset \mathcal{S}(\mathbf{D}) \cap \mathcal{S}(\mathbf{C}) = H$, a group we understand well. Indeed, $\ker \varepsilon \subset H \cap \mathcal{S}(\rho_F(\mathbf{C}))$, so it suffices to show that

$$(7.6.35) \quad H \cap \mathcal{S}(\rho_F(\mathbf{C})) = \{\pm I_3\}.$$

An easy calculation shows that the vertices of $\rho_F(\mathbf{C})$ are

$$\pm \begin{bmatrix} 0 \\ \phi \\ \Phi \end{bmatrix}, \pm \begin{bmatrix} \Phi \\ 0 \\ \phi \end{bmatrix}, \pm \begin{bmatrix} -\phi \\ \Phi \\ 0 \end{bmatrix}, \pm \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}.$$

And $\pm I_3$ are the only elements of H that preserve this set. Since $-I_3$ is orientation-reversing, $\varepsilon : \mathcal{O}(\mathbf{D}) \rightarrow \Sigma_5$ is an injection onto a 60-element subgroup of Σ_5 . So $\varepsilon(\mathcal{O}(\mathbf{D}))$ has index 2 in Σ_5 . Corollary 6.6.13 shows that A_5 is the only index 2 subgroup of Σ_5 , and (7.6.34) follows.

Finally, $\mathcal{O}(\mathbf{D})$ has index 2 in $\mathcal{S}(\mathbf{D})$ and $-I_3$ is orientation-reversing, so every element of $\mathcal{S}(\mathbf{D}) \setminus \mathcal{O}(\mathbf{D})$ has the form $A \cdot (-I_3)$ for $A \in \mathcal{O}(\mathbf{D})$. But $\varepsilon(A \cdot (-I_3)) = \varepsilon(A) \in A_5$, and the result follows. \square

7.6.5. Duality. The octahedron and the isosahedron are what's known as dual polyhedra to the cube and the dodecahedron, respectively:

Definition 7.6.28. Let \mathbf{P} be a 3-dimensional polytope. We write $\mathcal{F}(\mathbf{P})$ for the set of (2-dimensional) faces of \mathbf{P} . Similarly, we write $\mathcal{E}(\mathbf{P})$ and $\mathcal{V}(\mathbf{P})$ for the sets of edges and vertices of \mathbf{P} , respectively. For a face $F \in \mathcal{F}(\mathbf{P})$, we write $c(F)$ for its centroid and write

$$(7.6.36) \quad c\mathcal{F}(\mathbf{P}) = \{c(F) : F \in \mathcal{F}(\mathbf{P})\}.$$

We define the dual, $d(\mathbf{P})$, of \mathbf{P} to be the convex hull of the centroids of its faces:

$$(7.6.37) \quad d(\mathbf{P}) = \text{Conv}(c\mathcal{F}(\mathbf{P})).$$

We can make one observation immediately: since isometries carry faces to faces and carry the centroid of a face to the centroid of its image, we obtain an inclusion $\mathcal{S}(\mathbf{P}) \subset \mathcal{S}(c\mathcal{F}(\mathbf{P}))$ of subgroups of \mathcal{I}_3 . By Lemma 6.2.1 $\mathcal{S}(c\mathcal{F}(\mathbf{P}))$ is a subgroup of $\mathcal{S}(d(\mathbf{P}))$, even if the centroids $c(F)$ are not vertices of $d(\mathbf{P})$. We obtain:

Lemma 7.6.29. *Let \mathbf{P} be a 3-dimensional polytope in \mathbb{R}^3 . Then we have inclusions of subgroups*

$$\mathcal{S}(\mathbf{P}) \subset \mathcal{S}(c\mathcal{F}(\mathbf{P})) \subset \mathcal{S}(d(\mathbf{P})) \subset \mathcal{I}_3.$$

We shall not treat the theory of duals in any kind of general way, but will show for each of the Platonic solids \mathbf{P} that we've studied, the dual $d(\mathbf{P})$ is another Platonic solid. Moreover the above inclusion

$$\mathcal{S}(\mathbf{P}) \subset \mathcal{S}(d(\mathbf{P}))$$

is the identity for a Platonic solid \mathbf{P} , i.e., $\mathcal{S}(\mathbf{P}) = \mathcal{S}(d(\mathbf{P}))$.

Moreover, we shall show that the dual of a tetrahedron is a tetrahedron, and that the duals of the cube and dodecahedron are the octahedron and icosahedron, respectively. And these five solids exhaust the collection of Platonic solids.

The relationship between $\mathcal{S}(P)$ and $\mathcal{S}(d(P))$ has already been forecast in the calculation of the symmetries of the n -cube $\mathbf{C}^n = [-1, 1]^n$ in Proposition 6.2.17. It was shown that the symmetries of \mathbf{C}^n are determined by

their effect on the centroids of its $(n-1)$ -dimensional faces. These centroids form the vertex set of an n -dimensional analogue of the duals we study here, providing an interesting higher-dimensional analogue of the octahedron:

Definition 7.6.30. The n -cross, or n -orthoplex, is the convex hull of

$$\{\pm e_1, \dots, \pm e_n\} \subset \mathbb{R}^n$$

(i.e., each of e_i and $-e_i$ is in the convex generating set for $i = 1, \dots, n$). The 4-cross is also known as the 16-cell.

7.6.6. The octahedron. The standard regular octahedron is the dual of the standard balanced 3-dimensional cube \mathbf{C} . Let $O = \{\pm e_1, \pm e_2, \pm e_3\}$, where e_1, e_2, e_3 is the canonical basis of \mathbb{R}^3 and each of e_i and $-e_i$ is in O for $i = 1, 2, 3$. Then $O = c(\mathbf{C})$, and the octahedron \mathbf{O} is given by

$$(7.6.38) \quad \mathbf{O} = \text{Conv}(O),$$

Each of the six points $\pm e_1, \pm e_2, \pm e_3$ is a vertex of \mathbf{O} , as $\pm e_i$ are the extreme points in O of the linear function given by the projection onto the i -th factor of \mathbb{R}^3 . As in Proposition 6.2.17, the linear maps which permute O are precisely the signed permutations matrices $\text{O}(3, \mathbb{Z}) = \mathcal{S}(\mathbf{C})$. So

$$(7.6.39) \quad \mathcal{S}(\mathbf{C}) = \mathcal{S}(\mathbf{O})$$

as subgroups of \mathcal{I}_3 .

We can use the geometry of dualization to study the structure of the octahedron \mathbf{O} . First consider the following. Its proof is obvious.

Lemma 7.6.31. Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ be the linear function $f(x) = \langle x, v_0 \rangle$, where $v_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ as above. Then on O , we have $f(e_i) = 1$ and $f(-e_i) = -1$ for $i = 1, 2, 3$. Thus $f(\mathbf{O}) = [-1, 1]$ and

$(f|_{\mathbf{O}})^{-1}(1) = \text{Conv}(e_1, e_2, e_3)$ and $(f|_{\mathbf{O}})^{-1}(-1) = \text{Conv}(-e_1, -e_2, -e_3)$ are faces of \mathbf{O} .

Note that v_0 is a vertex of \mathbf{C} . We can repeat the argument with the other vertices of \mathbf{C} to obtain the following.

Lemma 7.6.32. The sets $\text{Conv}(\epsilon_1 e_1, \epsilon_2 e_2, \epsilon_3 e_3)$ are all faces of \mathbf{O} , where $\epsilon_1, \epsilon_2, \epsilon_3 \in \{-1, 1\}$. Indeed, $\epsilon_1 e_1, \epsilon_2 e_2$, and $\epsilon_3 e_3$ attain the maximum value on O of the linear function $f_v(x) = \langle x, v \rangle$, where $v = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}$ is a vertex of \mathbf{C} .

Thus, we obtain eight faces of \mathbf{O} , each isometric to $\text{Conv}(e_1, e_2, e_3) = \Delta^2$, an equilateral triangle in \mathbb{R}^3 . The symmetry group $\mathcal{S}(\mathbf{O})$ acts transitively on them.

Proof. The last statement follows since the signed permutations in $\text{O}(3, \mathbb{Z})$ permute the vertex sets of these faces, and hence acts on this set of faces. It is easy to check that the action is transitive. \square

We've constructed a face of $\mathbf{O} = d(\mathbf{C})$ corresponding to each vertex of \mathbf{C} , just as there is a vertex of \mathbf{O} corresponding to each face of \mathbf{C} . We show next that there are no other faces.

Lemma 7.6.33. *Let S be a proper subset of O with at least three elements that is not one of the vertex sets of the faces in Lemma 7.6.32. Then $\text{Conv}(S)$ contains 0 and hence is not a face of \mathbf{O} . In particular, all the faces of \mathbf{O} are given in Lemma 7.6.32.*

Proof. Under our hypotheses, S must contain both e_i and $-e_i$ for some i , so $\text{Conv}(S)$ contains 0. Since $S(0) = O$, it lies in the interior of \mathbf{O} . \square

Every segment $[\pm e_i, \pm e_j]$ with $i < j$, signs varying independently, is the intersection of two faces, one with additional vertex e_k , the other with additional vertex $-e_k$, where $k = \{1, 2, 3\} \setminus \{i, j\}$. In particular, these segments are edges. The other vertex segments are $[-e_i, e_i]$ which contain 0 and hence are not edges. We obtain:

Lemma 7.6.34. *The edges of \mathbf{O} are the segments $[\pm e_i, \pm e_j]$ with $i < j$, signs varying independently. Since there are $\binom{3}{2}$ such pairs $i < j$ and four choices of sign for each pair, this gives 12 edges.*

Finally, we note that the dual of \mathbf{O} is a rescaled version of \mathbf{C} . To see this, note that the centroid of $\text{Conv}(\epsilon_1 e_1, \epsilon_2 e_2, \epsilon_3 e_3)$ is $\frac{1}{3}(\epsilon_1 e_1 + \epsilon_2 e_2 + \epsilon_3 e_3) = \frac{1}{3} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}$, a rescaling by $\frac{1}{3}$ of the vertex $\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}$ of \mathbf{C} . Since this is true for all faces of \mathbf{O} , we obtain the following.

Proposition 7.6.35. *The double dual of \mathbf{C} is the rescaling of \mathbf{C} by a factor of $\frac{1}{3}$:*

$$d(d(\mathbf{C})) = d(\mathbf{O}) = \frac{1}{3}\mathbf{C}.$$

$\frac{1}{3}\mathbf{C}$ has exactly the same symmetry group as \mathbf{C} (and \mathbf{O}) as a subgroup of the linear isometries of \mathbb{R}^3 .

7.6.7. Dual of the tetrahedron. We compute of the dual of the standard regular tetrahedron \mathbf{T} . As above, we write S for the vertex set (7.6.2) of the cube \mathbf{C} and write $T \subset S$ for the vertices of \mathbf{T} given in (7.6.5):

$$T = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix} \right\} \subset S.$$

The vertex set $T' = S \setminus T$ is then the vertex set for the complementary tetrahedron \mathbf{T}' .

The four faces of \mathbf{T} are the convex hulls of each of the 3-element subsets of T . And each 3-element subset is determined by the vertex it does not contain: we write $T_v = T \setminus \{v\}$ for $v \in T$. Then the faces of \mathbf{T} are:

$$(7.6.40) \quad \mathcal{F}(\mathbf{T}) = \{\text{Conv}(T_v) : v \in T\}.$$

Lemma 7.6.36. *The centroid of $\text{Conv}(T_v)$ is $-\frac{1}{3}v$.*

Proof. The vertices in T add up to 0, so the sum of the vertices unequal to v is $-v$. The $\frac{1}{3}$ comes from taking their average. \square

The vertices in T' are the negatives of those in T , so the following is immediate.

Corollary 7.6.37. *The dual of the tetrahedron \mathbf{T} is the rescaling $\frac{1}{3}\mathbf{T}'$ of the tetrahedron \mathbf{T}' . Iterating this, we see that the double dual $d(d(\mathbf{T})) = \frac{1}{9}\mathbf{T}$, and similarly for \mathbf{T}' . All of these tetrahedra have the same group of isometries.*

7.6.8. The icosahedron. The icosahedron is the dual of the dodecahedron. Recall from Proposition 7.6.21 that the subgroup $H = \mathcal{S}(\mathbf{C}) \cap \mathcal{S}(\mathbf{D})$ acts transitively on the faces of \mathbf{D} . Here, (see Lemma 7.6.6 and Corollary 7.6.8) H is the group of linear isometries induced by the matrices

$$\left\{ \begin{bmatrix} \pm 1 & 0 & 0 \\ 0 & \pm 1 & 0 \\ 0 & 0 & \pm 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & \pm 1 \\ \pm 1 & 0 & 0 \\ 0 & \pm 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & \pm 1 & 0 \\ 0 & 0 & \pm 1 \\ \pm 1 & 0 & 0 \end{bmatrix} \right\},$$

where the signs vary independently of one another. (These are the signed cyclic permutation matrices. They cyclically permute the coordinates and add signs.) By (7.6.24), the centroid $\frac{1}{5}N$ of the face F we have studied is

$$(7.6.41) \quad \frac{1}{5}N = \frac{\Phi + 2}{5} \begin{bmatrix} 1 \\ 0 \\ \Phi \end{bmatrix}.$$

The set of all centroids of faces of \mathbf{D} is the orbit of this point under the action of H :

$$(7.6.42) \quad c\mathcal{F}(\mathbf{D}) = \left\{ \frac{\Phi + 2}{5} \begin{bmatrix} \pm 1 \\ 0 \\ \pm \Phi \end{bmatrix}, \frac{\Phi + 2}{5} \begin{bmatrix} 0 \\ \pm \Phi \\ \pm 1 \end{bmatrix}, \frac{\Phi + 2}{5} \begin{bmatrix} \pm \Phi \\ \pm 1 \\ 0 \end{bmatrix} \right\},$$

where the signs vary independently.

To simplify notation, we shall rescale and set

$$(7.6.43) \quad W = \left\{ \begin{bmatrix} \pm 1 \\ 0 \\ \pm \Phi \end{bmatrix}, \begin{bmatrix} 0 \\ \pm \Phi \\ \pm 1 \end{bmatrix}, \begin{bmatrix} \pm \Phi \\ \pm 1 \\ 0 \end{bmatrix} \right\},$$

and define the standard icosahedron \mathbf{I} to be the convex hull $\text{Conv}(W)$. Since the isometries of \mathbf{D} are all linear and preserve $c\mathcal{F}(\mathbf{D})$, we have

$$(7.6.44) \quad \mathcal{S}(\mathbf{D}) \subset \mathcal{S}(W) \subset \mathcal{S}(\mathbf{I}).$$

Of course, H acts transitively on W .

Let's give names to some vertices:

$$(7.6.45) \quad u_0 = \begin{bmatrix} 1 \\ 0 \\ \Phi \end{bmatrix}, \quad u_1 = \begin{bmatrix} \Phi \\ 1 \\ 0 \end{bmatrix}, \quad u_2 = \begin{bmatrix} 0 \\ \Phi \\ 1 \end{bmatrix}.$$

Taking u_0 as our base vertex, define the linear function $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ by

$$f(x) = \langle x, u_0 \rangle.$$

The following calculation is left to the reader.

Lemma 7.6.38. *The restriction of f to W is given by*

$$(7.6.46) \quad f(w) = \begin{cases} \Phi + 2 & \text{for } w = u_0, \\ \Phi & \text{for } w = \begin{bmatrix} \Phi \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \Phi \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ \Phi \end{bmatrix}, \begin{bmatrix} 0 \\ -\Phi \\ 1 \end{bmatrix}, \begin{bmatrix} \Phi \\ -1 \\ 0 \end{bmatrix}, \\ -\Phi & \text{for } w = \begin{bmatrix} -\Phi \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -\Phi \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ -\Phi \end{bmatrix}, \begin{bmatrix} 0 \\ \Phi \\ -1 \end{bmatrix}, \begin{bmatrix} -\Phi \\ 1 \\ 0 \end{bmatrix}, \\ -\Phi - 2 & \text{for } w = -u_0. \end{cases}$$

Thus, $u = (f|_W)^{-1}(\Phi + 2)$ is a vertex of \mathbf{I} , and since H acts transitively on W , $W = \mathcal{V}(\mathbf{I})$, the set of vertices of \mathbf{I} .

To find faces of \mathbf{I} we dot with $v_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$: let $g(x) = \langle x, v_0 \rangle$.

Lemma 7.6.39. *The values of the linear function g on W are given by*

$$(7.6.47) \quad g(w) = \begin{cases} \Phi + 1 & \text{for } w = u_0, u_1, u_2, \\ \phi & \text{for } w = \begin{bmatrix} -1 \\ 0 \\ \Phi \end{bmatrix}, \begin{bmatrix} 0 \\ \Phi \\ -1 \end{bmatrix}, \begin{bmatrix} \Phi \\ -1 \\ 0 \end{bmatrix}, \\ -\phi & \text{for } w = \begin{bmatrix} 0 \\ \Phi \\ -1 \end{bmatrix}, \begin{bmatrix} -\Phi \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -\Phi \\ 1 \end{bmatrix}, \\ -\Phi - 1 & \text{for } w = -u_0, -u_1, -u_2. \end{cases}$$

Thus, $G = \text{Conv}(u_0, u_1, u_2)$ is a face of \mathbf{I} with centroid

$$(7.6.48) \quad c(G) = \frac{\Phi + 1}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

(see (7.6.45) for this calculation). Since the orbit of v_0 under $\mathcal{S}(\mathbf{D})$ has 20 elements, there are 20 faces of the form $\alpha(G)$ with $\alpha \in \mathcal{S}(\mathbf{D})$.

Let us name some more vertices:

$$(7.6.49) \quad u_3 = \begin{bmatrix} -1 \\ 0 \\ \Phi \end{bmatrix}, \quad u_4 = \begin{bmatrix} 0 \\ -\Phi \\ 1 \end{bmatrix}, \quad u_5 = \begin{bmatrix} \Phi \\ -1 \\ 0 \end{bmatrix}.$$

The calculations in (7.6.46) now give us the following:

Lemma 7.6.40. *Let $w \in W$. Then*

$$(7.6.50) \quad d(u_0, w) = \begin{cases} 2 & \text{for } w = u_1, u_2, u_3, u_4, u_5, \\ 2\Phi & \text{for } w = -u_1, -u_2, -u_3, -u_4, -u_5, \\ 2\sqrt{\Phi + 2} & \text{for } w = -u_0. \end{cases}$$

Since $d(u_1, u_2) = 2$, we have:

Lemma 7.6.41. *G is an equilateral triangle with side length 2.*

The rotation ρ_F of Proposition 7.6.13 fixes u_0 and has order 5. Since it preserves distance, it must permute the set $\{u_1, u_2, u_3, u_4, u_5\}$. A little group theory shows this permutation must be cyclic. We can determine more with brute force calculation (left to the reader):

Lemma 7.6.42. ρ_F permutes $\{u_1, u_2, u_3, u_4, u_5\}$ as follows:

$$(7.6.51) \quad u_1 \mapsto u_2 \mapsto u_3 \mapsto u_4 \mapsto u_5 \mapsto u_1.$$

Since ρ_F is linear, it acts on their negatives by

$$(7.6.52) \quad -u_1 \mapsto -u_2 \mapsto -u_3 \mapsto -u_4 \mapsto -u_5 \mapsto -u_1.$$

Corollary 7.6.43. There are exactly five faces of the form $\alpha(G)$, $\alpha \in \mathcal{S}(\mathbf{D})$, with u_0 as a vertex: G_1, \dots, G_5 , where

$$(7.6.53) \quad G_i = \begin{cases} \text{Conv}(u_0, u_i, u_{i+1}) = \rho_F^{i-1}(G) & \text{for } i = 1, \dots, 4, \\ \text{Conv}(u_0, u_5, u_1) = \rho_F^4(G) & \text{for } i = 5. \end{cases}$$

Proof. Of course $G = G_1$. These five faces are immediate from Lemma 7.6.42. They are the only ones because they are the only equilateral triangles with vertices in W with side length 2 containing u_0 . Any other triple chosen from $\{u_0, \dots, u_5\}$ will contain a pair of vertices 2Φ apart. \square

Proposition 7.6.44. Let $v \neq w \in W$. Then $[v, w]$ is an edge of \mathbf{I} if and only if $d(v, w) = 2$. If $d(v, w) > 2$, then $(v, w) \subset \text{Int}(\mathbf{I})$.

Proof. Since $\mathcal{S}(\mathbf{D})$ acts transitively on W , we may assume $v = u_0$. If $d(u_0, w) = 2$, then $w \in \{u_1, u_2, u_3, u_4, u_5\}$. For $2 \leq i \leq 5$, $[u_0, u_i] = G_{i-1} \cap G_i$ and hence is an edge. Similarly, $[u_0, u_1] = G_5 \cap G_0$ is an edge.

If $d(u_0, w) = 2\sqrt{\Phi + 2}$, $w = -u_0$. Since $0 \in [u_0, -u_0]$ and the carrier of 0 is \mathbf{I} , the result follows.

If $d(u_0, w) = 2\Phi$, we rotate by powers of ρ_F until $w = -u_3 = \begin{bmatrix} 1 \\ 0 \\ -\Phi \end{bmatrix}$. But the midpoint of $[u_0, -u_3]$ is e_1 , and it suffices to show the carrier of e_1 is \mathbf{I} .

The midpoint of $[u_1, u_5]$ is Φe_1 , and the midpoint of $[-u_1, -u_5]$ is $-\Phi e_1$. Since both e_1 and 0 are in $(\Phi e_1, -\Phi e_1)$, e_1 and 0 have the same carrier, and the result follows. \square

Corollary 7.6.45. The twenty faces $\{\alpha(G) : \alpha \in \mathcal{S}(\mathbf{D})\}$ are the only faces of \mathbf{I} .

Proof. Let K be a face of \mathbf{I} . Then K has at least 3 vertices, each pair of which have distance 2 from one another. Moving one vertex to u_0 , we see that K must then contain one of the G_i , and hence be equal to it. \square

Corollary 7.6.46. The dual of the icosahedron \mathbf{I} is a rescaling of the dodecahedron \mathbf{D} . So $\mathcal{S}(\mathbf{I}) \subset \mathcal{S}(\mathbf{D})$. Since the reverse inclusion also holds, the two symmetry groups are equal.

Proof. By (7.6.48), the centroid of G is $\frac{\Phi+1}{3}v_0$, where v_0 is a vertex of \mathbf{D} . So for $\alpha \in \mathcal{S}(\mathbf{D})$, the centroid of $\alpha(G)$ is $\frac{\Phi+1}{3}\alpha(v_0)$. Since $\mathcal{S}(\mathbf{D})$ acts transitively on the vertices of \mathbf{D} , the dual of \mathbf{I} is the rescaling of \mathbf{D} by the factor $\frac{\Phi+1}{3}$. \square

7.7. Exercises.

1. Show that each face of the dodecahedron \mathbf{D} meets the cube \mathbf{C} in an edge and that the passage from a face F' of \mathbf{D} to $F' \cap \mathbf{C}$ gives a one-to-one correspondence from the faces of \mathbf{D} to the edges of \mathbf{C} .

8. Spheres and other manifolds

We now wish to generalize our study of Euclidean geometry to geometry in more general settings. The general setting in which isometries are best studied is that of Riemannian geometry. Everything is still based on the inner product, but the inner product is allowed to vary from point to point. The basic method is to study inner products of tangent vectors to curves. The inner product used will depend on the point on the curve at which that tangent is taken. That is the appropriate setting, for instance, to study the classical “non-Euclidean geometry” realized by hyperbolic space.

We will begin with the conceptually simpler case of spherical geometry. The simplest example is the 2-sphere

$$\mathbb{S}^2 = \{x \in \mathbb{R}^3 : \|x\| = 1\}.$$

This is the set of all unit vectors in \mathbb{R}^3 , and forms a model for the surface of the earth. This model is used for computing shortest flight paths between two cities. The “great circle routes” are what is used, and they come directly out of the geometry we develop here.

More generally, the $(n - 1)$ -sphere is the set of unit vectors in \mathbb{R}^n :

$$\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}.$$

In this chapter, we will show that \mathbb{S}^{n-1} is what’s called a smooth submanifold of \mathbb{R}^n of codimension 1.

8.1. Some advanced calculus. Advanced calculus is the foundation for the theory of smooth manifolds.

Recall that if $U \subset \mathbb{R}^n$ is open and if $f : U \rightarrow \mathbb{R}^m$, we write

$$f(x) = (f_1(x), \dots, f_m(x)) \in \mathbb{R}^m$$

and call f_i the i th coordinate function of f . Explicitly, $f_i = \pi_i \circ f$ where $\pi_i : \mathbb{R}^m \rightarrow \mathbb{R}$ is the projection onto the i th coordinate:

$$\pi_i((x_1, \dots, x_m)) = x_i.$$

The partial derivatives of f are defined by setting the partial of f_i with respect to x_j to be

$$(8.1.1) \quad \frac{\partial f_i}{\partial x_j}(x) = \frac{d}{dt}(f_i(x + te_j))|_{t=0}$$

whenever this derivative exists. Here, e_j is the j th canonical basis vector. Explicitly, $\frac{\partial f_i}{\partial x_j}(x)$ is the derivative at 0 of $f_i \circ \iota_j(x)$, where $\iota_j(x) : (-\epsilon, \epsilon) \rightarrow U$ is given by $\iota_j(x)(t) = x + te_j$. Since U is open in \mathbb{R}^n , this is defined for ϵ sufficiently small.

Definition 8.1.1. If $\frac{\partial f_i}{\partial x_j}(x)$ is well-defined for all $i = 1, \dots, m$ and $j = 1, \dots, n$, we say f is differentiable at x and we define the Jacobian matrix

of f at x by

$$Df(x) = \left(\frac{\partial f_i}{\partial x_j}(x) \right).$$

If f is differentiable at each $x \in U$ we say f is differentiable.

Let's now briefly review the idea of continuity. For simplicity, we will restrict attention to subspaces of Euclidean space.

Definition 8.1.2. Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$. We say that $f : X \rightarrow Y$ is continuous if for each $x_0 \in X$ and each $\epsilon > 0$ there exists $\delta > 0$ such that

$$\|x - x_0\| < \delta \quad \Rightarrow \quad \|f(x) - f(x_0)\| < \epsilon.$$

The following is immediate from the definition of continuity.

Lemma 8.1.3. *Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ and let $f : X \rightarrow Y$. Then f is continuous if and only if the composite $f : X \rightarrow Y \subset \mathbb{R}^m$ is continuous. Moreover, if $X \subset \hat{X}$ and $\hat{f} : \hat{X} \rightarrow Y$ is continuous with $\hat{f}|_X = f$, then f is continuous.*

The above is useful by the following basic result from the calculus of several variables.

Lemma 8.1.4. *Let $U \subset \mathbb{R}^n$ and let $f : U \rightarrow \mathbb{R}^m$ be differentiable. Then f is continuous.*

We can now discuss higher differentiability.

Definition 8.1.5. We say f is C^1 , or continuously differentiable, if the function $\frac{\partial f_i}{\partial x_j} : U \rightarrow \mathbb{R}$ is well-defined and continuous for all $i = 1, \dots, m$ and $j = 1, \dots, n$. By Proposition A.6.13, this is equivalent to saying the Jacobian matrix

$$Df : U \rightarrow M_{m,n}(\mathbb{R})$$

is continuous where $M_{m,n}(\mathbb{R})$ is the space of $m \times n$ matrices with coefficients in \mathbb{R} , which we identify with \mathbb{R}^{mn} in the usual way.

If each $\frac{\partial f_i}{\partial x_j} : U \rightarrow \mathbb{R}$ is itself C^1 , we say f is C^2 . Inductively, if each $\frac{\partial f_i}{\partial x_j} : U \rightarrow \mathbb{R}$ is C^r for some $r \geq 1$ we say f is C^{r+1} . In other words, all iterated partials of length $r + 1$ of all coordinate functions f_i of f are well-defined and continuous.

If f is C^r for all $r \geq 1$, we say f is C^∞ or smooth. It is also sometimes customary to write C^0 for a continuous function.

The Jacobian matrix is indeed the higher dimensional analogue of the derivative of a real valued function of one variable, and provides the best "linear" approximation at x to the function f . Specifically, f is best approximated near x_0 by the affine function

$$f(x_0) + Df(x_0) \cdot (x - x_0),$$

where the \cdot represents the matrix product of $Df(x_0)$ with the column vector $x - x_0$.

This approximation is used in the calculus of several variables to deduce properties of f from properties of Df . This, of course, is analogous to the first derivative test and mean value theorem for studying functions of a real variable.

In order for Df to be the best linear approximation to f , we must have the following.

Lemma 8.1.6. *Let $A = (a_{ij})$ be an $m \times n$ matrix and let $T_A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be the induced linear function. Then $DT_A(x) = A$ for all $x \in \mathbb{R}^n$.*

Proof. Write $A_i = [a_{i1} \ \dots \ a_{in}]$ for the i th row of A . Then the i th coordinate function of T_A is T_{A_i} . Now, $\frac{\partial T_{A_i}}{\partial x_j}$ is the derivative at 0 of the function $\gamma : \mathbb{R} \rightarrow \mathbb{R}$ given by

$$\gamma(t) = T_{A_i}(x + te_j) = T_{A_i}(x) + tT_{A_i}(e_j) = T_{A_i}(x) + ta_{ij}$$

by linearity and direct matrix multiplication. But $\gamma'(0) = a_{ij}$ and the result follows. \square

And now for some more examples of Jacobian matrices:

Example 8.1.7. A C^r curve in \mathbb{R}^n is a C^r map $\gamma : (a, b) \rightarrow \mathbb{R}^n$ with (a, b) an open interval in \mathbb{R} . The Jacobian matrix $D\gamma$ is given by

$$D\gamma(t) = \begin{bmatrix} \gamma'_1(t) \\ \vdots \\ \gamma'_n(t) \end{bmatrix},$$

where γ_i is the i th coordinate function of γ . We write $\gamma'(t)$ as a shorthand for $D\gamma(t)$.

An important example is the exponential map $\exp : \mathbb{R} \rightarrow \mathbb{R}^2$ given by $\exp(t) = \begin{bmatrix} \cos t \\ \sin t \end{bmatrix}$. We have

$$D \exp(t) = \begin{bmatrix} -\sin t \\ \cos t \end{bmatrix} = \exp(t)^\perp.$$

Note that both \exp and $D \exp$ take value in the unit circle \mathbb{S}^1 .

Example 8.1.8. Let $U \subset \mathbb{R}^n$ and let $f : U \rightarrow \mathbb{R}$ be C^r . Then

$$Df(x) = \left[\frac{\partial f}{\partial x_1}(x) \ \dots \ \frac{\partial f}{\partial x_n}(x) \right] = \nabla f(x),$$

the gradient of f at x . As in the one-variable case, this is important in finding local maxima and minima of f , which may occur only at critical points: points where $\nabla f(x) = 0$ (or undefined, if f is not C^1 everywhere).

An interesting example is given by $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with $f(x) = \langle x, x \rangle$ for all $x \in \mathbb{R}^n$. Here,

$$\begin{aligned} \frac{\partial}{\partial x_i}(x) &= \frac{d}{dt}(\langle x + te_i, x + te_i \rangle)|_{t=0} \\ &= \frac{d}{dt}(t^2 \langle e_i, e_i \rangle + 2t \langle e_i, x \rangle + \langle x, x \rangle)|_{t=0} \\ &= (2t + 2x_i)|_{t=0} = 2x_i. \end{aligned}$$

Thus,

$$\nabla f(x) = 2x$$

if we regard x as a row vector. So 0 is the only critical point of f , and gives the absolute minimum. Note that $f^{-1}(1) = \mathbb{S}^{n-1}$, the unit sphere in \mathbb{R}^n .

Another easy calculation is the following.

Lemma 8.1.9. *Let $y \in \mathbb{R}^n$ and let τ_y be the translation by y : $\tau_y(x) = x + y$ for all $x \in \mathbb{R}^n$. Then*

$$D\tau_y(x) = I_n$$

for every $x \in \mathbb{R}^n$.

Proof. The i th coordinate function $(\tau_y)_i$ is given by

$$(\tau_y)_i(x) = y_i + x_i,$$

where y_i and x_i are the i th coordinates of the vectors y and x , respectively. Thus

$$(\tau_y)_i(x + te_j) = \begin{cases} y_i + x_i & \text{if } i \neq j \\ y_i + x_i + t & \text{if } i = j. \end{cases}$$

Thus,

$$\begin{aligned} \frac{\partial (\tau_y)_i}{\partial x_j}(x) &= \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \\ &= \delta_{ij}. \end{aligned} \quad \square$$

A basic result from the calculus of several variables is:

Proposition 8.1.10 (Chain rule). *Let $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$ be open. Let $f : U \rightarrow V$ and $g : V \rightarrow \mathbb{R}^k$ be C^r , $1 \leq r \leq \infty$. Then $g \circ f$ is C^r with Jacobian matrix given by the matrix product of the differentials of g and f as follows:*

$$D(g \circ f)(x) = Dg(f(x))Df(x).$$

This allows us to calculate the Jacobian matrix of any isometry of \mathbb{R}^n . Recall that any isometry $\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^n$ may be written as a composite $\alpha = \tau_y \circ \beta$ where β is a linear isometry of \mathbb{R}^n . β , in turn, may be written as T_A where A is an $n \times n$ orthogonal matrix.

Corollary 8.1.11. *Let $\alpha = \tau_y \circ T_A$ be an isometry of \mathbb{R}^n with A an $n \times n$ orthogonal matrix. Then $D\alpha(x) = A$ for all $x \in \mathbb{R}^n$.*

The following concept is useful both for orientation theory and for developing the theory of smooth manifolds.

Definition 8.1.12. Let $U \subset \mathbb{R}^n$ open. A C^r map $f : U \rightarrow \mathbb{R}^m$, $1 \leq r \leq \infty$, is an immersion at $x \in U$ if the columns of $Df(x)$ are linearly independent. (In particular, if $m = n$, then $Df(x)$ is invertible.) f itself is an immersion if it is an immersion at every $x \in U$.

Example 8.1.13. The exponential map $\exp : \mathbb{R} \rightarrow \mathbb{R}^2$ is an immersion as $D \exp(x) \neq 0$ for all $x \in \mathbb{R}$.

A fundamental property of C^r immersions $\mathbb{R}^n \rightarrow \mathbb{R}^n$ is given by the inverse function theorem.

Theorem 8.1.14 (Inverse function theorem). *Let $U \subset \mathbb{R}^n$ be open and let $f : U \rightarrow \mathbb{R}^n$ be C^r , $1 \leq r \leq \infty$. Then $Df(x)$ is invertible if and only if there are open subsets $V \subset U$ and $W \subset \mathbb{R}^n$ with $x \in V$ such that $f|_V : V \rightarrow W$ is bijective and $f^{-1} : W \rightarrow V$ is C^r . Moreover,*

$$Df^{-1}(f(y)) = (Df(y))^{-1}$$

for all $y \in V$.

Proof. The “only if” part is one of the fundamental results of advanced calculus. See [12, Theorem I.5.2] for a proof. The “if” part is easy, and follows from the chain rule: if there are open subsets $V \subset U$ and $W \subset \mathbb{R}^n$ with $x \in V$ such that $f|_V : V \rightarrow W$ is bijective and $f^{-1} : W \rightarrow V$ is C^r , then

$$I_n = DI(x) = Df^{-1}(f(x))Df(x)$$

and

$$I_n = DI(f(x)) = Df(x)Df^{-1}(f(x)),$$

as $f^{-1}(f(x)) = x$. Here, I is the identity function of \mathbb{R}^n . Thus $Df(x)$ is invertible with inverse $Df^{-1}(f(x))$. \square

Definition 8.1.15. Let U and V be open subsets of \mathbb{R}^n . A C^r -isomorphism $f : U \rightarrow V$ is a bijective C^r map whose inverse function is also C^r . A C^∞ -isomorphism is called a diffeomorphism.

A C^r map $f : U \rightarrow V$ is a local C^r -isomorphism if each $x \in U$ is contained in a smaller open set U' such that $f|_{U'} : U' \rightarrow f(U')$ is a C^r -isomorphism onto an open subset of V . The inverse function theorem may be restated to say that if U is an open subset of \mathbb{R}^n , then a C^r map $f : U \rightarrow \mathbb{R}^n$ is an immersion if and only if it is a local C^r -isomorphism.

Corollary 8.1.16. *Let $U \subset \mathbb{R}^n$ open and let $f : U \rightarrow \mathbb{R}^n$ be a C^r immersion, $1 \leq r \leq \infty$. Then $f(U) \subset \mathbb{R}^n$ is open. If f is also one-to-one, then the inverse function $f^{-1} : f(U) \rightarrow U$ is also C^r , so $f : U \rightarrow f(U)$ is a C^r -isomorphism. When $r = \infty$, this says that if U is open in \mathbb{R}^n , then a one-to-one smooth immersion $f : U \rightarrow \mathbb{R}^n$ has open image and the map $f : U \rightarrow f(U)$ is a diffeomorphism.*

Proof. That $f(U)$ is open follows from the open sets W about each point in $f(U)$ obtained from the inverse function theorem. When f is one-to-one, the inverse function theorem also shows f^{-1} is C^r on each such W , and therefore is C^r on all of $f(U)$. \square

8.2. Orientation properties of nonlinear mappings in \mathbb{R}^n . We regard \mathbb{R}^n to have a natural orientation coming from the standard ordering of the standard basis. We can think of open subsets $U \subset \mathbb{R}^n$ as inheriting this orientation and ask when a C^r map preserves or reverses this orientation. Since a dimension-reducing linear map neither preserves nor reverses orientation, we shall ask this question only for immersions. This intuition may be fleshed out as follows:

Definition 8.2.1. Let $U \subset \mathbb{R}^n$ open and let $f : U \rightarrow \mathbb{R}^n$ be a C^1 immersion. Then f is orientation-preserving if $Df(x)$ has positive determinant for all $x \in U$.

We say f is orientation-reversing if $Df(x)$ has negative determinant for all $x \in U$.

This is, of course, consistent with the linear case by Lemma 8.1.6, and is also consistent with the definition we gave for isometries of the plane by Corollary 8.1.11. Indeed, Corollary 8.1.11 allows us to determine the orientation properties of isometries of \mathbb{R}^n . Since every linear isometry is induced by an orthogonal matrix, and since every orthogonal matrix has determinant ± 1 , the following is immediate.

Corollary 8.2.2. Let $\alpha \in \mathcal{I}_n$ and write $\alpha = \tau_y \circ T_A$ for A an $n \times n$ orthogonal matrix. Then α is orientation-preserving if $\det A = 1$ (i.e., if $A \in \text{SO}(n)$), and is orientation-reversing if $\det A = -1$. Recalling from (3.3.4) that $(\tau_y T_A) \circ (\tau_z T_B) = \tau_{y+Az} T_{AB}$ (and hence $(\tau_y T_A)^{-1} = \tau_{A^{-1}y} T_{A^{-1}}$), we see that the orientation-preserving isometries of \mathbb{R}^n form a subgroup, $\mathcal{O}_n \subset \mathcal{I}_n$.

Unlike the linear case, if U is not path-connected, it is possible for a C^1 immersion $f : U \rightarrow \mathbb{R}^n$ to be neither orientation-preserving nor orientation-reversing.

Example 8.2.3. Define $f : \mathbb{R} - \{0\} \rightarrow \mathbb{R}$ by $f(x) = x^2$ for all x in the domain. Then $Df(x) = [2x]$, a 1×1 matrix with entry $f'(x)$. Of course, $\det Df(x) = 2x$, so f is a smooth immersion, and is orientation-preserving on $(0, \infty)$ and orientation-reversing on $(-\infty, 0)$.

In some cases, it is possible to test for orientation properties at a single point. To see this we review some determinant theory. See [17] for the details.

Recall the sign homomorphism $\text{sgn} : \Sigma_n \rightarrow \{\pm 1\}$ of Definition 4.1.19. The determinant satisfies the following formula ([17, Corollary 10.2.6]).

Lemma 8.2.4. *Let $A = (a_{ij})$ be $n \times n$. Then*

$$(8.2.1) \quad \det A = \sum_{\sigma \in \Sigma_n} \operatorname{sgn}(\sigma) a_{\sigma(1)1} \cdots a_{\sigma(n)n}.$$

In particular, \det is a polynomial in the n^2 variables giving the coordinates of the matrix, and hence gives a smooth function

$$\det : M_n(\mathbb{R}) \rightarrow \mathbb{R}.$$

Here $M_n(\mathbb{R})$ denotes the space of $n \times n$ matrices with coefficients in \mathbb{R} , which we identify with \mathbb{R}^{n^2} .

As a bonus, we obtain the following. Recall that $\operatorname{GL}_n(\mathbb{R})$ is the group of invertible $n \times n$ matrices over \mathbb{R} and that a matrix A is invertible if and only if $\det A \neq 0$.

Corollary 8.2.5. $\operatorname{GL}_n(\mathbb{R}) = \det^{-1}(\mathbb{R} - \{0\})$ is an open subset of $M_n(\mathbb{R})$.

Proof. Differentiable maps are continuous. $\mathbb{R} - \{0\} = (-\infty, 0) \cup (0, \infty)$ is an open subset of \mathbb{R} . So $\det^{-1}(\mathbb{R} - \{0\})$ is open by Lemma A.1.12 below. \square

Definition 8.2.6. A subset $X \subset \mathbb{R}^n$ is path-connected if for each $x, y \in X$ there is a continuous map $\gamma : [0, 1] \rightarrow X$ with $\gamma(0) = x$ and $\gamma(1) = y$. Such a γ is called a path from x to y .

Convex sets are certainly path-connected, so we have plenty of examples.

Proposition 8.2.7. *Let $U \subset \mathbb{R}^n$ be open and path-connected and let $f : U \rightarrow \mathbb{R}^n$ be a C^1 immersion. Let $x \in U$. Then f is orientation-preserving if $Df(x)$ has positive determinant and is orientation-reversing if $Df(x)$ has negative determinant.*

Proof. In other words, we claim the sign of $\det Df$ is constant on U when U is path-connected and f is a C^1 immersion. To see this, note that f being C^1 says $Df : U \rightarrow M_n(\mathbb{R})$ is continuous, hence so is

$$\det \circ Df : U \rightarrow \mathbb{R}.$$

Let γ be a path from x to y . Then the composite

$$\det \circ Df \circ \gamma : [0, 1] \rightarrow \mathbb{R}$$

is continuous. Since f is an immersion, $\det(Df)$ is never zero. By the intermediate value theorem, the sign of $\det(Df(\gamma(t)))$ is constant. \square

8.3. Topological manifolds; \mathbb{S}^{n-1} . The basic objects of study in geometric topology are manifolds. We shall show that the unit sphere $\mathbb{S}^{n-1} \subset \mathbb{R}^n$ is an $(n - 1)$ -dimensional manifold.

The following is a basic topological concept.

Definition 8.3.1. Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$. A map $f : X \rightarrow Y$ is a homeomorphism if it is continuous, one-to-one and onto, and the inverse function $f^{-1} : Y \rightarrow X$ is also continuous. We write

$$f : X \xrightarrow{\cong} Y$$

for a homeomorphism and trust it will not be confused with an isomorphism of vector spaces.

Example 8.3.2. Define $f : \mathbb{R} \rightarrow \mathbb{R}$ by $f(x) = x^3$. Then f is a homeomorphism as the inverse function $f^{-1} : \mathbb{R} \rightarrow \mathbb{R}$ is given by $f^{-1}(x) = \sqrt[3]{x}$, a continuous function. Note that f^{-1} is not differentiable at 0 because $f'(0) = 0$.

Continuous bijections exist that are not homeomorphisms (because their inverse functions are not continuous). See Example 10.4.5 below.

Definition 8.3.3. Let $X \subset \mathbb{R}^m$. A subset $V \subset X$ is open in X if there is an open subset U of \mathbb{R}^m with $U \cap X = V$.

If $x \in X$ a neighborhood of x in X is simply an open set in X containing x .

Definition 8.3.4. A subset $M \subset \mathbb{R}^m$ is a topological n -manifold, or n -dimensional manifold, if for each $x \in M$ there is a neighborhood U of x in M and a homeomorphism $h : U \xrightarrow{\cong} V$ where V is an open subset of \mathbb{R}^n . The maps $h : U \xrightarrow{\cong} V$ (or their inverses, depending on one's convention) are called charts for M (and if $x \in U$, we call it a chart about x). If $h : U \xrightarrow{\cong} V$ is a chart, we call U a chart neighborhood (of each $x \in U$) in M .

Note that any open subset of a chart neighborhood is also a chart neighborhood by restriction.

We sometimes write M^n for M to emphasize that M has dimension n . The reader is warned not to confuse this with the cartesian product of n copies of M .

A 2-dimensional manifold is called a surface.

The following is implicit in the above definition, but is probably not intuitive to a beginner.

Remark 8.3.5. \mathbb{R}^0 is the 0-vector space, consisting of a single point. Its only open subspaces are \emptyset and itself. So a subset $M \subset \mathbb{R}^m$ is a 0-manifold if each point of M is a neighborhood of itself, i.e., if for each $x \in M$ there is an open set U of \mathbb{R}^m with $U \cap M = \{x\}$. Topologically, this says each point of M is open in the subspace topology inherited from \mathbb{R}^m . This says the subspace topology is what's known as the discrete topology.

Topological manifolds have many nice properties, but they are hard to verify, and the intuition in uncovering those properties came from the more obvious properties of smooth manifolds. We will show that the unit sphere \mathbb{S}^{n-1} of \mathbb{R}^n is a topological manifold and use it as the motivating example

in defining smooth manifolds. Consider e_n , the last of the canonical basis vectors of \mathbb{R}^n , and think of it as the north pole N of \mathbb{S}^{n-1} . The following shows that the complement of a point in \mathbb{S}^{n-1} is homeomorphic to \mathbb{R}^{n-1} .

Theorem 8.3.6. *Let $n \geq 2$ and identify \mathbb{R}^{n-1} with the equatorial $(n-1)$ -plane in \mathbb{R}^n :*

$$\mathbb{R}^{n-1} = \{(x_1, \dots, x_{n-1}, 0) : x_1, \dots, x_{n-1} \in \mathbb{R}\}.$$

Let $U = \mathbb{S}^{n-1} \setminus \{N\}$, the complement of the north pole in \mathbb{S}^{n-1} . Then there is a homeomorphism $h_U : U \rightarrow \mathbb{R}^{n-1}$ given by

$$(8.3.1) \quad h_U(x) = \frac{1}{1-x_n}(x_1, \dots, x_{n-1}, 0)$$

for $x = (x_1, \dots, x_n)$. The inverse function g_U of h_U is given by

$$(8.3.2) \quad g_U(x) = tx + (1-t)N \quad \text{for } t = \frac{2}{1+\langle x, x \rangle}$$

for $x = (x_1, \dots, x_{n-1}, 0)$.

Proof. We have defined h_U to take x to the unique point on the ray \overrightarrow{Nx} lying in \mathbb{R}^{n-1} : $h_U(x) = (1-t)N + tx$ for the unique t such that $1-t+tx_n = 0$.

g_U , in turn, takes x to the unique point of norm 1 on \overrightarrow{Nx} other than N .

h_U and g_U both extend to C^1 (in fact C^∞ functions) defined on open sets containing U and \mathbb{R}^{n-1} , respectively, and hence are continuous. Note that if $g_U(x) = (y_1, \dots, y_n)$, then $y_n = 1-t$ and $(y_1, \dots, y_{n-1}, 0) = tx$. Thus, $\frac{1}{1-y_n} = \frac{1}{t}$, so

$$h_U \circ g_U(x) = \frac{1}{t}(tx_1, \dots, tx_{n-1}, 0) = x.$$

Similarly, since $h_U(x) = \frac{1}{1-x_n}(x_1, \dots, x_{n-1}, 0)$, we have

$$\begin{aligned} \langle h_U(x), h_U(x) \rangle &= \left(\frac{1}{1-x_n} \right)^2 \sum_{i=1}^{n-1} x_i^2 \\ &= \left(\frac{1}{1-x_n} \right)^2 (1-x_n^2) \\ &= \frac{1+x_n}{1-x_n}. \end{aligned}$$

But then it's easy to see that $\frac{2}{1+\langle h_U(x), h_U(x) \rangle} = 1-x_n$, and that in turn shows $g_U \circ h_U = \text{id}$. \square

The map h_U is called the stereographic projection of U onto \mathbb{R}^{n-1} . The following may now be used to motivate the definition of smooth manifold and show that \mathbb{S}^{n-1} is one.

Corollary 8.3.7. \mathbb{S}^{n-1} is an $(n-1)$ -manifold for $n \geq 2$. Specifically, let $V = \mathbb{S}^{n-1} \setminus \{S\}$ where $S = -e_n$ is the south pole. Then there is a homeomorphism $h_V : V \rightarrow \mathbb{R}^{n-1}$ given by

$$h_V(x) = \frac{1}{1+x_n}(x_1, \dots, x_{n-1}, 0)$$

in the notations of Theorem 8.3.6. Since every point of \mathbb{S}^{n-1} lies in either U or V , this suffices to show \mathbb{S}^{n-1} is a manifold.

Note that $h_U(S) = 0 = h_V(N)$, and hence

$$h_U(U \cap V) = h_V(U \cap V) = \mathbb{R}^n \setminus \{0\}.$$

The composites

$$h_V \circ h_U^{-1} : \mathbb{R}^n \setminus \{0\} \rightarrow \mathbb{R}^n \setminus \{0\},$$

$$h_U \circ h_V^{-1} : \mathbb{R}^n \setminus \{0\} \rightarrow \mathbb{R}^n \setminus \{0\},$$

are both given by the same formula:

$$(8.3.3) \quad h_V \circ h_U^{-1}(x) = h_U \circ h_V^{-1}(x) = \frac{1}{\langle x, x \rangle} x.$$

Here, $h_U(U \cap V) = h_V(U \cap V) = \mathbb{R}^{n-1} - 0$.

Proof. We simply use the south pole in place of the north in the arguments for Theorem 8.3.6. Thus, we set $h_V(x)$ equal to the unique point on the ray \overrightarrow{Sx} lying in \mathbb{R}^{n-1} and set $g_U(x)$ equal to the unique point of norm 1 on \overrightarrow{Sx} other than S .

The formula for h_V is then clear. As above, $g_U(x) = (tx_1, \dots, tx_{n-1}, 1-t)$ for $t = \frac{2}{1+\langle x, x \rangle}$, hence

$$h_V \circ g_U(x) = \frac{t}{2-t} x.$$

The result then follows by calculating $\frac{t}{2-t}$. The same calculation works in the opposite direction. \square

In particular, \mathbb{S}^2 is a surface and \mathbb{S}^1 is a 1-manifold (the latter can also be shown using the exponential map).

Note that we have not yet considered the 0-sphere, $\mathbb{S}^0 = \{\pm 1\} \subset \mathbb{R}$, the unit sphere in \mathbb{R} . Since $\{-1\} = \mathbb{S}^0 \cap (-2, 0)$ and $\{1\} = \mathbb{S}^0 \cap (0, 2)$, \mathbb{S}^0 is a 0-manifold. We obtain:

Corollary 8.3.8. \mathbb{S}^n is an n -manifold for $n \geq 0$.

8.4. Smooth manifolds.

Definition 8.4.1. A smooth atlas \mathcal{A} on a topological n -manifold M is a family of charts

$$h : U \xrightarrow{\cong} h(U) \subset \mathbb{R}^n \quad h(U) \text{ open in } \mathbb{R}^n$$

such that:

The property of being a smooth map is a local one.

Lemma 8.4.4. *Let M and N be smooth manifolds and let $f : M \rightarrow N$. Then f is smooth if and only if for each $x \in M$ there exist charts h and k such that (8.4.2) is smooth on some smaller open set containing $h(x)$. Moreover, the rank of the Jacobian matrix Df_{kh} is independent of the choice of h and k .*

Proof. If we replace h and k by charts h' and k' about x and $f(x)$, respectively, then

$$f_{k'h'} = g_{k'k} \circ f_{hk} \circ g_{hh'}$$

near x . The result now follows from the chain rule. \square

Lemma 8.4.5. *If $f : M \rightarrow N$ is a diffeomorphism from an n -manifold to an m -manifold, then $n = m$.*

Proof. Suppose $U \cap f^{-1}(V) \neq \emptyset$. Since both f and f^{-1} are smooth, one can restrict the codomain of the maps f_{kh} of (8.4.2) to obtain a diffeomorphism

$$f_{kh} : h(U \cap f^{-1}(V)) \rightarrow k(V \cap f(U))$$

with inverse

$$f_{hk} : k(V \cap f(U)) \rightarrow h(U \cap f^{-1}(V)).$$

In particular, we may assume M an open subset of \mathbb{R}^n and N is an open subset of \mathbb{R}^m . Now

$$Df^{-1}(f(x))Df(x) = I_n \quad \text{and} \quad Df(x)Df^{-1}(f(x)) = I_m,$$

giving an isomorphism of vector spaces between \mathbb{R}^n and \mathbb{R}^m , so $n = m$. \square

Remark 8.4.6. The analogous result for homeomorphisms of topological manifolds is also true, as a consequence of invariance of domain [8, Theorem XVII.3.1]. Specifically, homeomorphic topological manifolds must have the same dimension.

Example 8.4.7. Define $f : \mathbb{R} \rightarrow \mathbb{R}$ by $f(x) = x^3$. Then f is a homeomorphism and a smooth bijection, but not a diffeomorphism as f^{-1} is not smooth at 0. While f is not a diffeomorphism, the domain and codomain are diffeomorphic. We could use f to define an atlas on \mathbb{R} giving a smooth structure different from the standard one, but diffeomorphic to the standard one via f .

Remarks 8.4.8.

- (1) John Milnor first showed there are smooth structures on a sphere, S^7 in the first examples, that are not diffeomorphic to the standard smooth structure. Kervaire and Milnor then classified all such structures on spheres of dimension ≥ 6 , laying the groundwork for a theory called surgery theory. Surgery theory may be used to classify the smooth structures on a topological manifold.

- (2) Kirby and Freedman, using work of Donaldson, showed there are smooth structures on \mathbb{R}^4 not diffeomorphic to the standard smooth structure. Note, then, that whole of \mathbb{R}^4 cannot be the domain of a chart for such a structure.

Definition 8.4.9. Let M be a topological manifold. We say two smooth atlases \mathcal{A} and \mathcal{B} on M are equivalent (and represent the same smooth structure on M) if the identity map of M gives a diffeomorphism between the two atlases. In particular, then, either atlas may be used to describe the smooth maps in or out of M .

If $U \subset M$ is open, it is a smooth manifold with an atlas given by the restriction to U of the charts of M , so we can talk about diffeomorphisms of U . If $h : U \rightarrow \mathbb{R}^n$ is such that $h : U \rightarrow h(U)$ is a diffeomorphism, we can add h to the atlas of M without changing any of the constructions we shall make below. We can also remove a chart if its domain is covered by other charts in the atlas. We will allow ourselves this flexibility without further discussion. In particular, if we say the smooth manifold M has an atlas with a particular property, we mean an atlas equivalent to the original one.¹⁴

The following concepts are useful.

Definition 8.4.10. An open cover \mathcal{U} of a space X is a set of open subsets of X such that

$$X = \bigcup_{U \in \mathcal{U}} U.$$

An open cover \mathcal{V} is a refinement of \mathcal{U} if each $V \in \mathcal{V}$ is contained in some $U \in \mathcal{U}$.

The following may be obtained simply by restricting the domains of charts.

Lemma 8.4.11. Let \mathcal{U} be any open cover of the smooth manifold M . Then M has an atlas \mathcal{A} that refines \mathcal{U} .

Remark 8.4.12. As shown in Exercises 2 or 3 below, every point in \mathbb{R}^n has arbitrary small neighborhoods diffeomorphic to \mathbb{R}^n . We can use these to alter any given atlas so that the charts actually give diffeomorphisms from their domains onto all of \mathbb{R}^n .

¹⁴One approach to defining the smooth structure on M is to insist that every atlas be “maximal” in the sense that every smooth embedding $h : U \rightarrow \mathbb{R}^n$ of an open subspace of M into \mathbb{R}^n be part of the atlas. This is a huge amount of redundancy in that we have an overabundance of open sets U , and for each such U and uncountable number of smooth embeddings $h : U \rightarrow \mathbb{R}^n$ represented in the atlas. The reason one might do this is that if \mathcal{A} and \mathcal{B} are maximal atlases on M and if the identity map gives a diffeomorphism between these atlases, then the atlases \mathcal{A} and \mathcal{B} are in fact equal. But that seems much too high a price to pay for uniqueness. In fact, we personally prefer atlases that are locally finite when possible, meaning that each point has a neighborhood that intersects only finitely many of the open sets in the atlas.

As shown in the Exercise 1 below, the displayed smooth structure on \mathbb{S}^{n-1} coincides with its structure as a submanifold of \mathbb{R}^n in the following sense:

Definition 8.4.13. Let N be a smooth n -manifold. A smooth m -submanifold M of N is a subset $M \subset N$ such that for each $x \in M$ there is a smooth chart $h : U \rightarrow \mathbb{R}^n = \mathbb{R}^m \times \mathbb{R}^k$ for N about x such that $U \cap M = h^{-1}(\mathbb{R}^m \times 0)$. The maps $h|_{U \cap M} : U \cap M \rightarrow \mathbb{R}^m$ then assemble to give a smooth atlas for M and the inclusion map $M \subset N$ has Jacobian matrices of rank m at every point in M .

The following is useful.

Proposition 8.4.14. Let M be a smooth m -submanifold of the smooth n -manifold N . Let S be an s -submanifold of the r -manifold R . Let $f : N \rightarrow R$ be a smooth map with $f(M) \subset S$. Then $f|_M : M \rightarrow S$ is smooth.

In the case $M = N$, if $f : M \rightarrow S \subset R$, then $f : M \rightarrow S$ is smooth if and only if $f : M \rightarrow R$ is smooth.

Proof. If $W \subset \mathbb{R}^n$ is open and $g : W \rightarrow \mathbb{R}^r$ is smooth and if

$$g(W \cap (\mathbb{R}^m \times 0)) \subset \mathbb{R}^s \times 0,$$

then $g|_{W \cap (\mathbb{R}^m \times 0)} : W \cap (\mathbb{R}^m \times 0) \rightarrow \mathbb{R}^s$ is smooth. In fact, its Jacobian matrix is a submatrix of the Jacobian matrix of g .

In the case $M = N$ and $f : M \rightarrow S \subset R$, the above argument shows that $f : M \rightarrow S$ is smooth if $f : M \rightarrow R$ is. But the converse follows by composite, as $S \subset R$ is smooth. \square

A vitally important application is the following.

Corollary 8.4.15. Let $\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a linear isometry. Then α restricts to a diffeomorphism $\alpha|_{\mathbb{S}^{n-1}} : \mathbb{S}^{n-1} \rightarrow \mathbb{S}^{n-1}$.

Proof. Since \mathbb{S}^{n-1} is the set of points of distance 1 from the origin, and since α preserves 0 and also preserves distance, $\alpha(\mathbb{S}^{n-1}) \subset \mathbb{S}^{n-1}$. But the same is true of α^{-1} , so $\alpha|_{\mathbb{S}^{n-1}} : \mathbb{S}^{n-1} \rightarrow \mathbb{S}^{n-1}$ is a smooth bijection with inverse $\alpha^{-1}|_{\mathbb{S}^{n-1}} : \mathbb{S}^{n-1} \rightarrow \mathbb{S}^{n-1}$. \square

In fact, every smooth manifold is a submanifold of some Euclidean space.

Theorem 8.4.16 (Whitney Embedding Theorem). Every smooth n -manifold is diffeomorphic to a smooth submanifold of \mathbb{R}^{2n} .

Of course, some n -manifolds are smooth submanifolds of much lower-dimensional Euclidean spaces. For instance \mathbb{S}^n is a smooth submanifold of \mathbb{R}^{n+1} . In fact, \mathbb{S}^n is what's called a regular hypersurface in \mathbb{R}^{n+1} , which gives it some important extra properties. We will now flesh out this notion.

Definition 8.4.17. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be smooth, with $n \geq m$. A point $y \in \mathbb{R}^m$ is a regular value of f if $Df(x)$ has rank m for each $x \in f^{-1}(y)$.

As shown in Corollary 10.4.10 below, this makes $M = f^{-1}(y)$ a smooth submanifold of \mathbb{R}^n of dimension $n - m$. We call it a regular submanifold of

\mathbb{R}^n . If $m = 1$, M is a codimension 1 submanifold of \mathbb{R}^n , otherwise known as a hypersurface in \mathbb{R}^n . When $n = 3$, then M is a surface, and hence a regular surface in \mathbb{R}^3 .

Example 8.4.18. \mathbb{S}^n is a regular hypersurface in \mathbb{R}^{n+1} , as $\mathbb{S}^n = f^{-1}(1)$, where $f : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ is given by $f(x) = \langle x, x \rangle$. As shown in Example 8.1.8,

$$(8.4.3) \quad Df(x) = \nabla f(x) = 2x$$

for all $x \in \mathbb{R}^{n+1}$, so 1 is a regular value of f . By Exercise 1 below, this smooth structure on \mathbb{S}^n agrees with the one given in Corollary 8.3.7 via stereographic projection.

We shall use this to obtain more information about the sphere after developing the notion of the tangent space of a manifold. We shall see (Corollary 10.2.12) that the tangent space $T_u(\mathbb{S}^n)$ to \mathbb{S}^n at a point $u \in \mathbb{S}^n$ is the orthogonal complement of $\text{span}(\nabla f(u))$ in \mathbb{R}^{n+1} , which is precisely $\{u\}^\perp$.

8.5. Products of manifolds. Let M and N be smooth manifolds of dimension m and n , respectively. Then smooth charts $h : U \rightarrow \mathbb{R}^m$ for M and $k : V \rightarrow \mathbb{R}^n$ for N combine to give a chart for $M \times N$:

$$h \times k : U \times V \rightarrow \mathbb{R}^m \times \mathbb{R}^n,$$

which we identify with \mathbb{R}^{m+n} in the usual way. The transition maps from $h \times k$ to $h' \times k'$ are then simply $g_{h'h} \times g_{k'k}$, which has Jacobian matrix

$$D(g_{h'h} \times g_{k'k}) = \left[\begin{array}{c|c} Dg_{h'h} & 0 \\ \hline 0 & Dg_{k'k} \end{array} \right].$$

The block sum of two square matrices as above is called the Whitney sum and has determinant $\det(Dg_{h'h}) \det(Dg_{k'k})$. In particular, if each of $g_{h'h}$ and $g_{k'k}$ preserves orientation, so does $g_{h'h} \times g_{k'k}$, but if each of $g_{h'h}$ and $g_{k'k}$ reverses orientation, then $g_{h'h} \times g_{k'k}$ preserves it. $g_{h'h} \times g_{k'k}$ only reverses orientation if $g_{h'h}$ and $g_{k'k}$ have orientation behavior opposite from one another.

We call this the standard smooth structure on $M \times N$. By construction, it satisfies the following.

Proposition 8.5.1. *Let M , N and P be smooth manifolds. Then a map $f : P \rightarrow M \times N$ is smooth if and only if its coordinate functions $f_1 = \pi_1 \circ f$ and $f_2 = \pi_2 \circ f$ are smooth, where $\pi_1 : M \times N \rightarrow M$ and $\pi_2 : M \times N \rightarrow N$ are the projections. In particular, for smooth functions $f_1 : P \rightarrow M$ and $f_2 : P \rightarrow N$ there is a unique smooth function $f = (f_1, f_2)$ such that the following diagram commutes:*

$$\begin{array}{ccccc} & & P & & \\ & f_1 \swarrow & \downarrow f & \searrow f_2 & \\ M & \xleftarrow{\pi_1} & M \times N & \xrightarrow{\pi_2} & N. \end{array}$$

Proof. For $U \subset \mathbb{R}^m$, $V \subset \mathbb{R}^n$, $W \subset \mathbb{R}^k$ and a map

$$f : W \rightarrow U \times V \subset \mathbb{R}^m \times \mathbb{R}^n \cong \mathbb{R}^{m+n},$$

the real valued coordinate functions of f when seen as taking value in \mathbb{R}^{m+n} are the coordinate functions of $f_1 : W \rightarrow U$ and $f_2 : W \rightarrow V$ taken in order, i.e.,

$$Df = \begin{bmatrix} Df_1 \\ Df_2 \end{bmatrix}. \quad \square$$

8.6. Oriented atlases. The following is a key in defining an orientation on a manifold.

Definition 8.6.1. Let M be a smooth manifold. An oriented atlas for M is an atlas \mathcal{A} such that the transition maps g_{kl} of (8.4.1) are orientation-preserving (i.e., their Jacobian matrices have positive determinant at every point).

Example 8.6.2. For $n \geq 1$, the atlas given for \mathbb{S}^n in Section 8.3 is not oriented. The transition map

$$h_V \circ h_U^{-1} : \mathbb{R}^n - \{0\} \rightarrow \mathbb{R}^n - \{0\}$$

is given by

$$h_V \circ h_U^{-1}(x) = \frac{x}{\langle x, x \rangle}.$$

This is easily seen to be the identity map on the unit sphere

$$\mathbb{S}^{n-1} \subset \mathbb{R}^n - \{0\},$$

and exchanges the open subsets $\{\|x\| > 1\}$ and $\{\|x\| < 1\}$ of $\mathbb{R}^n - \{0\}$. In particular, the transition map may be thought of as the reflection of $\mathbb{R}^n - \{0\}$ across the unit sphere. As such, when $n = 2$, it plays an important role in hyperbolic geometry. In Exercise 6 below, we show that $D(h_V \circ h_U^{-1})(e_n)$ has determinant -1 . When $n \geq 2$, this shows the transition map to be orientation-reversing by Proposition 8.2.7, as $\mathbb{R}^n - \{0\}$ is path-connected (Exercise 7).

For the case $n = 1$, $\mathbb{R} - \{0\}$ is the union of two path-connected pieces, the positive and negative reals. Each piece has an element (± 1) at which $D(h_V \circ h_U^{-1})$ has determinant -1 , so the transition map is again orientation-reversing by Proposition 8.2.7.

But now we can obtain an oriented atlas by composing the south pole chart with an orientation-reversing linear isomorphism of \mathbb{R}^n .

An oriented atlas provides an orientation of M , and indeed any orientation of M can be seen as coming from an oriented atlas. We shall develop this further in our discussion of tangent bundles below.

8.7. Exercises.

1. Let $f : \mathbb{R}^{n-1} \times (-\infty, 1) \rightarrow \mathbb{R}^{n-1} \times (-\infty, 1)$ be given by

$$f(x, t) = ((1-t)x_1, \dots, (1-t)x_{n-1}, t) = (1-t)(x, 0) + tN,$$

with N the north pole.

Let $k : \mathbb{R}^{n-1} \times (-\infty, 1) \rightarrow \mathbb{R}^{n-1} \times (-\infty, 1)$ be given by

$$\begin{aligned} k(x, t) &= (x, (1-u(x))t + u(x)), \\ &= (x, t + u(x)(1-t)) \end{aligned}$$

with

$$u(x) = 1 - \frac{2}{1 + \langle x, x \rangle} \in (-1, 1).$$

- Compute the Jacobian matrices of f and k and their determinants. Show that f and k are diffeomorphisms.
 - Show that $(f \circ k)|_{\mathbb{R}^{n-1} \times 0}$ coincides with the map g_U of Theorem 8.3.6.
 - Deduce from this and the analogous result for g_V that the smooth structure on \mathbb{S}^{n-1} given in Corollary 8.3.7 coincides with the structure of an embedded submanifold of \mathbb{R}^n .
2. Let $B_1(0)$ be the standard open ball of radius 1 in \mathbb{R}^n . Let $f : B_1(0) \rightarrow \mathbb{R}^n$ be given by

$$f(x) = \frac{1}{1 - \langle x, x \rangle} x.$$

Show that f is a diffeomorphism.

3. Let $f : (-\frac{\pi}{2}, \frac{\pi}{2})^n \rightarrow \mathbb{R}^n$ via

$$f(x_1, \dots, x_n) = (\arctan(x_1), \dots, \arctan(x_n)).$$

Show that f is a diffeomorphism.

4. Let $U \subset \mathbb{R}^n$, open, and let $f : U \rightarrow \mathbb{R}^m$ be smooth. Let $\gamma : (a, b) \rightarrow U$ be a smooth curve.

- Show that $D(f \circ \gamma)(t)$ depends only on $\gamma(t)$ and $\gamma'(t)$ for $t \in (a, b)$. In fact we refer to $D(f \circ \gamma)(t)$ as the directional derivative of f at $\gamma(t)$ in the direction $\gamma'(t)$.
- Show that if $\gamma'(t) = e_j$, the j th canonical basis vector, then $D(f \circ \gamma)(t)$ is the j th column of $Df(\gamma(t))$. In particular, if $m = 1$, then $D(f \circ \gamma)(t) = \frac{\partial f}{\partial x_j}(\gamma(t))$.

- Suppose now that $\gamma'(t) = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = c_1 e_1 + \dots + c_n e_n$ and that

$m = 1$. Show that

$$D(f \circ \gamma)(t) = c_1 \frac{\partial f}{\partial x_1}(\gamma(t)) + \dots + c_n \frac{\partial f}{\partial x_n}(\gamma(t)).$$

5. Let $n \geq 2$. For $j \in \{1, \dots, n-1\}$, define $\gamma_j : (-\frac{\pi}{2}, \frac{\pi}{2}) \rightarrow \mathbb{R}^n$ by

$$\gamma_j(t) = \cos(t)e_n + \sin(t)e_j.$$

- (a) Show that $\gamma_j(t)$ lies in the unit sphere \mathbb{S}^{n-1} for all $t \in (-\frac{\pi}{2}, \frac{\pi}{2})$.
 (b) Show that $\gamma_j(0) = e_n$ and $\gamma_j'(0) = e_j$.
 (c) Deduce from Problem 4 that if U is a neighborhood of e_n in \mathbb{R}^n and $f : U \rightarrow \mathbb{R}^m$ is smooth, and if we restrict the domain of γ_j to an interval $(-\epsilon, \epsilon)$ contained in $\gamma_j^{-1}(U)$, then $D(f \circ \gamma_j)(0)$ is the j th column of $Df(e_n)$.
 (d) Deduce that if $U \subset \mathbb{R}^n$ is an open set containing \mathbb{S}^{n-1} and if $f : U \rightarrow \mathbb{R}^m$ restricts to the identity on \mathbb{S}^{n-1} , then the j th column of $Df(e_n)$ is e_j for $j = 1, \dots, n-1$.
6. Let $n \geq 1$ and let $f : \mathbb{R}^n - \{0\} \rightarrow \mathbb{R}^n - \{0\}$ be given by $f(x) = \frac{x}{\langle x, x \rangle}$. Define $\gamma : (0, 2) \rightarrow \mathbb{R}^n - \{0\}$ by $\gamma(t) = te_n$. Write f_i for the i th coordinate function of f for $i = 1, \dots, n$.
- (a) Show that $f_i \circ \gamma$ is constant for $1 \leq i < n$ and that $f_n \circ \gamma(t) = \frac{1}{t}$ for all t .
 (b) Deduce that the last column of $Df(e_n)$ is $-e_n$.
 (c) Deduce from Problem 5 that if $n \geq 2$, $Df(e_n)$ is the diagonal matrix whose first $n-1$ diagonal entries are 1 and whose last diagonal entry is -1 .
 (d) Show that if $n = 1$, $Df(1) = Df(-1) = [-1]$.
7. For $n \geq 2$, show that $\mathbb{R}^n - \{0\}$ is path-connected.

9. Spherical geometry

The sphere \mathbb{S}^2 is a reasonably good model for the surface of the earth. Good enough that it can be used to calculate shortest flight paths for airplane flights. The point is that airplanes have to stay within a certain distance of the surface of the earth and cannot, for instance, tunnel through it. So the distance travelled along the surface is a good model for the total distance flown.

So how do we calculate distance on the surface of the earth? Like distance in the plane, it depends on the inner product. This time, we use the standard inner product in \mathbb{R}^3 and use it to calculate the arc length of curves on the sphere.

We may as well generalize this to studying the unit sphere \mathbb{S}^n of \mathbb{R}^{n+1} , as the shortest paths in \mathbb{S}^n may be developed in a similar way to those in \mathbb{S}^2 .

We shall see later in our discussion of Riemannian geometry that shortest paths may be given many different parametrizations, but the parametrizations whose velocity vectors have constant length play a special role. They are called geodesics. The geodesics in \mathbb{R}^n are the standard parametrizations of lines (Proposition 11.3.13). Geodesics in \mathbb{S}^n are the great circle routes, which give parametrizations of great circles (Definition 9.1.3). For this reason, we shall treat the great circles as the “lines” in our discussions of spherical geometry. We shall discuss lengths of line segments, angles between oriented lines, etc., just as we did for \mathbb{R}^n . And the calculation of shortest paths will allow us to show that $\mathcal{I}(\mathbb{S}^n)$, the group of isometries of \mathbb{S}^n , is isomorphic to the orthogonal group O_{n+1} , i.e., to the group of linear isometries of \mathbb{R}^{n+1} (Theorem 9.1.16).

The geometry of \mathbb{S}^n is easier to frame and understand than the geometry of a general Riemannian manifold because \mathbb{S}^n is a smooth submanifold of \mathbb{R}^{n+1} (e.g., by Exercise 1 in Chapter 8), and the geometry we care about for \mathbb{S}^n is induced by the inner product of \mathbb{R}^{n+1} . (This is called the *subspace metric* on the submanifold.) In particular, we do not need to develop the theory of geodesics (and the exponential map) in full generality to study \mathbb{S}^n . Instead, we shall devote some time to developing simpler, and hopefully more intuitive arguments that depend on the use of the subspace metric. That is the substance of Section 9.1.

9.1. Arc length and distance in \mathbb{S}^n ; isometries of \mathbb{S}^n . Let M be a smooth submanifold of \mathbb{R}^{n+1} . We shall use the inner product in \mathbb{R}^{n+1} to calculate distances in M . We calculate them in terms of the arc lengths of curves $\gamma : [a, b] \rightarrow M$.

Arc length is often studied in a first multivariate calculus class. One can calculate the arc length of a piecewise smooth curve in \mathbb{R}^{n+1} . Here, $\gamma : [a, b] \rightarrow \mathbb{R}^{n+1}$ is piecewise smooth if there is a partition $a = x_0 < x_1 < \dots < x_k = b$ of $[a, b]$ such that the restriction of γ to $[x_i, x_{i+1}]$ is smooth for

$i = 0, \dots, k - 1$. Here, a map from a closed interval is smooth if it could be extended to a smooth map on a slightly larger open interval.

Let $\gamma : (a, b) \rightarrow \mathbb{R}^{n+1}$ be smooth. Write

$$\gamma(t) = \begin{bmatrix} \gamma_1(t) \\ \vdots \\ \gamma_{n+1}(t) \end{bmatrix},$$

i.e., $\gamma(t) = \gamma_1(t)e_1 + \dots + \gamma_{n+1}(t)e_{n+1}$. Then the velocity vector $\gamma'(t)$ is the Jacobian matrix of γ at t :

$$\gamma'(t) = \begin{bmatrix} \gamma'_1(t) \\ \vdots \\ \gamma'_{n+1}(t) \end{bmatrix} = \gamma'_1(t)e_1 + \dots + \gamma'_{n+1}(t)e_{n+1}.$$

We define the arc length of a piecewise smooth curve $\gamma : [a, b] \rightarrow \mathbb{R}^{n+1}$ to be

$$(9.1.1) \quad \ell(\gamma) = \int_a^b \|\gamma'(t)\| dt,$$

the integral over $[a, b]$ of the length of the tangent vector of γ at each $t \in [a, b]$. The length $\|\gamma'(t)\|$ is often called the speed of γ at t . It depends, of course, on the inner product in \mathbb{R}^{n+1} :

$$\|\gamma'(t)\| = \sqrt{\langle \gamma'(t), \gamma'(t) \rangle}.$$

Allowing γ to be piecewise smooth rather than smooth permits traversing two sides of a triangle, for instance, and adding the distances travelled.

But, as discussed at the beginning of this chapter, if we want distances in M , we must study the arc lengths of piecewise smooth curves $\gamma : [a, b] \rightarrow M$, i.e., curves whose image lies in M , and it is natural to ask for smoothness in terms of the smooth structure on M . But by Proposition 8.4.14, there is no distinction between smooth maps into M and smooth maps into \mathbb{R}^{n+1} whose image lies in M : a map $f : N \rightarrow M$ is smooth if and only if the composite

$$N \xrightarrow{f} M \subset \mathbb{R}^{n+1}$$

is smooth.

Definition 9.1.1. Let M be a smooth, path-connected submanifold of \mathbb{R}^{n+1} and let $x, y \in M$. Then the distance from x to y in M (in the metric induced by the inclusion of M in \mathbb{R}^{n+1}) is given by

$$(9.1.2) \quad d(x, y) = \inf_{\gamma} \ell(\gamma),$$

where γ runs over the piecewise smooth paths from x to y in M . Here, a piecewise smooth curve $\gamma : [a, b] \rightarrow M$ is a path from x to y if $\gamma(a) = x$ and $\gamma(b) = y$.¹⁵

We say that the piecewise smooth path γ from x to y is distance minimizing if $d(x, y) = \ell(\gamma)$, i.e., if $\ell(\gamma) \leq \ell(\Delta)$ whenever Δ is a piecewise smooth path from x to y .

In (9.1.2), the arc length is calculated via (9.1.1), using the inner product in \mathbb{R}^{n+1} to calculate the lengths of the velocity vectors. With that assumption, (9.1.2) is the distance formula coming from the intrinsic Riemannian geometry for a submanifold of \mathbb{R}^{n+1} in the *subspace metric*. In Chapter 11, we will consider situations in which the inner product varies depending on the value of the point $\gamma(t)$, independent of any particular embedding of M in Euclidean space. That will give the general case of the distance formula in Riemannian geometry.

An obvious special case here is where $M = \mathbb{R}^{n+1}$. Here, we have been calculating the distance from x to y as $\|y - x\|$. It will be helpful to show that the distance formula (9.1.2) gives the same result here.

Proposition 9.1.2. *Let $x, y \in \mathbb{R}^{n+1}$. Then the straight line path*

$$\gamma(t) = x + t(y - x), \quad t \in [0, 1],$$

is distance minimizing. Its arc length is $\|y - x\|$.

Proof. Note that all paths from x to y are translates of paths from 0 to $y - x$ and that translation does not affect derivatives, and hence preserves arc length. Translation also preserves straight line paths. Thus, we may assume $x = 0$ and study the distance from 0 to some arbitrary point y .

Let $u_1 = \frac{y}{\|y\|}$, and extend it to a basis u_1, \dots, u_{n+1} of \mathbb{R}^{n+1} . Applying the Gram–Schmidt process, if necessary, we may assume u_1, \dots, u_{n+1} is an orthonormal basis of \mathbb{R}^{n+1} . Let $\gamma : [a, b] \rightarrow \mathbb{R}^{n+1}$ be a piecewise smooth path from 0 to y , and set

$$\gamma_i(t) = \langle \gamma(t), u_i \rangle$$

for $i = 1, \dots, n + 1$. Then (4.1.1) gives

$$\begin{aligned} \gamma(t) &= \langle \gamma(t), u_1 \rangle u_1 + \cdots + \langle \gamma(t), u_{n+1} \rangle u_{n+1} \\ &= \gamma_1(t) u_1 + \cdots + \gamma_{n+1}(t) u_{n+1}, \end{aligned}$$

and hence

$$\gamma'(t) = \gamma'_1(t) u_1 + \cdots + \gamma'_{n+1}(t) u_{n+1}$$

Moreover, the Pythagorean formula for orthonormal coordinates gives

$$\|\gamma'(t)\| = \sqrt{(\gamma'_1(t))^2 + \cdots + (\gamma'_{n+1}(t))^2}.$$

¹⁵Since M is path-connected, the Whitney approximation theorem ([13, Theorem 6.26]) shows there are smooth, and hence piecewise smooth paths from x to y .

Thus,

$$\|\gamma'(t)\| \geq \sqrt{(\gamma'_1(t))^2} = |\gamma'_1(t)|,$$

and hence

$$(9.1.3) \quad \ell(\gamma) = \int_0^1 \|\gamma'(t)\| dt \geq \int_0^1 \gamma'_1(t) dt = \gamma_1(1) - \gamma_1(0)$$

by the fundamental theorem of calculus. Note that $\gamma(0) = 0$ and

$$\gamma(1) = y = \|y\|u_1 + 0u_2 + \cdots + 0u_{n+1},$$

so $\gamma_1(1) = \|y\|$ and (9.1.3) gives $\ell(\gamma) \geq \|y\|$. But if β is the straight line path from 0 to y , $\beta(t) = ty$ for $t \in [0, 1]$, then $\beta'(t) = y$ for all t , and hence

$$\ell(\beta) = \int_0^1 \|y\| dt = \|y\|. \quad \square$$

This not only provides a reality check for this new notion of Riemannian distance, it helps us understand the properties of that distance. A useful notion of distance should satisfy the properties in Definition A.1.1, below: symmetry, positive-definiteness and the triangle inequality.

The triangle inequality is the easiest to show, as if $\gamma : [a, b] \rightarrow M$ is a piecewise smooth path from x to y and $\Delta : [b, c] \rightarrow M$ is a piecewise smooth path from y to z , then the arc length of the path obtained by traversing first γ and then Δ is the sum of the arc lengths of γ and Δ , demonstrating that

$$(9.1.4) \quad d(x, z) \leq d(x, y) + d(y, z) \quad \text{for any } x, y, z \in M.$$

This is the triangle inequality, and is the reason we have considered piecewise smooth paths instead of insisting on smooth ones.

Reflexivity says $d(x, y) = d(y, x)$. But it's an easy consequence of the chain rule that traversing a path in the opposite direction does not change arc length.

Positive-definiteness, the assertion that $d(x, y) > 0$ if $x \neq y$, is somewhat difficult to prove in general Riemannian manifolds, but is very easy for submanifolds $M \subset \mathbb{R}^{n+1}$ with the subspace metric. The point is that any piecewise smooth path from x to y in M is automatically a piecewise smooth path in \mathbb{R}^{n+1} , and therefore its arc length is greater than or equal to the distance from x to y in \mathbb{R}^{n+1} . In other words the Riemannian distance from x to y in M is greater than or equal to the distance from x to y in \mathbb{R}^{n+1} , a fact we observed intuitively for the distance flown by an airplane in getting from one city to another. Since the distance in \mathbb{R}^{n+1} is positive-definite, the same must be true for the Riemannian distance in M .

The distance minimizing paths in \mathbb{S}^n occur along great circles.

Definition 9.1.3. A great circle in \mathbb{S}^n is the intersection $\mathbb{S}^n \cap V$ for V a 2-dimensional linear subspace of \mathbb{R}^{n+1} . Recall from Lemma 7.2.5 that if v, w is an orthonormal basis for V of \mathbb{R}^{n+1} , then

$$\mathbb{S}^n \cap V = \{\cos tv + \sin tw : t \in \mathbb{R}\}.$$

Thus, for $v, w \in \mathbb{S}^n$ with $\langle v, w \rangle = 0$, the *great circle route*

$$(9.1.5) \quad \begin{aligned} \gamma_{v,w} : \mathbb{R} &\rightarrow \mathbb{S}^n \\ \gamma_{v,w}(t) &= \cos tv + \sin tw \end{aligned}$$

parametrizes the great circle $\mathbb{S}^n \cap \text{span}(v, w)$. We shall also refer to the restriction of $\gamma_{v,w}$ to a closed interval as a great circle route.

Great circles are the “lines” in spherical geometry, and the great circle routes are their standard parametrizations. Since a given 2-dimensional subspace V has many choices of orthonormal basis, there are numerous great circle routes parametrizing $V \cap \mathbb{S}^n$, but they are nicely related. The results in the following lemma follow from standard trigonometric formulae, e.g., the formulae for $\cos(\phi + \psi)$ and $\sin(\phi + \psi)$. The proof is left to the reader.

Lemma 9.1.4. *Let v, w be orthogonal vectors in \mathbb{S}^n . Then:*

- (1) *The traversal of $\gamma_{v,w}$ in the opposite direction satisfies*

$$(9.1.6) \quad \gamma_{v,w}(-t) = \cos tv - \sin tw = \gamma_{v,-w}(t).$$

Thus, $\gamma_{v,-w}$ provides the reverse orientation of the great circle $\mathbb{S}^n \cap \text{span}(v, w)$ to that given by $\gamma_{v,w}$. This corresponds to the fact that $v, -w$ gives the opposite orientation to $\text{span}(v, w)$ from that given by v, w .

- (2) *Let $x = \gamma_{v,w}(c)$ and let $y = \gamma_{v,w}(c + \frac{\pi}{2})$. Then*

$$(9.1.7) \quad \gamma_{v,w}(c + t) = \gamma_{x,y}(t)$$

for all $t \in \mathbb{R}$. Thus $\gamma_{v,w} \circ \tau_c = \gamma_{x,y}$ where τ_c is the translation of \mathbb{R} by c . By Lemma 7.2.5, such pairs x, y give all the orthonormal bases of $\text{span}(v, w)$ that induce the same orientation as v, w . Thus, as c varies, we obtain all the like-oriented great circle routes parametrizing $\mathbb{S}^n \cap \text{span}(v, w)$.

- (3) *Again by Lemma 7.2.5, the curves $\gamma_{x,-y}$ with x, y as in (2) give all great circle routes parametrizing $\mathbb{S}^n \cap \text{span}v, w$ with the opposite orientation. By (1),*

$$(9.1.8) \quad \gamma_{x,-y}(t) = \gamma_{x,y}(-t) = \gamma_{v,w}(c - t).$$

- (4) *The velocity vector to $\gamma_{v,w}$ at t satisfies*

$$(9.1.9) \quad \gamma'_{v,w}(t) = -\sin tv + \cos tw = \gamma_{v,w}\left(t + \frac{\pi}{2}\right).$$

Thus, $\gamma_{v,w}(t), \gamma'_{v,w}(t)$ is one of the bases of $\text{span}(v, w)$ specified in (2).

By (9.1.9), $\|\gamma'_{v,w}(t)\| = 1$ for all t . (Curves with this property are said to be parametrized by arc length.) Thus the speed of $\gamma_{v,w}$ is constantly equal to 1, and we obtain the following:

Corollary 9.1.5. *Let $v, w \in \mathbb{S}^n$ be orthogonal. Then the arc length of the restriction of $\gamma_{v,w}$ to a closed interval is given by*

$$(9.1.10) \quad \ell(\gamma_{v,w}|_{[a,b]}) = \int_a^b \|\gamma'_{v,w}(t)\| dt = \int_a^b dt = b - a.$$

Moreover, by Lemma 9.1.4, all other parametrizations of $\gamma_{v,w}|_{[a,b]}$ by great circle routes have the same arc length.

Proof. For the last statement, two parametrizations of $\gamma_{v,w}|_{[a,b]}$ by great circle routes differ by precomposition with an isometry of \mathbb{R} . \square

Note that $\gamma_{v,w}$ is periodic in the sense that $\gamma_{v,w}(t + 2\pi) = \gamma_{v,w}(t)$, so if $b - a > 2\pi$ then the length of $\gamma_{v,w}|_{[a,b]}$ is greater than the arc length of the full circle traced out by $\gamma_{v,w}$, which is 2π . In particular, for $b - a$ large enough, $\gamma_{v,w}|_{[a,b]}$ can wrap around the circle as many times as you like. We will see that $\gamma_{v,w}|_{[a,b]}$ is distance minimizing if and only if $b - a \leq \pi$.

When $b - a \leq \pi$ we can calculate the arc length of $\gamma_{v,w}|_{[a,b]}$ by a formula that only depends on the points it connects in \mathbb{S}^n . Indeed, once we show this path is distance minimizing, this will give us the Riemannian distance between these two points.

Lemma 9.1.6. *If $b - a \leq \pi$, then*

$$(9.1.11) \quad \ell(\gamma_{v,w}|_{[a,b]}) = \cos^{-1}(\langle \gamma_{v,w}(b), \gamma_{v,w}(a) \rangle).$$

Proof. For simplicity, consider the case $a = 0$. Then $\gamma_{v,w}(a) = v$ and $\gamma_{v,w}(b) = \cos bv + \sin bw$. Since v, w is an orthonormal set,

$$\langle v, \cos bv + \sin bw \rangle = \cos b.$$

If $0 \leq b \leq \pi$, the result follows from the standard properties of \cos^{-1} .

For the general case, expanding $\langle \cos av + \sin aw, \cos bv + \sin bw \rangle$ results in the usual expansion of $\cos(b - a)$. \square

We may as well at this point codify definitions about lines and line segments in \mathbb{S}^n , as the ideas are useful in showing great circle routes minimize arc length.

Definition 9.1.7. A line in \mathbb{S}^n is a great circle. A line segment in \mathbb{S}^n is a subset

$$(9.1.12) \quad \{\gamma_{u,v}(t) : t \in [a, b]\} \quad \text{with } u \perp v \in \mathbb{S}^n \text{ and } b - a < 2\pi.$$

Its endpoints are $x = \gamma_{u,v}(a)$ and $y = \gamma_{u,v}(b)$. We say the parametrization in (9.1.12) goes from x to y . Note that a parametrization from y to x is then given by $\gamma_{u,-v}|_{[-b,-a]}$. By Corollary 9.1.5, any two parametrizations of a line segment by great circle routes have the same arc length. We define the length of the segment to be this arc length.

The following is useful.

Lemma 9.1.8. *Let $u \neq \pm v \in \mathbb{S}^n$. Then there is a unique great circle containing u and v and exactly two line segments with u and v as endpoints. The shorter of these segments has length $\cos^{-1}(\langle u, v \rangle)$. We shall call it the “minor segment” with endpoints u and v . The longer has length $2\pi - \cos^{-1}(\langle u, v \rangle)$. We shall call it the “major segment”.*

If $u = -v$, then every great circle through one of them contains the other. All the great circle routes with u and v as endpoints have length $\cos^{-1}(\langle u, v \rangle) = \pi$.

Proof. Suppose $u \neq \pm v$. Since both have norm 1, they are linearly independent and $V = \text{span}(u, v)$ is a 2-dimensional subspace of \mathbb{R}^{n+1} . So $V \cap \mathbb{S}^n$ is the unique great circle containing u and v . Let u, w be an orthonormal basis of V . Then $v = \cos t u + \sin t w$ for some $t \in (0, 2\pi)$. Then $\gamma_{v,w}|_{[0,t]}$ gives one great circle route from u to v . Its length is t , and $\cos t = \langle u, v \rangle$. The two possible values of $t \in [0, 2\pi)$ with this cosine are $\cos^{-1}(\langle u, v \rangle)$ and $2\pi - \cos^{-1}(\langle u, v \rangle)$. The result now follows since $\gamma_{v,w}|_{[t,2\pi]}$ parametrizes the other line segment with endpoints u and v .

If $u = -v$, then every linear subspace containing one contains the other. If V is a 2-dimensional linear subspace containing u , simply choose an orthonormal basis u, w for V and calculate as above. \square

Showing that great circle routes of length $\leq \pi$ on \mathbb{S}^n are distance minimizing is surprisingly deep. Isometries will be useful, both for this and for further study of the geometry of the spheres.

Recall from Corollary 8.4.15 that if $\alpha : \mathbb{R}^{n+1} \rightarrow \mathbb{R}^{n+1}$ is a linear isometry, then it restricts to a diffeomorphism $\alpha_0 = \alpha|_{\mathbb{S}^n} : \mathbb{S}^n \rightarrow \mathbb{S}^n$. If $\gamma : [a, b] \rightarrow \mathbb{S}^n$ is piecewise smooth, so is $\alpha_0 \circ \gamma$ and we can compare their arc lengths. The following shows α_0 is an isometry in the same sense as the isometries of \mathbb{R}^n studied in previous chapters.

Proposition 9.1.9. *Let $\alpha : \mathbb{R}^{n+1} \rightarrow \mathbb{R}^{n+1}$ be a linear isometry and let $\alpha_0 : \mathbb{S}^n \rightarrow \mathbb{S}^n$ be its restriction to \mathbb{S}^n . Then α_0 preserves arc length, and hence preserves distance between points on the sphere.*

More specifically, if $\alpha = T_A$, the linear transformation induced by the $(n+1) \times (n+1)$ orthogonal matrix A , and if $\gamma : [a, b] \rightarrow \mathbb{S}^n$ is piecewise smooth, then the velocity vector $(\alpha_0 \circ \gamma)'(t)$ is given by the matrix product

$$(9.1.13) \quad (\alpha_0 \circ \gamma)'(t) = A \cdot \gamma'(t),$$

and hence has the same norm as $\gamma'(t)$.

Proof. Since the velocity vector for a map $\gamma : [a, b] \rightarrow \mathbb{S}^n$ is simply the derivative of the composite

$$[a, b] \xrightarrow{\gamma} \mathbb{S}^n \subset \mathbb{R}^{n+1},$$

we simply calculate our derivatives in \mathbb{R}^{n+1} . Thus

$$(\alpha_0 \circ \gamma)'(t) = (\alpha \circ \gamma)'(t) = D\alpha(\gamma(t)) \cdot \gamma'(t)$$

by the chain rule. Equation (9.1.13) is now immediate from Lemma 8.1.6. Since A is an orthogonal matrix,

$$\|(\alpha_0 \circ \gamma)'(t)\| = \|\gamma'(t)\|,$$

and hence the arc length integrals for $\alpha_0 \circ \gamma$ and γ are identical. Thus, α_0 preserves arc lengths of paths. But the inverse of α_0 is also the restriction to \mathbb{S}^n of a linear isometry of \mathbb{R}^{n+1} , so α_0 preserves the Riemannian distance between points in the sphere. \square

Remark 9.1.10. In fact, the proof of Proposition 9.1.9 shows that α_0 is an isometry in the stronger sense (Definition 11.1.5) used in general Riemannian manifolds, and therefore preserves angles between curves in \mathbb{S}^n (Lemma 11.2.12). We shall discuss angles in greater detail in our development of the geometry of \mathbb{S}^2 . For convenience, however, we shall use the following definition in this section.

Definition 9.1.11. An isometry of \mathbb{S}^n is a function $\alpha : \mathbb{S}^n \rightarrow \mathbb{S}^n$ that preserves the Riemannian distance. We write $\mathcal{I}(\mathbb{S}^n)$ for the set of all isometries of \mathbb{S}^n .

We shall see in Theorem 9.1.16 that every isometry of \mathbb{S}^n is the restriction to \mathbb{S}^n of a linear isometry of \mathbb{R}^{n+1} . Thus, every isometry of \mathbb{S}^n is a surjective diffeomorphism, and $\mathcal{I}(\mathbb{S}^n)$ is a group isomorphic to the orthogonal group O_{n+1} .

To show that great circle routes are distance minimizing, we shall introduce what turns out to be the exponential map for \mathbb{S}^n . Normally, the exponential map is defined using geodesics in a Riemannian manifold. The geodesics are specific parametrizations of distance minimizing curves. The great circle routes as defined above are in fact geodesics.

We take an approach here that does not use Riemannian geometry, and define a map that coincides with the Riemannian exponential map. We will call it the exponential map, but will define it directly and verify its properties directly.

Definition 9.1.12. Let $S = -e_{n+1}$, the negative of the last canonical basis vector of \mathbb{R}^{n+1} . We identify it with the south pole of \mathbb{S}^n . The exponential map for \mathbb{S}^n at S is the map $\exp : \mathbb{R}^n \rightarrow \mathbb{S}^n$ given by

$$(9.1.14) \quad \exp(v) = \begin{cases} S & \text{if } v = 0, \\ \cos \|v\| S + \sin \|v\| \frac{v}{\|v\|} & \text{if } v \neq 0. \end{cases}$$

Here, as in Theorem 8.3.6, we identify \mathbb{R}^n with the equatorial n -plane in \mathbb{R}^{n+1} : the points whose last coordinate is 0. This identifies the unit sphere in \mathbb{R}^n with the equator in \mathbb{S}^n : the set of points in \mathbb{S}^n orthogonal to S .

Note, then that if $\|v\| = 1$ in \mathbb{R}^n , we obtain

$$(9.1.15) \quad \exp(tv) = \cos t S + \sin t v = \gamma_{S,v}(t),$$

so $t \mapsto \exp(tv)$ is precisely the great circle route from S through v .

Proposition 9.1.13. *The exponential map*

$$\exp : \mathbb{R}^n \rightarrow \mathbb{S}^n$$

is smooth. Its restriction to $B_\pi(0) = \{w \in \mathbb{R}^n : \|w\| < \pi\}$ gives a diffeomorphism

$$(9.1.16) \quad \exp|_{B_\pi(0)} : B_\pi(0) \xrightarrow{\cong} \mathbb{S}^n \setminus \{e_{n+1}\}.$$

Proof. Elementary calculus comes to the rescue here. Let

$$f(x) = \begin{cases} \frac{\sin x}{x} & x \neq 0 \\ 1 & x = 0. \end{cases}$$

Then (9.1.14) gives

$$(9.1.17) \quad \exp(v) = \cos \|v\| S + f(\|v\|)v,$$

so \exp is smooth if $v \mapsto \cos \|v\|$ and $v \mapsto f(\|v\|)v$ are smooth. Now

$$\begin{aligned} \cos x &= \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k} = g(x^2), \\ f(x) &= \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k} = h(x^2), \end{aligned}$$

where $g(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^k$ and $h(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^k$ are smooth on all of \mathbb{R} by the ratio test. While $v \mapsto \|v\| = \sqrt{\langle v, v \rangle}$ is not smooth at 0 because of the square root, $v \mapsto \langle v, v \rangle$ is polynomial on the coordinates of v , and hence smooth on all of \mathbb{R}^n . Thus,

$$\exp(v) = g(\langle v, v \rangle)S + h(\langle v, v \rangle)v$$

is smooth on all of \mathbb{R}^n .

We claim next that $\exp|_{B_\pi(0)} : B_\pi(0) \rightarrow \mathbb{S}^n \setminus \{e_{n+1}\}$ is bijective. We construct an inverse function as follows. Let $u \in \mathbb{S}^n$. Then $u = x + sS$ with $x \in \mathbb{R}^n$ (i.e., with $\langle x, S \rangle = 0$). Thus, $x = u - \langle u, S \rangle S$. Since $\langle x, S \rangle = 0$, we have

$$1 = \langle u, u \rangle = \langle x, x \rangle + \langle sS, sS \rangle = \|x\|^2 + s^2.$$

Since $s \in [-1, 1]$, $t = \cos^{-1} s$ is the unique element of $[-\frac{\pi}{2}, \frac{\pi}{2}]$ with $s = \cos t$ and $\|x\| = \sin t$. If $u \neq \pm S$, $x \neq 0$ and $\cos t > 0$, so $v = \frac{x}{\|x\|}$ is the unique element of \mathbb{S}^n with $u = \exp(tv)$. Of course, $0 = \exp^{-1}(S)$.

Moreover, $t = \cos^{-1}(\langle u, S \rangle)$ and $v = \frac{u - \langle u, S \rangle S}{\|u - \langle u, S \rangle S\|}$ are smooth functions on

$$\{u \in \mathbb{R}^{n+1} : |\langle u, S \rangle| < 1\} \setminus \text{span}(S),$$

an open set in \mathbb{R}^{n+1} containing $\mathbb{S}^n \setminus \{\pm S\}$. Thus, the inverse function \exp^{-1} is smooth on $\mathbb{S}^n \setminus \{\pm S\}$.

Thus, it suffices to show \exp^{-1} is smooth on a neighborhood of S . By the inverse function theorem, this follows if the Jacobian matrix at 0 of the composite

$$B_0(\pi) \xrightarrow{\exp} \mathbb{S}^n \setminus \{e_{n+1}\} \xrightarrow{h} \mathbb{R}^n$$

is invertible, where h corresponds to the chart h_U for the smooth structure of \mathbb{S}^n given in Theorem 8.3.6. At this point, in calculating, Jacobian matrices, we should drop the $(n+1)$ -st coordinate of 0 that we have been carrying for points in \mathbb{R}^n . We may express h by

$$(9.1.18) \quad h(x_1e_1 + \cdots + x_{n+1}e_{n+1}) = \frac{1}{1 - x_{n+1}}(x_1e_1 + \cdots + x_n e_n).$$

Note this extends by the same formula to a smooth function

$$\bar{h} : \{x_1e_1 + \cdots + x_{n+1}e_{n+1} \in \mathbb{R}^{n+1} : x_{n+1} \neq 1\} \rightarrow \mathbb{R}^n$$

and that $h \circ \exp = \bar{h} \circ \exp$.

Write $f = \bar{h} \circ \exp$. Then the i th column of $Df(0)$ is the velocity vector at 0 of $f \circ \Delta$ where $\Delta : (-\pi, \pi) \rightarrow B_0(\pi)$ is given by $\Delta(t) = te_i$. (Here, $1 \leq i \leq n$.) By (9.1.15), $\exp \circ \Delta = \gamma_{S, e_i}$, so the chain rule gives

$$\begin{aligned} (f \circ \Delta)'(0) &= D\bar{h}(S) \cdot (\exp \circ \Delta)'(0) \\ &= D\bar{h}(S) \cdot \gamma'_{S, e_i}(0) \\ &= D\bar{h}(S) \cdot e_i, \end{aligned}$$

the i th column of $D\bar{h}(S)$. By (9.1.18), the j th coordinate function of \bar{h} is $\frac{x_j}{1-x_{n+1}}$ for $1 \leq i \leq n$. Thus, $D\bar{h}(S) \cdot e_i = \frac{1}{2}e_i$ for these values of i . Thus, $Df(0) = \frac{1}{2}I_n$. \square

The following is elaborated further in Corollary 10.2.12.

Lemma 9.1.14. *Let $\gamma : (a, b) \rightarrow \mathbb{S}^n$ be smooth. Then $\langle \gamma(t), \gamma'(t) \rangle = 0$ for all t .*

Proof. Let $f : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ be given by $f(x) = \langle x, x \rangle$. Then $\mathbb{S}^n = f^{-1}(1)$, so $f \circ \gamma$ is constant. Thus,

$$\begin{aligned} 0 &= (f \circ \gamma)'(t) = Df(\gamma(t)) \cdot \gamma'(t) \\ &= 2[\gamma_1(t) \cdots \gamma_{n+1}(t)] \cdot \gamma'(t) \\ &= 2\langle \gamma(t), \gamma'(t) \rangle. \end{aligned} \tag{8.4.3} \quad \square$$

Theorem 9.1.15. *Let $u \neq v \in \mathbb{S}^n$. Then the shortest path from u to v is a great circle route, and therefore has length $\cos^{-1}(\langle u, v \rangle)$. Thus, the Riemannian distance from u to v in \mathbb{S}^n is $\cos^{-1}(\langle u, v \rangle)$.*

Proof. $u \neq 0$, so we can find a basis u_1, \dots, u_{n+1} with $u_1 = u$. Applying Gram–Schmidt if necessary, we may assume u_1, \dots, u_{n+1} is an orthonormal basis of \mathbb{R}^{n+1} . Let $A = [u_2 | \cdots | u_{n+1}] - u_1$, the matrix with columns as listed. Then the transformation T_A induced by A takes the south pole S to

u . Noting that linear isometries preserve great circle routes, we may assume $u = S$ by Proposition 9.1.9.

By Lemma 9.1.8, it suffices to show that if $\gamma : [a, b] \rightarrow \mathbb{S}^n$ is a piecewise smooth path from S to some arbitrary w , then $\ell(\gamma) \geq \cos^{-1}(\langle S, w \rangle)$. If γ visits S or w more than once, we may shorten it, so we may assume $\gamma^{-1}(S) = \{a\}$ and $\gamma^{-1}(w) = \{b\}$.¹⁶

Now first consider the case $w = e_{n+1}$. Here, γ maps $[a, b)$ into $\mathbb{S}^n \setminus \{\pm S\}$ so

$$\exp^{-1} \circ \gamma : (a, b) \rightarrow B_\pi(0) \setminus \{0\}$$

is smooth. Thus, there are smooth functions

$$\begin{aligned} \phi &: (a, b) \rightarrow (0, \pi), \\ v &: (a, b) \rightarrow \mathbb{S}^{n-1} = \mathbb{S}^n \cap \mathbb{R}^n, \end{aligned}$$

with $\gamma(t) = \cos \phi(t)S + \sin \phi(t)v(t)$ for $t \in (a, b)$. But then

$$\gamma'(t) = -\sin \phi(t)\phi'(t)S + \cos \phi(t)\phi'(t)v(t) + \sin \phi(t)W(t).$$

Since $v(t) \in \mathbb{S}^{n-1} \subset \mathbb{R}^n$, $W(t) \in \mathbb{R}^n$, also. So S is orthogonal to both $v(t)$ and $W(t)$. Moreover, $v(t)$ is orthogonal to $W(t)$ by Lemma 9.1.14. Thus,

$$\begin{aligned} \|\gamma'(t)\| &= \sqrt{\sin^2 \phi(t)(\phi'(t))^2 + \cos^2 \phi(t)(\phi'(t))^2 + \sin^2 \phi(t)} \\ &\geq \sqrt{(\sin^2(t) + \cos^2(t))(\phi'(t))^2} \\ &\geq |\phi'(t)| \geq \phi'(t). \end{aligned}$$

Now $\phi(t) = \cos^{-1}(\langle \gamma(t), S \rangle)$ is continuous on $[a, b]$, and the potentially improper integral $\int_a^b \phi'(t) dt$ converges to $\phi(b) - \phi(a) = \pi$ by the fundamental theorem of calculus. Since $\|\gamma'(t)\| \geq \phi'(t)$ and since γ is piecewise smooth, $\ell(\gamma) \geq \pi$.

For $w \neq e_{n+1}$ we may now assume e_{n+1} is not in the image of γ : otherwise $\ell(\gamma) > \pi$. We may now repeat the argument above. \square

Recall that the isometries of \mathbb{S}^n are the (Riemannian) distance preserving functions $\alpha : \mathbb{S}^n \rightarrow \mathbb{S}^n$ and that the set of isometries of \mathbb{S}^n is denoted $\mathcal{I}(\mathbb{S}^n)$. We obtain the following converse to Proposition 9.1.9.

Theorem 9.1.16 (Isometries of \mathbb{S}^n are linear). *Let $\alpha_0 : \mathbb{S}^n \rightarrow \mathbb{S}^n$ be an isometry. Then there is a unique linear isometry, $\alpha : \mathbb{R}^{n+1} \rightarrow \mathbb{R}^{n+1}$ with $\alpha|_{\mathbb{S}^n} = \alpha_0$. We obtain an isomorphism of groups*

$$(9.1.19) \quad \begin{aligned} \iota : \mathcal{O}_{n+1} &\rightarrow \mathcal{I}(\mathbb{S}^n) \\ A &\mapsto T_A|_{\mathbb{S}^n}. \end{aligned}$$

¹⁶Technically, this requires an understanding of the behavior of closed subsets of a closed interval.

Proof. Let $\alpha_0 : \mathbb{S}^n \rightarrow \mathbb{S}^n$ be an isometry. By Theorem 9.1.15,

$$\langle \alpha_0(u), \alpha_0(v) \rangle = \langle u, v \rangle$$

for all $u, v \in \mathbb{S}^n$. Let $v_i = \alpha_0(e_i)$ for $i = 1, \dots, n+1$. Since all points on \mathbb{S}^n have norm 1, v_1, \dots, v_{n+1} is an orthonormal basis of \mathbb{R}^{n+1} , so

$$A = [v_1 | \dots | v_{n+1}]$$

is an orthogonal matrix. Moreover $T_A(e_i) = \alpha_0(e_i)$ for all i , so T_A is the only linear isometry of \mathbb{R}^{n+1} that could restrict to α_0 . To see that it does, let $u \in \mathbb{S}^n$. Since v_1, \dots, v_{n+1} is an orthonormal basis,

$$\begin{aligned} \alpha_0(u) &= \langle \alpha_0(u), v_1 \rangle v_1 + \dots + \langle \alpha_0(u), v_{n+1} \rangle v_{n+1} \\ &= \langle \alpha_0(u), \alpha_0(e_1) \rangle v_1 + \dots + \langle \alpha_0(u), \alpha_0(e_{n+1}) \rangle v_{n+1} \\ &= \langle u, e_1 \rangle v_1 + \dots + \langle u, e_{n+1} \rangle v_{n+1}, \end{aligned}$$

since α_0 preserves the inner product. But this is exactly $T_A(u)$, as

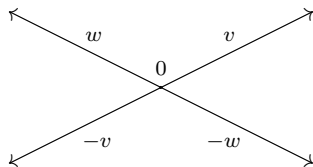
$$u = \langle u, e_1 \rangle e_1 + \dots + \langle u, e_{n+1} \rangle e_{n+1}. \quad \square$$

9.2. Lines and angles in \mathbb{S}^2 . We have defined lines in \mathbb{S}^n to be great circles. The great circle route

$$\gamma_{v,w}(t) = \cos tv + \sin tw$$

provides a parametrization for the great circle $\mathbb{S}^n \cap \text{span}(v, w)$ when $v, w \in \mathbb{S}^n$ are orthogonal. In particular, this gives an orientation for the great circle.

Recall that angles between intersecting lines are not well-defined in \mathbb{R}^n . For instance, in the following picture, the angle between $\text{span}(v)$ and $\text{span}(w)$ could be taken to be the angle from v to w or the angle from v to $-w$.



In absolute value, the two angles add up to π . For this reason, we defined angles as being between rays, rather than lines. Essentially, a ray is an oriented line, so on the sphere, we shall define angles as being between oriented lines.

Note first that in the picture above, the vectors v and w are velocity vectors for specific parametrizations of the lines in question. And if γ and Δ are parametrized curves in \mathbb{S}^n with $\gamma(t) = \Delta(s)$, then $\gamma'(t)$ and $\Delta'(s)$ are vectors in \mathbb{R}^{n+1} . We shall define the angle between γ and Δ at this intersection point to be the angle between their velocity vectors. In general, this is an unsigned angle, because there is no preferred orientation for the two-dimensional subspace these two velocity vectors span.

We shall specialize here to studying the angle between two oriented lines in \mathbb{S}^2 . In this case, the point of intersection will provide an orientation

for the plane containing the velocity vectors, allowing us to define directed angles. Let us set up some conventions.

Conventions 9.2.1. A pair of unit vectors $u, v \in \mathbb{S}^2$ are antipodes if $u = -v$.

Let ℓ be a line in \mathbb{S}^2 , say $\ell = V \cap \mathbb{S}^2$ with V a 2-dimensional subspace of \mathbb{R}^3 . Then V^\perp is 1-dimensional and has two unit vectors, say $\pm u$. A choice of unit vectors determines an orientation for V^\perp .

We shall refer to $\pm u$ as the poles of ℓ . Note that if u is a pole of ℓ , then $V = \{u\}^\perp$ and hence

$$(9.2.1) \quad \ell = \{v \in \mathbb{S}^2 : \langle u, v \rangle = 0\}.$$

As discussed in Remark 7.2.2, a choice of pole for ℓ determines an orientation of V (and vice versa, as the opposite pole determines the opposite orientation of V): if we choose u as our pole, then for any $v \in \ell$, the orientation of V induced by u corresponds to the unique orthonormal basis v, w of V such that $\det[u, v, w] = 1$, i.e., such that the orthonormal basis u, v, w of \mathbb{R}^3 determines the standard orientation of \mathbb{R}^3 .

Moreover, by Corollary 7.3.6, the cross product may be used to determine an orthonormal basis of V inducing the orientation of V determined by the pole u : For any $v \in \ell$, the orientation of V determined by u corresponds to the orthonormal basis $v, u \times v$ of V .

This orientation of V in turn orients ℓ via the great circle route

$$\gamma_{v, u \times v}(t) = \cos tv + \sin t(u \times v).$$

Choosing the opposite pole results in the reverse orientation as

$$(-u) \times v = -(u \times v).$$

Note we can tell these two orientations apart by the velocity vector to the great circle route:

$$(9.2.2) \quad \gamma'_{v, u \times v}(0) = u \times v = -\gamma'_{v, (-u) \times v}(0).$$

Of course, the same relationship will hold if v is replaced by any other point on ℓ . Thus, by Lemma 9.1.4,

$$(9.2.3) \quad \gamma'_{v, u \times v}(t) = -\gamma'_{v, (-u) \times v}(-t)$$

for all t . (The point here is that $\gamma_{v, u \times v}(t) = \gamma_{v, (-u) \times v}(-t)$, so we are taking velocity vectors at the same point.) Of course, this could also be verified directly.

Interestingly, there is no analogue of parallel lines in \mathbb{S}^2 .

Proposition 9.2.2. *Let ℓ and m be distinct lines in \mathbb{S}^2 . Then $\ell \cap m$ consists of the two points $\pm u$, where*

$$(9.2.4) \quad u = \frac{N \times M}{\|N \times M\|},$$

where N and M are poles for ℓ and m , respectively.

Proof. Let V and W be the planes in \mathbb{R}^3 whose intersection with \mathbb{S}^2 give ℓ and m , respectively. By Lemma 7.4.3, $V \cap W = \text{span}(u)$, with u as in (9.2.4). Since $\text{span}(u) \cap \mathbb{S}^2 = \{\pm u\}$, the result follows. \square

This now allows us to define directed angles between oriented lines in \mathbb{S}^2 . The setup here is as follows. Suppose the lines ℓ and m are oriented by the poles N and M , respectively. Let $u = \frac{N \times M}{\|N \times M\|}$. We shall analyze the directed angle from ℓ to m at u with respect to these orientations.

Let $v = N \times u$ and let $w = M \times u$. Then the orientations on ℓ and m are parametrized by $\gamma_{u,v}$ and $\gamma_{u,w}$, respectively. Note that by construction, v and w are orthogonal to u , and lie in $\{u\}^\perp$, a plane whose pole is u and may be oriented by u . As defined above, the unsigned angle between the oriented lines ℓ and m at u is given by the angle in \mathbb{R}^3 between $\gamma'_{u,v}(0) = v$ and $\gamma'_{u,w}(0) = w$. Since v and w lie in the subspace $\{u\}^\perp$, which is oriented by its pole u , we may calculate the directed angle as well:

Definition 9.2.3. Under the conventions above, the directed angle from ℓ to m is equal to the angle in the oriented subspace $\{u\}^\perp$ from v to w . Specifically, if $w = \cos t v + \sin t(u \times v)$, then the directed angle is t . In particular, $\cos t = \langle v, w \rangle$.

But in fact, we can calculate this angle from N and M .

Proposition 9.2.4. *Continuing the conventions above, the directed angle from ℓ to m at $u = \frac{N \times M}{\|N \times M\|}$ is $\cos^{-1}(\langle N, M \rangle)$. The directed angle at $-u$ is $-\cos^{-1}(\langle N, M \rangle)$.*

In particular, at either u or $-u$ the unsigned angle is $\cos^{-1}(\langle N, M \rangle)$.

Proof. First note that since $v = N \times u$,

$$\begin{aligned} u \times v &= -(N \times u) \times u \\ &= -[\langle N, u \rangle u - \langle u, u \rangle N] \\ &= N, \end{aligned}$$

where the second equality is from Proposition 7.3.4(6) and the third is from the fact that u, N is orthonormal. Similarly, $u \times w = M$.

Moreover, Proposition 7.3.4(7) gives

$$\begin{aligned} \langle v, w \rangle &= \langle N \times u, M \times u \rangle \\ &= \langle N, M \rangle \langle u, u \rangle - \langle u, M \rangle \langle N, u \rangle \\ &= \langle N, M \rangle. \end{aligned}$$

Now apply the proof of Proposition 7.4.4. \square

Corollary 9.2.5. *Let ℓ and m be lines in \mathbb{S}^2 with poles N and M , respectively. Then the following conditions are equivalent.*

- (1) ℓ and m are perpendicular, i.e., the unsigned angle between them is $\frac{\pi}{2}$.

- (2) $\langle N, M \rangle = 0$.
- (3) $N \in m$.
- (4) $M \in \ell$.

Proof. The equivalence of (1) and (2) is immediate from Proposition 9.2.4. The other conditions are equivalent via (9.2.1). \square

9.3. Spherical triangles. In \mathbb{R}^2 any three noncollinear points determine a triangle. The same is true in \mathbb{S}^2 . Here, u, v, w are noncollinear if they are not contained in a single spherical line. Since a spherical line is the intersection of \mathbb{S}^2 with a 2-dimensional linear subspace of \mathbb{R}^3 , this is equivalent to saying u, v, w are linearly independent. In particular, no two of u, v, w can be antipodes. The edges of the spherical triangle, $\Delta(u, v, w)$, determined by u, v, w are defined to be the three minor segments whose vertices lie in $\{u, v, w\}$. The following shows how to choose a pole for the minor segment from u to v appropriate for calculating the interior angle of $\Delta(u, v, w)$ at u .

Lemma 9.3.1. *Let $u \neq \pm v \in \mathbb{S}^2$ then $N = \frac{u \times v}{\|u \times v\|}$ is a unit vector orthogonal to both u and v , and hence is a pole for the unique spherical line ℓ containing u and v . Let $w = N \times u$, so that the orientation on $\text{span}(u, v)$ induced by N is given by the orthonormal basis u, w . Then*

$$v = \cos s u + \sin s w$$

for $s \in (0, \pi)$, and hence $\gamma_{u,v}|_{[0,s]}$ is the minor segment from u to v . Moreover,

$$(9.3.1) \quad \|u \times v\| = \sin s = \sin(d(u, v)),$$

where $d(u, v)$ is the Riemannian distance from u to v in \mathbb{S}^2 .

Proof. Since u and v are linearly independent, $u \times v \neq 0$, and hence N is a unit vector. It is orthogonal to u and v . Independently if this, there is a unique unit vector w orthogonal to u such that

$$v = \cos s u + \sin s w$$

for $s \in (0, \pi)$. By bilinearity of the cross product,

$$u \times v = \cos s (u \times u) + \sin s (u \times w) = \sin s (u \times w),$$

as $u \times u = 0$. Since $\sin s > 0$, $\sin s = \|u \times v\|$, and hence $u \times w = N$. But then, $N \times u = w$ by Proposition 7.3.4(6). The second equality in (9.3.1) follows as the distance minimizing path $\gamma_{u,v}|_{[0,s]}$ has length s . \square

The following is now immediate from Proposition 9.2.4.

Corollary 9.3.2. *The interior angle $\angle wuv$ of $\Delta(u, v, w)$ at u has unsigned measure*

$$(9.3.2) \quad m(\angle wuv) = \cos^{-1} \left\langle \frac{u \times w}{\|u \times w\|}, \frac{u \times v}{\|u \times v\|} \right\rangle.$$

Equation (9.3.2) allows us to make the following calculation.

Corollary 9.3.3 (Spherical law of cosines). *In a spherical triangle $\triangle(u, v, w)$,*

$$(9.3.3) \quad \cos(\angle wuv) = \frac{\cos(d(v, w)) - \cos(d(u, w)) \cos(d(u, v))}{\sin(d(u, w)) \sin(d(u, v))}.$$

Proof. Taking the cosine of both sides of (9.3.2) gives

$$\begin{aligned} \cos(\angle wuv) &= \frac{\langle u \times w, u \times v \rangle}{\|u \times w\| \|u \times v\|} \\ &= \frac{\langle u, u \rangle \langle v, w \rangle - \langle u, w \rangle \langle u, v \rangle}{\|u \times w\| \|u \times v\|} && \text{(Proposition 7.3.4(7))} \\ &= \frac{\langle v, w \rangle - \langle u, w \rangle \langle u, v \rangle}{\|u \times w\| \|u \times v\|}. \end{aligned}$$

The denominator agrees with that of the right-hand side of (9.3.3) by (9.3.1). The numerators agree by the distance formula in Theorem 9.1.15. \square

Corollary 9.3.2 is more than we need to analyze the following potentially interesting example.

Example 9.3.4. Consider the spherical triangle $\triangle(e_1, e_2, e_3)$. The poles for the edges of this triangle are $\pm e_1$, $\pm e_2$ and $\pm e_3$, depending on the orientations used to calculate the angles in question. Since these poles are pairwise orthogonal, the three interior angles are all right angles, and hence the sum of the interior angles is $\frac{3\pi}{2}$. In particular, the angle sum is greater than π . This is quite different from the behavior of triangles in the plane.

Note that the spherical triangle in this case may be identified with the intersection of \mathbb{S}^2 with the first octant of \mathbb{R}^3 .

Of course, the three edges in a spherical triangle are segments in three distinct spherical lines. We could ask a different question: what are the spherical triangles determined by three distinct spherical lines? As this last example illustrates, the three lines divide \mathbb{S}^2 into eight triangular regions.

To see this, note that each spherical line meets each other spherical line in two points. So each spherical line contains four vertices and hence is divided into four edges. The other two lines meet each other in two vertices which are antipodes of one another. So the original line meets four triangular faces in each of the two hemispheres into which it divides \mathbb{S}^2 .

Note that each of the eight triangles is congruent to one of the others: the one whose vertices are the antipodes of its own vertices. The congruence is induced by the linear isometry $-I_3$ of \mathbb{R}^3 .

9.4. Isometries of \mathbb{S}^2 . By Theorem 9.1.16, the isometries of \mathbb{S}^2 are isomorphic as a group to the linear isometries of \mathbb{R}^3 . The isomorphism is given by restricting a linear isometry to its effect on \mathbb{S}^2 .

We have calculated the linear isometries of \mathbb{R}^3 in Chapter 7. Therefore we have everything we need to study the isometries of \mathbb{S}^2 .

We have seen in Section 7.2 that every linear isometry of \mathbb{R}^3 of determinant one is a rotation $\rho_{(u,\theta)}$ about some unit vector u by the angle θ . This rotation has two important invariant subspaces. On $\text{span}(u)$, $\rho_{(u,\theta)}$ is the identity, so that $\text{span}(u)$ is contained in the eigenspace of $(\rho_{(u,\theta)}, 1)$. On the orthogonal complement $\text{span}(u)^\perp$, $\rho_{(u,\theta)}$ is the ordinary 2-dimensional rotation by θ with respect to the orientation of $\text{span}(u)^\perp$ induced by u . Since $-u$ induces the opposite orientation on $\text{span}(u)^\perp$, we have

$$(9.4.1) \quad \rho_{(u,\theta)} = \rho_{(-u,-\theta)}.$$

All linear isometries of \mathbb{R}^n have determinant ± 1 . The linear isometries of \mathbb{R}^3 of determinant -1 fall into two families: reflections (Section 7.4) and rotation-reflections (Section 7.5).

A reflection of \mathbb{R}^3 reflects across a 2-dimensional linear subspace V . Here, if N is a unit normal to V (i.e., a unit vector in V^\perp), then the reflection in V is given by

$$\sigma_V(x) = x - 2\langle x, N \rangle N.$$

This is independent of the choice of unit normal, as $-N$ gives the same function.

Of course, $\ell = V \cap \mathbb{S}^2$ is a line in \mathbb{S}^2 and N is a pole of ℓ . So we write

$$\sigma_\ell : \mathbb{S}^2 \rightarrow \mathbb{S}^2$$

for the restriction of σ_V to \mathbb{S}^2 . Proposition 7.4.4 and Proposition 9.2.4 now give:

Proposition 9.4.1. *Let $\ell \neq m$ be lines in \mathbb{S}^2 with poles N and M , respectively. Let $u = \frac{N \times M}{\|N \times M\|}$. Then*

$$(9.4.2) \quad \sigma_m \sigma_\ell = \rho_{(u, 2\cos^{-1}(\langle N, M \rangle))},$$

the rotation about u by twice the directed angle, measured at u , from ℓ to m with respect to the orientations induced by N and M , respectively.

Finally, as discussed in Section 7.5, a rotation-reflection is a composite

$$(9.4.3) \quad \rho_{(N,\theta)} \sigma_\ell = \sigma_\ell \rho_{(N,\theta)},$$

where ℓ is a spherical line with pole N and $\theta \in (0, 2\pi)$. N being a pole of ℓ is sufficient for $\rho_{(N,\theta)}$ and σ_ℓ to commute.

Recall from Remark 7.5.4, that for any spherical line ℓ with pole N , the composite $\rho_{(N,\pi)} \sigma_\ell$ is the isometry induced by the orthogonal matrix $-I_3$, so (9.4.3) does not in general determine the line ℓ . But in all other cases, $\text{span}(N)$ is the eigenspace of the linear isometry inducing the rotation-reflection, and hence ℓ is determined by the isometry.

Recall that $\mathcal{I}(\mathbb{S}^2)$ is the group of isometries of \mathbb{S}^2 . If $X \subset \mathbb{S}^2$, we write $S(X)$ for the symmetries of X :

$$S(X) = \{\alpha \in \mathcal{I}(\mathbb{S}^2) : \alpha(X) = X\}.$$

Since linear isometries of \mathbb{R}^n preserve the inner product, we obtain the following.

Lemma 9.4.2. *Let ℓ be a line in \mathbb{S}^2 with pole N . Then*

$$(9.4.4) \quad S(\ell) = \{\alpha \in \mathcal{I}(\mathbb{S}^2) : \alpha(N) = \pm N\} = S(\{\pm N\}).$$

Corollary 9.4.3. *Let ℓ be a line in \mathbb{S}^2 with pole N . Then*

$$(9.4.5) \quad S(\ell) = \{\rho_{(N,\theta)} : \theta \in \mathbb{R}\} \cup \{\rho_{(u,\pi)} : u \in \ell\} \cup \{\sigma_m : m \perp \ell\} \\ \cup \{\sigma_\ell \rho_{(N,\theta)} : \theta \in \mathbb{R}\} \cup \{(T_{-I_3})|_{\mathbb{S}^2}\}.$$

Proof. Let $A \in \text{O}_3$. By Lemma 9.4.2, $(T_A)|_{\mathbb{S}^2} \in S(\ell)$ if and only if N is an eigenvector for A . We first consider rotations. If θ is not a multiple of 2π , then the eigenspace of $(\rho_{(u,\theta)}, 1)$ is $\text{span}(u)$, while the eigenspace of $(\rho_{(u,\theta)}, -1) = \emptyset$ unless θ is an odd multiple of π . In this last case, the eigenspace of $(\rho_{(u,\theta)}, -1)$ intersects \mathbb{S}^2 in the line whose pole is u , which contains N if and only if $u \in \ell$ by Corollary 9.2.5.

We next consider reflections. Here, for a plane V in \mathbb{R}^3 with unit normal M , the eigenspace of $(\sigma_V, 1)$ is V . This contains N if and only if $\ell \perp (V \cap \mathbb{S}^2)$ by Corollary 9.2.5. The eigenspace of $(\sigma_V, -1)$ is $\text{span}(M)$, which contains N if and only if $\ell = V \cap \mathbb{S}^2$.

Finally, we consider rotation-reflections $\alpha = \sigma_W \rho_{(M,\theta)}$ with M a unit normal for W and $\theta \in (0, \pi)$. Here, the eigenspace of $(\alpha, 1)$ is empty, and if $\theta \neq \pi$ then the eigenspace of $(\alpha, -1)$ is $\text{span}(M)$, which contains N if and only if $\ell = W \cap \mathbb{S}^2$. In the remaining case $\theta = \pi$ and hence $\alpha = T_{-I_3}$. \square

Remark 9.4.4. As shown in Example 8.6.2, \mathbb{S}^n admits an oriented atlas, and hence is an orientable manifold in the sense of Definition 10.5.2, below. It is not too difficult to show that if $A \in \text{O}_{n+1}$, then $(T_A)|_{\mathbb{S}^n}$ is orientation-preserving if $\det A = 1$, and is orientation-reversing otherwise.

9.5. Perpendicular bisectors. Recall that if ℓ is a line in \mathbb{R}^2 then ℓ is the perpendicular bisector of the line segment between x and $\sigma_\ell(x)$ for any $x \in \mathbb{R}^2 \setminus \ell$. Moreover, ℓ then consists of the set of all points equidistant from x and $\sigma_\ell(x)$. This becomes useful in studying congruences of triangles.

Here, we will establish analogous results for reflections in \mathbb{S}^2 .

Proposition 9.5.1. *Let ℓ be a line in \mathbb{S}^2 with pole N and let $u \in \mathbb{S}^2 \setminus \ell$. Then ℓ perpendicularly bisects all segments with endpoints u and $\sigma_\ell(u)$.*

Proof. The keyword ‘‘all’’ is relevant in case $u = \pm N$, and then $\sigma_\ell(u) = -u$ and there are infinitely many segments with these as endpoints. Since each one contains N , it is perpendicular to ℓ by Corollary 9.2.5. For each one of them, its point of intersection with ℓ is orthogonal (as a vector) to each of $\pm u$, as it lies on ℓ , and hence its spherical distance from each of $\pm u$ is $\cos^{-1}(0) = \frac{\pi}{2}$. Thus, it bisects the segment.

If $u \neq \pm N$, u, N are linearly independent, and hence $N \times u \neq 0$. Now,

$$(9.5.1) \quad \sigma_\ell(u) \times u = (u - 2\langle u, N \rangle N) \times u = -2\langle u, N \rangle N \times u,$$

as $u \times u = 0$. Since $u \notin \ell$, $\langle u, N \rangle \neq 0$, and hence $M = \frac{\sigma_\ell(u) \times u}{\|\sigma_\ell(u) \times u\|} \in \mathbb{S}^2$ is a pole for the spherical line m containing u and $\sigma_\ell(u)$. Moreover, by (9.5.1), M is orthogonal to N , and hence ℓ is perpendicular to m at both points of intersection. But the next result shows that these points of intersection bisect the two segments in m with endpoints u and $\sigma_\ell(u)$. \square

Proposition 9.5.2. *Let ℓ be a line in \mathbb{S}^2 with pole N and let $u \in \mathbb{S}^2 \setminus \ell$. Then*

$$(9.5.2) \quad \ell = \{v \in \mathbb{S}^2 : d(u, v) = d(\sigma_\ell(u), v)\},$$

where d is the Riemannian distance.

Proof. Of course, $d(u, v) = d(\sigma_\ell(u), v)$ if and only if $\langle u, v \rangle = \langle \sigma_\ell(u), v \rangle$. Since $\sigma_\ell(u) = u - 2\langle u, N \rangle N$, this holds if and only if $\langle v, N \rangle = 0$, which is equivalent to $v \in \ell$. \square

9.6. Exercises.

- Give the vertices of an equilateral triangle in \mathbb{S}^2 whose angle sum is greater than $\frac{5\pi}{2}$.
 - What is the angle sum?
 - What are the lengths of its sides?
- Repeat the last problem for an equilateral triangle whose angle sum is less than $\frac{3\pi}{2}$.
- Prove the side-side-side theorem in \mathbb{S}^2 . Specifically, if we have triangles $\triangle ABC$ and $\triangle DEF$ with $AB = DE$, $BC = EF$ and $AC = DF$, then there is an isometry of \mathbb{S}^2 carrying A onto D , B onto E and C onto F .
- Prove the side-angle-side theorem in \mathbb{S}^2 . Specifically, if we have triangles $\triangle ABC$ and $\triangle DEF$ with $AB = DE$, $AC = DF$ and $m(\angle BAC) = m(\angle EDF)$, then there is an isometry of \mathbb{S}^2 carrying A onto D , B onto E and C onto F .
- Prove that a triangle in \mathbb{S}^2 with two equal sides has two equal angles.
- Write $\rho_{e_1, \frac{\pi}{2}} \rho_{e_2, \frac{\pi}{2}}$ as an explicit rotation around an explicit axis.
- Find the vertices of an inscribed regular tetrahedron in \mathbb{S}^2 one of whose vertices is e_3 . Call them e_3, v_1, v_2, v_3 (note the potential connection with problem 1).
 - What are the angles in the spherical equilateral triangles formed by the vertices of this tetrahedron?
 - Find the matrix for $\rho_{(e_3, \frac{2\pi}{3})}$ with respect to the standard basis and show it permutes the vertices v_1, v_2, v_3 .
 - Find the matrix for $\rho_{(v_1, \frac{2\pi}{3})}$ with respect to the standard basis and show it permutes the vertices e_3, v_2, v_3 .
 - Use these matrices to compute the composite $\rho_{(e_3, \frac{2\pi}{3})} \circ \rho_{(v_1, \frac{2\pi}{3})}$. What rotation is the composite? Write it in the form $\rho_{(w, \theta)}$ for specific w and θ .

- (e) Let σ_V be the reflection that interchanges e_3 and v_3 . Find the matrix of σ_V and show it permutes the vertices v_1, v_2 .
- (f) Let σ_W be the reflection that interchanges e_3 and v_2 . Find the matrix of σ_W and show it permutes the vertices v_1, v_3 .
- (g) Find the matrix of the composite $\sigma_V \circ \sigma_W$. What rotation does it represent? Write it in the form $\rho_{(w,\theta)}$ for specific w and θ .
- (h) Let $\beta = \rho_{(e_3, \frac{2\pi}{3})} \circ \sigma_V$ with V as above.
 - (i) Show that β is a rotation-reflection. Write it in the standard form $\beta = \rho_{(w,\theta)} \circ \sigma_Z$, where $Z = \{w\}^\perp$. (I.e., please specify w and θ . Z is then implicitly defined.)
 - (ii) Calculate the effect of β on the vertices e_3, v_1, v_2, v_3 .
 - (iii) What rotation is β^2 ? (Write it in the form $\rho_{(u,\phi)}$ for specific u and ϕ .) What is β^4 ?

10. Tangent bundles

The tangent bundle is the key concept underlying both differential topology and differential geometry. It is the underlying concept behind orientations and also behind the notion of straight lines in spherical and hyperbolic geometry.

The basic idea is this:

10.1. The local model. Let $U \subset \mathbb{R}^n$ be open. Then we can think of $U \times \mathbb{R}^n$ as a representation space for the tangent vectors of smooth curves in U . In particular, if $\gamma : (a, b) \rightarrow U$ is smooth, then for each $t \in U$, the pair $(\gamma(t), \gamma'(t))$ lies in $U \times \mathbb{R}^n$ and represents the tangent vector to γ at $\gamma(t)$: the tangent line at $\gamma(t)$ to the image of γ is the line $\gamma(t) + \text{span}(\gamma'(t))$. So the pair $(\gamma(t), \gamma'(t))$ does specify the tangent line. Moreover, every point in $U \times \mathbb{R}^n$ arises this way.

We take this as a local model for the tangent space of a smooth manifold. Specifically, we write $TU = U \times \mathbb{R}^n$ and write $\pi : TU \rightarrow U$ for the projection onto the first factor. In particular, $\pi^{-1}(x) = \{x\} \times \mathbb{R}^n$ is called the fibre over x and has the structure of a vector space of dimension $\dim U$. We call $U \times \mathbb{R}^n$ the tangent space of U and call π the tangent bundle of U . We write $T_x U$ for $\pi^{-1}(x)$ and call it the tangent space of U at x . We suppress the $\{x\}$ and write $T_x U = \mathbb{R}^n$.

Given open sets $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^k$ and a smooth map $f : U \rightarrow V$, we define the tangent map $Tf : TU \rightarrow TV$ by

$$(10.1.1) \quad Tf(x, y) = (f(x), Df(x)y).$$

Thus, the following diagram commutes:

$$\begin{array}{ccc} TU & \xrightarrow{Tf} & TV \\ \pi \downarrow & & \downarrow \pi \\ U & \xrightarrow{f} & V. \end{array}$$

In particular, Tf takes the tangent space of U at x to the tangent space of V at $f(x)$ by the linear transformation

$$(10.1.2) \quad T_x f : \mathbb{R}^n \rightarrow \mathbb{R}^k$$

induced by the Jacobian matrix $Df(x)$.

If $g : V \rightarrow W$ is smooth with $W \subset \mathbb{R}^m$, open, then $T(g \circ f) = Tg \circ Tf$ by the chain rule:

$$\begin{aligned} Tg \circ Tf(x, y) &= Tg(f(x), Df(x)y) = (g \circ f(x), [Dg(f(x))Df(x)]y) \\ &= (g \circ f(x), D[g \circ f](x)y) = T[g \circ f](x, y). \end{aligned}$$

Example 10.1.1. Let $\gamma : (a, b) \rightarrow U \subset \mathbb{R}^n$ be a smooth curve. The tangent space of (a, b) is $(a, b) \times \mathbb{R}$, as above, and for $(t, s) \in (a, b) \times \mathbb{R}$,

$$T\gamma(t, s) = (\gamma(t), \gamma'(t)s).$$

In particular, $T_t\gamma : \mathbb{R} \rightarrow \mathbb{R}^n$ is multiplication by the $n \times 1$ matrix $\gamma'(t)$. So the image of $T_t\gamma$ is $\text{span}(\gamma'(t))$. Moreover, $\gamma'(t) = T_t\gamma(1)$, the image of $1 \in \mathbb{R}$ under $T_t\gamma : \mathbb{R} \rightarrow \mathbb{R}^n$.

In the above example, t varies throughout (a, b) . Let's now fix $x \in U$ and insist $\gamma : (-\epsilon, \epsilon) \rightarrow U$ with $\gamma(0) = x$.

Lemma 10.1.2. *Let $x \in U \subset \mathbb{R}^n$ and let $v \in \mathbb{R}^n = T_xU$. Then there is an $\epsilon > 0$ and a smooth curve $\gamma : (-\epsilon, \epsilon) \rightarrow U$ with $\gamma(0) = x$ and $\gamma'(0) = T_0\gamma(1) = v$.*

Proof. Set $\gamma(t) = x + tv$. For ϵ small enough, the image of γ lies in U . \square

By the chain rule, if $f : U \rightarrow V$ is smooth, with $V \subset \mathbb{R}^k$, open, and if $v = T_0\gamma(1) \in T_xU$, then $T_x f(v) = T_0(f \circ \gamma)(1)$. Thus, we may use smooth curves to calculate $T_x f$. In fact, this is exactly what we did in Exercises 4–6 of Chapter 8.

10.2. The tangent bundle of a smooth manifold. Let \mathcal{A} be an atlas for the smooth n -manifold M . Then for each $h : U \xrightarrow{\cong} h(U) \subset \mathbb{R}^n$ in \mathcal{A} we can use $T(h(U)) = h(U) \times \mathbb{R}^n$ as a model for TU . We can then assemble these local models into a single space as follows: let

$$\tilde{T}M = \coprod_{h \in \mathcal{A}} h(U) \times \mathbb{R}^n,$$

The disjoint union over all the charts $h : U \xrightarrow{\cong} h(U)$ of the tangent spaces of the open subsets $h(U) \subset \mathbb{R}^n$. We put an equivalence relation \sim on $\tilde{T}M$ as follows. If $x \in U \cap V$ and if $h : U \xrightarrow{\cong} h(U)$ and $k : V \xrightarrow{\cong} k(V)$ are charts in \mathcal{A} , we set $(h(x), v) \in h(U) \times \mathbb{R}^n$ equivalent to $(k(x), Dg_{kh}(h(x))v) \in k(V) \times \mathbb{R}^n$. Here, g_{kh} is the transition map (8.4.1) from h to k .

The tangent space of M is given by

$$TM = \tilde{T}M / \sim = \left(\coprod_{h \in \mathcal{A}} h(U) \times \mathbb{R}^n \right) / \sim.$$

In Proposition A.7.7 below, we show that if \sim is the equivalence relation on $\coprod_{h \in \mathcal{A}} h(U)$ given by setting $h(x) \sim k(x)$ for $x \in U \cap V$ as above, then there is a homeomorphism

$$(10.2.1) \quad \bar{\eta} : \left(\coprod_{h \in \mathcal{A}} h(U) \right) / \sim \xrightarrow{\cong} M$$

that restricts on each $h(U)$ to h^{-1} , i.e., $\bar{\eta}(h(x)) = x$ for $h(x)$ in the image of the standard inclusion $h(U) \subset \coprod_{h \in \mathcal{A}} h(U)$.

We can relate these as follows. Since equivalence in $\tilde{T}(M)$ implies equivalence in $\coprod_{h \in \mathcal{A}} h(U)$, there is a commutative diagram

$$(10.2.2) \quad \begin{array}{ccc} \coprod_{h \in \mathcal{A}} h(U) \times \mathbb{R}^n & \longrightarrow & (\coprod_{h \in \mathcal{A}} h(U) \times \mathbb{R}^n) / \sim \\ \downarrow \pi & & \downarrow \tilde{\pi} \\ \coprod_{h \in \mathcal{A}} h(U) & \longrightarrow & (\coprod_{h \in \mathcal{A}} h(U)) / \sim, \end{array}$$

where the horizontal maps are the canonical maps for the equivalence relations in question, the left hand vertical is the disjoint union of the projections onto the first factor, and $\tilde{\pi}$ takes the equivalence class of $(h(x), v) \in h(U) \times \mathbb{R}^n$ to the equivalence class of $(h(x)) \in h(U)$.

The upper right corner of (10.2.2) is by definition TM . Define $\pi : TM \rightarrow M$ to be the composite $\pi = \bar{\eta} \circ \tilde{\pi}$, with $\bar{\eta}$ the mapping in (10.2.1). We refer to $\pi : TM \rightarrow M$ as the tangent bundle of M and refer to $\pi^{-1}(x) \subset TM$ as $T_x M$, the tangent space of M at $x \in M$. We also call it the fibre of the tangent bundle over x .

We have constructed a commutative diagram

$$(10.2.3) \quad \begin{array}{ccccc} h(U) \times \mathbb{R}^n & \xrightarrow{\hat{i}} & \left(\coprod_{h \in \mathcal{A}} h(U) \times \mathbb{R}^n \right) / \sim & = & TM \\ \downarrow \pi & & \downarrow \tilde{\pi} & & \downarrow \pi \\ h(U) & \xrightarrow{\iota} & \left(\coprod_{h \in \mathcal{A}} h(U) \right) / \sim & \xrightarrow[\cong]{\bar{\eta}} & M, \\ & \searrow & \text{---} & \nearrow & \\ & & h^{-1} & & \end{array}$$

where \hat{i} is induced by the standard inclusion $h(U) \times \mathbb{R}^n \subset \coprod_{h \in \mathcal{A}} h(U) \times \mathbb{R}^n$.

Proposition 10.2.1. *The map $\hat{i} : h(U) \times \mathbb{R}^n \rightarrow TM$ of (10.2.3) is a homeomorphism onto $\pi^{-1}U \subset TM$. Moreover, the following diagram commutes:*

$$(10.2.4) \quad \begin{array}{ccc} h(U) \times \mathbb{R}^n & \xrightarrow[\cong]{\hat{i}} & \pi^{-1}U \\ \pi \downarrow & & \downarrow \pi \\ h(U) & \xrightarrow[\cong]{h^{-1}} & U. \end{array}$$

In particular, \hat{i} restricts to a homeomorphism

$$(10.2.5) \quad \hat{i} : \{h(x)\} \times \mathbb{R}^n = T_{h(x)}U \xrightarrow{\cong} T_x M.$$

Proof. The equivalence relation defining TM does not identify distinct elements of $h(U) \times \mathbb{R}^n$, so \hat{i} is one-to-one. Moreover, if $x \in U$, then any element of $\pi^{-1}(x)$ is an equivalence class of the form $[(k(x), v)]$ with

$(k(x), v) \in k(V) \times \mathbb{R}^n$ for a chart $k : V \xrightarrow{\cong} k(V) \subset \mathbb{R}^n$ about x . But

$$(k(x), v) \sim (h(x), Dg_{hk}(k(x))v) \in h(U) \times \mathbb{R}^n,$$

so $\hat{\iota}$ maps onto $\pi^{-1}(U)$ as claimed. It suffices to show $\hat{\iota}$ is an open map. To see this, write

$$p : \coprod_{h \in \mathcal{A}} h(U) \times \mathbb{R}^n \rightarrow \left(\coprod_{h \in \mathcal{A}} h(U) \times \mathbb{R}^n \right) / \sim$$

for the canonical map taking each element to its equivalence class. (We call it p to avoid too many maps π .) We wish to show that for W open in $h(U) \times \mathbb{R}^n$, $p^{-1}p(W)$ is open in the disjoint union. Now, the intersection of $p^{-1}p(W)$ with $k(V) \times \mathbb{R}^n$ is the image of $W \cap (h(U \cap V) \times \mathbb{R}^n)$ under the map

$$(10.2.6) \quad \begin{aligned} \tilde{g}_{kh} : h(U \cap V) \times \mathbb{R}^n &\rightarrow k(U \cap V) \times \mathbb{R}^n \\ (h(x), v) &\mapsto (k(x), Dg_{kh}(h(x))v) \\ &= (g_{kh}(h(x)), Dg_{kh}(h(x))v). \end{aligned}$$

This map \tilde{g}_{kh} is in fact smooth if we regard its domain and codomain as open subsets of $\mathbb{R}^n \times \mathbb{R}^n = \mathbb{R}^{2n}$, as matrix multiplication is polynomial in the coefficients of the matrix and vector in question.

But that shows \tilde{g}_{kh} to be a diffeomorphism, as its inverse function is \tilde{g}_{hk} . Since a diffeomorphism is an open map, we are done. \square

In fact, we have also just shown:

Proposition 10.2.2. *TM is a smooth manifold with the charts given by*

$$\bar{h} = \hat{\iota}^{-1} : \pi^{-1}(U) \xrightarrow{\cong} h(U) \times \mathbb{R}^n \subset \mathbb{R}^{2n}.$$

The transition maps $g_{\bar{k}\bar{h}}$ are precisely the maps \tilde{g}_{kh} of (10.2.6). The tangent bundle projection map $\pi : TM \rightarrow M$ is smooth.

Note that (10.2.5) provides $T_x M$ with the structure of an n -dimensional vector space. In fact, it provides a basis, coming from the canonical basis of $\mathbb{R}^n = T_{h(x)}U$. To emphasize the dependence on the particular chart h , write

$$(10.2.7) \quad T_{h(x)}h^{-1} : T_{h(x)}U \xrightarrow{\cong} T_x M$$

for (10.2.5). This notation is meant to be compatible with (10.1.2). We will expand on it below.

If we vary the chart neighborhood to a chart V containing x , by (10.2.6), we get a commutative diagram

$$(10.2.8) \quad \begin{array}{ccc} \mathbb{R}^n & \xrightarrow{Dg_{kh}(h(x))} & \mathbb{R}^n \\ & \cong & \\ T_{h(x)}h^{-1} & \searrow & \swarrow T_{k(x)}k^{-1} \\ & T_x M & \end{array}$$

Since $Dg_{kh}(h(x))$ is a linear isomorphism, the induced vector space structure on T_xM is independent of the choice of chart. Moreover, if each $Dg_{kh}(h(x))$ has positive determinant, then different charts induce the same orientation on T_xM . We obtain:

Proposition 10.2.3. *An oriented atlas for M provides a linear orientation for each T_xM compatible with the isomorphisms (10.2.7).*

We now associate a map of tangent spaces to any smooth map.

Definition 10.2.4. Let M be a smooth n -manifold and N a smooth m -manifold. Let $f : M \rightarrow N$ be a smooth map. Then $Tf : TM \rightarrow TN$ is the smooth map defined as follows. Given smooth charts

$$\begin{aligned} h : U &\xrightarrow{\cong} h(U) \subset \mathbb{R}^n, \\ k : V &\xrightarrow{\cong} k(V) \subset \mathbb{R}^m, \end{aligned}$$

for M and N , respectively, Tf is defined on $\pi^{-1}(U \cap f^{-1}(V))$ to be the composite

$$\pi^{-1}(U \cap f^{-1}(V)) \xrightarrow{\bar{h}} h(U \cap f^{-1}(V)) \times \mathbb{R}^n \xrightarrow{f_{\bar{k}\bar{h}}} k(V) \times \mathbb{R}^m \xrightarrow{\bar{k}^{-1}} \pi^{-1}(V),$$

where $f_{\bar{k}\bar{h}}(h(x), v) = (kf(x), Df_{kh}(x)v)$, with f_{kh} the map of (8.4.2). This is easily seen to be compatible with the change of chart neighborhoods, and defines a smooth map $Tf : TM \rightarrow TN$ such that the following diagram commutes:

$$\begin{array}{ccc} TM & \xrightarrow{Tf} & TN \\ \pi \downarrow & & \downarrow \pi \\ M & \xrightarrow{f} & N. \end{array}$$

Moreover, the map $(Tf)_{\bar{k}\bar{h}}$ of (8.4.2) is just the map $f_{\bar{k}\bar{h}}$ above. Note that the map $T_x f : T_x M \rightarrow T_{f(x)} N$ obtained by restricting Tf is given in local coordinates by multiplication by the matrix $Df_{kh}(x)$ and hence is linear. In other words, $T_x f$ is the linear map whose matrix with respect to the bases provided by the charts h and k is $Df_{kh}(x)$.

Thus, the tangent map is a coordinate-free way of expressing the derivative of a smooth function f . The following is immediate from the chain rule.

Lemma 10.2.5. *Let $f : M \rightarrow M'$ and $g : M' \rightarrow M''$ be smooth maps of smooth manifolds. Then $T(g \circ f) = Tg \circ Tf$.*

Since $T(\text{id}_M)$ is the identity map of TM we obtain the following.

Corollary 10.2.6. *Let $f : M \rightarrow N$ be a diffeomorphism. Then Tf is a diffeomorphism and $T_x f : T_x M \rightarrow T_{f(x)} N$ is a linear isomorphism for each $x \in M$.*

Example 10.2.7. If M^n is a smooth submanifold of N^{n+k} and if $i : M \rightarrow N$ is the inclusion, then for each $x \in M$ we can choose a chart $k : V \rightarrow \mathbb{R}^{n+k}$ for N at x with $k^{-1}(\mathbb{R}^n \times 0) = V \cap M$, hence $h = k_{V \cap M} : V \cap M \rightarrow \mathbb{R}^n$ is a chart for M at x , and the Jacobian matrix Di_{kh} is just the matrix of the standard inclusion of \mathbb{R}^n in \mathbb{R}^{n+k} . Thus, $T_x i : T_x M \rightarrow T_x N$ is the inclusion of an n -dimensional linear subspace.

It is very useful to express the tangent map $T_x f$ in terms of tangents to curves in M . The following is consistent with the local model.

Definition 10.2.8. Let $\gamma : (a, b) \rightarrow M$ be a smooth curve in M . We define the tangent vector to γ at $t \in (a, b)$ by

$$(10.2.9) \quad \gamma'(t) = T_t \gamma(1) \in T_{\gamma(t)} M$$

The chain rule gives:

Lemma 10.2.9. Let $\gamma : (a, b) \rightarrow M$ be a smooth curve and let $f : M \rightarrow N$ be a smooth map. Then

$$(10.2.10) \quad T_{\gamma(t)} f(\gamma'(t)) = (f \circ \gamma)'(t)$$

for all $t \in (a, b)$.

Note that if $h : U \xrightarrow{\cong} h(U) \subset \mathbb{R}^n$ is a chart for M then $h^{-1} : h(U) \rightarrow U \subset M$ is smooth and the map \hat{i} of (10.2.4) is precisely Th^{-1} , thus justifying the notation (10.2.7). In fact, $h^{-1} : h(U) \rightarrow U$ is a diffeomorphism. (We note here that any open subset of a smooth manifold has a smooth structure induced by restriction of charts.) In particular, if $x \in U$ then a curve $\gamma : (-\epsilon, \epsilon) \rightarrow U$ with $\gamma(0) = x$ is smooth if and only if $h \circ \gamma : (-\epsilon, \epsilon) \rightarrow h(U)$ is smooth. So we can use the analysis of the tangents to curves in the local model to study $T_x M$. In particular, the following is immediate from Lemma 10.1.2.

Corollary 10.2.10. Let M be a smooth n -manifold. Let $x \in M$ and $v \in T_x M$. Then there is an $\epsilon > 0$ and a smooth curve $\gamma : (-\epsilon, \epsilon) \rightarrow M$ with $\gamma(0) = x$ and $\gamma'(0) = v$.

We can use this to study the tangent space of a regular hypersurface $S \subset \mathbb{R}^{n+1}$. Here, we are given a smooth map $f : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ and a regular value $y \in \mathbb{R}$ for f , meaning that for all $x \in f^{-1}(y)$, the Jacobian matrix $Df(x) = \nabla f(x)$ is nonzero (and hence of rank 1, as it is a row matrix). In this case $S = f^{-1}(y)$ is called a regular hypersurface in \mathbb{R}^{n+1} and is a smooth submanifold by Corollary 10.4.10 below.

If $i : S \subset \mathbb{R}^{n+1}$ is the inclusion of a regular hypersurface, we can use Corollary 10.2.10 to study $T_x i : T_x S \rightarrow T_x \mathbb{R}^{n+1} = \mathbb{R}^{n+1}$ for each $x \in S$, and to identify exactly which n -dimensional linear subspace of \mathbb{R}^{n+1} is the tangent space to S at x .

Proposition 10.2.11. *Let $S \subset \mathbb{R}^{n+1}$ be the regular hypersurface induced by the smooth map $f : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$, with $S = f^{-1}(y)$ for a regular value y of f . Then for $x \in S$, the tangent space $T_x S$ is the nullspace of the row matrix $\nabla f(x)$. (Implicitly, we have identified $T_x S$ with its image under $T_x i$, with i the inclusion of S in \mathbb{R}^{n+1} .)*

An equivalent description may be given as follows. Let $v(x)$ be the transpose of $\nabla f(x)$, so that $v(x)$ is a vector in \mathbb{R}^{n+1} , and is nonzero since y is a regular value. Let $N(x) = \frac{v(x)}{\|v(x)\|}$. Then $T_x(S)$ is the set of vectors orthogonal to $N(x)$, i.e.,

$$T_x S = \text{span}(N(x))^\perp.$$

Since f is smooth, so is $N : S \rightarrow \mathbb{R}^{n+1}$. We call it the unit normal to S induced by f . Note that the unit normal induced by $-f$ is $-N$, which gives the other unit normal to each $T_x S$.

Proof. Let $w \in T_x S$. Then $T_x i(w) \in T_x \mathbb{R}^{n+1} = \mathbb{R}^{n+1}$. We wish to show the matrix product

$$\nabla f(x) \cdot T_x i(w) = 0.$$

Let $\gamma : (-\epsilon, \epsilon) \rightarrow S$ be a smooth curve with $\gamma(0) = x$ and $\gamma'(0) = w$. By Lemma 10.2.9 and Example 10.1.1, $T_x i(w)$ is the standard velocity vector $\gamma'(0)$ to $\gamma : (a, b) \rightarrow \mathbb{R}^{n+1}$. But

$$\nabla f(x) \cdot \gamma'(0) = Df(\gamma(0)) \cdot \gamma'(0) = D(f \circ \gamma)(0) = (f \circ \gamma)'(0)$$

by the chain rule. But since the image of γ is contained in $S = f^{-1}(y)$, $f \circ \gamma$ is constant, so its derivative is 0, as desired.

Thus we have shown that $T_x(S)$ is contained in the nullspace of $\nabla f(x)$. Since S is an n -manifold, $T_x S$ is an n -dimensional subspace of \mathbb{R}^{n+1} . Since $\nabla f(x)$ is a nonzero $1 \times (n+1)$ row matrix, its nullspace also has dimension n , so the two subspaces must be equal.

The equivalent formulation just uses that $\nabla f(x) \cdot z = \langle v, z \rangle$ for any vector $z \in \mathbb{R}^{n+1}$. \square

Corollary 10.2.12. *Let $v \in \mathbb{S}^n$. Then the tangent space of \mathbb{S}^n at v is*

$$\text{span}(v)^\perp = \{v\}^\perp \subset \mathbb{R}^{n+1}.$$

Proof. \mathbb{S}^n is the regular hypersurface induced by $f : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$, $f(x) = \langle x, x \rangle$. Here $N(x) = x$ for $x \in \mathbb{S}^n$. \square

10.3. Tangent bundles of products. Let M and N be smooth manifolds. As shown in Section 8.5, the product $M \times N$ has a smooth structure whose charts are given by the product of a chart for M and a chart for N . In particular, the projection maps $\pi_1 : M \times N \rightarrow M$ and $\pi_2 : M \times N \rightarrow N$ are smooth submersions. The tangent maps $T\pi_1 : T(M \times N) \rightarrow TM$ and $T\pi_2 : T(M \times N) \rightarrow TN$ then give the coordinate functions for a map

$(T\pi_1, T\pi_2) : T(M \times N) \rightarrow TM \times TN$ making the following diagram commute:

$$(10.3.1) \quad \begin{array}{ccc} T(M \times N) & \xrightarrow{(T\pi_1, T\pi_2)} & TM \times TN \\ & \searrow \pi & \swarrow \pi \times \pi \\ & M \times N & \end{array}$$

Here, the maps π are the tangent bundles of $M \times N$, M and N , respectively. Under the assembly procedure (10.2.3), the diagram (10.3.1) is given locally by

$$(10.3.2) \quad \begin{array}{ccc} h(U) \times k(V) \times \mathbb{R}^{m+n} & \xrightarrow{\alpha} & (h(U) \times \mathbb{R}^m) \times (k(V) \times \mathbb{R}^n) \\ & \searrow \pi & \swarrow \pi \times \pi \\ & h(U) \times k(V) & \end{array}$$

where $\alpha(u, v, (x_1, \dots, x_{m+n})) = (u, (x_1, \dots, x_m), v, (x_{m+1}, \dots, x_{m+n}))$. In particular, α is a diffeomorphism and induces a linear isomorphism

$$T_{(u,v)}(M \times N) \rightarrow T_u M \oplus T_u N$$

on fibres. We can now assemble this to get global information:

Proposition 10.3.1. *Let M and N be smooth manifolds. Then the map $(T\pi_1, T\pi_2)$ of (10.3.1) is a diffeomorphism and induces an isomorphism of vector spaces*

$$(10.3.3) \quad T_{(u,v)}(M \times N) \rightarrow T_u M \oplus T_u N.$$

Let P be a smooth manifold and let $f : P \rightarrow M$ and $g : P \rightarrow N$ be smooth. Consider the smooth map

$$(f, g) : P \rightarrow M \times N$$

of Proposition 8.5.1. Then taking (10.3.3) as an identification, we have

$$T_x(f, g) = (T_x f, T_x g)$$

(in the notation of Proposition 1.7.4) for all $x \in P$.

10.4. Immersions and embeddings; submersions. We can now define immersions and submersions between smooth manifolds.

Definition 10.4.1. Let $f : M \rightarrow N$ be a smooth map from the m -manifold M to the n -manifold N . Then f is a smooth immersion at x if T_x has rank m and is a smooth submersion at x if T_x has rank n . We say f is a smooth immersion if it is so at each of the points in M , and similarly for submersions.

Note that if f is a smooth immersion then $m \leq n$ and if f is a smooth submersion then $m \geq n$. When $m = n$, then f is a smooth immersion if and only if it is a smooth submersion. This motivates another definition.

Definition 10.4.2. Let $f : M \rightarrow N$ be a smooth map from an m -manifold to an n -manifold. Then the codimension, $\text{codim } f$, of f is $n - m$.

Definition 10.4.3. A smooth embedding $f : M \rightarrow N$ is a one-to-one smooth immersion such that $f : M \xrightarrow{\cong} f(M)$ is a homeomorphism of M onto the image of f .

The inverse function theorem gives:

Corollary 10.4.4. *Let M be a smooth manifold. Then a one-to-one codimension 0 smooth immersion $f : N \rightarrow M$ is a diffeomorphism onto an open submanifold of M . Thus, a one-to-one codimension 0 smooth immersion is a smooth embedding.*

Proof. It suffices to consider a one-to-one codimension 0 smooth immersion $f : N \rightarrow M$. Let $k : V \rightarrow \mathbb{R}^n$ be a smooth chart of M and consider the maps f_{kh} of (8.4.2). By assumption, they have rank n and hence are invertible. By the inverse function theorem, the image of f_{kh} is open in $k(V)$. Taking the union over all charts of N and applying k^{-1} , we see that $f(N) \cap V$ is open in V and hence in M .

Taking the union over all chart neighborhoods V in M we see $f(N)$ is open in M . By the inverse function theorem, the inverse function $f^{-1} : f(N) \rightarrow N$ is smooth in each chart, and hence is smooth. Here, we give $f(N)$ the smooth structure given by restricting the charts of M to the open set $f(N)$. We see that $f : N \rightarrow f(N)$ is a diffeomorphism. \square

Smooth immersions in higher codimension can be more complicated.

Example 10.4.5. Let $f : (-3, -1) \cup (0, 1) \rightarrow \mathbb{R}^2$ be given by

$$f(x) = \begin{cases} (0, x + 2) & x \in (-3, -1), \\ (x, \sin \frac{1}{x}) & x \in (0, 1). \end{cases}$$

Then Df is never 0, so f is an immersion. Despite being one-to-one, f is not an embedding, as every point in $f((-3, -1))$ is the limit of a sequence of points in $f((0, 1))$, so f^{-1} is not continuous on the image of f . The closure in \mathbb{R}^2 of image of f is sometimes called the topologist's sine curve.

On the good side, here are two consequences of the inverse function theorem, whose proofs may be found in [2].

Theorem 10.4.6 (See [2, Theorem 7.1]). *Let $f : N^n \rightarrow M^m$ be smooth and an immersion at $x \in N$. Then there are smooth charts $h : U \xrightarrow{\cong} \mathbb{R}^n$ about x and $k : V \xrightarrow{\cong} \mathbb{R}^m$ about $f(x)$ such that the following diagram commutes:*

$$\begin{array}{ccc} U & \xrightarrow{f} & V \\ h \downarrow & & \downarrow k \\ \mathbb{R}^n & \xrightarrow{\iota} & \mathbb{R}^m, \end{array}$$

where ι is the standard inclusion of \mathbb{R}^n in \mathbb{R}^m :

$$\iota(x_1, \dots, x_n) = (x_1, \dots, x_n, 0, \dots, 0).$$

Note that what goes wrong in Example 10.4.5 is that for points $x \in (-3, -1)$ we cannot have $U = f^{-1}(V)$, as every point in $f((-3, -1))$ is a limit of points in $f((0, 1))$.

Theorem 10.4.7 (See [2, Theorem 7.3]). *Let $f : N^n \rightarrow M^m$ be smooth and a submersion at $x \in N$. Then there are smooth charts $h : U \xrightarrow{\cong} \mathbb{R}^n$ about x and $k : V \xrightarrow{\cong} \mathbb{R}^m$ about $f(x)$ such that $f(U) = V$ and the following diagram commutes:*

$$\begin{array}{ccc} U & \xrightarrow{f} & V \\ h \downarrow & & \downarrow k \\ \mathbb{R}^n & \xrightarrow{\pi} & \mathbb{R}^m, \end{array}$$

where π is the projection onto the first m coordinates.

There are some very nice consequences of these results.

Corollary 10.4.8. *Let $f : N \rightarrow M$ be a smooth embedding. Then $f(N)$ is a smooth submanifold of M and $f : N \rightarrow f(N)$ is a diffeomorphism.*

Proof. Let h and h be the charts from Theorem 10.4.6. Since $f : N \rightarrow f(N)$ is a homeomorphism, $f(U)$ is open in $f(N)$, so there is an open set W in M with $W \cap f(N) = f(U)$. Now cut down the chart k to $V \cap W$ and it satisfies the requirements of Definition 8.4.13. \square

A nice application of Theorem 10.4.7 comes from regular values.

Definition 10.4.9. Let $f : N \rightarrow M$ be smooth. An element $y \in M$ is a regular value for f if f is a submersion at every point in $f^{-1}(y)$.

Corollary 10.4.10. *Let y be a regular value for $f : N^n \rightarrow M^m$. Then $f^{-1}(y)$ is a smooth submanifold of dimension $n - m$.*

Proof. Let h and k be the charts given by Theorem 10.4.7. Then

$$h|_{h^{-1}(0 \times \mathbb{R}^{n-m})} : h^{-1}(0 \times \mathbb{R}^{n-m}) \xrightarrow{\cong} 0 \times \mathbb{R}^{n-m}$$

gives a chart for $f^{-1}(y)$ at x . Here, of course, $0 \times \mathbb{R}^{n-m}$ is the set of points in \mathbb{R}^n whose first m coordinates are 0. \square

As shown in Example 8.1.8, 1 is a regular value for the map $f : \mathbb{R}^n \rightarrow \mathbb{R}$ given by $f(x) = \langle x, x \rangle$. This gives another proof that $\mathbb{S}^{n-1} = f^{-1}(1)$ is a smooth submanifold of \mathbb{R}^n .

10.5. Orientation of manifolds.

Definition 10.5.1. An orientation for a smooth n -manifold is a choice of linear orientation for each $T_x M$ coming from some oriented atlas of M . If M admits such an atlas, we say M is orientable. We can then obtain the opposite orientation by composing each chart with an orientation-reversing linear isomorphism of \mathbb{R}^n .

A manifold M together with an orientation is said to be oriented. A codimension 0 smooth immersion $f : M \rightarrow N$ between oriented manifolds is then said to be orientation-preserving if $T_x f$ is orientation-preserving for all $x \in M$ and to be orientation-reversing if $T_x f$ is orientation-reversing for all $x \in M$.

Remark 10.5.2. As shown in Example 8.6.2, the sphere \mathbb{S}^n is orientable for $n \geq 1$. The example provides a specific orientation. We are not so lucky with the Klein bottle or the real projective space $\mathbb{R}P^2$, which are not orientable. Each of them contains an open Möbius band as an open subset. The Möbius band is often used as an intuitive illustration of nonorientability. (The open band is obtained by removing the boundary circle from the usual Möbius band.) It is easy to see that any open subset of an orientable manifold is orientable, as the restriction of an oriented atlas to an open submanifold is still oriented.

Remark 10.5.3. Let M and N be oriented smooth manifolds and let $f : M \rightarrow N$ be a codimension 0 smooth immersion. For $x \in M$, $T_x f$ is a linear isomorphism between oriented vector spaces, but we do not have preferred bases for either $T_x M$ or $T_{f(x)} N$. Instead, we have choices of basis, compatible with the orientations, coming from any choices of charts U about x and V about $f(x)$. These charts are assumed to come from our chosen oriented atlases of M and N , so if we choose different charts, the matrix for $T_x f$ coming from the new bases differs from the old one by multiplication by matrices of positive determinant. Thus, while $\det T_x f$ is not well-defined, its sign is well-defined. So we shall feel free discussing the “sign of $\det T_x f$ ” in the discussion below.

As shown in Example 8.2.3, a codimension 0 immersion may be neither orientation-preserving nor orientation-reversing, but if M is path-connected and M and N are oriented, then it must be one or the other:

Proposition 10.5.4. *Let $f : M \rightarrow N$ be a codimension 0 smooth immersion of oriented manifolds with M path-connected. Then f is either orientation-preserving or orientation-reversing.*

Proof. We wish to show the sign of $\det T_x f$ is independent of the choice of $x \in M$. Since M is path-connected, we are asking that if x and y are connected by a path in M then $\det T_x f$ and $\det T_y f$ have the same sign. If

M and N are open subsets of \mathbb{R}^n , this follows from the proof of Proposition 8.2.7. Thus, $\det T_x f$ and $\det T_y f$ have the same sign if they are connected by a path in a chart neighborhood U with the property that $f(U)$ is contained in a chart neighborhood V of N . Note that such U form an open cover of M .

We now apply a standard technique in topology called the Lebesgue number. Let $\gamma : [a, b] \rightarrow M$ be a path from x to y and consider the sets $\gamma^{-1}(U)$, where U is a chart neighborhood in M such that $f(U)$ is contained in a chart neighborhood of N . This forms an open cover of the closed interval $[a, b]$. By [8, Theorem XI.4.5], this cover has a Lebesgue number λ , with the property that any subinterval of $[a, b]$ whose width is less than λ is carried into a set in this cover.

Let $\frac{b-a}{n} < \lambda$, and let $x_k = a + k\frac{b-a}{n}$ for $k \in \{0, \dots, n\}$. Then $x_0 = x$, $x_n = y$ and for each $k \in \{0, \dots, n-1\}$, $\gamma([x_k, x_{k+1}])$ is contained in a chart neighborhood U of M with the property that $f(U)$ is contained in a chart neighborhood V of N . By the above, $\det T_{x_k} f$ and $\det T_{x_{k+1}} f$ have the same sign. So inductively, $\det T_x f$ and $\det T_y f$ have the same sign. \square

10.6. Vector fields.

Definition 10.6.1. A section of a function $f : X \rightarrow Y$ is a function $s : Y \rightarrow X$ such that $f \circ s$ is the identity map of Y . If X and Y are topological spaces and f is continuous, a continuous section is simply a section $s : Y \rightarrow X$ that is continuous. Similarly, if f is a smooth map between manifolds, we can ask that a section be smooth.

A vector field on a smooth manifold M is a smooth section of its tangent bundle $\pi : TM \rightarrow M$.

11. Riemannian manifolds

A smooth manifold, by itself, has topology but not geometry. For instance, the upper half plane $\mathbb{H} = \{[\frac{x}{y}] \in \mathbb{R}^2 : y > 0\}$ can be given a hyperbolic geometry that realizes the axioms of “non-Euclidean” geometry. But \mathbb{H} is diffeomorphic (i.e., equivalent as a smooth manifold) to \mathbb{R}^2 , which has the standard Euclidean geometry.¹⁷ The study of manifolds up to diffeomorphism is called differential topology.

We derived the standard Euclidean geometry of \mathbb{R}^2 from the standard inner product on \mathbb{R}^2 . We can derive the hyperbolic geometry on \mathbb{H} via a different inner product that varies from point to point.

11.1. Riemannian metrics. Recall that if M is a smooth n -manifold, then the tangent space $T_x(M)$ of M at a point $x \in M$ is a real vector space of dimension n , and if $h : U \xrightarrow{\cong} h(U) \subset \mathbb{R}^n$ is a smooth chart about x , then

$$T_{h(x)}h^{-1} : \mathbb{R}^n \rightarrow T_x(M)$$

is a linear isomorphism of vector spaces.

In particular, we can put an inner product on $T_x(M)$, meaning a bilinear, symmetric, positive-definite function

$$(11.1.1) \quad \begin{aligned} T_x(M) \times T_x(M) &\rightarrow \mathbb{R}, \\ (v, w) &\mapsto \langle v, w \rangle_x. \end{aligned}$$

In fact, we could put many different inner products on $T_x(M)$, and any one of them admits an orthonormal basis via the Gram–Schmidt process. So the result is linearly isometric to \mathbb{R}^n with the usual inner product, but not in an obvious way.

Note that if $f : V \rightarrow W$ is an injective homomorphism of vector spaces and if $\langle \cdot, \cdot \rangle$ is an inner product on W , then there is an induced inner product $\langle \cdot, \cdot \rangle^f$ on V given by

$$(11.1.2) \quad \langle v_1, v_2 \rangle^f = \langle f(v_1), f(v_2) \rangle$$

for $v_1, v_2 \in V$. We call this the pullback by f of the inner product on W .

Now suppose that $h : U \xrightarrow{\cong} h(U) \subset \mathbb{R}^n$ is a chart for the smooth n -manifold M . Then h is a diffeomorphism, and we get a commutative diagram

$$(11.1.3) \quad \begin{array}{ccc} T(U) & \xrightarrow[\cong]{Th} & T(h(U)) = h(U) \times \mathbb{R}^n \\ \pi \downarrow & & \downarrow \pi \\ U & \xrightarrow{h} & h(U), \end{array}$$

inducing a linear isomorphism of vector spaces

$$(11.1.4) \quad T_x h : T_x(U) \xrightarrow{\cong} T_{h(x)}(h(U)) = \mathbb{R}^n$$

¹⁷A diffeomorphism $f : \mathbb{H} \rightarrow \mathbb{R}^2$ is given by $f([\frac{x}{y}]) = [\frac{x}{\ln y}]$.

for each $x \in U$. Of course, $T_x(U) = T_x(M)$ as $T(U)$ is the full inverse image of U under the tangent bundle map $\pi : T(M) \rightarrow M$. In particular, if $\langle \cdot, \cdot \rangle_x$ is an inner product on $T_x(M)$ we can pull it back over $T_{h(x)}h^{-1}$ to obtain an inner product $\langle \cdot, \cdot \rangle_{h(x)}$ on $T_{h(x)}(h(U)) = \mathbb{R}^n$. Specifically, if $y = h(x)$, we have

$$(11.1.5) \quad \langle v, w \rangle_y = \langle T_y h^{-1}(v), T_y h^{-1}(w) \rangle_x$$

for $v, w \in \mathbb{R}^n$. With this as setup we can make the following definition.

Definition 11.1.1. A Riemannian metric on a smooth manifold M is a choice of inner product for the tangent space at each $x \in M$ that varies smoothly as x varies. This latter means that if $h : U \xrightarrow{\cong} h(U) \subset \mathbb{R}^n$ is a smooth chart for M , then the composite

$$(11.1.6) \quad \begin{aligned} h(U) \times \mathbb{R}^n \times \mathbb{R}^n &\rightarrow \mathbb{R} \\ (y, v, w) &\mapsto \langle v, w \rangle_y \end{aligned}$$

is smooth, where $\langle \cdot, \cdot \rangle_y$ is the inner product defined by (11.1.5).

Note that by (10.2.8), it suffices to show (11.1.6) is smooth as U varies over an open cover of M by chart neighborhoods. Moreover, (11.1.6) gives a Riemannian metric on $h(U)$. We call it the local model for the metric on M .

A smooth manifold equipped with a Riemannian metric is said to be a Riemannian manifold. The metric is said to put a Riemannian structure on M .

As in the case of \mathbb{H} , a given smooth manifold can have Riemannian structures with very different geometric properties. As in the case of Euclidean geometry, the study of distances and angles will be important in studying the geometry of a Riemannian manifold. The characterization of “straight lines” is surprisingly complicated, but is vital for understanding the geometry. Let’s begin by giving some examples and then study their geometric properties.

Examples 11.1.2.

- (1) \mathbb{R}^n with the standard inner product at each point is a Riemannian manifold, as

$$(11.1.7) \quad (x, u, v) \mapsto \langle u, v \rangle$$

is a smooth map from $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$ to \mathbb{R} .

- (2) If $i : M \subset \mathbb{R}^n$ is any smooth submanifold of \mathbb{R}^n . Then we can pull back the standard metric on \mathbb{R}^n by Ti as in (11.1.2). We call this the subspace metric on M .

More generally, if $\eta : N \rightarrow M$ is a smooth immersion, then any Riemannian metric on M may be pulled back by $T\eta$ to a Riemannian metric on N . Many other, nonrelated metrics are of course possible on N .

- (3) In particular, the restriction of the standard Euclidean metric to the unit sphere $\mathbb{S}^n \subset \mathbb{R}^{n+1}$ gives the standard Riemannian structure on the n -sphere, \mathbb{S}^n .
- (4) Consider the upper half-plane \mathbb{H} as a subset of the complex numbers. Thus,

$$(11.1.8) \quad \mathbb{H} = \{z = x + iy : x, y \in \mathbb{R}, y > 0\}.$$

here, it is customary to write $x = \operatorname{Re}(z)$ and $y = \operatorname{Im}(z)$, the real and imaginary parts of z , respectively. The standard metric on the hyperbolic space \mathbb{H} sets

$$(11.1.9) \quad \langle v, w \rangle_z = \frac{1}{\operatorname{Im}(z)^2} \langle v, w \rangle$$

for $z \in \mathbb{H}$, $v, w \in \mathbb{R}^2$, where the $\langle v, w \rangle$ on the right-hand side is the standard inner product of v, w in \mathbb{R}^2 . Thus, the standard inner product is scaled by the reciprocal of the square of the Euclidean distance from z to the x -axis. We think of the x -axis as lying on the boundary of \mathbb{H} (but the x -axis is not part of \mathbb{H}). Since $z \mapsto \operatorname{Im}(z)$ is smooth and since $\operatorname{Im}(z) \neq 0$ for all $z \in \mathbb{H}$, this prescribes a smooth metric on \mathbb{H} .

Definition 11.1.3. A smooth immersion (or embedding) $f : M \rightarrow N$ of Riemannian manifolds is isometric if for each $x \in M$, the metric on $T_x(M)$ induced by f (via pullback) coincides with the existing metric there, i.e., if

$$(11.1.10) \quad \langle v, w \rangle_x = \langle T_x f(v), T_x f(w) \rangle_{f(x)}$$

for all $v, w \in T_x(M)$. Note the dependence of this on Jacobian matrices. If $h : U \xrightarrow{\cong} h(U) \subset \mathbb{R}^m$ and $k : V \xrightarrow{\cong} k(V) \subset \mathbb{R}^n$ are charts about x and $f(x)$, respectively, this says

$$(11.1.11) \quad \langle v, w \rangle_y = \langle Df_{kh}(y)v, Df_{kh}(y)w \rangle_{f_{kh}(y)}$$

for all $y \in h(U \cap f^{-1}V)$ and $v, w \in \mathbb{R}^m$, where $f_{kh} = k \circ f \circ h^{-1}$ as in (8.4.2). Here, the metrics are the local metrics induced by those on M and N , respectively, as in (11.1.5).

John Nash (of *A Beautiful Mind* fame) proved the following. See [10] for a proof.

Theorem 11.1.4 (Nash embedding theorem). *Every Riemannian manifold m -manifold M embeds isometrically in \mathbb{R}^n for n sufficiently large. If M is compact, we may take $n = m(3m + 11)/2$. Otherwise, we may take $n = m(m + 1)(3m + 11)/2$.*

Our basic object of study here is isometries.

Definition 11.1.5. An isometry of a Riemannian manifold M is a diffeomorphism $f : M \rightarrow M$ that is isometric in the sense of Definition 11.1.3.

It is not immediately obvious that this coincides with the notion of surjective, distance-preserving maps we studied in our work on Euclidean geometry. We shall see below that any isometry in the sense of Definition 11.1.5 does preserve distance, and is certainly surjective, as all diffeomorphisms are. In the Euclidean case we have verified that every distance-preserving surjection $\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the composite of a translation and a linear isometry, each of which has been shown to preserve the Euclidean inner product in the sense of Definition 11.1.5. So our definition here does generalize the isometries of Euclidean space.

By the chain rule, if $f : M \rightarrow M$ is an isometry, so is $f^{-1} : M \rightarrow M$. Isometries are also closed under composition.

Definition 11.1.6. Let M be a Riemannian manifold. We write $\mathcal{I}(M)$ for the group (under composition) of isometries $f : M \rightarrow M$.

Isometries give us another source of examples of Riemannian manifolds. We say a group G acts by isometries on the Riemannian manifold M if for each $g \in G$ the map $\mu_g : M \rightarrow M$, $\mu_g(x) = gx$, is an isometry.

Proposition 11.1.7. *Let G act by isometries on the Riemannian manifold M . Suppose this action is free and properly discontinuous. Then there is a Riemannian metric on M/G given by setting*

$$(11.1.12) \quad \langle v, w \rangle_{\pi(x)} = \langle (T_x\pi)^{-1}(v)(T_x\pi)^{-1}(w) \rangle_x$$

for all $x \in M$ and $v, w \in T_{\pi(x)}(M/G)$. The map $\pi : M \rightarrow M/G$ is an isometric immersion.

Proof. The elements in $\pi^{-1}\pi(x)$ are those in the orbit Gx . Since G acts isometrically on M , the metric induced by (11.1.12) is independent of the choice of $x \in \pi^{-1}\pi(x)$. The induced metric is smooth, as π restricts to a diffeomorphism from a neighborhood of x onto a neighborhood of $\pi(x)$ (as, in fact, is true of any codimension 0 immersion). \square

In particular, our studies of the standard metric on \mathbb{R}^2 will enable us to study the geometry of the Klein bottle and \mathbb{T}^2 , including their isometries. Similarly, the Euclidean geometry of \mathbb{R}^n will give us geometric information about \mathbb{T}^n , and our studies of spherical geometry will have implications for projective geometry.

11.2. Arc length, distance and angles. Let M be a Riemannian manifold. We wish to study the lengths of curves in M . A natural thing to do would be to insist that we consider only smooth curves, but a goal here would be to calculate the perimeters of polygons in M . For instance, we'd like to show a triangle inequality, that says the length of one side of a triangle in M is less than or equal to the length of the path that traverses the other two sides. That is most easily done if we consider that traversal as a single path, rather than the sum of two separate lengths.

Thus, we assign an arc length to a piecewise smooth curve $\gamma : [a, b] \rightarrow M$. This means there is a partition $a = x_0 < x_1 < \cdots < x_k = b$ of $[a, b]$ such that the restriction $\gamma|_{[x_{i-1}, x_i]}$ of γ to $[x_{i-1}, x_i]$ is smooth for $i = 1, \dots, k$. Here, a map from a closed interval to a manifold is considered smooth if it may be extended to a smooth map on a slightly larger open interval.

Recall that $\gamma'(t)$ is our notation for $T_t\gamma(1) \in T_{\gamma(t)}M$. Here, in the local model, if $M = U \subset \mathbb{R}^n$, then $T_t\gamma : \mathbb{R} \rightarrow \mathbb{R}^n$ is multiplication by the velocity vector $\gamma'(t)$ (considered as a column matrix, hence a linear map from \mathbb{R} to \mathbb{R}^n), hence $T_t\gamma(1)$ is precisely that velocity vector, hence the notation.

In particular, we may define

$$\|\gamma'(t)\|_{\gamma(t)} = \sqrt{\langle \gamma'(t), \gamma'(t) \rangle_{\gamma(t)}},$$

the length of $\gamma'(t)$ evaluated in the inner product on $T_{\gamma(t)}(M)$ in the Riemannian metric.

Definition 11.2.1. Let $\gamma : [a, b] \rightarrow M$ be a piecewise smooth curve in the Riemannian manifold M . Then the arc length of γ is

$$(11.2.1) \quad \ell(\gamma) = \int_a^b \|\gamma'(t)\|_{\gamma(t)} dt.$$

Note that the piecewise smooth assumption implies that $t \mapsto \|\gamma'(t)\|_{\gamma(t)}$ is piecewise continuous, with at most finitely many jump discontinuities, and hence is Riemann integrable.

We now show the arc length is independent of the parametrization of γ . Recall that a function $f : [a, b] \rightarrow \mathbb{R}$ is increasing if $f(x) \leq f(y)$ whenever $x \leq y$, and is decreasing if $f(x) \geq f(y)$ whenever $x \leq y$. It is monotonic if it is either increasing or decreasing.

Lemma 11.2.2. Let $\gamma : [a, b] \rightarrow M$ be piecewise smooth and let $u : [c, d] \rightarrow [a, b]$ be surjective, piecewise smooth and monotonic. Then

$$\ell(\gamma) = \ell(\gamma \circ u).$$

Proof. Subdividing $[a, b]$ finely enough, we may assume γ is smooth and takes value in a chart neighborhood $U \subset M$. Thus, we may simply use the local model and assume $M = U \subset \mathbb{R}^n$. The inner product still varies over different points in U .

Suppose u is increasing. Then $u'(t) \geq 0$ for all t by the first derivative test. So

$$\begin{aligned} \ell(\gamma \circ u) &= \int_c^d \|(\gamma \circ u)'(t)\|_{\gamma \circ u(t)} dt \\ &= \int_c^d \|\gamma'(u(t))u'(t)\|_{\gamma \circ u(t)} dt \\ &= \int_c^d \|\gamma'(u(t))\|_{\gamma \circ u(t)} u'(t) dt \end{aligned}$$

$$= \int_{u(c)}^{u(d)} \|\gamma'(u)\|_{\gamma(u)} du.$$

But this is $\ell(\gamma)$ because $u : [c, d] \rightarrow [a, b]$ is surjective.

If u is decreasing, $\|u'(t)\| = -u'(t)$ by the first derivative test, so

$$\begin{aligned} \ell(\gamma \circ u) &= - \int_{u(c)}^{u(d)} \|\gamma'(u)\|_{\gamma(u)} du \\ &= - \int_b^a \|\gamma'(u)\|_{\gamma(u)} du = \ell(\gamma). \end{aligned} \quad \square$$

Thus, the arc length depends on the path traced out by γ . But note that γ can go around in circles, wrapping around a given circuit multiple times. So its arc length doesn't necessarily measure the size of its image.

Since we can reparametrize γ at will, it can be useful to do so in such a way that its velocity vector always has unit length.

Definition 11.2.3. We say $\gamma : [a, b] \rightarrow M$ is parametrized by arc length if $\|\gamma'(t)\|_{\gamma(t)} = 1$ for all $t \in [a, b]$. Of course, in this case,

$$\ell(\gamma|_{[a,t]}) = \int_a^t \|\gamma'(u)\|_{\gamma(u)} du = t - a$$

for all $t \in [a, b]$.

Definition 11.2.4. A curve $\gamma : [a, b] \rightarrow M$ is nonsingular if it is smooth and $\gamma'(t) \neq 0$ for all $t \in [a, b]$.

Of course, any curve parametrized by arc length is nonsingular. The following shows a nonsingular curve may be reparametrized by arc length.

Lemma 11.2.5. *Let $\gamma : [a, b] \rightarrow M$ be nonsingular. Then there is a smooth, increasing function τ such that $\gamma \circ \tau$ is parametrized by arc length. We call $\gamma \circ \tau$ the parametrization of γ by arc length.*

Proof. Define $s : [a, b] \rightarrow \mathbb{R}$ by

$$s(t) = \ell(\gamma|_{[a,t]}) = \int_a^t \|\gamma'(u)\|_{\gamma(u)} du.$$

Since $s'(t) = \|\gamma'(t)\|_{\gamma(t)} > 0$, s is strictly increasing with image $[0, d]$ with $d = \ell(\gamma)$. Let $\tau : [0, d] \rightarrow [a, b]$ be the inverse function of s . Then the inverse function theorem gives

$$\tau'(u) = \frac{1}{s'(\tau(u))} = \frac{1}{\|\gamma'(\tau(u))\|_{\gamma(\tau(u))}},$$

so

$$(\gamma \circ \tau)'(u) = \frac{\gamma'(\tau(u))}{\|\gamma'(\tau(u))\|_{\gamma(\tau(u))}}$$

is a unit vector in $T_{\gamma(\tau(u))}(M)$ for all $u \in [0, d]$. \square

We now show that isometries preserve arc length.

Lemma 11.2.6. *Let $f : M \rightarrow N$ be an isometric immersion of Riemannian manifolds and let $\gamma : [a, b] \rightarrow M$ be piecewise smooth. Then $\ell(f \circ \gamma) = \ell(\gamma)$.*

Proof. The isometric condition is that

$$\langle T_x f(v), T_x f(w) \rangle_{f(x)} = \langle v, w \rangle_x$$

for all $v, w \in T_x(M)$, and hence

$$\|T_x f(v)\|_{f(x)} = \|v\|_x$$

for all $v \in T_x(M)$. Thus,

$$\begin{aligned} \ell(f \circ \gamma) &= \int_a^b \|(f \circ \gamma)'(t)\|_{(f \circ \gamma)(t)} dt \\ &= \int_a^b \|(T_{f(\gamma(t))} f \circ T_t \gamma)(1)\|_{(f \circ \gamma)(t)} dt \\ &= \int_a^b \|T_t \gamma(1)\|_{\gamma(t)} dt = \ell(\gamma) \quad \square \end{aligned}$$

We can now define Riemannian distance.

Definition 11.2.7. Let M be a path-connected Riemannian manifold and let $x, y \in M$. A piecewise smooth path from x to y is a piecewise smooth curve $\gamma : [a, b] \rightarrow M$ with $\gamma(a) = x$ and $\gamma(b) = y$. Since M is path-connected, the Whitney approximation theorem ([13, Theorem 6.26]) shows there are smooth, and hence piecewise smooth curves from x to y .

The distance from x to y with respect to the Riemannian metric on M (otherwise known as the Riemannian distance from x to y) is defined by

$$(11.2.2) \quad d(x, y) = \inf_{\gamma} \ell(\gamma)$$

as γ varies over all the piecewise smooth paths from x to y .

We say the piecewise smooth path γ from x to y is distance minimizing if $d(x, y) = \ell(\gamma)$, i.e., if $\ell(\gamma) \leq \ell(\delta)$ for all piecewise smooth paths δ from x to y .

Remark 11.2.8. We will see that the distance minimizing curves in a Riemannian manifold are what's known as geodesics. The theory of geodesics is fundamental in the development of differential geometry. Geodesic curves are the analogue in Riemannian manifolds of the straight lines in \mathbb{R}^n . As such, they will form the edges of triangles in the appropriate analogue of Euclidean geometry for the manifold in question.

Having defined a distance function, it is natural to ask if it defines a metric for a topology as in Definition A.1.1. The triangle inequality is obvious from the piecewise smooth assumption, and symmetry follows from Lemma 11.2.2. The theory of geodesics will show it is positive-definite. We shall also show that the topology on M induced by this distance function coincides with its topology as a smooth manifold.

Proposition 11.2.9. *Let $f : M \rightarrow M$ be an isometry on a Riemannian manifold M . Then f preserves distance:*

$$(11.2.3) \quad d(x, y) = d(f(x), f(y))$$

for all $x, y \in M$.

Proof. Let $\gamma : [a, b] \rightarrow M$ be a piecewise smooth path from x to y . By Lemma 11.2.6, $\ell(\gamma) = \ell(f \circ \gamma) \geq d(f(x), f(y))$, so

$$d(x, y) \geq d(f(x), f(y)).$$

But $f^{-1} : M \rightarrow M$ is also an isometry, so the same argument shows $d(f(x), f(y)) \geq d(x, y)$. \square

Remark 11.2.10. Note that it is essential in Proposition 11.2.9 that f^{-1} also be an isometry. Indeed, if $f : M \rightarrow N$ is an isometric immersion, then

$$d(f(x), f(y)) \leq d(x, y)$$

by the argument given. Here, the distance on the left as in N and that on the right is in M . But these distances need not be equal, as there may be piecewise smooth paths in N shorter than those in M .

For instance if $f : \mathbb{S}^2 \rightarrow \mathbb{R}^3$ is the standard embedding and $v \in \mathbb{S}^2$, then the distance in \mathbb{S}^2 from v to $-v$ will be seen to be π , as any path from v to $-v$ in \mathbb{S}^2 must stay in \mathbb{S}^2 and hence go around the sphere. But in \mathbb{R}^3 we can cut through the interior in a straight line path, and we see that the distance from $f(v)$ to $f(-v)$ in \mathbb{R}^3 is 2.

We now define angles in a Riemannian manifold.

Definition 11.2.11. Let $\gamma_1 : (a, b) \rightarrow M$ and $\gamma_2 : (c, d) \rightarrow M$ be smooth curves through x in the Riemannian manifold M . Suppose $\gamma_1^{-1}(x) = \{s\}$ and $\gamma_2^{-1}(x) = \{t\}$ and that $\gamma_1'(s)$ and $\gamma_2'(t)$ are nonzero. Then we define the angle from γ_1 to γ_2 at x to be the angle from $\gamma_1'(s)$ to $\gamma_2'(t)$ in $T_x(M)$ with respect to the inner product there. This is a well-defined signed angle if $n = 2$ and M is oriented, but is unsigned otherwise.

Note by the proof of Lemma 11.2.2 that this angle is unchanged if we reparametrize our curves with respect to increasing functions with nonzero derivatives at the relevant points. But if we reparametrize one of the curves with respect to a decreasing function, it changes the orientation of that curve and adds π to the angle.

The chain rule again gives the following.

Lemma 11.2.12 (Isometries preserve angles). *Let $f : M \rightarrow N$ be an isometric embedding. Let $\gamma_1 : (a, b) \rightarrow M$ and $\gamma_2 : (c, d) \rightarrow M$ be smooth curves through $x \in M$ with $\gamma_1^{-1}(x) = \{s\}$ and $\gamma_2^{-1}(x) = \{t\}$ and $\gamma_1'(s)$ and $\gamma_2'(t)$ nonzero. Then the (unsigned) angle from $f \circ \gamma_1$ to $f \circ \gamma_2$ is equal to the (unsigned) angle from γ_1 to γ_2 .*

The case of signed angles is more delicate.

11.3. Geodesics. The geodesics in a Riemannian manifold provide distance-minimizing paths between points. Thus, they play a role like that of straight lines in \mathbb{R}^n . In particular, in \mathbb{R}^n with the standard metric, the straight lines are the geodesics.

A thorough treatment of geodesics requires more differential geometry than we wish to present here. A very good source is do Carmo's book [3]. We shall summarize the most important details so we can use the theory to study "straight lines" and angles in a Riemannian manifold.

11.3.1. Geodesics in the local model. The easiest way to understand geodesics is in terms of the local model. Let us study a Riemannian metric on an open subset $U \subset \mathbb{R}^n$, so that TU is just $U \times \mathbb{R}^n$ and the tangent bundle $\pi : U \times \mathbb{R}^n \rightarrow U$ is the projection map. For each $x \in U$ we have an inner product $\langle \cdot, \cdot \rangle_x : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ such that the map

$$\begin{aligned} U \times \mathbb{R}^n \times \mathbb{R}^n &\rightarrow \mathbb{R} \\ (x, v, w) &\mapsto \langle v, w \rangle_x \end{aligned}$$

is smooth.

In particular, this gives smooth functions $g_{ij} : U \rightarrow \mathbb{R}$ via

$$(11.3.1) \quad g_{ij}(x) = \langle e_i, e_j \rangle_x,$$

with e_i and e_j the canonical basis vectors. These functions in fact determine the Riemannian structure. Let $v = a_1e_1 + \cdots + a_n e_n$ and $w = b_1e_1 + \cdots + b_n e_n$. Then bilinearity gives

$$(11.3.2) \quad \langle v, w \rangle_x = \sum_{i,j=1}^n a_i g_{ij}(x) b_j.$$

This can be expressed as a matrix product as follows. Let $G(x)$ be the matrix whose ij th coordinate is $g_{ij}(x)$, then (11.3.2) just says

$$(11.3.3) \quad \langle v, w \rangle_x = v^T \cdot G(x) \cdot w,$$

where v^T is the transpose of v . Note the matrix $G(x)$ is symmetric, i.e., $G(x)^T = G(x)$ by the symmetry of the inner product. $G(x)$ is also invertible since the inner product is positive-definite: if $G(x)v = 0$, then

$$\langle v, v \rangle_x = v^T \cdot G(x) \cdot v = 0,$$

and hence $v = 0$.

The group $\text{GL}_n(\mathbb{R})$ of $n \times n$ invertible matrices over \mathbb{R} is an open subset of the n^2 -dimensional Euclidean space $M_n(\mathbb{R})$ (the space of all $n \times n$ matrices over \mathbb{R}), as $\text{GL}_n(\mathbb{R}) = \det^{-1}(\mathbb{R} \setminus \{0\})$, and $\det : M_n(\mathbb{R}) \rightarrow \mathbb{R}$ is a polynomial function in the n^2 variables ((8.2.4) — see [17, Corollary 10.2.6]). As an open subset of \mathbb{R}^{n^2} , $\text{GL}_n(\mathbb{R})$ is a manifold.

Definition 11.3.1. A Lie group G is a group which is also a smooth manifold, such that the following hold:

- (1) The multiplication $\mu : G \times G \rightarrow G$, $\mu(x, y) = x \cdot y$, is a smooth map.
- (2) The inverse map $\chi : G \rightarrow G$, $\chi(x) = x^{-1}$ is a smooth map.

We already know that the coefficients of the product of two matrices are polynomials in the coefficients of the matrices, so the multiplication map on $\text{GL}_n(\mathbb{R})$ is smooth. The following completes the proof that $\text{GL}_n(\mathbb{R})$ is a Lie group.

Proposition 11.3.2. *The map $\chi : \text{GL}_n(\mathbb{R}) \rightarrow \text{GL}_n(\mathbb{R})$ given by $\chi(A) = A^{-1}$ is smooth.*

Proof. It suffices to show that the coefficients of A^{-1} are quotients of polynomial functions (i.e., rational functions) in the coefficients of A . By [17, Corollary 10.3.6], the ij th coefficient of A^{-1} is

$$(-1)^{i+j} \frac{\det A_{ji}}{\det A},$$

where A_{ji} is the $(n-1) \times (n-1)$ matrix obtained by deleting the j th row and i th column of A . \square

Corollary 11.3.3. *Suppose given a Riemannian metric on the open set $U \subset \mathbb{R}^n$ and let $G(x) = (g_{ij}(x))$ be the matrix given by (11.3.1). Define $g^{ij} : U \rightarrow \mathbb{R}$ by setting $g^{ij}(x)$ equal to the ij th coordinate of $G(x)^{-1}$. Then g^{ij} is smooth.*

We use this to define important smooth functions in the local model that depend on the Riemannian metric.

Definition 11.3.4. Suppose given a Riemannian metric on the open set $U \subset \mathbb{R}^n$ and let $G(x) = (g_{ij}(x))$ be the matrix given by (11.3.1). For $i, j, k \in \{1, \dots, n\}$, the Christoffel symbol $\Gamma_{ij}^k : U \rightarrow \mathbb{R}$ for this metric is the smooth function given by

$$(11.3.4) \quad \Gamma_{ij}^k = \frac{1}{2} \sum_{l=1}^n \left(\frac{\partial g_{jl}}{\partial x_i} + \frac{\partial g_{li}}{\partial x_j} - \frac{\partial g_{ij}}{\partial x_l} \right) g^{lk},$$

where g^{lk} is the lk th coordinate of the inverse matrix G^{-1} .

Note that in \mathbb{R}^n with the standard metric, $G(x) = I_n$ for all x , so the partial derivatives of its coordinate functions are constantly 0. We obtain:

Lemma 11.3.5. *In \mathbb{R}^n with the standard metric, Γ_{ij}^k is constantly equal to 0 for all i, j, k .*

The Christoffel symbols are the coefficient functions for what is called the Riemannian connection obtained from the metric. They also occur in what is known as the covariant derivative, which we shall define here, as it is used in the definition of geodesics. We must first discuss vector fields over smooth curves.

Definition 11.3.6. Let $\gamma : (a, b) \rightarrow M$ be a smooth curve in a smooth manifold M . A smooth vector field over γ is a smooth map $V : (a, b) \rightarrow TM$ making the following diagram commute:

$$(11.3.5) \quad \begin{array}{ccc} & & TM \\ & \nearrow V & \downarrow \pi \\ (a, b) & \xrightarrow{\gamma} & M. \end{array}$$

We write $\mathfrak{X}(\gamma)$ for the set of smooth vector fields over γ . Since each $T_x M$ is a real vector space we can add vector fields pointwise: we set $(V+W)(t)$ to be the sum of $V(t)$ and $W(t)$ in $T_{\gamma(t)} M$. An examination of the local model shows that $V+W$ is a smooth vector field for V, W smooth. Similarly, if $V \in \mathfrak{X}(\gamma)$ and $c \in \mathbb{R}$ we set $(cV)(t) = cV(t)$. We see that $\mathfrak{X}(\gamma)$ is a real vector space under these operations.

In fact, if $V \in \mathfrak{X}(\gamma)$ and $f : (a, b) \rightarrow \mathbb{R}$ is smooth, we may define a vector field $fV \in \mathfrak{X}(\gamma)$ by setting $(fV)(t) = f(t)V(t)$, multiplication on $V(t)$ by the scalar $f(t)$. This operation satisfies the expected distributive laws:

$$\begin{aligned} (f+g)V &= fV + gV \\ f(V+W) &= fV + fW. \end{aligned}$$

The following obvious example is important.

Example 11.3.7. Recall that if $\gamma : (a, b) \rightarrow M$ is smooth and $t \in (a, b)$, we write $\gamma'(t) \in T_{\gamma(t)} M$ for $T_t \gamma(1)$. Here, we identify the tangent space of (a, b) with $(a, b) \times \mathbb{R}$, obtaining the tangent map $T\gamma$ in the following diagram:

$$(11.3.6) \quad \begin{array}{ccc} (a, b) \times \mathbb{R} & \xrightarrow{T\gamma} & TM \\ \pi \downarrow & & \downarrow \pi \\ (a, b) & \xrightarrow{\gamma} & M. \end{array}$$

So $\gamma'(t)$ is the tangent vector $T\gamma(t, 1)$ in the tangent space to $\gamma(t) \in M$. We define the tangent vector field γ' to γ to be the vector field $t \mapsto \gamma'(t)$. This is smooth, as $T\gamma$ is smooth.

In the local model, this really is the standard tangent vector field. Here, if $M = U \subset \mathbb{R}^n$, then $TM = U \times \mathbb{R}^n$ and we have

$$(11.3.7) \quad \begin{aligned} T\gamma : (a, b) \times \mathbb{R} &\rightarrow U \times \mathbb{R}^n \\ (t, s) &\mapsto (\gamma(t), \gamma'(t)s), \end{aligned}$$

where $\gamma'(t) \in \mathbb{R}^n$ is the usual tangent vector to γ at t . Specializing to $s = 1$, we see the tangent vector field γ' is given by

$$t \mapsto (\gamma(t), \gamma'(t)) \in U \times \mathbb{R}^n.$$

We now look more carefully at arbitrary vector fields over curves in the local model.

Notation 11.3.8. Let $U \subset \mathbb{R}^n$ be open. Then $TU = U \times \mathbb{R}^n$. Thus, if $\gamma : (a, b) \rightarrow U$ is smooth, then a smooth vector field over γ is a smooth map

$$(11.3.8) \quad \begin{aligned} V : (a, b) &\rightarrow U \times \mathbb{R}^n \\ t &\mapsto (\gamma(t), v(t)) \end{aligned}$$

with $v : (a, b) \rightarrow \mathbb{R}^n$ an arbitrary smooth map. In this context, we will write $V = (\gamma, v)$ and we note that both $\gamma(t)$ and $v(t)$ lie in \mathbb{R}^n . We write $\gamma_i(t)$ and $v_i(t)$ for the i th coordinates of $\gamma(t)$ and $v(t)$, respectively.

We can now make use of the Christoffel symbols.

Definition 11.3.9. Let U be an open subset of \mathbb{R}^n equipped with a Riemannian metric. Let $\gamma : (a, b) \rightarrow U$ be smooth. The covariant derivative on $\mathfrak{X}(\gamma)$ with respect to this metric is the operator

$$\mathfrak{D} : \mathfrak{X}(\gamma) \rightarrow \mathfrak{X}(\gamma)$$

given as follows. For $V = (\gamma, v)$ as above, $\mathfrak{D}(V) = (\gamma, w)$, where the coordinate functions of w are given by

$$(11.3.9) \quad w_i(t) = v'_i(t) + \sum_{j,k} v_k(t) \gamma'_j(t) \Gamma_{jk}^i.$$

Note that $w(t)$ is equal to the sum of $v'(t)$ with a term that depends on the metric. Since this latter term is expressed entirely in terms of the Christoffel symbols it vanishes for the standard metric on \mathbb{R}^n . We obtain:

Lemma 11.3.10. *The covariant derivative for the standard metric on \mathbb{R}^n is given by*

$$\mathfrak{D}(\gamma, v) = (\gamma, v').$$

We may now define geodesics in the local model.

Definition 11.3.11. Suppose given a Riemannian metric on the open subset $U \subset \mathbb{R}^n$ and let $\gamma : (a, b) \rightarrow U$ be a smooth curve in U . As above, write γ' for the tangent vector field to γ . Then γ is geodesic if $\mathfrak{D}(\gamma') = 0$, i.e., the covariant derivative of γ' is constantly 0.

The following is immediate from (11.3.9).

Lemma 11.3.12. *Suppose given a Riemannian metric on the open subset $U \subset \mathbb{R}^n$ and let $\gamma : (a, b) \rightarrow U$ be a smooth curve in U . Then γ is geodesic if and only if*

$$(11.3.10) \quad \gamma''_i(t) + \sum_{j,k=1}^n \Gamma_{jk}^i(\gamma(t)) \gamma'_j(t) \gamma'_k(t) = 0$$

for all $t \in (a, b)$ and for $i = 1, \dots, n$.

In the standard metric on \mathbb{R}^n , we have $\Gamma_{ij}^k = 0$, so this just says $\gamma'' = 0$, so that γ' is a constant vector, and hence there are vectors $v, x \in \mathbb{R}^n$ with $\gamma(t) = tv + x$ for all t . We obtain:

Proposition 11.3.13. *In \mathbb{R}^n with the standard metric, a geodesic is either a constant function or the standard parametrization of a line:*

$$\gamma(t) = tv + x.$$

11.3.2. Geodesics in general Riemannian manifolds. Now let M be an arbitrary Riemannian manifold and let $\gamma : (a, b) \rightarrow M$ be smooth. Let $h : U \rightarrow h(U) \subset \mathbb{R}^n$ be a chart for M with $\gamma(t) \in U$ and let $\epsilon > 0$ with $\gamma(t - \epsilon, t + \epsilon) \subset U$. Then (11.1.3) gives us a commutative diagram

$$(11.3.11) \quad \begin{array}{ccc} & T(U) & \xrightarrow{Th} h(U) \times \mathbb{R}^n \\ & \nearrow V|_{(t-\epsilon, t+\epsilon)} & \cong \downarrow \pi \\ (t-\epsilon, t+\epsilon) & \xrightarrow{\gamma|_{(t-\epsilon, t+\epsilon)}} U & \xrightarrow{h} h(U). \end{array}$$

Thus, $Th \circ V|_{(t-\epsilon, t+\epsilon)} \in \mathfrak{X}(\gamma|_{(t-\epsilon, t+\epsilon)})$ is a vector field over $\gamma|_{(t-\epsilon, t+\epsilon)}$ in the Riemannian metric on $h(U)$ induced by h^{-1} . Write

$$\mathfrak{D}_{h(U)} : \mathfrak{X}(\gamma|_{(t-\epsilon, t+\epsilon)}) \rightarrow \mathfrak{X}(\gamma|_{(t-\epsilon, t+\epsilon)})$$

for the covariant derivative induced by this metric. Note that

$$(11.3.12) \quad Th^{-1} \circ \mathfrak{D}_{h(U)}(V|_{(t-\epsilon, t+\epsilon)})$$

provides a smooth vector field over $\gamma|_{(t-\epsilon, t+\epsilon)}$.

The following now defines the covariant derivative in a general Riemannian manifold. It is proven in [3].

Theorem 11.3.14. *Let M be a Riemannian manifold and let $\gamma : (a, b) \rightarrow M$. Then there is an operator*

$$\mathfrak{D} : \mathfrak{X}(\gamma) \rightarrow \mathfrak{X}(\gamma)$$

obtained by setting $\mathfrak{D}(V)(t) = Th^{-1} \circ \mathfrak{D}_{h(U)}(V|_{(t-\epsilon, t+\epsilon)})(t)$ for any chart $h : U \rightarrow h(U)$ containing $\gamma(t)$ and for ϵ sufficiently small. In particular, this is independent of the choice of chart about $\gamma(t)$. Moreover, this covariant derivative \mathfrak{D} satisfies the following properties for $V, W \in \mathfrak{X}(\gamma)$ and $f : (a, b) \rightarrow \mathbb{R}$ smooth:

$$(11.3.13) \quad \mathfrak{D}(V + W) = \mathfrak{D}(V) + \mathfrak{D}(W)$$

$$(11.3.14) \quad \mathfrak{D}(fV) = f'V + f \mathfrak{D}(V)$$

$$(11.3.15) \quad \frac{d}{dt} \langle V, W \rangle = \langle \mathfrak{D}(V), W \rangle + \langle V, \mathfrak{D}(W) \rangle.$$

Here, as expected, $\langle V, W \rangle : (a, b) \rightarrow \mathbb{R}$ takes t to $\langle V(t), W(t) \rangle_{\gamma(t)}$.

Note that (11.3.13) and (11.3.14) are immediate from (11.3.9). Equation (11.3.15) is more work, and shows the relationship of the covariant derivative to the metric. The following is, of course, expected.

Definition 11.3.15. Let M be a Riemannian manifold. A smooth curve $\gamma : (a, b) \rightarrow M$ is geodesic if its tangent vector field γ' satisfies $\mathfrak{D}(\gamma') = 0$.

Theorem 11.3.14 gives the following.

Corollary 11.3.16. *Geodesics have constant speed: if $\gamma : (a, b) \rightarrow M$ is geodesic, then $t \mapsto \|\gamma'(t)\|_{\gamma(t)}$ is a constant function on (a, b) .*

Proof. By (11.3.15) $\frac{d}{dt} \langle \gamma', \gamma' \rangle = 0$. □

However, since any nonsingular curve can be parametrized by arc length, the converse is wildly false. Geodesics are in fact somewhat scarce, as can be seen from the following theorem. While there are many smooth curves through a particular point with a particular tangent vector, only one of them is geodesic.

Theorem 11.3.17. *Let M be a Riemannian manifold and let $v \in T_x(M)$. Then for some $\epsilon > 0$, there exists a unique geodesic $\gamma_{x,v} : (-\epsilon, \epsilon) \rightarrow M$ with:*

- (1) $\gamma_{x,v}(0) = x$.
- (2) $\gamma'_{x,v}(0) = v$.

Proof. The point here is that the covariant derivative is globally defined on M , independent of the choice of charts, and that $\mathfrak{D}(\gamma') = 0$ if and only if (11.3.10) holds in any given chart about x . The result now follows from the fundamental existence and uniqueness theorem for solutions of differential equations. □

We can now eliminate ϵ from the above.

Corollary 11.3.18. *Let $\gamma_1, \gamma_2 : (a, b) \rightarrow M$ be geodesics in M and suppose γ_1 and γ_2 have the same value and velocity vector at some $t_0 \in (a, b)$. Then $\gamma_1 = \gamma_2$ on all of (a, b) .*

Proof. Let $s = \inf\{t : \gamma_1 = \gamma_2 \text{ on } [t_0, t]\}$. Then $s > t_0$ by Theorem 11.3.17, and $\gamma_1 = \gamma_2$ on $[t_0, s)$. If $s < b$ then $\gamma_1 = \gamma_2$ on $[t_0, s]$ by continuity, and hence $\gamma'_1 = \gamma'_2$ on $[t_0, s]$ by the continuity of the velocity fields. But then $\gamma_1 = \gamma_2$ on $[t_0, s']$ for some $s' > s$ by Theorem 11.3.17, so $s = b$. Similarly $\gamma_1 = \gamma_2$ on $(a, t_0]$. □

Corollary 11.3.19. *Let M be a Riemannian manifold and let $x \in M$ and $v \in T_x(M)$. Then there is a unique largest interval containing 0 on which the geodesic $\gamma_{x,v}$ for which $\gamma_{x,v}(0) = x$ and $\gamma'_{x,v}(0) = v$ is defined.*

Proof. If (a, b) and (c, d) are two intervals about 0 on which such a geodesic is defined, then the two geodesics must agree on $(a, b) \cap (c, d)$ by Corollary 11.3.18. So they define a smooth curve on $(a, b) \cup (c, d)$, which is geodesic by Theorem 11.3.14. Similarly, we can pass to infinite unions of intervals about 0, and obtain the desired result. □

In particular, as shown in Proposition 11.3.13, the geodesics in \mathbb{R}^n are defined on all of \mathbb{R} .

Since a geodesic has constant speed, $\gamma_{x,v}$ is parametrized by arc length if and only if $\|v\|_x = 1$, i.e., if and only if v is a unit vector in the inner

product space $T_x(M)$. Since every vector $v \in T_x(M)$ has the form $v = cu$ with $\|u\|_x = 1$ (and $c = \|v\|_x$), the geodesics $\gamma_{x,u}$ actually determine all the others by the following:

Proposition 11.3.20. *Let M be a Riemannian manifold. Let u be a unit vector in the inner product space $T_x(M)$ and let (a, b) be the largest interval on which $\gamma_{x,u}$ is defined. Let $0 \neq c \in \mathbb{R}$. Then the largest interval for $\gamma_{x,cu}$ is $(\frac{a}{c}, \frac{b}{c})$, and*

$$(11.3.16) \quad \gamma_{x,cu}(t) = \gamma_{x,u}(ct)$$

for all $t \in (\frac{a}{c}, \frac{b}{c})$. In particular, if $c > 0$, then $\gamma_{x,u}$ is the parametrization of $\gamma_{x,cu}$ by arc length, while $\gamma_{x,-u}$ traverses the same arc with the opposite orientation.

Proof. The map $t \mapsto \gamma_{x,u}(ct)$ is geodesic by (11.3.10), and its tangent vector at 0 is cu . The result follows. \square

Of course, $\gamma_{x,0}$ is the constant path with image x , and is the unique singular geodesic through x .

The following, which is proven in [3], will help motivate studying geodesics.

Theorem 11.3.21. *Let γ be a piecewise-smooth distance minimizing path of constant speed from x to y in the Riemannian manifold M . Then γ is geodesic.*

The exponential map will provide a partial converse.

Our next example gives the geodesics in the n -sphere $\mathbb{S}^n \subset \mathbb{R}^{n+1}$. Recall that for $v \in \mathbb{S}^n$, the tangent space $T_v(\mathbb{S}^n)$ may be identified with $\{v\}^\perp$, the linear subspace of \mathbb{R}^{n+1} consisting of the vectors orthogonal to v . In particular, the unit vectors in $T_v(\mathbb{S}^n)$ are the vectors $w \in \mathbb{S}^n$ (i.e., w a unit vector) with $v \perp w$. The following is now immediate from Theorem 11.3.21 and Theorem 9.1.15. (We could also give a direct proof using Christoffel symbols and Proposition 9.1.13.)

Theorem 11.3.22. *Consider the n -sphere $\mathbb{S}^n \subset \mathbb{R}^{n+1}$ endowed with the subspace metric from \mathbb{R}^{n+1} . Its geodesics of unit speed are the great circle routes*

$$(11.3.17) \quad \begin{aligned} \gamma_{v,w} : \mathbb{R} &\rightarrow \mathbb{S}^n \\ \gamma_{v,w}(t) &= \cos t v + \sin t w, \end{aligned}$$

for $v, w \in \mathbb{S}^n$ with $v \perp w$.

11.4. The exponential map. Let M be a Riemannian manifold. We write $B_r(0)$ for the open ball of radius r about 0 in the inner product space $T_x(M)$:

$$B_r(0) = \{v \in T_x(M) : \|v\|_x < r\}.$$

The following is shown in [3].

Theorem 11.4.1. *Let M be a Riemannian manifold and $x \in M$. Then there exists $r > 0$ such that $\gamma_{x,v}(1)$ is defined for all $v \in B_r(0)$. Moreover, the exponential map $\exp_x : B_r(0) \rightarrow M$ defined by*

$$(11.4.1) \quad \exp_x(v) = \gamma_{x,v}(1)$$

is smooth.

The exponential map is often expressed best in “polar” coordinates, i.e., by writing $v \in B_r(0)$ in the form $v = cu$ with u a unit vector and $c \geq 0$ (so that $c = \|v\|_x$). As with polar coordinates in the plane, u is only a well-defined function of v when $v \neq 0$. Note that

$$B_r(0) = \{cu : \|u\|_x = 1 \text{ and } 0 \leq c < r\}.$$

The following is immediate from (11.3.16).

Lemma 11.4.2. *In polar coordinates, $\exp_x(cu) = \gamma_{x,u}(c) = \gamma_{x,u}(\|cu\|_x)$.*

Since $B_r(0)$ is an open subset of the inner product space $T_x(M)$, we can identify its tangent bundle with the projection map onto the first factor:

$$\pi : B_r(0) \times T_x(M) \rightarrow B_r(0).$$

Thus, we can identify $T_0(B_r(0))$ with $T_x(M)$, and we may ask about

$$T_0(\exp_x) : T_0(B_r(0)) = T_x(M) \rightarrow T_x(M).$$

The following is fundamental.

Lemma 11.4.3. *$T_0(\exp_x) : T_x(M) \rightarrow T_x(M)$ is the identity map.*

Proof. Let $0 \neq v \in T_x(M)$ and write $v = cu$ with u a unit vector and $c > 0$. Then for ϵ small enough, the map

$$\begin{aligned} \delta_v : (-\epsilon, \epsilon) &\rightarrow T_x(M) \\ t &\mapsto tv \end{aligned}$$

takes value in $B_r(0)$. By Lemma 11.4.2 and a second application of (11.3.16),

$$(\exp_x \circ \delta_v)(t) = \exp_x(tc) = \gamma_{x,u}(tc) = \gamma_{x,cu}(t) = \gamma_{x,v}(t)$$

Since $\delta_v'(0) = v$, we have

$$T_0(\exp_x)(v) = (\exp_x \circ \delta_v)'(0) = \gamma'_{x,v}(0) = v. \quad \square$$

Since $T_0(\exp_x)$ is an isomorphism, the inverse function theorem gives the following.

Corollary 11.4.4. *Let M be a Riemannian manifold and let $x \in M$. Then there exists $r > 0$ such that*

$$\exp_x : B_r(0) \rightarrow M$$

is a diffeomorphism onto a neighborhood of x .

Example 11.4.5. In \mathbb{S}^n , the geodesics $\gamma_{v,w}$ of (11.3.17) are defined on all of \mathbb{R} and the exponential map

$$\exp_v : T_v(\mathbb{S}^n) \rightarrow \mathbb{S}^n$$

is smooth on all of $T_v(\mathbb{S}^n)$. But \exp_v carries the entire sphere of radius π ,

$$S_\pi(0) = \{z \in T_v(\mathbb{S}^n) : \|z\|_v = \pi\},$$

onto the opposite pole, $-v$, to v , while

$$\exp_v : B_\pi(0) \rightarrow \mathbb{S}^n$$

maps $B_\pi(0)$ diffeomorphically onto $\mathbb{S}^n \setminus \{-v\}$. All this is easily verified using the analytic formula (11.3.17).

Note that

$$\exp_v^{-1} : \mathbb{S}^n \setminus \{-v\} \xrightarrow{\cong} B_\pi(0) \subset T_v(\mathbb{S}^n)$$

then provides a chart for the smooth structure on \mathbb{S}^n different from that given by stereographic projection, and quite useful for some purposes.

The connection between geodesics and shortest paths is given in the following, which is proven in [3].

Theorem 11.4.6. *Suppose the exponential map $\exp_x : B_r(0) \rightarrow M$ is a diffeomorphism onto a neighborhood of x in M . Let $y = \exp_x(v)$ for $v \in B_r(0)$. Then $\gamma_{x,v} : [0, 1] \rightarrow M$ is a distance minimizing path from x to y in M . Moreover, if $\delta : [a, b] \rightarrow M$ is another piecewise-smooth, distance minimizing path from x to y , then*

$$\gamma_{x,v}([0, 1]) = \delta([a, b]),$$

i.e., these two paths have the same image.

Recall from (11.2.2) that the Riemannian distance from x to y in M is given by $d(x, y) = \inf_\gamma \ell(\gamma)$ as γ varies over all the piecewise smooth paths from x to y .

Corollary 11.4.7. *Suppose the exponential map $\exp_x : B_r(0) \rightarrow M$ is a diffeomorphism onto a neighborhood of x in M . Let $y = \exp_x(v)$ for $v \in B_r(0)$. Then the Riemannian distance from x to y is $\|v\|_x$.*

Proof. $\gamma_{x,v}$ has constant speed $\|v\|_x$, so

$$\ell(\gamma_{x,v}|_{[0,1]}) = \int_0^1 \|\gamma'_{x,v}(t)\|_{\gamma_{x,v}(t)} dt = \int_0^1 \|v\|_x dt = \|v\|_x.$$

By Theorem 11.4.6, this is $d(x, y)$. □

We may use Corollary 11.4.7 to get an explicit calculation of the Riemannian distance function for the standard Riemannian metric on \mathbb{S}^n .

Corollary 11.4.8. *Let $v, w \in \mathbb{S}^n$. Then the Riemannian distance (with respect to the standard metric) from v to w is*

$$(11.4.2) \quad d(v, w) = \cos^{-1} \langle v, w \rangle,$$

where $\langle v, w \rangle$ is the standard inner product of v and w as vectors in \mathbb{R}^{n+1} .

Proof. First assume $w \neq -v$ so that $w = \exp_v(z)$ for some $z \in B_\pi(0)$. Note we may identify $T_v(\mathbb{S}^n)$ with $\{v\}^\perp = \{z \in \mathbb{R}^{n+1} : \langle v, z \rangle = 0\}$ and that the Riemannian metric on $T_v(\mathbb{S}^n)$ coincides with the restriction to this subspace of the standard inner product on \mathbb{R}^{n+1} .

Let $\|z\| = c$ and let $u = \frac{z}{c}$. Then the geodesic path from v to w parametrized by arc length is $\gamma_{v,u} : [0, c] \rightarrow \mathbb{S}^n$, where

$$\gamma_{v,u}(t) = \cos t v + \sin t u.$$

The length of this path is c , which is thus equal to $d(v, w)$. Moreover, $w = \gamma_{v,u}(c)$, so

$$\langle v, w \rangle = \langle v, \cos c v + \sin c u \rangle = \langle v, \cos c v \rangle = \cos c,$$

as $\langle v, u \rangle = 0$ because $u \in T_v(\mathbb{S}^n)$. So (11.4.2) follows.

The remaining case is $w = -v$. Let $\gamma : [a, b] \rightarrow \mathbb{S}^n$ be a piecewise-smooth path from v to $-v$. Then $t \mapsto \langle v, \gamma(t) \rangle$ is a path from 1 to -1 in \mathbb{R} . By the intermediate value theorem there is a $c \in (a, b)$ with $\langle v, \gamma(c) \rangle = 0$. Let $u = \gamma(c)$. Then

$$\ell(\gamma) = \ell(\gamma|_{[a,c]}) + \ell(\gamma|_{[c,b]}) \geq d(v, u) + d(u, -v) = \frac{\pi}{2} + \frac{\pi}{2} = \pi,$$

where the first equality comes from (11.4.2) applied at v and at $-v$. Of course, $\cos^{-1} \langle v, -v \rangle = \pi$, and it suffices to display a piecewise smooth path of length π from v to $-v$. But the geodesic path $\gamma_{v,z} : [0, \pi] \rightarrow \mathbb{S}^n$ will do, for any $z \in \mathbb{S}^n$ orthogonal to v . \square

Remark 11.4.9. Suppose $\exp_x : B_r(0) \rightarrow M$ is a diffeomorphism onto a neighborhood of x . Then for $v \in B_r(0)$, $d(\exp_x(0), \exp_x(v)) = \|v\|_x$, which is the distance from 0 to v in $T_x(M)$. It would be tempting to guess that $\exp_x : B_r(0) \rightarrow M$ is distance preserving with respect to the Riemannian distance, and perhaps even an isometric immersion. Here, we give $B_r(0)$ the subspace metric in $T_x(M)$.

More advanced differential geometry says that \exp_x is not an isometric immersion, as the curvature in $T_x(M)$ is 0, while the curvature near x in M need not be.

But we can show directly that $\exp_x : B_r(0) \rightarrow M$ does not, in general preserve Riemannian distance. For instance, in \mathbb{S}^2 , consider

$$\exp_{e_1} : B_\pi(0) \rightarrow \mathbb{S}^2.$$

By Corollary 11.4.8, $\exp_{e_1}(\frac{\pi}{2}e_2) = e_2$ and $\exp_{e_1}(\frac{\pi}{2}e_3) = e_3$, with e_1, e_2, e_3 the canonical basis. Thus,

$$d\left(\exp_{e_1}\left(\frac{\pi}{2}e_2\right), \exp_{e_1}\left(\frac{\pi}{2}e_3\right)\right) = d(e_2, e_3) = \frac{\pi}{2},$$

but the distance from $\frac{\pi}{2}e_2$ to $\frac{\pi}{2}e_3$ in $T_{e_1}(\mathbb{S}^2)$ is $\sqrt{2}\frac{\pi}{2}$, via the straight-line path between them (or simply by the usual distance formula in $T_{e_1}(\mathbb{S}^2) \subset \mathbb{R}^3$).

Theorems 11.4.6 and 11.3.21, work together to give very powerful result.

Theorem 11.4.10. *Let $f : M \rightarrow M$ be an isometry of the Riemannian manifold M . Then f preserves geodesics, i.e., if $\gamma : (a, b) \rightarrow M$ is geodesic, so is $f \circ \gamma$.*

Proof. By Theorem 11.3.14, being geodesic is a local property in the sense that if γ is geodesic on a small subinterval around each $t \in (a, b)$ then γ is geodesic on all of (a, b) .

By Theorem 11.4.6, γ is distance minimizing on both $[t - \epsilon, t]$ and $[t, t + \epsilon]$ for some $\epsilon > 0$. By Proposition 11.2.9 (and its proof), $f \circ \gamma$ is distance minimizing on these same intervals, and hence geodesic on these intervals by Theorem 11.3.21.

But $f \circ \gamma$ is smooth, so the derivatives of its restrictions to these two subintervals must agree at t . By uniqueness of geodesics, $f \circ \gamma$ is geodesic on $[t - \epsilon, t + \epsilon]$. \square

We obtain the following.

Corollary 11.4.11. *Let $f : M \rightarrow M$ be an isometry of the Riemannian manifold M . Suppose that $\exp_x : B_r(0) \rightarrow M$ is a diffeomorphism onto a neighborhood of x . Then $\exp_{f(x)} : B_r(0) \rightarrow M$ is a diffeomorphism onto a neighborhood of $f(x)$ (for the same r) and the following diagram commutes:*

$$(11.4.3) \quad \begin{array}{ccc} B_r(0) & \xrightarrow{T_x f} & B_r(0) \\ \exp_x \downarrow & & \downarrow \exp_{f(x)} \\ M & \xrightarrow{f} & M. \end{array}$$

Of course, the $B_r(0)$ on the left is the ball of radius r in $T_x(M)$, while that on the right is in $T_{f(x)}(M)$. In other words, $f \circ \exp_x = \exp_{f(x)} \circ T_x f$.

Proof. $T_x f$ is a linear isometric isomorphism of inner product spaces and f is a diffeomorphism, so it suffices to show the diagram commutes. Let $v \in B_r(0) \subset T_x(M)$. Then $f \circ \gamma_{x,v}$ is the geodesic taking 0 to $f(x)$ whose velocity vector at 0 is $T_x f(v)$. The result follows. \square

Using a little point-set topology we can generalize the invocation to the intermediate value theorem in the argument for Corollary 11.4.8 to obtain a strengthening of Corollary 11.4.7. See [8] for details on the theory.

Theorem 11.4.12. *Suppose the exponential map $\exp_x : B_r(0) \rightarrow M$ is a diffeomorphism onto a neighborhood of x in M . Then*

$$(11.4.4) \quad \exp_x(B_r(0)) = \{y \in M : d(x, y) < r\}.$$

Thus, the Riemannian distance is positive-definite (i.e., $d(x, y) = 0$ implies $x = y$), and defines a metric for a topology as in as in Definition A.1.1. Moreover the topology induced by this distance agrees with the original one.

Proof. We use the general definition of a topological manifold given in Definition A.5.5. In particular, we use the Hausdorff property. We first show that if $y \in M \setminus \exp_x(B_r(0))$, then $d(x, y) \geq r$. This implies (11.4.4).

Let $\epsilon > 0$ and let $s = r - \epsilon$. Let $\bar{B}_s(0)$ be the closed ball of radius s in $T_x(M)$:

$$\bar{B}_s(0) = \{v : \|v\| \leq s\}.$$

Then $\bar{B}_s(0)$ is closed and bounded in $T_x(M)$, and hence is compact by the Heine–Borel theorem. Since $\exp_x : B_r(0) \rightarrow M$ is a diffeomorphism onto a neighborhood of x in M , $\exp_x(\bar{B}_s(0))$ is a compact subset of M . Since M is Hausdorff, $\exp_x(\bar{B}_s(0))$ is closed in M . So the subset $V = M \setminus \exp_x(\bar{B}_s(0))$ is open in M , as is $U = \exp_x(B_s(0))$. U and V are disjoint, with $x \in U$ and $y \in V$. Since $x \in U$ and $y \in V$, no continuous path from x to y in M can lie entirely in $U \cup V$ by the connectedness of a closed interval. In particular, any piecewise smooth path $\gamma : [a, b] \rightarrow M$ from x to y must meet

$$M \setminus (U \cup V) = \exp_x(S_s(0)),$$

where $S_s(0)$ is the sphere of radius s in $T_x(M)$:

$$S_s(0) = \{v : \|v\| = s\}.$$

In particular, given γ as above, let $c \in (a, b)$ with $\gamma(c) = \exp_x(v)$ with $\|v\| = s$. Then

$$\ell(\gamma) = \ell(\gamma|_{[a, c]}) + \ell(\gamma|_{[c, b]}) \geq \ell(\gamma|_{[a, c]}) \geq d(x, \gamma(c)) = s,$$

with the last equality coming from Corollary 11.4.7. As observed in Remark 11.2.8, this implies the Riemannian distance satisfies the properties for a topological metric in Definition A.1.1.

By (11.4.4), for a given x and for small enough ϵ , the ball of radius ϵ with respect to the metric induced by d coincides with $\exp_x(B_\epsilon(0))$, which is open in M , so open sets in the metric topology induced by d are open in M and vice versa, as any open set of M containing x must contain $\exp_x(B_\epsilon(0))$ for some $\epsilon > 0$. \square

12. Hyperbolic geometry

We use the upper half plane model \mathbb{H} for hyperbolic space. Here,

$$(12.0.5) \quad \mathbb{H} = \{z = x + iy \in \mathbb{C} : x, y \in \mathbb{R} \text{ and } y > 0\}.$$

We write $x = \operatorname{Re}(z)$ and $y = \operatorname{Im}(z)$. We give \mathbb{H} the Riemannian metric

$$(12.0.6) \quad \langle v, w \rangle_z = \frac{1}{\operatorname{Im}(z)^2} (v \cdot w),$$

where $v \cdot w$ is the ordinary dot product of the vectors $v, w \in \mathbb{R}^2$. In other words, we scale the usual Euclidean inner product so it gets larger and larger as the point of origin of the two vectors approaches the x -axis.

In particular, if $\gamma : (a, b) \rightarrow \mathbb{H}$ is a piecewise smooth curve, its speed is given by

$$(12.0.7) \quad \begin{aligned} \|\gamma'(t)\|_{\gamma(t)} &= \sqrt{\langle \gamma'(t), \gamma'(t) \rangle_{\gamma(t)}} \\ &= \frac{\sqrt{\gamma'(t) \cdot \gamma'(t)}}{\operatorname{Im}(\gamma(t))} \\ &= \frac{\sqrt{(\gamma'_1(t))^2 + (\gamma'_2(t))^2}}{\gamma_2(t)}, \end{aligned}$$

when $\gamma(t) = \gamma_1(t) + i\gamma_2(t)$ with $\gamma_i(t)$ real for $i = 1, 2$. Because of the denominator, this depends strongly on the value of $\gamma(t)$. For instance, consider

$$\begin{aligned} \gamma : (0, \infty) &\rightarrow \mathbb{H} \\ \gamma(t) &= it. \end{aligned}$$

Then $\gamma'(t) = e_2$ for all t , but $\|\gamma'(t)\|_{\gamma(t)} = \frac{1}{t}$, and the velocity vector gets longer and longer as $t \rightarrow 0$.

For a piecewise smooth curve $\gamma : [a, b] \rightarrow \mathbb{H}$, we call $\|\gamma'(t)\|_{\gamma(t)}$ the hyperbolic length of $\gamma'(t)$, in distinction to its usual Euclidean length. We use it to calculate the hyperbolic arc length of γ :

$$(12.0.8) \quad \ell(\gamma) = \int_a^b \|\gamma'(t)\|_{\gamma(t)} dt$$

as in (11.2.1). As in spherical geometry, the hyperbolic distance between two points in \mathbb{H} is then given as

$$(12.0.9) \quad d_{\mathbb{H}}(z, w) = \inf_{\gamma} \ell(\gamma),$$

as γ ranges over all the piecewise smooth paths in \mathbb{H} from z to w , and such a path γ is said to be distance minimizing if $d_{\mathbb{H}}(z, w) = \ell(\gamma)$. Distance minimizing paths are what's known as geodesics, as studied in Chapter 11.

We will calculate these geodesics. They are the correct analogue of straight lines in studying the geometry of hyperbolic space. And using them to make a geometry of lines produces results very different from the geometry of ordinary lines in Euclidean space. Indeed, the hyperbolic geodesics do not

satisfy the parallel postulate of Euclidean geometry. Instead, they satisfy its negation. So \mathbb{H} provides an analytic geometric realization of “non-Euclidean geometry”.

Our approach to developing the geometry here will be by starting with the group of isometries.

12.1. Boundary of \mathbb{H} and compactification of \mathbb{C} . It is useful to think about the hyperbolic space \mathbb{H} as having a boundary. One might be tempted to use the x -axis for this purpose, but it is more useful to close it up so the boundary is a circle.

First, let us adjoin a formal point, ∞ , to \mathbb{C} and write

$$\bar{\mathbb{C}} = \mathbb{C} \cup \{\infty\}.$$

We shall interpret this via the stereographic projection map of (8.3.1):

$$h_U : U \xrightarrow{\cong} \mathbb{C}.$$

Here, $U = \mathbb{S}^2 \setminus \{N\}$, $N = e_3$ is the north pole, and we are identifying \mathbb{R}^2 with \mathbb{C} so that

$$h_U(x_1e_1 + x_2e_2 + x_3e_3) = \frac{1}{1 - x_3}(x_1 + ix_2)$$

by (8.3.1). We use this to identify $\bar{\mathbb{C}}$ with \mathbb{S}^2 via the function $\tilde{h}_U : \mathbb{S}^2 \rightarrow \bar{\mathbb{C}}$,

$$(12.1.1) \quad \tilde{h}_U(x) = \begin{cases} h_U(x) & x \neq N, \\ \infty & x = N. \end{cases}$$

Since \tilde{h}_U is bijective and restricts to a diffeomorphism of U onto \mathbb{C} , we may use it to endow $\bar{\mathbb{C}}$ with the structure of a smooth manifold.¹⁸ In other words we shall identify $\bar{\mathbb{C}}$ with the smooth manifold \mathbb{S}^2 via \tilde{h}_U . Since $h_U : U \xrightarrow{\cong} \mathbb{C}$ is a diffeomorphism, this agrees with the usual smooth structure on \mathbb{C} .

Thus, we shall declare a map $f : \bar{\mathbb{C}} \rightarrow \bar{\mathbb{C}}$ to be smooth if the composite

$$(12.1.2) \quad \mathbb{S}^2 \xrightarrow{\tilde{h}_U} \bar{\mathbb{C}} \xrightarrow{f} \bar{\mathbb{C}} \xrightarrow{\tilde{h}_U^{-1}} \mathbb{S}^2$$

is smooth. This agrees with the usual notion of smoothness on $f|_{f^{-1}(\mathbb{C})} : f^{-1}(\mathbb{C}) \rightarrow \mathbb{C}$.

Let $W \subset \mathbb{S}^2$ be given by

$$(12.1.3) \quad W = \{w \in \mathbb{S}^2 : \langle w, e_2 \rangle > 0\}.$$

Then

$$(12.1.4) \quad \tilde{h}_U|_W : W \xrightarrow{\cong} \mathbb{H}$$

is a diffeomorphism. Moreover, The open hemisphere W has an obvious boundary in \mathbb{S}^2 : the great circle Σ with pole e_2 .

$$\Sigma = \{w \in \mathbb{S}^2 : \langle w, e_2 \rangle = 0\}.$$

¹⁸The topology on $\bar{\mathbb{C}}$ induced by this identification coincides with the *one-point compactification* construction in point-set topology. See [8]

Note that

$$(12.1.5) \quad h_U|_{\Sigma \setminus \{N\}} : \Sigma \setminus \{N\} \xrightarrow{\cong} \mathbb{R}$$

is a diffeomorphism of $\Sigma \setminus \{N\}$ onto the x -axis. We define the boundary $\partial\mathbb{H}$ of \mathbb{H} by

$$(12.1.6) \quad \partial\mathbb{H} = \mathbb{R} \cup \{\infty\} \subset \bar{\mathbb{C}},$$

the union of the x -axis with the point at infinity. Then

$$(12.1.7) \quad \tilde{h}_U|_{\Sigma} : \Sigma \xrightarrow{\cong} \partial\mathbb{H}.$$

We write

$$(12.1.8) \quad \bar{\mathbb{H}} = \mathbb{H} \cup \partial\mathbb{H},$$

and set $\bar{W} = W \cup \Sigma = \{w \in \mathbb{S}^2 : \langle w, e_2 \rangle \geq 0\}$, the closed hemisphere containing W . Then

$$(12.1.9) \quad \tilde{h}_U|_{\bar{W}} : \bar{W} \xrightarrow{\cong} \bar{\mathbb{H}}.$$

In order to the smoothness of the composites (12.1.2), we need to consider the charts in \mathbb{S}^2 defined near the north pole. The setting we have already considered works as follows in this context. We set $V = \mathbb{S}^2 \setminus \{S\}$ with $S = -e_3$, the south pole. Corollary 8.3.7 gives a chart $h_V : V \xrightarrow{\cong} \mathbb{C}$, given by

$$h_V(x_1e_1 + x_2e_2 + x_3e_3) = \frac{1}{1+x_3}(x_1 + ix_2).$$

This, actually, would be quite sufficient for our purposes, as (8.3.3) then gives

$$h_U \circ h_V^{-1}(z) = \frac{z}{\langle z, z \rangle},$$

where $\langle z, z \rangle$ is the real inner product of z with itself, which, if $z = x + iy$ with $x, y \in \mathbb{R}$, is $x^2 + y^2$. But this is the result of complex multiplication $z\bar{z}$, where \bar{z} is the complex conjugate of z : $\bar{z} = x - iy$. We obtain

$$(12.1.10) \quad h_U \circ h_V^{-1}(z) = \frac{z}{\langle z, z \rangle} = \frac{z}{z\bar{z}} = \frac{1}{\bar{z}},$$

with the identical formula holding for $h_V \circ h_U^{-1}$.

This would indeed be sufficient for our purposes, but it is nicer to replace $h_V(x)$ with its complex conjugate. Write $\Gamma : \mathbb{C} \rightarrow \mathbb{C}$ for complex conjugation: $\Gamma(z) = \bar{z}$. Note that Γ is smooth, as its Jacobian matrix is $D\Gamma(z) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ for all $z \in \mathbb{C}$. Thus, Γ is a diffeomorphism of \mathbb{C} whose square is the identity, so we can replace h_V by

$$(12.1.11) \quad k_V = \Gamma \circ h_V : V \xrightarrow{\cong} \mathbb{C}$$

$$k_V(x_1e_1 + x_2e_2 + x_3e_3) = \frac{1}{1+x_3}(x_1 - ix_2).$$

A straightforward computation from (12.1.10) now gives:

Proposition 12.1.1. $\{h_U : U \xrightarrow{\cong} \mathbb{C}, k_V : V \xrightarrow{\cong} \mathbb{C}\}$ gives an atlas for the standard smooth structure on \mathbb{S}^2 with the property that the transition maps are given by

$$(12.1.12) \quad h_U \circ k_V^{-1}(z) = k_V \circ h_U^{-1}(z) = \frac{1}{z}$$

for all $z \in \mathbb{C} \setminus \{0\}$.

$\mathbb{C} \setminus \{0\}$ is, of course the domain as well as the codomain for each of the two transition maps. The advantage of using these charts is that the transition maps are holomorphic (complex differentiable) functions. We shall review a bit of complex analysis.

If $f : U \rightarrow \mathbb{C}$ with $U \subset \mathbb{C}$ open, we say f is complex differentiable at z if

$$f'(z) = \lim_{h \rightarrow 0} \frac{f(z+h) - f(z)}{h}$$

exists. We say f is holomorphic on U if it is complex differentiable at each $z \in U$.

In this case, f is smooth and its real Jacobian matrix can be computed from what are called the Cauchy–Riemann equations. These say that if we write $f(z) = u(z) + iv(z)$, with $u(z), v(z)$ real, then

$$\begin{aligned} \frac{\partial u}{\partial x} &= \operatorname{Re}(f'(z)) & \frac{\partial u}{\partial y} &= -\operatorname{Im}(f'(z)) \\ \frac{\partial v}{\partial x} &= \operatorname{Im}(f'(z)) & \frac{\partial v}{\partial y} &= \operatorname{Re}(f'(z)), \end{aligned}$$

so the Jacobian matrix is given by

$$(12.1.13) \quad Df(z) = \begin{bmatrix} \operatorname{Re}(f'(z)) & -\operatorname{Im}(f'(z)) \\ \operatorname{Im}(f'(z)) & \operatorname{Re}(f'(z)) \end{bmatrix}.$$

There is a nicer, more conceptual way to express this. Think of $[f'(z)]$ as a 1×1 complex matrix, so it induces a \mathbb{C} -linear function from \mathbb{C} to itself by matrix multiplication (which is simply ordinary multiplication in the 1×1 case. But any complex linear vector space is a real vector space by restriction of the ground field. And any \mathbb{C} -linear function between complex vector spaces is \mathbb{R} -linear. Moreover, if v_1, \dots, v_n is a \mathbb{C} -basis for a complex vector space, then $v_1, iv_1, \dots, v_n, iv_n$ is an \mathbb{R} -basis for its underlying real vector space. We obtain a ring homomorphism

$$\rho : M_n(\mathbb{C}) \rightarrow M_{2n}(\mathbb{R})$$

by expressing the linear function induced by an $n \times n$ complex matrix in terms of the \mathbb{R} -basis $e_1, ie_1, \dots, e_n, ie_n$. Here, ρ stands for “realification”.

What the Cauchy–Riemann equations say is that if $f : U \rightarrow \mathbb{C}$ is holomorphic, with U an open subset of \mathbb{C} , then if we regard it as a smooth function $f : U \rightarrow \mathbb{R}^2$, then

$$(12.1.14) \quad Df(z) = \rho([f'(z)]),$$

the realification of the complex matrix $[f'(z)]$.

Complex differentiation is complex linear on functions and satisfies the usual product, quotient and chain rules, so we can compute the complex derivative of a rational function (a quotient of two polynomials) in the usual way. As in the real case, the complex derivative of a convergent complex power series $\sum_{k=0}^{\infty} a_k x^k$ is the term by term derivative $\sum_{k=1}^{\infty} k a_k x^{k-1}$.

One big advantage of the atlas given in Proposition 12.1.1 is that the transition maps are holomorphic. This gives \mathbb{S}^2 the structure of a complex manifold (a structure we will not investigate in detail here) and allows us to say that a function $f : \mathbb{S}^2 \rightarrow \mathbb{S}^2$ is holomorphic if all the composites given as in (8.4.2) (as we allow the charts to vary) are holomorphic (and not just smooth).

12.2. Möbius transformations. As an open subset of $\mathbb{C} = \mathbb{R}^2$, hyperbolic space has an intrinsic orientation. We shall see that the Möbius transformations are the orientation-preserving isometries of \mathbb{H} . Recall that

$$\mathrm{SL}_2(\mathbb{R}) = \{A \in \mathrm{GL}_2(\mathbb{R}) : \det A = 1\}.$$

Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2(\mathbb{R})$. We define the Möbius transformation $\varphi_A : \bar{\mathbb{C}} \rightarrow \bar{\mathbb{C}}$ by

$$(12.2.1) \quad \varphi_A(z) = \begin{cases} \frac{az+b}{cz+d} & z \neq -\frac{d}{c}, \infty, \\ \infty & z = -\frac{d}{c}, \\ \frac{a}{c} & z = \infty. \end{cases}$$

Such a map is sometimes called a fractional linear transformation. We write $\mathrm{Möb}$ for the set of all Möbius transformations.

Lemma 12.2.1. *Let $A, B \in \mathrm{SL}_2(\mathbb{R})$. Then,*

$$(12.2.2) \quad \varphi_A \circ \varphi_B = \varphi_{AB}.$$

Proof. Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and $B = \begin{bmatrix} r & s \\ t & u \end{bmatrix}$. Then

$$\varphi_A \circ \varphi_B(z) = \frac{a \left(\frac{rz+s}{tz+u} \right) + b}{c \left(\frac{rz+s}{tz+u} \right) + d} = \frac{a(rz+s) + b(tz+u)}{c(rz+s) + d(tz+u)}.$$

This simplifies to $\varphi_{AB}(z)$. The other cases may be computed by hand. \square

Corollary 12.2.2. *Möb is a group under composition. There is a group homomorphism*

$$\varphi : \mathrm{SL}_2(\mathbb{R}) \rightarrow \mathrm{Möb}$$

via $\varphi(A) = \varphi_A$. The kernel of φ is $\{\pm I_2\}$. Thus, $\varphi(A) = \varphi(-A)$ for all $A \in \mathrm{SL}_2(\mathbb{R})$.

Proof. Composition of functions is associative, with identity element $\text{id} = \varphi_{I_2}$. Now $\varphi_A \circ \varphi_{A^{-1}} = \varphi_{A^{-1}} \circ \varphi_A = \varphi_{I_2}$. So Möb is a group, and φ is a homomorphism. $\pm I_2$ are obviously contained in its kernel. The next result shows nothing else is. \square

Complex derivatives will allow us to calculate the Jacobian matrices of Möbius transformations.

Lemma 12.2.3. *Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \text{SL}_2(\mathbb{R})$. Then the complex derivative of φ_A is given by*

$$(12.2.3) \quad \varphi'_A(z) = \frac{1}{(cz + d)^2}$$

for $z \neq \infty, -\frac{d}{c}$. Thus, φ_A is holomorphic on \mathbb{H} , with nonzero complex derivative.

Proof. Just apply the quotient rule.

$$\varphi'_A(z) = \frac{a(cz + d) - c(az + b)}{(cz + d)^2}$$

and apply that $\det A = 1$. Regarding the last statement, neither ∞ nor $-\frac{d}{c}$ lies in \mathbb{H} . \square

Recall that we may identify $\bar{\mathbb{C}}$ with \mathbb{S}^2 via $\tilde{h}_U : \mathbb{S}^2 \xrightarrow{\cong} \bar{\mathbb{C}}$. So we can ask and answer the following.

Corollary 12.2.4. *Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \text{SL}_2(\mathbb{R})$. Then $\varphi_A : \bar{\mathbb{C}} \rightarrow \bar{\mathbb{C}}$ is holomorphic.*

Proof. By Lemma 12.2.3, it suffices to show φ_A is holomorphic at ∞ and $-\frac{d}{c}$. Near $-\frac{d}{c}$ we can consider the composite

$$\bar{\mathbb{C}} \xrightarrow{\varphi_A} \bar{\mathbb{C}} \xrightarrow{\tilde{h}_U^{-1}} \mathbb{S}^2 \xrightarrow{k_V} \mathbb{C}.$$

This takes z to $\frac{cz+d}{az+b}$, which is holomorphic near $-\frac{d}{c}$ as A is invertible: $-\frac{d}{c} \neq -\frac{b}{a}$.

Near ∞ we study the composite

$$\mathbb{C} \xrightarrow{k_V} \mathbb{S}^2 \xrightarrow{\tilde{h}_U} \bar{\mathbb{C}} \xrightarrow{\varphi_A} \bar{\mathbb{C}}.$$

near $0 \in \mathbb{C}$. This composite takes z to

$$\frac{a\frac{1}{z} + b}{c\frac{1}{z} + d} = \frac{a + bz}{c + dz}.$$

For $c \neq 0$, this is covered by Lemma 12.2.3. For $c = 0$, we apply the other chart now in the target, getting $\frac{c+dz}{a+bz}$, and the result follows. \square

In particular, as maps from $\bar{\mathbb{C}}$ to $\bar{\mathbb{C}}$, Möbius transformations are continuous.

Definition 12.2.5. Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2(\mathbb{R})$. Write $\mathrm{tr}(A)$ for the trace of A .

- We say φ_A is *parabolic* if $|\mathrm{tr}(A)| = 2$.
- We say φ_A is *hyperbolic* if $|\mathrm{tr}(A)| > 2$.
- We say φ_A is *elliptic* if $|\mathrm{tr}(A)| < 2$.

Note this is well-defined as $\ker \varphi = \{\pm I_2\}$. This implies $\varphi_A = \varphi_B$ if and only if $B = \pm A$. Note also that the identity map has been classified as parabolic.

These three types of transformations can be told apart by their fixed-points. We first characterize the Möbius transformations fixing ∞ .

Proposition 12.2.6. *The Möbius transformations fixing ∞ have the form*

$$(12.2.4) \quad \mathrm{Möb}_\infty = \left\{ \varphi_A : A = \begin{bmatrix} a & b \\ 0 & \frac{1}{a} \end{bmatrix} \text{ with } a, b \in \mathbb{R}, a > 0 \right\}.$$

Moreover, for A in (12.2.4) we have

$$(12.2.5) \quad \varphi_A(z) = a^2 z + ab.$$

Indeed, any degree 1 polynomial of the form $f(z) = rz + s$ with $r, s \in \mathbb{R}$ and $r > 0$ may be written in this form.

For $a = 1$, $\varphi_A(z) = z + b$ is parabolic. When $b \neq 0$, ∞ is the only fixed-point of φ_A .

For $a \neq 1$, φ_A is hyperbolic. Its fixed-points are ∞ and $\frac{ab}{1-a^2}$. In particular, the Möbius transformations fixing 0 and ∞ have the form $f(z) = cz$ for $1 \neq c > 0$ in \mathbb{R} .

In the general case of (12.2.4), we have

$$(12.2.6) \quad A = \begin{bmatrix} 1 & ab \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & \frac{1}{a} \end{bmatrix} = A_1 A_2.$$

So $\varphi_A = \varphi_{A_1} \circ \varphi_{A_2}$, the composite of a parabolic transformation fixing ∞ and a hyperbolic transformation fixing 0 and ∞ (or the identity, if $a = 1$).

Proof. For a general matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2(\mathbb{R})$, we have $\varphi_A(\infty) = \frac{a}{c}$, so φ_A fixes ∞ if and only if $c = 0$. Since $\det A = 1$, this also forces $d = \frac{1}{a}$. Finally, since $\varphi_A = \varphi_{-A}$, We may assume $a > 0$. This establishes (12.2.4), and (12.2.5) follows by direct calculation.

When $a = 1$, $\mathrm{tr}(A) = 2$, hence φ_A is parabolic and $\varphi_A(z) = z + b$ by (12.2.5). On \mathbb{C} , this is just translation by $b \in \mathbb{R}$, so if $b \neq 0$, there are no fixed-points in \mathbb{C} .

For a general A in (12.2.4), $\mathrm{tr}(A) = \frac{a^2+1}{a}$. Consider $g : (0, \infty) \rightarrow (0, \infty)$, given by $g(a) = \frac{a^2+1}{a}$. Then $g'(a) = \frac{a^2-1}{a^2}$ is negative on $(0, 1)$ and positive on $(1, \infty)$, so g has an absolute minimum at 1. In particular $\mathrm{tr}(A) > 2$ for $a \neq 1$, hence φ_A is hyperbolic for $a \neq 1$. In this case, for $z \in \mathbb{C}$, $\varphi_A(z) = z$ if and only if $(a^2 - 1)z = -ab$ by (12.2.5), so the fixed-point calculation is as stated. \square

Proposition 12.2.7. *Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2(\mathbb{R})$, with $A \neq \pm I_2$. Then φ_A has either one or two fixed-points in $\bar{\mathbb{C}}$:*

- (1) *If φ_A is parabolic, then φ_A has exactly one fixed-point in $\bar{\mathbb{C}}$, either ∞ or a point in the x -axis, \mathbb{R} (i.e., the fixed-point is on $\partial\mathbb{H}$).*
- (2) *If φ_A is hyperbolic, then φ_A has two fixed-points in $\bar{\mathbb{C}}$, both on $\partial\mathbb{H}$.*
- (3) *If φ_A is elliptic, then there are two fixed-points in $\bar{\mathbb{C}}$. They are complex conjugates, and one lies in \mathbb{H} and the other in $\mathbb{C} \setminus \bar{\mathbb{H}}$.*

Proof. The case $c = 0$ (i.e., ∞ is fixed) is treated in Proposition 12.2.6. So we assume $c \neq 0$. In this case, $\varphi_A(z) = z$ if and only if $az + b = cz^2 + dz$. This gives a quadratic whose solutions are

$$z = \frac{(a-d) \pm \sqrt{a^2 - 2ad + d^2 + 4bc}}{2c}.$$

Since $\det A = 1$, $bc = ad - 1$, and this is equivalent to

$$(12.2.7) \quad z = \frac{(a-d) \pm \sqrt{(a+d)^2 - 4}}{2c}.$$

Thus, there are two distinct real roots if $|\mathrm{tr}(A)| > 2$, a single, repeated real root if $|\mathrm{tr}(A)| = 2$, and a pair of complex conjugate roots if $|\mathrm{tr}(A)| < 2$. \square

There are no situations above in which two fixed-points lie in \mathbb{H} , so we obtain the following.

Corollary 12.2.8. *Any Möbius transformation fixing more than one element of \mathbb{H} is the identity. Similarly any Möbius transformation fixing single element of \mathbb{H} along with at least one point of $\partial\mathbb{H}$ is the identity.*

This now allows the following.

Corollary 12.2.9. *Let f and g be Möbius transformations that agree on two distinct points in \mathbb{H} . Then $f = g$. Similarly, if f and g agree on one point of \mathbb{H} and one point of $\partial\mathbb{H}$, then $f = g$.*

Proof. If f and g agree on both z and w , then gf^{-1} is a Möbius transformation fixing z and w . Apply Corollary 12.2.8. \square

We now show that the Möbius transformations act on hyperbolic space.

Proposition 12.2.10. *Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2(\mathbb{R})$ with $c \neq 0$. Then we can factor A as a composite*

$$(12.2.8) \quad A = \begin{bmatrix} 1 & \frac{a}{c} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{c} & 0 \\ 0 & c \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & \frac{d}{c} \\ 0 & 1 \end{bmatrix} = A_1 A_2 A_3 A_4.$$

The transformations φ_{A_1} and φ_{A_4} are parabolic fixing ∞ . If $c \neq \pm 1$, φ_{A_2} is hyperbolic fixing ∞ and 0 (otherwise it is the identity). φ_{A_3} is elliptic fixing $\pm i$. Each φ_{A_k} , $k = 1, \dots, 4$, gives a diffeomorphism

$$\varphi_{A_k} : \mathbb{H} \xrightarrow{\cong} \mathbb{H}.$$

We also have $\varphi_{A_k} : \bar{\mathbb{H}} \rightarrow \bar{\mathbb{H}}$. Thus, the same is true for φ_A .

Proof. A direct calculation shows $A_1 A_2 A_3 A_4 = \begin{bmatrix} a & \frac{ad-1}{c} \\ c & d \end{bmatrix}$. Since $\det A = 1$, this is A . We now study these matrices case by case. For $B = \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix}$, φ_B translates the plane parallel to the x -axis, thus preserving \mathbb{H} , the x -axis and ∞ .

For $B = \begin{bmatrix} a & 0 \\ 0 & \frac{1}{a} \end{bmatrix}$, $\varphi_B(z) = a^2 z$. This preserves the sign of the pure imaginary part of z , and hence preserves \mathbb{H} and $\bar{\mathbb{H}}$, fixing ∞ . the map $\varphi_B : \mathbb{H} \rightarrow \mathbb{H}$ is a diffeomorphism, as B^{-1} has the same form.

The most interesting case is $B = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. Here $\varphi_B(z) = -\frac{1}{z} = -\frac{\bar{z}}{z\bar{z}}$, with \bar{z} the complex conjugate of z . An easy calculation shows $\text{Im}(-\frac{1}{z}) = \frac{\text{Im}(z)}{z\bar{z}}$. Since $z\bar{z} > 0$ for $z \neq 0$, this has the same sign as $\text{Im}(z)$. Again φ_B preserves \mathbb{H} and $\bar{\mathbb{H}}$. The map $\varphi_B : \mathbb{H} \rightarrow \mathbb{H}$ is a diffeomorphism as $B^{-1} = -B$, hence $\varphi_B^2 = \text{id}$. The fixed-points of φ_B are clear by direct calculation. \square

Notation 12.2.11. The Möbius transformations in Proposition 12.2.6 are important and will recur often in discussion, so we establish notation for them.

For $a \in \mathbb{R}$, we write $p_a : \bar{\mathbb{H}} \rightarrow \bar{\mathbb{H}}$ for the (parabolic) Möbius transformation induced by $\begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix}$:

$$(12.2.9) \quad p_a(z) = z + a.$$

For $0 < a \in \mathbb{R}$ write $h_a : \bar{\mathbb{H}} \rightarrow \bar{\mathbb{H}}$ for the Möbius transformation induced by $\begin{bmatrix} \sqrt{a} & 0 \\ 0 & \frac{1}{\sqrt{a}} \end{bmatrix}$:

$$(12.2.10) \quad h_a(z) = az.$$

This is hyperbolic when $a \neq 1$ and the identity for $a = 1$.

The output of Proposition 12.2.6 is that each element of Möb_∞ may be written uniquely in the form

$$(12.2.11) \quad \begin{aligned} f &= p_b h_a \\ z &\mapsto az + b \end{aligned}$$

for $b \in \mathbb{R}$ and $0 < a \in \mathbb{R}$, and that $p_b h_a$ is parabolic if $a = 1$ and hyperbolic if $a \neq 1$. Moreover, the Möbius transformations fixing both 0 and ∞ are precisely $\{h_a : 0 < a \in \mathbb{R}\}$.

Remarks 12.2.12.

- (1) As we shall see, the elliptic Möbius transformations fixing $z \in \mathbb{H}$ are precise analogues of the rotations about a particular point in the Euclidean plane.
- (2) An orientation-preserving isometry of \mathbb{R}^2 that is not a rotation is automatically a translation, and has no fixed-points. In the hyperbolic case, there are two families of Möbius transformations of \mathbb{H} with no fixed-points in \mathbb{H} : the parabolic ones and the hyperbolic ones. Neither is a precise analogue of a Euclidean translation, but perhaps the hyperbolic ones are closer. Parabolic transformations do not fix

any lines. Hyperbolic ones have a unique fixed line, and “translate” that line along itself. For instance, the hyperbolic transformation h_a preserves the line $\ell_0 = \{ti : t > 0\}$. As we shall see, this line is geodesic in \mathbb{H} .

- (3) The hyperbolic transformations h_a exhibit different behavior at their two fixed-points. The behavior depends on whether $a > 1$ or $a < 1$. Suppose $a > 1$. Then ∞ is what’s known as an attractor for h_a in the sense that for $0 \neq z \in \bar{\mathbb{H}}$,

$$\lim_{n \rightarrow \infty} h_a^n(z) = \infty,$$

where h_a^n is the composite of h_a with itself n times. Here, the limit may be taken in the usual sense of we regard $\bar{\mathbb{H}}$ as a subset of $\bar{\mathbb{C}} = \mathbb{S}^2$. That is equivalent to saying that $\lim_{n \rightarrow \infty} \|h_a^n(z)\| = \infty$, using the usual norm in \mathbb{C} .

The fixed-point 0, on the other hand, is a repulsor when $a > 1$: the iterates $h_a^n(z)$ get farther and farther away from 0 as n increases.)

When $a < 1$ it is 0 that is an attractor for h_a . For $z \in \bar{\mathbb{H}} \setminus \{\infty\}$,

$$\lim_{n \rightarrow \infty} h_a^n(z) = 0.$$

In this case, ∞ is a repulsor.

- (4) The fixed-point, ∞ , of the parabolic transformations p_a is also an attractor. For $z \in \bar{\mathbb{H}}$,

$$\lim_{n \rightarrow \infty} p_a^n(z) = \infty.$$

But some points start by moving away from ∞ before they move toward it.

The following makes the analogy between elliptic transformations and rotations explicit when the fixed-point is i .

Lemma 12.2.13. *The isotropy subgroup, Möb_i , of the complex number $i \in \mathbb{H}$ under the action of the Möbius transformations is*

$$(12.2.12) \quad \text{Möb}_i = \{\varphi_{R_\theta} : \theta \in \mathbb{R}\},$$

where $R_\theta = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$, the standard rotation matrix in SO_2 . Of course, $\varphi_{R_\theta} \varphi_{R_\psi} = \varphi_{R_{\theta+\psi}}$.

Proof. This amounts to solving for A in (12.2.7) when $z = i$. Here $\text{Re}(z) = 0$ when $a = d$. So the discriminant is $4a^2 - 4$ and must equal $-4c^2$. This reduces to $a^2 + c^2 = 1$. Since $a = d$ and $\det A = 1$, this forces $b = -c$ and the result follows. \square

We shall see later that the elements of finite order in Möb are all conjugate to elements of Möb_i . Note that $\varphi_{R_\pi} = \varphi_{-I_2} = \text{id}$, so $\theta \mapsto \varphi_{R_\theta}$ goes around the circle double-time. We have the following.

Lemma 12.2.14. *The transformation φ_{R_θ} has finite order if and only if θ is a rational multiple of π . If n and k are relatively prime, $\varphi_{R_{\frac{k}{n}\pi}}$ has order n .*

Proof. $(\varphi_{R_\theta})^m = \varphi_{R_{m\theta}}$ is the identity if and only if $R_{m\theta} = \pm I_2$. This holds if and only if $m\theta$ is an integral multiple of π . \square

Remark 12.2.15. We now have what we need to understand the isotropy subgroup

$$(12.2.13) \quad \text{Möb}_z = \{f \in \text{Möb} : f(z) = z\}$$

for any $z \in \bar{\mathbb{H}}$. Lemma 12.2.13 computes Möb_i and Proposition 12.2.6 computes Möb_∞ . By Lemma 3.6.4, if $f \in \text{Möb}$ and $z \in \bar{\mathbb{H}}$, then

$$(12.2.14) \quad \text{Möb}_{f(z)} = f \text{Möb}_z f^{-1},$$

the conjugate subgroup to Möb_z by f . Note that Möb_i and Möb_∞ are not conjugate, as the nonidentity elements of Möb_i are all elliptic, while the elements of Möb_∞ are all either hyperbolic or parabolic. But each type of Möbius transformation, be it elliptic, parabolic or hyperbolic, is preserved by conjugation, as matrix conjugation preserves the trace, and

$$\varphi_B \varphi_A \varphi_B^{-1} = \varphi_{BAB^{-1}}.$$

This does not contradict Lemma 3.6.4, as no Möbius transformation takes ∞ , which lies in $\partial\mathbb{H}$, to any element in \mathbb{H} . However, these calculations are sufficient to understand every isotropy group:

- Every element in $\partial\mathbb{H}$ has the form $f(\infty)$ for some Möbius transformation f . For $a \in \mathbb{R}$,

$$(12.2.15) \quad \varphi \begin{bmatrix} a & 0 \\ 1 & \frac{1}{a} \end{bmatrix} (\infty) = a.$$

- Every element of \mathbb{H} has the form $f(i)$ for some Möbius transformation f . For $a, b \in \mathbb{R}$ with $b > 0$,

$$(12.2.16) \quad \varphi \begin{bmatrix} \sqrt{b} & \frac{a}{\sqrt{b}} \\ 0 & \frac{1}{\sqrt{b}} \end{bmatrix} (i) = a + bi.$$

By (12.2.16), every elliptic transformation is conjugate to φ_{R_θ} for some θ . By (12.2.15), every parabolic or hyperbolic transformation f is conjugate to $p_b h_a$ for $a, b \in \mathbb{R}$ with $a > 0$. In particular if f is parabolic, it is conjugate to p_b for some $b \in \mathbb{R}$. If f is hyperbolic, we shall show in Proposition 12.4.21 that f is conjugate to h_a for some $a > 0$.

12.3. Isometric properties of Möbius transformations. We show that Möbius transformations are Riemannian isometries of \mathbb{H} (i.e., isometries in the strong sense of Definition 11.1.5). We could then deduce from Proposition 11.2.9 that Möbius transformations preserve arc length and Riemannian distance with respect to the hyperbolic metric (12.0.6). But since it is easy to do so, we shall prove this directly.

In other words, Möbius transformations are distance-preserving, and hence are isometries in the naive sense we've been using in Euclidean and spherical geometry.

To see that Möbius transformations are Riemannian isometries, we make use of the factorizations in (12.2.8) and (12.2.6). The only hard step concerns the elliptic transformation $f = \varphi_{A_3}$, with $A_3 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. Here,

$$f(z) = -\frac{1}{z} = -z^{-1}.$$

In particular, its complex derivative is given by $f'(z) = z^{-2}$. We shall apply this to our situation using the following trick. The proof is left to the reader.

Lemma 12.3.1. *Identify \mathbb{C} with \mathbb{R}^2 in the usual way so that $x + iy$ is identified with $\begin{bmatrix} x \\ y \end{bmatrix}$ for $x, y \in \mathbb{R}$. Then for $\zeta, \omega \in \mathbb{C}$, the standard real inner product $\langle \zeta, \omega \rangle$ may be calculated by*

$$(12.3.1) \quad \langle \zeta, \omega \rangle = \operatorname{Re}(\zeta \bar{\omega}),$$

where $\bar{\omega}$ is the complex conjugate of ω .

We use this in the following.

Lemma 12.3.2. *Let $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. Then $\varphi_A : \mathbb{H} \rightarrow \mathbb{H}$ is an isometry in the sense of Definition 11.1.5.*

Proof. Let $f(z) = \varphi_A(z) = -z^{-1}$. We must show that for $z \in \mathbb{H}$ and for v, w tangent vectors at z , we have

$$(12.3.2) \quad \langle Df(z)v, Df(z)w \rangle_{f(z)} = \langle v, w \rangle_z,$$

i.e.,

$$\frac{1}{(\operatorname{Im}(f(z)))^2} \langle Df(z)v, Df(z)w \rangle = \frac{1}{(\operatorname{Im}(z))^2} \langle v, w \rangle,$$

where this time the inner products are the ordinary inner products in \mathbb{R}^2 . Identifying the tangent space with \mathbb{C} and applying (12.3.1) along with our formula for f and hence f' . this says

$$\frac{1}{(\operatorname{Im}(-z^{-1}))^2} \operatorname{Re}(z^{-2} \zeta \bar{z}^{-2} \bar{\omega}) = \frac{1}{(\operatorname{Im}(z))^2} \operatorname{Re}(\zeta \bar{\omega})$$

for all $\zeta, \omega \in \mathbb{C}$. Since $z^{-2} \bar{z}^{-2}$ is real, ζ and ω drop out and it suffices to show that

$$(z \bar{z})^2 \operatorname{Im}(-z^{-1})^2 = (\operatorname{Im}(z))^2.$$

But $-z^{-1} = \frac{-\bar{z}}{z\bar{z}}$, so $\operatorname{Im}(-z^{-1}) = \frac{\operatorname{Im}(z)}{z\bar{z}}$, and the result follows. \square

This was the key step in the following.

Proposition 12.3.3. *Let $f : \mathbb{H} \rightarrow \mathbb{H}$ be Möbius transformation: $f = \varphi_A$ for $A \in \operatorname{SL}_2(\mathbb{R})$. Then f is an isometry in the sense of Definition 11.1.5. Specifically,*

$$(12.3.3) \quad \langle Df(z)v, Df(z)w \rangle_{f(z)} = \langle v, w \rangle_z,$$

for all $z \in \mathbb{H}$ and v, w tangent vectors at z .

Proof. It suffices to show this when A is one of the matrices in the factorizations (12.2.8) and (12.2.6). Lemma 12.3.2 treats the most difficult case. The others induce parabolic transformations of the form $p_b(z) = z + b$, $b \in \mathbb{R}$, or hyperbolic transformations of the form $h_a(z) = az$ for $a > 0$.

If $f = p_b$, then $f(z)$ has the same pure imaginary part as z , so the hyperbolic inner products at z and $f(z)$ are identical. Since the Jacobian matrix of p_b is the identity, p_b is an isometry.

If $f = h_a$ then $\text{Im}(\varphi_A(z)) = a \text{Im}(z)$. The Jacobian matrix of h_a induces multiplication by a , so the result follows as in Lemma 12.3.2. \square

We could now deduce a lot of important results from our chapter on Riemannian geometry, but we prefer to prove some of the easier results directly. The point is that \mathbb{H} is an open subset of \mathbb{R}^2 , and its geometry is given by the “local model” in which many proofs are easier.

First, we deduce that Möbius transformations preserve hyperbolic distance and are therefore isometries in the naive sense.

Definition 12.3.4. A function $f : \mathbb{H} \rightarrow \mathbb{H}$ is distance-preserving, or a *naive isometry* if

$$d_{\mathbb{H}}(f(z), f(w)) = d_{\mathbb{H}}(z, w) \quad \text{for all } z, w \in \mathbb{H}.$$

We write $\mathcal{I}(\mathbb{H})$ for the set of naive isometries of \mathbb{H} .

Remark 12.3.5. Note that while composition provides $\mathcal{I}(\mathbb{H})$ with an associative multiplication with identity element id , we do not yet know it is a group, as we have not assumed the naive isometries are surjective. (It is easy to show that if $f : \mathbb{H} \rightarrow \mathbb{H}$ is a distance-preserving surjection, then it is bijective and its inverse function is distance-preserving.)

We studied the naive isometries in our analysis of Euclidean and spherical geometry, and showed there that being a naive isometry is equivalent to being an isometry in the stronger sense of Definition 11.1.5. We shall show in Theorem 12.9.5 that the same is true in the hyperbolic case. As a result, we will see that the maps in $\mathcal{I}(\mathbb{H})$ are surjective and that $\mathcal{I}(\mathbb{H})$ is a group.

The first step is showing that Möbius transformations are naive isometries. This would follow from Proposition 12.3.3 and Proposition 11.2.9, but as discussed above, we prefer to give a direct proof.

Corollary 12.3.6. *Möbius transformations preserve arc length: if f is a Möbius transformation and $\gamma : [a, b] \rightarrow \mathbb{H}$ is piecewise smooth, then the arc lengths of γ and $f \circ \gamma$ are equal. By (12.0.9), $f \in \mathcal{I}(\mathbb{H})$. Thus, $\text{Möb} \subset \mathcal{I}(\mathbb{H})$.*

Proof. It suffices to show $\|(f \circ \gamma)'(t)\|_{f \circ \gamma(t)} = \|\gamma'(t)\|_{\gamma(t)}$. This is immediate from the chain rule and (12.3.3). \square

12.4. Hyperbolic lines and geodesics. We will first define the hyperbolic lines and then show they are parametrized by geodesic curves.

Definition 12.4.1. Let $a \in \mathbb{R}$. The hyperbolic line ℓ_a is given by

$$(12.4.1) \quad \ell_a = \{z \in \mathbb{H} : \operatorname{Re}(z) = a\} = \{a + bi : 0 < b \in \mathbb{R}\}.$$

We shall refer to the points $a, \infty \in \partial H$ as the *endpoints* or *points at infinity* of ℓ_a .

Let $a, r \in \mathbb{R}$ with $r > 0$. The hyperbolic line $\mathcal{C}_r(a)$ is given by

$$(12.4.2) \quad \mathcal{C}_r(a) = \{z \in \mathbb{H} : \|z - a\| = r\} = \{(a + r \cos t) + ir \sin t : 0 < t < \pi\},$$

the intersection of \mathbb{H} with the circle of radius r with center a . The endpoints, or points at infinity of $\mathcal{C}_r(a)$ are the points $a \pm r \in \mathbb{R} \subset \partial H$.

For ℓ one of the lines above, we write $\partial \ell$ for its endpoints and write

$$\bar{\ell} = \ell \cup \partial \ell.$$

Thus, for instance, $\bar{\ell}_0 = \{ti : t \geq 0\} \cup \{\infty\}$. Note that the endpoints are never in \mathbb{H} and never in ℓ . We have $\partial \ell = \bar{\ell} \cap \partial \mathbb{H}$ and $\ell = \bar{\ell} \cap \mathbb{H}$.

In particular ℓ_0 is the intersection of \mathbb{H} with the y -axis and $\mathcal{C}_1(0)$ is the intersection of \mathbb{H} with the unit circle.

We now wish to invoke the theory of geodesics. As we've seen earlier, in the context of spherical geometry, these give parametrizations of the "straight lines" appropriate for the geometry in question. We wish to use the following results from Chapter 11.

- (1) Distance minimizing curves are geodesic (Theorem 11.3.21).
- (2) There is a unique geodesic through a given point with a given velocity vector there (Theorem 11.3.17). This geodesic has a largest interval on which it is defined and geodesic, and is unique on that interval (Corollary 11.3.19).

We will prove everything else we need directly.

The following gives a geodesic parametrization of ℓ_0 .

Proposition 12.4.2. Define $\gamma_{i,i} : \mathbb{R} \rightarrow \mathbb{H}$ by

$$(12.4.3) \quad \gamma_{i,i}(t) = ie^t.$$

Then $\gamma_{i,i}$ parametrizes ℓ_0 with constant speed 1 and is distance minimizing between any two points of ℓ_0 . Thus, by Theorem 11.3.21 $\gamma_{i,i}$ is geodesic.

Proof. $\gamma'_{i,i}(t) = \gamma_{i,i}(t)$. By (12.0.7)

$$\|\gamma'_{i,i}(t)\|_{\gamma_{i,i}(t)} = \frac{e^t}{e^t} = 1.$$

For $a < b \in \mathbb{R}$, $\gamma_{i,i}|_{[\ln a, \ln b]}$ is a path from ai to bi . Its length is

$$(12.4.4) \quad \int_{\ln a}^{\ln b} \|\gamma'_{i,i}(t)\|_{\gamma_{i,i}(t)} dt = \ln b - \ln a.$$

It suffices to show this is minimal for the hyperbolic arc length of piecewise smooth paths from ai to bi . Thus, let $\gamma : [c, d] \rightarrow \mathbb{H}$ be such a path, and

write $\gamma(t) = \gamma_1(t) + i\gamma_2(t)$ with $\gamma_i(t)$ real for $i = 1, 2$. By (12.0.7), the arc length of γ is

$$\begin{aligned} \int_c^d \frac{\sqrt{(\gamma_1'(t))^2 + (\gamma_2'(t))^2}}{\gamma_2(t)} dt &\geq \int_c^d \frac{|\gamma_2'(t)|}{\gamma_2(t)} dt \geq \int_c^d \frac{\gamma_2'(t)}{\gamma_2(t)} dt \\ &= \ln \gamma_2(d) - \ln \gamma_2(c) = \ln b - \ln a. \quad \square \end{aligned}$$

Corollary 12.4.3. $\gamma_{i,i} : \mathbb{R} \rightarrow \mathbb{H}$ is the unique hyperbolic geodesic with $\gamma_{i,i}(0) = i$ and $\gamma'_{i,i}(0) = i$. (Here, we identify the tangent vectors at any point of \mathbb{H} with \mathbb{C} , rather than \mathbb{R}^2 .)

Notation 12.4.4. More generally, for $z \in \mathbb{H}$ and $w \in \mathbb{C}$, $\|w\|_z = 1$, we write $\gamma_{z,w}$ for the unique hyperbolic geodesic with $\gamma_{z,w}(0) = z$ and $\gamma'_{z,w}(0) = w$. We will see below that the domain of $\gamma_{z,w}$ is all of \mathbb{R} .

Since Möbius transformations preserve arc length and distance, we obtain the following.

Corollary 12.4.5. Let $f : \mathbb{H} \rightarrow \mathbb{H}$ be a Möbius transformation. Then $f \circ \gamma_{i,i}$ is geodesic, with image $f(\ell_0)$.

Proof. By Corollary 12.3.6, $f \circ \gamma_{i,i}$ is distance minimizing. \square

Remark 12.4.6. The situation here is similar to the parametrization of lines in \mathbb{R}^n by geodesics. In \mathbb{R}^n , a line has the form $\ell = x + \text{span}(u)$ for $x, u \in \mathbb{R}^n$ with $\|u\| = 1$. A parametrization of ℓ by a geodesic is given by

$$\begin{aligned} \varepsilon_{x,u} : \mathbb{R} &\rightarrow \mathbb{R}^n, \\ \varepsilon_{x,u}(t) &= x + tu. \end{aligned}$$

This is the unique geodesic ε with $\varepsilon(0) = x$ and $\varepsilon'(0) = u$. It maps onto ℓ , and the geodesic $\varepsilon_{x,-u}$ parametrizes ℓ with the opposite orientation. (Here, we are using ε to denote Euclidean geodesics.)

Consider now what happens when x is replaced by a different point, $y \in \ell$. Then $y = x + su$ for some $s \in \mathbb{R}$, and

$$\varepsilon_{y,u}(t) = x + su + tu = x + \tau_s(t)u = \varepsilon_{x,u} \circ \tau_s(t),$$

where $\tau_s : \mathbb{R} \rightarrow \mathbb{R}$ is translation by s . Thus, $\varepsilon_{y,u} = \varepsilon_{x,u} \circ \tau_s$.

Now consider the analogous situation for the parametrization

$$\gamma_{i,i} : \mathbb{R} \rightarrow \ell_0.$$

Let $\sigma_0 : \mathbb{R} \rightarrow \mathbb{R}$ be multiplication by -1 and consider the composite

$$\gamma_{i,i} \circ \sigma_0(t) = ie^{-t}.$$

By the chain rule,

$$(\gamma_{i,i} \circ \sigma_0)'(t) = \gamma'_{i,i}(-t),$$

so $\gamma_{i,i} \circ \sigma_0$ has unit speed. By Lemma 11.2.2, a nonsingular reparametrization does not change arc length, so $\gamma_{i,i} \circ \sigma_0$ is geodesic. Now $(\gamma_{i,i} \circ \sigma_0)'(0) = -i$, so

$$(12.4.5) \quad \gamma_{i,i} \circ \sigma_0 = \gamma_{i,-i}.$$

It parametrizes ℓ_0 with the opposite orientation to that given by $\gamma_{i,i}$. There are only two possible unit tangent vectors at i to curves parametrizing ℓ_0 , so these are the only two unit unit speed geodesics parametrizing ℓ_0 and taking 0 to i .

Now consider the reparametrization $\gamma_{i,i} \circ \tau_s$. By the chain rule, this is again a unit speed geodesic, this time taking 0 to $\gamma_{i,i}(s)$. Since $\gamma_{i,i} : \mathbb{R} \rightarrow \ell_0$ is onto, all the unit speed geodesics parametrizing ℓ_0 are either obtained in this way or by precomposing these with σ_0 . In fact, we have shown the following.

Lemma 12.4.7. *The geodesic parametrizations of ℓ_0 are precisely the composites $\gamma_{i,i} \circ \alpha$, with $\alpha \in \mathcal{I}_1$, the Euclidean isometries of \mathbb{R} . Thus, if f is a Möbius transformation, the geodesic parametrizations of $f(\ell_0)$ are the composites $f \circ \gamma_{i,i} \circ \alpha$ for $\alpha \in \mathcal{I}_1$.*

Proof. The Euclidean isometries are composites $\tau_s \circ \beta$, with β a linear isometry. The linear isometries are induced by orthogonal matrices, which in the 1×1 case are just $[\pm 1]$. Of course $[1]$ induces the identity and $[-1]$ induces σ_0 . \square

We will show that every hyperbolic line is the image of ℓ_0 under a Möbius transformation, and therefore is a geodesic line in \mathbb{H} via the parametrizations given in Lemma 12.4.7. Note first that as shown in the proof of Proposition 12.2.10, the map $f(z) = rz + a$ is Möbius for $r, a \in \mathbb{R}$ with $r > 0$. We have $f(\mathcal{C}_1(0)) = \mathcal{C}_r(a)$ and $f(\ell_0) = \ell_a$, so it suffices to show there is a Möbius transformation g with $g(\ell_a) = \mathcal{C}_r(b)$ for some a, r, b .

We shall also show that Möbius transformations take hyperbolic lines to hyperbolic lines. The key for understanding the effects of the Möbius transformations on the hyperbolic lines will be studying $g(z) = -z^{-1}$.

It is useful to give new formulae for hyperbolic lines. The proof of the following is immediate from $z + \bar{z} = 2 \operatorname{Re}(z)$ and $\|z - a\|^2 = (z - a)(\bar{z} - a)$.

Lemma 12.4.8. *The hyperbolic line ℓ_a is given by*

$$(12.4.6) \quad \ell_a = \{z \in \mathbb{H} : z + \bar{z} = 2a\}.$$

The hyperbolic line $\mathcal{C}_r(a)$ is given by

$$(12.4.7) \quad \mathcal{C}_r(a) = \{z \in \mathbb{H} : z\bar{z} - a(z + \bar{z}) + a^2 = r^2\}.$$

Proposition 12.4.9. *Let g be the Möbius transformation $g(z) = -z^{-1}$. Then $g(\ell_0) = \ell_0$ and for $a \neq 0$, $g(\ell_a) = \mathcal{C}_{\frac{1}{2|a|}}(-\frac{1}{2a})$, the hyperbolic line with endpoints $-\frac{1}{a} = g(a)$ and $0 = g(\infty)$.*

Since $g^2 = \operatorname{id}$ this also computes the effect of g on all hyperbolic lines having 0 as an endpoint. If neither 0 nor ∞ is an endpoint, then the line has the form $\mathcal{C}_r(a)$ with $a \neq \pm r$. In this case, $g(\mathcal{C}_r(a)) = \mathcal{C}_{\frac{r}{|r^2 - a^2|}}(\frac{a}{r^2 - a^2})$, the hyperbolic line whose endpoints are $\frac{1}{r-a}$ and $-\frac{1}{r+a}$, the images under g of the endpoints of $\mathcal{C}_r(a)$.

Proof. Write $g(z) = w$, so that $z = -\frac{1}{\bar{w}}$. Then the formula for $g(\ell_a)$ is obtained by solving

$$-\frac{1}{w} - \frac{1}{\bar{w}} = 2a.$$

This may be rewritten as

$$2aw\bar{w} + (w + \bar{w}) = 0.$$

If $a = 0$, this says $\operatorname{Re}(w) = 0$, hence $g(\ell_0) = \ell_0$. For $a \neq 0$ we can divide by $2a$, getting

$$w\bar{w} + \frac{1}{2a}(w + \bar{w}) + \frac{1}{4a^2} = \frac{1}{4a^2},$$

the formula for $\mathcal{C}_{\frac{1}{2|a|}}(-\frac{1}{2a})$.

For $a \neq \pm r$, we calculate $g(\mathcal{C}_r(a))$ by plugging $-\frac{1}{\bar{w}}$ into (12.4.7). A similar calculation achieves the stated result. \square

An immediate corollary is the calculation of which hyperbolic lines are preserved by g .

Corollary 12.4.10. *Let g be the Möbius transformation $g(z) = -z^{-1}$ and let ℓ be a hyperbolic line. Then $g(\ell) = \ell$ if and only if $i \in \ell$.*

Proof. The only line ℓ_a containing i is ℓ_0 , which is also the only line ℓ_a preserved by g . The line $\mathcal{C}_r(a)$ is preserved by g if and only if $r^2 - a^2 = 1$, i.e., $r^2 = a^2 + 1$. But $a^2 + 1 = (a - i)(a - \bar{i}) = \|a - i\|^2$, and the result follows. \square

A more significant corollary is the following.

Corollary 12.4.11. *Let ℓ be the hyperbolic line with endpoints $a, b \in \partial\mathbb{H}$ and let f be a Möbius transformation. Then $f(\ell)$ is the hyperbolic line with endpoints $f(a)$ and $f(b)$.*

Moreover, every hyperbolic line is the image of ℓ_0 under a Möbius transformation. Thus every hyperbolic line is the image of a geodesic. Up to precomposition with a translation, there are exactly two geodesics parametrizing each hyperbolic line, one with each possible orientation.

Proof. If f fixes ∞ , then Proposition 12.2.6 gives $f(z) = cz + d$ with $c, d \in \mathbb{R}$ with $c > 0$, and $f(\ell)$ is the line with endpoints $f(a)$ and $f(b)$ by a case by case inspection.

If f does not fix ∞ , the factorization (12.2.8) gives $f = f_3 \circ f_2 \circ f_1$ where f_1 and f_3 fix ∞ and $f_2(z) = -z^{-1}$. The identification of $f(\ell)$ follows by composition and Proposition 12.4.9.

The lines ℓ_a are all images of ℓ_0 under Möbius transformations fixing ∞ . Similarly, the lines $\mathcal{C}_r(a)$ are all images of $\mathcal{C}_1(0)$ under Möbius transformations, so they are also images of ℓ_0 by Proposition 12.4.9. \square

Recall that if $X \subset \bar{\mathbb{H}}$, that

$$(12.4.8) \quad \mathcal{S}_{\text{Möb}}(X) = \{f \in \text{Möb} : f(X) = X\},$$

the group of all Möbius transformations that preserve X .

Corollary 12.4.12. *Let ℓ be a hyperbolic line with endpoints a and b . Then*

$$(12.4.9) \quad \mathcal{S}_{\text{Möb}}(\ell) = \mathcal{S}_{\text{Möb}}(\{a, b\}),$$

i.e., f preserves ℓ if and only if f preserves the two-point set $\{a, b\}$.

An important subgroup of $\mathcal{S}_{\text{Möb}}(\ell)$ is the set of Möbius transformations fixing each endpoint:

Definition 12.4.13. For $z \neq w \in \bar{\mathbb{H}}$, we write

$$(12.4.10) \quad \text{Möb}_{z,w} = \text{Möb}_z \cap \text{Möb}_w,$$

the set of elements in Möb that fix both z and w .

The following is proven in Proposition 12.2.6.

Corollary 12.4.14. *The Möbius transformations fixing 0 and ∞ are the transformations $h_a(z) = az$:*

$$(12.4.11) \quad \text{Möb}_{0,\infty} = \{h_a : a > 0\}.$$

These are hyperbolic for $a \neq 1$. The multiplication is given by

$$(12.4.12) \quad h_a h_b = h_{ab}.$$

These transformations deserve further study.

Lemma 12.4.15. *The hyperbolic transformation h_a acts on lines as follows:*

$$(12.4.13) \quad h_a(\mathcal{C}_r(b)) = \mathcal{C}_{ar}(ab), \quad h_a(\ell_b) = \ell_{ab}.$$

Thus if $a \neq 1$, ℓ_0 is the only hyperbolic line preserved by h_a . On ℓ_0 we can think of h_a as a translation. We can see this in its effect on geodesics:

$$(12.4.14) \quad h_a \circ \gamma_{i,i} = \gamma_{i,i} \circ \tau_{\ln a}.$$

We shall refer to ℓ_0 as the axis of translation for h_a .

Proof. The effect of h_a on lines is an easy computation. For (12.4.14), we have

$$h_a \circ \gamma_{i,i}(t) = iae^t = ie^{t+\ln a}. \quad \square$$

This will allow us to understand the full group $\mathcal{S}_{\text{Möb}}(\ell_0)$.

Proposition 12.4.16. *$\text{Möb}_{0,\infty}$ is an index 2 subgroup of $\mathcal{S}_{\text{Möb}}(\ell_0)$. The elements of $\mathcal{S}_{\text{Möb}}(\ell_0) \setminus \text{Möb}_{0,\infty}$ form the coset*

$$(12.4.15) \quad (\text{Möb}_{0,\infty})g = \{h_a g : a \in \mathbb{R}\},$$

where $g(z) = -z^{-1}$. The multiplication of $\mathcal{S}_{\text{Möb}}(\ell_0)$ is determined by

$$(12.4.16) \quad gh_a g^{-1} = h_{\frac{1}{a}} = h_a^{-1}.$$

Thus,

$$(12.4.17) \quad h_a g h_b = h_{\frac{a}{b}} g,$$

and the rest of the multiplication follows. In particular, $(h_a g)^2 = h_{\frac{a}{a}} g^2 = \text{id}$, so each $h_a g$ has order 2.

The element $h_a g$ is elliptic with fixed-point $i\sqrt{a} \in \ell_0$. Thus, every element of $\mathcal{S}_{\text{Möb}}(\ell_0) \setminus \text{Möb}_{0,\infty}$ is elliptic with its fixed-point in ℓ_0 . Moreover, every element of ℓ_0 is the fixed-point of a unique element of $\mathcal{S}_{\text{Möb}}(\ell_0) \setminus \text{Möb}_{0,\infty}$.

On geodesics,

$$(12.4.18) \quad h_a g \circ \gamma_{i,i} = \gamma_{i,i} \circ \tau_{\ln a} \sigma_0,$$

where $\sigma_0(t) = -t$.

Proof. By Corollary 12.4.12, $f \in \mathcal{S}_{\text{Möb}}(\ell_0) \setminus \text{Möb}_{0,\infty}$ if and only if f interchanges 0 and ∞ . But then $f g^{-1}$ fixes both 0 and ∞ , and hence $f g^{-1} = h_a$ for some $a \in \mathbb{R}$. Hence $f = h_a g$. (12.4.16) and the calculation of the fixed-point are straightforward.

(12.4.18) follows from Lemma 12.4.15 once we show that $g \circ \gamma_{i,i} = \gamma_{i,i} \circ \sigma_0$. But

$$g \circ \gamma_{i,i}(t) = e^{-t} i,$$

and the result follows. □

Proposition 12.2.6 also determines which parabolic Möbius transformations fix ∞ .

Corollary 12.4.17. *The parabolic Möbius transformations fixing ∞ form the subgroup*

$$(12.4.19) \quad \mathcal{P}_\infty = \{p_b : b \in \mathbb{R}\} \subset \text{Möb}_\infty.$$

Their multiplication is given by

$$(12.4.20) \quad p_a p_b = p_{a+b}.$$

Indeed, we may deduce the group structure on Möb_∞ in the same way we derived the group structure on \mathcal{I}_n from the translations and the linear isometries. In both cases, the group in question is what's known as a semidirect product. The proof of the following is left to the reader.

Proposition 12.4.18. *Each Möbius transformation fixing ∞ may be written uniquely in the form $f = p_b h_a$ with $a, b \in \mathbb{R}$ and $a > 0$. The multiplication is given by*

$$p_{b_1} h_{a_1} p_{b_2} h_{a_2} = p_{b_1+a_1 b_2} h_{a_1 a_2}.$$

Behavior on lines is an important difference between hyperbolic and parabolic Möbius transformations. The proof of the following is obvious.

Lemma 12.4.19. *The parabolic transformations in \mathcal{P}_∞ act on lines as follows. For $b \in \mathbb{R}$, $p_b(\ell_a) = \ell_{a+b}$ and $p_b(C_r(a)) = C_r(a+b)$. Thus, if $b \neq 0$, no hyperbolic line is fixed by p_b .*

We may now determine the lines preserved by an arbitrary hyperbolic Möbius transformation.

Lemma 12.4.20. *Let $a, b \in \partial\mathbb{H}$. Then there is a Möbius transformation f with $f(\infty) = a$ and $f(0) = b$.*

Proof. If $a, b \in \mathbb{R}$, take $f = \varphi_A$ for $A = \begin{bmatrix} \frac{a}{a-b} & b \\ \frac{1}{a-b} & 1 \end{bmatrix}$. For $a = \infty$, take $f = p_b$. For $b = \infty$, take $f = p_a \circ g$ for $g(z) = -z^{-1}$. \square

Recall that a nonidentity Möbius transformation is hyperbolic if and only if it has exactly two fixed-points in $\bar{\mathbb{H}}$, both of which must then lie on $\partial\mathbb{H}$. Any other nonidentity Möbius transformation has one fixed-point in $\bar{\mathbb{H}}$. If the fixed-point lies on $\partial\mathbb{H}$, the transformation is parabolic. If it lies in \mathbb{H} , the transformation is elliptic.

Proposition 12.4.21. *The hyperbolic Möbius transformations fixing $a, b \in \partial\mathbb{H}$ are the nonidentity elements in the subgroup $\text{Möb}_{a,b}$. If f is a Möbius transformation taking ∞ to a and 0 to b (e.g., from Lemma 12.4.20), then*

$$(12.4.21) \quad \text{Möb}_{a,b} = f \text{Möb}_{0,\infty} f^{-1}.$$

Let ℓ be the hyperbolic line with endpoints a and b . Then ℓ is the unique line preserved by a given nonidentity element $g \in \text{Möb}_{a,b}$. We refer to ℓ as the axis of translation for g , and we have

$$(12.4.22) \quad \mathcal{S}_{\text{Möb}}(\ell) = f \mathcal{S}_{\text{Möb}}(\ell_0) f^{-1}.$$

Thus, $\text{Möb}_{a,b}$ has index 2 in $\mathcal{S}_{\text{Möb}}(\ell)$, and every element of $\mathcal{S}_{\text{Möb}}(\ell) \setminus \text{Möb}_{a,b}$ is an elliptic transformation of order 2 fixing an element of ℓ . Each element of ℓ arises as such a fixed-point.

Nonidentity parabolic transformations, on the other hand, preserve no lines.

Proof. Since conjugation induces a bijection from Möb to itself, it preserves intersections, so

$$\begin{aligned} f \text{Möb}_{0,\infty} f^{-1} &= f(\text{Möb}_0 \cap \text{Möb}_\infty) f^{-1} = (f \text{Möb}_0 f^{-1}) \cap (f \text{Möb}_\infty f^{-1}) \\ &= \text{Möb}_b \cap \text{Möb}_a = \text{Möb}_{a,b}. \end{aligned}$$

By (3.6.8), the Möbius transformations preserving $f(m)$ are the conjugates by f of the Möbius transformations preserving m . Since no line other than ℓ_0 is preserved by the nonidentity elements of $\text{Möb}_{0,\infty}$, no line other than $f(\ell_0) = \ell$ can be preserved by the nonidentity elements of $\text{Möb}_{a,b}$. Since $f(\ell_0) = \ell$, (12.4.22) follows from (3.6.8).

Regarding parabolic transformations, every parabolic transformation fixes some $a \in \partial\mathbb{H}$, and hence is conjugate by this same transformation f to one fixing ∞ . The same argument applies. \square

Remark 12.4.22. As an open subset of \mathbb{R}^2 , \mathbb{H} has a natural orientation in the sense of Definition 10.5.1. We simply choose the standard identification of the tangent space at each $z \in \mathbb{H}$ with \mathbb{R}^2 . That gives a linear orientation of the tangent space.

Recall, then, that a diffeomorphism $f : \mathbb{H} \rightarrow \mathbb{H}$ is orientation-preserving if $\det Df(z) > 0$ for all $z \in \mathbb{H}$ and is orientation-reversing if $\det Df(z) < 0$ for all $z \in \mathbb{H}$.

The following shows that Möbius transformations are orientation-preserving.

Lemma 12.4.23. *Let $U \subset \mathbb{C}$ be open and let $f : U \rightarrow \mathbb{C}$ be holomorphic. Suppose the complex derivative $f'(z) \neq 0$. Then the real Jacobian matrix Df has positive determinant.*

Thus, Möbius transformations are orientation-preserving.

Proof. $Df(z)$ is the realification of $[f'(z)]$. If $f'(z) = a + ib$ with $a, b \in \mathbb{R}$, then

$$Df(z) = \begin{bmatrix} a & -b \\ b & a \end{bmatrix},$$

so $\det Df = a^2 + b^2 = \|f'(z)\|^2$. □

Lemma 12.2.3 allows us to make explicit our connection between elliptic Möbius transformations and rotations. Recall from Lemma 12.2.13 that the Möbius transformations preserving i are precisely the transformations φ_{R_θ} , where $R_\theta = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$, is the standard rotation matrix in SO_2 . Explicitly,

$$\varphi_{R_\theta}(z) = \frac{\cos \theta z - \sin \theta}{\sin \theta z + \cos \theta}.$$

Substituting $z = i$ into (12.2.3) gives the following.

Corollary 12.4.24. $\varphi'_{R_\theta}(i) = e^{-2i\theta}$. *Thus, the Jacobian matrix of φ_{R_θ} at i is*

$$(12.4.23) \quad D\varphi_{R_\theta}(i) = R_{-2\theta} = \begin{bmatrix} \cos(-2\theta) & -\sin(-2\theta) \\ \sin(-2\theta) & \cos(-2\theta) \end{bmatrix}.$$

Proof. The displayed matrix is just the realification of the complex 1×1 matrix $[e^{-2i\theta}]$. □

Recall our geodesic $\gamma_{i,i} : \mathbb{R} \rightarrow \ell_0$ given by $\gamma_{i,i}(t) = ie^t$. We saw that $\|\gamma'_{i,i}(t)\|_{\gamma_{i,i}(t)} = 1$ for all t . Note that $\gamma'_{i,i}(0) = i$. We shall now explicitly compute the velocity vectors for the composites $\varphi_{R_\theta} \circ \gamma_{i,i}$ and show this exhausts the unit speed geodesics emanating from i .

Let us first replace $\gamma_{i,i}$ with the unit speed geodesic along $\mathcal{C}_1(0)$. We first use (12.2.8) to map ℓ_0 onto $\mathcal{C}_1(0)$.

Lemma 12.4.25. *The elliptic transformation $\varphi_{R_{\frac{\pi}{4}}}$ carries ℓ_0 onto $\mathcal{C}_1(0)$. Consider the geodesic $\gamma = \varphi_{R_{\frac{\pi}{4}}} \circ \gamma_{i,i}$. Then $\gamma(0) = i$ and $\gamma'(0)$ is the complex number 1. Thus, $\gamma = \gamma_{i,1}$.*

Proof. The factorization (12.2.8) gives $\varphi_{R_{\frac{\pi}{4}}} = f_3 \circ f_2 \circ f_1$ with $f_1(z) = z + 1$, $f_2(z) = -z^{-1}$ and $f_3(z) = 2z + 1$. f_1 takes ℓ_0 to ℓ_1 , which is then carried to $\mathcal{C}_{-\frac{1}{2}}(-\frac{1}{2})$ by f_2 (Proposition 12.4.9), and then on to $\mathcal{C}_1(0)$ by f_3 . The calculation of the velocity vector follows from (12.4.23) and the chain rule. \square

Note that when viewed via the factorization (12.2.8), the sign in (12.4.23) comes from the orientation of $\mathcal{C}_{-\frac{1}{2}}(-\frac{1}{2})$ induced by f_2 : f_2 converts the upward pointing orientation of the vertical line ℓ_1 to the left-to-right orientation of the semicircle. The following is now immediate from (12.4.23).

Proposition 12.4.26.

- (1) Let $\theta \in \mathbb{R}$. Then $\varphi_{R_{-\frac{\theta}{2}}} \circ \gamma_{i,1} = \gamma_{i,e^{i\theta}}$, the unique geodesic taking 0 to i with velocity vector $e^{i\theta}$ there. It parametrizes a hyperbolic line by Corollary 12.4.11.

Since this accounts every possible unit vector at i as θ varies, we see that every geodesic in \mathbb{H} through i parametrizes a hyperbolic line.

- (2) Similarly, $\varphi_{R_{-\frac{\theta}{2}}} \circ \gamma_{i,e^{i\psi}} = \gamma_{i,e^{i(\theta+\psi)}}$, so $\varphi_{R_{-\frac{\theta}{2}}}$ rotates the line parametrized by $\gamma_{i,e^{i\psi}}$ by the angle θ about i . The elliptic Möbius transformations fixing i are indeed rotations about i .

We have justified the following.

Definition 12.4.27. The rotation $\rho_{(i,\theta)}$ of \mathbb{H} about i by θ is given by

$$(12.4.24) \quad \rho_{(i,\theta)} = \varphi_{R_{-\frac{\theta}{2}}}.$$

The following summarized what we know about these rotations.

Proposition 12.4.28. The isotropy group Möb_i is given by

$$(12.4.25) \quad \text{Möb}_i = \{\rho_{(i,\theta)} : \theta \in \mathbb{R}\}.$$

This is an Abelian group as

$$(12.4.26) \quad \rho_{(i,\theta)}\rho_{(i,\psi)} = \rho_{(i,\theta+\psi)}.$$

Moreover $\rho_{(i,\theta)}$ has finite order if and only if θ is a rational multiple of 2π . If $\theta = \frac{2\pi k}{n}$ with $(k,n) = 1$, then $\rho_{(i,\theta)}$ has order n . In particular, $\rho_{(i,\theta)} = \text{id}$ if and only if θ is a multiple of 2π .

We now generalize this to elliptic transformations that fix points other than i , obtaining that all hyperbolic geodesics parametrize hyperbolic lines. The following is an easy calculation.

Lemma 12.4.29. Let $w \in \mathbb{H}$ and write $w = a + ib$ with $a, b \in \mathbb{R}$. Then $f(z) = bz + a$ is a Möbius transformation with $f(i) = w$ and $f(\mathcal{C}_1(0)) = \mathcal{C}_{\|w\|}(a)$. Moreover $f \circ \gamma_{i,1} = \gamma_{w,b}$, the unique unit speed hyperbolic geodesic whose value at 0 is w and whose velocity vector at 0 is a positive multiple of 1. It parametrizes $\mathcal{C}_{\|w\|}(a)$.

Proposition 12.4.30. *Let $w \in \mathbb{H}$ and let $f \in \text{Möb}$ with $f(i) = w$. By Lemma 3.6.4, the isotropy group Möb_w is given by*

$$(12.4.27) \quad \text{Möb}_w = \{f\varphi_{R_\theta}f^{-1} : \theta \in \mathbb{R}\}.$$

If f is the Möbius transformation of Lemma 12.4.29, then

$$(12.4.28) \quad f\varphi_{R_{-\frac{\theta}{2}}}f^{-1} \circ \gamma_{w,b} = \gamma_{w,be^{i\theta}}.$$

Thus, every unit speed geodesic through w is the image of $\gamma_{w,b}$ under a Möbius transformation, and hence parametrizes a hyperbolic line.

Since w is arbitrary, every hyperbolic geodesic is the composite of a Möbius transformation with $\gamma_{i,1}$ and parametrizes a hyperbolic line.

More generally,

$$(12.4.29) \quad f\varphi_{R_{-\frac{\theta}{2}}}f^{-1} \circ \gamma_{w,be^{i\psi}} = \gamma_{w,be^{i(\theta+\psi)}},$$

So $f\varphi_{R_{-\frac{\theta}{2}}}f^{-1}$ rotates the hyperbolic lines through w by the angle θ about w .

Proof. Since $f^{-1} \circ \gamma_{w,b} = \gamma_{i,1}$, (12.4.28) follows from Proposition 12.4.26 and the chain rule. \square

The following is a useful observation.

Lemma 12.4.31. *Let $f, g \in \text{Möb}$ with $f(i) = g(i) = w$. Then*

$$(12.4.30) \quad f\rho_{(i,\theta)}f^{-1} = g\rho_{(i,\theta)}g^{-1}$$

for all θ .

Proof.

$$f\rho_{(i,\theta)}f^{-1} = g\rho_{(i,\theta)}g^{-1} \Leftrightarrow (g^{-1}f)\rho_{(i,\theta)}(g^{-1}f)^{-1} = \rho_{(i,\theta)}.$$

But $g^{-1}f$ fixes i , so $g^{-1}f = \rho_{(i,\psi)}$ for some ψ and hence commutes with $\rho_{(i,\theta)}$. The result follows. \square

In particular, the following is now justified.

Definition 12.4.32. Let f be any Möbius transformation with $f(i) = w$. Then

$$(12.4.31) \quad \rho_{(w,\theta)} = f\rho_{(i,\theta)}f^{-1}.$$

In summary, we obtain:

Proposition 12.4.33. *Let $w \in \mathbb{H}$. The isotropy group Möb_w is given by*

$$(12.4.32) \quad \text{Möb}_w = \{\rho_{(w,\theta)} : \theta \in \mathbb{R}\}.$$

This is an Abelian group as

$$(12.4.33) \quad \rho_{(w,\theta)}\rho_{(w,\psi)} = \rho_{(w,\theta+\psi)}.$$

Moreover $\rho_{(w,\theta)}$ has finite order if and only if θ is a rational multiple of 2π . If $\theta = \frac{2\pi k}{n}$ with $(k, n) = 1$, then $\rho_{(w,\theta)}$ has order n . In particular, $\rho_{(w,\theta)} = \text{id}$ if and only if θ is a multiple of 2π .

Remark 12.4.34. Note that the hyperbolic transformations h_a ($a \neq 1$) all have infinite order. Since every hyperbolic transformation is conjugate to some h_a , all hyperbolic Möbius transformations have infinite order.

Similarly, the parabolic transformations p_b , $b \neq 0$, all have infinite order. Since every parabolic transformation is conjugate to some p_b , all parabolic Möbius transformations have infinite order.

Thus, the only nonidentity Möbius transformations of finite order are elliptic, and we've identified their orders in Proposition 12.4.33. In particular, we have the following.

Lemma 12.4.35. *The Möbius transformations of order 2 are precisely the rotations $\rho_{(w,\pi)}$ for $w \in \mathbb{H}$.*

This leads to a reinterpretation of part of Proposition 12.4.21.

Corollary 12.4.36. *Let ℓ be a hyperbolic line with $\partial\ell = \{a, b\}$. Then the nonidentity, nonhyperbolic elements of $\mathcal{S}_{\text{Möb}}(\ell)$ are given by*

$$(12.4.34) \quad \mathcal{S}_{\text{Möb}}(\ell) \setminus \text{Möb}_{a,b} = \{\rho_{(w,\pi)} : w \in \ell\}.$$

A consequence of this is the following.

Corollary 12.4.37. *An elliptic Möbius transformation of order unequal to 2 preserves no hyperbolic line. An elliptic Möbius transformation of order 2 preserves every line containing its center of rotation (and no others).*

12.5. Incidence relations and transitivity properties. The following is a direct analogue of the Euclidean case.

Proposition 12.5.1. *Two points determine a line in \mathbb{H} . Given $z \neq w \in \mathbb{H}$, there is a unique hyperbolic line containing z and w .*

Proof. If $\text{Re}(z) = \text{Re}(w) = a$, then ℓ_a is the unique vertical line containing z and w . Since the semicircles $\mathcal{C}_r(a)$ are all graphs of functions of x , none of them contain both z and w .

Now suppose $\text{Re}(z) \neq \text{Re}(w)$. Then no vertical line contains both z and w , and both points lie in $\mathcal{C}_r(a)$ if and only if

$$(12.5.1) \quad \begin{aligned} (z - a)(\bar{z} - a) &= (w - a)(\bar{w} - a) \\ z\bar{z} - a(z + \bar{z}) + a^2 &= w\bar{w} - a(w + \bar{w}) + a^2 \\ z\bar{z} - a(2\text{Re}(z)) &= w\bar{w} - a(2\text{Re}(w)). \end{aligned}$$

Since $\text{Re}(z) \neq \text{Re}(w)$, this allows us to solve for a and hence r . □

We now analyze how the Möbius transformations act on pairs of points.

Proposition 12.5.2. *Let $\zeta \neq \omega \in \mathbb{H}$. Then there is a unique Möbius transformation f such that:*

- (1) $f(\zeta)$ and $f(\omega)$ lie in ℓ_0 .
- (2) $f(\zeta) = i$.
- (3) $\text{Im}(f(\omega)) > 1$.

Proof. By Proposition 12.5.1, there is a hyperbolic line ℓ containing ζ and ω . By Corollary 12.4.11, there is a Möbius transformation f_1 taking ℓ onto ℓ_0 , obtaining (1).

We can now apply a hyperbolic Möbius transformation h_a ($h_a(z) = az$) such that $h_a \circ f_1(\zeta) = i$, obtaining (2).

Now note that if $g(z) = -z^{-1}$, then $g(ti) = \frac{1}{t}i$ for $t \in \mathbb{R}$. So g interchanges the subsets of ℓ_0 with pure imaginary parts in $(0, 1)$ and $(1, \infty)$. Thus, composing with g if necessary, we obtain (3).

Uniqueness is proven in Corollary 12.2.9. \square

In particular, note that the value of $f(\omega)$ is forced by the hyperbolic distance from ζ to ω . Specifically, (12.4.4) gives the following.

Addendum 12.5.3. *The value of $f(w)$ in Proposition 12.5.2 is ie^d , where $d = d_{\mathbb{H}}(z, w)$.*

We obtain the following.

Corollary 12.5.4. *Let $z \neq w$ and $\zeta \neq \omega$ be two pairs of points in \mathbb{H} with*

$$d_{\mathbb{H}}(z, w) = d_{\mathbb{H}}(\zeta, \omega),$$

where d is hyperbolic distance. Then there is a unique Möbius transformation f with $f(z) = \zeta$ and $f(w) = \omega$.

Proof. Let $d = d_{\mathbb{H}}(z, w)$. Then there are Möbius transformations g, h , with $g(z) = h(\zeta) = i$ and $g(w) = h(\omega) = ie^d$. Let $f = h^{-1}g$. Uniqueness is proven in Corollary 12.2.9. \square

Remark 12.5.5. There are analogous results in the plane. The correct analogue is between the Möbius transformations and the orientation-preserving isometries $\mathcal{O}_2 = \mathcal{O}(\mathbb{R}^2)$ (we have not yet considered the orientation-reversing isometries of \mathbb{H}).

The elements of \mathcal{O}_2 are all translations and rotations. Nonidentity translations have no fixed-points and nonidentity rotations have exactly one fixed-point. So any two orientation-preserving isometries of \mathbb{R}^2 that agree on two points must be equal.

It is then easy to see that given $x \neq y \in \mathbb{R}^2$, there is a unique orientation-preserving isometry $f \in \mathcal{O}_2$ with $f(x) = 0$ and $f(y)$ on the positive x -axis. As a consequence if $x \neq y \in \mathbb{R}^2$ and if $z, w \in \mathbb{R}^2$ with

$$d(x, y) = d(z, w),$$

there is a unique $f \in \mathcal{O}_2$ with $f(x) = z$ and $f(y) = w$.

In hyperbolic space we have an extra layer of information coming from the boundary points.

Lemma 12.5.6. *Let $w \in \mathbb{H}$ and let $a \in \partial\mathbb{H}$, then there is a unique hyperbolic line ℓ containing w and having a as one of its boundary points.*

Proof. This is very much like the proof of Proposition 12.5.1. If $a = \infty$ or $a = \operatorname{Re}(w)$ then $\ell = \ell_{\operatorname{Re}(w)}$. Otherwise, $\operatorname{Re}(w)$ and a are distinct real numbers and $\ell = \mathcal{C}_r(b)$ for some r and b . We must have

$$(a - b)^2 = (w - b)(\bar{w} - b) = w\bar{w} - 2b\operatorname{Re}(w) + b^2.$$

Solving this, we get $b = \frac{a^2 - w\bar{w}}{2(a - \operatorname{Re}(w))}$ and $r = |a - b|$. \square

We obtain the following.

Proposition 12.5.7. *Let ℓ be a hyperbolic line and let $w \in \ell$. Let $\partial\ell = \{a, b\}$. Then there are exactly two Möbius transformations taking ℓ to ℓ_0 and w to i . One of them, say f , takes a to ∞ . The other is gf , where $g(z) = -z^{-1}$. gf takes b to ∞ and a to 0 .*

Proof. The proof is much like that of Proposition 12.5.2. First find a Möbius transformation f_1 taking ℓ onto ℓ_0 . Then compose with a hyperbolic Möbius transformation h_a so that $h_a f_1(w) = i$. Then compose with g , if necessary so that a goes to ∞ .

Uniqueness comes from Corollary 12.2.9. \square

The following is an immediate consequence.

Corollary 12.5.8. *Let ℓ and m be hyperbolic lines with $z \in \ell$ and $w \in m$. Then there are exactly two Möbius transformations taking ℓ to m and z to w . They are determined by their behavior on $\partial\ell$. If f is one of them, then $\rho_{(w, \pi)} f$ is the other.*

12.6. Hyperbolic line segments.

Definition 12.6.1. Let $\zeta \neq \omega \in \mathbb{H}$. A geodesic path from ζ to ω is a path $\gamma|_{[a, b]} : [a, b] \rightarrow \mathbb{H}$ with γ a hyperbolic geodesic of unit speed with $\gamma(a) = \zeta$ and $\gamma(b) = \omega$. Note that if $d = d_{\mathbb{H}}(\zeta, \omega)$ and f is the Möbius transformation given by Proposition 12.5.2, then $f^{-1} \circ \gamma_{i, i}|_{[0, d]}$ is one such path. By Corollary 12.4.11, any other such path is obtained from that one by precomposition with a translation $\tau_s : [-s, d - s] \rightarrow [0, d]$ for $s \in \mathbb{R}$.

We write $[\zeta, \omega]$ for the image of such a path and call it the geodesic segment between ζ and ω . Note that precomposition of $f^{-1} \circ \gamma_{i, i}|_{[0, d]}$ by $t \mapsto d - t$ gives a geodesic path from ω to ζ , hence $[\zeta, \omega] = [\omega, \zeta]$. We refer to d as the length of this segment.

Of course, we set $[\zeta, \zeta] = \{\zeta\}$.

Geodesic line segments in \mathbb{H} play a role similar to that of ordinary line segments in \mathbb{R}^n . Indeed, they form the edges of hyperbolic triangles. They also satisfy the following important property.

Proposition 12.6.2. *Let $\zeta \neq \omega \in \mathbb{H}$. Then*

$$(12.6.1) \quad [\zeta, \omega] = \{z \in \mathbb{H} : d_{\mathbb{H}}(\zeta, z) + d_{\mathbb{H}}(z, \omega) = d_{\mathbb{H}}(\zeta, \omega)\}.$$

Proof. Let $z \in [\zeta, \omega]$ and let $\gamma : [a, b] \rightarrow \mathbb{H}$ be a geodesic path from ζ to ω with $\gamma(c) = z$. Then $\gamma|_{[a,c]}$ and $\gamma|_{[c,b]}$ are geodesic paths from ζ to z and from z to ω , respectively. We have

$$d_{\mathbb{H}}(\zeta, \omega) = \ell(\gamma) = \ell(\gamma|_{[a,c]}) + \ell(\gamma|_{[c,b]}) = d_{\mathbb{H}}(\zeta, z) + d_{\mathbb{H}}(z, \omega).$$

Conversely, suppose $d_{\mathbb{H}}(\zeta, z) + d_{\mathbb{H}}(z, \omega) = d_{\mathbb{H}}(\zeta, \omega)$. Let $\delta : [a, c] \rightarrow \mathbb{H}$ be a unit speed geodesic path from ζ to z and $\varepsilon : [c, b] \rightarrow \mathbb{H}$ be a unit speed geodesic path from z to ω (as can be arranged via a translation to get the endpoints right). Define $\gamma : [a, b] \rightarrow \mathbb{H}$ by

$$\gamma(t) = \begin{cases} \delta(t) & \text{for } t \in [a, c], \\ \varepsilon(t) & \text{for } t \in [c, b]. \end{cases}$$

Then γ is a unit speed distance minimizing path, and hence is geodesic. \square

Corollary 12.6.3. *Let $f \in \mathcal{I}(\mathbb{H})$ (i.e., $f : \mathbb{H} \rightarrow \mathbb{H}$ is distance-preserving). Then*

$$(12.6.2) \quad f([z, w]) = [f(z), f(w)] \quad \text{for all } z, w \in \mathbb{H}.$$

Indeed, if $\gamma : [a, b] \rightarrow [z, w]$ is a unit speed geodesic parametrization of $[z, w]$, then $f(\gamma(t))$ is the unique point on $[f(z), f(w)]$ of distance $d_{\mathbb{H}}(z, \gamma(t))$ from $f(z)$. Thus, $f \circ \gamma$ is the unique unit speed geodesic parametrization of $[f(z), f(w)]$ taking a to $f(z)$.

Proof. Let γ be as in the statement. Then

$$\begin{aligned} d_{\mathbb{H}}(f(z), f(w)) &= d_{\mathbb{H}}(z, w) = d_{\mathbb{H}}(z, \gamma(t)) + d_{\mathbb{H}}(\gamma(t), w) \\ &= d_{\mathbb{H}}(f(z), f(\gamma(t))) + d_{\mathbb{H}}(f(\gamma(t)), f(w)). \end{aligned}$$

So $f(\gamma(t)) \in [f(z), f(w)]$ by Proposition 12.6.2. \square

Corollary 12.6.4. *Let $f \in \mathcal{I}(\mathbb{H})$. Then $f(\ell)$ is a hyperbolic line for each hyperbolic line ℓ . Moreover, if $\gamma : \mathbb{R} \rightarrow \mathbb{H}$ is a unit speed geodesic parametrization of ℓ , then $f \circ \gamma$ is a unit speed geodesic parametrization of $f(\ell)$.*

Proof. Let z, ζ, w be any three points on ℓ . By Corollary 12.6.3 $f(z), f(\zeta)$, and $f(w)$ all lie on the same hyperbolic line. Since two distinct points determine a hyperbolic line and since z, ζ, w are arbitrary, $f(\ell)$ must be contained in a single hyperbolic line m . Moreover, since f is distance-preserving, $f(\ell)$ must contain the two points of distance d from $f(z)$ for each $d > 0$, so $f : \ell \rightarrow m$ is onto.

The result now follows since geodesics are distance-preserving, so $f \circ \gamma$ must coincide with a geodesic parametrization of m . \square

Corollary 12.6.5. *If $f \in \mathcal{I}(\mathbb{H})$ is the identity on two points, say z and w , then f is the identity on the line ℓ containing z and w .*

Proof. By Corollary 12.6.4, $f(\ell)$ is a line. But then $f(\ell)$ must equal ℓ as it contains z and w . Moreover, if γ is the unit speed geodesic parametrization of ℓ taking 0 to z , and with w on its “positive side”, then $f \circ \gamma = \gamma$. Since ℓ the image of γ , f is the identity there. \square

12.7. Parallels and perpendiculars. We use the naive notion of parallel lines:

Definition 12.7.1. Two hyperbolic lines are parallel if they do not intersect. We say they are hyperparallel if they also do not share a common boundary point.

Note that the lines ℓ_a all share ∞ as a boundary point, so they are parallel, but not hyperparallel. Similarly, the lines $\mathcal{C}_{|a|}(a)$ all share the boundary point 0. They are pairwise parallel, but not hyperparallel. (We may add ℓ_0 to this last family and retain these properties.)

Indeed, for any $a \in \partial\mathbb{H}$, the distinct lines having a as one of their boundary points are pairwise parallel but not hyperparallel. We can see this by applying a parabolic transformation to the preceding example.

On the other hand, the lines $\mathcal{C}_r(0)$ are pairwise hyperparallel as r varies.

Hyperbolic space satisfies the antithesis of the Parallel Postulate.

Proposition 12.7.2. *Let ℓ be a hyperbolic line and let $w \in \mathbb{H} \setminus \ell$. Then there are infinitely many hyperbolic lines through w parallel to ℓ .*

Proof. We first assume $\ell = \ell_0 = \{z \in \mathbb{H} : \operatorname{Re}(z) = 0\}$. Let $w \in \mathbb{H} \setminus \ell_0$. Write $w = x + iy$ with $x, y \in \mathbb{R}$. For simplicity, assume $x > 0$. The argument for negative x is similar.

Of course $\ell_x = \{z \in \mathbb{H} : \operatorname{Re}(z) = x\}$ is parallel to ℓ_0 and contains w . We now ask which of the lines $\mathcal{C}_r(a)$ contain w . Of course, for $a \in \mathbb{R}$, there is a unique such line: the one with $r = \|w - a\|$. So we now ask whether $\mathcal{C}_{\|w-a\|}(a)$ intersects ℓ_0 .

An intermediate value theorem argument shows that $\mathcal{C}_{\|w-a\|}(a) \cap \ell_0 = \emptyset$ if and only if $a \geq \|w - a\|$. An easy calculation shows this is equivalent to

$$a \geq \frac{x^2 + y^2}{2x},$$

leaving infinitely many possibilities for a . Note that of the infinitely many lines through z parallel to ℓ_0 only two of them fail to be hyperparallel to ℓ_0 : ℓ_x and $\mathcal{C}_{\|w-a\|}(a)$ for $a = \frac{x^2+y^2}{2x}$.

For a general line ℓ , let f be a Möbius transformation with $f(\ell_0) = \ell$. Then ℓ and m are parallel if and only if ℓ_0 and $f^{-1}(m)$ are parallel. So apply the preceding argument with w replaced by $f^{-1}(w)$ and take the images of the resulting lines under f . \square

Just as in \mathbb{R}^n and \mathbb{S}^n , the geodesics provide parametrizations, and hence orientations of hyperbolic lines. We use them to calculate directed angles between oriented lines in \mathbb{H} . Note that since any pair of distinct points in \mathbb{H} is contained in a unique hyperbolic line, two nonparallel hyperbolic lines intersect in exactly one point (precisely as was the case in \mathbb{R}^2 , but different from the behavior in \mathbb{S}^2).

Definition 12.7.3. Let ℓ and m be intersecting lines in \mathbb{H} , with $\ell \cap m = z = x + iy$ with $x, y \in \mathbb{R}$. Let $\gamma : \mathbb{R} \rightarrow \ell$ and $\delta : \mathbb{R} \rightarrow m$ be unit speed geodesics parametrizing ℓ and m , respectively, with $\gamma(t) = \delta(s) = z$. Then $\|\gamma'(t)\|_z = \|\delta'(s)\|_z = 1$, so $\gamma'(t) = ye^{i\theta}$ and $\delta'(s) = ye^{i\varphi}$ for $\theta, \varphi \in \mathbb{R}$. We define the directed angle from ℓ to m to be $\varphi - \theta$.

Note this is precisely the Euclidean directed angle between the oriented tangent lines to ℓ and m , where the orientations of the tangent lines are given by the velocity vectors to the geodesics chosen. Just as in the Euclidean case, a direct calculation of dot products shows the following.

Lemma 12.7.4. *Keeping the notations above, the unsigned angle between ℓ and m with respect to these orientations is $\cos^{-1}(\langle \gamma'(t), \delta'(s) \rangle_z)$.*

Definition 12.7.5. The hyperbolic lines ℓ and m are perpendicular (written $\ell \perp m$) if the unsigned angle between them is $\frac{\pi}{2}$. Note this is independent of orientations as $\frac{\pi}{2} = \pi - \frac{\pi}{2}$.

In particular, perpendicularity depends only on the tangent lines and not on their orientation:

Corollary 12.7.6. *Two hyperbolic lines are perpendicular if and only if their tangent lines are perpendicular at their point of intersection.*

A key here is the observation we made in spherical geometry that if $\gamma : (a, b) \rightarrow \mathbb{S}^n$ is smooth, then $\gamma'(t)$ is orthogonal to $\gamma(t)$ for all t . Applying this to circles in \mathbb{R}^2 and allowing the center and radius to vary, we obtain the following.

Lemma 12.7.7. *Let $z \in \mathcal{C}_r(a)$. Then the tangent line to $\mathcal{C}_r(a)$ at z is the Euclidean line through z perpendicular to the Euclidean segment \overline{za} (i.e., the radial segment in $\mathcal{C}_r(a)$ ending at z).*

Of course, the tangent line to ℓ_a at any point is ℓ_a .

Corollary 12.7.8. *ℓ_a is perpendicular to $\mathcal{C}_r(b)$ if and only if $a = b$.*

Proof. ℓ_a is perpendicular to $\mathcal{C}_r(b)$ if and only if the tangent line to $\mathcal{C}_r(b)$ at $z = \ell_a \cap \mathcal{C}_r(b)$ has slope 0. The standard Euclidean parametrization of $\mathcal{C}_r(z)$ shows this to be the case if and only if $\operatorname{Re}(z) = a$. \square

Of course if $a \neq b$, then $\ell_a \cap \ell_b = \emptyset$, and the two lines are not perpendicular.

Corollary 12.7.9. *There is a unique hyperbolic line perpendicular to ℓ_a through any point $z \in \mathbb{H}$. It is the line $\mathcal{C}_{\|z-a\|}(a)$.*

Proof. This is the only line of the form $\mathcal{C}_r(a)$ containing z . \square

Lemma 12.7.7 has another important and immediate consequence.

Corollary 12.7.10. *Let $\mathcal{C}_r(a)$ and $\mathcal{C}_s(b)$ be intersecting hyperbolic lines with $z = \mathcal{C}_r(a) \cap \mathcal{C}_s(b)$. Then $\mathcal{C}_r(a)$ and $\mathcal{C}_s(b)$ are perpendicular if and only if the line segments \overline{az} and \overline{bz} are perpendicular.*

A very important result is the following. For unsigned angles, this would follow from Lemma 11.2.12. We prove the signed case here directly.

Proposition 12.7.11. *Möbius transformations preserve directed angles: if ℓ and m be nonparallel hyperbolic lines f is a Möbius transformation then the directed angle from ℓ to m is equal to the directed angle from $f(\ell)$ to $f(m)$.*

Proof. Let γ and δ be geodesics parametrizing ℓ and m respectively, with $\gamma(t) = \delta(s) = z$. Then $f \circ \gamma$ and $f \circ \delta$ are geodesics parametrizing $f(\ell)$ and $f(m)$, respectively.

The result now follows from the chain rule and the simple fact that f is holomorphic (complex differentiable). The Jacobian matrix $Df(z)$ is the realification of the complex matrix $f'(z)$. If $f'(z) = re^{i\psi}$, then $Df = rR_\psi$, the scalar r times the rotation matrix R_ψ . Multiplication by R_ψ preserves directed angles, as does scalar multiplication. \square

And now we get a nice bonus from our clear picture of the perpendiculars to ℓ_0 . In words: we can “drop a perpendicular” from any point in \mathbb{H} to any hyperbolic line. Perpendiculars are unique in \mathbb{H} in a way parallels are not.

Corollary 12.7.12. *Let ℓ be a hyperbolic line and let $z \in \mathbb{H}$. Then there is a unique hyperbolic line through z perpendicular to ℓ .*

Proof. Let f be a Möbius transformation taking ℓ_0 onto ℓ . Then m is a perpendicular to ℓ containing z if and only if $f^{-1}(m)$ is a perpendicular to ℓ_0 containing $f^{-1}(z)$. \square

Perpendicularity may be used to characterize hyperparallel lines.

Proposition 12.7.13. *Let ℓ and m be hyperparallel hyperbolic lines. Then there is a unique hyperbolic line n perpendicular to both ℓ and m .*

In the special case that $\ell = \ell_0$ and $m = \mathcal{C}_r(a)$ we can calculate n explicitly:

$$(12.7.1) \quad n = \mathcal{C}_{\sqrt{a^2 - r^2}}(0).$$

Proof. We first consider the special case. Here, since $n \perp \ell_0$, it must have the form $\mathcal{C}_t(0)$ for some $t > 0$. Let $z = \mathcal{C}_t(0) \cap \mathcal{C}_r(a)$. By Corollary perpendicularcircles, 0 , a and z form the vertices of a right triangle with right angle at z . By the Pythagorean theorem,

$$\|a - 0\|^2 = \|z - 0\|^2 + \|z - a\|^2.$$

But $\|z\| = t$ and $\|z - a\| = r$, hence $a^2 = t^2 + r^2$, as claimed. This gives the uniqueness and formula for n . The existence follows from the assumption that ℓ_0 and $\mathcal{C}_r(0)$ are hyperparallel, as then $r < |a|$, so we may realize this equation with an actual line $\mathcal{C}_t(0)$.

The general case follows from this one by applying a Möbius transformation f with $f(\ell_0) = \ell$. Since $f^{-1}(m)$ is hyperparallel to $f^{-1}(\ell) = \ell_0$, it cannot have the form ℓ_a , so the special case does apply. \square

This now characterizes hyperparallel lines.

Corollary 12.7.14. *Two hyperbolic lines are hyperparallel if and only if they have a common perpendicular.*

Proof. It suffices to show that if ℓ and m have a common perpendicular, n , then they are hyperparallel. Let f be a Möbius transformation with $f(\ell_0) = n$. Then $f^{-1}(\ell)$ and $f^{-1}(m)$ are both perpendicular to ℓ_0 and hence may be expressed as $\mathcal{C}_r(0)$ and $\mathcal{C}_s(0)$ for $r \neq s$. But then $f^{-1}(\ell)$ and $f^{-1}(m)$ are hyperparallel, hence ℓ and m are also. \square

12.8. Reflections. Möbius transformations are holomorphic and expressed as fractional linear transformations in the complex variable z . Hyperbolic reflections all involve complex conjugation. Since perpendicularity to the hyperbolic line ℓ_0 is so easy to understand, we will start by defining the reflection across ℓ_0 .

Definition 12.8.1. The hyperbolic reflection, σ_{ℓ_0} , of \mathbb{H} across ℓ_0 is defined by

$$(12.8.1) \quad \sigma_{\ell_0}(z) = -\bar{z}.$$

In particular, if $z = x + iy$ with $x, y \in \mathbb{R}$, then $\sigma_{\ell_0}(z) = -x + iy$.

The following is immediate.

Lemma 12.8.2. *The reflection σ_{ℓ_0} preserves \mathbb{H} and satisfies $\sigma_{\ell_0}^2 = \text{id}$. The Jacobian matrix of σ_{ℓ_0} is*

$$(12.8.2) \quad D\sigma_{\ell_0}(z) = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

for all $z \in \mathbb{H}$. Thus, $\sigma_{\ell_0} : \mathbb{H} \rightarrow \mathbb{H}$ is a diffeomorphism.

The fixed-points of σ_{ℓ_0} are precisely the elements $z \in \mathbb{H}$ with $\text{Re}(z) = 0$. In other words, $\mathbb{H}^{\sigma_{\ell_0}} = \ell_0$.

Continuing the discussion in Remark 12.4.22, that $D\sigma_{\ell_0}(z) < 0$ for all z implies the following.

Corollary 12.8.3. $\sigma_{\ell_0} : \mathbb{H} \rightarrow \mathbb{H}$ is orientation-reversing.

Note that σ_{ℓ_0} is \mathbb{R} -linear and equal to its Jacobian matrix at any point. In fact, $\sigma_{\ell_0} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is an orientation-reversing linear isometry. Since it preserves the pure imaginary part of a complex number z , it is also a hyperbolic isometry:

Lemma 12.8.4. *The reflection σ_{ℓ_0} is an isometry of \mathbb{H} in the sense of Definition 11.1.5:*

$$(12.8.3) \quad \langle D\sigma_{\ell_0}(z)v, D\sigma_{\ell_0}(z)w \rangle_{\sigma_{\ell_0}(z)} = \langle v, w \rangle_z$$

for all $z \in \mathbb{H}$ and all tangent vectors v, w at z . Moreover, σ_{ℓ_0} preserves arc length and hyperbolic distance. Thus $\sigma_{\ell_0} \circ \gamma$ is geodesic for every geodesic $\gamma : \mathbb{R} \rightarrow \mathbb{H}$ and $\sigma_{\ell_0} \in \mathcal{I}(\mathbb{H})$.

The fixed-point set $\mathbb{H}^{\sigma_{\ell_0}}$ of σ_{ℓ_0} is precisely ℓ_0 , i.e.,

$$(12.8.4) \quad \ell_0 = \{z \in \mathbb{H} : \sigma_{\ell_0}(z) = z\}.$$

The effect of σ_{ℓ_0} on hyperbolic lines is easily computed:

$$(12.8.5) \quad \sigma_{\ell_0}(\ell_a) = \ell_{-a} \quad \text{and} \quad \sigma_{\ell_0}(\mathcal{C}_r(a)) = \mathcal{C}_r(-a).$$

In all cases $\sigma_{\ell_0}(\ell)$ is the line with boundary points $\sigma_{\ell_0}(\partial\ell)$.

Finally, if ℓ and m are oriented lines, then the directed angle from $\sigma_{\ell_0}(\ell)$ to $\sigma_{\ell_0}(m)$ is the negative of the directed angle from ℓ to m .

Proof. Let $A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$. For (12.8.3), if $z = x + iy$ with $x, y \in \mathbb{R}$, then

$$\langle D\sigma_{\ell_0}(z)v, D\sigma_{\ell_0}(z)w \rangle_{\sigma_{\ell_0}(z)} = \frac{1}{y^2}(Av \cdot Aw) = \frac{1}{y^2}(v \cdot w) = \langle v, w \rangle_z.$$

Here, \cdot is the standard dot product and the second equality is because A is an orthogonal matrix.

Let $\gamma : [a, b] \rightarrow \mathbb{H}$ be piecewise smooth. Then (12.8.3) implies that

$$\|(\sigma_{\ell_0} \circ \gamma)'(t)\|_{\sigma_{\ell_0} \circ \gamma(t)} = \|\gamma'(t)\|_{\gamma(t)}$$

for all t , hence $\sigma_{\ell_0} \circ \gamma$ and γ have the same arc length.

(12.8.4) and (12.8.5) are easy calculations. Regarding angles, let $\ell \cap m = z = x + iy$. Let γ and δ be unit speed geodesic parametrizations of ℓ and m , respectively, with $\gamma(t) = \delta(s) = z$. Write $\gamma'(t) = ye^{i\theta}$ and $\delta'(s) = ye^{i\varphi}$. Then the chain rule shows that $(\sigma_{\ell_0} \circ \gamma)'(t) = ye^{i(\pi-\theta)}$ and $(\sigma_{\ell_0} \circ \delta)'(s) = ye^{i(\pi-\varphi)}$, and the angle is reversed, as claimed. \square

Corollary 12.8.5. *Let ℓ be a hyperbolic line. Then σ_{ℓ_0} preserves ℓ if and only if either $\ell = \ell_0$ or $\ell \perp \ell_0$.*

Proof. $\mathcal{C}_r(a) \perp \ell_0$ if and only if $a = 0$. \square

Definition 12.8.6. The perpendicular bisector of a geodesic segment $[z, w]$ is the unique hyperbolic line through the midpoint M of $[z, w]$ perpendicular to the hyperbolic line containing z and w .

Lemma 12.8.7. *Let $z \in \mathbb{H} \setminus \ell_0$. Then ℓ_0 is the perpendicular bisector of $[z, \sigma_{\ell_0}(z)]$.*

Proof. Note first that $\|z\| = \|\sigma_{\ell_0}(z)\|$, as σ_{ℓ_0} merely alters the sign on the real part of z . Thus, $\sigma_{\ell_0}(z) \in \mathcal{C}_{\|z\|}(0)$, the unique line through z perpendicular to ℓ_0 . Let $M = i\|z\| = \mathcal{C}_{\|z\|}(0) \cap \ell_0$. Then M is fixed by σ_{ℓ_0} . Since σ_{ℓ_0} is an isometry,

$$(12.8.6) \quad d_{\mathbb{H}}(z, M) = d_{\mathbb{H}}(\sigma_{\ell_0}(z), M).$$

Since M lies on the geodesic segment between z and $\sigma_{\ell_0}(z)$ and since geodesics minimize arc length,

$$(12.8.7) \quad d_{\mathbb{H}}(z, \sigma_{\ell_0}(z)) = d_{\mathbb{H}}(z, M) + d_{\mathbb{H}}(M, \sigma_{\ell_0}(z)) = 2d_{\mathbb{H}}(z, M),$$

so M is the midpoint of that segment. \square

The following is by now an expected property of perpendicular bisectors.

Proposition 12.8.8. *Let $z \in \mathbb{H} \setminus \ell_0$. Then*

$$(12.8.8) \quad \ell_0 = \{\zeta \in \mathbb{H} : d_{\mathbb{H}}(z, \zeta) = d_{\mathbb{H}}(\sigma_{\ell_0}(z), \zeta)\}.$$

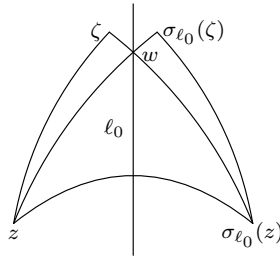
Proof. Let $\zeta \in \ell_0$. Then ζ is fixed by σ_{ℓ_0} . Since σ_{ℓ_0} is an isometry, we get

$$d_{\mathbb{H}}(z, \zeta) = d_{\mathbb{H}}(\sigma_{\ell_0}(z), \sigma_{\ell_0}(\zeta)) = d_{\mathbb{H}}(\sigma_{\ell_0}(z), \zeta),$$

as claimed.

Conversely, suppose $d_{\mathbb{H}}(z, \zeta) = d_{\mathbb{H}}(\sigma_{\ell_0}(z), \zeta)$, and suppose $\zeta \notin \ell_0$. Then $\operatorname{Re}(\zeta) \neq 0$. Then $\operatorname{Re}(\zeta)$ has the same sign as exactly one of z and $\sigma_{\ell_0}(z)$. Suppose both $\operatorname{Re}(z)$ and $\operatorname{Re}(\zeta)$ are negative. The other cases are similar.

By the intermediate value theorem, $[z, \sigma_{\ell_0}(\zeta)]$ intersects ℓ_0 , by necessity in one point, w , as two distinct hyperbolic lines have at most one point of intersection. And $[z, \zeta]$ cannot intersect ℓ_0 , as nonvertical hyperbolic lines are implicitly functions of x .



By assumption,

$$d_{\mathbb{H}}(z, \zeta) = d_{\mathbb{H}}(\sigma_{\ell_0}(z), \zeta) = d_{\mathbb{H}}(\sigma_{\ell_0}(z), w) + d_{\mathbb{H}}(w, \zeta) = d_{\mathbb{H}}(z, w) + d_{\mathbb{H}}(w, \zeta),$$

as σ_{ℓ_0} is an isometry fixing w . But this forces $w \in [z, \zeta]$ by Proposition 12.6.2, contradicting that $[z, \zeta] \cap \ell_0 = \emptyset$. \square

We can use this detailed understanding of σ_{ℓ_0} to study the hyperbolic reflection across an arbitrary hyperbolic line.

Proposition 12.8.9. *Let ℓ be a hyperbolic line and let f be a Möbius transformation with $f(\ell_0) = \ell$. Then the fixed-point set $\mathbb{H}^{f\sigma_{\ell_0}f^{-1}}$ is equal to ℓ . Moreover, let $z \in \mathbb{H} \setminus \ell$ and let m be the perpendicular to ℓ containing z . Let $M = m \cap \ell$ and let w be the unique point on m unequal to z with*

$$d_{\mathbb{H}}(w, M) = d_{\mathbb{H}}(M, z).$$

Then $f\sigma_{\ell_0}f^{-1}(z) = w$ and ℓ is the perpendicular bisector of $[z, w]$. As was the case for ℓ_0 ,

$$(12.8.9) \quad \ell = \{\zeta \in \mathbb{H} : d_{\mathbb{H}}(z, \zeta) = d_{\mathbb{H}}(\sigma_{\ell}(z), \zeta)\}.$$

Note this is independent of the choice of f : if g is another Möbius transformation with $g(\ell_0) = \ell$, then $g\sigma_{\ell_0}g^{-1}$ will have exactly the same effect on each element of \mathbb{H} as $f\sigma_{\ell_0}f^{-1}$ does. I.e.,

$$(12.8.10) \quad f\sigma_{\ell_0}f^{-1} = g\sigma_{\ell_0}g^{-1} \quad \text{if} \quad f(\ell_0) = g(\ell_0) = \ell.$$

For such an f , we write

$$(12.8.11) \quad \sigma_\ell = f\sigma_{\ell_0}f^{-1},$$

and call it the reflection of \mathbb{H} across ℓ .

Proof. That $\mathbb{H}^{f\sigma_{\ell_0}f^{-1}} = f(\mathbb{H}^{\sigma_{\ell_0}})$ is shown in the proof of Lemma 5.5.3. Of course, $f(\mathbb{H}^{\sigma_{\ell_0}}) = f(\ell_0) = \ell$.

Let z and m be as in the statement. Since Möbius transformations preserve angles, $f^{-1}(m)$ is the perpendicular to $f^{-1}(\ell) = \ell_0$ containing $f^{-1}(z)$. Now $\sigma_{\ell_0}(f^{-1}(z))$ is the unique point on $f^{-1}(m)$ other than $f^{-1}(z)$ with

$$d_{\mathbb{H}}(\sigma_{\ell_0}(f^{-1}(z)), f^{-1}(m)) = d_{\mathbb{H}}(f^{-1}(m), f^{-1}(z)).$$

Now, apply f to this picture, and the result follows. \square

The situation of Proposition 12.8.9 is generic for perpendicular bisectors.

Corollary 12.8.10. *Let $z \neq w \in \mathbb{H}$ and let ℓ be the perpendicular bisector of $[z, w]$. Then $\sigma_\ell(z) = w$. Thus,*

$$(12.8.12) \quad \ell = \{\zeta \in \mathbb{H} : d_{\mathbb{H}}(z, \zeta) = d_{\mathbb{H}}(w, \zeta)\}.$$

Proof. This comes out of the description of σ_ℓ in Proposition 12.8.9. \square

Proposition 12.8.9 does not give an explicit calculation of σ_ℓ , but that can be done as an exercise. Since Möbius transformations are angle-preserving isometries and σ_{ℓ_0} is an angle-reversing isometry, we obtain the following.

Corollary 12.8.11. *Let ℓ be a hyperbolic line. Then σ_ℓ is an isometry of \mathbb{H} in the sense of Definition 11.1.5:*

$$(12.8.13) \quad \langle D\sigma_\ell(z)v, D\sigma_\ell(z)w \rangle_{\sigma_\ell(z)} = \langle v, w \rangle_z$$

for all $z \in \mathbb{H}$ and all tangent vectors v, w at z . Moreover, σ_ℓ preserves arc length and hyperbolic distance. Thus $\sigma_\ell \circ \gamma$ is geodesic for every geodesic $\gamma : \mathbb{R} \rightarrow \mathbb{H}$ and $\sigma_\ell \in \mathcal{I}(\mathbb{H})$.

If n and m are oriented lines, then the directed angle from $\sigma_\ell(n)$ to $\sigma_\ell(m)$ is the negative of the directed angle from n to m .

The following is a consequence of Corollary 12.8.5.

Corollary 12.8.12. *Let ℓ and m be hyperbolic lines. Then σ_ℓ preserves m if and only if either $m = \ell$ or $m \perp \ell$.*

Proof. Let f be a Möbius transformation with $f(\ell_0) = \ell$. We claim that $f\sigma_{\ell_0}f^{-1}$ preserves $f(n)$ if and only if σ_{ℓ_0} preserves n :

$$f\sigma_{\ell_0}f^{-1}(f(n)) = f(n) \quad \Leftrightarrow \quad f\sigma_{\ell_0}(n) = f(n)$$

$$\Leftrightarrow \sigma_{\ell_0}(n) = n,$$

where the last equivalence is obtained by applying f^{-1} to the previous one. Thus, $f\sigma_{\ell_0}f^{-1}$ preserves m if and only if σ_{ℓ_0} preserves $f^{-1}(m)$. \square

12.9. Generalized Möbius transformations. We define a larger class of transformations that includes both the Möbius transformations and the hyperbolic reflections.

Definition 12.9.1. Let

$$\tilde{\text{SL}}_2(\mathbb{R}) = \{A \in \text{GL}_2(\mathbb{R}) : \det A = \pm 1\}.$$

For $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \tilde{\text{SL}}_2(\mathbb{R})$ with $\det A = -1$, define $\varphi_A : \bar{\mathbb{C}} \rightarrow \bar{\mathbb{C}}$ by

$$(12.9.1) \quad \varphi_A(z) = \begin{cases} \frac{a\bar{z} + b}{c\bar{z} + d} & z \neq -\frac{d}{c}, \infty \\ \infty & z = -\frac{d}{c} \\ \frac{a}{c} & z = \infty. \end{cases}$$

This extends the definition of φ_A for $A \in \text{SL}_2(\mathbb{R})$. The generalized Möbius transformations are now given by

$$(12.9.2) \quad \widetilde{\text{Möb}} = \{\varphi_A : A \in \tilde{\text{SL}}_2(\mathbb{R})\}.$$

The following is a straightforward calculation, similar to the one for Lemma 12.2.1. The only difference is that conjugation gets applied at most twice in the calculation, depending on the determinants of the two matrices.

Lemma 12.9.2. Let $A, B \in \tilde{\text{SL}}_2(\mathbb{R})$. Then,

$$(12.9.3) \quad \varphi_A \circ \varphi_B = \varphi_{AB}.$$

Thus, there is a group homomorphism $\varphi : \tilde{\text{SL}}_2(\mathbb{R}) \rightarrow \widetilde{\text{Möb}}$ via $\varphi(A) = \varphi_A$.

Corollary 12.9.3. Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \tilde{\text{SL}}_2(\mathbb{R})$ with $\det A = -1$. Then

$$(12.9.4) \quad \varphi_A = \varphi_B \sigma_{\ell_0} \quad \text{for} \quad B = \begin{bmatrix} -a & b \\ -c & d \end{bmatrix}.$$

Here, $\det B = 1$, hence φ_B is Möbius.

Thus, $\widetilde{\text{Möb}}$ has index 2 in $\widetilde{\text{Möb}}$, with

$$(12.9.5) \quad \widetilde{\text{Möb}} = \text{Möb} \cup \text{Möb} \sigma_{\ell_0}.$$

Proof. Let $J = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$. Then $\varphi_J = \sigma_{\ell_0}$, so (12.9.4) is immediate from Lemma 12.9.2. Thus, any element of $\widetilde{\text{Möb}} \setminus \text{Möb}$ must lie in the right coset $\text{Möb} \sigma_{\ell_0}$, and (12.9.5) follows. In particular $[\widetilde{\text{Möb}} : \text{Möb}] = 1$ or 2, depending on whether σ_{ℓ_0} is Möbius. But nonidentity Möbius transformations have at most one fixed-point in \mathbb{H} , and σ_{ℓ_0} has an entire hyperbolic line of fixed-points, so σ_{ℓ_0} is not Möbius, hence $[\widetilde{\text{Möb}} : \text{Möb}] = 2$. \square

Another proof that the index is 2 and not 1 comes from the fact that every element of Möb is angle-preserving, while every element of Möb σ_{ℓ_0} is angle-reversing. This, of course, is a result of the fact that Möbius transformations are orientation-preserving, while σ_{ℓ_0} is orientation-reversing.

Corollary 12.9.4. *Generalized Möbius transformations induce diffeomorphisms from \mathbb{H} to \mathbb{H} and extend to mappings from $\overline{\mathbb{H}}$ to $\overline{\mathbb{H}}$. In addition, they are distance preserving and hence lie in $\mathcal{I}(\mathbb{H})$.*

For any generalized Möbius transformation f and any hyperbolic line ℓ , $f(\ell)$ is the line with boundary points $f(\partial\ell)$.

Proof. Both Möbius transformations and σ_{ℓ_0} have these properties. \square

In particular, $\widetilde{\text{Möb}} \subset \mathcal{I}(\mathbb{H})$. We now show the two are equal, and hence that $\mathcal{I}(\mathbb{H})$ is a group, as claimed. Note that, just as in the Euclidean and spherical cases, we have not assumed the elements of $\mathcal{I}(\mathbb{H})$ are continuous. So the following is striking, as before.

Theorem 12.9.5. *Every distance-preserving transformation of \mathbb{H} is a generalized Möbius function.*

Proof. Let $f \in \mathcal{I}(\mathbb{H})$. Since f preserves distance, $d_{\mathbb{H}}(f(i), f(2i)) = d_{\mathbb{H}}(i, 2i)$. By Corollary 12.5.4, there is a Möbius transformation g with $g(i) = f(i)$ and $g(2i) = f(2i)$. Replacing f by $g^{-1}f$, we may assume $f(i) = i$ and $f(2i) = 2i$. By Corollary 12.6.5 this implies f is the identity on ℓ_0 .

By Corollary 12.8.10, if f is the identity on two points z and w and if ℓ is the perpendicular bisector $[z, w]$, then $f(\ell) = \ell$. Note that each hyperbolic line of the form $\mathcal{C}_r(0)$ is the perpendicular bisector of the geodesic segment joining a pair of points on ℓ_0 . Thus $f(\mathcal{C}_r(0)) = \mathcal{C}_r(0)$ for all $r > 0$.

For a fixed r and $z \in \mathcal{C}_r(0) \setminus \ell_0$, ir is the midpoint of the segment $[z, \sigma_{\ell_0}(z)]$. Since ir is fixed by f , $f(z) \in \{z, \sigma_{\ell_0}(z)\}$. In particular, one of f and $\sigma_{\ell_0} \circ f$ is the identity on $\ell_0 \cup \{z\}$, and hence on $\mathcal{C}_r(0)$ as well, as it is the identity on two points of that line. Replacing f by $\sigma_{\ell_0} \circ f$, if necessary, we may assume f is the identity on $\ell_0 \cup \mathcal{C}_r(0)$.

If f were continuous, this would be enough to deduce that f is the identity on all of \mathbb{H} . But the same argument holds for every r : for $r > 0$, f coincides either with the identity or with σ_{ℓ_0} on $\mathcal{C}_r(0)$.

There are, of course, infinitely many possible r to choose from. We claim it is enough to know that f is the identity on exactly two of them. Suppose this is the case. Say f is the identity on $\mathcal{C}_r(0)$ and $\mathcal{C}_s(0)$ with $0 < r < s$. Then for each $a \in (-r, r)$, ℓ_a meets each of these lines in a point, and hence f is the identity on two points of ℓ_a . By Corollary 12.6.5, f is the identity on all of ℓ_a . But each line $\mathcal{C}_t(0)$ then has infinitely many points on which it is the identity, so f is the identity on $\mathcal{C}_t(0)$. Since each $z \in \mathbb{H}$ lies on $\mathcal{C}_{\|z\|}(0)$, f is the identity everywhere.

But, of course, if f were the identity on only one line $\mathcal{C}_r(0)$, then it must coincide with σ_{ℓ_0} on all the others, leading to a contradiction, as then $\sigma_{\ell_0} \circ f$ is the identity everywhere by the argument above. \square

Recall that hyperbolic reflections have the form $f\sigma_{\ell_0}f^{-1}$ for f Möbius, and hence are orientation-reversing generalized Möbius transformations. So what do the other generalized Möbius transformations look like? As in the Euclidean case, there are also orientation-reversing hyperbolic isometries that have no fixed-points in \mathbb{H} . Indeed, they are enough like glide reflections that we'll call them that.

Definition 12.9.6. A hyperbolic glide reflection is a composite $h\sigma_{\ell}$ with h a hyperbolic transformation whose translation axis is ℓ . We call this its standard form and call ℓ its axis.

As usual, we shall study these by first considering the case where $\ell = \ell_0$. The following is an easy calculation and is left to the reader.

Lemma 12.9.7. *Let $a > 0$, $a \neq 1$. Then the hyperbolic Möbius transformation h_a commutes with σ_{ℓ_0} . Let $f = h_a\sigma_{\ell_0}$, a glide reflection with axis ℓ_0 . Then*

$$\bar{\mathbb{H}}^f = \{0, \infty\}.$$

On lines, we have $f(\ell_b) = \ell_{-ab}$ and $f(\mathcal{C}_r(b)) = \mathcal{C}_{ar}(-ab)$. Thus, the only line preserved by f is ℓ_0 . On ℓ_0 , f agrees with h_a .

Since h_a commutes with σ_{ℓ_0} , we have

$$(12.9.6) \quad f^2 = (h_a\sigma_{\ell_0})^2 = h_a^2\sigma_{\ell_0}^2 = h_a^2 = h_{a^2},$$

a hyperbolic Möbius transformation with axis ℓ_0 .

Corollary 12.9.8. *Let $f = h\sigma_{\ell}$ be a hyperbolic glide reflection in standard form (so that h is a hyperbolic Möbius transformation with translation axis ℓ). Then h commutes with σ_{ℓ} and the only line preserved by f is ℓ . Moreover,*

$$(12.9.7) \quad \mathbb{H}^f = \partial\ell.$$

Since h commutes with σ_{ℓ} , we have

$$(12.9.8) \quad f^2 = (h\sigma_{\ell})^2 = h^2\sigma_{\ell}^2 = h^2,$$

a hyperbolic Möbius transformation with axis ℓ .

Proof. Let g be a Möbius transformation with $g(\ell_0) = \ell$. Then $g^{-1}hg$ is hyperbolic with translation axis ℓ_0 and $g^{-1}\sigma_{\ell}g = \sigma_{\ell_0}$. So $g^{-1}hg = h_a$ for some $a \neq 1$ and it commutes with σ_{ℓ_0} . Moreover, $g^{-1}fg = h_a\sigma_{\ell_0}$ and the result follows by conjugating that by g . \square

The following observation is useful.

Lemma 12.9.9. *An orientation-reversing isometry of \mathbb{H} fixes exactly two points of $\partial\mathbb{H}$.*

Proof. The orientation-reversing isometries have the form $f = \varphi_A$ with $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \tilde{\text{SL}}_2(\mathbb{R})$ of determinant -1 . For $x \in \mathbb{R} \subset \partial\mathbb{H}$, $x \neq -\frac{d}{c}$,

$$(12.9.9) \quad f(x) = \frac{ax + b}{cx + d}.$$

As in the Möbius case, f fixes ∞ if and only if $c = 0$, in which case $f(x) = x$ when $ax + b = dx$. Since $ad = \det A = -1$, $a \neq d$ and there is a unique solution for $x \in \mathbb{R}$.

If ∞ is not fixed, the real solutions of $f(x) = x$ are the solutions of

$$ax + b = cx^2 + dx,$$

which are the roots of the quadratic $cx^2 + (d-a)x - b = 0$. The discriminant here is $(d-a)^2 + 4bc$. Since $\det A = -1$, $4bc = 4ad + 4$, so the discriminant is $\text{tr}(A)^2 + 4$, which is strictly positive, providing two real roots. \square

We obtain the following analogue of the Euclidean case.

Theorem 12.9.10. *Every orientation-reversing isometry f of \mathbb{H} is either a reflection or a glide reflection.*

Proof. By Lemma 12.9.9, f fixes exactly two points of $\partial\mathbb{H}$. Let ℓ be the hyperbolic line with these two points as its endpoints. Then $f\sigma_\ell$ is a Möbius transformation fixing the two boundary points $\partial\ell$. In particular, either $f\sigma_\ell$ is the identity (in which case $f = \sigma_\ell$), or $f\sigma_\ell$ is a nonidentity Möbius transformation fixing these two points. In this case, $f\sigma_\ell$ must be a hyperbolic Möbius transformation h with translation axis ℓ . But then $f = h\sigma_\ell$ is a hyperbolic glide reflection with axis ℓ . \square

The following will be useful.

Corollary 12.9.11. *Let $f \neq \text{id}$ be a hyperbolic isometry with $\ell \subset \mathbb{H}^f$ for some hyperbolic line ℓ . Then $f = \sigma_\ell$.*

Proof. Reflections are the only nonidentity isometries of \mathbb{H} with more than one fixed-point in \mathbb{H} . Moreover, $\mathbb{H}^{\sigma_\ell} = \ell$. \square

The following could also be proven directly, using Proposition 12.8.9 and the fact that $\rho_{(i, \frac{\pi}{2})}(\ell_0) = \mathcal{C}_1(0)$.

Corollary 12.9.12. *The reflection across $\mathcal{C}_r(0)$ is given by*

$$(12.9.10) \quad \sigma_{\mathcal{C}_r(0)}(z) = \frac{r^2}{\bar{z}}.$$

The reflection across ℓ_a is given by

$$(12.9.11) \quad \sigma_{\ell_a}(z) = 2a - \bar{z}.$$

Proof. $\mathcal{C}_r(0) = \{z \in \mathbb{H} : z\bar{z} = r^2\}$. $\ell_a = \{z \in \mathbb{H} : z + \bar{z} = 2a\}$. \square

12.10. Calculus of isometries. As in the Euclidean case, we show every orientation-preserving isometry of \mathbb{H} is the product of two reflections. As usual, we start with examples we can analyze easily and extend by conjugating.

Lemma 12.10.1. *The composite of the reflections in two distinct lines centered at 0 is given by*

$$(12.10.1) \quad \sigma_{C_r(0)}\sigma_{C_s(0)} = h_{(\frac{r}{s})^2},$$

a hyperbolic Möbius transformation with translation axis ℓ_0 . Given positive real numbers a and r , there are unique positive real numbers s and t with

$$(12.10.2) \quad h_a = \sigma_{C_r(0)}\sigma_{C_s(0)} = \sigma_{C_t(0)}\sigma_{C_r(0)}.$$

The composite of two reflections in vertical lines is given by

$$(12.10.3) \quad \sigma_{\ell_a}\sigma_{\ell_b} = p_{2(a-b)},$$

a parabolic Möbius transformation fixing ∞ . Given real numbers s and a , there are unique real numbers b and c with

$$(12.10.4) \quad p_s = \sigma_{\ell_a}\sigma_{\ell_b} = \sigma_{\ell_c}\sigma_{\ell_a}.$$

Proof. These are easy calculations based on Corollary 12.9.12. \square

We can immediately extend this to general cases. Recall from Proposition 12.7.13 that if ℓ and m are hyperparallel, there is a unique line perpendicular to both.

Proposition 12.10.2. *Let ℓ and m be hyperparallel lines and let n be the unique line perpendicular to both. Then $\sigma_\ell\sigma_m$ is a hyperbolic Möbius transformation with n as its axis of translation.*

Given a hyperbolic Möbius transformation h and a line ℓ perpendicular to its translation axis, n , there are unique hyperbolic lines m_1 and m_2 perpendicular to n such that

$$(12.10.5) \quad h = \sigma_\ell\sigma_{m_1} = \sigma_{m_2}\sigma_\ell.$$

Proof. If ℓ and m are hyperparallel and n is perpendicular to both, then n is preserved by both σ_ℓ and σ_m , and hence is preserved by their composite. Their composite is orientation-preserving, and hence is hyperbolic with n as its translation axis by Proposition 12.4.21.

Now given a hyperbolic Möbius transformation h and a line ℓ perpendicular to its translation axis, n , let f be a Möbius transformation with $f(\ell_0) = n$. The existence and uniqueness of m_1 and m_2 follow by applying (12.10.2) to $h_a = f^{-1}hf$ and $C_r(0) = f^{-1}(\ell)$; then apply f to the results. \square

The parabolic case is the following.

Proposition 12.10.3. *Let ℓ and m be distinct parallel hyperbolic lines that are not hyperparallel. Let $\bar{\ell} \cap \bar{m} = x \in \bar{\mathbb{H}}$. Then $\sigma_\ell\sigma_m$ is a parabolic Möbius transformation fixing x .*

Given a parabolic transformation p fixing x and a hyperbolic line ℓ with $x \in \partial\ell$, there are unique hyperbolic lines m and n with $\partial m \cap \partial n = x$ such that

$$(12.10.6) \quad p = \sigma_\ell \sigma_m = \sigma_n \sigma_\ell.$$

Proof. Let f be a Möbius transformation taking ∞ to x . Apply f^{-1} to the lines, replace p by $f^{-1}pf$, and apply Lemma 12.10.1. \square

Composition of reflections across intersecting lines has the expected behavior.

Proposition 12.10.4. *Let ℓ and m be hyperbolic lines with $\ell \cap m = w$. Then*

$$(12.10.7) \quad \sigma_\ell \sigma_m = \rho_{(w, \theta)},$$

where θ is twice the directed angle from m to ℓ .

As in the Euclidean case, since we're doubling the angle, it doesn't matter how we orient the two lines when calculating the angle. This will come out in the proof.

Proof of Proposition 12.10.4. The composite $f = \sigma_\ell \sigma_m$ is orientation-preserving, and hence is a Möbius transformation. Since both σ_ℓ and σ_m fix w , so does f . So $f = \rho_{(w, \theta)}$ for some θ and it suffices to calculate θ .

To do so, we choose orientations of ℓ and m . These will determine the directed angles we need. The angle of rotation is then given by the directed angle from m to $f(m)$. Since m is fixed by σ_m , $f(m) = \sigma_\ell(m)$.

Since angles are determined mod 2π , the angle from m to $\sigma_\ell(m)$ is the sum of the angle from m to ℓ and the angle from ℓ to $\sigma_\ell(m)$. The latter is the angle from $\sigma_\ell(\ell)$ to $\sigma_\ell(m)$, which in turn is the negative of the angle from ℓ to m , as reflections reverse angle measure. But the negative of the angle from ℓ to m is the angle from m to ℓ , and the result follows. \square

12.11. Exercises.

1. Let ℓ be a hyperbolic line and let $w \in \mathbb{H} \setminus \ell$. Show there is a one-to-one correspondence between the lines through w hyperparallel to ℓ and the lines perpendicular to ℓ .
2. Show that $f(z) = \frac{1}{\bar{z}}$ is the reflection across $\mathcal{C}_1(0)$.
3. Find the formula for the reflection across $\mathcal{C}_r(a)$.
4. Find the formula for the reflection across ℓ_a .
5. Let T be the hyperbolic triangle with vertices at i , $1+2i$ and $-1+2i$.
 - (a) What are the measures of the angles in T ?
 - (b) What is the angle sum as a fraction of π ?
 - (c) What are the hyperbolic lengths of the sides of T ?
 - (d) How do the above measurements compare to the Euclidean lengths and angles on the Euclidean triangle connecting those three points?

6. Show that i , $1 + i$ and $-1 + i$ are not collinear in \mathbb{H} despite being collinear in the Euclidean plane.
7. Find a hyperbolic line ℓ perpendicular to both $\mathcal{C}_1(0)$ and $\mathcal{C}_2(5)$.
 - (a) What are the values of its intersections with $\mathcal{C}_1(0)$ and $\mathcal{C}_2(5)$?
 - (b) Find the eigenvalues of the matrix $A \in \mathrm{SL}_2(\mathbb{R})$ inducing

$$\sigma_{\mathcal{C}_1(0)}\sigma_{\mathcal{C}_2(5)}.$$

8. Let T be the hyperbolic triangle with vertices $\frac{1}{\sqrt{3}}i$, $\frac{1}{\sqrt{3}} + \frac{2}{\sqrt{3}}i$ and $-\frac{1}{\sqrt{3}} + \frac{2}{\sqrt{3}}i$.
 - (a) Show that T is equilateral with centroid i .
 - (b) Calculate the angles of T .
 - (c) What is the perpendicular bisector, ℓ , of the geodesic segment $[\frac{1}{\sqrt{3}}i, \frac{1}{\sqrt{3}} + \frac{2}{\sqrt{3}}i]$?
 - (d) Find the formula for σ_ℓ and verify that it interchanges the endpoints of the above segment.
 - (e) Calculate $\sigma_\ell\sigma_{\ell_0}$.
9. Show that $\mathcal{S}_{\mathrm{Möb}}(\ell_0)$ is isomorphic to the group $\mathcal{O}(\ell)$ of orientation-preserving symmetries of a Euclidean line in \mathbb{R}^2 .
10. Show that $\mathcal{S}_{\mathcal{I}(\mathbb{H})}(\ell_0)$ is isomorphic to the group $\mathcal{S}(\ell)$ of symmetries of a Euclidean line in \mathbb{R}^2 .
11. Show that $\widetilde{\mathrm{Möb}}_i = \mathcal{S}_{\mathcal{I}(\mathbb{H})}(\{i\})$ is isomorphic to $\mathcal{O}(2) \cong \mathcal{S}(\{0\})$, the group of Euclidean symmetries of $\{0\}$ in \mathbb{R}^2 .
12. Let ℓ and m be hyperbolic lines. Show that σ_ℓ and σ_m commute if and only if either $\ell = m$ or $\ell \perp m$.

Appendix A. Spaces with identifications

The topological spaces we study in this book will all be quotient spaces of metric spaces. We shall only be as general as we need to be in defining these.

A.1. Metric topology.

Definition A.1.1. A metric on a set X is a distance function

$$d : X \times X \rightarrow [0, \infty) \subset \mathbb{R}$$

with the following properties.

- (1) d is symmetric: $d(x, y) = d(y, x)$ for all $x, y \in X$.
- (2) d is positive-definite: $d(x, y) = 0$ if and only if $x = y$.
- (3) The triangle inequality holds: $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in X$.

Properties (1) and (2) and a strengthened version of (3) are satisfied by the Euclidean distance function by Proposition 2.3.8.

A topological space consists of a set X together with a notion of which subsets $U \subset X$ will be considered open.

Definition A.1.2. A metric space consists of a set X with a metric d as above, and we declare that $U \subset X$ is open if for each $x \in U$ there exists $\epsilon \in (0, \infty)$ such that

$$d(x, y) < \epsilon \quad \Rightarrow \quad y \in U.$$

We write

$$B_\epsilon(x) = \{y \in X : d(x, y) < \epsilon\}$$

and call it the open ball about x of radius ϵ . Thus, our definition of open set in a metric space X is that U is open if and only if for each $x \in U$ there is an open ball about x contained in U .

Example A.1.3. In the real numbers \mathbb{R} , the distance between points x and y is $|x - y|$. Thus, $B_\epsilon(x)$ is the open interval

$$(A.1.1) \quad B_\epsilon(x) = (x - \epsilon, x + \epsilon).$$

Now let $a < b$ and let $x \in (a, b)$. Set $\epsilon = \min(|x - a|, |x - b|)$. Then $B_\epsilon(x) \subset (a, b)$. So (a, b) is open.

Similarly, (a, ∞) and $(-\infty, b)$ are open.

We shall need the following.

Lemma A.1.4. $B_\epsilon(x)$ is open in X .

Proof. This is a consequence of the triangle inequality. If $y \in B_\epsilon(x)$ and let $d = d(x, y)$. Then $d < \epsilon$, so $\epsilon - d > 0$. We claim that $B_{\epsilon-d}(y) \subset B_\epsilon(x)$, and that will complete the proof that $B_\epsilon(x)$ is open.

Thus, let $z \in B_{\epsilon-d}(y)$. Then

$$d(x, z) < d(x, y) + d(y, z)$$

$$< d + (\epsilon - d) = \epsilon$$

by the triangle inequality. So $z \in B_\epsilon(x)$. □

By the definition of open set, an arbitrary union of open sets is open. We obtain the following.

Corollary A.1.5. *A subset of a metric space is open if and only if it is a union of open balls.*

More general topological spaces are required to satisfy the following axioms.

Definition A.1.6. A topology for a space X consists of a collection \mathcal{U} of subsets of X , the *open subsets* of X , satisfying:

- (1) \emptyset and X are open (i.e., lie in \mathcal{U}).
- (2) The union of an arbitrary collection of open sets is open.
- (3) The intersection of any finite collection of open sets is open.

X , together with a topology, is a topological space. The elements of X are called its points.

For a point $x \in X$, a neighborhood of x is an open set containing it.

Of course, inductively, (3) is equivalent to saying that the intersection of any two open sets is open.

Definition A.1.7. The open sets in a metric space clearly satisfy (1)–(3), and hence form a topology on X . We call it the topology induced by the metric d . Different metrics on X may induce different topologies. But sometimes different metrics will induce the same topology.

We call a topological space *metrizable* if its topology is induced by some (often unspecified) metric on X .

Examples A.1.8. There are two obvious topologies one could place on a set X .

- (1) The discrete topology on X is the one in which every subset of X is open. Since arbitrary unions of open sets must be open, this is equivalent to saying each point in X is open.
- (2) The indiscrete topology on X is the one in which the only open sets are \emptyset and X .

If X has more than one point, then the indiscrete topology on X does not satisfy the following property.

Definition A.1.9. Let X be a topological space and let $x, y \in X$. We say x and y may be separated in X if there are open sets U and V with $x \in U$ and $y \in V$ such that $U \cap V = \emptyset$. We call a choice of such U and V a separation of x and y in X .

We say X is *Hausdorff* if each pair of distinct points $x, y \in X$ may be separated in X .

Lemma A.1.10. *A metric space X is Hausdorff.*

Proof. If x and y are distinct and $d = d(x, y)$ then $B_{\frac{d}{2}}(x) \cap B_{\frac{d}{2}}(y) = \emptyset$ by the triangle inequality. \square

Topological spaces were developed to study the notion of continuity, which is a key ingredient in the intermediate value theorem. Thus, continuity plays a role in basic calculus. The reader is probably aware of the metric definition of continuity (Definition 8.1.2). We give a definition here appropriate for general topological spaces.

Definition A.1.11. A function $f : X \rightarrow Y$ between topological spaces is continuous if $f^{-1}(U)$ is open in X whenever U is open in Y . A homeomorphism of topological spaces is a continuous, one-to-one, onto map $f : X \rightarrow Y$ whose inverse function is continuous. We write $f : X \xrightarrow{\cong} Y$ for a homeomorphism $f : X \rightarrow Y$.

In particular, homeomorphic spaces are topologically indistinguishable in the same way that isomorphic vector spaces are indistinguishable as vector spaces. So the topological setting strips away notions of differentiability until they are “added back” with smooth atlases, etc.

We should at least show that our new notion of continuity coincides with the metric notion.

Lemma A.1.12. *Let $f : X \rightarrow Y$ be a function between the metric spaces X and Y . Then f is continuous in the topological sense if and only if it satisfies the (ϵ, δ) definition of continuity that for each $x \in X$ and each $\epsilon > 0$, there exists $\delta > 0$ such that*

$$d(x, y) < \delta \quad \Rightarrow \quad d(f(x), f(y)) < \epsilon,$$

i.e., $B_\delta(x) \subset f^{-1}B_\epsilon(f(x))$.

Proof. Suppose f satisfies the (ϵ, δ) definition of continuity, and let $U \subset Y$ open. We wish to show $f^{-1}(U)$ is open in X . Let $x \in f^{-1}(U)$. We wish to find an open ball about x contained in $f^{-1}(U)$. Since U is open in Y and $f(x) \in U$, there exists $\epsilon > 0$ with $B_\epsilon(f(x)) \subset U$. But the (ϵ, δ) definition of continuity now says there exists $\delta > 0$ with $B_\delta(x) \subset f^{-1}B_\epsilon(f(x))$. But this in turn is contained in $f^{-1}(U)$, so $f^{-1}(U)$ is open as desired.

Conversely, if the inverse image of every open set is open, we can simply apply this to each $B_\epsilon(f(x))$, which is open by Lemma A.1.4. And that, in turn, provides an open ball around x contained in $f^{-1}(B_\epsilon(f(x)))$. \square

Note that a metric on X restricts to a metric on any subset of X , and hence induces a metric topology as above. To avoid ambiguity we write

$$(A.1.2) \quad B_\epsilon(x, Y) = \{y \in Y : d(x, y) < \epsilon\}$$

for the ϵ -ball in Y about $x \in Y$ (as differentiated from

$$B_\epsilon(x, X) = \{y \in X : d(x, y) < \epsilon\},$$

the ϵ -ball about x in X).

Definition A.1.13. A subset C of a topological space X is closed if its complement, $X \setminus C$, is open.

Lemma A.1.14. *The closed sets in a space X satisfy the following:*

- (1) \emptyset and X are closed.
- (2) The intersection of an arbitrary family of closed sets is closed.
- (3) The union of finitely many closed sets is closed.

Proof. This is immediate from Definition A.1.6 via de Morgan's laws: if $\{Y_\alpha : \alpha \in A\}$ is an arbitrary family of subsets of a set X , then:

$$(A.1.3) \quad \begin{aligned} X \setminus \bigcap_{\alpha \in A} Y_\alpha &= \bigcup_{\alpha \in A} (X \setminus Y_\alpha), \\ X \setminus \bigcup_{\alpha \in A} Y_\alpha &= \bigcap_{\alpha \in A} (X \setminus Y_\alpha). \end{aligned} \quad \square$$

There are as many examples of closed sets as there are of open sets. Some are particularly valuable.

Lemma A.1.15. *Let X be a metric space. Let $x \in X$ and let $r > 0$. Then the closed ball $\bar{B}_r(x) = \{y \in X : d(x, y) \leq r\}$ is closed in X .*

Proof. We show the complement of $\bar{B}_r(x)$ is open. Let $y \in X \setminus \bar{B}_r(x)$. Then $d(x, y) = s > r$. Then $B_{s-r}(y)$ is disjoint from $\bar{B}_r(x)$: if z were in $B_{s-r}(y) \cap \bar{B}_r(x)$, we would have

$$\begin{aligned} d(x, y) &\leq d(x, z) + d(z, y) \\ &\leq r + d(z, y) \\ &< r + (s - r) = s, \end{aligned}$$

a contradiction. □

Closed sets are vital to a number of geometric questions. They can be used to study continuity questions.

Lemma A.1.16. *A function $f : X \rightarrow Y$ between topological spaces is continuous if and only if $f^{-1}(C)$ is closed in X for every closed subset C of Y .*

Proof. $X \setminus f^{-1}(C) = f^{-1}(Y \setminus C)$. □

Our intuition about spaces comes from metric spaces. So the following may seem obvious.

Lemma A.1.17. *Let X be a Hausdorff space. Then every point $x \in X$ is closed in X (i.e., the one-point subset $\{x\}$ is a closed subset of X).*

Proof. For $y \in X \setminus \{x\}$, we can find open sets U_y and V_y with $x \in U_y$ and $y \in V_y$ such that $U_y \cap V_y = \emptyset$. In particular, $y \in V_y \subset (X \setminus \{x\})$, so

$$X \setminus \{x\} = \bigcup_{y \in (X \setminus \{x\})} V_y$$

is a union of open sets, and hence is open. \square

Thus, Hausdorff spaces are T_1 :

Definition A.1.18. A topological space is T_1 if each of its points is closed.

There do exist spaces that are not T_1 . Indeed they can be quotient spaces of metric spaces.

Recall from Proposition 2.8.10 that a subset $H \subset \mathbb{R}^n$ is an affine subspace if and only if there is a linear map $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ for some m and an element $y \in \mathbb{R}^m$ such that $H = f^{-1}(y)$. In particular, because points are closed in \mathbb{R}^m and linear maps are continuous, we obtain the following.

Lemma A.1.19. *An affine subspace $H \subset \mathbb{R}^n$ is closed in \mathbb{R}^n .*

A.2. Subspace topology. A subset Y of a topological space X inherits a topology from X .

Definition A.2.1. Let X be a space and let $Y \subset X$. In the subspace topology on Y , a subset is open if it is the intersection of Y with an open subset of X .

In particular, if X is a metric space and $Y \subset X$ we can then ask if the topology Y inherits as a subspace coincides with the topology induced by the metric of X . The answer is affirmative:

Lemma A.2.2. *Let X be a metric space and let $Y \subset X$. Then the subspace topology and the metric topology on Y coincide.*

Proof. Let $y \in Y$. Then the ϵ ball about y in the metric space Y is the intersection of Y with the ϵ ball about y in the metric space X . \square

The subspace topology gives the subspace Y the minimal number of open sets for the inclusion map $Y \subset X$ to be continuous.

Lemma A.2.3. *Let Y be a subset of the topological space X , and give it the subspace topology. Then the inclusion $i : Y \subset X$ is continuous.*

Proof. For an open set $U \subset X$, $i^{-1}(U) = U \cap Y$ is open by the definition of the subspace topology. \square

There is a good recognition principle for closed sets in subspaces.

Lemma A.2.4. *Let X be a space and let $Y \subset X$. Then the closed subsets in the subspace topology on Y are precisely the sets $Y \cap C$ with C closed in X .*

Proof. An open subset of Y in the subspace topology has the form $Y \cap U$ with U open in X . But $Y \cap (X \setminus U) = Y \setminus (Y \cap U)$. \square

Heredity properties of subspaces are important.

Lemma A.2.5.

- (1) *Let U be open in X . Then the open sets in the subspace topology on U are exactly the open sets in X contained in U . In particular, an open subspace of an open subspace is open.*
- (2) *Let C be closed in X . Then the closed sets in the subspace topology on C are exactly the closed sets in X contained in C . In particular, a closed subspace of a closed subspace is closed.*

Proof. (1) An open subspace of U (in the subspace topology) is a set of the form $V \cap U$, where V is open in X . But intersections of open sets are open, so $V \cap U$ is open in X .

The proof of (2) is analogous, using Lemma A.2.4. \square

A.3. Quotient topology. It is fair to say that topology grew out of the study of manifolds and was developed to prove theorems about manifolds. Under minor hypotheses, manifolds are subspaces of \mathbb{R}^n and are therefore metric spaces. But some constructions of manifolds, e.g., quotients of appropriate group actions, are not naturally subsets of \mathbb{R}^n nor are they obviously metric. And while the tangent bundle of a smoothly embedded submanifold of \mathbb{R}^n is obviously a subspace of \mathbb{R}^{2n} , it is useful to have a model for the tangent bundle independent of the embedding.

Let us establish some notation.

Notation A.3.1. The unit square is

$$I^2 = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^2 : x, y \in [0, 1] \right\}.$$

Its interior is

$$\text{Int}(I^2) = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^2 : x, y \in (0, 1) \right\}.$$

Its boundary is

$$\partial I^2 = I^2 - \text{Int}(I^2) = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \in I^2 : \text{at least one of } x \text{ and } y \text{ is in } \{0, 1\} \right\}.$$

The idea behind a quotient space is that we wish to make identifications between certain points on the space X (i.e., view them as the same point in a new space).

Example A.3.2. The Klein bottle, K , is the space obtained from the unit square I^2 by making certain identifications on the boundary. (We shall give it the quotient topology as described in Definition A.3.4, below.) We identify opposite edges of the square with a flip in one direction but not the other. Thus, we identify the top edge with the bottom edge so the single arrow heads coincide in Figure A.3.1, and identify the left edge with the right edge so the double arrow heads there coincide.

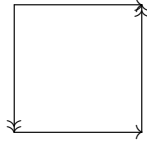


FIGURE A.3.1. Identifications on the unit square to create the Klein bottle.

So there is a twist in the identification of the vertical edges, but not for the horizontal ones. Note that if you only make the prescribed identifications on the vertical edges (but not the horizontal ones) you get a Möbius band, while, if you only make the prescribed identifications on the horizontal edges, you get a cylinder.

Though it might not be immediately apparent, the result is a 2-dimensional manifold, or surface. It may be easier to see the surface structure if you first identify the top and bottom edges, obtaining a cylinder. If we then make the identification of the two boundary circles of the cylinder, the gluing is by a homeomorphism between the circles, so near the glued circle K looks like the product of the circle with an open interval. What's deceptive is that the result cannot be embedded in \mathbb{R}^3 . You have to add a dimension to avoid the surface passing through itself. when you glue the opposite ends of the cylinder.

A similar construction gives us the standard torus.

Example A.3.3. The 2-torus \mathbb{T}^2 is obtained from I^2 by identifying opposite edges by translations: the bottom edge is identified to the top edge by τ_{e_2} , while the left edge is identified with the right edge by τ_{e_1} . The identifications are as given in Figure A.3.2.

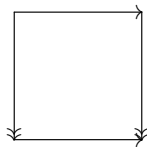


FIGURE A.3.2. Identifications on the unit square to create the 2-torus \mathbb{T}^2 .

We've described the Klein bottle and the 2-torus as result of making identifications on I^2 but have not yet described the induced topology on them. What we have so far is a set of gluing instructions, producing sets K and \mathbb{T}^2 together with functions $f : I^2 \rightarrow K$ and $g : I^2 \rightarrow \mathbb{T}^2$ taking each point in I^2 its image after gluing. The topology we shall place on K and \mathbb{T}^2 is the quotient topology, defined in Definition A.3.4(2).

Definition A.3.4.

(1) Let X and Y be spaces and let $f : X \rightarrow Y$ be continuous. We say Y has the quotient topology induced by f if:

(a) f is surjective.

(b) A subset $U \subset Y$ is open if and only if $f^{-1}(U)$ is open in X .

In this case, we call f a quotient map or identification map.

(2) More generally, let X be a space and Y a set. Let $f : X \rightarrow Y$ be a surjective function. Then we can impose a topology on Y by declaring $U \subset Y$ to be open if $f^{-1}(U)$ is open in X . Again, we call this the quotient topology induced by f and refer to f a quotient map or identification map.

Since $U \mapsto f^{-1}(U)$ preserves arbitrary unions and intersections of subsets (whereas $V \mapsto f(V)$ does not), this is easily seen to be a topology on Y .

The quotient topology has an important property.

Proposition A.3.5. *Let $f : X \rightarrow Y$ be a quotient map and let $g : Y \rightarrow Z$ be any function, where Z is a space. Then g is continuous if and only if $g \circ f$ is continuous.*

Proof. Suppose $g \circ f$ is continuous and $U \subset Z$ is open. Then $f^{-1}(g^{-1}(U))$ is open in X , so $g^{-1}(U)$ is open in Y . \square

Remark A.3.6. The quotient topology induced by a function $f : X \rightarrow Y$ can be somewhat bizarre if the function f is badly behaved. For instance, we could take $X = I = [0, 1]$ and take Y to consist of exactly 3 points, say $Y = \{0, z, 1\}$. If we then define $f : X \rightarrow Y$ by

$$f(t) = \begin{cases} 0 & \text{if } t = 0 \\ z & \text{if } t \in (0, 1) \\ 1 & \text{if } t = 1, \end{cases}$$

then the point z is open in the quotient topology on Y , but is not closed. We shall avoid examples like this, but the reader should be aware they exist.

Quotient topologies can be described in terms of closed sets as well as open ones.

Lemma A.3.7. *Let $f : X \rightarrow Y$ be a quotient map. Then a subset $C \subset Y$ is closed if and only if $f^{-1}(C)$ is closed in X .*

Conversely, if $f : X \rightarrow Y$ is surjective and we topologize Y by declaring a $C \subset Y$ to be closed if and only if $f^{-1}(C)$ is closed, then f is a quotient map with this topology.

Proof. The correspondence $A \mapsto f^{-1}(A)$ respects complements. \square

It is useful to be able to recognize when a map is a quotient map.

Definition A.3.8. An open map is a continuous map $f : X \rightarrow Y$ such that $f(U)$ is open in Y whenever U is open in X .

A closed map is a continuous map $f : X \rightarrow Y$ such that $f(C)$ is closed in Y whenever C is closed in X .

Lemma A.3.9. *A surjective open map is a quotient map. So is a surjective closed map.*

Proof. If $f : X \rightarrow Y$ is surjective, then $A = f(f^{-1}(A))$ for any $A \subset Y$. \square

We give some examples of spaces constructed as quotients via Definition A.3.4(2).

Example A.3.10. We can generalize \mathbb{T}^2 to n dimensions. We consider the n -cube

$$I^n = \{x_1e_1 + \cdots + x_ne_n : x_i \in I \text{ for all } i\}.$$

The interior of I^n is

$$\text{Int}(I^n) = \{x_1e_1 + \cdots + x_ne_n : x_i \in (0, 1) \text{ for all } i\},$$

and the boundary ∂I^n is the set of points at least one of whose coordinates is in $\{0, 1\}$. We have inclusions $\iota_i^\epsilon : I^{n-1} \rightarrow \partial I^n$, $\epsilon = 0$ or 1 , given by

$$\iota_i^\epsilon \left(\begin{bmatrix} x_1 \\ \vdots \\ x_{n-1} \end{bmatrix} \right) = \begin{bmatrix} x_1 \\ \vdots \\ x_{i-1} \\ \epsilon \\ x_i \\ \vdots \\ x_{n-1} \end{bmatrix},$$

inserting $\epsilon = 0$ or 1 into the i th slot. We write $\partial_i^\epsilon(I^n) = \text{Im}(\iota_i^\epsilon)$, a face of ∂I^n . The boundary of I^n is the union of these faces.

We define the n -torus \mathbb{T}^n to be obtained from I^n by identifying $\iota_i^0(x)$ with $\iota_i^1(x)$ for each $x \in I^{n-1}$ and $i \in \{1, \dots, n\}$. Thus, a point in $\text{Int}(I^n)$ is not identified to any other point in I^n , while a boundary point that is contained in exactly k faces is part of a group of 2^k elements identified with one another. This generalizes the identifications made to create \mathbb{T}^2 .

The process of making identifications on a space X is often most easily described by specifying an equivalence relation on X , as that gives a nice clean description of the set Y and the map $f : X \rightarrow Y$. In particular, we shall take Y to be the set of equivalence classes under the relation. If \sim is an equivalence relation on X and if $\pi : X \rightarrow X/\sim$ is the canonical map to its set of equivalence classes, then the equivalence classes of \sim are the point inverses of π by Lemma 0.0.7. This description relates nicely to the general surjective map $f : X \rightarrow Y$ of Definition A.3.4.

Example A.3.11. If $f : X \rightarrow Y$ is a function then there is an equivalence relation \sim on X defined by $x \sim y$ if $f(x) = f(y)$. We refer to \sim as the equivalence relation induced by f .

Indeed, if \approx is an equivalence relation on Y we may define an equivalence relation on X by setting $x \sim y$ if $f(x) \approx f(y)$.

The canonical map $\pi : X \rightarrow X/\sim$ has an important universal property.

Proposition A.3.12. *Let $f : X \rightarrow Y$ and let \sim be an equivalence relation on X . Then there is a function $\bar{f} : X/\sim \rightarrow Y$ making the diagram*

$$(A.3.1) \quad \begin{array}{ccc} X & \xrightarrow{f} & Y \\ & \searrow \pi & \nearrow \bar{f} \\ & X/\sim & \end{array}$$

commute if and only if $x \sim y$ implies $f(x) = f(y)$. Such an \bar{f} , if it exists, is unique, and is given by $\bar{f}([x]) = f(x)$. In this case, the image of \bar{f} is the image of f , and \bar{f} is one-to-one if and only if \sim coincides with the equivalence relation induced by f , i.e., if and only if

$$x \sim y \quad \Leftrightarrow \quad f(x) = f(y).$$

Proof. The uniqueness and formula for \bar{f} , if it exists, are forced by π being onto. The formula then shows \bar{f} exists if and only if $f(x) = f(y)$ implies $[x] = [y]$, but that in turn is equivalent to $x \sim y$. Finally, \bar{f} is one-to-one if and only if $f(x) = f(y)$ is equivalent to $[x] = [y]$. \square

In particular, Proposition A.3.12 determines all functions out of X/\sim , as, if $g : X/\sim \rightarrow Y$, then $g \circ \pi : X \rightarrow Y$ is a function, and $g = \overline{g \circ \pi}$ by uniqueness.

Corollary A.3.13. *Let $f : X \rightarrow Y$ be surjective and let \sim be the equivalence relation induced by f . Then $\bar{f} : X/\sim \rightarrow Y$ is bijective, and we can identify \bar{f} with π .*

Of course, we can ask how the results above relate to topologies. For any equivalence relation \sim on X the quotient topology induced by π is a natural topology to put on X/\sim .

Proposition A.3.14. *Let $f : X \rightarrow Y$ be a continuous map. Let \sim be an equivalence relation on X such that $x \sim y \Rightarrow f(x) = f(y)$. Regard X/\sim as a topological space via the quotient topology induced by $\pi : X \rightarrow X/\sim$. Then the induced map $\bar{f} : X/\sim \rightarrow Y$ is continuous.*

Suppose now that f is onto and that \sim is the equivalence relation induced by f . Then \bar{f} is bijective. If, in addition, Y has the quotient topology induced by f , then \bar{f} is a homeomorphism.

Proof. Let $U \subset Y$ be open. Then $\pi^{-1}(\bar{f}^{-1}(U)) = f^{-1}(U)$ is open in X since f is continuous. So $\bar{f}^{-1}(U)$ is open in X/\sim as π is a quotient map. Thus, \bar{f} is continuous if f is.

If f is onto and \sim is the equivalence relation induced by f , then \bar{f} is bijective. By Corollary A.3.13. Assume also that Y has the quotient topology induced by f . It suffices to show that $U \subset Y$ is open if and only if $\bar{f}^{-1}(U)$ is open in X/\sim . But $\pi^{-1}(\bar{f}^{-1}(U)) = f^{-1}(U)$, so the result follows since both f and π are quotient maps. \square

The point of using equivalence relations instead of simply using functions is that the equivalence relation can make it easier to define $f : X \rightarrow Y$ from the data present. The most pertinent example is from a group action.

A.4. Group actions and orbit spaces.

Definition A.4.1. An action of a group G on a space X is a function

$$\begin{aligned} G \times X &\rightarrow X \\ (g, x) &\mapsto gx \end{aligned}$$

such that:

- (1) $g(hx) = (gh)x$ for all $g, h \in G$ and $x \in X$.
- (2) $1 \cdot x = x$ for all $x \in X$, where 1 is the identity element of G .
- (3) For each $g \in G$ the map $\mu_g : X \rightarrow X$ given by $\mu_g(x) = gx$ is continuous.

A space X together with an action of G is called a G -space. If X and Y are G -spaces, a G -map $f : X \rightarrow Y$ is a continuous function with the property that $f(gx) = gf(x)$ for all $g \in G$ and $x \in X$. A G -homeomorphism is a G -map $f : X \rightarrow Y$ that is also a homeomorphism.

If X is a smooth manifold and each μ_g is smooth, we say G acts smoothly on X or that the action $G \times X \rightarrow X$ is smooth. If G acts smoothly on both X and Y , a G -diffeomorphism $f : X \rightarrow Y$ is a G -map that is also a diffeomorphism.

As the reader may easily verify, if $f : X \rightarrow Y$ is a G -homeomorphism, then the inverse function f^{-1} is a G -map, and hence also a G -homeomorphism. Similarly, the inverse function to a G -diffeomorphism is a G -diffeomorphism.

Lemma A.4.2. *Let X be a G -space and $g \in G$. Then $\mu_g : X \rightarrow X$ is a homeomorphism.*

Proof. By properties (1) and (2) of the definition of G -space, μ_g is a bijection whose inverse function is $\mu_{g^{-1}}$. Since both μ_g and $\mu_{g^{-1}}$ are continuous, the result follows. \square

There is an important equivalence relation defined on a G -space.

Lemma A.4.3. *Let X be a G -space. Then there is an equivalence relation on X defined by setting $x \sim gx$ for all $x \in X$ and $g \in G$.*

Proof. Reflexivity comes from $1x = x$. Symmetry comes from $g^{-1}gx = x$. Transitivity comes from $g(hx) = (gh)x$. \square

Definition A.4.4. Let X be a G -space and let \sim be the equivalence relation of Lemma A.4.3. We refer to its equivalence classes as the orbits of the action and write $Gx = \{gx : g \in G\} = [x]$, the equivalence class of x . We write X/G for X/\sim and refer to the canonical map $\pi : X \rightarrow X/G$ as the orbit map. $\pi(x) = Gx$ for all $x \in X$. We give X/G the quotient topology induced by π .

We can use the orbit space to give a different construction for the Klein bottle. This new construction is useful in showing the Klein bottle is a smooth manifold, in part because the action of a wallpaper group on \mathbb{R}^2 is smooth.

Example A.4.5. We describe a wallpaper group \mathcal{K} as follows. Let ℓ be the line $y = \frac{1}{2}$ and let γ be the glide reflection $\gamma = \tau_{e_1}\sigma_\ell$ with $e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, the first canonical basis vector. Note that γ makes the desired identification of the left and right edges of the unit square, while τ_{e_2} makes the desired identification on the top and bottom edges, $e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Define the Klein group by

$$(A.4.1) \quad \mathcal{K} = \{\tau_{e_2}^m \gamma^n : m, n \in \mathbb{Z}\} \subset \mathcal{I}_2.$$

\mathcal{K} is a group because $\gamma\tau_{e_2}\gamma^{-1} = \sigma_\ell\tau_{e_2}\sigma_\ell^{-1} = \tau_{e_2}^{-1}$. So

$$\begin{aligned} \tau_{e_2}^m \gamma^n \tau_{e_2}^r \gamma^s &= \tau_{e_2}^{m+(-1)^n r} \gamma^{n+s}, \\ (\tau_{e_2}^m \gamma^n)^{-1} &= \tau_{e_2}^{(-1)^{n+1} m} \gamma^{-n}. \end{aligned}$$

We shall provide a homeomorphism from the Klein bottle, K , to \mathbb{R}^2/\mathcal{K} . We first provide the map and show it is a continuous bijection.

To obtain this, we first show I^2 is a fundamental region for \mathcal{K} in the sense that:

- (1) $\mathbb{R}^2 = \bigcup_{g \in \mathcal{K}} g(I^2)$.
- (2) For all $g \in \mathcal{K}$, $I^2 \cap g(I^2) \subset \partial I^2$.

To see this, note that each element of \mathcal{K} carries I^2 onto a unique tile of the form

$$\left\{ \begin{bmatrix} x \\ y \end{bmatrix} : x \in [i, i + 1], y \in [j, j + 1] \right\}$$

for $i, j \in \mathbb{Z}$. In fact, for $g = \tau_{e_2}^m \gamma^n$, $i = n$ and $j = m$. Thus, if $g \notin \{\gamma^{\pm 1}, \tau_{e_2}^{\pm 1}\}$, $I^2 \cap g(I^2) = \emptyset$, while if $g \in \{\gamma^{\pm 1}, \tau_{e_2}^{\pm 1}\}$, $I^2 \cap g(I^2)$ is one of the edges of I^2 .

The action of \mathcal{K} on the plane is *free* in the sense that no point is fixed by any nonidentity element of \mathcal{K} .

Now consider the composite

$$(A.4.2) \quad \begin{array}{ccc} I^2 & \xrightarrow{i} & \mathbb{R}^2 \xrightarrow{\pi} \mathbb{R}^2/\mathcal{K}, \\ & \searrow & \nearrow \\ & & j \end{array}$$

with i the standard inclusion of I^2 in \mathbb{R}^2 . We claim that j is onto and that the equivalence relation on I^2 induced by j produces precisely the identifications used to define K . Indeed, by (1), every element $x \in \mathbb{R}^2$ has the form $x = gy$ for $y \in I^2$ and $g \in \mathcal{K}$. Thus, j is onto.

By (2), if $j(y) = j(z)$ for $y \neq z$, then y and z must both lie on ∂I^2 . Moreover, an examination of the effects of $\{\gamma^{\pm 1}, \tau_{e_2}^{\pm 1}\}$ shows that y and z must be related by one of the identifications specified in the definition of K in Example A.3.2. We obtain a commutative diagram

$$(A.4.3) \quad \begin{array}{ccc} I^2 & \xrightarrow{i} & \mathbb{R}^2 \\ \pi \downarrow & \searrow j & \downarrow \pi \\ K & \xrightarrow{\bar{j}} & \mathbb{R}^2/\mathcal{K} \end{array}$$

where the maps π are the quotient maps. By Proposition A.3.14, \bar{j} is a continuous bijection.

Using basic point-set topology, it is easy to show \bar{j} is a homeomorphism. One need only show that I^2 is compact and that \mathbb{R}^2/\mathcal{K} is Hausdorff. We shall use more elementary arguments involving the concept of a basis for a topology. Meanwhile, we have some more examples.

Example A.4.6. The standard translation lattice

$$\mathcal{T}_{\Lambda_{\mathcal{E}}} = \langle \tau_{e_1}, \tau_{e_2} \rangle = \{ \tau_{ke_1 + \ell e_2} : k, \ell \in \mathbb{Z} \}$$

is another wallpaper group, giving one of the realizations of \mathcal{W}_1 . The orbit space $\mathbb{R}^2/\mathcal{T}_{\Lambda_{\mathcal{E}}}$ gives an alternative model for the 2-torus \mathbb{T}^2 . Indeed, the unit square I^2 gives a fundamental region for the action of $\mathcal{T}_{\Lambda_{\mathcal{E}}}$ on \mathbb{R}^2 , satisfying (1) and (2) above.

The same argument as above provides a commutative diagram

$$(A.4.4) \quad \begin{array}{ccc} I^2 & \xrightarrow{i} & \mathbb{R}^2 \\ \pi \downarrow & \searrow j & \downarrow \pi \\ \mathbb{T}^2 & \xrightarrow{\bar{j}} & \mathbb{R}^2/\mathcal{T}_{\Lambda_{\mathcal{E}}} \end{array}$$

with \bar{j} a continuous bijection.

Example A.4.7. The preceding example generalizes to the n -torus \mathbb{T}^n . Here, we use the standard n -dimensional translation lattice

$$\mathcal{T}_{\Lambda_{\mathcal{E}}} = \langle \tau_{e_1}, \dots, \tau_{e_n} \rangle = \{ \tau_{k_1 e_1 + \dots + k_n e_n} : k_1, \dots, k_n \in \mathbb{Z} \}$$

on \mathbb{R}^n . Lemma 6.12.6 shows that I^n is a fundamental region for this action, and once again, the identifications on the boundary of I^n induced by this action are precisely the ones we used to define \mathbb{T}^n . We obtain a commutative

diagram

$$(A.4.5) \quad \begin{array}{ccc} I^n & \xrightarrow{i} & \mathbb{R}^n \\ \pi \downarrow & \searrow j & \downarrow \pi \\ \mathbb{T}^n & \xrightarrow{\bar{j}} & \mathbb{R}^n / \mathcal{T}_{\Lambda_\varepsilon} \end{array}$$

where the vertical maps are the quotient maps and \bar{j} is a continuous bijection.

Example A.4.8. Note that the $(n + 1) \times (n + 1)$ matrix $-I_{n+1}$ induces a linear isometry of \mathbb{R}^{n+1} and hence preserves the n -sphere, inducing a smooth map (Corollary 8.4.15) $\alpha : \mathbb{S}^n \rightarrow \mathbb{S}^n$, $\alpha(u) = -u$ for $u \in \mathbb{S}^n$. Let $G = \{\text{id}, \alpha\}$, the group with two elements. We define the n -dimensional real projective space by

$$\mathbb{RP}^n = \mathbb{S}^n / G.$$

We shall show that \mathbb{RP}^n has a natural Riemannian metric that realizes what is called projective geometry.

A.5. Basis for a topology.

Definition A.5.1. A basis \mathcal{B} for the topology of a space X is a collection of open sets \mathcal{B} such that for each open set U of X and each $x \in U$, there exists $V \in \mathcal{B}$ with $x \in V \subset U$.

Example A.5.2. For a metric space X the collection

$$\mathcal{B} = \bigcup_{\substack{\epsilon > 0 \\ x \in X}} B_\epsilon(x)$$

is a basis for the topology of X by the very definition of that topology. We can get a smaller basis by restricting to

$$(A.5.1) \quad \mathcal{B}' = \bigcup_{\substack{n > 0 \\ x \in X}} B_{\frac{1}{n}}(x),$$

as, for each $\epsilon > 0$ there exists n with $\frac{1}{n} < \epsilon$.

Definition A.5.3. A space X is *first-countable* if each $x \in X$ has what's called a countable neighborhood base, i.e., a countable set \mathcal{B}_x of open subsets containing x , such that every open subset containing x contains an element of \mathcal{B}_x . In particular, (A.5.1) shows every metric space is first-countable.

A subset $Y \subset X$ is *dense in X* if every open subset of X contains an element of Y . Decimal approximation shows that the set \mathbb{Q} of rational numbers is dense in \mathbb{R} . Similarly, \mathbb{Q}^n is dense in \mathbb{R}^n .

A space X is *separable* if it has a countable dense subset. A standard argument shows a finite product of countable sets is countable, so \mathbb{Q}^n is countable, and hence \mathbb{R}^n is separable.

A space X is *second-countable* if it has a countable basis \mathcal{B} for its topology.

Lemma A.5.4. *Every subspace of a separable metric space is second-countable. Indeed, we have the following:*

- (1) *A separable metric space X is second-countable.*
- (2) *Every subspace of a second-countable space X is second-countable.*

Thus, every subspace of \mathbb{R}^n is second-countable.

Proof. (1) Let $Y \subset X$ be countable and dense. We claim that

$$\mathcal{B} = \bigcup_{\substack{y \in Y \\ n > 0}} B_{\frac{1}{n}}(y)$$

is a basis for X . To see this, let U be an open set containing X . Then $B_{\frac{1}{n}}(x) \subset U$ for some $n > 0$. But then $B_{\frac{1}{2n}}(x) \subset U$ as well. Since Y is dense and $B_{\frac{1}{2n}}(x)$ is open, there exists $y \in B_{\frac{1}{2n}}(x) \cap Y$. But then

$$x \in B_{\frac{1}{2n}}(y) \subset B_{\frac{1}{n}}(x) \subset U$$

by the symmetry of the metric and the triangle inequality.

(2) if $\mathcal{B} = \{U_i : i > 0\}$ is a basis for X , then $\{U_i \cap Y : i > 0\}$ is a basis for $Y \subset X$. \square

The definition of topological manifold given in Definition 8.3.4 required that M be contained in \mathbb{R}^n for some n . This makes M a metric space, which we used to discuss continuity issues. By Lemma A.5.4 is also makes M second-countable. Since metric spaces are also Hausdorff, M satisfies the following more general definition.

Definition A.5.5 (Topological manifold: general definition). A topological n -manifold M is a second-countable Hausdorff space with the property that every point $x \in M$ has a neighborhood homeomorphic to an open subset of \mathbb{R}^n .

Topological manifolds under this definition turn out to be metrizable. We shall not need this here. We are most interested in the smooth case. The general definition of smooth manifold is precisely the one given in Section 8.4. We do not use the metric there, and need only assume M is a topological manifold in the sense of Definition A.5.5. This is useful in discussing the smooth structure on the Klein bottle and other orbit spaces of properly discontinuous smooth actions, as there is no obvious metric on the orbit space, nor any obvious embedding in Euclidean space. But in fact, this seemingly more general definition of smooth manifold does embed smoothly in Euclidean space:

Theorem A.5.6 (Whitney embedding theorem. See [13, Theorem 6.19]). *Every smooth n -manifold embeds smoothly into \mathbb{R}^{2n} .*

We now give a couple more general results and apply them to the Klein bottle and the torus.

Definition A.5.7. A map $f : X \rightarrow Y$ is open if $f(U)$ is open in Y for each open subset U of X .

Note that a continuous bijection $f : X \rightarrow Y$ is a homeomorphism if and only if it is an open map.

Lemma A.5.8. Let \mathcal{B} be a basis for the topology of X . Then a subset U is open if and only if it is the union of some family of basis elements.

If $f : X \rightarrow Y$ is a map, then f is an open map if and only if $f(V)$ is open in Y for each basis element V of X .

Proof. Basis elements are open, and an arbitrary union of open sets is open, so any union of basis elements is open.

Conversely if U is open, then any $x \in U$ is contained in a basis element contained in U , so U is the union of all basis elements contained in it.

The statement about open mappings follows as

$$f\left(\bigcup_{V \in S} V\right) = \bigcup_{V \in S} f(V)$$

for any set S of subsets of X . □

We shall now construct a bases for the topology of the Klein bottle and the 2-torus and use them to show the continuous bijections \bar{j} of Examples A.4.5 and A.4.6 are open maps, and hence homeomorphisms.

The following concept is useful.

Definition A.5.9. Let $f : X \rightarrow Y$. A subset $U \subset X$ is saturated under f if $U = f^{-1}(f(U))$. In particular, if f is an quotient map then $U \mapsto f^{-1}(U)$ gives a one-to-one correspondence from the open sets of Y to the saturated open sets of X .

We now specify a basis for the Klein bottle K .

Lemma A.5.10. Let $\pi : I^2 \rightarrow K$ be the canonical map. Then there is a basis \mathcal{B} for K consisting of the images under π of the following collections of π -saturated subsets of I^2 :

- (1) all ϵ -balls contained entirely in the interior of I^2 ;
- (2) the unions $B_\epsilon([0_t], I^2) \cup B_\epsilon([1_{1-t}], I^2)$ for all $\epsilon \leq \min(t, 1-t)$ and all $t \in (0, 1)$;
- (3) the unions $B_\epsilon([t_0], I^2) \cup B_\epsilon([t_1], I^2)$ for all $\epsilon \leq \min(t, 1-t)$ and all $t \in (0, 1)$;
- (4) the unions $B_\epsilon([0_0], I^2) \cup B_\epsilon([0_1], I^2) \cup B_\epsilon([1_0], I^2) \cup B_\epsilon([1_1], I^2)$ for all $\epsilon \leq \frac{1}{2}$.

Proof. The displayed sets are obviously open and saturated. Moreover, any open set containing $\pi^{-1}(x)$ for some $x \in K$ must contain one of the sets in (1)–(4). □

The following is the key step in showing $\bar{j} : K \rightarrow \mathbb{R}^2/\mathcal{K}$ is open, and hence a homeomorphism.

Proposition A.5.11. *The canonical map $\pi : \mathbb{R}^2 \rightarrow \mathbb{R}^2/\mathcal{K}$ is an open map.*

Proof. This is an immediate consequence of Theorem A.7.1 below. \square

Corollary A.5.12. *The map $\bar{j} : K \rightarrow \mathbb{R}^2/\mathcal{K}$ of Example A.4.5 is a homeomorphism.*

Proof. To avoid confusion, write $\pi' : I^2 \rightarrow K$ and $\pi : \mathbb{R}^2 \rightarrow \mathbb{R}^2/\mathcal{K}$ for the respective canonical maps. It suffices to show that if V is in the basis for K given in Lemma A.5.10, then $\bar{j}(V)$ is open in \mathbb{R}^2/\mathcal{K} . In case (1), that is immediate from π being an open map. In case (2), it follows as

$$\begin{aligned} \bar{j}\pi'(B_\epsilon([\tfrac{0}{t}], I^2) \cup B_\epsilon([\tfrac{1}{1-t}], I^2)) \\ = \pi(B_\epsilon([\tfrac{1}{1-t}], I^2) \cup \gamma(B_\epsilon([\tfrac{0}{t}], I^2))) \\ = \pi(B_\epsilon([\tfrac{1}{1-t}], \mathbb{R}^2)), \end{aligned}$$

and we again use that π is open. Case (3) is similar, using τ_{e_2} in place of γ , and getting the open set $\pi(B_\epsilon([\tfrac{t}{1}], \mathbb{R}^2))$. In case (4) we apply isometries to three of the four pieces to obtain $\pi(B_\epsilon([\tfrac{1}{1}], \mathbb{R}^2))$. \square

The argument for the following is almost identical.

Corollary A.5.13. *The map $\bar{j} : \mathbb{T}^2 \rightarrow \mathbb{R}^2/\mathcal{T}_{\Lambda_\epsilon}$ of Example A.4.6 is a homeomorphism.*

Proof. The only changes needed are to replace (2) with

$$(2') \text{ the unions } B_\epsilon([\tfrac{0}{t}], I^2) \cup B_\epsilon([\tfrac{1}{t}], I^2) \text{ for all } \epsilon \leq \min(t, 1-t) \text{ and all } t \in (0, 1)$$

and replace γ with τ_{e_1} . \square

The proof of the following generalization is left to the reader.

Corollary A.5.14. *The map $\bar{j} : \mathbb{T}^n \rightarrow \mathbb{R}^n/\mathcal{T}_{\Lambda_\epsilon}$ of Example A.4.7 is a homeomorphism.*

A.6. Properly discontinuous actions. We can in fact prove something much stronger than Proposition A.5.11, and in process show the Klein bottle is a smooth manifold.

Definition A.6.1. An action of G on X is free and properly discontinuous if for each $x \in X$ there is an open set U containing x such that

$$(A.6.1) \quad g(U) \cap U = \emptyset \quad \text{for all } 1 \neq g \in G.$$

Lemma A.6.2. *The action of the Klein group \mathcal{K} on \mathbb{R}^2 is free and properly discontinuous.*

Proof. Let U be any region of the form

$$\left\{ \begin{bmatrix} x \\ y \end{bmatrix} : x \in (x_0, x_0 + 1), y \in (y_0, y_0 + 1) \right\}.$$

Since γ^n moves points by $|n|$ units in the x -direction and $\tau_{e_2}^m$ moves points by $|m|$ units in the y -direction while keeping the x -coordinate fixed, $g(U) \cap U = \emptyset$ for each nonidentity element $g \in \mathcal{K}$. \square

Similarly, we have the following.

Lemma A.6.3. *The action of the torus group $\mathcal{T}_{\Lambda_\varepsilon}$ on \mathbb{R}^2 is free and properly discontinuous.*

In fact, this generalizes to the n -torus.

Lemma A.6.4. *The action of the standard n -dimensional lattice $\mathcal{T}_{\Lambda_\varepsilon}$ on \mathbb{R}^n is free and properly discontinuous.*

Proof. Here, we use the sets

$$U_y = \{x_1 e_1 + \cdots + x_n e_n : x_i \in (y_i, y_i + 1) \text{ for } i = 1, \dots, n\}$$

for $y = y_1 e_1 + \cdots + y_n e_n$ arbitrary. \square

The action on \mathbb{S}^n whose orbit space is $\mathbb{R}\mathbb{P}^n$ is also properly discontinuous.

Lemma A.6.5. *Let $\alpha : \mathbb{S}^n \rightarrow \mathbb{S}^n$ be given by $\alpha(u) = -u$ for all $u \in \mathbb{S}^n$. Then $G = \{\text{id}, \alpha\}$ acts freely and properly discontinuously on \mathbb{S}^n .*

Proof. Let $u \in \mathbb{S}^n$. Then $u \in U = \{v \in \mathbb{S}^n : \langle u, v \rangle > 0\}$. Since $\alpha(U) = \{v \in \mathbb{S}^n : \langle u, v \rangle < 0\}$, the result follows. \square

We wish to show next that if G acts freely and properly discontinuously on X and if U satisfies (A.6.1), then the saturation $\pi^{-1}\pi(U)$ is homeomorphic to $G \times U$. But, of course, we have not yet discussed the topology on a product nor discussed the discrete topology, which is the appropriate topology for G , here.

Definition A.6.6. The discrete topology on a set S is the one in which every subset of S is open. Since arbitrary unions of open sets are open, this is equivalent to saying that every point is open. Thus, the points of a discrete space form a basis for its topology.

A.6.1. Product topology.

Definition A.6.7. Let X and Y be spaces. The product topology on $X \times Y$ is the one specified by the basis

$$\{U \times V : U \subset X, V \subset Y \text{ are open}\}.$$

In other words, a subset $W \subset X \times Y$ is open if and only if for each $(x, y) \in W$ there are open sets U of X and V of Y with

$$(x, y) \in U \times V \subset W.$$

Lemma A.6.8. *Let \mathcal{B} be a basis of X and \mathcal{B}' a basis of Y . Then*

$$\{U \times V : U \in \mathcal{B}, V \in \mathcal{B}'\}$$

is a basis for the product topology of $X \times Y$.

Proof. If U and V are arbitrary open sets of X and Y , respectively, and if $(x, y) \in U \times V$, there exist $W \in \mathcal{B}$, $W' \in \mathcal{B}'$ with $x \in W \subset U$, $y \in W' \subset V$, so

$$(x, y) \in W \times W' \subset U \times V. \quad \square$$

Corollary A.6.9. *Let A be a subspace of X and B a subspace of Y . Then the product topology on $A \times B$ coincides with its subspace topology in $X \times Y$.*

Proof. Either topology has a basis given by the sets

$$(U \times V) \cap (A \times B) = (U \cap A) \times (V \cap B)$$

as U and V range over bases of X and Y , respectively. \square

Definition A.6.10. Let X and Y be metric spaces. Then the product metric on $X \times Y$ is given by

$$(A.6.2) \quad d((x, y), (z, w)) = \max(d(x, z), d(y, w)).$$

Lemma A.6.11. *Let X and Y be metric spaces. Then the product topology on $X \times Y$ is induced by the product metric (A.6.2). Moreover, with respect to this metric, we have*

$$(A.6.3) \quad B_\epsilon((x, y)) = B_\epsilon(x) \times B_\epsilon(y).$$

Proof. (A.6.3) is immediate from the definition of the metric, as

$$\max(d(x, z), d(y, w)) < \epsilon \quad \Leftrightarrow \quad d(x, z) < \epsilon \text{ and } d(y, w) < \epsilon.$$

The rest of the argument is similar to that for Lemma A.6.8. If $(x, y) \in U \times V$, we can find a single ϵ with $B_\epsilon(x) \subset U$ and $B_\epsilon(y) \subset V$, so the result follows. \square

Of course, we can identify $\mathbb{R}^n \times \mathbb{R}^m$ with \mathbb{R}^{n+m} via the map

$$(A.6.4) \quad f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^{n+m} \\ ((x_1, \dots, x_n), (y_1, \dots, y_m)) \mapsto (x_1, \dots, x_n, y_1, \dots, y_m).$$

Viewing $\mathbb{R}^n \times \mathbb{R}^m$ as a vector space in the usual way (direct sum) f is a linear isomorphism.

Let's take f as an identification and write (v, w) for an arbitrary element of \mathbb{R}^{n+m} , with $v \in \mathbb{R}^n$ and $w \in \mathbb{R}^m$. This gives us two different metrics on \mathbb{R}^{n+m} : the product metric d_\times and the usual metric d . It is useful to compare these two metrics. Let us study them in greater detail. Each of them comes from a norm.

Remark A.6.12. The metric obtained from a norm $\|\cdot\|$ sets

$$(A.6.5) \quad d(x, y) = \|y - x\|.$$

It is then immediate that

$$(A.6.6) \quad B_\epsilon(x) = \tau_x(B_\epsilon(0)) = \tau_x(\{z : \|z\| < \epsilon\}).$$

Recall that any inner product induces a norm: $\|x\| = \sqrt{\langle x, x \rangle}$. In our setting, \mathbb{R}^n and \mathbb{R}^m are orthogonal complements in \mathbb{R}^{n+m} , so the inner product is given by

$$(A.6.7) \quad \langle (v_1, w_1), (v_2, w_2) \rangle = \langle v_1, v_2 \rangle + \langle w_1, w_2 \rangle,$$

where the inner products on the right are the standard inner products in \mathbb{R}^n and \mathbb{R}^m , respectively (which can be viewed as the restriction of the inner product in \mathbb{R}^{n+m} to the subspaces \mathbb{R}^n and \mathbb{R}^m).

Thus, the standard norm is given by

$$(A.6.8) \quad \|(v, w)\| = \sqrt{\langle (v, w), (v, w) \rangle} = \sqrt{\langle v, v \rangle + \langle w, w \rangle} = \sqrt{\|v\|^2 + \|w\|^2},$$

where the norms on the right-hand side can be taken to be the standard norms in \mathbb{R}^n and \mathbb{R}^m , respectively, and coincide with the subspace norms coming from \mathbb{R}^{n+m} .

The product norm is given by

$$(A.6.9) \quad \|(v, w)\|_\times = \max(\|v\|, \|w\|),$$

where the norms on the right are the standard norms in \mathbb{R}^n and \mathbb{R}^m , respectively. It is then immediate that the product metric is induced by the product norm:

$$(A.6.10) \quad d_\times((v_1, w_1), (v_2, w_2)) = \|(v_1, w_1) - (v_2, w_2)\|_\times$$

for $(v_1, w_1), (v_2, w_2) \in \mathbb{R}^{m+n}$. Moreover, (A.6.8) shows that $\|(v, w)\| \geq \|v\|$ and $\|(v, w)\| \geq \|w\|$. Thus,

$$(A.6.11) \quad \|(v, w)\| \geq \|(v, w)\|_\times \quad \text{for all } (v, w) \in \mathbb{R}^{n+m}.$$

And this immediately implies that

$$(A.6.12) \quad B_\epsilon(0, d) \subset B_\epsilon(0, d_\times).$$

Here, the second argument indicates the metric in use.

On the other hand if $\|v\| < \epsilon$ and $\|w\| < \epsilon$, then

$$\|(v, w)\| = \sqrt{\|v\|^2 + \|w\|^2} < \sqrt{2\epsilon^2} = \sqrt{2}\epsilon,$$

so

$$(A.6.13) \quad B_\epsilon(0, d_\times) \subset B_{\sqrt{2}\epsilon}(0, d).$$

We obtain the following.

Proposition A.6.13. *The standard topology agrees with the product topology on $\mathbb{R}^n \times \mathbb{R}^m = \mathbb{R}^{n+m}$.*

Proof. A subset U is open in the standard topology if for each $x \in U$, we have $B_\epsilon(x, d) \subset U$ for some $\epsilon > 0$. But this contains $B_{\frac{\epsilon}{\sqrt{2}}}(x, d_\times)$, so it is open in the product topology as well. The converse is similar. \square

The product topology on an arbitrary pair of spaces has a useful universal property. First note that a function $f : Z \rightarrow X \times Y$ is specified by its coordinate functions $f_1 : Z \rightarrow X$ and $f_2 : Z \rightarrow Y$:

$$f(z) = (f_1(z), f_2(z)).$$

The universal property of the product topology is the following.

Lemma A.6.14. *Let Z , X and Y be spaces and let $f : Z \rightarrow X \times Y$ be a function. Then f is continuous if and only if f_1 and f_2 are continuous. We obtain a one-to-one correspondence*

$$(A.6.14) \quad \text{Map}(Z, X \times Y) \rightarrow \text{Map}(Z, X) \times \text{Map}(Z, Y) \\ f \mapsto (f_1, f_2),$$

where $\text{Map}(Z, W)$ denotes the set of continuous maps from Z to W . This may be restated as saying that for any pair of continuous maps $g : Z \rightarrow X$ and $h : Z \rightarrow Y$ there is a unique continuous map $f = (g, h) : Z \rightarrow X \times Y$ such that the following diagram commutes:

$$(A.6.15) \quad \begin{array}{ccccc} & & Z & & \\ & g \swarrow & \downarrow h & \searrow h & \\ X & \xleftarrow{\pi_1} & X \times Y & \xrightarrow{\pi_2} & Y. \end{array}$$

Here, $\pi_1((x, y)) = x$ and $\pi_2((x, y)) = y$.

Proof. For $U \subset X$ open, $\pi_1^{-1}(U) = U \times Y$ is a basis element for $X \times Y$, and hence is open, so π_1 is continuous. Similarly π_2 is continuous. So if $f : Z \rightarrow X \times Y$ is continuous, then the composites $f_1 = \pi_1 \circ f$ and $f_2 = \pi_2 \circ f$ are continuous.

Conversely, if f_1 and f_2 are continuous, let $U \times V$ be a basis element for $X \times Y$. Then

$$f^{-1}(U \times V) = f_1^{-1}(U) \cap f_2^{-1}(V)$$

is open. Since the inverse image of a union is the union of the inverse images, a map is continuous if and only if the inverse image of every element of some basis is open. \square

Corollary A.6.15. *Let X and Y be spaces and let $x \in X$. Let $\iota_x : Y \rightarrow X \times Y$ be given by $\iota_x(y) = (x, y)$ for $y \in Y$. Then ι_x is a homeomorphism of Y onto the subspace $x \times Y$ of $X \times Y$.*

Proof. ι_x is continuous by Lemma A.6.14. The inverse function of $\iota_x : Y \rightarrow x \times Y$ is given by the projection of $X \times Y$ onto Y . \square

A.6.2. Disjoint unions. The disjoint union of topological spaces has an important universal property called the *coproduct* in the language of category theory.

Definition A.6.16. We first define the disjoint union of sets. If

$$\{X_s : s \in S\}$$

are sets, we assume there is a set X containing each of the X_s . We then define their disjoint union by

$$\coprod_{s \in S} X_s = \{(s, x) \in S \times X : x \in X_s\}.$$

In particular, we identify X_s with the subset $\{s\} \times X_s \subset \coprod_{s \in S} X_s$.

Now let $\{X_s : s \in S\}$ be a set of spaces. Their disjoint union, $\coprod_{s \in S} X_s$, as spaces is defined to be the topology on their disjoint union as sets given by the basis

$$\{\{s\} \times U : U \text{ open in } X_s\}.$$

Identifying X_s with the subset $\{s\} \times X_s$ as above, we see that the subspace topology on X_s coincides with the original topology and that a subset

$$U \subset \coprod_{s \in S} X_s$$

is open if and only if its intersection with each X_s is open.

Lemma A.6.17. *Let S have the discrete topology and let X be a space. Then the product topology on $S \times X$ coincides with the disjoint union topology on $\bigcup_{s \in S} (\{s\} \times X)$.*

Proof. Since S is discrete, $S \times X$ has basis $\{\{s\} \times U : U \text{ open in } X\}$. \square

A.7. Topology of the orbit space. Now, finally, we shall tackle the orbit spaces of free and properly discontinuous actions.

Theorem A.7.1. *Let $\pi : X \rightarrow X/G$ be the canonical map to the orbit space of a G -action on X . Then π is an open map.*

Now let $U \subset X$ have the property that

$$(A.7.1) \quad gU \cap U = \emptyset \quad \text{for all } 1 \neq g \in G.$$

Then there is a G -homeomorphism

$$\mu : G \times U \rightarrow \pi^{-1}\pi(U)$$

given by $\mu((g, x)) = gx$ for all $g \in G$ and $x \in U$. Here, G has the discrete topology. Moreover, $\pi|_U : U \rightarrow \pi(U)$ is a homeomorphism and the following diagram commutes:

$$(A.7.2) \quad \begin{array}{ccc} G \times U & \xrightarrow{\mu} & \pi^{-1}\pi(U) \\ \searrow \pi \circ \pi_2 & & \swarrow \pi \\ & & \pi(U). \end{array}$$

Here π_2 is projection onto the second factor. In consequence

$$\pi|_{g(U)} : g(U) \rightarrow \pi(U)$$

is a homeomorphism for each $g \in G$.

Finally, if the action of G on X is free and properly discontinuous, then the set of open sets satisfying (A.7.1) is a basis of X .

Proof. Let X be a G -space. For any $x \in X$, $\pi^{-1}\pi(x) = Gx = \{gx : g \in G\}$, the orbit of x . So, for any $Y \subset X$,

$$\pi^{-1}\pi(Y) = \bigcup_{g \in G} g(Y).$$

Suppose Y is open in X . Then so is each $g(Y)$, since multiplication by g is a homeomorphism, so $\pi^{-1}\pi(Y)$, as a union of open subsets, is open. Since the orbit space has the quotient topology, $\pi(Y)$ is open. Thus, π is an open map.

Now suppose U satisfies (A.7.1). Then if $g \neq h \in G$, $g(U) \cap h(U) = \emptyset$, as otherwise $h^{-1}g(U) \cap U$ would be nonempty. But that makes μ one-to-one. As shown above for $U = Y$, the image of μ is $\pi^{-1}\pi(U)$. Moreover, μ is an open map, as it carries the basis elements $\{g\} \times V$ with V open in U onto the open sets $g(V)$. Thus, μ is a homeomorphism, and the commutativity of the diagram follows from the fact that $\pi(gx) = \pi(x)$ for all $x \in X$ and $g \in G$. Since μ is a G -map, it is a G -homeomorphism.

By (A.7.1), $\pi|_U$ is one-to-one, and since π is an open map,

$$\pi|_U : U \xrightarrow{\cong} \pi(U).$$

The same holds for $\pi|_{g(U)}$ as $\pi|_{g(U)}$ is the composite of π with multiplication by g^{-1} from $g(U)$ to U .

The last statement holds as if (A.7.1) holds for U , it holds for $g(V)$ for any open subset V of U and any $g \in G$. \square

Corollary A.7.2. *If G acts freely and properly discontinuously on the topological manifold M , then M/G is a topological manifold.*

Proof. Since M is a manifold, it has a basis \mathcal{B} of chart neighborhoods. Since the action is free and properly discontinuous, it has a basis \mathcal{B}' of sets satisfying (A.7.1). The set

$$\mathcal{B}'' = \{U \cap V : U \in \mathcal{B} \text{ and } V \in \mathcal{B}'\}$$

is a basis of chart neighborhoods satisfying (A.7.1). In particular, for $W \in \mathcal{B}''$, $\pi(W)$ is a chart neighborhood of M/G . The collection $\{\pi(W) : W \in \mathcal{B}''\}$ forms a basis for M/G . \square

When G acts smoothly on M we obtain a smooth structure on the orbit space:

Proposition A.7.3. *Let G act smoothly, freely and properly discontinuously on the smooth n -manifold M . Then M/G is a smooth n -manifold and $\pi : M \rightarrow M/G$ is smooth. Moreover, the maps π_{kh} in (8.4.2) have Jacobian matrices invertible at each point. So π is both an immersion and a submersion.*

If $f : M \rightarrow N$ is smooth and if $f(gx) = f(x)$ for each $g \in G$ and $x \in M$ then induced map $\bar{f} : M/G \rightarrow N$ in the unique factorization

$$(A.7.3) \quad \begin{array}{ccc} M & \xrightarrow{f} & N \\ & \searrow \pi & \nearrow \bar{f} \\ & M/G & \end{array}$$

is smooth and the rank of $T_{\pi(x)}\bar{f} : T_{\pi(x)}(M/G) \rightarrow T_{f(x)}N$ is equal to the rank of $T_x f$.

Finally, if M is oriented and if each $\mu_g : M \rightarrow M$ is orientation-preserving, then we may give M/G an orientation such that π is orientation-preserving.

Proof. By Lemma 8.4.11, M has an atlas \mathcal{A} whose chart neighborhoods satisfy (A.7.1). For each $h : U \xrightarrow{\cong} h(U)$ in \mathcal{A} , the composite

$$\pi(U) \xrightarrow{\pi^{-1}} U \xrightarrow{h} h(U)$$

\bar{h}

is a homeomorphism, and hence defines a chart for M/G .

The transition maps are somewhat more complicated. Let $h : U \xrightarrow{\cong} h(U)$ and $k : V \xrightarrow{\cong} h(V)$ be charts in \mathcal{A} . Then $\pi^{-1}\pi(V)$ is the disjoint union over $g \in G$ of $g(V)$. In particular,

$$U \cap \pi^{-1}\pi(V) = \bigcup_{g \in G} U \cap g(V),$$

and the union is disjoint. Since $\pi : U \rightarrow \pi(U)$ is a homeomorphism,

$$\pi(U) = \bigcup_{g \in G} \pi(U \cap g(V)),$$

and again the union is disjoint. On $\pi(U \cap g(V))$ the transition map $g_{\bar{k}\bar{h}}$ is given by the composite

$$\pi(U \cap g(V)) \xrightarrow{\pi^{-1}} U \cap g(V) \xrightarrow{\mu_{g^{-1}}} V \xrightarrow{k} k(V),$$

a diffeomorphism onto its image because $\mu_{g^{-1}}$ is a diffeomorphism of M .

The statement about Jacobian matrices is immediate from the construction: with appropriate choices of charts, the Jacobian matrix for π is the identity.

Let $f : M \rightarrow N$ be as stated. Then for the choices of charts for M/G we've just made the maps \bar{f}_{kh} of (8.4.2) coincide with the maps f_{kh} for analyzing Tf , and the result follows.

Finally, if M is oriented, we may choose \mathcal{A} to be an oriented atlas. Since the Jacobian matrices for the transition maps in M/G all come from the Jacobian matrices of the transition maps in M , the resulting atlas on M/G is oriented. By construction, π preserves orientation. \square

Since isometries of \mathbb{R}^n are diffeomorphisms we obtain:

Corollary A.7.4. *Let G act freely and properly discontinuously on \mathbb{R}^n by isometries. Then \mathbb{R}^n/G is a smooth manifold and $\pi : \mathbb{R}^n \rightarrow \mathbb{R}^n/G$ is smooth with the maps (8.4.2) having invertible Jacobian matrices at each point.*

We also know that $G = \{\text{id}, \alpha\}$ acts smoothly on \mathbb{S}^n . We obtain:

Corollary A.7.5. *The Klein bottle, the n -torus \mathbb{T}^n and $\mathbb{R}\mathbb{P}^n$ are smooth manifolds. The canonical maps $\pi : \mathbb{R}^2 \rightarrow K$, $\pi : \mathbb{R}^n \rightarrow \mathbb{T}^n$ and $\pi : \mathbb{S}^n \rightarrow \mathbb{R}\mathbb{P}^n$ are smooth immersions (and submersions).*

This now allows us to prove the following.

Corollary A.7.6. *Define $f : \mathbb{R}^n \rightarrow (\mathbb{S}^1)^n$ by*

$$(A.7.4) \quad f(x_1e_1 + \cdots + x_ne_n) = (\exp(2\pi x_1), \dots, \exp(2\pi x_n)),$$

where $\exp : \mathbb{R} \rightarrow \mathbb{S}^1$ is given by $\exp(x) = \begin{bmatrix} \cos x \\ \sin x \end{bmatrix}$. Then f is an immersion and submersion and factors through \mathbb{T}^n :

$$(A.7.5) \quad \begin{array}{ccc} \mathbb{R}^n & \xrightarrow{f} & (\mathbb{S}^1)^n \\ & \searrow \pi & \nearrow \bar{f} \\ & \mathbb{T}^n & \end{array}$$

The induced map $\bar{f} : \mathbb{T}^n \rightarrow (\mathbb{S}^1)^n$ is a diffeomorphism.

Proof. For $\alpha \in \mathcal{T}_{\Lambda_\varepsilon}$ and $x \in \mathbb{R}^n$, $f(\alpha x) = f(x)$, so f factors as stated. Moreover, f factors a product of n copies of $x \mapsto \exp(2\pi x)$. The latter is both an immersion and a submersion as it is a map between 1-manifolds with nowhere vanishing derivative. So f itself is both an immersion and a submersion.

The restriction of f to $[0, 1)^n \subset I^n$ is bijective. But so is the restriction of π to $[0, 1)^n$, so \bar{f} is bijective. Since π is both an immersion and a submersion, the result follows. \square

The following is a key in defining the tangent bundle.

Proposition A.7.7. *Let M be a smooth manifold with atlas \mathcal{A} . Let*

$$\eta : \coprod_{h \in \mathcal{A}} h(U) \rightarrow M$$

restrict on each $h(U)$ to h^{-1} . Then η is a quotient map, and the equivalence relation on $\coprod_{h \in \mathcal{A}} h(U)$ induced by η is given by setting $h(x) \in h(U)$ equivalent to $k(x) \in k(V)$ whenever $x \in U \cap V$. In particular, if we denote this equivalence relation by \sim , we obtain a homeomorphism

$$\bar{\eta} : \coprod_{h \in \mathcal{A}} h(U) / \sim \xrightarrow{\cong} M.$$

Proof. Because each $h^{-1} : h(U) \rightarrow U$ is a homeomorphism onto an open subset of M , η is an open map. Continuous, open maps are always quotient maps. \square

Appendix B. Compactness

B.1. Heine–Borel.

Definition B.1.1.

- (1) An open cover of a space X is a collection $\mathcal{U} = \{U_\alpha : \alpha \in A\}$ of open subsets U_α of X , that covers X in the sense that

$$(B.1.1) \quad \bigcup_{\alpha \in A} U_\alpha = X.$$

- (2) A subcover of an open cover $\{U_\alpha : \alpha \in A\}$ of X is a subcollection of these sets that still cover X , i.e., the subcover is $\{U_\alpha : \alpha \in B\}$ for some $B \subset A$, such that

$$\bigcup_{\alpha \in B} U_\alpha = X.$$

We say this subcover is finite if B is a finite set.

- (3) A space X is compact if every open cover of X has a finite subcover.
 (4) A subset Y of a metric space X is bounded if there exists $z \in X$ and $r > 0$ such that Y is contained in the open ball of radius r about z : $Y \subset B_r(z)$.

Example B.1.2. The real numbers \mathbb{R} is not compact: the open cover

$$\{(n-1, n+1) : n \in \mathbb{Z}\}$$

does not admit a finite subcover because the intervals $(n-1, n+1)$ are bounded. But the triangle inequality and induction show that any finite union of bounded sets is bounded.

Example B.1.3. Any topology on a finite set X gives a compact space, because there are only finitely many subsets of X in the first place.

The Heine–Borel theorem (see, e.g., [8]) is one of the most important theorems about the topology of the real line. It gives us our first nontrivial example of a compact space.

Theorem B.1.4 (Heine–Borel). *The closed intervals $[a, b] \subset \mathbb{R}$ are compact.*

We can now obtain a number of more exotic compact spaces from the following.

Proposition B.1.5. *A closed subspace of a compact space is compact.*

Proof. Let X be compact and let C be a closed subspace of X . An open cover of C has the form $\mathcal{U} = \{U_\alpha : \alpha \in A\}$, where $U_\alpha = V_\alpha \cap C$, with V_α open in X . So

$$C \subset \bigcup_{\alpha \in A} V_\alpha,$$

an open set in X . Since C is closed in X , $\{V_\alpha : \alpha \in A\} \cup \{X \setminus C\}$ is an open cover of X . Since X is compact, there is a finite subset $S \subset A$ such

that $\{V_\alpha : \alpha \in S\} \cup \{X \setminus C\}$ covers X . But then $\{U_\alpha : \alpha \in S\}$ is a finite subcover of our original cover \mathcal{U} . \square

Example B.1.6. Let $C = \{\frac{1}{n} : n > 0\} \cup \{0\} \subset I = [0, 1]$. Then

$$I \setminus C = \bigcup_{n>0} \left(\frac{1}{n+1}, \frac{1}{n} \right)$$

is open in \mathbb{R} and hence also in I . So C is a closed subspace of the compact space I and hence is compact.

Note that C consists of a decreasing sequence of positive real numbers that converges to 0, together with its limit point 0.

Example B.1.7. The Cantor set is obtained from $I = [0, 1]$ by performing a sequence of deletions. We first delete the open middle third, $(\frac{1}{3}, \frac{2}{3})$, obtaining the disjoint union of two closed intervals of width $\frac{1}{3}$:

$$C_1 = \left[0, \frac{1}{3} \right] \cup \left[\frac{2}{3}, 1 \right].$$

We then delete the open middle third of each of these intervals, obtaining the disjoint union of four intervals of width $\frac{1}{9}$:

$$C_2 = \left[0, \frac{1}{9} \right] \cup \left[\frac{2}{9}, \frac{1}{3} \right] \cup \left[\frac{2}{3}, \frac{7}{9} \right] \cup \left[\frac{8}{9}, 1 \right].$$

We iterate this process, each time removing the open middle third of each of the intervals, obtaining C_n . the disjoint union of 2^n closed intervals of width $(\frac{1}{3})^n$.

The Cantor set C , is the intersection of all these sets:

$$(B.1.2) \quad C = \bigcap_{n>0} C_n.$$

Each C_n is closed in C_{n-1} , as its complement there is open. So the infinite intersection C is closed in I and hence is compact. Note that $(\frac{1}{3})^n \in C$ for all $n > 0$, so C is infinite. In fact, C is uncountable.

The next lemma is useful enough in geometric topology that it is sometimes given a name. Arunas Liulevicius, who taught me smooth manifold theory, calls it the Hotdog lemma, because of the obvious pictures you can draw when proving it.

Lemma B.1.8. *Let X and Y be spaces. Let A and B be compact subspaces of X and Y , respectively. Let W be an open subset of $X \times Y$ with $A \times B \subset W$.*

(1) *or each $x \in A$, there are open sets U_x and V_x of X and Y , respectively, with*

$$(B.1.3) \quad (x \times B) \subset (U_x \times V_x) \subset W.$$

(2) *There are open subspaces U and V of X and Y , respectively, such that*

$$(B.1.4) \quad (A \times B) \subset (U \times V) \subset W.$$

Proof. (1) For each $y \in B$, there exist open subspaces U_y and V_y of X and Y , respectively, with

$$(x, y) \in (U_y \times V_y) \subset W,$$

as W is open in the product topology. But then $\{V_y \cap B : y \in B\}$ is an open cover of B . Since B is compact, there is a finite subset $S_x \subset B$ such that

$$B \subset \bigcup_{y \in S_x} V_y.$$

Now simply set $U_x = \bigcap_{y \in S_x} U_y$ and $V_x = \bigcup_{y \in S_x} V_y$. These sets satisfy (B.1.3).

(2) We now have an open cover $\{U_x \cap A : x \in A\}$ of A . Since A is compact, there is a finite subset S of A such that

$$A \subset U = \bigcup_{x \in S} U_x.$$

Now set $V = \bigcap_{x \in S} V_x$ and (B.1.4) holds for this U and V . \square

The next proposition is a special case of the famous Tikhonov theorem.¹⁹

Proposition B.1.9. *Let X and Y be compact. Then $X \times Y$ (with the product topology) is compact.*

Proof. Let $\mathcal{U} = \{U_\alpha : \alpha \in A\}$ be an open cover of $X \times Y$. For each $x \in X$, the subset $x \times Y$ is compact by Corollary A.6.15, so there is a finite subset $S_x \subset A$ such that $x \times Y \subset W_x = \bigcup_{\alpha \in S_x} U_\alpha$. By Lemma B.1.8(1), there is an open set U_x in X such that $U_x \times Y \subset W_x$.

Now $\{U_x : x \in X\}$ is an open cover of X . So there is a finite subset $S \subset X$ such that $X = \bigcup_{x \in S} U_x$. But each $U_x \times Y$ is contained in a finite union of elements of \mathcal{U} . Since X is a finite union of sets of the form U_x , $X \times Y$ is contained in a finite union of elements of \mathcal{U} . \square

The following is now useful.

Proposition B.1.10. *A compact subspace C of a Hausdorff space X is closed in X .*

¹⁹The Tikhonov theorem says that if $\{X_\alpha : \alpha \in A\}$ are compact spaces then the infinite product $\prod_{\alpha \in A} X_\alpha$ is compact. This is more interesting than it may seem: if each X_α is finite and nonempty, then $\prod_{\alpha \in A} X_\alpha$ is homeomorphic to the Cantor set; the product $\prod_{0 < n \in \mathbb{Z}} I$ is called the Hilbert cube and has played an important role in geometric topology.

Proof. We show that $X \setminus C$ is open. It suffices to show that for each $y \in X \setminus C$, there is an open set $V \subset X \setminus C$ containing y .

Let $y \in X \setminus C$. Since X is Hausdorff, for each $x \in C$, we can find open sets U_x and V_y containing x and y , respectively, such that $U_x \cap V_y = \emptyset$. Since C is compact, there is a finite set $S \subset C$ such that $C \subset \bigcup_{x \in S} U_x$. But $\bigcup_{x \in S} U_x$ is disjoint from $V = \bigcap_{x \in S} V_x$, an open set in $X \setminus C$ containing y . \square

Corollary B.1.11. *Let A and B be subspaces of the Hausdorff space X with A compact and B closed. Then $A \cap B$ is compact. In particular, the intersection of two compact subspaces of a Hausdorff space is compact.*

Proof. Since B is closed in X , $A \cap B$ is closed in A . But closed subspaces of compact spaces are compact. \square

Example B.1.12. It is not true that compact subspaces of non-Hausdorff spaces are necessarily closed. There are many examples involving exotic topologies on finite sets. One such example is called Sierpinski space: X consists of two points, say 0 and 1, where the open sets are $\{\emptyset, \{0\}, X\}$. This, $\{0\}$ is open but not closed, and $\{1\}$ is closed but not open. $\{0\}$ is a compact subspace, but not closed.

This example actually arises in cases of interest, as it is the prime spectrum of the p -adic integers (and also of the integers localized at p). It is also the quotient space $[0, 1]/[0, 1)$.

We can now characterize the compact subsets of \mathbb{R}^n .

Theorem B.1.13. *A subspace $X \subset \mathbb{R}^n$ is compact if and only if it is closed and bounded (with respect to the standard metric on \mathbb{R}^n).*

Proof. By the triangle inequality, a subspace is bounded if and only if it is contained in $B_r(0)$ for some $r > 0$. And $B_r(0)$ is contained in the compact subspace $[-r, r]^n$. So a closed, bounded subspace is a closed subspace of the compact space $[-r, r]^n$, and hence is compact.

Now let $C \subset \mathbb{R}^n$ be compact. By Proposition B.1.10 it is closed in \mathbb{R}^n . To show it is bounded, consider the open cover $\{B_n(0) \cap C : 0 < n \in \mathbb{Z}\}$. Since C is compact, it has a finite subcover. Since the balls in this cover are nested, the largest ball in this finite subcover contains C . \square

This now gives tons of examples of compact spaces, some exotic and some not. One of the most important ones is the standard $(n - 1)$ -simplex $\Delta^{n-1} = \text{Conv}(e_1, \dots, e_n) \subset \mathbb{R}^n$.

Proposition B.1.14. *The standard $(n - 1)$ -simplex $\Delta^{n-1} \subset \mathbb{R}^n$ is compact.*

Proof. The vertices e_1, \dots, e_n are contained in the compact, convex subset $I^n = [0, 1]^n \subset \mathbb{R}^n$. So their convex hull Δ^{n-1} must also be contained in I^n . Indeed, by (2.8.2), $\Delta^{n-1} = I^n \cap f^{-1}(1)$, where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the linear map $f(x) = \langle x, \xi \rangle$, with $\xi = e_1 + \dots + e_n$. Since f^{-1} is closed in \mathbb{R}^n , Δ^{n-1} is closed in I^n and hence compact. \square

A simpler application of Theorem B.1.13 is:

Corollary B.1.15. *For $x \in \mathbb{R}^n$ and $r > 0$, the closed ball $\bar{B}_r(x)$ is compact.*

Proof. Lemma A.1.15 shows $\bar{B}_r(x)$ to be closed. \square

Since the closed ball is also convex, this is an important example. Here is another:

Corollary B.1.16. *The sphere $\mathbb{S}_r(x)$ of radius r about $x \in \mathbb{R}^n$ is compact for $r > 0$. Here*

$$(B.1.5) \quad \mathbb{S}_r(x) = \bar{B}_r(x) \setminus B_r(x) = \{y \in \mathbb{R}^n : d(x, y) = r\}.$$

Proof. $B_r(x)$ is open in \mathbb{R}^n , so $\mathbb{S}_r(x)$ is a closed subspace of the compact space $\bar{B}_r(x)$. \square

A more exotic example is what's called the Hawaiian earring. It has an interesting fundamental group.

Example B.1.17. For $x \in \mathbb{R}^2$, write $C_r(x)$ for the circle (1-dimensional sphere) of radius r about x . Define the Hawaiian earring by

$$(B.1.6) \quad H = \bigcup_{1 < n \in \mathbb{Z}} C_{\frac{1}{n}}\left(\frac{1}{n}e_1\right).$$

Thus, H is the union of infinitely many circles, any two of which intersect only at the origin, and all tangent to the y -axis. Each circle is compact by Corollary B.1.16, but there are infinitely many of them, so we need further argument to show H is compact.

Of course, H is contained in the compact subset $\bar{B}_{\frac{1}{2}}(\frac{1}{2}e_1)$, so it suffices to show it is closed in it. But

$$\bar{B}_{\frac{1}{2}e_1}\left(\frac{1}{2}e_1\right) \setminus H = \bigcup_{1 < n \in \mathbb{Z}} \left(B_{\frac{1}{n}}\left(\frac{1}{n}e_1\right) \setminus \bar{B}_{\frac{1}{n+1}}\left(\frac{1}{n+1}e_1\right) \right),$$

a union of open sets in \mathbb{R}^2 .

B.2. Maps out of compact spaces.

Lemma B.2.1. *Let $f : X \rightarrow Y$ be continuous, with X compact. Then $f(X)$ is compact.*

Proof. If $\{U_\alpha : \alpha \in A\}$ is an open cover of $f(X)$, then $\{f^{-1}(U_\alpha) : \alpha \in A\}$ is an open cover of X . So there is a finite subset $S \subset A$ such that

$$\{f^{-1}(U_\alpha) : \alpha \in S\}$$

covers X . Since $U_\alpha \subset f(X)$, $U_\alpha = f(f^{-1}(U_\alpha))$, so $\{U_\alpha : \alpha \in S\}$ covers $f(X)$. \square

We have an immediate and important corollary.

Corollary B.2.2. *Polytopes are compact.*

Proof. Let \mathbf{P} be a polytope. By Corollary 2.8.22, there is a surjective, continuous map from a standard simplex Δ^{k-1} onto \mathbf{P} . By Corollary B.1.14, Δ^{k-1} is compact, so the result follows from Lemma B.2.1. \square

Corollary 2.8.39 enumerates the convex subsets of a line ℓ . Of them, only the line segments $[x, y]$ and the individual points $\{x\}$ are closed and bounded. We also know that if \mathbf{P} is a one-dimensional polytope, then $\text{Aff}(\mathbf{P})$ is a line. We obtain the following.

Corollary B.2.3.

- (1) *The compact, convex subsets of a line $\ell \subset \mathbb{R}^n$ consist of the segments $[x, y]$ with $x \neq y \in \ell$, together with the singleton points $x \in \ell$.*
- (2) *A one-dimensional polytope in \mathbb{R}^n is a line segment $[x, y]$ with $x \neq y \in \mathbb{R}^n$.*

Proposition B.2.4. *Let $f : X \rightarrow Y$ be continuous, with X compact and Y Hausdorff. Then f is a closed map, i.e., $f(C)$ is closed in Y whenever C is closed in X . Moreover, each such $f(C)$ is compact.*

Proof. Let C be closed in X . Then C is compact by Proposition B.1.5. So $f(C)$ is compact by Lemma B.2.1. Since Y is Hausdorff, $f(C)$ is closed in Y by Proposition B.1.10. \square

This gives a very important corollary.

Corollary B.2.5. *Let $f : X \rightarrow Y$ be injective and continuous, with X compact and Y Hausdorff. Then*

$$f : X \xrightarrow{\cong} f(X)$$

is a homeomorphism onto a closed, compact subspace of Y .

Proof. $f : X \rightarrow f(X)$ is continuous and bijective, so it suffices to show $f^{-1} : f(X) \rightarrow X$ is continuous. But if $C \subset X$ is closed, then $f(C)$ is closed in $f(X)$ by Proposition B.2.4, and the result follows. \square

B.3. Cones and convex bodies.

B.3.1. Cones. The cone construction has a variety of uses in topology, from geometry to homotopy theory. Cones work best in the context of compactness.

Definition B.3.1. The cone CX on a space X is the quotient space

$$CX = X \times I / X \times 1,$$

where $I = [0, 1]$. This can be described as the identification space $X \times I / \sim$, where $(x, 1) \sim (y, 1)$ for $x, y \in X$. We write $\pi : X \times I \rightarrow CX$ for the quotient map. It is customary to write $[x, t]$ for $\pi((x, t))$.

The basic idea of a cone is that we stretch X out and pinch it to a point. The following illustrates the utility of this idea in geometry.

Proposition B.3.2. Write \mathbb{D}^n for the unit disk in \mathbb{R}^n and write \mathbb{S}^{n-1} for the unit sphere. Then there is a homeomorphism

$$f : C\mathbb{S}^{n-1} \rightarrow \mathbb{D}^n \\ [x, t] \mapsto (1 - t)x.$$

Proof. f is continuous because the composite $f \circ \pi : \mathbb{S}^{n-1} \times I \rightarrow \mathbb{D}^n$ is continuous. And f is bijective because each element $z \in \mathbb{D}^n \setminus \{0\}$ may be written uniquely in the form $z = sx$ with $s \in (0, 1]$ and $x \in \mathbb{S}^{n-1}$: $s = \|z\|$ and $x = \frac{z}{\|z\|}$. So $f \circ \pi$ is a bijection from $\mathbb{S}^{n-1} \times [0, 1)$ to $\mathbb{D}^n \setminus \{0\}$ and maps $\mathbb{S}^{n-1} \times 1$ onto 0 .

Since $\mathbb{S}^{n-1} \times I$ is compact, so is its quotient space $C\mathbb{S}^{n-1}$. So f is a continuous bijection from a compact space onto the Hausdorff space \mathbb{D}^n , and hence is a homeomorphism by Corollary B.2.5. \square

Of course, the geometry of \mathbb{D}^n as a compact, convex subset of \mathbb{R}^n is lost when we view it as a cone, but Proposition B.3.2 gives us a way to compare it topologically to other geometric objects. Topology is weaker than geometry, but can be easier to measure and understand.

B.3.2. Convex bodies.

Definition B.3.3. A convex body in \mathbb{R}^n is a compact, convex subset C in \mathbb{R}^n of dimension n , i.e., $\text{Aff}(C) = \mathbb{R}^n$.

Examples B.3.4. The standard regular polygon P_n , $n \geq 3$, is a convex body in \mathbb{R}^2 . The Platonic solids are convex bodies in \mathbb{R}^3 . The unit disk \mathbb{D}^n is a convex body in \mathbb{R}^n .

Remark B.3.5. Let C be a convex body in \mathbb{R}^n . By Corollary 2.9.27, the interior $\text{Int}(C)$ is nonempty, and by Lemma 2.9.16, $x \in \text{Int}(C)$ if and only if the open ball $B_\epsilon(x) \subset C$ for some $\epsilon > 0$.

This, of course, shows that $\text{Int}(C)$ is an open subspace of \mathbb{R}^n , so the boundary

$$\partial C = C \setminus \text{Int}(C)$$

is closed in C , and hence is compact.²⁰

What we are really studying here is compact, convex subspaces of \mathbb{R}^n . The ones that are also n -dimensional have a name, and have special prominence for a number of reasons. But the following nonstandard definition may provide a useful perspective.

Definition B.3.6. Let H be an affine subspace of \mathbb{R}^n . A convex body in H is a compact, convex subset $C \subset H$ whose affine hull is H .

In particular, every compact, convex subspace of \mathbb{R}^n is a convex body in its affine hull. Since polytopes are compact, we obtain the following.

²⁰The basic topology of \mathbb{R}^n guarantees that $\text{Int}(C)$ cannot be closed in \mathbb{R}^n and hence ∂C is nonempty (as C is closed in \mathbb{R}^n). We shall give a simpler argument that $\partial C \neq \emptyset$.

Lemma B.3.7. *Every polytope is a convex body in its affine hull.*

Corollary 2.9.27 and Lemma 2.9.16 again give the following.

Lemma B.3.8. *Let C be a convex body in the affine subspace $H \subset \mathbb{R}^n$. Then*

$$(B.3.1) \quad \text{Int}(C) = \{x \in C : B_\epsilon(x, H) \subset C \text{ for some } \epsilon > 0\}$$

is an open subset of H , and

$$\partial C = C \setminus \text{Int}(C)$$

is compact.

A very useful technique in convex geometry is intersecting a convex subset with an affine subspace. Faces of polytopes are obtained in this way. So are hyperplane sections. Intersection with lines is important in what follows.

Lemma B.3.9. *Let C be a compact, convex subset of \mathbb{R}^n and let H be an affine subspace. Then $C \cap H$ is a compact, convex subset.*

Proof. Intersections of convex subsets are convex, so it suffices to show $C \cap H$ is a compact. By Lemma A.1.19, affine subspaces are closed. Now apply Corollary B.1.11. \square

We shall make use of the following:

Lemma B.3.10. *Let C be a convex subset of \mathbb{R}^n . Let $x \in \text{Int}(C)$ and $y \in C$. Then every element in the half-open line segment $[x, y)$ lies in $\text{Int}(C)$.*

Proof. Let $H = \text{Aff}(C)$. Then a point z lies in $\text{Int}(C)$ if and only if $B_\epsilon(z, H) \subset C$ for some $\epsilon > 0$. Here, if V is the linear base of H ,

$$B_\epsilon(z, H) = \tau_z(B_\epsilon(0, V)) = \{z + v : v \in B_\epsilon(0, V)\}.$$

Let $\gamma_t : H \rightarrow H$ be given by $\gamma_t(w) = (1-t)w + ty$. Then z lies in $[x, y)$ if and only if $z = \gamma_t(x)$ for some $t \in [0, 1)$. But

$$\gamma_t(x + v) = (1-t)x + ty + (1-t)v = z + (1-t)v,$$

so $\gamma_t(B_\epsilon(x, H)) = B_{(1-t)\epsilon}(z, H)$. Since C is convex, $B_{(1-t)\epsilon}(z, H) \subset C$, and hence $z \in \text{Int}(C)$. \square

We can now show the following.

Corollary B.3.11. *Let C be a compact, convex subset of \mathbb{R}^n and let $x \in \text{Int}(C)$. Let ℓ be a line in \mathbb{R}^n containing x . Then:*

- (1) *If ℓ is not contained in the affine hull $\text{Aff}(C)$, then $\ell \cap C = \{x\}$.*
- (2) *If ℓ is contained in $\text{Aff}(C)$, then $\ell \cap C$ is a line segment $[y, z]$ with $(y, z) \subset \text{Int}(C)$ and $y, z \in \partial C$.*

Proof. (1) If $\ell \cap C$ contains a point $y \neq x$, then ℓ is the unique line containing both x and y . Since both x and y are in $\text{Aff}(C)$, so is this line.

(2) Let $x \neq w \in \ell$ and consider the affine map $\gamma : \mathbb{R} \rightarrow \ell$ given by $\gamma(t) = (1-t)x + tw$. Since $x \in \text{Int}(C)$, $B_\epsilon(x) \subset C$ for some $\epsilon > 0$. Since affine maps are continuous, $B_\delta(0) \subset \gamma^{-1}(B_\epsilon(x))$ for some $\delta > 0$. So $\gamma(B_\delta(0)) \subset \ell \cap C$. By Lemma B.3.9, $\ell \cap C$ is compact and convex. Since it consists of more than one point, it must be a line segment $[y, z]$ for $y \neq z$ by Corollary B.2.3. Moreover, $\gamma(B_\delta(0)) \subset [y, z]$, so $x \in (y, z)$. By Lemma B.3.10, all of (y, z) is contained in $\text{Int}(C)$.

But y cannot lie in $\text{Int}(C)$: if it did there would be an open interval about $\gamma^{-1}(y)$ mapping into C . But $\gamma : \mathbb{R} \rightarrow \ell$ is an affine isomorphism, so $\gamma^{-1}(C) = \gamma^{-1}([x, y])$ is a closed interval. So $y \in \partial C$, as is z by the same argument. \square

We can now generalize Proposition B.3.2 to arbitrary compact, convex subsets of \mathbb{R}^n .

Proposition B.3.12. *Let C be a compact, convex subset of \mathbb{R}^n and let $x \in \text{Int}(C)$. Then there is a homeomorphism*

$$(B.3.2) \quad h : C(\partial C) \xrightarrow{\cong} C \\ [y, t] \mapsto (1-t)y + tx,$$

from the cone on ∂C to C .

Proof. Let $x \neq z \in C$. By Corollary B.3.11, the ray \overrightarrow{xz} meets ∂C in exactly one point, say y . Again by Corollary B.3.11, we must have that $z \in [x, y]$. But $[x, y]$ is precisely the image of the restriction of f to $\{[y, t] : t \in I\}$, on which f is one-to-one. So f is a bijection. Since $C(\partial C)$ is compact and C is Hausdorff, f is a homeomorphism by Corollary B.2.5. \square

But now we can use the topology of cones to show that any two compact, convex subsets of \mathbb{R}^n of the same dimension are homeomorphic. We first treat the case of convex bodies.

Theorem B.3.13. *Let C be a convex body in \mathbb{R}^n . Then ∂C is homeomorphic to \mathbb{S}^{n-1} , and hence C is homeomorphic to \mathbb{D}^n .*

Proof. After applying a translation, if necessary, we may assume the origin is an interior point of C . Now consider the lines through the origin. By Corollary B.3.11, each line through the origin intersects ∂C in exactly two points, one on each side of the origin. Thus, the function

$$f : \partial C \rightarrow \mathbb{S}^{n-1} \\ x \mapsto \frac{x}{\|x\|}$$

is bijective. Since both ∂C and \mathbb{S}^{n-1} are compact Hausdorff spaces, f is a homeomorphism.

Since homeomorphic spaces have homeomorphic cones, C is homeomorphic to \mathbb{D}^n . (A compactness argument could be used here, as well.) \square

This argument generalizes easily:

Theorem B.3.14. *Let C be a compact, convex subset of \mathbb{R}^n with $\dim C = k$. Then ∂C is homeomorphic to \mathbb{S}^{k-1} and hence C is homeomorphic to \mathbb{D}^k .*

Proof. Let $H = \text{Aff}(C)$ and let V be its linear base. Let $y \in \text{Int}(C)$. Translating by $-y$, we may assume $H = V$ and $0 \in \text{Int}(C)$. Then as in the proof of Theorem B.3.13, the map $x \mapsto \frac{x}{\|x\|}$ is a homeomorphism from ∂C to the unit sphere of V . But an orthonormal basis of V provides a linear isometry from \mathbb{R}^k to V , inducing an isometric homeomorphism from \mathbb{S}^{k-1} to the unit sphere of V . \square

This is a nice, abstract setting in which to have worked, but the result is actually striking when applied to polytopes.

Corollary B.3.15. *Let \mathbf{P} be a polytope in \mathbb{R}^n with $\dim \mathbf{P} = k$. Then $\partial \mathbf{P}$ is homeomorphic to \mathbb{S}^{k-1} , and \mathbf{P} is homeomorphic to \mathbb{D}^k .*

In particular, by Corollary 8.3.8, $\partial \mathbf{P}$ is a manifold, which is rather striking, given the definitions.

References

- [1] ANTON, HOWARD; RORRES, CHRIS. Elementary linear algebra: applications version. Tenth edition. *John Wiley & Sons, New York*, 2010.
- [2] BREDON, GLEN E. Topology and geometry. Graduate Texts in Mathematics, 139. *Springer-Verlag, New York*, 1993. ISBN: 0-387-97926-3. [MR1224675](#) (94d:55001).
- [3] CARMO, MANFREDO P. DO. Riemannian geometry. Translated by Francis Flaherty. *Birkhäuser, Boston*, 1992.
- [4] CARMO, MANFREDO P. DO. Differential geometry of curves and surfaces. *Prentice-Hall, Inc., Englewood Cliffs, NJ*, 1976.
- [5] CHILDS, LINDSAY N. A concrete introduction to higher algebra. Third edition. Undergraduate Texts in Mathematics. *Springer, Berlin*, 2009. ISBN: 978-0-387-74527-5. [MR2464583](#) (2009i:00001).
- [6] CONWAY, JOHN H.; BURGIEL, HEIDI; GOODMAN-STRAUSS, CHAIM. The symmetries of things. A K Peters, Ltd., Wellesley, MA, 2008. xviii+426 pp. ISBN: 1-56881-220-5. [MR2410150](#) (2009c:00002), [Zbl 1173.00001](#).
- [7] CONWAY, J. H.; SLOANE, N. J. A. Sphere packings, lattices and groups. With additional contributions by E. Bannai, J. Leech, S. P. Norton, A. M. Odlyzko, R. A. Parker, L. Queen and B. B. Venkov. Grundlehren der Mathematischen Wissenschaften, 290. *Springer-Verlag, New York*, 1988. xxviii+663 pp. ISBN: 0-387-96617-X. [MR0920369](#) (89a:11067), [Zbl 0634.52002](#).
- [8] DUGUNDJI, JAMES. Topology. Reprinting of the 1966 original. Allyn and Bacon Series in Advanced Mathematics. *Allyn and Bacon, Inc., Boston, Mass.*, 1978. xv+447 pp. ISBN: 0-205-00271-4. [MR0478089](#) (57 #17581).
- [9] GUGGENHEIMER, HEINRICH W. Plane geometry and its groups. *Holden-Day, Inc., San Francisco*, 1967. xii+288 pp. [MR0213943](#) (35 #4796), [Zbl 0147.38801](#).
- [10] HAN, QING; HONG, JIA-XING. (2006), Isometric embedding of Riemannian manifolds in Euclidean spaces. *American Mathematical Society, Providence, RI*, 2006. ISBN 0-8218-4071-1
- [11] LANG, SERGE. Algebra. Revised third edition. Graduate Texts in Mathematics, 211. *Springer-Verlag, New York*, 2002. xvi+914 pp. ISBN: 0-387-95385-X. [MR1878556](#) (2003e:00003).
- [12] LANG, SERGE. Introduction to differentiable manifolds. Second edition. Universitext. *Springer-Verlag, New York*, 2002. xii+250 pp. ISBN: 0-387-95477-5. [MR1931083](#) (2003h:58002).
- [13] LEE, JOHN M. Introduction to smooth manifolds. Second edition. Graduate Texts in Mathematics, 218. *Springer, New York*, 2013. xvi+708 pp. ISBN: 978-1-4419-9981-8. [MR2954043](#), [Zbl 1258.53002](#).
- [14] MARTIN, GEORGE. Transformation geometry. An introduction to symmetry. Undergraduate Texts in Mathematics. *Springer-Verlag, New York*, 1982. XII+237pp. [Zbl 0484.51001](#).
- [15] MUNKRES, JAMES R. Elements of algebraic topology. Advanced Book Program. *Perseus Books, Cambridge, MA*, 1984.
- [16] RYAN, PATRICK J. Euclidean and non-Euclidean geometry. An analytical approach. *Cambridge University Press, Cambridge*, 1986. xviii+215 pp. ISBN: 0-521-25654-2; 0-521-27635-7. [MR0854104](#) (87i:51001), [Zbl 0592.51013](#).
- [17] STEINBERGER, MARK. Algebra. <http://www.albany.edu/~mark/algebra.pdf>.
- [18] THORPE, JOHN A. Elementary topics in differential geometry. Undergraduate Texts in Mathematics. *Springer-Verlag, NY*, 1979.

DEPARTMENT OF MATHEMATICS AND STATISTICS, UNIVERSITY AT ALBANY, ALBANY, NY 12222

<http://albany.edu/~mark/>