RHODES UNIVERSITY
Grahamstown 6140, South Africa

Lecture Notes CCR

M4.3 - Geometry

Claudiu C. Remsing

DEPT. of MATHEMATICS (Pure and Applied) **2006**

"Where there is matter, there is geometry".

J. Kepler

"Imagination is more important than knowledge".

A. Einstein

Contents

Chapter 1

Geometric Transformations

Topics :

1. Euclidean 3-Space
2. Linear Transformations
3. Translations and Affine Transformations
4. Isometries
5. Galilean Transformations
6. Lorentz Transformations

1.1 Euclidean 3-Space

The Euclidean space, points, and vectors

Three-dimensional *visual space* $\mathcal{S}$ is often used in mathematics without being formally defined. The "elements" of $\mathcal{S}$ are called *points*. In the usual sense, we introduce *Cartesian coordinates* by fixing a point, called the *origin*, and three (mutually orthogonal) *coordinate axes*. The choice of origin and of axes is arbitrary, but once it has been fixed, three real numbers (or coordinates) p_1, p_2, p_3 can be *measured* to describe the position of each point p.

The one-to-one correspondence

$$p \in \mathcal{S} \mapsto (p_1, p_2, p_3) \in \mathbb{R}^3$$

makes possible the *identification* of $\mathcal{S}$ with the set $\mathbb{R}^3$ of all ordered triplets of real numbers. In other words, instead of saying that three numbers *describe the position* of a point, we define them to *be* the point.

We make the following definition.

1.1.1 Definition. The (standard) **Euclidean 3-space** is the set $\mathbb{R}^3$ together with the *Euclidean distance* between points $p = (p_1, p_2, p_3)$ and $q = (q_1, q_2, q_3)$ given by

$$d(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + (q_3 - p_3)^2}.$$

Note : Euclidean 3-space $\mathbb{R}^3$ is a *model* for the physical space. There are other models for our Universe. The question of what is the most convenient geometry with which to model physical space is an open one, and is the subject of intense contemporary investigation and speculation.

Let $p = (p_1, p_2, p_3)$ and $q = (q_1, q_2, q_3)$ be two points of $\mathbb{R}^3$, and let λ be a *scalar* (real number). The **sum** of p and q is the point

$$p + q := (p_1 + q_1, p_2 + q_2, p_3 + q_3)$$

and the **scalar multiple** of p by λ is the point

$$\lambda p := (\lambda p_1, \lambda p_2, \lambda p_3).$$

Under these two operations (the usual addition and scalar multiplication), $\mathbb{R}^3$ is a *vector space* over $\mathbb{R}$.

NOTE : The origin $o = (0, 0, 0)$ plays the role of *identity* (with respect to addition). The sum $p + (-1)q$ is usually written $p - q$.

We shall consider now the relationship between *points* and *geometric vectors* in Euclidean 3-space $\mathbb{R}^3$.

NOTE : The concept of *vector* originated in physics from such notions as velocity, acceleration, force, and angular momentum. These physical quantities are supplied with length and direction; they can be added and multiplied by scalars.

Intuitively, a geometric vector v in $\mathbb{R}^3$ is represented by a directed line segment (or "arrow") $\overrightarrow{pq}$. Here we take the view that a geometric vector is really the same thing as a *translation* in space.

NOTE : We can also take the view that we can describe an "arrow" (located at some point) by giving the starting point and the *change* necessary to reach its terminal point. This approach leads to the concept of (geometric) *tangent vector* and will be considered in the next chapter.

We make the following definition.

1.1.2 DEFINITION. A (geometric) **vector** in Euclidean 3-space $\mathbb{R}^3$ is a mapping

$$v : \mathbb{R}^3 \to \mathbb{R}^3, \quad p \mapsto v(p)$$

such that for any two points p and q, the midpoint of $\overline{pv(q)}$ is equal to the midpoint of $\overline{qv(p)}$.

Thus, if v is a vector and p, q are two points, then the quadrilateral $\square\, pqv(q)v(p)$ is a *parallelogram* (proper or degenerate).

$\diamond$ **Exercise 1** Show that given two points p and q, there is exactly one vector v such that $v(p) = q$.

This unique vector is denoted by $\overrightarrow{pq}$. A vector $\overrightarrow{pq}$ is sometimes called a **free vector**.

NOTE : An alternative description is the following. Two directed line segments $\overrightarrow{pq}$ and $\overrightarrow{p'q'}$ (or, if one prefers, two ordered pairs of points (p, q) and (p', q')) are *equivalent* if the line segments $\overrightarrow{pq}$ and $\overrightarrow{p'q'}$ are of the same length and are parallel in the same sense. This relation, being reflexive, symmetric, and transitive, is a genuine *equivalence relation*. Such an equivalence class of directed line segments (or, if one prefers, of ordered pairs of points) is a *vector*. We denote the vector $[\overrightarrow{pq}]$ simply by $\overrightarrow{pq}$. If $p = (p_1, p_2, p_3)$ and $q = (q_1, q_2, q_3)$, the *components* of the vector are $q_1 - p_1, q_2 - p_2,$ and $q_3 - p_3$. Two vectors are *equal* if and only if they have the same components.

◇ **Exercise 2** Show that two directed line segments $\overrightarrow{pq}$ and $\overrightarrow{p'q'}$ are *equivalent* if and only if $p + q' = p' + q$.

If $p = (p_1, p_2, p_3)$ and $q = (q_1, q_2, q_3)$, it is customary to represent the vector $v = \overrightarrow{pq}$ by the 3×1 matrix

$$\begin{bmatrix} q_1 - p_1 \\ q_2 - p_2 \\ q_3 - p_3 \end{bmatrix}.$$

Let o be the origin of the Euclidean 3-space $\mathbb{R}^3$. Any point $p \in \mathbb{R}^3$ can be described by means of the vector $\overrightarrow{op}$ (the *position vector* of the point p). Each point has a unique position vector, and each position vector describes a unique point. Hence we set up a one-to-one correspondence between points and geometric vectors in $\mathbb{R}^3$. It is convenient to *identify*

$$\text{the point } (p_1, p_2, p_3) \text{ with the vector } \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}.$$

NOTE : An element of Euclidean 3-space $\mathbb{E}^3$ can be considered (or represented) either as an ordered triplet of real numbers or as a column 3-matrix with real entries. In other words, *we can think of the Euclidean 3-space as either the set of all its points or the set of all its (geometric) vectors.*

◇ **Exercise 3** Explain why the *identification* of the vector $v = \overrightarrow{pq}$ with the point $q - p$ is legitimate.

The (vector) space $\mathbb{R}^3$ has a built-in standard *inner product* (i.e., a non-degenerate symmetric bilinear form). For $v, w \in \mathbb{R}^3$, the **dot product** of (the vectors) v and w is the number (scalar)

$$v \bullet w := v_1\, w_1 + v_2\, w_2 + v_3\, w_3.$$

The dot product is a *positive definite* inner product; that is, it has the following three properties (for $v, v', w \in \mathbb{R}^3$ and $\lambda, \lambda' \in \mathbb{R}$) :

(IP1) $(\lambda v + \lambda' v') \bullet w = \lambda(v \bullet w) + \lambda'(v' \bullet w)$ (linearity);

(IP2) $v \bullet w = w \bullet v$ (symmetry);

(IP3) $v \bullet v \geq 0$, and $v \bullet v = 0 \iff v = 0$ (positivity).

◇ **Exercise 4** Given $v, w \in \mathbb{R}^3$, show that

$$(v \bullet w)^2 \leq (v \bullet v)(w \bullet w).$$

This inequality is called the *Cauchy-Schwarz inequality.*

Write

$$\|v\| := \sqrt{v \bullet v} = \sqrt{v_1^2 + v_2^2 + v_3^2}$$

and call it the **norm** (or **length**) of (the vector) v. A vector with unit norm is called a **unit vector**.

NOTE : In view of our definition, we can rewrite the Cauchy-Schwarz inequality in the form

$$|v \bullet w| \leq \|v\|\, \|w\|.$$

The norm (more precisely, the *norm function* $v \in \mathbb{R}^3 \mapsto \|v\| \in \mathbb{R}$) has the following properties (for $v, w \in \mathbb{R}^3$ and $\lambda \in \mathbb{R}$) :

(N1) $\|v\| \geq 0$, and $\|v\| = 0 \iff v = 0$ (positivity);

(N2) $\|\lambda v\| = |\lambda|\, \|v\|$ (homogeneity);

(N3) $\|v + w\| \leq \|v\| + \|w\|$ (the triangle inequality).

◇ **Exercise 5** Let $v, w \in \mathbb{R}^3$. Verify the following properties.

(a) *Polarization identity* : $v \bullet w = \frac{1}{4}\left(\|v+w\|^2 - \|v-w\|^2\right)$, which expresses the standard inner product in terms of the norm.

(b) *Parallelogram identity* : $\|v+w\|^2 + \|v-w\|^2 = 2\left(\|v\|^2 + \|w\|^2\right)$. That is, the sum of the squares of the diagonals of a parallelogram equals the sum of the squares of the sides.

◇ **Exercise 6** Given $v, w \in \mathbb{R}^3$, prove the *Pythagorean property*

$$v \bullet w = 0 \iff \|v \pm w\|^2 = \|v\|^2 + \|w\|^2.$$

In terms of the norm we get a compact version of the (Euclidean) distance formula :

$$d(p,q) = \|v - w\| \quad \text{with} \quad v = \overrightarrow{op} \ \text{ and } \ w = \overrightarrow{oq}.$$

In other words, $\|v - w\|$ represents the distance between two points with position vectors v and w.

◇ **Exercise 7** Verify that the Euclidean distance satisfies the following properties (the axioms for a *metric*):

(M1) $d(p,q) \geq 0$, and $d(p,q) = 0 \iff p = q$;

(M2) $d(p,q) = d(q,p)$;

(M3) $d(p,r) \leq d(p,q) + d(q,r)$.

Relation (M3) is also known as the *triangle inequality*.

NOTE : Euclidean 3-space $\mathbb{R}^3$ is not only a vector space. It is also a *metric space*. It is important to realize that the Euclidean distance is completely determined by the dot product; indeed,

$$d(p,q) = \sqrt{(q-p) \bullet (q-p)} \qquad (p, q \in \mathbb{R}^3).$$

However, not any distance function is associated with an inner product. A (real) vector space endowed with a specific (positive definite) inner product is called an *inner product space*.

Let v and w be two *nonzero* vectors of $\mathbb{R}^3$. The Cauchy-Schwarz inequality permits us to define the cosine of the *angle* θ, $0 \leq \theta \leq \pi$ between v and w by the equation

$$v \bullet w = \|v\| \, \|w\| \, \cos\theta.$$

Thus the dot product of two vectors is the product of their lengths times the cosine of the angle between them. If $\theta = 0$ or $\theta = \pi$, the vectors v and w are said to be **collinear**, whereas if $\theta = \frac{\pi}{2}$, the vectors are called **orthogonal**.

NOTE : We regard the zero vector as both collinear with and orthogonal to *every* vector. Clearly, vectors v and w are orthogonal if and only if $v \bullet w = 0$.

$\diamond$ **Exercise 8** Given a nonzero vector w, show that vectors v and w are collinear if and only if $v = \lambda w$ for some $\lambda \in \mathbb{R}$.

There is another product on the Euclidean 3-space $\mathbb{R}^3$, second in importance only to the dot product. For $v, w \in \mathbb{R}^3$, the **cross product** of v and w is the vector

$$v \times w := \begin{bmatrix} v_2 w_3 - v_3 w_2 \\ v_3 w_1 - v_1 w_3 \\ v_1 w_2 - v_2 w_1 \end{bmatrix}.$$

An easy way to remember this formula is to compute the "determinant"

$$v \times w = \begin{vmatrix} e_1 & e_2 & e_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}$$

by *formal* expansion along the first row. Here e_1, e_2, e_3 denote the *standard unit vectors*

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \text{and} \quad e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

NOTE : The vectors e_1, e_2, e_3 are lineary independent, and hence form a (orthonormal) *basis* of the vector space $\mathbb{R}^3$. Any vector $v \in \mathbb{R}^3$ can be expressed uniquely as a linear combination of the standard unit vectors e_1, e_2, e_3 :

$$v = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = v_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + v_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + v_3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = v_1 e_1 + v_2 e_2 + v_3 e_3.$$

Familiar properties of determinants show that the cross product (also called *vector product*) is a skew-symmetric bilinear mapping; that is, it has the following properties (for $v, v', w \in \mathbb{R}^3$ and $\lambda \in \mathbb{R}$) :

(VP1) $(v + v') \times w = v \times w + v' \times w$ (additivity);

(VP2) $\lambda(v \times w) = (\lambda v) \times w$ (homogeneity);

(VP3) $v \times w = -w \times v$ (skew-symmetry).

Hence, in particular, $v \times v = 0$.

◇ **Exercise 9** Show that

$$v \bullet (v \times w) = 0 \quad \text{and} \quad w \bullet (v \times w) = 0.$$

Therefore, the cross product of two vectors is a vector orthogonal to both of them.

◇ **Exercise 10** Verify (by tedious computation) the following formula known as the *Lagrange identity* :

$$\|v \times w\|^2 = \|v\|^2 \|w\|^2 - (v \bullet w)^2.$$

NOTE : The geometric usefulness of the cross product is based mostly on this result. A more intuitive description of the length of a cross product is

$$\|v \times w\| = \|v\| \, \|w\| \sin\theta$$

where θ is the angle between v and w. The *direction* of $v \times w$ on the straight line orthogonal to v and w is given, for practical purposes, by the so-called "right-hand rule": if the fingers of the right hand point in the direction of the shortest rotation of v to w, then the thumb points in the direction of $v \times w$.

◇ **Exercise 11** Show that vectors v and w are collinear if and only if $v \times w = 0$.

Combining the dot and cross product, we get the *triple scalar product* of three vectors $u, v,$ and w : $u \bullet v \times w$. Parantheses are unnecessary : $u \bullet (v \times w)$ is the only possible meaning.

◇ **Exercise 12** Given vectors $u, v,$ and w, show that

$$u \bullet v \times w = v \bullet w \times u = w \bullet u \times v = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}.$$

◇ **Exercise 13** Let $v, w \in \mathbb{R}^3$. Show that the only vector $x \in \mathbb{R}^3$ such that $u \bullet x$ is equal to the determinant $\det \begin{bmatrix} u & v & w \end{bmatrix}$ for all $u \in \mathbb{R}^3$ is $x = v \times w$.

◇ **Exercise 14** Given vectors $u, v,$ and w, show that

$$(u \times v) \times w = (u \bullet w)v - (v \bullet w)u.$$

Deduce the *Jacobi identity* :

$$(u \times v) \times w + (v \times w) \times u + (w \times u) \times v = 0.$$

◇ **Exercise 15** Let $v_1, v_2, w_2, w_2 \in \mathbb{R}^3$. Verify the following identities.

(a) $(v_1 \times v_2) \bullet (w_1 \times w_2) = \det \begin{bmatrix} v_1 \times v_2 & w_1 & w_2 \end{bmatrix}$.

(b) $(v_1 \times v_2) \times (w_1 \times w_2) = \det \begin{bmatrix} v_1 & w_1 & w_2 \end{bmatrix} v_2 - \det \begin{bmatrix} v_2 & w_1 & w_2 \end{bmatrix} v_1$.

Geometric transformations

One of the most important concepts in geometry is that of a *transformation.*

NOTE : Moving geometric figures around is an ancient and natural approach to geometry. However, the Greek emphasis on synthetic geometry and constructions and much later the development of analytic geometry overshadowed transformational thinking. The study of polynomials and their roots in the early nineteenth century led to algebraic transformations and abstract groups. At the same time, AUGUST FERDINAND MÖBIUS (1790-1868) began studying geometric transformations. In the late nineteenth century, FELIX KLEIN (1849-1925) and SOPHUS LIE (1842-1899) showed the central importance of both groups and transformations for geometry.

Generally speaking, a geometric transformation is merely a mapping between two sets. However, these sets are assumed to be, in a certain sense, geometrical; they are equipped with some additional structure and are usually referred to as "spaces". We shall find it convenient to use the word *transformation* ONLY IN THE SPECIAL SENSE of a bijective mapping of a set (space) onto itself. *Groups* of transformations form the heart of geometry.

We make the following definition.

1.1.3 DEFINITION. A (geometric) **transformation** on $\mathbb{R}^3$ is a mapping from $\mathbb{R}^3$ to itself that is one-to-one and onto.

NOTE : Hereafter, in this chapter, all the definitions and results hold for $\mathbb{R}^3$ as well as for the *Euclidean plane* $\mathbb{R}^2$. We shall only discuss the case of $\mathbb{R}^3$, and consider the case of $\mathbb{R}^2$ as a special case.

Let T be a transformation on $\mathbb{R}^3$. Then T can be visualized as "moving" (or transforming) *each* point $p \in \mathbb{R}^3$ to its *unique* image $T(p) \in \mathbb{R}^3$. Given two transformations T and S, their *composition* (T followed by S)

$$ST : \mathbb{R}^3 \to \mathbb{R}^3, \quad p \mapsto S(T(p))$$

is called the **product** of S with T.

◇ **Exercise 16** Verify that the product of two transformations is a transformation.

The **identity transformation** I is defined by

$$I : \mathbb{R}^3 \to \mathbb{R}^3, \quad p \mapsto p.$$

For any transformation T on $\mathbb{R}^3$, $TI = IT = T$. Every transformation T has a unique *inverse* T^{-1}.

◇ **Exercise 17** Given two transformations T and S, show that

$$(ST)^{-1} = T^{-1}S^{-1}.$$

The set of all transformations on $\mathbb{R}^3$ is a (transformation) *group*. Various sets of transformations corespond to important geometric properties and also form groups.

NOTE : FELIX KLEIN in his famous *Erlanger Programm* (1872) used groups of transformations to give a definition of geometry : *Geometry is the study of those properties of a set that are preserved under a group of transformations on that set.* KLEIN showed that various non-Euclidean geometries, projective geometry, and Euclidean geometry were closely related, not competing subjects. He realized that we can, for example, investigate the properties of Euclidean geometry by studing *isometries* (i.e., distance-preserving transformations).

1.2 Linear Transformations

Linear transformations (on $\mathbb{R}^3$) are structure-preserving transformations on the *vector space* $\mathbb{R}^3$. The structure that must be preserved is that of vector addition and scalar multiplication (of which the geometric analogues are the parallelograms with one vertex at the origin and straight lines through the origin, respectively).

1.2.1 DEFINITION. A transformation $T : \mathbb{R}^3 \to \mathbb{R}^3$ is a **linear transformation** if, for all $x, y \in \mathbb{R}^3$ and all $\lambda \in \mathbb{R}$,

(L1) $T(x + y) = T(x) + T(y)$;

(L2) $T(\lambda x) = \lambda T(x)$.

NOTE : The terms *function, mapping, map*, and *transformation* are commonly used interchangeably. However, in studying geometric objects (particularly, on smooth manifolds), it is often convenient to make slight distinctions between them. Thus, we will reserve the term "function" for a map whose range is $\mathbb{R}$ (i.e., a real-valued map), whereas the terms "map" or "mapping" can mean any type of map. Furthermore, invertible maps (or mappings) – on some structured sets – will be referred to as "transformations". Typical transformations are structure-preserving bijections on structured sets of a certain kind. (In modern algebraic parlance, such transformations are usually called *automorphisms*.)

Addition and *scalar multiplication* of linear transformations are defined in the usual way. That is, for (linear) transformations S, T and scalar $\lambda \in \mathbb{R}$,

$$\begin{aligned} (S + T)(x) &:= S(x) + T(x) \\ (\lambda T)(x) &:= \lambda T(x). \end{aligned}$$

◇ **Exercise 18** Is the sum of any two linear transformations a linear transformation ? Justify your answer.

◇ **Exercise 19** Verify that, under the usual product, the set of all linear transformations on $\mathbb{R}^3$ is a (transformation) *group*.

Let $\{e_1, e_2, e_3\}$ be the *standard basis* of $\mathbb{R}^3$ and let T be a linear transformation on $\mathbb{R}^3$. Then we have, uniquely,

$$T(e_i) = a_{1i}e_1 + a_{2i}e_2 + a_{3i}e_3, \quad i = 1, 2, 3.$$

So we can associate to T a 3×3 matrix with real entries

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

Notice that the image of e_i under T is the i^{th} column of the matrix A; that is, $A = \begin{bmatrix} T(e_1) & T(e_2) & T(e_3) \end{bmatrix}$. We can write

$$\begin{aligned} T(x) &= T(x_1e_1 + x_2e_2 + x_3e_3) = x_1T(e_1) + x_2T(e_2) + x_3T(e_3) \\ &= x_1(a_{11}e_1 + a_{21}e_2 + a_{31}e_3) + x_2(a_{12}e_1 + a_{22}e_2 + a_{32}e_3) + \\ &\quad x_3(a_{13}e_1 + a_{23}e_2 + a_{33}e_3) \\ &= (a_{11}x_1 + a_{12}x_2 + a_{13}x_3)e_1 + (a_{21}x_1 + a_{22}x_2 + a_{23}x_3)e_2 + \\ &\quad (a_{31}x_1 + a_{32}x_2 + a_{33}x_3)e_3 \\ &= \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= Ax. \end{aligned}$$

Consider now two linear transformations T and S with associated matrices (with respect to the standard basis of $\mathbb{R}^3$) $A = \begin{bmatrix} a_{ij} \end{bmatrix}$ and $B = \begin{bmatrix} b_{ij} \end{bmatrix}$, respectively. Then the product ST is a linear transformation whose associated matrix is $C = BA$ (the matrix product of B and A). Indeed, we have

$$Cx = ST(x) = S\left(T(x)\right) = S(Ax) = B(Ax) = (BA)x.$$

◇ **Exercise 20** Show that the matrix associated with a linear transformation is *nonsingular* (i.e., invertible).

Let $\mathsf{GL}(3, \mathbb{R})$ be the set of all nonsingular 3×3 matrices with real entries. Under the usual matrix multiplication, $\mathsf{GL}(3, \mathbb{R})$ is a *group*.

◇ **Exercise 21** Show that the group of all linear transformations on $\mathbb{R}^3$ is *isomorphic* to the group $\mathsf{GL}(3,\mathbb{R})$.

Either one of these groups is called the **general linear group**. Given a matrix $A \in \mathsf{GL}(3,\mathbb{R})$, the transformation $T,\ x \mapsto Ax$ is the only linear transformation whose associated matrix is A. We say that the matrix A *represents* the linear transformation T. It is convenient to *identify*

the linear transformation $T,\ x \mapsto Ax$ with the (nonsingular) matrix A.

Henceforth, the same symbol will be used to denote a linear transformation and its associated matrix. Thus, for instance, I will denote the identity transformation $x \mapsto x$ as well as the *identity matrix*

$$\left[\delta_{ij}\right] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

NOTE : The notation Ax stands for both the image of (the point) x under the linear transformation A and the matrix product of the (nonsingular) matrix A by the column matrix (vector) x.

Orthogonal transformations

Recall that the Euclidean 3-space $\mathbb{R}^3$ has a built-in inner product. Inner-product-preserving transformations form an important class of (linear) transformations.

1.2.2 DEFINITION. A linear transformation $A,\ x \mapsto Ax$ is an **orthogonal transformation** if it preserves the inner-product between any two vectors; that is, for all $x, y \in \mathbb{R}^3$,

$$Ax \bullet Ay = x \bullet y.$$

Let A and B be two orthogonal transformations. Then their product BA is also an orthogonal transformation. Indeed, for all vectors $x, y \in \mathbb{R}^3$,

$$(BA)x \bullet (BA)y = B(Ax) \bullet B(Ay) = Ax \bullet Ay = x \bullet y.$$

◇ **Exercise 22** Verify that the inverse of an orthogonal transformation is also an orthogonal transformation.

The set of all orthogonal transformations on $\mathbb{R}^3$ is a (transformation) *group*.

1.2.3 DEFINITION. A 3×3 matrix (with real entries) A is called **orthogonal** if

$$A^\top A = I,$$

where $A^\top$ is the *transpose* of A.

NOTE : If the matrix $A = \left[a_{ij}\right]$ is orthogonal, then (and only then)

$$a_{1i}a_{1j} + a_{2i}a_{2j} + a_{3i}a_{3j} = \delta_{ij}, \quad i,j = 1,2,3.$$

Thus the vectors (the columns of the matrix)

$$a_i := \begin{bmatrix} a_{1i} \\ a_{2i} \\ a_{3i} \end{bmatrix}, \quad i = 1,2,3$$

have unit length and are orthogonal to one another :

$$\|a_1\| = \|a_2\| = \|a_3\| = 1 \quad \text{and} \quad a_i \bullet a_j = 0 \quad (i \neq j).$$

(This can be written, in a more compact form, as $a_i \bullet a_j = \delta_{ij}, \quad i,j = 1,2,3.$)

Hence $\{a_1, a_2, a_3\}$ is an *orthonormal basis* for $\mathbb{R}^3$.

◇ **Exercise 23** Show that any orthogonal matrix is nonsingular.

◇ **Exercise 24** Let $A \in \mathsf{GL}(3,\mathbb{R})$. Show that

$$A^\top A = I \iff AA^\top = I \iff A^{-1} = A^\top.$$

Let $\mathsf{O}(3)$ be the set of all orthogonal matrices. Thus

$$\mathsf{O}(3) := \{A \in \mathsf{GL}(3,\mathbb{R}) \mid A^\top A = I\}.$$

◇ **Exercise 25** Show that $\mathsf{O}(3)$ is a *subgroup* of the general linear group $\mathsf{GL}(3,\mathbb{R})$.

1.2.4 PROPOSITION. *A linear transformation A, $x \mapsto Ax$ is an orthogonal transformation if and only if the matrix A is orthogonal.*

PROOF : ($\Rightarrow$) Suppose the transformation A, $x \mapsto Ax$ is orthogonal. Then we have

$$\begin{aligned} \delta_{ij} &= e_i \bullet e_j = Ae_i \bullet Ae_j \\ &= (Ae_i)^\top Ae_j = e_i^\top (A^\top A) e_j \\ &= (A^\top A)_{ij} \end{aligned}$$

and hence the matrix A is orthogonal.

($\Leftarrow$) Conversely, suppose the matrix A is orthogonal. Then

$$Ax \bullet Ay = (Ax)^\top Ay = x^\top (A^\top A) y = x^\top I y = x \bullet y$$

and thus the transformation $x \mapsto Ax$ is orthogonal. □

The group of all orthogonal transformations is *isomorphic* to the group $\mathsf{O}(3)$. Either one of these groups is called the **orthogonal group**. Those elements of $\mathsf{O}(3)$ which have determinant equal to $+1$ form a subgroup of $\mathsf{O}(3)$, denoted by $\mathsf{SO}(3)$ and called the **special orthogonal group**.

1.2.5 PROPOSITION. *An orthogonal transformation A, $x \mapsto Ax$ preserves the distance between any two points; that is, for all $x, y \in \mathbb{R}^3$,*

$$d(Ax, Ay) = d(x, y).$$

PROOF : First we show that A preserves norms. By definition, $\|x\|^2 = x \bullet x$ and hence

$$\|Ax\|^2 = Ax \bullet Ax = x \bullet x = \|x\|^2.$$

Thus $\|Ax\| = \|x\|$ for all (vectors) $x \in \mathbb{R}^3$. Since A is linear, it follows that

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

□

NOTE : The orthogonal groups $\mathsf{O}(2)$ and $\mathsf{O}(3)$ were first studied, by the number theorists of the eighteenth century, as the groups of transformations preserving the quadratic form $\xi_1^2 + \xi_2^2$ or $\xi_1^2 + \xi_2^2 + \xi_3^2$, respectively.

Rotations and reflections

If $A \in \mathsf{O}(2)$, then the columns of A are unit vectors and are orthogonal to one another. Suppose

$$A = \begin{bmatrix} a_1 & a_3 \\ a_2 & a_4 \end{bmatrix}.$$

Then the point (a_1, a_2) lies on the unit circle $\mathbb{S}^1$ giving

$$a_1 = \cos\theta \quad \text{and} \quad a_2 = \sin\theta$$

for some θ satisfying $0 \leq \theta < 2\pi$. As the vector $\begin{bmatrix} a_3 \\ a_4 \end{bmatrix}$ is at right angles to $\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$ and (as a point) also lies on the unit circle $\mathbb{S}^1$, we have

$$a_3 = \cos\varphi \quad \text{and} \quad a_4 = \sin\varphi$$

where either $\varphi = \theta + \frac{\pi}{2}$ or $\varphi = \theta - \frac{\pi}{2}$. In the first case we obtain

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

which is an element of $\mathsf{SO}(2)$ and represents a *rotation* about the origin (more precisely, a counterclockwise rotation about the origin through the angle θ). The second case gives

$$\begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$$

which has determinant -1 and represents a *reflection* in a line through the origin (more precisely, a reflection in a line through the origin at angle $\frac{\theta}{2}$ to the positive x_1-axis).

Therefore, *a 2×2 orthogonal matrix represents either a rotation of the plane about the origin or a reflection in a line through the origin, and the matrix has determinant $+1$ precisely when it represents a rotation.*

NOTE : The group $\mathsf{SO}(2)$ is often referred to as the *rotation group*. $\mathsf{SO}(2)$ is in fact the *unit circle* $\mathbb{S}^1$ in disguise. (Each point on the unit circle has the form $e^{i\theta}$, where $0 \leq \theta < 2\pi$ and hence corresponds to an angle.)

◇ **Exercise 26** Show that the mapping

$$e^{i\theta} \mapsto \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

is an *isomorphism* from $\mathbb{S}^1$ to $\mathsf{SO}(2)$. (Here $\mathbb{S}^1 = \{z \in \mathbb{C} \,|\, |z| = 1\}$ is considered as a subgroup of the multiplicative group $\mathbb{C}^\times$ of complex numbers.)

◇ **Exercise 27** Let

$$A_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad \text{and} \quad B_\varphi = \begin{bmatrix} \cos\varphi & \sin\varphi \\ \sin\varphi & -\cos\varphi \end{bmatrix}.$$

(a) Verify that

$$A_\theta A_\varphi = A_{\theta+\varphi}, \quad A_\theta B_\varphi = B_{\theta+\varphi}, \quad B_\theta A_\varphi = B_{\theta-\varphi}, \quad B_\theta B_\varphi = A_{\theta-\varphi}$$

where the angles in the matrices are read modulo 2π. Interpret these results geometrically.

(b) Work out the products

$$A_\theta B_\varphi A_\theta^{-1}, \quad B_\varphi A_\theta B_\varphi, \quad A_\theta B_\varphi A_\theta^{-1} B_\varphi.$$

Evaluate each of these when $\theta = \frac{\pi}{3}$ and $\varphi = \frac{\pi}{2}$.

Rotations of the Euclidean space about any one of the coordinate axes are similar to those of the plane (about the origin). The three basic types are (realized by the following orthogonal matrices) :

$$\begin{aligned} R(e_1,\theta) = R_1(\theta) &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \\ R(e_2,\theta) = R_2(\theta) &= \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \\ R(e_3,\theta) = R_3(\theta) &= \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

NOTE : The minus sign appears above the (main) diagonal in R_1 and R_3, but below the diagonal in R_2. This is *not* a "mistake" : it is due to the *orientation* of the positive x_1-axis with respect to the x_2x_3-plane. Clearly, $R_i(\theta) \in \mathsf{SO}(3)$, $i = 1, 2, 3$.

It can be shown that *any rotation* $x \mapsto Ax$ *of* $\mathbb{R}^3$ *which fixes the origin can be written as a product of just three of these elementary rotations :*

$$A = R_1(\theta)R_2(\varphi)R_3(\psi).$$

(The independent parameters θ, φ, ψ are called the *Euler angles* for the given rotation.)

It follows that

1.2.6 PROPOSITION. *Every rotation of* $\mathbb{R}^3$ *which fixes the origin can be represented by a matrix in* $\mathsf{SO}(3)$.

Now suppose that $A \in \mathsf{SO}(3)$. The *characteristic polynomial* $\mathrm{char}_A(\lambda) = \det(\lambda I - A)$ is cubic and therefore must have at least one real root. That is to say, A has a real eigenvalue. As the product of the eigenvalues of a matrix is the determinant of the matrix, we *see* that $+1$ is an eigenvalue of A.

◇ **Exercise 28** Show that every $A \in \mathsf{SO}(3)$ has an eigenvalue equal to $+1$.

NOTE : The other two eigenvalues are complex conjugate and have absolute value 1, so they can be written as $e^{i\theta}$ and $e^{-i\theta}$ for some $\theta \in \mathbb{R}$.

If w is a corresponding *eigenvector* (i.e., $Aw = w$), the line through the origin determined by w is invariant under (the linear transformation) A. Also since A preserves right angles, it must send the *plane* which is orthogonal to w, and which contains the origin, to itself.

◇ **Exercise 29** Check that the set (plane) $w^\perp = \{y \in \mathbb{E}^3 \mid y \bullet w = 0\}$ is invariant under (the orthogonal transformation) A; that is, $A(w^\perp) = w^\perp$.

Construct an *orthonormal* basis for $\mathbb{R}^3$ which has the unit vector $\frac{1}{\|w\|}w$ as first member. The matrix of $x \mapsto Ax$ with respect to this new basis will be

of the form

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & a_1 & a_3 \\ 0 & a_2 & a_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & R \end{bmatrix}.$$

Since $R \in \mathsf{SO}(2)$, $x \mapsto Ax$ is a *rotation* with axis determined by w.

Therefore, *each matrix in* $\mathsf{SO}(3)$ *represents a rotation of* $\mathbb{R}^3$ *about an axis which passes through the origin.*

NOTE : Every element (rotation) $A \in \mathsf{SO}(3)$ can be written as

$$\begin{aligned} A &= R(w, \theta) \\ &= P\, R(e_1, \theta)\, P^{-1} \end{aligned}$$

for some w, θ, and $P \in \mathsf{SO}(3)$. (We say that A and $R(e_1, \theta)$ are *conjugate* in $\mathsf{SO}(3)$.) The eigenvector w determines the *axis* of the rotation (i.e., the unique line through the origin which is left fixed). The *angle* of rotation is obtained from the other two eigenvectors. (In fact, θ is given by the eigenvalue $e^{i\theta}$.)

◇ **Exercise 30** Let $A = R(w, \theta) \in \mathsf{SO}(3)$.

(a) Show that

$$A - A^\top = \begin{bmatrix} 0 & c & b \\ -c & 0 & a \\ -b & -a & 0 \end{bmatrix} \quad \text{and} \quad w \in \ker\left(A - A^\top\right).$$

Hence deduce that (for $\theta \neq 0, \pi$)

$$w = \lambda \begin{bmatrix} -a \\ b \\ -c \end{bmatrix}.$$

(b) Show that

$$\operatorname{tr} A = 1 + 2\cos\theta.$$

(So we can solve for $\cos\theta$ from the trace of A. However, we don't know without further investigation if the rotation is clockwise or counterclockwise about w.)

◇ **Exercise 31** Show that the matrices

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2/3 & 1/3 & 2/3 \\ -2/3 & 2/3 & 1/3 \\ -1/3 & -2/3 & 2/3 \end{bmatrix}$$

both represent rotations, and then find axes and angles for these rotations.

◇ **Exercise 32** Let $A \in \mathsf{SO}(3)$ and $w \in \mathbb{E}^3$ such that $\|w\| = 1$. Show that, for all $x, y \in \mathbb{E}^3$ and $\theta \in \mathbb{R}$,

(a) $Ax \times Ay = A(x \times y)$;

(b) $w \bullet (R(w,\theta)x - x) = 0$;

(c) $A\,R(w,\theta)\,A^{-1} = R(Aw,\theta)$.

(HINT : The cross product $x \times y$ can be characterized as the unique vector such that

$$w \bullet x \times y = \det \begin{bmatrix} w & x & y \end{bmatrix}$$

for every vector w. See **Exercise 13**.)

NOTE : The group $\mathsf{SO}(3)$ is often referred to as the *rotation group*. $\mathsf{SO}(3)$ and the *sphere* $\mathbb{S}^3$ are *not* the "same" (i.e., they are not isomorphic groups). It is an interesting fact that $\mathbb{S}^0 = \{-1, 1\}$, $\mathbb{S}^1$, and $\mathbb{S}^3$ are the *only* spheres which can be groups.

If A lies in $\mathsf{O}(3)$ but *not* in $\mathsf{SO}(3)$, then $AS \in \mathsf{SO(3)}$ where

$$S = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

The matrix S represents a *reflection* in the x_1x_2-plane (identified with the Euclidean plane $\mathbb{R}^2$). We write

$$A = (AS)S.$$

As above, the transformation $x \mapsto ASx$ is a rotation. Consequently, A is a reflection (in the x_1x_2-plane) followed by a rotation.

◇ **Exercise 33** Complete the entries in the matrix

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \cdot \\ 0 & 1 & \cdot \\ -\frac{1}{\sqrt{2}} & 0 & \cdot \end{bmatrix}$$

to give an element of $\mathsf{SO}(3)$, and to give an element of $\mathsf{O}(3) \setminus \mathsf{SO}(3)$. Describe the (linear) trasformations represented by these matrices.

◇ **Exercise 34** Let $c \in \mathbb{R}^3$ such that $\|c\| = 1$. Prove that the correspondence

$$x \mapsto c \times x + (c \bullet x)c$$

defines an orthogonal transformation. Describe its general effect on $\mathbb{R}^3$.

1.3 Translations and Affine Transformations

Let $c \in \mathbb{R}^3$ be a vector and let T_c be the mapping that adds c to every point of $\mathbb{R}^3$. This mapping is one-to-one and onto and hence a transformation.

1.3.1 DEFINITION. The transformation

$$T_c : \mathbb{R}^3 \to \mathbb{R}^3, \quad x \mapsto x + c$$

is called the **translation** by vector c.

NOTE : A nonidentity translation is *not* a linear transformation.

◇ **Exercise 35** Show that given two points $p, q \in \mathbb{R}^3$, there exists a unique translation T such that $T(p) = q$.

The *inverse* of the translation $T_c,\ x \mapsto x + v$ is the translation $T_c^{-1},\ x \mapsto x - c$. Thus,

$$T_c^{-1} = T_{-c}.$$

◇ **Exercise 36** Verify that the product of two translations is also a translation.

The set of all translations on $\mathbb{R}^3$ is a (transformation) *group*. This group is *isomorphic* to the additive group (also denoted by $\mathbb{R}^3$) of (the vectors of) $\mathbb{R}^3$. Either one of these groups is called the **translation group**.

1.3.2 PROPOSITION. *A translation* $T = T_c,\ x \mapsto x + c$ *preserves the distance between any two points; that is, for all* $x, y \in \mathbb{R}^3$,

$$d(T(x), T(y)) = d(x, y).$$

PROOF : We have

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

□

1.3.3 DEFINITION. An **affine transformation** F on $\mathbb{R}^3$ is a linear transformation followed by a translation; that is, a transformation of the form

$$F : \mathbb{R}^3 \to \mathbb{R}^3, \quad F = TA$$

where $A,\ x \mapsto Ax$ is a linear transformation, and $T = T_c,\ x \mapsto x + c$ is a translation. A is called the *linear part* of F, and T the *translation part* of F.

For every $x \in \mathbb{R}^3$,

$$F(x) = Ax + c.$$

NOTE : The pair $(c, A) \in \mathbb{R}^3 \times \mathsf{GL}(3, \mathbb{R})$ *represents* the affine transformation $F,\ x \mapsto Ax + c$.

Affine transformations $F,\ x \mapsto Ax + c$, include the linear transformations (with $c = 0$) and the translations (with $A = I$). Let $F,\ x \mapsto Ax + c$ and $G,\ x \mapsto Bx + d$ be two affine transformations. Then (for $x \in \mathbb{R}^3$)

$$GF(x) = G(F(x)) = B(Ax + c) + d = (BA)x + Bc + d$$

and thus the product of G with F is also an affine transformation.

◇ **Exercise 37** Show that the inverse of an affine transformation is also an affine transformation.

The set of all affine transformations on $\mathbb{R}^3$ is a (transformation) *group*, which contains as subgroups the general linear group $\mathsf{GL}(3, \mathbb{R})$ and the translation group $\mathbb{R}^3$.

NOTE : Affine transformations preserve lines, parallelism, betweeness, and proportions on lines. Affine transformations can distort shapes. However, there is a limit to the amount of distortion : a convex set is always mapped to a convex set. (The converse holds as well : Transformations on $\mathbb{R}^3$ that preserve convexity are affine transformations.)

Any affine transformation (on $\mathbb{R}^3$) is represented by a pair $(c, A) \in \mathbb{R}^3 \times \mathsf{GL}(3, \mathbb{R})$ which we can write further as a 4×4 matrix $\begin{bmatrix} 1 & 0 \\ c & A \end{bmatrix}$. Call such a matrix an **affine matrix**. Let $\mathsf{GA}(3, \mathbb{R})$ be the set of all affine matrices. Thus

$$\mathsf{GA}(3, \mathbb{R}) := \left\{ \begin{bmatrix} 1 & 0 \\ c & A \end{bmatrix} \mid c \in \mathbb{R}^3,\ A \in \mathsf{GL}(3, \mathbb{R}) \right\}.$$

◇ **Exercise 38** Show that $\mathsf{GA}(3, \mathbb{R})$ is a *group*.

The group of all affine transformations on $\mathbb{R}^3$ is *isomorphic* to the group $\mathsf{GA}(3, \mathbb{R})$. Either of these groups is called the **general affine group**.

NOTE : In $\mathbb{R}^3$, of special interest are the affine transformations $x \mapsto Ax + c$ with $\det A = 1$. These transformations also form a group..

1.4 Isometries

Isometries (on Euclidean 3-space $\mathbb{R}^3$) are distance-preserving transformations on the *metric space* $\mathbb{R}^3$. They do not change the distance between points as the transformations move these points. Isometries are the dynamic counterpart to the Euclidean notion of *congruence*.

1.4.1 DEFINITION. A transformation $F : \mathbb{R}^3 \to \mathbb{R}^3$ is an **isometry** (or **rigid motion**) if it preserves the distance between any two points; that is, for all $x, y \in \mathbb{R}^3$,

$$d(F(x), F(y)) = d(x, y).$$

Orthogonal transformations and translations are isometries. If F is an isometry, then (for $x, y \in \mathbb{R}^3$)

$$d(F^{-1}(x), F^{-1}(y)) = d(FF^{-1}(x), FF^{-1}(y)) = d(x, y)$$

and thus the inverse F^{-1} is also an isometry.

◇ **Exercise 39** Verify that the product of two isometries is also an isometry.

The set of all isometries on $\mathbb{R}^3$ is a (transformation) *group*, which contains as subgroups the orthogonal group $\mathsf{O}(3)$ and the translation group $\mathbb{R}^3$.

1.4.2 PROPOSITION. *If F is an isometry on $\mathbb{R}^3$ such that $F(0) = 0$, then F is an orthogonal transformation.*

PROOF : For any (vector) $x \in \mathbb{R}^3$,

$$\|F(x)\| = d(0, F(x)) = d(F(0), F(x)) = d(0, x) = \|x\|.$$

Let $x, y \in \mathbb{R}^3$. Then we have

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

which implies

$$(F(x) - F(y)) \bullet (F(x) - F(y)) = (x - y) \bullet (x - y)$$

or

$$\|F(x)\|^2 - 2F(x) \bullet F(y) + \|F(y)\|^2 = \|x\|^2 - 2x \bullet y + \|y\|^2.$$

Thus we have

$$F(x) \bullet F(y) = x \bullet y$$

so that F preserves the inner product of any two vectors.

It remains to prove that F is a linear transformation. Let $x \in \mathbb{R}^3$. Then (with respect to the standard basis)

$$x = x_1 e_1 + x_2 e_2 + x_3 e_3.$$

Since $\{e_1, e_2, e_3\}$ is an orthonormal basis (and F preserves the inner product of any two vectors), it follows that $\{F(e_1), F(e_2), F(e_3)\}$ is also an orthonormal basis so that

$$F(x) = \bar{x}_1 F(e_1) + \bar{x}_2 F(e_2) + \bar{x}_3 F(e_3).$$

Taking the inner product of both sides with $F(e_i)$, we get

$$\bar{x}_i = F(x) \bullet F(e_i) = x \bullet e_i = x_i, \quad i = 1, 2, 3.$$

Hence

$$F(x) = x_1 F(e_1) + x_2 F(e_2) + x_3 F(e_3)$$

and we can easily check the linearity conditions (L_1) and (L_2). □

1.4.3 THEOREM. *If F is an isometry on $\mathbb{R}^3$, then there exists a unique orthogonal transformation A, $x \mapsto Ax$ and a unique translation $T = T_c$, $x \mapsto x + c$ such that*

$$F = TA.$$

A is called the orthogonal part *of F, and T the* translation part *of F.*

PROOF : Let T be the translation by vector $c = F(0)$. Then T^{-1} is the translation by vector $-c = -F(0)$, and so $T^{-1}F$ is an isometry. Furthermore,

$$T^{-1}F(0) = T^{-1}(F(0)) = F(0) - F(0) = 0.$$

Thus $T^{-1}F$ is an orthogonal transformation, say $T^{-1}F = A$, from which follows immediately $F = TA$.

To prove the required uniqueness, suppose $F = \bar{T}\bar{A}$, where $\bar{T}$ is a translation and $\bar{A}$ is an orthogonal transformation. Then

$$TA = \bar{T}\bar{A}$$

so that $A = T^{-1}\bar{T}\bar{A}$. Since A and $\bar{A}$ are linear transformations, $A(0) = \bar{A}(0) = 0$. It follows that $T^{-1}\bar{T} = I$ (the identity transformation), so that $\bar{T} = T$, which implies $\bar{A} = A$. □

NOTE : We see that an isometry on $\mathbb{R}^3$ is a special affine transformation. Intermediate between isometries and affine transformations are *similarities*, the transformations corresponding to similar figures. Similarities preserve betweeness, segments, angle measure, and the proportions of all distances. The set of all similarities on $\mathbb{R}^3$ is a subgroup $\mathsf{GE}(3)$ of the general affine group, called the *general Euclidean group*. A similarity that is not an isometry is either a *dilation* or a *dilative rotation (spiral)*.

The group of all isometries on $\mathbb{R}^3$ is (*isomorphic* to) a subgroup of the general affine group $\mathsf{GA}(3,\mathbb{R})$, denoted by $\mathsf{E}(3)$. We have

$$\mathsf{E}(3) = \left\{ \begin{bmatrix} 1 & 0 \\ c & A \end{bmatrix} \mid c \in \mathbb{R}^3, \;\; A \in \mathsf{O}(3) \right\}.$$

Either one of these groups is called the **Euclidean group**.

NOTE : The Euclidean group $\mathsf{E}(3)$ is generated by *reflections*. Each isometry on $\mathbb{R}^3$ is exactly one of the following : *translation, rotation, glide rotation (screw), reflection, glide reflection*, or *rotary reflection*. In the case of the plane, an element of $\mathsf{E}(2)$ is exactly one of the following : *translation, rotation, reflection,* or *glide reflection*.

Orientation

We now come to one of the most interesting and elusive ideas in geometry. Intuitively, it is *orientation* that distinguishes between a right-handed glove and a left-handed glove in ordinary space. We shall not formalize this concept now.

NOTE : To handle the concept of *orientation* mathematically, we replace "gloves" by orthonormal bases (in fact, *frames*) and separate all these orthonormal bases of $\mathbb{R}^3$ into two classes : positively-oriented (or right-handed) and negatively-oriented (or left-handed).

Let $F,\ x \mapsto Ax + c$ be an isometry on $\mathbb{R}^3$. Since (the matrix) A is orthogonal, its determinant is either $+1$ or -1. We define the **sign** of F to be the determinant of A, with notation

$$\operatorname{sgn} F := \det A.$$

1.4.4 DEFINITION. An isometry $F,\ x \mapsto Ax + c$ is said to be

- **direct** (or **orientation-preserving**) if $\operatorname{sgn} F = +1$;
- **opposite** (or **orientation-reversing**) if $\operatorname{sgn} F = -1$.

All translations are orientation-preserving. Intuitively this is clear. In fact, the orthogonal part of a translation T is just the identity transformation I, and so $\operatorname{sgn} T = \det I = +1$.

$\diamond$ **Exercise 40** Consider the orthogonal transformation $R_1(\theta)$ represented by the matrix

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

Show that $R_1(\theta)$ is orientation-preserving.

1.4.5 EXAMPLE. One can (literally) see reversal of orientation by using a mirror. Suppose the x_2x_3-plane of $\mathbb{R}^3$ is the mirror. If one looks toward the plane, the point $p = (p_1, p_2, p_3)$ appears to be located at the point

$$S(p) = (-p_1, p_2, p_3).$$

The transformation S, $p \mapsto S(p)$ is the *reflection* in the x_2x_3-plane. Evidently, S is an orthogonal transformation represented by the (orthogonal) matrix

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus S is an orientation-reversing isometry, as confirmed by the experimental fact that the mirror image of the right hand is a left hand.

Recall that an isometry is also called a *rigid motion*. If this is the case, a direct isometry is referred to as a **proper rigid motion**. The set of all direct isometries (or proper rigid motions) on $\mathbb{R}^3$ is a (transformation) *group*. This group is (*isomorphic* to) a subgroup of the Euclidean group $\mathsf{E}(3)$, denoted by $\mathsf{SE}(3)$. We have

$$\mathsf{SE}(3) := \left\{ \begin{bmatrix} 1 & 0 \\ c & A \end{bmatrix} \mid c \in \mathbb{R}^3, \;\; A \in \mathsf{SO}(3) \right\}.$$

Either one of these groups is called the **special Euclidean group**.

NOTE : The orientation-preserving isometries on $\mathbb{R}^3$ are precisely the *translations, rotations,* and *glide rotations (screws).* In the case of the plane, the elements of the special Euclidean group $\mathsf{SE}(2)$ are the *translations* and the *rotations.*

1.5 Galilean Transformations

Galilean spacetime

Newtonian mechanics takes place in a *Galilean spacetime.* Let $\mathbb{R}^3$ be the Euclidean 3-space and let $\mathbb{R}\times\mathbb{R}^3$ denote the (standard) **Galilean spacetime**. Elements of $\mathbb{R}\times\mathbb{R}^3$ are called **events**.

NOTE : $\mathbb{R}\times\mathbb{R}^3$ is a *model* for spatio-temporal world of Newtonian mechanics. The Newtonian world is comprised of objects sitting in a Universe (i.e., a Galilean spacetime) and interacting with one another in a way consistent with the *Galilean relativity principle* (which states that for a closed system in Galilean spacetime the governing physical laws are invariant under Galilean transformations). In particular, determinacy principle says that to "see" what will happen in the Universe, one need only specify initial conditions for the ODEs of Newtonian mechanics, and all else follows, at least in principle.

Given two events $\xi = (t, x) = (t, (x_1, x_2, x_3))$ and $\xi' = (t', x') = (t', (x'_1, x'_2, x'_3))$, the *time* between these events is

$$\mathfrak{t}(\xi, \xi') := t' - t.$$

The *distance* between **simultaneous events** (t, x) and (t, x') is then

$$\mathfrak{d}((t, x), (t, x')) := \|x' - x\| = \sqrt{(x'_1 - x_1)^2 + (x'_2 - x_2)^2 + (x'_3 - x_3)^2},$$

where $\|\cdot\|$ is the (standard) Euclidean norm on $\mathbb{R}^3$.

NOTE : Distance between events that are *not* simultaneous *cannot* be measured. In particular, it does not make sense to talk about two non-simultaneous events as ocurring in the same place (i.e., as separated by zero distance). The picture one should have in mind for a Galilean spacetime is of it being *a union of simultaneous events, nicely stacked together.* We write

$$\mathbb{R}\times\mathbb{R}^3 = \bigcup_{t\in\mathbb{R}} \{t\}\times\mathbb{R}^3 := \bigcup_{t\in\mathbb{R}} \mathbb{R}^3_t.$$

That one cannot measure distance between non-simultaneous events reflects there being *no* natural direction transverse to the *stratification* by simultaneous events.

Galilean transformations

Galilean transformations are structure-preserving transformations on the Galilean spacetime. They preserve simultaneity of events and do not change the distance between simultaneous events.

1.5.1 Definition. An affine transformation $F : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R} \times \mathbb{R}^3$ is a **Galilean transformation** if it preserves the time between any two events and the distance between any two simultaneous events; that is, for all $\xi, \xi' \in \mathbb{R} \times \mathbb{R}^3$,

$$\mathfrak{t}\,(F(\xi), F(\xi')) = \mathfrak{t}\,(\xi, \xi')$$

and, for all $t \in \mathbb{R}$ and $\xi, \xi' \in \mathbb{R}^3_t$,

$$\mathfrak{d}\,(F(\xi), F(\xi')) = \mathfrak{d}\,(\xi, \xi').$$

Let $F,\ (t, x) \mapsto A(t, x) + (\zeta, c)$ be a Galilean transformation. Let us write

$$\begin{aligned} A(t,x) + (\zeta, c) &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} t \\ x \end{bmatrix} + \begin{bmatrix} \zeta \\ c \end{bmatrix} \\ &= \begin{bmatrix} A_{11}t + A_{12}x + \zeta \\ A_{21}t + A_{22}x + c \end{bmatrix} \\ &= (A_{11}t + A_{12}x + \zeta, A_{21}t + A_{22}x + c) \end{aligned}$$

where $A_{11} \in \mathbb{R}$ and $A_{22} \in \mathsf{GL}\,(3, \mathbb{R})$.

◇ **Exercise 41** Show that if

$$(t, x) \mapsto (A_{11}t + A_{12}x + \zeta,\ A_{21}t + A_{22}x + c)$$

is a Galilean transformation, then

$$A_{11} = 1, \quad A_{12} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}, \quad A_{21} = v \in \mathbb{R}^{3\times 1}, \quad A_{22} \in \mathsf{O}\,(3).$$

Any Galilean transformation

$$(t, x) \mapsto (t + \zeta,\ Rx + tv + c)$$

where $\zeta \in \mathbb{R},\ c, v \in \mathbb{R}^{3\times 1}$, and $R \in \mathsf{O}(3)$, may be written in matrix form as

$$\begin{bmatrix} t \\ x \end{bmatrix} \mapsto \begin{bmatrix} 1 & 0 \\ v & R \end{bmatrix} \begin{bmatrix} t \\ x \end{bmatrix} + \begin{bmatrix} \zeta \\ c \end{bmatrix}.$$

◇ **Exercise 42** Show that the set of all Galilean transformations is a (transformation) *group*.

The following basic Galilean transformations

- $(t, x) \mapsto (t + \zeta, x + c)$ (*shift* of origin);
- $(t, x) \mapsto (t, x + tv)$ (velocity *boost*);
- $(t, x) \mapsto (t, Rx)$ ("rotation" of reference frame)

can be used to *generate* the whole set (group) of Galilean transformations.

NOTE : The names given to these basic Galilean transformations are suggestive. A shift of the origin (in fact a spacetime translation) may be thought of as moving the origin to a new position and resetting the clock, but maintaining the same orientation in space. A (Galilean) velocity boost means the origin maintains its "orientation" and uses the same clock, but now moves with a certain velocity with respect to the previous origin. Finally, the "rotation" of reference frame (in fact an orthogonal transformation or linear isometry) means the origin stays in the same place and uses the same clock, but rotates the "point of view".

Any Galilean transformation is represented by a quadruple $(\zeta, c, v, R) \in \mathbb{R} \times \mathbb{R}^3 \times \mathbb{R}^3 \times \mathsf{O}(3)$ which we can write further as a 5×5 matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ \zeta & 1 & 0 \\ c & v & R \end{bmatrix}.$$

Let Gal be the set of all such matrices. Thus

$$\mathsf{Gal} := \left\{ \begin{bmatrix} 1 & 0 & 0 \\ \zeta & 1 & 0 \\ c & v & R \end{bmatrix} \mid \zeta \in \mathbb{R},\ c, v \in \mathbb{R}^3,\ R \in \mathsf{O}(3) \right\}.$$

◇ **Exercise 43** Show that Gal is a *group*.

The group of all Galilean transformations is *isomorphic* to the group Gal. Either one of these groups is called the **Galilean group**

We saw that the elements of Gal are products of spacetime translations, velocity boosts, and spatial orthogonal transformations (in particular, rotations). Various *subgroups* of Gal are of particular interest in applications (including some familiar transformation groups). For instance,

- the subgroup of *isochronous* Galilean transformations consists of those Galilean transformations (represented by the quadruple (ζ, c, v, R)) for which $\zeta = 0$;
- the subgroup of *unboosted* Galilean transformations consists of those Galilean transformations (represented by the quadruple (ζ, c, v, R)) for which $v = 0$;
- the subgroup of *anisotropic* Galilean transformations consists of those Galilean transformations (represented by the quadruple (ζ, c, v, R)) for which $R = I$;
- the subgroup of *homogeneous* Galilean transformations consists of those Galilean transformations (represented by the quadruple (ζ, c, v, R)) for which $\zeta = 0,\ c = 0$.

◇ **Exercise 44** Identify the following subgroups of Gal.

(a) The subgroup of Gal consisting of those Galilean transformations (represented by the quadruple (ζ, c, v, R)) for which $\zeta = 0,\ v = 0$.

(b) The subgroup of Gal consisting of those Galilean transformations (represented by the quadruple (ζ, c, v, R)) for which $v = 0,\ R = I$.

(c) The subgroup of Gal consisting of those Galilean transformations (represented by the quadruple $\zeta, c, v, R)$) for which $\zeta = 0,\ v = 0,\ R = I$.

(d) The subgroup of Gal consisting of those Galilean transformations (represented by the quadruple (ζ, c, v, R)) for which $c = 0,\ v = 0,\ R = I$.

1.6 Lorentz Transformations

Minkowski spacetime

The geometric setting for EINSTEIN's *Special Theory of Relativity* is provided by *Minkowski spacetime.*

NOTE : A *spacetime* is simply the mathematical version of a universe that, like our own physical universe, has dimensions both of space and of time. A *flat* spacetime is a spacetime with no *gravity*, since gravitation tends to "bend" spacetime. Flat spacetimes are the simplest kind of spacetimes; they stand in the same relation to curved spacetimes as a flat Euclidean plane does to a curved surface.

We make the following definition.

1.6.1 DEFINITION. The (standard) **Minkowski spacetime** $\mathbb{R}^{1,3}$ is the vector space $\mathbb{R}^4$ together with the *Minkowski product* between vectors $v = (v_0, v_1, v_2, v_3)$ and $w = (w_0, w_1, w_2, w_3)$ given by

$$v \odot w := -v_0 w_0 + v_1 w_1 + v_2 w_2 + v_3 w_3.$$

The Minkowski product is an inner product; that is, it has the following three properties (for $v, v', w \in \mathbb{R}^{1,3}$ and $\lambda, \lambda' \in \mathbb{R}$) :

(IP1) $(\lambda v + \lambda' v') \odot w = \lambda(v \odot w) + \lambda'(v' \odot w)$;

(IP2) $v \odot w = w \odot v$;

(IP4) $v \odot w = 0$ for all v implies $w = 0$.

We can write (for $v, w \in \mathbb{R}^{1,3}$)

$$v \odot w = v^\top Q w$$

where

$$Q = \operatorname{diag}(-1, 1, 1, 1) = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The elements (vectors) of $\mathbb{R}^{1,3}$ are also called **events**.

NOTE : We can use t in place of v_0 since in *relativity theory* this coordinate is related to the time measurements while the others are related to the spatial ones. Hence we can write elements of the Minkowski spacetime $\mathbb{R}^{1,3}$ in the form

$$\xi = (t, x) = \begin{bmatrix} t \\ x \end{bmatrix} \qquad (t \in \mathbb{R},\ x \in \mathbb{R}^3).$$

Then (for $\xi = (t, x)$ and $\xi' = (t', x')$) $\xi \odot \xi' = -tt' + x \bullet x'$.

Two vectors $v, w \in \mathbb{R}^{1,3}$ are **Minkowski-orthogonal** provided $v \odot w = 0$.

1.6.2 EXAMPLE. Since the Minkowski product is *not* positive definite, there exist nonzero elements (vectors) $v \in \mathbb{R}^{1,3}$ for which $v \odot v = 0$. For instance, such a vector is $v = (1, 0, 1, 0)$. Such vectors are said to be *null* and $\mathbb{R}^{1,3}$ actually has bases which consist exclusively of this type of vector. A *null basis* cannot consist of mutually (Minkowski-)orthogonal vectors, however.

◇ **Exercise 45** Show that two null vectors v, w are Minkowski-orthogonal if and only if they are linearly dependent (i.e., $v = \lambda w$ for some $\lambda \in \mathbb{R}$).

We make the following definitions (this terminology derives from *relativity theory*).

1.6.3 DEFINITION. A nonzero vector $v \in \mathbb{R}^{1,3}$ is called

- **spacelike** provided $v \odot v > 0$;
- **timelike** provided $v \odot v < 0$;
- **null** (or **lightlike**) provided $v \odot v = 0$.

◇ **Exercise 46** Show that if a nonzero vector is Minkowski-orthogonal to a timelike vector, then it must be spacelike.

NOTE : Let $\mathcal{Q}$ denote the quadratic form associated with the Minkowski product on $\mathbb{R}^{1,3}$; that is, the mapping

$$\mathcal{Q} : \mathbb{R}^{1,3} \to \mathbb{R}, \qquad v \mapsto v \odot v.$$

Consider two distinct events ξ and ξ_0 for which the *displacement vector* $v := \xi - \xi_0$ from ξ_0 to ξ is null (i.e., $\mathcal{Q}(\xi - \xi_0) = 0$). Then we can define the *null cone* (or *light cone* $\mathcal{C}_N(\xi_0)$) at ξ_0 by

$$\mathcal{C}_N(\xi_0) := \left\{\xi \in \mathbb{R}^{1,3} \,|\, \mathcal{Q}(\xi - \xi_0) = 0\right\}.$$

$\mathcal{C}_N(\xi_0)$ consists of all those events in $\mathbb{R}^{1,3}$ that are "connectible to ξ_0 by a light ray".

Let $\mathcal{T}$ denote the collection of all timelike vectors in $\mathbb{R}^{1,3}$ and define a relation $\sim$ on $\mathcal{T}$ as follows :

$$v \sim w \iff v \odot w < 0.$$

This is an equivalence relation and hence $\mathcal{T}$ is the union of two disjoint subsets (equivalence classes) $\mathcal{T}^+$ and $\mathcal{T}^-$, called *time cones*, and there is no intrinsic way to distinguish one from the other. We think of the elements of $\mathcal{T}^+$ (and $\mathcal{T}^-$) as having the same time orientation. More specifically, we select (arbitrarily) $\mathcal{T}^+$ and refer to its elements as *future-directed* timelike vectors, whereas the vectors in $\mathcal{T}^-$ we call *past-directed*.

For each ξ_0 in $\mathbb{R}^{1,3}$ we define the *time cone* $\mathcal{C}_T(\xi_0)$, *future time cone* $\mathcal{C}_T^+(\xi_0)$, and *past time cone* $\mathcal{C}_T^-(\xi_0)$ at ξ_0 by

$$\begin{aligned}
\mathcal{C}_T(\xi_0) &:= \left\{\xi \in \mathbb{R}^{1,3} \,|\, \mathcal{Q}(\xi - \xi_0) < 0\right\} \\
\mathcal{C}_T^+(\xi_0) &:= \left\{\xi \in \mathbb{R}^{1,3} \,|\, \xi - \xi_0 \in \mathcal{T}^+\right\} = \mathcal{C}_T(\xi_0) \cap \mathcal{T}^+ \\
\mathcal{C}_T^-(\xi_0) &:= \left\{\xi \in \mathbb{R}^{1,3} \,|\, \xi - \xi_0 \in \mathcal{T}^-\right\} = \mathcal{C}_T(\xi_0) \cap \mathcal{T}^-.
\end{aligned}$$

We picture $\mathcal{C}_T(\xi_0)$ as the interior of the null cone $\mathcal{C}_N(\xi_0)$. It is the (disjoint) union of $\mathcal{C}_T^+(\xi_0)$ and $\mathcal{C}_T^-(\xi_0)$.

The notion of time-orientation can be extended to null vectors. We say that a null vector n is *future-directed* if $n \odot v < 0$ for all $v \in \mathcal{T}^+$ and *past-directed* if $n \odot v > 0$ for all $v \in \mathcal{T}^+$. For any event ξ_0 we define the *future null cone* $\mathcal{C}_N^+(\xi_0)$ and the *past null cone* $\mathcal{C}_N^-(\xi_0)$ at ξ_0 by

$$\begin{aligned}
\mathcal{C}_N^+(\xi_0) &:= \{\xi \in \mathcal{C}_N(\xi_0) \,|\, \xi - \xi_0 \text{ is future-directed}\} \\
\mathcal{C}_N^-(\xi_0) &:= \{\xi \in \mathcal{C}_N(\xi_0) \,|\, \xi - \xi_0 \text{ is past-directed}\}.
\end{aligned}$$

Physically, event ξ is in $\mathcal{C}_N^+(\xi_0)$ if ξ_0 and ξ can be regarded as the emission and reception of a light signal, respectively. Consequently, $\mathcal{C}_N^+(\xi_0)$ may be thought of as the history in spacetime of a spherical electromagnetic wave (photons in all directions) whose emission event is ξ_0.

For a vector $v = (v_0, v_1, v_2, v_3) \in \mathbb{R}^{1,3}$ we write

$$\|v\| := \sqrt{v \odot v} = \sqrt{\left|-v_0^2 + v_1^2 + v_2^2 + v_3^2\right|}$$

and call it the **Minkowski norm** (or **length**) of v. A *unit vector* is a vector v with Minkowski norm 1 : $v \odot v = \pm 1$.

NOTE : This is a funny kind of "length" since null vectors have zero length (even though they are not zero). For any timelike vector v, the Minkowski norm $\|v\|$ is commonly referred to as the *duration* of v. If $v = \xi - \xi_0$ is the displacement vector between two events ξ, ξ_0, then $\|v\|$ is to be interpreted physically as the *time separation* of ξ_0 and ξ (in any admissible frame of reference in which both events occur at the same spatial location).

Many features of Euclidean 3-space $\mathbb{R}^3$ (which is a positive definite inner product space) have counter-intuitive analogues in the Minkowski case. For example, analogues of the basic inequalities (like the Cauchy-Schwarz inequality and the triangle inequality) are generally reversed.

◇ **Exercise 47** Show that if v and w are timelike vectors, then

$$(v \odot w)^2 \geq (v \odot v)(w \odot w)$$

and equality holds if and only if v and w are linearly dependent.

1.6.4 PROPOSITION. *Let v and w be timelike vectors in the same time cone (i.e., with the same time orientation : $v \odot w < 0$). Then*

$$\|v + w\| \geq \|v\| + \|w\|$$

and equality holds if and only if v and w are linearly dependent.

PROOF : Since $v \odot v < 0$, $v + w \in \mathcal{T}$ and (by **Exercise 47**)

$$\|v\| \, \|w\| \leq -v \odot w.$$

Hence

$$\begin{aligned} (\|v\| + \|w\|)^2 &= \|v\|^2 + 2\|v\|\|w\| + \|w\|^2 \\ &\leq \|v\|^2 - 2v \odot w + \|w\|^2 \\ &\leq -(v + w) \odot (v + w) \\ &= \|v + w\|^2 \end{aligned}$$

and the equality holds if and only if $\|v\|\,\|w\| = -v \odot w$. The conclusion now follows from **Exercise 47**. □

Lorentz transformations

Lorentz transformations are structure-preserving transformations on the Minkowski spacetime.

1.6.5 DEFINITION. A linear transformation (on $\mathbb{R}^{1,3}$) $L,\ v \mapsto Lv$ is an **orthogonal transformation** if it preserves the Minkowski product between any two vectors; that is, for all $v, w \in \mathbb{R}^{1,3}$,

$$Lv \odot Lw = v \odot w.$$

The set of all orthogonal transformations on $\mathbb{R}^{1,3}$ is a (transformation) *group*.

◇ **Exercise 48** Let $L : \mathbb{R}^{1,3} \to \mathbb{R}^{1,3}$ be a linear transformation. Then show that the following statements are equivalent :

(a) L is an orthogonal transformation.

(b) L preserves the quadratic form on $\mathbb{R}^{1,3}$ (i.e., $\mathcal{Q}(Lv) = \mathcal{Q}(v)$ for all $v \in \mathbb{R}^{1,3}$).

(c) (The matrix of) L satisfies the condition

$$L^\top Q L = Q$$

where $Q = \operatorname{diag}(-1, 1, 1, 1)$. (HINT : To prove that $(b) \Rightarrow (a)$ compute $L(v+w) \odot L(v+w) - L(v-w) \odot L(v-w)$.)

Any such linear transformation $L,\ v \mapsto Lv$ on $\mathbb{R}^{1,3}$ is called a **general (homogeneous) Lorentz transformation**.

NOTE : If $L = \left[l_{ij}\right]$ is a 4×4 matrix such that $L^\top Q L = Q$, where $Q = \operatorname{diag}(-1, 1, 1, 1)$, then its columns are *mutually Minkowski-orthogonal unit vectors*.

Let Lor_{GH} be the set of all such 4×4 matrices (i.e. matrices representing general homogeneous Lorentz transformations). Thus

$$\mathsf{Lor}_{GH} := \{L \in \mathsf{GL}(4, \mathbb{R}) \mid L^\top Q L = Q\}.$$

◇ **Exercise 49** Show that Lor_{GH} is a *subgroup* of the general linear group $\mathsf{GL}(4,\mathbb{R})$.

The group of all (Minkowski-)orthogonal transformations is *isomorphic* to the group Lor_{GH}. Either one of these groups is called the **general (homogeneous) Lorentz group.**

Let $L = \left[l_{ij}\right] \in \mathsf{Lor}_{GH}$ (i.e., $L^\top Q L = Q$). Then, in particular, we have

$$l_{11}^2 = 1 + l_{12}^2 + l_{13}^2 + l_{14}^2 \geq 1$$

so that

$$l_{11} \geq 1 \quad \text{or} \quad l_{11} \leq -1.$$

L is said to be *orthocronous* if $l_{11} \geq 1$ and *nonorthocronous* if $l_{11} \leq -1$. Nonorthocronous Lorentz transformations have certain "unsavory" characteristics; for instance, they always *reverse* time orientation (and so presumably relate reference frames in which someone's clock is running backwards). For this reason, *it is commom practice to restrict attention to the orthocronous elements of* Lor_{GH}.

There is yet one more restriction we would like to impose on our Lorentz transformations.

◇ **Exercise 50** Show that if $L \in \mathsf{Lor}_{GH}$, then

$$\det L = 1 \quad \text{or} \quad \det L = -1.$$

We shall say that a Lorentz transformation $v \mapsto Lv$ is *proper* if $\det L = 1$ and *improper* if $\det L = -1$. The set Lor of all proper, orthocronous Lorentz transformations is a *subgroup* of Lor_{GH}. Generally, we shall refer to Lor simply as the **Lorentz group** and its elements as Lorentz transformations with the understanding that they are all proper and orthocronous.

NOTE : Ocasionally, it is convenient to enlarge the group Lor to include spacetime translations, thereby obtaining the so-called **inhomogeneous Lorentz group** (or **Poincaré group**). Physically, this amounts to allowing "admissible" observers to use different spacetime origins.

The Lorentz group Lor has an important subgroup consisting of those elements of the form

$$R = \begin{bmatrix} 1 & 0 \\ 0 & A \end{bmatrix}$$

where $A \in \mathsf{SO}(3)$ (i.e., $A^\top = A^{-1}$ and $\det A = 1$). Such elements are called (spatial) *rotations* (in Lor).

A Lorentz transformation $v \mapsto L(\beta)v$ of the form

$$L(\beta) := \begin{bmatrix} \gamma & 0 & 0 & -\beta\gamma \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\beta\gamma & 0 & 0 & \gamma \end{bmatrix}$$

where $-1 < \beta < 1$ (and $\gamma := \frac{1}{\sqrt{1-\beta^2}} \geq 1$) is called a **special Lorentz transformation**. The matrix $L(\beta)$ is often called a (Lorentz) *boost* in the x_1-direction.

NOTE : Likewise, one can define matrices (representing) boosts in the x_2- and x_3-directions. One can also define a boost in an arbitrary direction by first rotating, say, the positive x_1-axis into that direction and then applying $L(\beta)$.

◇ **Exercise 51** Suppose $-1 < \beta_1 \leq \beta_2 < 1$. Show that :

(a) $\left|\dfrac{\beta_1 + \beta_2}{1 + \beta_1\beta_2}\right| < 1$.

(b) $L(\beta_2)L(\beta_1) = L\left(\dfrac{\beta_1 + \beta_2}{1 + \beta_1\beta_2}\right)$.

(HINT : Show that if a is a constant, then the function $x \mapsto \dfrac{x + a}{1 + ax}$ is *increasing* for $-1 \leq x \leq 1$.)

It follows from **Exercise 51** that the product of two boosts in the x_1-direction is another boost in the x_1-direction. Since $L(\beta)^{-1} = L(-\beta)$, *the collection of all such special Lorentz transformations forms a subgroup of* Lor. We point out, however, that the product of two boosts in two different directions is, in general, *not* equivalent to a single boost in any direction.

NOTE : A simple computation shows that if we put $\beta = \tanh\theta$, then the Lorentz transformation $L(\beta)$ takes the *hyperbolic form* :

$$L(\theta) = \begin{bmatrix} \cosh\theta & 0 & 0 & -\sinh\theta \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\sinh\theta & 0 & 0 & \cosh\theta \end{bmatrix}.$$

It is remarkable that *all of the physically interesting behaviour of (proper, orthochronous) Lorentz transformations is exhibited by the special Lorentz transformations* : any element of Lor differs from some $L(\beta)$ only by at most two rotations (in Lor); that is, for $L \in \mathsf{Lor}$ there is some (real) number θ and (spatial) rotations $R_1, R_2 \in \mathsf{Lor}$, such that

$$L = R_1 L(\theta) R_2.$$

Chapter 2

Curves

Topics :

1. Tangent Vectors and Frames
2. Directional Derivatives
3. Curves in Euclidean 3-Space $\mathbb{R}^3$
4. Serret-Frenet Formulas
5. The Fundamental Theorem for Curves
6. Some Remarks

2.1 Tangent Vectors and Frames

Tangent vectors

The basic method used to investigate *curves* (in Euclidean 3-space $\mathbb{R}^3$) consists in assigning at each point (along the curve) a certain *frame* (i.e., a set of three mutually orthogonal unit vectors) and then express the rate of change of the frame in terms of the frame itself. In a real sense, the geometry of curves is merely a corollary of these basic results.

NOTE : A frame consists of vectors *located* at some specific point. These vectors are *not* free vectors (viewed as translations) but *fixed* vectors. We need to make this distinction precise by "re-thinking" the representation of a geometric vector. To obtain a concept that is both practical and precise, we shall describe an "arrow" by giving the starting (fixed) point p and the change (vector) v, necessary to reach its terminal point $p + v$.

We make the following definition.

2.1.1 DEFINITION. A **tangent vector** to $\mathbb{R}^3$ at a point p, denoted by v_p, is an ordered pair (p, v). p is the *point of application* of v_p, and v is the *vector part*.

NOTE : $p + v$ is considered as the position vector of a point.

We shall always picture v_p as the arrow (directed line segment) from the point p to the point $p + v$.

2.1.2 EXAMPLE. If $p = (1, 1, 3)$ and $v = (2, 3, 2)$ (in fact, $v = \begin{bmatrix} 2 \\ 3 \\ 2 \end{bmatrix}$), then v_p "runs" from $(1, 1, 3)$ to $(3, 4, 5)$.

We emphasize that tangent vectors v_p and w_q are *equal* if and only if they have the same vector part and the same point of application :

$$v_p = w_q \iff (v = w \quad \text{and} \quad p = q).$$

Tangent vectors v_p and v_q with the same vector part, but different points of application, are said to be **parallel**.

NOTE : It is essential to recognize that v_p and v_q are *different* tangent vectors if $p \neq q$. In physics, the concept of moment of a force shows this clearly : the same force v applied at different points p and q of a rigid body can produce quite different rotational effects.

Let p be a point of $\mathbb{R}^3$. The set $T_p\mathbb{R}^3$ of all tangent vectors to $\mathbb{R}^3$ at p is called the **tangent space** of $\mathbb{R}^3$ at p.

NOTE : $\mathbb{R}^3$ has a *different* tangent space at each and every one of its points.

Since all the tangent vectors in a given tangent space have the same point of application, we can borrow the vector addition and scalar multiplication of $\mathbb{R}^3$ to turn $T_p\mathbb{R}^3$ into a *vector space*. Explicitly, we define (for $v_p, w_p \in T_p\mathbb{R}^3$ and $\lambda \in \mathbb{R}$)

$$v_p + w_p := (v + w)_p \quad \text{and} \quad \lambda v_p := (\lambda v)_p.$$

This is just the usual "parallelogram law" for addition of vectors, and scalar multiplication by λ merely stretches a tangent vector by a factor $|\lambda|$, reversing its direction if $\lambda < 0$.

◇ **Exercise 52** Show that, for a fixed point p, the vector spaces $\mathbb{R}^3$ and $T_p\mathbb{R}^3$ are *isomorphic*.

Vector fields

2.1.3 DEFINITION. A **vector field** X on $\mathbb{R}^3$ is a mapping

$$p \in \mathbb{R}^3 \mapsto X(p) \in T_p\mathbb{R}^3.$$

Let X and Y be vector fields on $\mathbb{R}^3$. Then we can define $X + Y$ to be the vector field on $\mathbb{R}^3$ such that

$$(X + Y)(p) := X(p) + Y(p)$$

for all $p \in \mathbb{R}^3$. Similarly, if f is a (real-valued) function on $\mathbb{R}^3$ and X is a vector field on $\mathbb{R}^3$, then we can define fX to be the vector field on $\mathbb{R}^3$ such that

$$(fX)(p) := f(p)X(p)$$

for all $p \in \mathbb{R}^3$.

NOTE : Both operations were defined "pointwise". This scheme is general. For convenience, we shall call it the *pointwise principle* : if a certain operation can be performed on the values of two functions at each point, then that operation can be extended to the functions themselves; simply apply it to their values at each point. By means of the pointwise principle we can automatically extend other operations on individual tangent vectors (like dot product and cross product) to operations on vector fields.

Let E_1, E_2, E_3 be the vector fields on $\mathbb{R}^3$ such that

$$E_1(p) := (1,0,0)_p, \quad E_2(p) := (0,1,0)_p, \quad \text{and} \quad E_3(p) := (0,0,1)_p$$

at each point p of $\mathbb{R}^3$. Thus E_i is the unit vector field in the positive x_i-direction. We shall refer to the ordered set $\underline{E} = (E_1, E_2, E_3)$ as the **natural frame field** on $\mathbb{R}^3$.

2.1.4 PROPOSITION. *If X is a vector field on $\mathbb{R}^3$, then there exist uniquely determined real-valued functions X_1, X_2, X_3 on $\mathbb{R}^3$ such that*

$$X = X_1E_1 + X_2E_2 + X_3E_3.$$

PROOF : By definition, vector field X assigns to each point p a tangent vector $X(p)$ at p. Thus the vector part of $X(p)$ depends on p, so we write it

$$(X_1(p), X_2(p), X_3(p))\,.$$

This defines X_1, X_2 and X_3 as (real-valued) *functions* on $\mathbb{R}^3$. Hence

$$\begin{aligned} X(p) &= (X_1(p), X_2(p), X_3(p))_p \\ &= X_1(p)(1,0,0)_p + X_2(p)(0,0,1)_p + X_3(p)(0,0,1)_p \\ &= X_1(p)E_1(p) + X_2(p)E_2(p) + X_3(p)E_3(p) \end{aligned}$$

for each point p. This means that the vector fields X and $X_1E_1 + X_2E_2 + X_3E_3$ have the same (tangent vector) value at each point, and hence they are equal. □

The functions $X_1, X_2,$ and X_3 are called the **Euclidean coordinate functions** of X. We write

$$X = (X_1, X_2, X_3) \quad \text{or sometimes} \quad X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}.$$

NOTE : A vector field X on $\mathbb{R}^3$ is a mapping *not* from $\mathbb{R}^3$ to $\mathbb{R}^3$ but from $\mathbb{R}^3$ to (the union) $\bigcup_{p \in \mathbb{R}^3} T_p\mathbb{R}^3$. So $X(p) = (p, (X_1(p), X_2(p), X_3(p))) = (X_1(p), X_2(p), X_3(p))_p$.

Computations involving vector fields may always be expressed in terms of their Euclidean coordinate functions. A vector field X is *differentiable* provided its Euclidean coordinate functions are differentiable. Henceforth, we shall understand "vector field" to mean "differentiable vector field".

NOTE : Since the subscript notation v_p for a tangent vector is somewhat cumbersome, from now on we shall omit the point of application p from the notation if no confusion is caused. However, in many situations the point of application is crucial, and will be indicated by using the old notation v_p or the phrase "a tangent vector v to $\mathbb{R}^3$ at p".

◇ **Exercise 53** Sketch the following vector fields on (the Euclidean plane) $\mathbb{R}^2$:

(a) $X(p) = (1, 0)$;

(b) $X(p) = p$;

(c) $X(p) = -p$;

(d) $X(x_1, x_2) = (x_2, x_1)$;

(e) $X(x_1, x_2) = (-x_2, x_1)$.

Frames

Using the isomorphism $v \mapsto v_p$ between $\mathbb{R}^3$ and $T_p\mathbb{R}^3$, the dot product on $\mathbb{R}^3$ may be transferred to each of its tangent spaces.

2.1.5 DEFINITION. The **dot product** of tangent vectors v_p and w_p at the same point of $\mathbb{E}^3$ is the number

$$v_p \bullet w_p := v \bullet w.$$

NOTE : This definition provides a dot product on each tangent space $T_p(\mathbb{R}^3)$ with the same properties as the original dot product on $\mathbb{R}^3$. In particular, each tangent vector v_p has a *norm* (or *length*) $\|v_p\| := \|v\|$. A vector of length 1 is called a *unit tangent vector*. Two tangent vectors v_p and w_p are *orthogonal* if and only if $v_p \bullet w_p = 0$.

2.1.6 DEFINITION. An ordered set $\underline{u} = (u_1, u_2, u_3)$ of three mutually orthogonal unit tangent vectors to $\mathbb{R}^3$ at the point p is called a **frame** (at p).

Thus $\underline{u} = (u_1, u_2, u_3)$ is a frame if and only if

$$u_i \bullet u_j = \delta_{ij}, \quad i, j = 1, 2, 3.$$

◇ **Exercise 54** Check that the tangent vectors

$$u_1 = \frac{1}{\sqrt{6}}(1, 2, 1)_p, \quad u_2 = \frac{1}{2\sqrt{2}}(-2, 0, 2)_p, \quad \text{and} \quad u_3 = \frac{1}{\sqrt{3}}(1, -1, 1)_p$$

constitute a frame at p. Express $v = (6, 1, -1)_p$ as a linear combination of these vectors. (Check the result by direct computation.)

2.1.7 EXAMPLE. At each point $p \in \mathbb{R}^3$, the tangent vectors

$$E_1(p) := (1, 0, 0)_p, \quad E_2(p) := (0, 1, 0)_p, \quad E_3(p) := (0, 0, 1)_p$$

constitute a frame, called the **natural frame** (at p).

If v is a tangent vector to $\mathbb{R}^3$ at some point p, then

$$v = (v_1, v_2, v_3)_p = v_1 E_1(p) + v_2 E_2(p) + v_3 E_3(p).$$

◇ **Exercise 55** Let $v \in T_p\mathbb{R}^3$ and let (u_1, u_2, u_3) be a frame (at p). Show that

$$v = (v_1, v_2, v_3)_p = (v \bullet u_1)u_1 + (v \bullet u_2)u_2 + (v \bullet u_3)u_3.$$

The numbers $v \bullet u_i$ $(i = 1, 2, 3)$ are the *coordinates* of the tangent vector v with respect to the frame $\underline{u} = (u_1, u_2, u_3)$.

2.1.8 DEFINITION. The **cross product** of tangent vectors v_p and w_p at the same point $p \in \mathbb{R}^3$ is the tangent vector (at p)

$$\begin{aligned} v_p \times w_p &:= \begin{vmatrix} E_1(p) & E_2(p) & E_3(p) \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} \\ &= (v_2w_3 - v_3w_2)E_1(p) + (v_3w_1 - v_1w_3)E_2(p) + (v_1w_2 - v_2w_1)E_3(p). \end{aligned}$$

NOTE : Likewise, this definition provides a cross product on each tangent space $T_p\mathbb{R}^3$ with the same properties as the original cross product on $\mathbb{R}^3$. In particular, two tangent vectors v_p and w_p are *collinear* if and only if $v_p \times w_p = 0$.

◇ **Exercise 56** If (u_1, u_2, u_3) is a frame, show that

$$u_1 \bullet u_2 \times u_3 = \pm 1.$$

Let F, $x \mapsto Ax + c$ be an isometry on $\mathbb{R}^3$. Then its orthogonal part A defines a mapping F_* that carries each tangent vector at p to a tangent vector at $F(p)$. The mapping

$$F_* = F_{*,p} : T_p\mathbb{R}^3 \to T_{F(p)}\mathbb{R}^3, \quad v_p \mapsto (Av)_{F(p)}$$

is called the **tangent mapping** of F (at p). In terms of Euclidean coordinates, we have

$$\begin{aligned} F_* \left(v_1E_1(p) + v_2E_2(p) + v_3E_3(p)\right) &= (a_{11}v_1 + a_{12}v_2 + a_{13}v_3)E_1(F(p)) + \\ &\quad (a_{21}v_1 + a_{22}v_2 + a_{23}v_3)E_2(F(p)) + \\ &\quad (a_{31}v_1 + a_{32}v_2 + a_{33}v_3)E_3(F(p)) \\ &= \sum_{i,j=1}^{3} (a_{ij}v_j)E_i(F(p)). \end{aligned}$$

◇ **Exercise 57** If T is a translation on $\mathbb{R}^3$, then for every tangent vector $v \in T_p\mathbb{R}^3$ show that $T_*(v)$ is *parallel* to v.

◇ **Exercise 58** If F and G are two isometries on $\mathbb{R}^3$, show that

$$(GF)_* = G_*F_* \quad \text{and} \quad \left(F^{-1}\right)_* = (F_*)^{-1}.$$

◇ **Exercise 59** Given an isometry F on $\mathbb{R}^3$, show that its tangent mapping F_* preserves the dot product of any two (tangent) vectors.

Since dot products are preserved, it follows automatically that derived concepts such as norm and orthogonality are preserved. Explicitly, if F is an isometry, then $\|F_*(v)\| = \|v\|$, and if v and w are orthogonal, so are $F_*(v)$ and $F_*(w)$. Thus frames are also preserved.

◇ **Exercise 60** If $\underline{u} = (u_1, u_2, u_3)$ is a frame at some point $p \in \mathbb{R}^3$ and F is an isometry on $\mathbb{R}^3$, show that $F_*(\underline{u}) = (F_*(u_1), F_*(u_2), F_*(u_3))$ is a frame at $F(p)$.

Recall that two *points* uniquely determine a translation. We now show that two *frames* uniquely determine an isometry.

2.1.9 THEOREM. *Given any two frames on $\mathbb{R}^3$, say $\underline{u} = (u_1, u_2, u_3)$ at the point p and $\underline{w} = (w_1, w_2, w_3)$ at the point q, there exists a unique isometry F on $\mathbb{R}^3$ such that*

$$F_*(u_i) = w_i, \quad i = 1, 2, 3.$$

PROOF : First we show that there is such an isometry. Let u_1, u_2, u_3 and w_1, w_2, w_3 be the points of $\mathbb{R}^3$ corresponding to (the vector parts of) the elements in the two frames. Let A be the *unique* linear transformation on $\mathbb{R}^3$ such that $A(u_i) = w_i$, $i = 1, 2, 3$.

◇ **Exercise 61** Check that the transformation (matrix) A is orthogonal.

Let T be the translation by (the vector) $q - A(p)$. We claim that the isometry $F = TA$ carries the frame $\underline{u} = (u_1, u_2, u_3)$ to the frame $\underline{w} = (w_1, w_2, w_3)$. First observe that

$$F(p) = TA(p) = q - A(p) + A(p) = q.$$

Then we get

$$F_*(u_i) = (Au_i)_{F(p)} = (w_i)_{F(p)} = (w_i)_q = w_i, \quad i = 1, 2, 3.$$

To prove uniqueness, we observe that the choice of A is the *only* possibility for the orthogonal part of the required isometry. The translation part is then completely determined also, since it must carry p to q. Hence the isometry $F = TA$ is uniquely determined. □

The isometry $F = TA$ (that carries the frame $\underline{u} = (u_1, u_2, u_3)$ to the frame $\underline{w} = (w_1, w_2, w_3)$) can be computed explicitly as follows. Let

$$u_i = (u_{1i}, u_{2i}, u_{3i})_p \quad \text{and} \quad w_i = (w_{1i}, w_{2i}, w_{3i})_q, \quad i = 1, 2, 3.$$

Then we form the 3×3 matrices (called the *attitude matrices* of the frames)

$$U := \begin{bmatrix} u_1 & u_2 & u_3 \end{bmatrix} = \begin{bmatrix} u_{ij} \end{bmatrix} \quad \text{and} \quad W := \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} = \begin{bmatrix} w_{ij} \end{bmatrix}.$$

◇ **Exercise 62** Show that the attitude matrix of a frame is orthogonal.

We claim that (the orthogonal matrix) A is WU^T. To verify this it suffices to check that

$$WU^\top(u_i) = w_i, \quad i = 1, 2, 3$$

since this uniquely characterizes A. For $i = 1$, we have

$$WU^\top(u_1) = WU^\top \begin{bmatrix} u_{11} \\ u_{21} \\ u_{31} \end{bmatrix} = W \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} w_{11} \\ w_{21} \\ w_{31} \end{bmatrix} = w_1.$$

That is, $WU^\top(u_1) = w_1$. The cases $i = 2, 3$ are similar; hence

$$A = WU^\top \left(= WU^{-1}\right).$$

As noted above, T is then necessarily the translation by $q - A(p)$.

◇ **Exercise 63** In each case decide whether F is an isometry on $\mathbb{R}^3$. If isometry exists, find the translation and orthogonal parts.

(a) $F(x) = -x$;

(b) $F(x) = (x \bullet a)a$ where $\|a\| = 1$;

(c) $F(x) = (x_3 - 3, x_2 - 2, x_1 + 1)$;

(d) $F(x) = (x_1, x_2, 2)$.

◇ **Exercise 64** Identify the *isometry* F, $x \mapsto -x$ on $\mathbb{R}^3$.

◇ **Exercise 65** Show that the matrix

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

represents a *reflection* in a plane. Find the plane.

◇ **Exercise 66** Given the frame

$$u_1 = \frac{1}{3}(2,2,1)_p, \quad u_2 = \frac{1}{3}(-2,1,2)_p, \quad u_3 = \frac{1}{3}(1,-2,2)_p$$

at $p = (0,1,0)$ and the frame

$$w_1 = \frac{1}{\sqrt{2}}(1,0,1)_q, \quad w_2 = (0,1,0)_q, \quad w_3 = \frac{1}{\sqrt{2}}(1,0,-1)_q$$

at $q = (3,-1,1)$, find c and A such that the isometry $F = T_c A$ carries the frame $\underline{u} = (u_1, u_2, u_3)$ to the frame $\underline{w} = (w_1, w_2, w_3)$.

Frame fields

2.1.10 Definition. Vector fields U_1, U_2, U_3 on $\mathbb{R}^3$ constitute a **frame field** on $\mathbb{R}^3$ provided

$$U_i \bullet U_j = \delta_{ij}, \quad i,j = 1,2,3.$$

Thus at each point $p \in \mathbb{R}^3$ the (tangent) vectors $U_1(p), U_2(p), U_3(p)$ form a frame.

◇ **Exercise 67** If X and Y are vector fields on $\mathbb{R}^3$ that are linearly independent at each point, show that

$$U_1 = \frac{X}{\|X\|}, \quad U_2 = \frac{\widetilde{Y}}{\|\widetilde{Y}\|}, \quad U_3 = U_1 \times U_2$$

is a frame field, where $\widetilde{Y} = Y - (Y \bullet U_1)U_1$.

Let (U_1, U_2, U_3) be a frame field on $\mathbb{R}^3$. If X is a vector field on $\mathbb{R}^3$, then

$$X = f_1 U_1 + f_2 U_2 + f_3 U_3,$$

where the (differentible) functions $f_i = X \bullet U_i$ are called the *coordinate functions* of X with respect to the frame (U_1, U_2, U_3). If

$$X = f_1U_1 + f_2U_2 + f_3U_3 \quad \text{and} \quad Y = g_1U_1 + g_2U_2 + g_3U_3$$

are two vector fields on $\mathbb{R}^3$, then

$$X \bullet Y = f_1g_1 + f_2g_2 + f_3g_3.$$

In particular,

$$\|X\| = \sqrt{f_1^2 + f_2^2 + f_3^2}.$$

NOTE : A given vector field X has a different set of coordinates functions with respect to each choice of a frame field (U_1, U_2, U_3). The *Euclidean* coordinate functions, of course, come from the natural frame field (E_1, E_2, E_3). In studying curves in $\mathbb{R}^3$ we shall be able to choose a frame field *specifically* adapted to the problem at hand. Not only does this simplify computations, but it gives a clearer understanding of geometry than if we had insisted on using the same frame field in every situation.

Orientation

Let $\underline{u} = (u_1, u_2, u_3)$ be a frame at a point $p \in \mathbb{R}^3$. Recall that associated with each frame $\underline{u}$ is its *attitude matrix* U.

NOTE : $u_1 \bullet u_2 \times u_3 = \det U = \pm 1$.

We make the following definition.

2.1.11 DEFINITION. The frame $\underline{u} = (u_1, u_2, u_3)$ is said to be

- **positively-oriented** (or **right-handed**) provided $u_1 \bullet u_2 \times u_3 = +1$;
- **negatively-oriented** (or **left-handed**) provided $u_1 \bullet u_2 \times u_3 = -1$.

At each point p of $\mathbb{R}^3$, the natural frame (e_1, e_2, e_3) is positively-oriented.

◇ **Exercise 68** Show that a frame (u_1, u_2, u_3) is positively-oriented if and only if $u_1 \times u_2 = u_3$. Thus the orientation of a frame can be determined, for practical purposes, by the "right-hand rule".

We know that the tangent mapping of an isometry carries frames to frames. The following result tells what happens to their orientations.

2.1.12 PROPOSITION. *If $\underline{u} = (u_1, u_2, u_3)$ is a frame at a point $p \in \mathbb{R}^3$ and F is an isometry on $\mathbb{R}^3$, then*

$$F_*(u_1) \bullet F_*(u_2) \times F_*(u_3) = (\operatorname{sgn} F)\, u_1 \bullet u_2 \times u_3.$$

PROOF : If

$$u_j = u_{1j}E_1(p) + u_{2j}E_2(p) + u_{3j}E_3(p), \quad j = 1, 2, 3$$

then we have

$$F_*(u_j) = \sum_{i,k=1}^{3} a_{ik}u_{kj}E_i(F(p)),$$

where $A = \left[a_{ik}\right]$ is the orthogonal part of F. Thus the attitude matrix of the frame $F_*(\underline{u}) = (F_*(u_1), F_*(u_2), F_*(u_3))$ is the matrix

$$\left[\sum_{k=1}^{3} a_{ik}u_{kj}\right] = AU.$$

But the triple scalar product of a frame is the determinant of its attitude matrix, and hence

$$\begin{aligned} F_*(u_1) \bullet F_*(u_2) \times F_*(u_3) &= \det(AU) \\ &= \det A \cdot \det U \\ &= (\operatorname{sgn} F)\, u_1 \bullet u_2 \times u_3. \end{aligned}$$

□

This result shows that if the isometry F is direct (i.e., $\operatorname{sgn} F = +1$), then F_* carries positively-oriented frames to positively-oriented frames and carries negatively-oriented frames to negatively-oriented frames. On the other hand, if the isometry F is opposite (i.e., $\operatorname{sgn} F = -1$), then positive goes to negative and negative to positive.

NOTE : Direct isometries preserve orientation (of frames), and opposite isometries reverse it. For this very reason, direct isometries are also called *orientation-preserving isometries*, whereas opposite isometries are called *orientation-reversing isometries*.

Both dot and cross product were originally defined in terms of *Euclidean* coordinates. It is easy to see that the dot product is given by the same formula, no matter what frame (u_1, u_2, u_3) is used to get coordinates. Indeed, we have

$$\begin{aligned} v \bullet w &= (v_1u_1 + v_2u_2 + v_3u_3) \bullet (w_1u_1 + w_2u_2 + w_3u_3) \\ &= \sum_{i,j=1}^{3} (v_iw_j)\, u_i \bullet u_j \\ &= \sum_{i,j=1}^{3} \delta_{ij}\, v_iw_j \\ &= v_1w_1 + v_2w_2 + v_3w_3. \end{aligned}$$

Almost the same result holds for cross products, but orientation is now involved.

◇ **Exercise 69** Let (u_1, u_2, u_3) be a frame at a point $p \in \mathbb{R}^3$. If $v = \sum v_iu_i$ and $w = \sum w_iu_i$, show that

$$v \times w = \epsilon \begin{vmatrix} e_1 & e_2 & e_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix},$$

where $\epsilon = u_1 \bullet u_2 \times u_3 = \pm 1$.

It follows immediately that the effect of an isometry on cross products also involves orientation. Explicitly, if v and w are tangent vectors to $\mathbb{R}^3$ at p, and F is an isometry on $\mathbb{R}^3$, then

$$F_*(v \times w) = (\operatorname{sgn} F)\, F_*(v) \times F_*(w).$$

2.2 Directional Derivatives

Associated with each tangent vector v_p to $\mathbb{R}^3$ is the *line* $t \mapsto p + tv$. If f is a differentiable function on $\mathbb{R}^3$, then

$$t \mapsto f(p + tv)$$

is an ordinary differentiable function $\mathbb{R} \to \mathbb{R}$.

NOTE : The derivative of this function at $t = 0$ tells the *initial rate of change* of f as p moves in the v direction.

We make the following definition.

2.2.1 DEFINITION. Given a differentiable function $f : \mathbb{R}^3 \to \mathbb{R}$ and a tangent vector $v_p \in T_p\mathbb{R}^3$, the number

$$v_p[f] := \frac{d}{dt} f(p + tv)\Big|_{t=0}$$

is called the **directional derivative** of f with respect to v_p.

NOTE : This definition appears in elementary calculus with the additional restriction that v_p be a unit vector. Even though we do *not* impose this restriction, we shall nevertheless refer to $v_p[f]$ as a *directional derivative*.

2.2.2 EXAMPLE. We compute (the directional derivative) $v_p[f]$ for the function $f(x_1, x_2, x_3) = x_1^2 x_2 x_3$ with $p = (1, 1, 0)$ and $v = (1, 0, -3)$. Then

$$p + tv = (1, 1, 1) + t(1, 0, -3) = (1 + t, 1, -3t)$$

describes the line through p in the v direction. Evaluating f along this line, we get

$$f(p + tv) = (1 + t)^2 \cdot 1 \cdot (-3t) = -3t - 6t^2 - 3t^3.$$

Now

$$\frac{d}{dt} f(p + tv) = -3 - 12t - 9t^2$$

and hence, at $t = 0$, we find $v_p[f] = -3$. Thus, in particular, the function f is initially *decreasing* as p moves in the v direction.

◇ **Exercise 70** Compute the directional derivative of the function $f(x_1, x_2, x_3) = x_1 x_2 + x_3$ with respect to $v_p = (1, -4, 2)_p$, where $p = (1, 1, 0)$.

The following result shows how to compute $v_p[f]$ in general, in terms of the partial derivatives of f at the point p.

2.2.3 PROPOSITION. *If* $v_p = (v_1, v_2, v_3)_p$ *is a tangent vector to* $\mathbb{R}^3$*, then*

$$v_p[f] = v_1 \frac{\partial f}{\partial x_1}(p) + v_2 \frac{\partial f}{\partial x_2}(p) + v_3 \frac{\partial f}{\partial x_3}(p).$$

PROOF : Let $p = (p_1, p_2, p_3)$. Then

$$p + tv = (p_1 + tv_1, p_2 + tv_2, p_3 + tv_3).$$

We use the *chain rule* to compute the derivative at $t = 0$ of the function

$$f(p + tv) = f(p_1 + tv_1, p_2 + tv_2, p_3 + tv_3).$$

We obtain

$$\begin{aligned} v_p[f] &= \frac{d}{dt} f(p + tv)\Big|_{t=0} \\ &= \sum_{i=1}^{3} \frac{d}{dt}(p_i + tv_i) \frac{\partial f}{\partial x_i}(p) \\ &= v_1 \frac{\partial f}{\partial x_1}(p) + v_2 \frac{\partial f}{\partial x_2}(p) + v_3 \frac{\partial f}{\partial x_3}(p). \end{aligned}$$

□

NOTE : We can write the directional derivative $v_p[f]$ in *matrix form* :

$$v_p[f] = \begin{bmatrix} \frac{\partial f}{\partial x_1}(p) & \frac{\partial f}{\partial x_2}(p) & \frac{\partial f}{\partial x_3}(p) \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}.$$

The main properties of this notion of derivative are as follows.

◇ **Exercise 71** Let f and g be differentiable functions on $\mathbb{R}^3$, v_p and w_p tangent vectors at a point $p \in \mathbb{R}^3$, and λ and μ real numbers. Show that :

(a) $(\lambda v_p + \mu w_p)[f] = \lambda v_p[f] + \mu w_p[f]$.

(b) $v_p[\lambda f + \mu g] = \lambda v_p[f] + \mu v_p[f]$.

(c) $v_p[fg] = v_p[f]g(p) + f(p)v_p[g]$.

The first two properties may be summarized by saying that (the mapping) $(v_p, f) \mapsto v_p[f]$ *is linear in* v_p *and* f. The third property is essentially just the usual *Leibniz rule* for differentiation of a product.

NOTE : No matter what form *differentiation* may take, it will always have suitable *linear* and *Leibnizian* properties.

◇ **Exercise 72** Given two tangent vectors v_p, w_p to $\mathbb{R}^3$, show that if

$$v_p[f] = w_p[f]$$

for *every* differentiable function f on $\mathbb{R}^3$, then $v_p = w_p$.

We now use the pointwise principle to define the *operation of a vector field on a function.* Let X be a vector field and f a differentiable function on $\mathbb{R}^3$. Then we define the function $X[f]$ (or simply Xf) by

$$X[f](p) := X(p)[f].$$

That is, the value of $X[f]$ at the point p is the directional derivative of f with respect to the tangent vector $X(p)$ at p.

◇ **Exercise 73** If X and Y are vector fields, and f, g, h are differentiable functions on $\mathbb{R}^3$, then show that (for $\lambda, \mu \in \mathbb{R}$) :

(a) $(fX + gY)[h] = fX[h] + gY[h]$.

(b) $X[\lambda f + \mu g] = \lambda X[f] + \mu X[g]$.

(c) $X[fg] = X[f] \cdot g + f \cdot X[g]$.

In particular, if (E_1, E_2, E_3) is the natural frame field on $\mathbb{R}^3$, then

$$E_i[f] = \frac{\partial f}{\partial x_i} \qquad (i = 1, 2, 3).$$

This is an immediate consequence of PROPOSITION 2.2.3. For example, $E_1(p) = (1, 0, 0)_p$ and hence (for all points $p = (p_1, p_2, p_3)$)

$$E_1(p)[f] = \frac{d}{dt} f(p_1 + t, p_2, p_3)\Big|_{t=0} = \frac{\partial f}{\partial x_1}(p)\cdot$$

If $X = (X_1, X_2, X_3)$ is a vector field on $\mathbb{R}^3$ we can write

$$\begin{aligned} X &= X_1E_1 + X_2E_2 + X_3E_3 \\ &= X_1\frac{\partial}{\partial x_1} + X_2\frac{\partial}{\partial x_2} + X_3\frac{\partial}{\partial x_3}. \end{aligned}$$

This notation makes it a simple matter to carry out explicit computations.

2.2.4 EXAMPLE. For (the vector field)

$$X = x_1\frac{\partial}{\partial x_1} - x_2^2\frac{\partial}{\partial x_3}$$

and (the differentiable function) $f = x_1^2x_2 + x_3^3$ we compute

$$\begin{aligned} X[f] &= x_1\frac{\partial}{\partial x_1}\left(x_1^2x_2 + x_3^3\right) - x_2^2\frac{\partial}{\partial x_3}\left(x_1^2x_2 + x_3^3\right) \\ &= x_1(2x_1x_2) - x_2^2(3x_3^2) \\ &= 2x_1^2x_2 - 3x_2^2x_3^2. \end{aligned}$$

◇ **Exercise 74** Given a vector field X, show that

$$X = X[x_1]\frac{\partial}{\partial x_1} + X[x_2]\frac{\partial}{\partial x_2} + X[x_3]\frac{\partial}{\partial x_3}$$

where $x \mapsto x_i$, $i = 1, 2, 3$ are the natural *coordinate functions*.

◇ **Exercise 75** Let

$$\begin{aligned} X &= \frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_2} \\ Y &= x_1\frac{\partial}{\partial x_1} - \frac{\partial}{\partial x_3} \\ f &= x_1x_2. \end{aligned}$$

Compute $X[f], Y[f], X[f^2], X[X[f]]$, and $X[Y[f]] - Y[X[f]]$.

2.3 Curves in Euclidean 3-Space $\mathbb{R}^3$

Parametrized curves

We want to characterize certain subsets of Euclidean 3-space $\mathbb{R}^3$ (to be called *curves*) that are, in a certain sense, one-dimensional and to which the methods of calculus can be applied.

NOTE : There are various notions of a curve in $\mathbb{R}^3$. We shall deal here with only one such notion. The definition is not entirely satisfactory but sufficient for our purposes.

A convenient way of defining such subsets is through differentiable maps. Let J be an *interval* on the real line (the interval may be open or closed, finite, semi-infinite or the entire real line). One can picture a curve in $\mathbb{R}^3$ as a trip taken by a moving particle α. At each "time" t, α is located at the point $\alpha(t) = (\alpha_1(t), \alpha_2(t), \alpha_3(t))$ in $\mathbb{R}^3$. We make the following definition.

2.3.1 DEFINITION. A (parametrized) **curve** in $\mathbb{R}^3$ is a *smooth* map

$$\alpha : J \to \mathbb{R}^3, \quad t \mapsto (\alpha_1(t), \alpha_2(t), \alpha_3(t)).$$

The curve is **regular** if $\dot{\alpha}(t) = \frac{d\alpha}{dt}(t) \neq 0$ for all $t \in J$.

The functions $\alpha_1, \alpha_2, \alpha_3$ are called the **Euclidean coordinate functions** and t is called the **parameter** of α. We write $\alpha = (\alpha_1, \alpha_2, \alpha_3)$.

NOTE : More generally, we can speak of differentiability of order k (or class C^k). One then requires the appropriate order of differentiability in each definition and theorem. To focus more on the geometry than the analysis we have ignored this subtlety by assuming curves to be smooth (i.e., of class C^∞).

The image set $\alpha(J) \subset \mathbb{R}^3$ is called the **trace** of α, which is the geometric object of interest. One should carefully distinguish a parametrized curve, which is a map, from its trace, which is a subset of $\mathbb{R}^3$.

NOTE : A given trace may be the image set (or route) of many (parametrized) curves. In this setting, it may be appropriate to call the common trace a *geometric curve* and refer to the curves as *parametrizations* (or *parametric representations*).

2.3.2 EXAMPLE. (The *line*) A (straight) line is the simplest type of geometric curve in $\mathbb{R}^3$. We know that two points determine a line. For two points $p, q \in \mathbb{R}^3$, the line $\overleftrightarrow{pq}$ may be described as follows. To attain the line, add

the *vector* p. To travel along the line, use the *direction vector* $q-p$ since this is the direction from p to q. A *parameter* t tells exactly how far along $q-p$ to go. Putting these steps together produces

$$\alpha : \mathbb{R} \to \mathbb{R}^3, \quad t \mapsto p + t(q-p), \quad q \neq p$$

which gives a parametrization of the **line** through the points p and q (or, if one prefers, the line through the point p with direction vector $q-p$).

◇ **Exercise 76** Find a parametrization of the line through the points $(-1, 0, 5)$ and $(3, -1, -2)$.

2.3.3 Example. (The *circle*) The **circle** of radius a with centre $p = (p_1, p_2, 0) \in \mathbb{R}^2$ is the set (locus) of points x in the plane $\mathbb{R}^2$ (i.e., the x_1x_2-plane of $\mathbb{R}^3$) such that

$$\|x - p\| = a$$

(the distance between x and p is the fixed positive real number a). A natural *parametric representation* is

$$\alpha : \mathbb{R} \to \mathbb{R}^3, \quad t \mapsto (p_1 + a\cos t, p_2 + a\sin t, 0).$$

◇ **Exercise 77** Find a curve $\alpha : \mathbb{R} \to \mathbb{R}^3$ whose trace is the *unit circle* $x_1^2 + x_2^2 = 1$, $x_3 = 0$ and such that $\alpha(t)$ runs clockwise around the circle with $\alpha(0) = (0, 1)$.

2.3.4 Example. (The *helix*) A (circular) **helix** is a geometric curve represented (given) parametrically by

$$\alpha : \mathbb{R} \to \mathbb{R}^3, \quad t \mapsto (a\cos t, a\sin t, bt); \quad a > 0,\ b \neq 0.$$

It rises (when $b > 0$) or falls (when $b < 0$) at a constant rate on the (circular) *cylinder* $x_1^2 + x_2^2 = a^2$. The x_3-axis is called the *axis*, and $2\pi\, b$ the *pitch* of the helix.

2.3.5 Example. The curve

$$\alpha : \mathbb{R} \to \mathbb{R}^3, \quad t \mapsto (t^3, t^2, 0).$$

is *not* regular because $\dot{\alpha}(0) = (0, 0, 0)$. The trace has a *cusp* at the origin.

2.3.6 EXAMPLE. The curve

$$\alpha : \mathbb{R} \to \mathbb{R}^3, \quad t \mapsto (\cosh t, \sinh t, t)$$

is known as the **hyperbolic helix**. (Recall that the *hyperbolic trigonometric functions* are defined by the formulas

$$\cosh t = \frac{e^t + e^{-t}}{2}, \quad \sinh t = \frac{e^t - e^{-t}}{2}, \quad \tanh t = \frac{e^t - e^{-t}}{e^t + e^{-t}}.$$

We have the fundamental identity $\cosh^2 t - \sinh^2 t = 1$.)

If we visualize a (parametrized) curve α in $\mathbb{R}^3$ as a moving particle, then at every time t there is a tangent vector at the point $\alpha(t)$ which gives the *instantaneous velocity* of α at that time.

2.3.7 DEFINITION. Let $\alpha : J \to \mathbb{E}^3$ be a curve in $\mathbb{R}^3$ with $\alpha = (\alpha_1, \alpha_2, \alpha_3)$. For each number $t \in J$, the **velocity vector** of α at t is the tangent vector

$$\dot{\alpha}(t) := \left(\frac{d\alpha_1}{dt}(t), \frac{d\alpha_2}{dt}(t), \frac{d\alpha_3}{dt}(t) \right)_{\alpha(t)}$$

at the point $\alpha(t) \in \mathbb{R}^3$.

If α is a regular curve, all its velocity vectors are different from zero. A regular curve can have no *corners* or *cusps*.

2.3.8 EXAMPLE. The velocity vector of the (straight) line $\alpha(t) = p + t(q - p)$ is

$$\dot{\alpha}(t) = (q_1 - p_1, q_2 - p_2, q_3 - p_3)_{\alpha(t)} = (q - p)_{\alpha(t)}.$$

The fact that α is "straight" is reflected in the fact that all its velocity vectors are *parallel*; only the point of application changes as t changes.

2.3.9 EXAMPLE. For the helix represented by $\alpha(t) = (a \cos t, a \sin t, bt)$, the velocity vector at t is

$$\dot{\alpha}(t) = (-a \sin t, a \cos t, b)_{\alpha(t)}.$$

The fact that the helix "rises" constantly is shown by the constancy of the x_3-coordinate of $\dot{\alpha}(t)$.

NOTE : The line, the circle, the ellipse, and the helix (circular or hyperbolic) are all *regular* curves.

◇ **Exercise 78** For a fixed t, the *tangent line* to a regular curve $\alpha : J \to \mathbb{R}^3$ at the point $\alpha(t)$ is the line $u \mapsto \alpha(t) + u\dot{\alpha}(t)$. Find the tangent line to the helix

$$\alpha(t) = (2\cos t, 2\sin t, t)$$

at the points $p = \alpha(0)$ and $q = \alpha\left(\frac{\pi}{4}\right)$.

◇ **Exercise 79** Find the curve $\alpha : \mathbb{R} \to \mathbb{R}^3$ such that

$$\alpha(0) = (2, 3, 0) \quad \text{and} \quad \dot{\alpha}(t) = (e^t, -2t, t^2).$$

Let $\alpha : J \to \mathbb{R}^3$ be a curve. The *norm* of the velocity vector $\dot{\alpha}(t)$ of α at t

$$\|\dot{\alpha}(t)\| := \sqrt{\dot{\alpha}(t) \bullet \dot{\alpha}(t)} = \sqrt{\left(\frac{d\alpha_1}{dt}(t)\right)^2 + \left(\frac{d\alpha_2}{dt}(t)\right)^2 + \left(\frac{d\alpha_3}{dt}(t)\right)^2}$$

is called the **speed** of α at t. Again, thinking of α as the path of a moving particle and t as time, we see that the length of the velocity vector is precisely the speed of the particle at the given time.

NOTE : A regular curve has speed always greater than zero.

◇ **Exercise 80** If $\alpha : J \to \mathbb{R}^3$ is a curve, its *acceleration vector* at t is given by

$$\ddot{\alpha}(t) := \left(\frac{d^2\alpha_1}{dt^2}(t), \frac{d^2\alpha_2}{dt^2}(t), \frac{d^2\alpha_3}{dt^2}(t)\right)_{\alpha(t)}.$$

What can be said about α if its acceleration is identically zero ?

◇ **Exercise 81** Verify that the curve $\alpha(t) = (\cos t, \sin t, 1)$ has constant speed, but nonzero acceleration.

◇ **Exercise 82** For the curve $\alpha(t) = \left(2t, t^2, \frac{t^3}{3}\right)$, find the velocity, speed, and acceleration for arbitrary t, and at $t = 1$.

◇ **Exercise 83** Show that (the trace of) the curve $\alpha(t) = (t\cos t, t\sin t, t)$ lies on a *cone* in $\mathbb{R}^3$. Find the velocity, speed, and acceleration of α at the vertex of the cone.

Other examples of curves

The following *plane* parametrized curves arise naturally throughout the physical sciences and mathematics.

2.3.10 EXAMPLE. (The *catenary*) Let $f : J \to \mathbb{R}$ be any smooth function. The *graph* of f is the set of all points $(t, f(t)) \in \mathbb{R}^2$ with $t \in J$, so is the trace of the (regular) curve

$$\alpha : J \to \mathbb{R}^2, \quad t \mapsto (t, f(t)).$$

In particular, for $f(t) = a \cosh \frac{t}{a}$, we get the **catenary** (from the Latin for "chain").

NOTE : The catenary is of historical interest, representing the form (shape) adopted by a perfect inextensible chain of uniform density suspended by its ends and acted upon by gravity. It was studied first by GALILEO GALILEI (1564-1642), who mistook it for a parabola, and later by GOTTFRIED LEIBNIZ (1646-1716), CHRISTIAAN HUYGENS (1629-1695), and JOHANN BERNOULLI (1667-1748). (They were responding to the challenge put out by JAKOB (JACQUES) BERNOULLI (1654-1705) to find the equation of the "chain-curve".) It is also of contemporary mathematics interest, being a plane section of the minimal surface (a soap film catenoid) spanning two circular discs, the *only minimal surface of revolution.*

2.3.11 EXAMPLE. (The *cycloid*) Suppose a circle of radius a sits on the x_1-axis making contact at the origin. Let the circle to roll (without slipping) along the positive x_1-axis. The figure (path) described by the point on the circle, originally in contact with the x_1-axis, is a geometric curve called **cycloid**. It can be shown that a parametric reprezentation of the cycloid is

$$\alpha : \mathbb{R} \to \mathbb{E}^2, \quad t \mapsto (a(t - \sin t), a(1 - \cos t))$$

where (the parameter) t is the angle formed by the (new) point of contact with the axis, the centre of the circle, and the original point of contact.

◇ **Exercise 84** Draw a picture (i.e., sketch the graph) of the cycloid.

We can see that the cycloid has infinitely many *cusps* (corresponding to $t = 2k\pi,\ k \in \mathbb{Z}$) : the arc of the cycloid between any two consecutive cusps is called an *arch*. Generally, when a curve rolls (without slipping) along another fixed curve, any point which moves with the moving curve describes a curve, called a *roulette* (from the French for "small wheel"). Consider now the roulette of a tracing point carried by a (moving) circle of radius a rolling along a line (the x_1-axis, say). It is assumed that in the initial configuration the (moving) circle is tangent to the x_1-axis at the origin and that the tracing point is the point on the x_2-axis distance ha from the centre of the circle. The resulting roulette, also known as a *cycloid*, has the parametrization

$$x_1(t) = a(t - h\sin t), \quad x_2(t) = a(1 - h\cos t).$$

The form of the cycloid depends on whether the tracing point is *inside* ($h < 1$), *on* ($h = 1$) or *outside* ($h > 1$) the moving circle. For $h < 1$ we obtain a "shortened" cycloid reminiscent of the sine curve. For $h > 1$ we obtain an "extended" cycloid with infinitely many self crossings. Finally, for $h = 1$ we get the (standard or "cuspidal") cycloid as introduced above.

NOTE : The cycloid has two additional names and a lot of interesting history. The other two names are the *tautochrone* and the *brachistochrone*. CHRISTIAAN HUYGENS (1629-1695) discovered a remarkable property of the cycloid : it is the only curve such that a body falling under its own weight is guided by this curve so as to oscillate with a period that is independent of the initial point where the body is released. Therefore, he called this curve (i.e. the cycloid) the *tautochrone* (from the Greek for "same time"; $\tau\alpha\upsilon\tau\acute{o}\varsigma$: equal, and $\chi\rho\acute{o}\nu o\varsigma$: time).

In 1696 JOHANN BERNOULLI (1667-1784) posed a question (problem) and invited his fellow mathematicians to solve it. The problem (which he had solved and which he considered very beautiful and very difficult), called the *brachistochrone problem*, is the following : *Given two points p and q in a vertical plane (with q below and to the right of p) find, among all (smooth) curves with endpoints p and q, the curve such that a particle which slides without friction along the curve under the influence of gravity will travel from the one point to the other in the least possible time.* (JOHANN BERNOULLI solved the problem ingeniously by employing *Fermat's principle* that light travels to minimize time together with *Snell's law of refraction*. The other solvers included Johann's brother, JAKOB BERNOULLI, as well as GOTTFRIED

Leibniz (1646-1716), Isaac Newton (1642-1727), and L'Hôpital (1661-1704).) This problem is important because it led to the systematic consideration of similar problems; the new discipline which developed thereby is called the *calculus of variations*. Moreover, Bernoulli's problem is a true minimum time problem of the kind that is studied today in *optimal control theory*. Bernoulli called the "fastest path" the *brachistochrone* (from the Greek for "shortest time"; $\beta\rho\acute{\alpha}\chi\iota\sigma\tau\varsigma$: shortest, and $\chi\rho\acute{o}\nu\varsigma$: time).

2.3.12 Example. (The *astroid*) A parametric reprezentation of the curve called the **astroid** is

$$\alpha : [0, 2\pi] \to \mathbb{E}^2, \quad t \mapsto \left(a\cos^3 t,\, a\sin^3 t\right).$$

The definition of the astroid is very similar to that of the cycloid. For the astroid, however, a circle is rolled (without slipping), not on a line, but *inside* another (fixed) circle. More precisely, let a circle of radius $\frac{a}{4}$ roll inside a large circle of radius a (and centered at the origin say). For concreteness, suppose we start the little circle at $(a, 0)$ and follow the path of the point originally in contact with $(a, 0)$ as the circle rolls up. Let t denote the angle from the centre of the large circle to the new contact point. One can show that, with respect to the origin, the rolling point moves to

$$\alpha(t) = \left(\frac{3a}{4}\cos t + \frac{a}{4}\cos 3t,\, \frac{3a}{4}\sin t - \frac{a}{4}\sin 3t\right).$$

◇ **Exercise 85** Show that the formula for the astroid may be reduced to

$$\alpha(t) = \left(a\cos^3 t,\, a\sin^3 t\right)$$

with *implicit* form

$$x_1^{\frac{2}{3}} + x_2^{\frac{2}{3}} = a^{\frac{2}{3}}.$$

Note : Recall that when one curve rolls along another fixed curve, any point which moves with the moving curve describes a curve, called a roulette. When the curves are circles the resulting roulette is called a *trochoid*. Trochoids occur naturally in the physical sciences. Assume that the fixed circle has centre o at the origin and radius $a > 0$, and that the moving circle has centre o' and "radius" $a' \neq 0$. The case $a' > 0$ is interpreted as the moving circle rolling on the outside of the fixed one

(an *epitrochoid*), while $a' < 0$ is interpreted as the moving circle rolling on the inside of it (an *hypotrochoid*). Suppose the tracing point is distant $a'h$ from the centre a' of the moving circle. Write t for the angle between the line oo' and the x_1-axis, and assume (without loss of generality) that when $t = 0$ the tracing point lies on the x_1-axis. It can be shown that the following parametric reprezentation results

$$t \mapsto ((\lambda+1)\cos t - h\cos(\lambda+1)t,\ (\lambda+1)\sin t - h\sin(\lambda+1)t)$$

where $\lambda := \frac{a}{a'} \neq 0$ and $h > 0$. The case $h = 1$ (i.e. when the tracing point lies on the moving circle) is of special significance. Various names have been assigned traditionally to the curves (trochoids) arising by taking certain values of λ; for example, for $\lambda = 1$: the *cardioid* (the "heart shaped" curve), for $\lambda = -3$: the *deltoid*, or for $\lambda = -4$: the *astroid*. *Ellipses are special case of trochoids.* (Consider the special case when $\lambda = -2$: the moving circle rolls *inside* the fixed circle, and has half the radius. For $0 < h < 1$ this is an ellipse; for $h = 0$ the ellipse becomes a circle, concentric with the fixed circle, and of half its radius.)

Another special case is obtained when $\lambda = 1$: the moving circle rolls *outside* the fixed circle, and has the same radius. The (epi)trochoid is then a *limancon* with parametric reprezentation

$$(x_1(t), x_2(t)) = (2\cos t - h\cos 2t, 2\sin t - h\sin 2t)\,.$$

The form of the limancon depends on the value of h. (When $h = 1$ we get the cardioid.)

2.3.13 Example. (The *tractrix*) The trace of the parametrized curve

$$\alpha : \mathbb{R} \to \mathbb{E}^2, \quad t \mapsto \left(t - \tanh t, \frac{1}{\cosh t}\right)$$

is called the **tractrix**.

◇ **Exercise 86** Show that the tractrix has the following remarkable property : the length of the line segment of the tangent of the tractrix between the point of tangency and the x_1-axis is constantly equal to 1.

There is another way of saying this : the circle of unit radius centered at the point $(t, 0)$ passes through the point $x(t)$ (on the tractrix), and the tangent line to the circle at $x(t)$ is orthogonal to the tangent line to the tractrix at that point. Thus the tractrix has the property that *it meets all circles of unit*

radius centered on the x_1-axis orthogonally. (For that reason the tractrix is described as an "orthogonal trajectory" of that family of circles.)

NOTE : The tractrix gives rise to an interesting example in the elementary geometry of surfaces : the surface of revolution obtained by rotating it about the x_1-axis is the *pseudosphere*, distinguished by the property of having constant negative (Gaussian) curvature. (Intuitively, the *curvature* of a surface is a number κ that measures the extent to which the surface "bends". In general, the curvature κ varies from point to point, being close to zero at points where the surface is rather flat, large at points where the surface bends sharply. For some surfaces the curvature is the same at all points, so naturally these are called *surfaces of constant curvature* κ.)

2.3.14 EXAMPLE. (The standard *conics*) A general **conic** is a set of points defined by the vanishing of a polynomial of degree two in two variables :

$$Ax_1^2 + 2Bx_1x_2 + Cx_2^2 + 2Dx_1 + 2Ex_2 + F = 0,$$

where $A, B, C, D, E, F \in \mathbb{R}$ and not all of $A, B,$ and C are zero. A class of conics arises from the following classical construction. One is given a line $\mathcal{D}$ (the *directrix*), a point f (the *focus*) not on $\mathcal{D}$, and a variable point p subject to the constraint that its distance from f is proportional to its distance from $\mathcal{L}$. Write p' for the (orthogonal) projection of p onto $\mathcal{D}$. Then the constraint reads

$$\|p - f\| = \epsilon \, \|p - p'\|$$

for some positive constant of proportionality ϵ, known as the *eccentricity.* (The line through f and its projection f' onto $\mathcal{D}$ is an *axis of symmetry.*) The *locus* of p is a **parabola** when $\epsilon = 1$, an **ellipse** when $\epsilon < 1$, or a **hyperbola** when $\epsilon > 1$. The "standard conics" arise from the special case when $\mathcal{D}$ is parallel to one of the coordinate axes (the x_2-axis say), and the focus lies on the other coordinate axis (the x_1-axis say). Then we get an equation of the form

$$(1 - \epsilon^2)x_1^2 + 2\beta x_1 + x_2^2 + \gamma = 0.$$

(Circles cannot be constructed in this way since the eccentricity is positive.) Convenient forms for the equations of the three "standard conics" can be now obtained easily.

(i) Consider first the case $\epsilon = 1$ of a parabola. The equation of the conic reduces to that of a *standard* parabola

$$x_2^2 = 4ax_1$$

with directrix the line $x_1 = -a$ and focus the point $f = (a, 0)$. The x_1-axis is the axis of symmetry of the parabola, and the point where it meets the parabola (in this case, the origin) is the *vertex*. A *standard parametrization* of this parabola is

$$x_1 = at^2, \quad x_2 = 2at.$$

(ii) Consider next the case $\epsilon < 1$ of an ellipse. The equation of the conic reduces to that of a *standard* ellipse

$$\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} = 1$$

where $b^2 = a^2(1 - \epsilon^2)$ and $0 < b < a$. The coordinate axes are axes of symmetry of a standard ellipse. The points $(0, \pm b), (\pm a, 0)$ where the axes meet the ellipse provide the four *vertices*. It is traditional to refer to a as the *major semiaxis* and b as the *minor semiaxis*. The directrix $\mathcal{D}^-$ is the line $x_1 = -\frac{a}{\epsilon}$, and the focus is the point $f^- = (-a\epsilon, 0)$. The symmetry of the equation shows that there is a second directrix $\mathcal{D}^+$ with equation $x_1 = \frac{a}{\epsilon}$ having a corresponding focus $f^+ = (a\epsilon, 0)$. The *centre* of a standard ellipse is the mid-point of the line segment joining the two foci (in this case, the origin).

NOTE : Despite the fact that the circle does not appear as a standard conic, it is profitable to think of a circle (centered at the origin) as the limiting case of standard ellipses as $b \to a$ (which corresponds to $\epsilon \to 0$).

A *standard parametrization* of this ellipse is

$$(x_1(t), x_2(t)) = (a\cos t, b\sin t).$$

(iii) Finally, consider the case $\epsilon > 1$ of a hyperbola. The equation of the conic reduces to that of a *standard* hyperbola

$$\frac{x_1^2}{a^2} - \frac{x_2^2}{b^2} = 1$$

where $b^2 = a^2(\epsilon^2 - 1)$ and $0 < a, b$. The coordinate axes are axes of symmetry of a standard hyperbola. Only the x_1-axis meets the hyperbola, at the *vertices* $(\pm a, 0)$. Again we have directrix lines $x_1 = -\frac{a}{\epsilon}$, $x_1 = \frac{a}{\epsilon}$ with corresponding foci $(\pm a\epsilon, 0)$. The *centre* of a standard hyperbola is the mid-point of the line segment joining the two foci (i.e. the origin). The lines $x_2 = \pm\frac{b}{a}x$ are the *asymptotes* of the hyperbola. (The asymptotes are orthogonal if and only if $a = b$; this corresponds to the case when the eccentricity $\epsilon = \sqrt{2}$.)

NOTE : A point $p = (x_1, x_2)$ satisfying the equation of a standard hyperbola is subject only to the constraint that $x_1 \geq a$ or $x_1 \leq -a$. Thus the key feature of a (standard) hyperbola is that it splits into two "branches" : the *positive branch* (defined for $x_1 \geq a$) and the *negative branch* (defined for $x_1 \leq -a$).

◇ **Exercise 87** Find parametrizations for each of the two branches of this hyperbola.

NOTE : A general conic

$$Ax_1^2 + 2Bx_1x_2 + Cx_2^2 + Dx_1 + Ex_2 + F = 0$$

(where not all of $A, B,$ and C are zero) represents one of the following eight types of *loci* (geometric curves) : an ellipse, a hyperbola, a parabola, a pair of intersecting lines, a pair of parallel lines, a line "counted twice", a single point, or the empty set. Moreover, the cases that can occur are governed by the sign of the expression $AC - B^2$ as follows :

- If $AC - B^2 > 0$, the possibilities are an ellipse, a single point, or the empty set.
- If $AC - B^2 = 0$, the possibilities are a parabola, two parallel lines, a single line, or the empty set.
- If $AC - B^2 < 0$, the possibilities are a hyperbola or two intersecting lines.

2.3.15 EXAMPLE. (General *algebraic* curves)

. . . . . .

The shortest distance between two points

Let's consider the following question : *What is the shortest distance between two points* $p, q \in \mathbb{R}^3$ *?* We have been taught since childhood that the answer is the (straight) line, but now we can see why our intuition is correct.

Let $\alpha : J \to \mathbb{R}^3$ be a curve. Let $[t_0, t_1] \subset J$ and set $\alpha(t_0) = p$ and $\alpha(t_1) = q$. We defined the speed of α at t as the length of the velocity vector $\alpha'(t)$. Thus speed is a real-valued (continuous) function on the interval J. In physics, the distance traveled by a moving particle is determined by integrating its speed with respect to time. Thus we define the **arc length** of α, from p to q, to be the *number*

$$L(\alpha) := \int_{t_0}^{t_1} \|\dot{\alpha}(t)\|\, dt.$$

We are able to answer the question of which route between two given points gives the *shortest distance.*

2.3.16 THEOREM. *The line is the curve of least arc length between two points.*

PROOF : Consider two points $p, q \in \mathbb{R}^3$. The line *segment* between them may be parametrized by

$$\lambda : [0, 1] \to \mathbb{E}^3, \quad t \mapsto p + t(q - p)$$

where $q - p$ gives the *direction.* Then

$$\dot{\lambda}(t) = q - p \quad \text{and} \quad \|\dot{\lambda}(t)\| = \|q - p\|.$$

Therefore

$$L(\lambda) = \int_0^1 \|\dot{\lambda}(t)\|\, dt = \|q - p\| \int_0^1 dt = \|q - p\|$$

and the length of the line segment (or direction vector) from p to q is the distance from p to q (as of course expected). Now we consider another curve

segment $\alpha : [t_0, t_1] \to \mathbb{R}^3$ which joins p and q; that is, $\alpha(t_0) = p$ and $\alpha(t_1) = q$.

We want to show that $L(\alpha) > L(\lambda)$ and, since α is arbitrary, this will say that the straight line minimizes distance.

Now, why should α be longer than λ ? One intuitive explanation is to say that α starts off in the wrong direction. That is, $\dot{\alpha}(t_0)$ is not "pointing toward" q. How can we measure this deviation ? The angle between the unit vector u in the direction of $q - p$ and the velocity vector $\dot{\alpha}(t_0)$ of α at p may be calculated by taking the dot product $\dot{\alpha}(t) \bullet u$. The total deviation may be added up by integration to give us an idea of why $L(\alpha) > L(\lambda)$ should hold.

Let $u = \frac{1}{\|q-p\|}(q - p)$. We have

$$\frac{d}{dt}(\alpha(t) \bullet u) = \dot{\alpha}(t) \bullet u + \alpha(t) \bullet \dot{u} = \dot{\alpha}(t) \bullet u.$$

Now we compute the integral

$$\int_{t_0}^{t_1} \dot{\alpha}(t) \bullet u \, dt$$

in two different ways to obtain the inequality. On the one hand, we have

$$\begin{aligned}
\int_{t_0}^{t_1} \dot{\alpha}(t) \bullet u \, dt &= \int_{t_0}^{t_1} \frac{d}{dt}(\alpha(t) \bullet u) \, dt = \alpha(t_1) \bullet u - \alpha(t_0) \bullet u \\
&= q \bullet u - p \bullet u \\
&= (q - p) \bullet u \\
&= \frac{(q - p) \bullet (q - p)}{\|q - p\|} \\
&= \|q - p\| \\
&= L(\lambda).
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
\int_{t_0}^{t_1} \dot{\alpha}(t) \bullet u \, dt &\leq \int_{t_0}^{t_1} \|\dot{\alpha}(t)\| \, \|u\| \, dt \quad \text{(by the Cauchy-Schwarz inequality)} \\
&= \int_{t_0}^{t_1} \|\dot{\alpha}(t)\| \, dt \\
&= L(\alpha).
\end{aligned}$$

Consequently,

$$L(\lambda) = \int_{t_0}^{t_1} \dot{\alpha}(t) \bullet u \, dt \leq L(\alpha).$$

Observe that

$$\dot{\alpha}(t) \bullet u = \|\dot{\alpha}(t)\| \, \|u\|$$

only when $\cos\theta = 1$, or $\theta = 0$. That is, the vector $\dot{\alpha}(t)$ must be *collinear* with $q - p$ for all t. In this case α is (a parametrization of) the line segment from p to q. Therefore we have strict inequality $L(\lambda) < L(\alpha)$ unless α is a line segment. □

◇ **Exercise 88**

(a) Find the arc length of the *catenary* $t \mapsto (t, \cosh t)$ from $t = 0$ to $t = t_1$.

(b) Show that the curve $t \mapsto (3t^2, t - 3t^3)$ has a unique self crossing, determine the corresponding parameters a and b, and then find the arc length from $t = a$ to $t = b$.

(c) Find the arc length of the *astroid* $t \mapsto (\cos^3 t, \sin^3 t)$ from $t = 0$ to $t = \frac{\pi}{2}$ and then from $t = 0$ to $t = \pi$. Compare the results.

(d) Show that the arc length of the *parabola* $t \mapsto (t^2, 2t)$ from $t = 0$ to $t = t_1$ is given by

$$L = t_1\sqrt{1 + t_1^2} + \ln\left(t_1 + \sqrt{1 + t_1^2}\right).$$

◇ **Exercise 89** Find an expression for the arc length of the *cycloid* $t \mapsto (a(t - \sin t), a(1 - \cos t))$ from $t = 0$ to $t = t_0$, where $0 \leq t_0 \leq 2\pi$. Deduce that the arc length from $t = 0$ to $t = 2\pi$ is 8.

Arc length parametrizations

Given a (parametrized) curve α, we can construct many new (parametrized) curves that follow the same route (i.e., have the same trace) as α but travel at different speeds. Any such alteration is called a *reparametrization*. More precisely, we have the following definition.

2.3.17 DEFINITION. Let J and J' be intervals on the real line. Let $\alpha : J \to \mathbb{R}^3$ be a curve and let $h : J' \to J$ be a smooth function (usually with smooth inverse). Then the composite function

$$\beta = \alpha(h) : J' \to \mathbb{E}^3, \quad \beta(s) = \alpha(h(s))$$

is a curve called the **reparametrization** of α by h. (The function h is the **change of parameter**.)

The curves β and α pass through the same points in $\mathbb{R}^3$, but they reach any of these points in different "times" (s and t).

◇ **Exercise 90** A smooth function

$$h : J_\theta \to J_t, \quad \theta \mapsto t = h(\theta)$$

is said to be an *allowable change of parameter* on (the interval) J_θ if it is *onto* and $h'(\theta) \neq 0$ on J_θ. Show that if h is an allowable change of parameter, then h is invertible and its inverse h^{-1} is also an allowable change of parameter (on J_t).

2.3.18 EXAMPLE. The (smooth) function $s \mapsto s + s^3$ defines an allowable change of parameter on $\mathbb{R}$. On the other hand, the function $s \mapsto s^3$ does *not* (since its derivative vanishes at $s = 0$).

◇ **Exercise 91** Check that

(a) The function $\theta \mapsto t = \dfrac{\theta - a}{b - a}$ is an allowable change of parameter which takes the interval $[a, b]$ onto $[0, 1]$.

(b) The function $\theta \mapsto t = \dfrac{1}{\pi}\left(\dfrac{\pi}{2} + \arctan\theta\right)$ is an allowable change of parameter which takes the interval $(-\infty, \infty)$ onto $(0, 1)$.

(c) The function $\theta \mapsto t = \dfrac{\arctan\theta - \arctan a}{\frac{\pi}{2} - \arctan a}$ is an allowable change of parameter which takes the interval $[a, \infty)$ onto $[0, 1)$.

2.3.19 EXAMPLE. The reparametrization of the curve

$$\alpha : (0, 4) \to \mathbb{R}^3, \quad t \mapsto (\sqrt{t}, t\sqrt{t}, 1 - t)$$

by (the change of parameter) $h : (0, 2) \to (0, 4), \quad s \mapsto s^2$ is

$$\beta(s) = \alpha(h(s)) = \alpha(s^2) = (s, s^3, 1 - s^2), \quad s \in (0, 2).$$

◇ **Exercise 92** Reparametrize

(a) the *circle*

$$t \mapsto (a\cos t, a\sin t), \quad t \in [-\pi, \pi]$$

by $h : [-1, 1] \to [-\pi, \pi], \quad \theta \mapsto 4 \arctan \theta.$

(b) the positive branch of the (standard) *hyperbola*

$$t \mapsto (a\cosh t, b\sinh t), \quad t \in \mathbb{R}$$

by $h : (0, \infty) \to \mathbb{R}, \quad \theta \mapsto \ln \theta.$

(c) the *tractrix*

$$t \mapsto \left(t - \tanh t, \frac{1}{\cosh t}\right), \quad t \in \mathbb{R}$$

by $h : (0, \pi) \to \mathbb{R}, \quad \theta \mapsto \ln \tan \frac{\theta}{2}.$

The following result relates the velocities of a curve and of a reparametrization.

2.3.20 Proposition. *If β is a reparametrization of α by h, then*

$$\dot{\beta}(s) = \frac{dh}{ds}(s) \cdot \dot{\alpha}(h(s)).$$

Proof : If $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, then

$$\beta(s) = \alpha(h(s)) = (\alpha_1(h(s)), \alpha_2(h(s)), \alpha_3(h(s))) .$$

By the chain-rule we obtain

$$\begin{aligned} \dot{\beta}(s) &= \frac{d\beta}{ds}(s) = \frac{d\alpha(h)}{ds}(s) \\ &= \left(\frac{d\alpha_1}{ds}(h(s)) \cdot \frac{dh}{ds}(s), \frac{d\alpha_2}{ds}(h(s)) \cdot \frac{dh}{ds}(s), \frac{d\alpha_3}{ds}(h(s)) \cdot \frac{dh}{ds}(s)\right) \\ &= \frac{dh}{ds}(s) \cdot \dot{\alpha}(h(s)). \end{aligned}$$

□

◇ **Exercise 93** Recall that the *arc length* of a curve $\alpha : [a, b] \to \mathbb{R}^3$ is given by $L(\alpha) = \int_a^b \|\dot{\alpha}(t)\| \, dt$. Let $\beta : [c, d] \to \mathbb{R}^3$ be a reparametrization of a by (the change of parameter) $h : [c, d] \to [a, b]$. Show that the arc length does *not* change under reparametrization.

◇ **Exercise 94** Let $\alpha : J \to \mathbb{R}^3$ be a curve and $f : \mathbb{R}^3 \to \mathbb{R}$ a (smooth) function on $\mathbb{R}^3$. Show that (for $t \in J$)

$$\dot{\alpha}(t)[f] = \frac{d}{dt} f(\alpha(t)).$$

This simple result shows that *the rate of change of f along the line through the point $\alpha(t)$ in the direction $\alpha'(t)$ is the same as along the curve α itself.*

Sometimes one is interested only in the trace of a curve and not in the particular speed at which it is covered. One way to ignore the speed of a curve α is to reparametrize to a curve β which has *unit speed*. Then β represents a "standard trip" along the trace of α.

2.3.21 THEOREM. *If $\alpha : J \to \mathbb{R}^3$ is a regular curve, then there exists a reparametrization β of α such that β has unit speed.*

PROOF : Fix a number t_0 in J and consider the *arc length function*

$$s(t) := \int_{t_0}^{t} \|\dot{\alpha}(u)\| \, du.$$

Since α is regular, the FUNDAMENTAL THEOREM OF CALCULUS implies

$$\frac{ds}{dt} = \|\dot{\alpha}(t)\| > 0.$$

By the MEAN VALUE THEOREM, s is strictly increasing on J and so is one-to-one. Therefore s has an inverse function $t = t(s)$ and the respective derivatives are inversely related :

$$\frac{dt}{ds}(s) = \frac{1}{\frac{ds}{dt}(t(s))} > 0.$$

Let $\beta(s) = \alpha(t(s))$ be the reparametrization of α. We claim that β has unit speed. Indeed, we have

$$\dot{\beta}(s) = \frac{dt}{ds}(s) \cdot \dot{\alpha}(t(s))$$

and hence

$$\begin{aligned} \|\dot{\beta}(s)\| &= \left|\frac{dt}{ds}(s)\right| \cdot \|\dot{\alpha}(t(s))\| \\ &= \frac{dt}{ds}(s) \cdot \frac{ds}{dt}(t(s)) \\ &= 1. \end{aligned}$$

□

Without loss of generality, suppose β is defined on the interval $[0,1]$. Consider the arc length of the reparametrization β out to a certain parameter value s_0

$$L(\beta) = \int_0^{s_0} \|\dot{\beta}(s)\|\, ds = \int_0^{s_0} ds = s_0.$$

Thus β has its arc length as parameter. We sometimes call the unit speed curve β the **arc length parametrization** of α.

NOTE : Observe that a curve β parametrized by arc length has speed given by

$$\|\dot{\beta}(s)\| = \left|\frac{dt}{ds}\right| \cdot \|\dot{\alpha}(t)\| = \frac{\|\dot{\alpha}(t)\|}{\|\dot{\alpha}(t)\|} = 1.$$

Hence we speak of a unit speed curve as a curve parametrized by arc length. We reserve the variable s for the arc length parameter when it is convenient and t for an arbitrary parameter.

2.3.22 EXAMPLE. If $\alpha(t) = p + t(q-p)$, then $\dot{\alpha}(t) = (q-p)_{\alpha(t)}$ and hence

$$\|\dot{\alpha}(t)\| = \|q-p\|.$$

Then

$$s(t) = \int_0^t \|\dot{\alpha}(t)\|\, dt = \int_0^t \|q-p\|\, dt = \|q-p\|t$$

and the inverse function is $t(s) = \frac{1}{\|q-p\|}s$. So an arc length parametrization is given by

$$\beta(s) = p + s\frac{q-p}{\|q-p\|}.$$

Note that $\|\dot{\beta}(s)\| = 1$.

◇ **Exercise 95** Find an arc length parametrization of the circle $x_1^2 + x_2^2 = a^2$.

2.3.23 EXAMPLE. Consider the helix $\alpha(t) = (a\cos t, a\sin t, bt)$ with $\dot{\alpha}(t) = (-a\sin t, a\cos t, b)_{\alpha(t)}$. We have

$$\|\dot{\alpha}(t)\|^2 = \dot{\alpha}(t) \bullet \dot{\alpha}(t) = a^2\sin^2 t + a^2\cos^2 t + b^2 = a^2 + b^2.$$

Thus α has constant speed $k = \sqrt{a^2 + b^2}$. Then

$$s(t) = \int_0^t \|\dot{\alpha}(t)\|\, dt = \int_0^t k\, dt = kt.$$

Hence $t(s) = \frac{s}{k}$. Substituting in the formula for α, we get the unit speed reparametrization

$$\beta(s) = \alpha\left(\frac{s}{k}\right) = \left(a\cos\frac{s}{k}, a\sin\frac{s}{k}, \frac{bs}{k}\right).$$

It is easy to check directly that $\|\dot{\beta}(s)\| = 1$.

NOTE : If a curve α has constant speed, then it may be parametrized by arc length explicitly. For general curves, however, the integral defining s may be impossible to compute in *closed form.*

2.3.24 EXAMPLE. The curve $\alpha(t) = (a\cos t, b\sin t)$ gives an *ellipse* in (the x_1x_2-plane of $\mathbb{R}^3$, identified with) $\mathbb{R}^2$. Furthermore, $\dot{\alpha}(t) = (-a\sin t, b\cos t)_{\alpha(t)}$ and

$$\|\dot{\alpha}(t)\| = \sqrt{a^2\sin^2 t + b^2\cos^2 t} = \sqrt{a^2 + (b^2 - a^2)\cos^2 t}.$$

The resulting length-function

$$s(t) = \int_0^t \sqrt{a^2 + (b^2 - a^2)\cos t}\, dt$$

is *not* generally expressible in terms of elementary functions (it is an example of an *elliptic integral*).

Vector fields (on curves)

The general notion of vector field can be adapted to curves as follows.

2.3.25 DEFINITION. A **vector field** X on a curve $\alpha : J \to \mathbb{R}^3$ is a (differentible) mapping

$$t \in J \mapsto X(t) \in T_{\alpha(t)}\mathbb{R}^3.$$

We have already met such a vector field : for any curve α, its velocity $\dot{\alpha}$ clearly satisfies this definition.

NOTE : It is important to realize that unlike $\dot{\alpha}$, arbitrary vector fields on α need *not* be tangent to (the trace of) the curve a, but may point in any direction.

A vector field X on a curve α is a *unit vector field* if each vector $\alpha'(t)$ (which is a tangent vector to $\mathbb{E}^3$ at $\alpha(t)$) is a unit vector.

The properties of vector fields on curves are analogous to those of vector fields on $\mathbb{R}^3$. For example, if X is a vector field on the curve $\alpha : J \to \mathbb{R}^3$, then for each $t \in J$ we can write

$$\begin{aligned} X(t) &= (X_1(t), X_2(t), X_3(t))_{\alpha(t)} \\ &= X_1(t)E_1(\alpha(t)) + X_2(t)E_2(\alpha(t)) + X_3(t)E_3(\alpha(t)) \\ &= X_1(t)\left.\frac{\partial}{\partial x_1}\right|_{\alpha(t)} + X_2(t)\left.\frac{\partial}{\partial x_2}\right|_{\alpha(t)} + X_3(t)\left.\frac{\partial}{\partial x_3}\right|_{\alpha(t)}. \end{aligned}$$

We have thus defined real-valued functions X_1, X_2, X_3 on J, called the *Euclidean coordinate functions* of X. These will always be assumed to be differentiable (in fact, smooth).

NOTE : The composite function $t \mapsto E_i(\alpha(t)) = \left.\frac{\partial}{\partial x_i}\right|_{\alpha(t)}$ is a vector field on α. Where it seems to be safe to do so, we shall often write merely E_i (or $\frac{\partial}{\partial x_i}$) insted of $E_i(\alpha(t))$.

The operations of addition, scalar multiplication, dot product, and cross product of vector fields (on the same curve) are all defined in the usual pointwise fashion.

2.3.26 EXAMPLE. Given

$$X(t) = t^2\frac{\partial}{\partial x_1} - t\frac{\partial}{\partial x_3}, \quad Y(t) = (1-t^2)\frac{\partial}{\partial x_2} + t\frac{\partial}{\partial x_3}, \quad \text{and} \quad f(t) = \frac{1+t}{t}$$

we obtain the vector fields

$$\begin{aligned}(X+Y)(t) &= t^2\frac{\partial}{\partial x_1}+(1-t^2)\frac{\partial}{\partial x_2}\\ (fX)(t) &= t(t+1)\frac{\partial}{\partial x_1}-(t+1)\frac{\partial}{\partial x_3}\\ (X\times Y)(t) &= \begin{vmatrix} E_1 & E_2 & E_3\\ t^2 & 0 & -t\\ 0 & 1-t^2 & t\end{vmatrix}\\ &= t(1-t^2)\frac{\partial}{\partial x_1}-t^3\frac{\partial}{\partial x_2}+t^2(1-t^2)\frac{\partial}{\partial x_3}\end{aligned}$$

and the real-valued function

$$(X\bullet Y)(t)=-t^2.$$

To differentiate a vector field on α one simply differentiate its Euclidean coordinate functions, thus obtaining a new vector field on α. Explicitly, if

$$X=X_1\frac{\partial}{\partial x_1}+X_2\frac{\partial}{\partial x_2}+X_3\frac{\partial}{\partial x_3}$$

then

$$\dot{X}=\dot{X}_1\frac{\partial}{\partial x_1}+\dot{X}_2\frac{\partial}{\partial x_2}+\dot{X}_3\frac{\partial}{\partial x_3}\cdot$$

In particular, the derivative $\ddot{\alpha}$ of the velocity vector field $\dot{\alpha}$ is called the *acceleration* of α.

NOTE : By contrast with velocity, acceleration is generally *not* tangent to the curve.

The following basic differentiation rules hold (for X and Y vector fields on $\mathbb{R}^3$, f a real-valued (differentiable) function on $\mathbb{R}^3$, and λ and μ real numbers) :

$$\begin{aligned}(\lambda X+\mu Y)^{\cdot} &= \lambda\dot{X}+\mu\dot{Y}\\ (fX)^{\cdot} &= \dot{f}X+f\dot{X}\\ (X\bullet Y)^{\cdot} &= \dot{X}\bullet Y+X\bullet\dot{Y}.\end{aligned}$$

If the function $X\bullet Y$ is constant, the last formula shows that

$$\dot{X}\bullet Y+X\bullet\dot{Y}=0.$$

◇ **Exercise 96** Show that a curve has constant speed if and only if its acceleration is everywhere orthogonal to its velocity.

◇ **Exercise 97** Let X be a vector field on the helix $\alpha(t) = (\cos t, \sin t, t)$. In each of the following cases, express X in the form $\sum X_i \frac{\partial}{\partial x_i}$.

(a) $X(t)$ is the vector from $\alpha(t)$ to the origin of $\mathbb{R}^3$;

(b) $X(t) = \dot{\alpha}(t) - \ddot{\alpha}(t)$;

(c) $X(t)$ has unit length and is orthogonal to both $\dot{\alpha}(t)$ and $\ddot{\alpha}(t)$;

(d) $X(t)$ is the vector from $\alpha(t)$ to $\alpha(t + \pi)$.

Recall that tangent vectors are *parallel* if they have the same vector parts. We say that a vector field X on a curve is **parallel** provided all its (tangent vector) values are parallel. In this case, if the common vector part is (c_1, c_2, c_3), then

$$X(t) = (c_1, c_2, c_3)_{\alpha(t)} = c_1 \frac{\partial}{\partial x_1} + c_2 \frac{\partial}{\partial x_2} + c_3 \frac{\partial}{\partial x_3}$$

for all t. The parallelism for a vector field is equivalent to the constancy of its Euclidean coordinate functions.

◇ **Exercise 98** Let $\alpha, \bar{\alpha} : J \to \mathbb{R}^3$ be two curves such that $\dot{\alpha}(t)$ and $\dot{\bar{\alpha}}(t)$ are parallel (same Euclidean coordinates) at each t. Show that α and $\bar{\alpha}$ are *parallel* in the sense that there is a point $p \in \mathbb{R}^3$ such that $\bar{\alpha}(t) = \alpha(t) + p$ for all t.

2.4 Serret-Frenet Formulas

The geometry of a curve (i.e., its turning and twisting) may be (completely) described by attaching a "moving trihedron" (or *moving frame)* along the curve. The variation of its elements is described by the so-called *Serret-Frenet formulas*, which are fundamental in the study of (differential) geometry of curves in $\mathbb{R}^3$. We start by deriving mathematical measurements of the *turning* and *twisting* of a curve.

The Serret-Frenet frame

Let $\beta : J \to \mathbb{R}^3$ be a unit speed curve, so $\|\dot{\beta}(s)\| = 1$ for all $s \in J$.

2.4.1 DEFINITION. The vector field $T := \dot{\beta}$ is called the **unit tangent vector field** on β.

Since T has constant length 1, its derivative $\dot{T} = \ddot{\beta}$ measures only the rate of change of $T's$ direction (i.e., measures the way the curve is turning in $\mathbb{R}^3$). Hence $\dot{T}$ is a good choice to detect some of the geometry of β.

2.4.2 DEFINITION. The vector field $\dot{T}$ is called the **curvature vector field** on β.

Differentiation of $T \bullet T = 1$ gives

$$0 = (T \bullet T)^{\cdot} = \dot{T} \bullet T + T \bullet \dot{T} = 2\, T \bullet \dot{T}.$$

Hence $T \bullet \dot{T} = 0$ and, therefore, $\dot{T}$ is orthogonal to T. We say that $\dot{T}$ is *normal* to β. The length of the curvature vector field $\dot{T}$ gives a numerical measurement of the turning of β.

2.4.3 DEFINITION. The real-valued function $\kappa : I \to \mathbb{R}$ given by

$$\kappa(s) := \|\dot{T}(s)\|$$

is called the **curvature function** of β.

Of course $\kappa \geq 0$ and κ increases as β turns more sharply.

NOTE : If $\kappa = 0$, then (as we will see in THEOREM 2.4.10 below) we know everything about the curve β already.

We assume that κ is never zero, so $\kappa > 0$. Then the vector field $N = \frac{1}{\kappa}\dot{T}$ on β tells the *direction* in which β is turning at each point.

2.4.4 DEFINITION. The vector field

$$N := \frac{1}{\kappa}\dot{T}$$

is called the **principal normal vector field** on β.

We need to introduce a third vector field on β as part of our "moving trihedron" along the curve and this vector field should be orthogonal to both T and N (just as T and N are to each other).

2.4.5 DEFINITION. The vector field

$$B := T \times N$$

is called the **binormal vector field** on β.

2.4.6 PROPOSITION. *Let $\beta : J \to \mathbb{R}^3$ be a unit speed curve with nonzero curvature. Then the three vector fields $T, N,$ and B on β are unit vector fields which are mutually orthogonal at each point.*

PROOF : By definition, $\|T\| = 1$. Since $\kappa = \|\dot{T}\| > 0$, we have

$$\|N\| = \frac{1}{\kappa}\|\dot{T}\| = \frac{\|\dot{T}\|}{\|\dot{T}\|} = 1.$$

We saw that T and N are orthogonal; that is, $T \bullet N = 0$. Now $B = T \times N$ is orthogonal to both T and N, and we have

$$\|B\| = \|T \times N\| = \sqrt{\|T\|^2\|N\|^2 - (T \bullet N)^2} = \sqrt{1 - 0} = 1.$$

□

The ordered set (T, N, B) is a *frame field*, called the **Serret-Frenet frame field**, on (the unit speed curve) β. The Serret-Frenet frame field on β is full of information about β.

NOTE : The moving trihedron (with its curvature and torsion functions) was introduced in 1847 by JEAN-FRÉDÉRIC FRENET (1816-1900) and independently by JOSEPH SERRET (1819-1885) in 1851.

The Serret-Frenet formulas

Let $\beta : J \to \mathbb{R}^3$ be a unit speed curve with nonzero curvature (i.e., $\kappa > 0$) and consider the associated Serret-Frenet frame field (T, N, B). The measurement of how $T, N,$ and B vary as we move along (the trace of) the curve β will tell us how the curve itself turns and twists through space. The variation of $T, N,$ and B will be determined by calculating the derivatives $\dot{T}, \dot{N},$ and $\dot{B}$. We already know

$$\dot{T} = \kappa N$$

by definition of N. So the curvature κ describes $T's$ variation in direction.

◇ **Exercise 99** Show that $\dot{B} \bullet B = 0$ and $\dot{B} \bullet T = 0$.

Because $\dot{B}$ is orthogonal to both B and T, it follows that, at each point, $\dot{B}$ is a scalar multiple of N.

2.4.7 DEFINITION. The real-valued function $\tau : J \to \mathbb{R}^3$ given by

$$\dot{B} = -\tau N$$

is called the **torsion function** of β. The minus sign is traditional.

NOTE : By contrast with curvature, there is no restriction on the values of τ : it can be positive, negative, or zero at various points of I. We shall show that the torsion function τ does measure the twisting (or torsion) of the curve β.

For a unit speed curve $\beta : J \to \mathbb{R}^3$, the associated collection

$$\{\kappa, \tau, T, N, B\}$$

is called the **Serret-Frenet apparatus** of β.

2.4.8 EXAMPLE. Consider the arc length parametrization of a circle of radius a

$$\beta(s) = \left(a \cos \frac{s}{a}, a \sin \frac{s}{a}, 0\right).$$

The unit tangent vector field is given by

$$T(s) = \dot{\beta}(s) = \left(-\sin \frac{s}{a}, \cos \frac{s}{a}, 0\right)$$

and

$$\dot{T}(s) = \ddot{\beta}(s) = -\frac{1}{a}\left(\cos \frac{s}{a}, \sin \frac{s}{a}, 0\right).$$

Hence

$$\kappa(s) = \|\dot{T}(s)\| = \frac{1}{a}\cdot$$

It follows that

$$N(s) = \frac{1}{\kappa(s)}\dot{T}(s) = \left(-\cos \frac{s}{a}, -\sin \frac{c}{a}, 0\right).$$

To compute the binormal vector field B, we take the cross product :

$$B(s) = T(s) \times N(s) = e_3 = (0, 0, 1).$$

Hence $-\tau(s)N(s) = \dot{B}(s) = 0$, and therefore $\tau = 0$.

NOTE : For a circle of radius a, the curvature function is *constant* and is equal to $\frac{1}{a}$. This makes sense intuitively since, as a increases, the circle becomes less curved. The limit $\kappa = \frac{1}{a} \to 0$ reflects this. Moreover, the circle has zero torsion. We shall see a general reason for this fact shortly.

◇ **Exercise 100** Compute the *Serret-Frenet apparatus* of the unit speed curve (the *helix*)

$$\beta(s) = \left(a\cos\frac{s}{k}, a\sin\frac{s}{k}, \frac{bs}{k} \right) \quad \text{with} \quad k = \sqrt{a^2 + b^2}.$$

2.4.9 THEOREM. (THE SERRET-FRENET THEOREM) *If* $\beta : J \to \mathbb{R}^3$ *is a unit speed curve with nonzero curvature, then*

$$\begin{aligned} \dot{T} &= \kappa N \\ \dot{N} &= -\kappa T + \tau B \\ \dot{B} &= -\tau N. \end{aligned}$$

PROOF : The first and the third formulas are essentially just the definitions of curvature and torsion. To prove the second formula, we express $\dot{N}$ in terms of $T, N,$ and B :

$$\dot{N} = (\dot{N} \bullet T)T + (\dot{N} \bullet N)N + (\dot{N} \bullet B)B.$$

These coefficients are easily found. Differentiating $N \bullet T = 0$, we get $\dot{N} \bullet T + N \bullet \dot{T} = 0$, and hence

$$\dot{N} \bullet T = -N \bullet \dot{T} = -N \bullet (\kappa N) = -\kappa.$$

As usual, $\dot{N} \bullet N = 0$, since N is a unit vector field. Finally,

$$\dot{N} \bullet B == -N \bullet \dot{B} = -N \bullet (-\tau N) = \tau.$$

□

NOTE : We can record the *Serret-Frenet formulas* more succinctly in the matrix expression

$$\begin{bmatrix} \dot{T} \\ \dot{N} \\ \dot{B} \end{bmatrix} = \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{bmatrix} \begin{bmatrix} T \\ N \\ B \end{bmatrix}$$

or, equivalently,

$$\begin{bmatrix} \dot{T} & \dot{N} & \dot{B} \end{bmatrix} = \begin{bmatrix} T & N & B \end{bmatrix} \begin{bmatrix} 0 & -\kappa & 0 \\ \kappa & 0 & -\tau \\ 0 & \tau & 0 \end{bmatrix}.$$

◇ **Exercise 101** If a rigid body moves along a (unit speed) curve β, then the motion of the body consists of translation along (the trace of) β and rotation about (the trace of) β. The rotation is determined by an *angular velocity vector* ω which satisfies

$$\dot{T} = \omega \times T, \quad \dot{N} = \omega \times N, \quad \text{and} \quad \dot{B} = \omega \times B.$$

The vector ω is called the *Darboux vector*. Show that ω, in terms of $T, N,$ and B, is given by $\omega = \tau T + \kappa B$. (HINT : Write $\omega = aT + bN + cB$ and take cross products with $T, N,$ and B to determine $a, b,$ and c.)

◇ **Exercise 102** Show that

$$\dot{T} \times \ddot{T} = \kappa^2 \, \omega$$

where ω is the *Darboux vector*.

Constraints on curvature and torsion

Constraints on curvature and torsion produce constraints on the geometry of the curve. The simplest constraints are contained in the following two results.

2.4.10 THEOREM. *Let $\beta : J \to \mathbb{R}^3$ be a unit speed curve. Then*

$$\kappa = 0 \quad \textit{if and only if} \quad \beta \textit{ is a (part of a) line.}$$

PROOF : ($\Rightarrow$) Suppose $\kappa = 0$. Then $\dot{T} = 0$ by the Serret-Frenet formulas, and so $T = v$ is a constant (with $\|v\| = 1$ since β has unit speed). But

$$\dot{\beta}(s) = T = v \;\Rightarrow\; \beta(s) = p + sv$$

with p a constant of integration. Hence β is (the arc length parametrization of) a line.

($\Leftarrow$) Suppose β is (the arc length parametrization of) a line. Then $\beta(s) = p + sv$ with $\|v\| = 1$ (so β has unit speed). It follows that

$$T(s) = \dot{\beta}(s) = v = \text{ constant}$$

and so $\dot{T} = 0 = \kappa N$, and hence $\kappa = 0$. □

A **plane curve** in $\mathbb{R}^3$ is a curve that *lies* in a single *plane* of $\mathbb{R}^3$. That is, the trace of the curve is a subset of a certain plane of $\mathbb{R}^3$. Clearly, the straight line and the circle are plane curves.

2.4.11 THEOREM. *Let $\beta : J \to \mathbb{R}^3$ be a unit speed curve with nonzero curvature. Then*

$$\tau = 0 \quad \textit{if and only if} \quad \beta \textit{ is a plane curve.}$$

PROOF : ($\Rightarrow$) Suppose $\tau = 0$. Then, by the Serret-Frenet formulas, $\dot{B} = 0$ and so B is constant (parallel). But this means that $\beta(s)$ should always lie in the plane through $\beta(0)$ orthogonal to B. We show this.

Take the plane determined by the point $\beta(0)$ and the normal vector B. Recall that a point p is in this plane if $(p - \beta(0)) \bullet B = 0$. Consider the real-valued function

$$f(s) := (\beta(s) - \beta(0)) \bullet B$$

for all s. Then

$$\dot{f}(s) = (\beta(s) - \beta(0))^{\cdot} \bullet B + (\beta(s) - \beta(0)) \bullet \dot{B} = \dot{\beta}(s) \bullet B = T \bullet B = 0.$$

Hence $f(s) =$ constant. To identify the constant, evaluate

$$f(0) = (\beta(0) - \beta(0)) = 0.$$

Then (for all s) $(\beta(s) - \beta(0)) \bullet B = 0$ and hence $\beta(s)$ is in the plane determined by $\beta(0)$ and the (constant) vector B.

($\Leftarrow$) Suppose β lies in a plane. Then the plane is determined by a point p and a normal vector $n \neq 0$. Since β lies in the plane, we have (for all s)

$$(\beta(s) - p) \bullet n = 0.$$

By differentiating, we get

$$\dot{\beta}(s) \bullet n = \ddot{\beta}(s) \bullet n = 0.$$

That is, $T \bullet n$ and $\kappa N \bullet n = 0$. These equations say that n is orthogonal to both T and N. Thus n is collinear to B and

$$B = \pm \frac{1}{\|n\|} n.$$

Hence $\dot{B} = 0$ and the Serret-Frenet formulas then give $\tau = 0$. □

We now see that curvature measures the deviation of a curve from being a (straight) line and torsion the deviation of a curve from being contained in a plane. We know that the standard circle of radius a in the x_1x_2-plane in $\mathbb{R}^3$ has $\kappa = \frac{1}{a}$ and $\tau = 0$. To see that a circle located anywhere in $\mathbb{R}^3$ has these properties we have two choices. We could give a parametrization for an arbitrary circle in $\mathbb{R}^3$ or we could use the familiar definition of a circle as the *locus* of points in a plane equally distanced from a fixed point (in the plane). In order to emphasize geometry, we take the latter approach.

2.4.12 THEOREM. *Let $\beta : J \to \mathbb{R}^3$ be a unit speed curve. Then (the trace of) β is a part of a circle if and only if $\kappa > 0$ is constant and $\tau = 0$.*

PROOF : ($\Rightarrow$) Suppose (the trace of) β is part of a circle. By definition, β is a plane curve, so $\tau = 0$. Also by definition (for all s) $\|\beta(s) - p\| = r$. Squaring both sides gives $(\beta(s) - p) \bullet (\beta(s) - p) = a^2$. If we differentiate this expression, we get (for $T = \dot{\beta}$)

$$2T \bullet (\beta(s) - p) = 0 \quad \text{or} \quad T \bullet (\beta(s) - p) = 0.$$

If we differentiate again, then we obtain

$$\begin{aligned} \dot{T} \bullet (\beta(s) - p) + T \bullet T &= 0 \\ \kappa N \bullet (\beta(s) - p) + 1 &= 0 \qquad (*) \\ \kappa N \bullet (\beta(s) - p) &= -1. \end{aligned}$$

This means, in particular, that $\kappa > 0$ and $N \bullet (\beta(s) - p) \neq 0$. Now differentiating $(*)$ produces

$$\begin{aligned} \frac{d\kappa}{ds} N \bullet (\beta(s) - p) + \kappa \dot{N} \bullet (\beta(s) - p) + \kappa N \bullet T &= 0 \\ \frac{d\kappa}{ds} N \bullet (\beta(s) - p) + \kappa(-\kappa T + \tau B) \bullet (\beta(s) - p) + 0 &= 0. \end{aligned}$$

Since $\tau = 0$ and $T \bullet (\beta(s) - p) = 0$ by above, we have

$$\frac{d\kappa}{ds} N \bullet (\beta(s) - p) = 0.$$

Also $N \bullet (\beta(s) - p) \neq 0$ by above, and so $\dfrac{d\kappa}{ds} = 0$. This means, of course, that $\kappa > 0$ is constant.

$(\Leftarrow)$ Suppose now that $\tau = 0$ and $\kappa > 0$ is constant. To show $\beta(s)$ is part of a circle we must show that each $\beta(s)$ is a fixed distance from a fixed point.

For the standard circle, from any point on the circle to the center we proceed in the normal direction a distance equal to the radius. That is, we go $aN = \frac{1}{\kappa} N$. We do the same here.

Let γ denote the curve

$$\gamma(s) := \beta(s) + \frac{1}{\kappa} N.$$

Since we want γ to be a single *point* (the center of the desired circle), we must have $\dot{\gamma}(s) = 0$. Computing, we obtain

$$\begin{aligned} \dot{\gamma}(s) &= \dot{\beta}(s) + \frac{1}{\kappa} \dot{N} \\ &= T + \frac{1}{\kappa}(-\kappa T + \tau B) \\ &= T - T \\ &= 0. \end{aligned}$$

Hence $\gamma(s)$ is a constant p. Then we have

$$\|\beta(s) - p\| = \left\| -\frac{1}{\kappa} N \right\| = \frac{1}{\kappa}$$

so p is the center of a circle $\beta(s)$ of radius $\frac{1}{\kappa}$. $\square$

◇ **Exercise 103** Compute the *Serret-Frenet apparatus* of the unit speed curve

$$\beta(s) = \left(\frac{4}{5}\cos s, 1 - \sin s, -\frac{3}{5}\cos s\right).$$

◇ **Exercise 104** Let β be a unit speed curve which lies entirely on the sphere of radius a centered at the origin. Show that the curvature κ is such that $\kappa \geq \frac{1}{a}$. (HINT : Differentiate $\beta \bullet \beta = a^2$ and use the Serret-Frenet formulas to get $\kappa\, \beta \bullet N = -1$.)

◇ **Exercise 105** Let β be a unit speed curve which lies entirely on the sphere of center p and radius a. Show that, if $\tau \neq 0$, then

$$\beta(s) - p = -\frac{1}{\kappa}N - \left(\frac{1}{\kappa}\right)^{\cdot}\frac{1}{\tau}B \quad \text{and} \quad a^2 = \left(\frac{1}{\kappa}\right)^2 + \left(\left(\frac{1}{\kappa}\right)^{\cdot}\frac{1}{\tau}\right)^2.$$

◇ **Exercise 106** Show that, if

$$\left(\frac{1}{\kappa}\right)^{\cdot} \neq 0 \quad \text{and} \quad \left(\frac{1}{\kappa}\right)^2 + \left(\left(\frac{1}{\kappa}\right)^{\cdot}\frac{1}{\tau}\right)^2 \quad \text{is a constant}$$

then the unit speed curve β lies entirely on a sphere. (HINT : Show that the "center curve" $\gamma(s) := \beta(s) + \left(\frac{1}{\kappa}\right)^{\cdot}\frac{1}{\tau}B$ is constant.)

◇ **Exercise 107** Find the curvature κ and torsion τ for the curve

$$\beta(s) = \left(\frac{1}{\sqrt{2}}\cos s, \sin s, \frac{1}{\sqrt{2}}\cos s\right).$$

Identify the curve.

2.5 The Fundamental Theorem for Curves

Recall the notion of a vector field on a curve. If X is a vector field on $\alpha : J \to \mathbb{R}^3$ and F is an isometry, then $\widetilde{X} = F_*(X)$ is a vector field on the image curve $\widetilde{\alpha} = F(\alpha)$. In fact, for each $t \in J$, $X(t)$ is a tangent vector to $\mathbb{E}^3$ at the point $\alpha(t)$. But then $\widetilde{X}(t) = F_*(X(t))$ is a tangent vector to $\mathbb{E}^3$ at the point $F(\alpha(t)) = \widetilde{\alpha}(t)$.

Isometries preserve the *derivatives* of such vector fields.

◇ **Exercise 108** If X is a vector field on a curve α in $\mathbb{R}^3$ and F is an isometry on $\mathbb{R}^3$, then $\widetilde{X} = F_*(X)$ is a vector field on $\widetilde{\alpha} = F(\alpha)$. Show that

$$\dot{\widetilde{X}} = F_*(\dot{X}).$$

It follows immediately that (if we set $X = \dot{\alpha}$)

$$\ddot{\widetilde{\alpha}} = \dot{\widetilde{X}} = F_*(\dot{X}) = F_*(\ddot{\alpha}).$$

That is, *isometries preserve acceleration.* Now we show that the Serret-Frenet apparatus of a curve is preserved by isometries.

NOTE : This is certainly to be expected on intuitive grounds, since a rigid motion ought to carry one curve into another that turns and twists in exactly the same way. And this is what happens *when the isometry is orientation-preserving.*

2.5.1 PROPOSITION. *Let β be a unit speed curve on $\mathbb{R}^3$ with nonzero curvature and let $\widetilde{\beta} = F(\beta)$ be the image curve of β under the isometry F on $\mathbb{R}^3$. Then*

$$\begin{aligned} \widetilde{\kappa} &= \kappa, \quad \widetilde{T} = F_*(T) \\ \widetilde{\tau} &= (\operatorname{sgn} F)\,\tau, \quad \widetilde{N} = F_*(N) \\ \widetilde{B} &= (\operatorname{sgn} F)\,F_*(B). \end{aligned}$$

PROOF : Observe first that $\widetilde{\beta}$ is also a unit speed curve, since

$$\|\dot{\widetilde{\beta}}\| = \|F_*(\dot{\beta})\| = \|\dot{\beta}\| = 1.$$

Thus

$$\widetilde{T} = \dot{\widetilde{\beta}} = F_*(\dot{\beta}) = F_*(T).$$

Since F_* preserves both acceleration and norms, it follows from the definition of curvature that

$$\widetilde{\kappa} = \|\ddot{\widetilde{\beta}}\| = \|F_*(\ddot{\beta})\| = \|\ddot{\beta}\| = \kappa.$$

To get the full Serret-Frenet frame, we now use the hypothesis $\kappa > 0$ (which implies $\widetilde{\kappa} > 0$, since $\widetilde{\kappa} = \kappa$). By definition, $N = \frac{1}{\kappa}\ddot{\beta}$ and hence

$$\widetilde{N} = \frac{\ddot{\widetilde{\beta}}}{\widetilde{\kappa}} = \frac{F_*(\ddot{\beta})}{\kappa} = F_*\left(\frac{\ddot{\beta}}{\kappa}\right) = F_*(N).$$

It remains only to prove the interesting cases B and τ. We have

$$\widetilde{B} = \widetilde{T} \times \widetilde{N} = F_*(T) \times F_*(N) = (\operatorname{sgn} F)\,F_*(T \times N) = (\operatorname{sgn} F)\,F_*(B).$$

Furthermore,

$$\widetilde{\tau} = \widetilde{B} \bullet \dot{\widetilde{N}} = (\operatorname{sgn} F)\, F_*(B) \bullet F_*(\dot{N}) = (\operatorname{sgn} F)\, B \bullet \dot{N} = (\operatorname{sgn} F)\, \tau.$$

□

NOTE : The presence of $\operatorname{sgn} F$ in the formula for the torsion of $F(\beta)$ shows that the torsion of a curve gives more subtle description of the curve than has been apparent so far. *The sign of τ measures the orientation of the twisting of the curve.*

We have seen that curvature and torsion, individually and in combination, tell us a great deal about the geometry of a curve. In fact, in a very real sense, they tell us everything. Precisely, if two unit speed curves have the same curvature and torsion functions, then there is a *rigid motion* of $\mathbb{R}^3$ taking one curve onto another. Furthermore, given specified curvature and torsion functions, there is a curve which realizes them as its own curvature and torsion. These results are, essentially, theorems about existence and uniqueness of solutions of systems of differential equations. The Serret-Frenet formulas provide the system and the unique solution provides the curve.

2.5.2 THEOREM. (THE FUNDAMENTAL THEOREM) *Let $\kappa, \tau : (a, b) \to \mathbb{R}$ be (smooth) functions with $\kappa > 0$. Then there exists a regular curve $\beta : (a, b) \to \mathbb{R}^3$ parametrized by arc length such that κ is the curvature function and τ is the torsion function of β. Moreover, any other curve $\bar{\beta}$ satisfying the same conditions, differs from β by a proper rigid motion; that is, there exists a direct isometry F, $x \mapsto Ax + c$ such that (for all s)*

$$\beta(s) = A\bar{\beta}(s) + c.$$

PROOF : Consider the matrix-valued function

$$g(s) := \begin{bmatrix} 0 & \kappa(s) & 0 \\ -\kappa(s) & 0 & \tau(s) \\ 0 & -\tau(s) & 0 \end{bmatrix} = \left[a_{ij}\right].$$

If we write

$$\xi_1 = T, \quad \xi_2 = N, \quad \text{and} \quad \xi_3 = B$$

then the Serret-Frenet formulas give us the system of differential equations

$$\begin{aligned}\dot{\xi}_1 &= a_{11}\xi_1 + a_{12}\xi_2 + a_{13}\xi_3\\ \dot{\xi}_2 &= a_{21}\xi_1 + a_{22}\xi_2 + a_{23}\xi_3\\ \dot{\xi}_3 &= a_{31}\xi_1 + a_{32}\xi_2 + a_{33}\xi_3\end{aligned}$$

or, equivalently, the vector differential equation

$$\begin{bmatrix}\dot{\xi}_1\\ \dot{\xi}_2\\ \dot{\xi}_3\end{bmatrix} = g(s)\begin{bmatrix}\xi_1\\ \xi_2\\ \xi_3\end{bmatrix}.$$

It is known that if the matrix-valued function $s \mapsto g(s)$ is *continuous*, then the diferential equation

$$\dot{\xi} = g(s)\xi, \quad s \in (a,b)$$

has solutions $\xi : (a,b) \to \mathbb{E}^3$.

Thus there is a *solution* $(\xi_1(s), \xi_2(s), \xi_3(s))$ dependent upon the initial conditions. For a value $s_0 \in (a,b)$, we may take $(\xi_1(s_0), \xi_2(s_0), \xi_3(s_0))$ to be a choice of a positively-oriented frame on $\mathbb{R}^3$. We next show that (for every s) the solution $(\xi_1(s), \xi_2(s), \xi_3(s))$ is a frame. Observe that

$$(\xi_i \bullet \xi_j)^{\cdot} = \sum_{k=1}^{3}(a_{ik}\xi_j \bullet \xi_k + a_{jk}\xi_i \bullet \xi_k), \quad i,j = 1,2,3.$$

Let $\xi_{ij} := \xi_i \bullet \xi_j$. We obtain the system of differential equations

$$\dot{\xi}_{ij} = \sum_{k=1}^{3}(a_{ik}\xi_{jk} + a_{jk}\xi_{ik}), \quad i,j = 1,2,3 \qquad (*)$$

with initial conditions

$$\xi_{ij} = \delta_{ij} \quad \text{(the Kronecker delta function)}.$$

In order to have a frame we need to show that $\xi_{ij}(s) = \delta_{ij}$ holds for all s. But notice

$$\dot{\delta}_{ij} = 0 = a_{ij} + a_{ji} = \sum_{k=1}^{3}(a_{ik}\delta_{jk} + a_{jk}\delta_{ik}), \quad i,j = 1,2,3$$

which holds by the skew-symmetry of the matrix $g(s)$. Thus $\delta_{ij},\ i,j=1,2,3$ satisfies the differential equation $(*)$ and so, by the uniqueness of solutions to differential equations, we have (for all s) a frame.

To define the curve $\beta : (a,b) \to \mathbb{R}^3$ we integrate

$$\beta(s) = \int_{s_0}^{s} \xi_1(\sigma)\, d\sigma.$$

Then $\dot{\beta}(s) = \xi_1(s) = T(s)$ and so

$$\ddot{\beta} = \dot{\xi}_1 = \kappa\, \xi_2 = \kappa\, N.$$

Thus $\kappa(s)$ is the curvature of β (at s) and hence $\kappa_\beta = \kappa$.

$\diamond$ **Exercise 109** Show that if β is a unit speed curve with nonzero curvature, then

$$\tau = \frac{\dot{\beta} \bullet \ddot{\beta} \times \dddot{\beta}}{\kappa^2}.$$

(HINT : Compute $\dot{\beta} \times \ddot{\beta}$ and $\dddot{\beta}$ (in terms of $T, N,$ and B) and then dot them.)

It follows immediately that $\tau(s)$ is the torsion of β (at s) and hence $\tau_\beta = \tau$.

Now assume that two (unit speed) curves β and $\bar{\beta}$ satisfy the conditions

$$\kappa_\beta = \kappa_{\bar{\beta}} = \kappa \quad \text{and} \quad \tau_\beta = \tau_{\bar{\beta}} = \tau.$$

Let (T_0, N_0, B_0) and $(\bar{T}_0, \bar{N}_0, \bar{B}_0)$ be the Serret-Frenet frames at $s_0 \in I = (a,b)$ of β and $\bar{\beta}$, respectively. By THEOREM 2.1.9, there is a (proper) rigid motion $F,\ x \mapsto Ax + c$ on $\mathbb{E}^3$ which takes $\bar{\beta}(s_0)$ into $\beta(s_0)$ and $(\bar{T}_0, \bar{N}_0, \bar{B}_0)$ into (T_0, N_0, B_0).

Denote the Serret-Frenet apparatus of $\widetilde{\beta} = F(\bar{\beta})$ by $\{\widetilde{k}, \widetilde{\tau}, \widetilde{T}, \widetilde{N}, \widetilde{B}\}$. Then (from PROPOSITION 2.5.1 and the information above)

$$\begin{aligned}
\widetilde{\beta}(s_0) &= \beta(s_0) \\
\widetilde{\kappa} &= \kappa, \quad \widetilde{T}(s_0) = T_0 \\
\widetilde{\tau} &= \tau, \quad \widetilde{N}(s_0) = N_0 \\
\widetilde{B}(s_0) &= B_0.
\end{aligned}$$

We need to show that the curves β and $\widetilde{\beta}$ *coincide*; that is (for all s)

$$\beta(s) = \widetilde{\beta}(s) = F(\bar{\beta}(s)) = A\bar{\beta}(s) + c.$$

We shall show that $T = \widetilde{T}$; that is, the curves β and $\widetilde{\beta}$ are *parallel*.

◇ **Exercise 110** Show that if two curves $\beta, \widetilde{\beta} : J \to \mathbb{R}^3$ are parallel and $\beta(s_0) = \widetilde{\beta}(s_0)$ for some $s_0 \in I$, then $\beta = \widetilde{\beta}$.

Consider the real-valued function (on the interval J)

$$f = T \bullet \widetilde{T} + N \bullet \widetilde{N} + B \bullet \widetilde{B}.$$

Since these are *unit* vector fields, the Cauchy-Schwarz inequality shows that

$$T \bullet \widetilde{T} \leq 1.$$

Furthermore, $T \bullet \widetilde{T} = 1$ if and only if $T = \widetilde{T}$. Similar remarks hold for the other two terms in f. Thus *it suffices to show that* f has constant value 3.

◇ **Exercise 111** Show that the real-valued function

$$f = T \bullet \widetilde{T} + N \bullet \widetilde{N} + B \bullet \widetilde{B}$$

has constant value 3. (HINT : Compute $\dot{f} = 0$ and observe that $f(s_0) = 3$.)

□

2.6 Some Remarks

Arbitrary speed curves

Let $\alpha : J \to \mathbb{R}^3$ be a regular curve that does not necessarily have unit speed. We may reparametrize α to get a unit speed curve $\bar{\alpha}$ and then transfer to α the Serret-Frenet apparatus of $\bar{\alpha}$. Explicitly, if s is an arc length function for α, then

$$\alpha(t) = \bar{\alpha}(s(t)) \quad \text{for } t \in J.$$

Let $\{\bar{\kappa}, \bar{\tau}, \bar{T}, \bar{N}, \bar{B}\}$ be the Serret-Frenet apparatus of $\bar{\alpha}$. We now make the following definition (for $\bar{\kappa} > 0$).

2.6.1 DEFINITION. We define (for the regular curve α)

- the **curvature** function : $\kappa(t) := \bar{\kappa}(s(t))$;
- the **torsion** function : $\tau(t) := \bar{\tau}(s(t))$;
- the **unit tangent** vector field : $T(t) := \bar{T}(s(t))$;
- the **principal normal** vector field : $N(t) := \bar{N}(s(t))$;
- the **binormal** vector field : $B(t) := \bar{B}(s(t))$.

NOTE : In general, κ and $\bar{\kappa}$ are different functions, defined on different intervals. But *they give exactly the same description of the turning of the common route of* α *and* $\bar{\alpha}$, since at any point $\alpha(t) = \bar{\alpha}(s(t))$ the numbers $\kappa(t)$ and $\bar{\kappa}(s(t))$ are by definition the same. Similarly with the rest of the Serret-Frenet apparatus; since only the change of parametrization is involved, its fundamental geometric meaning is the same as before.

For purely theoretical work, this simple transference is often all that is needed. Data about α converts into data about the unit speed reparametrization $\bar{\alpha}$; results about $\bar{\alpha}$ convert to results about α. However, for explicit numerical computations – and occasionally for the theory as well – this transference is impractical, since it is rarely possible to find explicit formulas for $\bar{\alpha}$.

The Serret-Frenet formulas are valid *only* for unit speed curves; they tell the rate of change of the frame field (T, N, B) *with respect to arc length.* However, the speed ν of the curve is the proper correction factor in the general case.

2.6.2 PROPOSITION. *If* α *is a regular curve with nonzero curvature, then*

$$\begin{aligned} \dot{T} &= \kappa\nu\, N \\ \dot{N} &= -\kappa\nu\, T + \tau\nu\, B \\ \dot{B} &= -\tau\nu\, N. \end{aligned}$$

PROOF : The *speed* of α is

$$\nu(t) = \|\dot{\alpha}(t)\| = \frac{ds}{dt}.$$

Let $\bar{\alpha}$ be a unit speed reparametrization of α. Then $T(t) = \bar{T}(s(t))$. The chain rule and the usual Serret-Frenet equations give

$$\begin{aligned} \dot{T} = \frac{dT}{dt} &= \frac{d\bar{T}}{ds}\frac{ds}{dt} \\ &= \bar{\kappa}\bar{N}\nu \\ &= \kappa\nu\, N \end{aligned}$$

so the first formula is proved. For the second and third,

$$\begin{aligned} \dot{N} = \frac{dN}{dt} &= \frac{d\bar{N}}{ds}\frac{ds}{dt} \\ &= (-\bar{\kappa}\bar{T} + \bar{\tau}\bar{B})\nu \\ &= -\kappa\nu\, T + \tau\nu\, B \end{aligned}$$

and

$$\begin{aligned} \dot{B} = \frac{dB}{dt} &= \frac{d\bar{B}}{ds}\frac{ds}{dt} \\ &= -\bar{\tau}\bar{N}\nu \\ &= -\tau\nu\, N. \end{aligned}$$

□

Recall that only for a *constant speed* curve is acceleration everywhere orthogonal to velocity. In the general case, we analyze velocity and acceleration by expressing them in terms of the Serret-Frenet frame field.

2.6.3 PROPOSITION. *If α is a regular curve with speed function ν, then the velocity and acceleration of α are given by*

$$\dot{\alpha} = \nu\, T \quad \textit{and} \quad \ddot{\alpha} = \frac{d\nu}{dt}T + \kappa\nu^2\, N.$$

PROOF : Since $\alpha(t) = \bar{\alpha}(s(t))$, the first calculation is

$$\begin{aligned} \dot{\alpha} = \frac{d\alpha}{dt} &= \frac{d\bar{\alpha}}{ds}\frac{ds}{dt} \\ &= \nu\,\bar{T} \\ &= \nu\, T \end{aligned}$$

while the second is

$$\begin{aligned}\ddot{\alpha} = \frac{d\dot{\alpha}}{dt} &= \frac{d\nu}{dt}T + \nu\dot{T} \\ &= \frac{d\nu}{dt}T + \nu\kappa\nu\, N \\ &= \frac{d\nu}{dt}T + \kappa\nu^2\, N.\end{aligned}$$

□

NOTE : The formula $\dot{\alpha} = \nu\, T$ is to be expected since $\dot{\alpha}$ and T are each tangent to the curve and T has unit length, while $\|\dot{\alpha}\| = \nu$. The formula for acceleration is more interesting. By definition, $\ddot{\alpha}$ is the rate of change of the velocity $\dot{\alpha}$, and in general both the length and the direction of $\dot{\alpha}$ are changing. The *tangential component* $(d\nu/dt)T$ of $\ddot{\alpha}$ measures the rate of change of the length of $\dot{\alpha}$ (i.e. of the speed of α). The *normal component* $\kappa\nu^2\, N$ measures the rate of change of the direction of $\dot{\alpha}$. Newton's laws of motion show that these components may be experienced as *forces*.

We now find effectively computable expressions for the Serret-Frenet apparatus. Clearly we have (for an arbitrary speed curve)

$$T = \frac{\dot{\alpha}}{\|\dot{\alpha}\|} \quad \text{and} \quad N = B \times T.$$

We also have

2.6.4 PROPOSITION. *For any regular curve α (with positive curvature)*

$$\begin{aligned}(1) \quad B &= \frac{\dot{\alpha} \times \ddot{\alpha}}{\|\dot{\alpha} \times \ddot{\alpha}\|}; \\ (2) \quad \kappa &= \frac{\|\dot{\alpha} \times \ddot{\alpha}\|}{\|\dot{\alpha}\|^3}; \\ (3) \quad \tau &= \frac{(\dot{\alpha} \times \ddot{\alpha}) \bullet \dddot{\alpha}}{\|\dot{\alpha} \times \ddot{\alpha}\|^2}.\end{aligned}$$

PROOF : For (1), we use the formulas of PROPOSITION 2.6.3 to get

$$\begin{aligned}\dot{\alpha} \times \ddot{\alpha} &= (\nu\, T) \times \left(\frac{d\nu}{dt}T + \kappa\nu^2\, N\right) \\ &= \nu\frac{d\nu}{dt}T \times T + \kappa\nu^3\, T \times N \\ &= 0 + \kappa\nu^3\, B.\end{aligned}$$

Hence $\|\dot\alpha \times \ddot\alpha\| = \kappa\,\nu^3$ and so

$$B = \frac{\dot\alpha \times \ddot\alpha}{\|\dot\alpha \times \ddot\alpha\|}.$$

For (2), we use the expression for $\ddot\alpha$ in PROPOSITION 2.6.3, take cross product with T and note that $T \times T = 0$ to isolate the curvature

$$\begin{aligned} T \times \ddot\alpha &= 0 + \kappa\nu^2\, T \times N \\ \frac{\dot\alpha \times \ddot\alpha}{\|\dot\alpha\|} &= \kappa\nu^2\, T \times N. \end{aligned}$$

We get (by taking norms)

$$\frac{\|\dot\alpha \times \ddot\alpha\|}{\nu} = \kappa\nu^2\,\|B\|$$

and hence

$$\frac{\|\dot\alpha \times \ddot\alpha\|}{\nu^3} = \kappa.$$

For (3), we take the third derivative

$$\begin{aligned} \dddot\alpha &= \left(\frac{d\nu}{dt}T + \kappa\nu^2\, N\right)^{\cdot} \\ &= \frac{d^2\nu}{dt^2}T + \frac{d\nu}{dt}\dot T + \frac{d\kappa}{dt}\nu^2 N + 2\kappa\nu\,\frac{d\nu}{dt}N + \kappa\nu^2\dot N. \end{aligned}$$

Therefore, since $\dot T = \kappa\, N$ and B is othogonal to T and N, we get

$$\begin{aligned} B \bullet \dddot\alpha &= \kappa\nu^2\, B \bullet \dot N \\ &= \kappa\nu^2\, B \bullet (-\kappa\nu\, T + \tau\nu\, B) \\ &= \kappa\tau\nu^3. \end{aligned}$$

Now $\dot\alpha \times \ddot\alpha = \kappa\nu^3\, B$, and so

$$\begin{aligned} (\dot\alpha \times \ddot\alpha) \bullet \dddot\alpha &= \kappa\nu^3\, B \bullet \dddot\alpha \\ &= \kappa\nu^3(\kappa\tau\nu^3) \\ &= \kappa^2\nu^6\tau. \end{aligned}$$

Of course $\|\dot\alpha \times \ddot\alpha\| = \kappa^2\nu^6$, so we have

$$\tau = \frac{(\dot\alpha \times \ddot\alpha) \bullet \dddot\alpha}{\kappa^2\nu^6} = \frac{(\dot\alpha \times \ddot\alpha) \bullet \dddot\alpha}{\|\dot\alpha \times \ddot\alpha\|^2}.$$

□

NOTE : Equation (2) shows that for regular curves, $\|\dot{\alpha} \times \ddot{\alpha}\| > 0$ is equivalent to the usual condition $\kappa > 0$.

2.6.5 EXAMPLE. We compute the Serret-Frenet apparatus of the regular curve

$$\alpha(t) = \left(3t - t^3, 3t^2, 3t + t^3\right).$$

The derivatives are

$$\begin{aligned} \dot{\alpha} &= 3(1 - t^2, 2t, 1 + t^2) \\ \ddot{\alpha} &= 6(-t, 1, t) \\ \dddot{\alpha} &= 6(-1, 0, 1). \end{aligned}$$

We have $\dot{\alpha} \bullet \dot{\alpha} = 18(1 + 2t^2 + t^4)$, and so

$$\nu = \|\dot{\alpha}\| = 3\sqrt{2}(1 + t^2).$$

Next

$$\dot{\alpha} \times \ddot{\alpha} = 18 \begin{vmatrix} E_1 & E_2 & E_3 \\ 1 - t^2 & 2t & 1 + t^2 \\ -t & 1 & t \end{vmatrix} = 18(-1 + t^2, -2t, 1 + t^2)$$

and hence

$$\|\dot{\alpha} \times \ddot{\alpha}\| = 18\sqrt{2}(1 + t^2).$$

The expressions above for $\dot{\alpha} \times \ddot{\alpha}$ and $\dddot{\alpha}$ yield

$$(\dot{\alpha} \times \ddot{\alpha}) \bullet \dddot{\alpha} = 216.$$

It remains only to substitute this data into the formulas above, with N being computed by another cross product. The final results are

$$\begin{aligned} T &= \frac{(1 - t^2, 2t, 1 + t^2)}{\sqrt{2}(1 + t^2)} \\ N &= \frac{(-2t, 1 - t^2, 0)}{1 + t^2} \\ B &= \frac{(-1 + t^2, -2t, 1 + t^2)}{\sqrt{2}(1 + t^2)} \\ \kappa &= \tau = \frac{1}{3(1 + t^2)}. \end{aligned}$$

◇ **Exercise 112** Compute the Serret-Frenet apparatus for each of the following (regular) curves :

(a) $\alpha(t) = (e^t \cos t, e^t \sin t, e^t)$.

(b) $\beta(t) = (\cosh t, \sinh t, t)$ (the hyperbolic helix).

◇ **Exercise 113** If α is a regular curve with *constant* speed $c > 0$, show that

$$T = \frac{1}{c}\,\dot{\alpha}\,; \quad N = \frac{\ddot{\alpha}}{\|\ddot{\alpha}\|}\,; \quad B = \frac{\dot{\alpha} \times \ddot{\alpha}}{c\,\|\ddot{\alpha}\|}\,; \quad \kappa = \frac{1}{c^2}\|\ddot{\alpha}\|\,; \quad \tau = \frac{(\dot{\alpha} \times \ddot{\alpha}) \bullet \dddot{\alpha}}{c^2\,\|\ddot{\alpha}\|^2}\cdot$$

◇ **Exercise 114** Consider the unit speed helix

$$\beta(s) = \left(a\cos\frac{s}{k}, a\sin\frac{s}{k}, \frac{bs}{k}\right), \quad k = \sqrt{a^2 + b^2}$$

and define the curve $\sigma := \dot{\beta}$, the *spherical image* of β. (For every s, the point $\sigma(s)$ lies on the unit sphere $\mathbb{S}^2$, and the *motion* of σ represents the *turning* of β.) Show that

$$\kappa_\sigma = \sqrt{1 + \left(\frac{\tau_\beta}{\kappa_\beta}\right)^2} \geq 1$$

(and thus depends *only* on the ratio of torsion to curvature for the original curve).

Some implications of curvature and torsion

There are instances in which the ratio of torsion to curvature (for a certain curve) plays an important role (see **Exercise 107**). This ratio can be used to characterize an entire class of regular curves, called *cylindrical helices*.

◇ **Exercise 115** Consider the standard *helix*

$$\alpha(t) = (a\cos t, a\sin t, bt)\,.$$

Verify that the angle θ between the unit tangent vector $T = \frac{\dot{\alpha}}{\|\dot{\alpha}\|}$ of α and the standard unit vector e_3 is constant.

A cylindrical helix is a generalization of a standard (or circular) helix. We make the following definition.

2.6.6 Definition. A regular curve $\alpha : J \to \mathbb{R}^3$ is called a **cylindrical helix** provided the unit tangent vector T of α has constant angle θ with some fixed unit vector u; that is, $T(t) \bullet u = \cos\theta$ for all $t \in J$.

The condition is not altered by reparametrization, so without loss of generality we may assume that cylindrical helices have unit speed. Cylindrical helices can be identified by a simple condition on torsion and curvature.

2.6.7 PROPOSITION. *Let $\beta : J \to \mathbb{R}^3$ be a unit speed curve with nonzero curvature. Then*

$$\beta \text{ is a cylindrical helix} \quad \text{if and only if} \quad \frac{\tau}{\kappa} \text{ is constant.}$$

PROOF : ($\Rightarrow$) If β is a cylindrical helix with $T \bullet u = \cos\theta$, then

$$0 = (T \bullet u)^{\cdot} = \dot{T} \bullet u = \kappa\, N \bullet u$$

so $N \bullet u = 0$ since $\kappa > 0$. The unit vector u is orthogonal to N and hence

$$\begin{aligned} u &= (u \bullet T)T + (u \bullet B)B \\ &= \cos\theta\, T + \sin\theta\, B. \end{aligned}$$

By differentiating we obtain

$$\begin{aligned} 0 &= \cos\theta\, \dot{T} + \sin\theta\, \dot{B} \\ &= \kappa \cos\theta\, N - \tau \sin\theta\, N \\ &= (\kappa\cos\theta - \tau\sin\theta)\, N. \end{aligned}$$

Thus $\kappa\cos\theta - \tau\sin\theta = 0$ which gives

$$\frac{\tau}{\kappa} = \cot\theta = \text{ constant.}$$

($\Leftarrow$) Now suppose $\dfrac{\tau}{\kappa}$ is constant. Choose an angle θ such that $\cot\theta = \dfrac{\tau}{\kappa}$. Define $U := \cos\theta\, T + \sin\theta\, B$ to get

$$\dot{U} = (\kappa\cos\theta - \tau\sin\theta)\, N = 0.$$

This parallel vector field U then determines a unit vector u such that

$$T \bullet u = \cos\theta.$$

Thus β is a cylindrical helix. □

NOTE : A regular curve with nonzero curvature is a *circular* helix if and only if both τ and κ are constant. Also, it can be shown that *a regular curve is a cylindrical helix if and only if its spherical image is part of a circle.*

◇ **Exercise 116** Show that the curve

$$\alpha(t) = \left(at,\, bt^2,\, t^3\right)$$

is a cylindrical helix if and only if $4b^4 = 9a^2$. In this case, find the vector u such that $T \bullet u = \text{constant}$.

Plane Curves

Recall that a *plane curve* in $\mathbb{R}^3$ is a curve that lies entirely in a single plane of $\mathbb{R}^3$. The theory of plane curves can be viewed as a special case of the theory of curves in $\mathbb{R}^3$.

NOTE : The Euclidean plane $\mathbb{R}^2$ can be *embedded* in $\mathbb{R}^3$ and thus identified with a subset (plane) of $\mathbb{R}^3$. For instance, we can think of $\mathbb{R}^2$ as the x_1x_2-plane of $\mathbb{R}^3$; that is, we identify the Euclidean plane $\mathbb{R}^2$ with the plane $\{x = (x_1, x_2, x_3) \,|\, x_3 = 0\} \subset \mathbb{R}^3$. Another option is to identify $\mathbb{R}^2$ with the plane $\{x = (x_1, x_2, x_3) \,|\, x_1 = 1\} \subset \mathbb{R}^3$. In this case, it is convenient to represent the point (p_1, p_2) of $\mathbb{R}^2$ as the column matrix (vector) $\begin{bmatrix} 1 \\ p_1 \\ p_2 \end{bmatrix}$.

We can also give an independent treatment of plane curves; this approach has the advantage that the plane $\mathbb{R}^2$ can be taken to be oriented. An *orientation* of $\mathbb{R}^2$ may be given by fixing a *positive* frame at a point $p \in \mathbb{R}^2$; an obvious choice is the *natural frame* (e_1, e_2) at the origin.

Let $\beta : J \to \mathbb{R}^2$ an *oriented* unit speed curve and denote by T the **unit tangent** vector field on β : $T := \dot{\beta}$. We define the **normal** vector field (on β) N by requiring the oriented frame field (T, N) to have the "same orientation" as the plane $\mathbb{R}^2$.

Then the Serret-Frenet formulas become

$$\begin{aligned} \dot{T} &= \kappa N \\ \dot{N} &= -\kappa T \end{aligned}$$

where the real-valued function $\kappa : J \to \mathbb{R}$ is the (signed) **curvature** of β.

NOTE : The curvature κ might be either positive or negative. It is clear that $|\kappa|$ agrees with the curvature in the case of space curves and that κ changes sign when we change either the orientation of β or the orientation of $\mathbb{R}^2$.

◇ **Exercise 117** Show that if $\beta = (\beta_1, \beta_2)$ is a unit speed curve in $\mathbb{R}^2$, then

$$T(s) = (\dot{\beta}_1(s), \dot{\beta}_2(s)) \quad \text{and} \quad N(s) = (-\dot{\beta}_2(s), \dot{\beta}_1(s)).$$

◇ **Exercise 118** Show that the regular plane curve $\alpha(t) = (x_1(t), x_2(t))$ has curvature

$$\kappa(t) = \frac{\dot{x}_1(t)\ddot{x}_2(t) - \ddot{x}_1(t)\dot{x}_2(t)}{(\dot{x}_1^2(t) + \dot{x}_2^2(t))^{3/2}}$$

at $\alpha(t)$.

2.6.8 EXAMPLE. The curvature of the (standard) *ellipse* $\alpha(t) = (a\cos t, b\sin t)$ is given by

$$\kappa(t) = \frac{ab}{(a^2\sin^2 t + b^2\cos^2 t)^{3/2}} = \frac{ab}{\sqrt{(a^2 + (b^2 - a^2)\cos^2 t)^3}}.$$

Recall that $0 < b < a$ and hence observe that that the curvature achieves a minimum when $t = \pm\frac{\pi}{2}$ and a maximum when $t = 0$ or π.

Consider the example of a (regular) plane curve given by the graph of a (differentiable) function

$$\alpha(t) = (t, f(t)).$$

Here $\|\dot{\alpha}(t)\| = \sqrt{1 + \dot{f}^2}$ and so we can compute

$$\begin{aligned} T &= \left(\frac{1}{\sqrt{1 + \dot{f}^2(t)}}, \frac{\dot{f}(t)}{\sqrt{1 + \dot{f}^2(t)}} \right) \\ N &= \left(\frac{-\dot{f}(t)}{\sqrt{1 + \dot{f}^2(t)}}, \frac{1}{\sqrt{1 + \dot{f}^2(t)}} \right) \\ \dot{T} &= \frac{\ddot{f}(t)}{1 + \dot{f}^2(t)} \left(\frac{-\dot{f}(t)}{\sqrt{1 + \dot{f}^2(t)}}, \frac{1}{\sqrt{1 + \dot{f}^2(t)}} \right). \end{aligned}$$

Hence

$$\kappa(t) = \frac{\ddot{f}(t)}{\left(1 + \dot{f}^2(t)\right)^{3/2}}.$$

Observe that the sign of the curvature is determined by the second derivative $\ddot{f}(t)$, which is positive if $f(t)$ is concave up, negative if $f(t)$ is concave down. Since *any curve in* $\mathbb{R}^2$ *is locally the graph of a function*, we see that the signed curvature at a point is positive if the curve turns to left of the tangent, negative if to the right.

◇ **Exercise 119** Compute the curvature of the *semicircle*

$$x_2 = \sqrt{a^2 - x_1^2}.$$

◇ **Exercise 120** Show that the curvature of the (cuspidal) *cycloid* $t \mapsto (t - \sin t, 1 - \cos t)$ (at a regular value t) is given by

$$\kappa(t) = -\frac{1}{4 \sin \frac{t}{2}}.$$

◇ **Exercise 121** Find a formula for the curvature of the *parabola* $x_1 = at^2$, $x_2 = 2at$ (with $a > 0$). Show that the vertex is the unique point on the parabola where the curvature assumes a maximal value.

In the case of a plane curve, the proof of the FUNDAMENTAL THEOREM is actually very simple.

◇ **Exercise 122** Given a smooth function $\kappa : (a, b) \to \mathbb{R}$, show that the plane curve $\beta : (a, b) \to \mathbb{R}^2$, parametrized by arc length and having κ as (directed) curvature function, is given by

$$\beta(s) = \left(\int \cos \theta(s)\, ds + c_1, \int \sin \theta(s)\, ds + c_2 \right)$$

where

$$\theta(s) = \int \kappa(s)\, ds + \varphi.$$

Furthermore, any other curve curve $\bar{\beta}$ satisfying the same conditions, differs from β by a rotation of angle φ followed by a translation by vector $c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$.

Chapter 3

Submanifolds

Topics :

1. Euclidean m-Space
2. Linear Submanifolds
3. The Inverse Mapping Theorem
4. Smooth Submanifolds

3.1 Euclidean m-Space

Let $\mathbb{R}$ be the *set* of real numbers and let $\mathbb{R}^m$ $(m \geq 1)$ denote the Cartesian product of m copies of $\mathbb{R}$. Clearly, $\mathbb{R}^1 = \mathbb{R}$. The elements of $\mathbb{R}^m$ are ordered m-tuples of real numbers. Under the usual addition and scalar multiplication, $\mathbb{R}^m$ is a *vector space* over $\mathbb{R}$.

NOTE : The set $\mathbb{R}^m$ may be equipped with various "natural" structures (e.g., group structure, vector space structure, topological structure, etc.) thus yielding various *spaces* (having the same underlying set) $\mathbb{R}^m$. We must usually decide from the context which structure is intended. We shall find it convenient to refer to the vector space $\mathbb{R}^m$ equipped with its *canonical topology* as the **Cartesian m-space**.

For $0 < \ell < m$ the *canonical inclusion* $\mathbb{R}^\ell \hookrightarrow \mathbb{R}^m$ is defined as the map $(x_1, \ldots, x_\ell) \mapsto (x_1, \ldots, x_\ell, 0, \ldots, 0)$. Similarly, the map $(x_1, \ldots, x_\ell, \ldots x_m) \mapsto ((x_1, \ldots, x_\ell), (x_{\ell+1}, \ldots x_m))$ defines a *canonical isomorphism* between (vector spaces) $\mathbb{R}^m$ and $\mathbb{R}^\ell \times \mathbb{R}^{m-\ell}$. We write $\mathbb{R}^m = \mathbb{R}^\ell \times \mathbb{R}^{m-\ell}$.

The concept of Euclidean (2- or 3-dimensional) space extends straightforwardly to higher dimensions. We make the following definition.

3.1.1 DEFINITION. The (standard) **Euclidean m-space** is the set $\mathbb{R}^m$ together with the *Euclidean distance* between points $x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_m)$ given by

$$d(x, y) = \sqrt{(y_1 - x_1)^2 + \cdots + (y_m - x_m)^2}.$$

The distance function $d : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}, \quad (x, y) \mapsto d(x, y)$ is a *metric* (see **Exercise 7**) and hence *Euclidean m-space $\mathbb{R}^m$ is a metric space.*

NOTE : Any metric space is a topological space and so any (standard) Euclidean space is, by definition, a Cartesian space. It is important to realize that these two structures are *distinct* : a Euclidean space has "more structure" than a Cartesian space; this distinction will subsequently play an important role.

We denote the *open ball* of center p and radius $\rho > 0$ by

$$\mathcal{B}(x, \rho) := \{x \in \mathbb{R}^m \mid d(x, p) < \rho\}.$$

It turns out that the *open sets* are exactly the (arbitrary) unions of such open balls. In the usual sense one can introduce concepts like *closed sets*, *connected sets*, *convergence* (of sequences), *completeness*, *compact sets*, etc. Also, one can speak of *continuous mappings*.

Under the usual addition and scalar multiplication, Euclidean m-space $\mathbb{R}^m$ is a *vector space*. This vector space is rather special in the sense that it has a built-in positive definite *inner product* (i.e., a positive definite symmetric bilinear form), the so-called **dot product**,

$$x \bullet y := x_1y_1 + x_2y_2 + \cdots + x_my_m$$

and an *orthonormal basis*

$$\{e_1, e_2, \ldots, e_m\} \quad \text{with} \quad e_i \bullet e_j = \delta_{ij}.$$

NOTE : (1) The Euclidean metric d can be defined using the standard inner product on $\mathbb{R}^m$. We define $\|x\|$, the *norm* of the element (vector) x, by $\|x\| = \sqrt{x \bullet x}$. Then we have

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

This notation is frequently useful even when we are dealing with the Euclidean m-space $\mathbb{R}^m$ as a metric space and not using its vector space structure. In particular, $\|x\| = d(x, 0)$.

(2) An *abstract* concept of Euclidean space (i.e., a space satisfying the *axioms* of Euclidean geometry) can be introduced. It is defined as a structure $\left(\mathcal{E}, \vec{E}, \varphi\right)$, consisting of a (nonempty) *set* $\mathcal{E}$, an associated *standard vector space* (which is a real vector space equipped with an arbitrary positive definite inner product $\langle \cdot, \cdot \rangle$), and a *structure map*

$$\varphi : \mathcal{E} \times \mathcal{E} \to \vec{E}, \quad (p, q) \mapsto \overrightarrow{pq}$$

such that

(AS1) $\overrightarrow{pq} + \overrightarrow{qr} = \overrightarrow{pr}$ for every $p, q, r \in \mathcal{E}$;

(AS2) For every $o \in \mathcal{E}$ and every $v \in \vec{E}$, there is a unique $p \in \mathcal{E}$ such that $\overrightarrow{op} = v$.

Elements of $\mathcal{E}$ are called *points*, whereas elements of $\vec{E}$ are called *vectors*. ($\overrightarrow{op}$ is the position vector of p with the initial point o.) The *dimension* of $\mathcal{E}$ is the dimension of (the vector space) $\vec{E}$. It turns out that

(i) if we fix an arbitrary point $o \in \mathcal{E}$, there is a one-to-one correspondence between $\mathcal{E}$ and $\vec{E}$ (the mapping $p \mapsto \overrightarrow{op}$ is a bijection);

(ii) in addition, if we fix an arbitrary orthonormal basis $e_1, e_2, \ldots, e_m$ of $\vec{E}$, the (inner product) spaces $\vec{E}$ and $\mathbb{R}^m$ are *isomorphic*. (In other words, the inner product on $\vec{E}$ "is" a dot product : for $v, w \in \vec{E}$,

$$\begin{aligned} \langle v, w \rangle &= \langle v_1 e_1 + \cdots + v_m e_m, w_1 e_1 + \cdots + w_m e_m \rangle \\ &= v_1 w_1 + \cdots + v_m w_m .) \end{aligned}$$

In this sense, we *identify* the (abstract) m-dimensional Euclidean space $\mathcal{E} = \mathcal{E}^m$ with the (concrete) standard Euclidean m-space $\mathbb{R}^m$.

Elements of Euclidean m-space $\mathbb{R}^m$, when thought of as *points*, will be written as m-tuples. When thought of as *vectors*, they will be written as column m-matrices. Euclidean 1-space $\mathbb{R}^1 = \mathbb{R}$ will be referred to as the *real line*.

Let $U \subseteq \mathbb{R}^m$. Let $x = (x_1, \ldots, x_m)$ denote the general (variable) point of U and let $p = (p_1, \ldots, p_m)$ be a fixed but arbitrary point of U. U is an *open set* if (and only if) for each point $x \in U$ there is an open ball $\mathcal{B}(x, \rho) \subset U$; intuitively, this means that points in U are entirely surrounded by points of U (or that points sufficiently close to points of U still belong to U). Let $\emptyset \neq A \subseteq \mathbb{R}^m$. An *open neighborhood* of A is an open set containing A, and a *neighborhood* of A is any set containing an open neighborhood of A. In particular, a neighborhood of a set $\{p\}$ is also called a neighborhood of the point p.

Henceforth, throughout this chapter, U will denote an open set.

Continuity

A mapping $F : U \subseteq \mathbb{R}^m \to \mathbb{R}^n$ is *continuous* at $p \in U$ if (and only if) given $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$F\left(\mathcal{B}(p, \delta)\right) \subseteq \mathcal{B}(F(p), \varepsilon).$$

In other words, F is continuous at p if (and only if) points arbitrarily close to $F(p)$ are images of points sufficiently close to p. We say that F is **continuous** provided it is continuous at each $p \in U$.

Given a mapping $F : U \subseteq \mathbb{R}^m \to \mathbb{R}^n$, we can determine n *functions* (of m variables) as follows. Let $x = (x_1, \ldots, x_m) \in U$ and $F(x) = (y_1, \ldots, y_n)$. Then we can write

$$y_1 = f_1(x_1, \ldots, x_m), \quad y_2 = f_2(x_1, \ldots, x_m), \quad \ldots, \quad y_n = f_n(x_1, \ldots, x_m).$$

The functions $f_i : U \to \mathbb{R}$, $i = 1, 2, \ldots, n$ are the **component functions** of F. The continuity of the mapping F is equivalent to the continuity of its component functions.

◇ **Exercise 123** Prove that a mapping $F : U \subseteq \mathbb{R}^m \to \mathbb{R}^n$ is continuous if and only if each component function $f_i : U \subseteq \mathbb{R}^m \to \mathbb{R}$, $i = 1, 2, \ldots, n$ is continuous.

The following results are standard (and easy to prove).

3.1.2 PROPOSITION. *Let $F, G : U \subseteq \mathbb{R}^m \to \mathbb{R}^n$ be continuous mappings and let $\lambda \in \mathbb{R}$. Then $F + G$, λF, and $F \bullet G$ are each continuous. If $n = 1$ and $G(x) \neq 0$ for all $x \in U$, then the quotient $\frac{F}{G}$ is also continuous.*

3.1.3 PROPOSITION. *Let $F : U \subseteq \mathbb{R}^\ell \to \mathbb{R}^m$ and $G : V \subseteq \mathbb{R}^m \to \mathbb{R}^n$ be continuous mappings, where U and V are open sets such that $F(U) \subseteq V$. Then $G \circ F$ is a continuous mapping.*

◇ **Exercise 124** Show that the following mappings (or functions) are continuous.

(a) The *identity mapping* $id : \mathbb{R}^m \to \mathbb{R}^m, \quad x \mapsto x$.

(b) The *norm function* $\nu : \mathbb{R}^m \to \mathbb{R}, \quad x \mapsto \|x\|$.

(c) The i^{th} *natural projection* $\mathrm{pr}_i : \mathbb{R}^m \to \mathbb{R}, \quad x \mapsto x_i$.

Hence derive that every *polynomial function* (in several variables)

$$p_k : \mathbb{R}^m \to \mathbb{R}, \quad x = (x_1, \ldots, x_m) \mapsto \sum_{\substack{i_1, \ldots, i_m = 0 \\ i_1 + \cdots + i_m \leq k}}^{k} a_{i_1 \ldots i_m} x_1^{i_1} \ldots x_m^{i_m}$$

is continuous.

NOTE : More generally, every *rational function* (i.e., a quotient of two polynomial functions) is continuous. In can be shown that *elementary* functions like exp, log, sin, and cos are also continuous.

Mappings $L : \mathbb{R}^m \to \mathbb{R}^n$ that preserve the linear structure of the Euclidean space (i.e., *linear* mappings) play an important role in differentiation. Such mappings are continuous (see also **Exercise 128**).

◇ **Exercise 125** Show that every linear mapping $L : \mathbb{R}^m \to \mathbb{R}^n$ is continuous.

In most applications it is convenient to express continuity in terms of neighborhoods instead of open balls.

◇ **Exercise 126** Prove that a mapping $F : U \subseteq \mathbb{R}^m \to \mathbb{R}^n$ is continuous at $p \in U$ if and only if given a neighborhood $\mathcal{N}$ of $F(p)$ in $\mathbb{R}^n$ there exists a neighborhood $\mathcal{M}$ of p in $\mathbb{R}^m$ such that $F(\mathcal{M}) \subseteq \mathcal{N}$.

It is often necessary to deal with mappings (or functions) defined on arbitrary (i.e., not necessarily open) sets. To extend the previous ideas to this situation, we shall proceed as follows.

Let $F : A \subseteq \mathbb{R}^m \to \mathbb{R}^n$ be a mapping, where A is an *arbitrary* set. We say that F is **continuous** on A provided there exists an open set $U \subseteq \mathbb{R}^m$, $A \subseteq U$, and a continuous mapping $\bar{F} : U \to \mathbb{R}^n$ such that the restriction $\bar{F}\big|_A = F$. In other words, F is continuous (on A) if it is the restriction of a continuous mapping defined on an open set containing A.

NOTE : It is clear that if $F : A \subseteq \mathbb{R}^m \to \mathbb{R}^n$ is continuous, then given a neighborhood $\mathbb{N}$ of $F(p)$ in $\mathbb{R}^n$, $p \in A$, there exists a neighborhood $\mathcal{M}$ of p in $\mathbb{R}^m$ such that $F(\mathcal{M} \cap A) \subseteq \mathcal{N}$. For this reason, it is convenient to call the set $W \cap A$ a *neighborhood* of p in A.

We say that a continuous mapping $F : A \subseteq \mathbb{R}^m \to \mathbb{R}^m$ is a **homeomorphism** onto $F(A)$ if F is one-to-one and the inverse $F^{-1} : F(A) \subseteq \mathbb{R}^m \to \mathbb{R}^m$ is continuous. In this case A and $F(A)$ are *homeomorphic* sets.

3.1.4 EXAMPLE. Let $F : \mathbb{R}^3 \to \mathbb{R}^3$ be given by

$$F(x_1, x_2, x_3) = (ax_1, bx_2, cx_3).$$

F is clearly continuous, and the restriction of F to the (unit) *sphere*

$$\mathbb{S}^2 = \left\{x = (x_1, x_2, x_3) \in \mathbb{R}^3 \,|\, x_1^2 + x_2^2 + x_3^2 = 1\right\}$$

is a continuous mapping $\widetilde{F} : \mathbb{S}^2 \to \mathbb{R}^3$. Observe that $\widetilde{F}(\mathbb{S}^2) = E$, where E is the *ellipsoid*

$$E = \left\{ x = (x_1, x_2, x_3) \in \mathbb{R}^3 \mid \frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} + \frac{x_3^2}{c^2} = 1 \right\}.$$

It is also clear that F is one-to-one and that

$$F^{-1}(x_1, x_2, x_3) = \left(\frac{x_1}{a}, \frac{x_2}{b}, \frac{x_3}{c} \right).$$

Thus $\widetilde{F}^{-1} = F^{-1}\big|_E$ is continuous. Therefore, $\widetilde{F}$ is a *homeomorphism* of the sphere $\mathbb{S}^2$ onto the ellipsoid E.

Differentiability

A function $f : U \subseteq \mathbb{R}^m \to \mathbb{R}$ is *differentiable* at $p \in U$ if there exists a linear functional $L_p : \mathbb{R}^m \to \mathbb{R}$ such that

$$\lim_{x \to p} \frac{f(x) - f(p) - L_p(x - p)}{\|x - p\|} = 0$$

or, equivalently, if there exist a linear functional $L_p : \mathbb{R}^m \to \mathbb{R}$ and a function $R(\cdot, p)$, defined on an open neighborhood V of $p \in U$, such that

$$f(x) = f(p) + L_p(x - p) + \|x - p\| \cdot R(x, p), \quad x \in V$$

and

$$\lim_{x \to p} R(x, p) = 0.$$

Then L_p is called a *derivative* (or differential) of f at p. We say that f is **differentiable** provided it is differentiable at each $p \in U$.

NOTE : We think of a derivative L_p as a linear approximation of f near p. By the definition, the error involved in replacing $f(x)$ by $L_p(x - p)$ is negligible compared to the distance from x to p, provided that this distance is sufficiently small.

If $L_p(x) = b_1 x_1 + \cdots + b_m x_m$ is a derivative of f at p, then

$$b_i = \frac{\partial f}{\partial x_i}(p) := \lim_{t \to 0} \frac{1}{t} \left(f(p + te_i) - f(p) \right), \quad i = 1, 2, \ldots, m.$$

In particular, if f is differentiable at p, these *partial derivatives* exist and the derivative L_p is *unique*. We denote by $Df(p)$ (or sometimes $f'(p)$) *the* derivative of f at p, and write (by a slight abuse of notation)

$$Df(p) = \frac{\partial f}{\partial x_1}(p)(x_1 - p_1) + \frac{\partial f}{\partial x_2}(p)(x_2 - p_2) + \cdots + \frac{\partial f}{\partial x_n}(p)(x_m - p_m).$$

◇ **Exercise 127** Show that any linear functional $f : \mathbb{R}^m \to \mathbb{R}$ is differentiable and $Df(p) = f$ for all $p \in \mathbb{R}^m$.

◇ **Exercise 128** Prove that any differentiable function $f : U \subseteq \mathbb{R}^m \to \mathbb{R}$ is continuous.

NOTE : Mere existence of partial derivatives is *not* sufficient for differentiability (of the function f). For example, the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x_1, x_2) = \frac{x_1 x_2}{x_1^2 + x_2^2} \quad \text{and} \quad f(0,0) = 0$$

is *not* continuous at $(0,0)$, yet both partial derivatives are defined there. However, *if all partial derivatives* $\frac{\partial f}{\partial x_i}$, $i = 1, 2, \ldots, m$ *are defined and continuous in a neighborhood of* $p \in U$, *then* f *is differentiable at* p.

If the function $f : U \subseteq \mathbb{R}^m \to \mathbb{R}$ has all partial derivatives continuous (on U) we say that f is **continuously differentiable** (or of *class* C^1) on U. We denote this class of functions by $C^1(U)$. (The class of continuous functions on U is denoted by $C^0(U)$.)

NOTE : We have seen that

$$f \in C^1(U) \Rightarrow f \text{ is differentiable (on } U) \Rightarrow \text{all partial derivatives } \frac{\partial f}{\partial x_i} \text{ exist (on } U)$$

but the converse implications may fail. Many results actually need f to be of class C^1 rather than differentiable.

If $r \geq 1$, the class $C^r(U)$ of functions $f : U \subseteq \mathbb{R}^m \to \mathbb{R}$ that are r-*fold continuously differentiable* (or C^r functions) is specified inductively by requiring that the partial derivatives of f exist and belong to $C^{r-1}(U)$. If f is of class C^r for all r, then we say that f is of *class* C^∞ or simply **smooth**. The class of smooth functions on U is denoted by $C^\infty(U)$.

NOTE : If $f \in C^r(U)$, then (at any point of U) the value of the partial derivatives of order k, $1 < k \leq r$ is independent of the order of differentiation; that is, if $(j_1, \ldots, j_k)$ is a permutation of $(i_1, \ldots, i_k)$, then

$$\frac{\partial^k f}{\partial x_{i_1} \ldots \partial x_{i_k}} = \frac{\partial^k f}{\partial x_{j_1} \ldots \partial x_{j_k}}.$$

We are now interested in extending the notion of differentiability to mappings $F : U \subseteq \mathbb{R}^m \to \mathbb{R}^n$. We say that F is *differentiable* at $p \in U$ if (and only if) its component functions are differentiable at p; that is, by writing

$$F(x_1, \ldots, x_m) = (f_1(x_1, \ldots, x_m), \ldots, f_n(x_1, \ldots, x_m))$$

the functions $f_i : U \to \mathbb{R}$, $i = 1, 2, \ldots, n$ have partial derivatives at $p \in U$. F is **differentiable** provided it is differentiable at each $p \in U$.

The class $C^r(U, \mathbb{E}^n)$, $1 \leq r \leq \infty$ of C^r-mappings $F : U \subseteq \mathbb{R}^m \to \mathbb{R}^n$ is defined in the obvious way. We will be concerned primarily with *smooth* (i.e., of class C^∞) mappings. So if F is a smooth mapping, then its component functions f_i, $i = 1, 2, \ldots, n$ have continuous partial derivatives of all orders and each such derivative is independent of the order of differentiation.

NOTE : For the case $m = 1$, we obtain the notion of (parametrized) *smooth curve* in Euclidean n-space $\mathbb{R}^n$. In Chapter 2, we have already seen such an object in $\mathbb{E}^3$. (Most of the concepts introduced in Chapter 2 can be extended to higher dimensions; in particular, the concept of *tangent vector*.)

Let $T_p\mathbb{R}^m$ be the *tangent space* to $\mathbb{R}^m$ at p; this vector space can be *identified* with $\mathbb{R}^m$ via

$$v_1 \left.\frac{\partial}{\partial x_1}\right|_p + \cdots + v_m \left.\frac{\partial}{\partial x_m}\right|_p \mapsto (v_1, \cdots, v_m).$$

Let $\alpha : U \subseteq \mathbb{R} \to \mathbb{R}^m$ be a smooth (parametrized) curve with component functions $\alpha_1, \ldots, \alpha_m$. The *velocity vector* (or tangent vector) to α at $t \in U$ is the element

$$\dot{\alpha}(t) := \left(\frac{d\alpha_1}{dt}(t), \cdots, \frac{d\alpha_m}{dt}(t)\right) \in T_{\alpha(t)}\mathbb{R}^m.$$

3.1.5 EXAMPLE. Given a point $p \in U \subseteq \mathbb{R}^m$ and a (tangent) vector $v \in T_p\mathbb{R}^m$, we can always find a smooth curve $\alpha : (-\varepsilon, \varepsilon) \to U$ with $\alpha(0) = p$ and $\dot{\alpha}(0) = v$. Simply define $\alpha(t) = p + tv,\ t \in (-\varepsilon, \varepsilon)$. By writing $p = (p_1, \ldots, p_m)$ and $v = (v_1, \ldots, v_m)$, the component functions of α are $\alpha_i(t) = p_i + tv_i,\ i = 1, 2, \ldots, m$. Thus α is smooth, $\alpha(0) = p$ and

$$\dot{\alpha}(0) = \left(\frac{d\alpha_1}{dt}(0), \cdots, \frac{d\alpha_m}{dt}(0) \right) = (v_1, \ldots, v_m) = v.$$

We shall now introduce the concept of *derivative* (or differential) of a differentiable mapping. Let $F : U \subseteq \mathbb{R}^m \to \mathbb{R}^n$ be a differentiable mapping. To each $p \in U$ we associate a linear mapping

$$DF(p) : \mathbb{R}^m = T_p\mathbb{R}^m \to \mathbb{R}^n = T_{F(p)}\mathbb{R}^n$$

which is called the **derivative** (or *differential*) of F at p and is defined as follows. Let $v \in T_p\mathbb{E}^m$ and let $\alpha : (-\varepsilon, \varepsilon) \to U$ be a differentiable curve such that $\alpha(0) = p$ and $\dot{\alpha}(0) = v$. By the chain rule (for functions), the curve $\beta = F \circ \alpha : (-\varepsilon, \varepsilon) \to \mathbb{E}^n$ is also differentiable. Then

$$DF(p) \cdot v := \dot{\beta}(0).$$

NOTE : The above definition of $DF(p)$ does not depend on the choice of the curve which passes through p with tangent vector v, and $DF(p)$ is, in fact, linear. So

$$DF(p) \cdot v = \frac{d}{dt} F(\alpha(t)) \Big|_{t=0} \in T_{F(p)}\mathbb{R}^n = \mathbb{R}^n.$$

The derivative $DF(p)$ is also denoted by $F_{*,p}$ and called the *tangent mapping* of F at p (see Section 2.1 for the special case when F is an isometry on Euclidean 3-space $\mathbb{R}^3$).

The matrix of the linear mapping $DF(p)$ (relative to bases $\left(\frac{\partial}{\partial x_1}\Big|_p, \ldots, \frac{\partial}{\partial x_m}\Big|_p \right)$ of $T_p\mathbb{R}^m$ and $\left(\frac{\partial}{\partial y_1}\Big|_{F(p)}, \ldots, \frac{\partial}{\partial y_n}\Big|_{F(p)} \right)$ of $T_{F(p)}\mathbb{R}^n$) is the **Jacobian matrix**

$$\frac{\partial F}{\partial x}(p) = \frac{\partial(f_1, \ldots, f_n)}{\partial(x_1, \ldots, x_m)}(p) := \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(p) & \cdots & \frac{\partial f_1}{\partial x_m}(p) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(p) & \cdots & \frac{\partial f_n}{\partial x_m}(p) \end{bmatrix} \in \mathbb{R}^{n \times m}$$

of F at p. When $m = n$ this is a square matrix and its determinant is then defined. This determinant is called the **Jacobian** of F at p and is denoted by $J_F(p)$. Thus

$$J_F(p) = \left| \frac{\partial F}{\partial x}(p) \right| := \det \frac{\partial F}{\partial x}(p).$$

$\diamond$ **Exercise 129** Let $f : I \to \mathbb{R}$ and $g : J \to \mathbb{R}$ be differentiable functions, where I and J are open intervals such that $f(I) \subseteq J$. Show that the function $g \circ f$ is differentiable and (for $t \in I$)

$$(g \circ f)'(t) = g'(f(t)) \cdot f'(t).$$

The standard *chain rule* (for functions) extends to mappings.

3.1.6 PROPOSITION. (THE GENERAL CHAIN RULE) *Let $F : U \subseteq \mathbb{R}^\ell \to \mathbb{R}^m$ and $G : V \subseteq \mathbb{R}^m \to \mathbb{R}^n$ be differentiable mappings, where U and V are open sets such that $F(U) \subseteq V$. Then $G \circ F$ is a differentiable mapping and (for $p \in U$)*

$$D(G \circ F)(p) = DG(F(p)) \circ DF(p).$$

PROOF : The fact that $G \circ F$ is differentiable is a consequence of the chain rule for functions. Now, let $v \in T_p\mathbb{E}^\ell$ be given and let us consider a (differentiable) curve $\alpha : (-\varepsilon, \varepsilon) \to U$ with $\alpha(0) = p$ and $\dot{\alpha}(0) = v$. Set $DF(p) \cdot v = w$ and observe that

$$DG(F(p)) \cdot w = \frac{d}{dt}(G \circ F \circ \alpha)\Big|_{t=0}.$$

Then

$$\begin{aligned} D(G \circ F)(p) \cdot v &= \frac{d}{dt}(G \circ F \circ \alpha)\Big|_{t=0} \\ &= DG(F(p)) \cdot w \\ &= DG(F(p)) \circ DF(p) \cdot v. \end{aligned}$$

□

NOTE : In terms of Jacobian matrices, the general chain rule can be written

$$\frac{\partial (G \circ F)}{\partial x}(p) = \frac{\partial G}{\partial y}(F(p)) \cdot \frac{\partial F}{\partial x}(p).$$

Thus if $H = G \circ F$ and $y = F(x)$, then

$$\frac{\partial H}{\partial x} = \begin{bmatrix} \frac{\partial g_1}{\partial y_1} & \cdots & \frac{\partial g_1}{\partial y_m} \\ \vdots & & \vdots \\ \frac{\partial g_n}{\partial y_1} & \cdots & \frac{\partial g_n}{\partial y_m} \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_\ell} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_\ell} \end{bmatrix}$$

where $\frac{\partial g_1}{\partial y_1}, \ldots, \frac{\partial g_n}{\partial y_m}$ are evaluated at $y = F(x)$ and $\frac{\partial f_1}{\partial x_1}, \cdots, \frac{\partial F_m}{\partial x_\ell}$ at x. Writing this out, we obtain

$$\frac{\partial h_i}{\partial x_j} = \frac{\partial g_i}{\partial y_1}\frac{\partial y_1}{\partial x_j} + \cdots + \frac{\partial g_i}{\partial y_m}\frac{\partial y_m}{\partial x_j} \qquad (i = 1, 2, \ldots, n\,;\, j = 1, 2, \ldots, \ell).$$

◇ **Exercise 130** Let

$$F(x_1, x_2) = (x_1^2 - x_2^2 + x_1x_2, x_2^2 - 1) \quad \text{and} \quad G(y_1, y_2) = (y_1 + y_2, 2y_1, y_2^2).$$

(a) Show that F and G are differentiable, and that $G \circ F$ exists.

(b) Compute $D(G \circ F)(1, 1)$

i. directly

ii. using the chain rule.

◇ **Exercise 131** Show that

(a) if $\sigma : \mathbb{R}^2 \to \mathbb{R}$ is defined by $\sigma(x, y) = x + y$, then $D\sigma(a, b) = \sigma$.

(b) if $\pi : \mathbb{R}^2 \to \mathbb{R}$ is defined by $\pi(x, y) = x \cdot y$, then $D\pi(a, b) \cdot (x, y) = bx + ay$.

Hence deduce that if the functions $f, g : U \subseteq \mathbb{R}^m \to \mathbb{R}$ are differentiable at $p \in U$, then

$$\begin{aligned} D(f + g)(p) &= DF(p) + Dg(p) \\ D(f \cdot g)(p) &= g(p)DF(p) + f(p)DG(p). \end{aligned}$$

If moreover $g(p) \neq 0$, then

$$D\left(\frac{f}{g}\right) = \frac{g(p)DF(p) - f(p)DG(p)}{(g(p))^2}.$$

NOTE : The precise sense in which the derivative $DF(p)$ of the (differentiable) mapping F at p is a linear approximation of F near p is given by the following

result (in which $DF(p)$ is interpreted as a linear mapping from $\mathbb{R}^m$ to $\mathbb{R}^n$) : *If the mapping* $F : U \subseteq \mathbb{R}^m \to \mathbb{R}^n$ *is differentiable, then for each* $p \in U$,

$$\lim_{x \to p} \frac{F(x) - F(p) - DF(p) \cdot (x - p)}{\|x - p\|} = 0.$$

.

If $A \subseteq \mathbb{R}^m$ is an *arbitrary* set, then $C^\infty(A)$ denotes the set of all functions $f : A \to \mathbb{R}$ such that $f = \bar{f}\big|_A$, where $\bar{f} : U \to \mathbb{R}$ is a smooth function on some open neighborhood U of A.

3.2 Linear Submanifolds

Smooth curves in Euclidean 3-space $\mathbb{R}^3$ represent an important class of "geometrically interesting" subsets that are one-dimensional and can be thoroughly studied with the methods of calculus (and linear algebra). The simplest type of such geometric curve is the *line*, which is "straight". A two-dimensional analogue of the line is the *plane*, which is "flat". We shall briefly discuss these two simple cases before considering their natural higher-dimensional analogues, the *linear submanifolds*.

Lines and planes in $\mathbb{R}^3$

Let $p \in \mathbb{R}^3$ and $0 \neq v \in T_p\mathbb{R}^3 = \mathbb{R}^3$. The **line** through the point p with *direction vector* v is the subset

$$L := p + \operatorname{span}\{v\} \subset \mathbb{R}^3.$$

We can write

$$L = \{p + \lambda v \,|\, \lambda \in \mathbb{R}\}$$

and refer to the equation

$$x = p + \lambda v, \quad \lambda \in \mathbb{R}$$

as the **vector equation** of the line.

NOTE : In the vector equation of line L, the elements $x, p,$ and v are all viewed as geometric vectors, hence written as column matrices :

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} + \lambda \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}, \quad \lambda \in \mathbb{R}.$$

The vector equation is equivalent to the following set of three scalar equations :

$$\begin{aligned} x_1 &= p_1 + \lambda v_1 \\ x_2 &= p_2 + \lambda v_2 \\ x_3 &= p_3 + \lambda v_3, \quad \lambda \in \mathbb{R} \end{aligned}$$

called **parametric equations** for the line L. Alternatively, the line L can be viewed as the image set of the *linear mapping*

$$G : \mathbb{R} \to \mathbb{R}^3, \quad t \mapsto (p_1 + tv_1, p_2 + tv_2, p_3 + tv_3)\,.$$

Now let $p \in \mathbb{R}^3$ and consider two linearly independent vectors $v, w \in T_p\mathbb{R}^3 = \mathbb{R}^3$. The **plane** through the point p with *direction subspace* $\vec{P} = \text{span}\{v, w\}$ is the subset

$$P := p + \text{span}\{v, w\} \subset \mathbb{R}^3.$$

Likewise, we can write

$$P = \{p + \lambda v + \mu w \,|\, \lambda, \mu \in \mathbb{R}\}$$

and refer to the equation

$$x = p + \lambda v + \mu w, \quad \lambda, \mu \in \mathbb{R}$$

as the **vector equation** of the plane. The vector equation is equivalent to the following set of three scalar equations :

$$\begin{aligned} x_1 &= p_1 + \lambda v_1 + \mu w_1 \\ x_2 &= p_2 + \lambda v_2 + \mu w_2 \\ x_3 &= p_3 + \lambda v_3 + \mu w_3, \quad \lambda, \mu \in \mathbb{R} \end{aligned}$$

called **parametric equations** for the plane P.

NOTE : The fact that the vectors v and w are linearly independent is equivalent to the following rank condition :

$$\operatorname{rank}\begin{bmatrix} v & w \end{bmatrix} = \operatorname{rank}\begin{bmatrix} v_1 & w_1 \\ v_2 & w_2 \\ v_3 & w_3 \end{bmatrix} = 2.$$

Alternatively, the plane P can be viewed as the image set of the *linear mapping*

$$G' : \mathbb{R}^2 \to \mathbb{R}^3, \quad (s,t) \mapsto (p_1 + sv_1 + tw_1, p_2 + sv_2 + tw_2, p_3 + sv_3 + tw_3)\,.$$

◇ **Exercise 132** Show that the system of linear equations (in unknowns λ and μ)

$$\begin{aligned} \lambda v_1 + \mu w_1 &= x_1 - p_1 \\ \lambda v_2 + \mu w_2 &= x_2 - p_2 \\ \lambda v_3 + \mu w_3 &= x_3 - p_3 \end{aligned}$$

(where $\operatorname{rank}\begin{bmatrix} v & w \end{bmatrix} = 2$) is consistent if and only if

$$\begin{vmatrix} x_1 - p_1 & x_2 - p_2 & x_3 - p_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} = 0.$$

(HINT : A system of linear equations $Ax = b$ is consistent if and only if $\operatorname{rank}\begin{bmatrix} A & b \end{bmatrix} = \operatorname{rank}(A)$.)

◇ **Exercise 133** Show that the condition $x - p = \lambda v + \mu w$ (where $\operatorname{rank}\begin{bmatrix} v & w \end{bmatrix} = 2$) is equivalent to

$$(x - p) \bullet v \times w = 0.$$

The plane

$$P = p + \vec{P} = p + \operatorname{span}\{v, w\}, \quad \operatorname{rank}\begin{bmatrix} v & w \end{bmatrix} = 2$$

can be described by the scalar equation

$$a_1(x_1 - p_1) + a_2(x_2 - p_2) + a_3(x_3 - p_3) = 0$$

or by the so-called (general) **Cartesian equation**

$$a_1x_1 + a_2x_1 + a_3x_3 + c = 0, \quad a_1^2 + a_2^2 + a_3^2 \neq 0.$$

(Here

$$a_1 = \begin{vmatrix} v_2 & v_3 \\ w_2 & w_3 \end{vmatrix}, \quad a_2 = \begin{vmatrix} v_3 & v_1 \\ w_3 & w_1 \end{vmatrix}, \quad a_3 = \begin{vmatrix} v_1 & v_2 \\ w_1 & w_2 \end{vmatrix}.)$$

◇ **Exercise 134** Show that any equation of the form

$$a_1x_1 + a_2x_2 + a_3x_3 + c = 0, \quad a_1^2 + a_2^2 + a_3^2 \neq 0$$

represents a plane P in $\mathbb{R}^3$.

NOTE : The Cartesian equation for the plane P can be put into the form

$$u \bullet x + c = 0$$

where $u = v \times w$ and $c = -p \bullet v \times w$. The (nonzero) vector u defines the *normal direction* of P. We can see that the line with vector direction $u = v \times w$ is *orthogonal* to the plane with vector subspace $\text{span}\{v, w\}$.

Let P_1 and P_2 be two planes (not necessarily distinct) in $\mathbb{R}^3$. So

$$P_i = p_i + \vec{P}_i, \quad i = 1, 2$$

and it is easy to see that

$$P_1 = P_2 \iff p_2 - p_1 \in \vec{P}_1 = \vec{P}_2.$$

Hence

$$P_1 \neq P_2 \iff \left(\vec{P}_1 \neq \vec{P}_2 \quad \text{or} \quad p_2 - p_1 \notin \vec{P}_1 = \vec{P}_2\right).$$

It turns out that condition $p_2 - p_1 \notin \vec{P}_1 = \vec{P}_2$ is equivalent to $P_1 \cap P_2 = \emptyset$; in this case, we say that the planes P_1 and P_2 are *strictly parallel* : $P_1 \parallel P_2$ but $P_1 \neq P_2$. Otherwise, P_1 and P_2 are two intersecting planes.

On intuitive grounds we "know" that the intersection of two distinct planes is either the empty set (when the planes are strictly parallel) or a line.

3.2.1 PROPOSITION. *The intersection of two distinct, intersectiong planes is a line.*

PROOF : Let P_1 and P_2 be two distinct, intersecting planes. We can describe each of these planes by a Cartesian equation of the form

$$a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 + c_i = 0$$

where each set of coefficients is such that $a_{i1}^2 + a_{i2}^2 + a_{i3}^2 \neq 0, \quad i = 1, 2.$ The facts that the planes are *distinct* and are *not* parallel translate into the following rank condition :

$$\text{rank} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} = 2.$$

But this means that the system of two linear equations in three unknowns $x_1, x_2,$ and x_3

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= -c_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= -c_2 \end{aligned}$$

is *consistent* and, moreover, there is one basic variable ($v \neq 0$) and one free variable (λ). As a result, the general solution has the form

$$x = p + \lambda v, \quad \lambda \in \mathbb{R}$$

which represents a line. □

◇ **Exercise 135** Show that any line can be represented as an intersection of two (distinct) planes. (HINT : Write the parametric equations of your line in "symmetric form" :

$$\frac{x_1 - p_1}{v_1} = \frac{x_2 - p_2}{v_2} = \frac{x_3 - p_3}{v_3}\cdot)$$

NOTE : Any line can be represented as the intersection of an arbitrary family of planes. Indeed, given a line L described by the (Cartesian) equations

$$\begin{aligned} (P) \quad a_1x_1 + a_2x_2 + a_3x_3 + c &= 0 \\ (P') \quad a'_1x_1 + a'_2x_2 + a'_3x_3 + c' &= 0 \end{aligned}$$

where the coefficients satisfy the rank condition

$$\operatorname{rank}\begin{bmatrix} a_1 & a_2 & a_3 \\ a_1' & a_2' & a_3' \end{bmatrix} = 2$$

(i.e., the line L is represented as an intersection of two planes : $L = P \cap P'$), then the *family* of planes

$$\nu_1 (a_1x_1 + a_2x_2 + a_3x_3 + c) + \nu_2 (a_1'x_1 + a_2'x_2 + a_3'x_3 + c') = 0, \quad \nu_1, \nu_2 \in \mathbb{R}$$

contains *all* planes through the line L. (For $\nu_1 = 0$ we get the plane P. If $\nu_1 \neq 0$, put $\nu := \dfrac{\nu_2}{\nu_1}$ and we may write our family of planes - excluding the plane P' - as follows

$$a_1x_1 + a_2x_2 + a_3x_3 + c + \nu (a_1'x_1 + a_2'x_2 + a_3'x_3 + c') = 0, \quad \nu \in \mathbb{R}.$$

So

$$L = P \cap P' = \bigcap_{\nu \in \mathbb{R}} P_\nu.$$

Clearly, $P = P_0 \in (P_\nu)_{\nu \in \mathbb{R}}$ but $P' \notin (P_\nu)_{\nu \in \mathbb{R}}$. The "exclusion" of the plane P' can be easily fixed by simply putting $P' = P_\infty := \lim_{\nu \to \infty} P_\nu$. Hence any subfamily, finite or infinite, of $(P_\nu)_{\nu \in \overline{\mathbb{R}}}$, $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ has the desired property.

◇ **Exercise 136** Show that the Cartesian equation of the plane through three *noncolinear* points p, q, r can be put into the form

$$\begin{vmatrix} x_1 & x_2 & x_3 & 1 \\ p_1 & p_2 & p_3 & 1 \\ q_1 & q_2 & q_3 & 1 \\ r_1 & r_2 & r_3 & 1 \end{vmatrix} = 0.$$

What do we get when the points are collinear ?

◇ **Exercise 137** Prove that the lines

$$(L) \quad x = p + \lambda v \quad \text{and} \quad (L') \quad x = p' + \mu v'$$

lie in the same plane if and only if

$$\begin{vmatrix} p_1 - p_1' & p_2 - p_2' & p_3 - p_3' \\ v_1 & v_2 & v_3 \\ v_1' & v_2' & v_3' \end{vmatrix} = 0.$$

ℓ-Planes in $\mathbb{R}^m$

Higher-dimensional analogues of lines and planes can be now defined without difficulty.

3.2.2 DEFINITION. A (nonempty) subset $L \subseteq \mathbb{R}^m$ of the form

$$L = p + \vec{L},$$

where $p \in \mathbb{R}^m$ and $\vec{L}$ is a *vector subspace* of $T_o\mathbb{R}^m = \mathbb{R}^m$, is said to be a **linear submanifold** of Euclidean m-space $\mathbb{R}^m$.

The vector subspace $\vec{L}$ is called the **direction subspace** of the linear submanifold L. If the dimension of $\vec{L}$ (as a vector subspace) is ℓ, then we say that L is a linear submanifold of *dimension* ℓ (or, simply, a linear **ℓ-submanifold**); in this case, $m - \ell$ is referred to as the *codimension* of L.

NOTE : A linear submanifold $L = p + \vec{L}$ is the result of "shifting" a vector subspace $\vec{L}$ by (a vector) p. In this vein, linear ℓ-submanifolds are also called ℓ-planes (or even ℓ-flats).

3.2.3 EXAMPLE. Vector subspaces of $\mathbb{R}^m$ are linear submanifolds. Indeed, if $p \in \vec{L}$ (in particular, if $p = o$), then $L = \vec{L}$.

3.2.4 EXAMPLE. A linear 0-submanifold is simply a point (in fact, a singleton). In this case, $L = p + \vec{0} = p$, hence $L = \{p\}$.

3.2.5 EXAMPLE. A linear 1-submanifold is a line (in $\mathbb{R}^m$).

A linear submanifold of dimension $m - 1$ is called a **hyperplane**. A hyperplane has codimension 1. What about linear submanifolds of codimension zero ? There is only one such linear submanifold, the space itself. Indeed, in this case,

$$L = p + \operatorname{span}\{v_1, v_2, \ldots, v_m\} = p + \mathbb{R}^m = \mathbb{R}^m.$$

◇ **Exercise 138** Let $L = p + \vec{L}$ be a linear submanifold and let $q \in L$. Show that

$$L = q + \vec{L}.$$

Hence deduce that a linear submanifold L is a vector subspace if and only if $o \in L$.

◇ **Exercise 139** Prove that if $p + \vec{L} = p' + \vec{L}'$, then $\vec{L} = \vec{L}'$.

◇ **Exercise 140** Let $(L_\alpha)_{\alpha\in\mathfrak{A}}$ be a family of linear submanifolds such that $\bigcap_{\alpha\in\mathfrak{A}} L_\alpha \neq \emptyset$. Show that the subset $L = \bigcap_{\alpha\in\mathfrak{A}} L_\alpha$ is a linear submanifolds. Hence deduce that

$$\dim(L) = \dim \bigcap_{\alpha\in\mathfrak{A}} \vec{L}_\alpha.$$

3.2.6 PROPOSITION. *Given two distinct points $p, q \in \mathbb{R}^m$, there exists a unique line $\overleftrightarrow{pq}$ containing p and q.*

PROOF : (Existence) The line $p + \text{span}\{q - p\}$ contains both points p, q.

(Uniqueness) Let L be a line such that $p, q \in L$. We must show that

$$L = p + \text{span}\{q - p\}.$$

We have

$$L = p + \vec{L}$$

and so

$$q \in p + \vec{L}.$$

Thus the 1-dimensional vector subspace $\vec{L}$ contains the nonzero vector $q - p$. Hence

$$\vec{L} = \text{span}\{q - p\}.$$

□

NOTE : The line $\overleftrightarrow{pq}$ throught the points p and q can be expressed as follows

$$\overleftrightarrow{pq} = \{(1 - \lambda)p + \lambda q \,|\, \lambda \in \mathbb{R}\}.$$

We can now characterize linear submanifolds in terms of lines.

3.2.7 THEOREM. *A subset $\emptyset \neq L \subseteq \mathbb{R}^m$ is a linear submanifold if and only if for every two distinct points $x, y \in \mathbb{R}^m$, the line $\overleftrightarrow{xy}$ is contained in L.*

PROOF : Observe that this condition is equivalent to

$$(x, y \in L,\ \lambda \in \mathbb{R}) \Rightarrow (1 - \lambda)x + \lambda y \in L.$$

($\Rightarrow$) Let $x, y \in L$. Then $L = x + \vec{L}$, so $y - x \in \vec{L}$ and hence

$$\lambda(y - x) \in \vec{L}.$$

We have

$$(1 - \lambda)x + \lambda y = x + \lambda(y - x) \in x + \vec{L} = L.$$

($\Leftarrow$) Let $p \in L$ and denote $\vec{L} := L - p$. Let

$$y_1 = x_1 - p \in \vec{L} \quad \text{and} \quad y_2 = x_2 - p \in \vec{L}.$$

Then

$$\begin{aligned} (1 - \lambda)y_1 + \lambda y_2 &= (1 - \lambda)(x_1 - p) + \lambda(x_2 - p) \\ &= (1 - \lambda)x_1 + \lambda x_2 - p \in L - p. \end{aligned}$$

Hence

$$\Big(y_1, y_2 \in \vec{L},\ \lambda \in \mathbb{R}\Big) \Rightarrow (1 - \lambda)y_1 + \lambda y_2 \in \vec{L}.$$

In particular, for $y_1 = 0$, we get

$$\Big(y \in \vec{L},\ \lambda \in \mathbb{R}\Big) \Rightarrow \lambda y \in \vec{L}.$$

Now let $\mu \in \mathbb{R} \setminus \{0, 1\}$ and let $y, y' \in \vec{L}$. Then $y_1 = \dfrac{1}{1 - \mu}y$, $y_2 = \dfrac{1}{\mu}y' \in \vec{L}$ and thus

$$\begin{aligned} y + y' &= (1 - \mu)\frac{1}{1 - \mu}y + \mu\frac{1}{\mu}y' \\ &= (1 - \mu)y_1 + \mu y_2 \in \vec{L}. \end{aligned}$$

Hence

$$y, y' \in \vec{L} \Rightarrow y + y' \in \vec{L}.$$

It follows that $\vec{Y}$ is a vector subspace of $\mathbb{R}^m$. But $L = p + \vec{L}$, which proves the result. □

This result can be easily generalized.

◇ **Exercise 141** Prove that a subset $\emptyset \neq L \subseteq \mathbb{R}^m$ is a linear submanifold if and only if

$$\left(x_1, \ldots, x_m \in L,\ \lambda_1, \ldots, \lambda_m \in \mathbb{R},\ \sum_{i=1}^{m} \lambda_i = 1 \right) \Rightarrow \sum_{i=1}^{m} \lambda_i x_i \in L.$$

NOTE : A linear combination $\sum \lambda_i x_i$ where the coefficients λ_i satisfy the condition $\sum \lambda_i = 1$ is called an *affine combination*. A linear submanifold can be characterized by the condition that it contains all the affine combinations of any (finite collection) of its elements; such special subsets (of some "affine space") are called *affine subspaces*. So linear submanifolds are just affine subspaces of $\mathbb{R}^m$.

In general, the *union* of two linear submanifolds is *not* a linear submanifold. Let L_1 and L_2 be two linear submanifolds of Euclidean m-space $\mathbb{R}^m$. Then the *set* $L_1 \cup L_2$ does generate a linear submanifold, denoted by $L_1 \vee L_2$, by taking the intersection of all linear submanifolds of $\mathbb{R}^m$ that contain $L_1 \cup L_2$. Thus

$$L_1 \vee L_2 := \bigcap_{L_1 \cup L_2 \subseteq L} L \subseteq \mathbb{R}^m.$$

NOTE : $L_1 \vee L_2$ is the smallest linear submanifold that contains (as subsets) L_1 and L_2. It is sometimes referred to as the *affine span* of $L_1 \cup L_2$. It turns out that for $L_i = p_i + \vec{L}_i,\ i = 1, 2$ one has

$$L_1 \vee L_2 = p_1 + \vec{L}_1 + \vec{L}_2 + \operatorname{span}\{p_2 - p_1\}.$$

(Here $L_1 + L_2$ denotes the *sum* of the vector subspaces L_1 and L_2.)

◇ **Exercise 142** Given linear submanifolds $L_i = p_i + \vec{L}_i,\ i = 1, 2$, show that

$$L_1 \cap L_2 \neq \emptyset \iff \operatorname{span}\{p_2 - p_1\} \subseteq \vec{L}_1 + \vec{L}_2.$$

Hence deduce that if $p \in L_1 \cap L_2$, then

$$\begin{aligned} L_1 \cap L_2 &= p + \vec{L}_1 \cap \vec{L}_2 \\ L_1 \vee L_2 &= p + \vec{L}_1 + \vec{L}_2. \end{aligned}$$

3.2.8 THEOREM. (DIMENSION THEOREM) *Let* $L_i = p_i + \vec{L}_i,\ i = 1, 2$ *be linear submanifolds.*

(a) If $L_1 \cap L_2 \neq \emptyset$, *then*

$$\dim(L_1 \vee L_2) = \dim L_1 + \dim L_2 - \dim(L_1 \cap L_2).$$

(b) If $L_1 \cap L_2 = \emptyset$, *then*

$$\dim(L_1 \vee L_2) = \dim\left(\vec{L}_1 + \vec{L}_2\right) + 1.$$

PROOF : (a) We have (see **Exercise 142**)

$$\begin{aligned} \dim(L_1 \vee L_2) &= \dim\left(\vec{L}_1 + \vec{L}_2\right) \\ \dim(L_1 \cap L_2) &= \dim\left(\vec{L}_1 \cap \vec{L}_2\right). \end{aligned}$$

But

$$\dim\left(\vec{L}_1 + \vec{L}_2\right) = \dim \vec{L}_1 + \dim \vec{L}_2 - \dim\left(\vec{L}_1 \cap \vec{L}_2\right)$$

and the first result follows.

(b) We have

$$\begin{aligned} \dim(L_1 \vee L_2) &= \dim\left(\vec{L}_1 + \vec{L}_2 + \operatorname{span}\{p_2 - p_1\}\right) \\ &= \dim\left(\vec{L}_1 + \vec{L}_2\right) + 1. \end{aligned}$$

□

3.2.9 EXAMPLE. The linear submanifold $L_1 \vee L_2$ generated by the *lines* L_1 and L_2

- is a plane if $L_1 \cap L_2 = \{p\}$.
- is a plane if $L_1 \cap L_2 = \emptyset$ and $\vec{L}_1 = \vec{L}_2$.
- has dimension 3 (i.e., is a 3-flat) if $L_1 \cap L_2 = \emptyset$ and $\vec{L}_1 \neq \vec{L}_2$.

◇ **Exercise 143** In Euclidean 4-space $\mathbb{R}^4$, write (parametric) equations for the linear submanifold generated by te lines

$$\frac{x_1}{2} = \frac{x_2 - 1}{1} = \frac{x_3 + 1}{-1} = \frac{x_4}{3}$$

and

$$\frac{x_1 - 1}{3} = \frac{x_2}{2} = \frac{x_3}{1} = \frac{x_4 - 2}{-1}.$$

Consider an *affine map*

$$F : \mathbb{R}^m \to \mathbb{R}^n, \quad x \mapsto Ax + c.$$

(Here A is an $n \times m$ matrix and c a column n-matrix, both with real entries.) We can see that such a map preserves affine combinations of points.

3.2.10 PROPOSITION. *Let $L = p + \vec{L}$ be a linear submanifold of $\mathbb{R}^m$. Then the image of L under the affine map F, $x \mapsto Ax + c$ is also a linear submanifold (of $\mathbb{R}^n$).*

PROOF : We shall show that

$$F(L) = F(p) + A(\vec{L}).$$

Let $y = F(x)$, $x \in L$; then $x - p \in \vec{L}$ and hence

$$\begin{aligned} y - F(p) &= F(x) - F(p) \\ &= A(x - p) \in A(\vec{L}). \end{aligned}$$

Thus $F(L) \subseteq F(p) + A(\vec{L})$.

Conversely, let $y - F(p) \in A(\vec{L})$. Then

$$y - F(p) = A(x - p)$$

for some $x \in L$. This implies $y = F(x)$ and thus $F(L) \supseteq F(p) + A(\vec{L})$. The result now follows. □

◇ **Exercise 144** Given a linear submanifold $L = p + \vec{L}$ of $\mathbb{R}^m$ and an affine map $F : \mathbb{R}^m \to \mathbb{R}^n$, $x \mapsto Ax + c$, show that the inverse image of any $y \in F(L)$ under F is a linear submanifold. (The direction subspace of $F^{-1}(y)$ is $\ker(A) \subseteq \vec{L}$.)

NOTE : The linear submanifold $F^{-1}(y)$, $y \in F(\mathbb{R}^m) = \operatorname{im}(F)$ may be referred to as the *fibre* of (the affine map) F over (the point) y. All the fibres of F have the same direction subspace. So the space $\mathbb{R}^m$ decomposes into a family of parallel submanifolds of the same dimension :

$$\mathbb{R}^m = \bigcup_{y \in \operatorname{im}(F)} F^{-1}(y), \qquad \dim F^{-1}(y) = \dim \ker(A).$$

Recall that, for an $n \times m$ matrix A, the following basic relation holds :

$$\dim \ker (A) + \dim \operatorname{im} (A) = m$$

(the rank-nullity formula). Geometrically, this means that, for the linear map $x \mapsto Ax$, the *nullity* of $A\,(= \dim \ker (A))$ counts for the number of dimensions that collapse as we perform A and the *rank* of $A\,(= \dim \operatorname{im} (A))$ counts for the number of dimensions that survive after we perform A.

It follows that the dimension of any of the fibres of the affine map $F,\ x \mapsto Ax + c$ is $m - \operatorname{rank} (A)$.

A function $f : \mathbb{R}^m \to \mathbb{R}$ of the form

$$x = (x_1, x_2, \ldots, x_m) \mapsto a_1 x_1 + a_2 x_2 + \cdots + a_m x_m + c$$

is called an **affine functional** on $\mathbb{R}^m$. We shall find it convenient to assume that not all the coefficients $a_1, \ldots, a_m$ are zero; so, in other words, we rule out the constant function $x \mapsto c$.

NOTE : A nonconstant affine functional is an affine map (function)

$$f : \mathbb{R}^m \to \mathbb{R}, \quad x \mapsto Ax + c$$

with

$$\operatorname{rank} (A) = \operatorname{rank} \begin{bmatrix} a_1 & a_2 & \cdots & a_m \end{bmatrix} = 1.$$

Hence the fibres of f are linear submanifolds (of $\mathbb{R}^m$) of dimension $m - 1$ (i.e., hyperplanes).

The Cartesian equation

$$a_1 x_1 + a_2 x_2 + \cdots + a_m x_m + c = 0 \quad \text{with} \quad \operatorname{rank} \begin{bmatrix} a_1 & \cdots & a_m \end{bmatrix} = 1$$

represents the hyperplane $f^{-1}(0) \subseteq \mathbb{R}^m$.

◇ **Exercise 145** Show that any nonconstant affine functional $f : \mathbb{R}^m \to \mathbb{R}$ is surjective.

NOTE : A system of linear equations (in unknowns $x_1, x_2, \ldots, x_m$)

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + \cdots + & a_{1m}x_m & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + \cdots + & a_{2m}x_m & = & b_2 \\ \multicolumn{9}{c}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots} \\ \multicolumn{9}{c}{a_{m-\ell,1}x_1 + a_{m-\ell,2}x_2 + \cdots + a_{m-\ell,m}x_m = b_{m-\ell}} \end{array}$$

with

$$\operatorname{rank}(A) = \operatorname{rank}\begin{bmatrix} a_{11} & \dots & a_{1m} \\ \vdots & & \vdots \\ a_{m-\ell,1} & \dots & a_{m-\ell,m} \end{bmatrix} = m - \ell$$

represents (geometrically) the intersection of $m - \ell$ hyperplanes in $\mathbb{R}^m$.

Let

$$L = p + \vec{L} = p + \operatorname{span}\{v_1, v_2, \dots, v_\ell\}$$

be a linear submanifold of dimension ℓ. (It is assumed, of course, that the vectors $v_1, v_2, \dots, v_\ell$ are linearly independent.) Then we can write

$$L = \{p + \lambda_1 v_1 + \cdots + \lambda_\ell v_\ell \,|\, \lambda_1, \dots, \lambda_\ell \in \mathbb{R}\}$$

and refer to the equation

$$x = p + \lambda_1 v_1 + \cdots + \lambda_\ell v_\ell, \quad \lambda_1, \dots, \lambda_\ell \in \mathbb{R}$$

as the **vector equation** of the linear submanifold.

Equivalently, we can express (in coordinates) the linear submanifold L by the following set of m scalar equations

$$\begin{aligned} x_1 &= p_1 + \lambda_1 v_{11} + \lambda_2 v_{12} + \cdots + \lambda_\ell v_{1\ell} \\ x_2 &= p_2 + \lambda_1 v_{21} + \lambda_2 v_{22} + \cdots + \lambda_\ell v_{2\ell} \\ &\vdots \\ x_m &= p_m + \lambda_1 v_{m1} + \lambda_2 v_{m2} + \cdots + \lambda_\ell v_{m\ell}, \quad \lambda_1, \dots, \lambda_\ell \in \mathbb{R} \end{aligned}$$

called **parametric equations** for L. (Here $v_i = \begin{bmatrix} v_{1i} \\ \vdots \\ v_{mi} \end{bmatrix}$, $i = 1, 2, \dots, \ell.$)
Alternatively, the linear submanifold L can be viewed as the image set of the following affine mapping

$$(t_1, \dots, t_\ell) \mapsto (p_1 + t_1 v_{11} + \cdots + t_\ell v_{1l}, \dots, p_m + t_1 v_{m1} + \cdots + t_\ell v_{m\ell}).$$

NOTE : Linear submanifolds are in fact solution sets for (consistent) systems of linear equations. More precisely, let $Ax = b$ (where $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^{n\times 1}$) be a

system of n linear equations in m unknows $x_1, x_2, \ldots, x_n$. Suppose that $\text{rank}\,(A) = k$ with $0 < k \leq \min\{m, n\}$. The system is *consistent* (i.e., it has at least one solution) if and only if the rank of the augmented matrix of the system equals the rank of the coefficient matrix (Kronecker-Capelli) :

$$\text{rank}\begin{bmatrix} A & b \end{bmatrix} = \text{rank}\,(A).$$

(When $b = 0$, the system is said to be *homogeneous* and, clearly, it is consistent. A homogeneous system possesses a unique solution - the trivial solution - if and only if $\text{rank}\,(A) = m$.) Reducing the matrix $\begin{bmatrix} A & b \end{bmatrix}$ to a row echelon form using Gaussian elimination and then solving for the *basic variables* in terms of the *free variables* leads to the **general solution**

$$x = p + \lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_{m-k} v_{m-k}.$$

As the free variables λ_i range over all possible values, this general solution generates all possible solutions of the system. (p is a particular solution of the nonhomogeneous system, whereas the expression $\lambda_1 v_1 + \cdots + \lambda_{m-k} v_{m-k}$ is the general solution of the associated homogeneous system.) We see that the solution set S of the system (assumed to be consistent) is a linear submanifold of dimension $m - k$:

$$S = p + \text{span}\{v_1, \ldots, v_{m-k}\} \subset \mathbb{R}^m.$$

(The basic vectors form a basis of the direction subspace of S.) This algebraic viewpoint makes it clear that linear submanifolds can be studied, at least in principle, only by (linear) algebraic means. On the other hand, the alternative geometric viewpoint offers a broader perspective : linear submanifolds are simple, special cases of nonlinear objects/subspaces, the so-called *smooth submanifolds*; these are the natural higher-dimensional analogues of regular curves.

We can interpret the parametric equations for (the linear ℓ-submanifold) L as the general solution of a system of linear equations (in unknowns $x_1, x_2, \ldots, x_m$). If we write down one such system (i.e., if we *eliminate* the parameters $\lambda_1, \ldots, \lambda_\ell$) we get **Cartesian equations** for L :

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m + c_1 &=& 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m + c_2 &=& 0 \\ \multicolumn{3}{c}{\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots} \\ a_{m-\ell,1}x_1 + a_{m-\ell,2}x_2 + \cdots + a_{m-\ell,m}x_m + c_{m-\ell} &=& 0 \end{array}$$

with

$$\operatorname{rank}\begin{bmatrix} a_{11} & \dots & a_{1m} \\ \vdots & & \vdots \\ a_{m-\ell,1} & \dots & a_{m-\ell,m} \end{bmatrix} = m-\ell.$$

(The linear ℓ-submanifold L is represented as an intersection of $m-\ell$ distinct hyperplanes.)

We can summarize all these characterizations of a linear submanifold in the following

3.2.11 THEOREM. *Let $\emptyset \neq L$ be a subset of $\mathbb{R}^m$ and assume $0 \leq \ell \leq m$. The following statements are equivalent.*

(i) L is a linear ℓ-submanifold of $\mathbb{R}^m$.

(ii) There exist linearly independent affine functions

$$f_i : \mathbb{R}^m \to \mathbb{R}, \quad (x_1, \dots, x_m) \mapsto a_{i1}x_1 + \dots + a_{im}x_m + c_i \ \ (i = 1, 2, \dots, m-\ell)$$

(i.e., the row matrices $a_i = \begin{bmatrix} a_{i1} & \cdots & a_{im} \end{bmatrix}$, $i = 1, 2, \dots, m-\ell$ are linearly independent) such that

$$L = \bigcap_{i=1}^{m-\ell} f_i^{-1}(0).$$

(iii) There exists an affine mapping

$$F : \mathbb{R}^m \to \mathbb{R}^{m-\ell}, \quad x \mapsto Ax + c$$

with $\operatorname{rank}(A) = m-\ell$ such that

$$L = F^{-1}(0).$$

(iv) There exist affine functions

$$h_i : \mathbb{R}^{m-\ell} \to \mathbb{R}, \quad i = 1, 2, \dots, m-\ell$$

such that (possibly after a permutation of coordinates) L is the graph of the mapping

$$H = (h_1, \dots, h_{m-\ell}) : \mathbb{R}^{m-\ell} \to \mathbb{R}^{m-\ell} \subseteq \mathbb{R}^m$$

(under the canonical isomorphism).

(v) There exists an affine mapping

$$G : \mathbb{R}^{m-\ell} \to \mathbb{R}^m, \quad t = (t_1, \ldots, t_{m-\ell}) \mapsto Bt + d$$

with $\operatorname{rank}(B) = m - \ell$ *such that* L *is the image set of* G.

NOTE : In (ii) we think of a linear submanifold as *an intersection of hyperplanes*, in (iii) as the *zero-set* of a certain affine mapping, in (iv) as a *graph*, and in (v) as the *image set* of a certain affine mapping (i.e., a parametrized set).

Parallelism and orthogonality

Let $L_i = p_i + \vec{L}_i,\ i = 1, 2$ be linear submanifolds of $\mathbb{R}^m$.

3.2.12 DEFINITION. We say that L_1 and L_2 are **parallel**, denoted $L_1 \parallel L_2$, provided $\vec{L}_1 \subseteq \vec{L}_2$ or $\vec{L}_2 \subseteq \vec{L}_1$.

◇ **Exercise 146** Show that if $L_1 \parallel L_2$, then either $L_1 \subseteq L_2$ or $L_2 \subseteq L_1$ or $L_1 \cap L_2 = \emptyset$.

◇ **Exercise 147** Given two planes

$$\begin{aligned} (P) \quad & a_1x_1 + a_2x_2 + a_3x_3 + c = 0 \\ (P') \quad & a'_1x_1 + a'_2x_2 + a'_3x_3 + c' = 0 \end{aligned}$$

in Euclidean 3-space $\mathbb{R}^3$, show that a necessary and sufficient condition for them to be *parallel* is

$$\frac{a_1}{a'_1} = \frac{a_2}{a'_2} = \frac{a_3}{a'_3}.$$

(The convention is made that if a denominator is zero, the corresponding numerator is also zero.)

◇ **Exercise 148** Show that a necessary and sufficient condition for the plane

$$a_1x_1 + a_2x_2 + a_3x_3 + c = 0$$

and the line

$$\begin{aligned} x_1 &= p_1 + tu_1 \\ x_2 &= p_2 + tu_2 \\ x_3 &= p_3 + tu_3, \quad t \in \mathbb{R} \end{aligned}$$

to be *parallel* is

$$a_1u_1 + a_2u_2 + a_3u_3 = 0.$$

3.2.13 PROPOSITION. *Let L and H be an arbitrary linear submanifold an a hyperplane (i.e., a linear submanifold of codimension 1), respectively. If $L \cap H = \emptyset$, then $L \parallel H$.*

PROOF : Let $L = p + \vec{L}$ and $H = q + \vec{H}$. It is clear that

$$\dim(L \vee H) = m.$$

Since

$$\dim(L \vee H) = \dim\left(\vec{L} + \vec{H}\right) + 1,$$

it follows that

$$\dim\left(\vec{L} + \vec{H}\right) = m - 1 = \dim \vec{H}.$$

We have $\vec{H} \subseteq \vec{L} + \vec{H}$ and thus

$$\vec{H} = \vec{L} + \vec{H}.$$

Hence $\vec{L} \subseteq \vec{H}$. This shows that $L \parallel H$. □

. . . . . .

3.3 The Inverse Mapping Theorem

One of the most important results of differential calculus is the so-called *inverse mapping theorem*. (Another fundamental result is the existence theorem for ordinary differential equations.) In order to simplify the terminology of this and later sections we introduce first the notion of *diffeomorphism* (or differentiable homeomorphism) between two spaces.

NOTE : This concept can have no meaning unless the spaces are such that *differentiability* is defined, which – at the present moment – means that they must be subsets of Euclidean spaces.

Let $U \subseteq \mathbb{R}^m$ and $V \subseteq \mathbb{R}^n$ be open sets. We say that a mapping $F : U \to V$ is a C^r **diffeomorphism** $(1 \leq r \leq \infty)$ if F is a homeomorphism and both F and F^{-1} are of class C^r. (When $r = 1$ we simply say *diffeomorphism.*)

NOTE : A diffeomorphism is thus necessarily bijective, but a differentiable bijective mapping may not be a diffeomorphism. For example, the function $f : \mathbb{R} \to \mathbb{R},\ t \mapsto t^3$ is a homeomorphism and f is differentiable (in fact, smooth), but $f^{-1} : \mathbb{R} \to \mathbb{R},\ s \mapsto \sqrt[3]{s}$ is *not* differentiable (since it has no derivative at $s = 0$).

$\diamond$ **Exercise 149** Let A be an $n \times m$ matrix and B an $m \times n$ matrix. Prove that if $BA = I_m$ and $AB = I_n$, then $m = n$ and A is invertible with inverse B. (HINT : Show that if $BA = I_m$, then $\operatorname{rank}(A) = \operatorname{rank}(B) = m$.)

3.3.1 PROPOSITION. *If $F : U \to V$ is a diffeomorphism (of an open subset of $\mathbb{R}^m$ onto an open subset of $\mathbb{R}^n$) and $p \in U$, then the derivative $DF(p) : \mathbb{R}^m = T_p\mathbb{R}^m \to \mathbb{R}^n = T_{F(p)}\mathbb{R}^n$ is a linear isomorphism. In particular, $m = n$.*

PROOF : Since

$$F^{-1} \circ F = id_U = id_{\mathbb{R}^m}|_U$$

($id_{\mathbb{R}^m}$ is a linear mapping), we have

$$D(F^{-1} \circ F)(p) = id_{\mathbb{R}^m}$$

or, by the general chain rule,

$$DF^{-1}(F(p)) \circ DF(p) = id_{\mathbb{R}^m}.$$

Likewise,

$$DF(p) \circ DF^{-1}(F(p)) = id_{\mathbb{R}^n}.$$

(It is safe to identify

$$T_pU = T_p\mathbb{R}^m = \mathbb{R}^m \quad \text{and} \quad T_{F(p)}V = T_{F(p)}\mathbb{R}^n = \mathbb{R}^n.)$$

It follows that the linear mapping $DF(p)$ is invertible with inverse $D(F^{-1})(F(p))$. $\square$

NOTE : It would *not* be possible to have a diffeomorphism between open subsets of Euclidean spaces of *different* dimensions; indeed, a famous (and deep) result of algebraic topology – *Brouwer's theorem on invariance of domain* – asserts that even homeomorphisms between open subsets of Euclidean spaces of different dimensions is impossible. (In fact, the result says that *if* $U \subseteq \mathbb{R}^m$ *is open and* $f : U \to \mathbb{R}^n$ *is continuous and one-to-one, then* $f(U)$ *is open.* It is then easy to derive the fact that if $U \subseteq \mathbb{R}^m$ and $V \subseteq \mathbb{R}^n$ are open subsets such that U is homeomorphic to V, then $m = n$.)

We have seen that if the mapping $F : U \to V$ is a diffeomorphism between open subsets of $\mathbb{R}^m$, then the Jacobian matrix $\frac{\partial F}{\partial x}(p)$ is nonsingular (or, equivalently, the Jacobian $J_F(p) \neq 0$) for every $p \in U$. While the converse is not exactly true, it is true *locally.* The following fundamental result holds.

3.3.2 THEOREM. (INVERSE MAPPING THEOREM) *Let* $U \subseteq \mathbb{R}^m$ *be an open set and let* $F : U \to \mathbb{R}^m$ *be of class* C^r $(1 \leq r \leq \infty)$. *Let* $p \in U$ *and suppose that* $DF(p)$ *is a linear isomorphism (i.e., the Jacobian matrix* $\frac{\partial F}{\partial x}(p)$ *is non-singular). Then there exists an open neighborhood* W *of* p *in* U *such that* $F|_W : W \to F(W)$ *is a* C^r *diffeomorphism. Moreover, for* $y \in F(W)$ *we have the following formula for the derivatives of* F^{-1} *at* y *:*

$$DF^{-1}(y) = (DF(x))^{-1}, \qquad \text{where } y = F(x).$$

This is a remarkable result. From a single piece of linear information at one point, it concludes to information in a whole neighborhood of that point. The proof is quite involved and will be omitted.

NOTE : The following two results are consequences of the inverse mapping theorem :

- If DF is invertible at every point of U, then F is an *open mapping* (i.e., it carries U and open subsets of $\mathbb{R}^m$ contained in U into open subsets of $\mathbb{R}^m$).
- A necessary and sufficient condition for the C^1 mapping F to be a diffeomorphism (from U to $F(U)$) is that it be one-to-one and DF be invertible at every point of U.

◇ **Exercise 150** Let $F : \mathbb{R}^2 \to \mathbb{R}^2$ be given by

$$F(x_1, x_2) = (e^{x_1} \cos x_2, e^{x_1} \sin x_2).$$

Show that the (smooth) mapping F is *locally invertible*, but *not* invertible.

◇ **Exercise 151** Show that the system

$$\begin{aligned} y_1 &= x_1^3 x_2 + x_2^2 \\ y_2 &= \ln(x_1 + x_2) \end{aligned}$$

has a unique solution $x_1 = f(y_1, y_2),\ x_2 = g(y_1, y_2)$ in a neighborhood of $(6, \ln 3)$ with $f(6, \ln 3) = 1$ and $g(6, \ln 3) = 2$. Find

$$\frac{\partial f}{\partial y_1},\ \frac{\partial f}{\partial y_2},\ \frac{\partial g}{\partial y_1}, \quad \text{and} \quad \frac{\partial g}{\partial y_2}.$$

There is a generalization of THEOREM 3.3.2, called the *constant rank theorem*, which is actually equivalent to the inverse function theorem.

A C^1 mapping $F : U \subseteq \mathbb{R}^m \to \mathbb{R}^n$ has *constant rank* k if the rank of the linear mapping $DF(x) : \mathbb{R}^m = T_x\mathbb{R}^m \to \mathbb{R}^n = T_{F(x)}\mathbb{R}^n$ is k at every point $x \in U$. Equivalently, the Jacobian matrix $\frac{\partial F}{\partial x}$ has constant rank k on U.

NOTE : In linear algebra, the *rank* of a matrix $A \in \mathbb{R}^{n\times m}$ is defined in three equivalent ways : (i) the dimension of the subspace of $\mathbb{R}^m$ spanned by the rows, (ii) the dimension of the subspace of $\mathbb{R}^n$ spanned by the columns, or (iii) the maximum order of any nonvanishing minor determinant. We see at once from (i) and (ii) that $\operatorname{rank}(A) \leq m, n$.

The *rank* of a linear mapping is defined to be the dimension of the image, and one proves that this is the rank of any matrix which represents the mapping. From this it follows that, if P and Q are nonsingular matrices, then $\operatorname{rank}(PAQ) = \operatorname{rank}(A)$. When $F : U \subseteq \mathbb{R}^m \to \mathbb{R}^n$ is a C^1 mapping, then the linear mapping $DF(x)$ has a rank at each $x \in U$. Because the value of the determinant is a continuous function of its entries, we see from (iii) that if $\operatorname{rank}(DF(p)) = k$, then for some neighborhood V of p, $\operatorname{rank}(DF(x)) \geq k$; and, if $k = \min\{m, n\}$, then $\operatorname{rank}(DF(x)) = k$ on V. We shall refer to the rank of $DF(x)$ as the *rank* of F at x.

If we compose F with diffeomorphisms, then the facts cited and the general chain rule imply that the rank of the composition is the rank of F, since diffeomorphisms have nonsingular Jacobian matrices.

3.3.3 EXAMPLE. Consider the composition

$$\mathbb{R}^k \times \mathbb{R}^{m-k} \xrightarrow{\pi} \mathbb{R}^k \xrightarrow{i} \mathbb{R}^n \quad (1 \leq k < m, n)$$

where

$$\begin{aligned} \pi(x_1, \ldots, x_k, y_1, \ldots, y_{m-k}) &= (x_1, \ldots, x_k) \\ i(x_1, \ldots, x_k) &= (x_1, \ldots, x_k, 0, \ldots, 0). \end{aligned}$$

The Jacobian matrix of $i \circ \pi$ is constantly the matrix

$$\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{n \times m}.$$

The rank is constantly k.

The constant rank theorem asserts that, in a certain precise sense, mappings of constant rank k locally "look like" the above example.

3.3.4 THEOREM. (CONSTANT RANK THEOREM) *Let $U \subseteq \mathbb{R}^m$ and $V \subseteq \mathbb{R}^n$ be open sets and let $F : U \to V$ be of class C^r $(1 \leq r \leq \infty)$. Let $p \in U$ and suppose that, in some neighborhood of p, F has constant rank k. Then there are open neighborhoods W of p in U and $Z \supseteq F(W)$ of $F(p)$ in V, respectively, together with C^r diffeomorphisms*

$$G : W \to \widetilde{W} \subseteq \mathbb{R}^m \quad \textit{and} \quad H : Z \to \widetilde{Z} \subseteq \mathbb{R}^n$$

such that (on $\widetilde{W}$)

$$H \circ F \circ G^{-1}(z_1, \ldots, z_m) = (z_1, \ldots, z_k, 0, \ldots, 0).$$

NOTE : The diffeomorphisms $G : W \to \widetilde{W}$ and $H : Z \to \widetilde{Z}$ should be thought of as *changes of coordinates* in these open sets. For instance, one could write

$$\begin{aligned} z_1 &= g_1(x_1, \ldots, x_m) \\ z_2 &= g_2(x_1, \ldots, x_m) \\ &\vdots \\ z_m &= g_m(x_1, \ldots, x_m) \end{aligned}$$

viewing $(z_1, \ldots, z_m)$ as new coordinates of the point $(x_1, \ldots, x_m)$. The new coordinates depend differentiably on the original ones and, G being a diffeomorphism,

the original coordinates depend differentiably on the new ones. Thus, all of calculus, formulated in the coordinates x_i has a completely equivalent formulation in the coordinates z_i. (The specific formulas change, but the "realities" they express do not.) According to this philosophy, the point of the constant rank theorem is that *the most general mapping of constant rank can be expressed locally using the same formula as the simple* EXAMPLE 3.3.3, provided the coordinates in the domain and the range are suitably changed.

The immersion and submersion theorems

There are two important special cases of THEOREM 3.3.4, the *immersion theorem* and the *submersion theorem.* A C^r mapping $F : U \subseteq \mathbb{R}^m \to V \subseteq \mathbb{R}^n$ is

- an **immersion** if it has constant rank m
- a **submersion** if it has constant rank n

on U.

NOTE : If F is an immersion, then $m \leq n$. If it is a submersion, then $m \geq n$. If it is both an immersion and a submersion, then $n = m$ and F is *locally* a diffeomorphism (such a mapping is also said to be *regular*).

◇ **Exercise 152** Let $F : U \subseteq \mathbb{R}^m \to V \subseteq \mathbb{R}^n$ be a C^r mapping (between open sets), and $m \leq n$. Show that F is an *immersion* if and only if the derivative $DF(x)$ is *one-to-one* at every point $x \in U$.

When $m = 1$, let U be an open interval $J \subseteq \mathbb{R}$. In this case, the mapping $F : J \to \mathbb{R}^n$ is a parametrized curve in the Euclidean space $\mathbb{R}^n$. To verify that F is an immersion it is necessary to check that the Jacobian matrix of F has rank 1 (i.e., one of the derivatives, with respect to t, of the components of F differs from zero for every $t \in J$).

◇ **Exercise 153** Verify that the following mappings are immersions.

(a) $F_1 : \mathbb{R} \to \mathbb{R}^3, \quad t \mapsto (\cos t, \sin t, t)$.
(The image of F_1 is a circular helix.)

(b) $F_2 : \mathbb{R} \to \mathbb{R}^2, \quad t \mapsto (\cos t, \sin t)$.
(The image of F_2 is the unit circle $\mathbb{S}^1$.)

(c) $F_3 : (1, \infty) \to \mathbb{R}^2, \quad t \mapsto \left(\frac{1}{t}\cos(2\pi t), \frac{1}{t}\sin(2\pi t)\right)$.
(The image of F_3 is a curve spiraling to the origin as $t \to \infty$ and tending to the point $(1, 0)$ as $t \to 1$.)

3.3.5 Corollary. (Immersion Theorem) *Let $F : U \to V$ be a C^r immersion. Then there are open neighborhoods W of p in U and $Z \supseteq F(W)$ of $F(p)$ in V, respectively, together with C^r diffeomorphisms*

$$G : W \to \widetilde{W} \subseteq \mathbb{R}^m \quad \textit{and} \quad H : Z \to \widetilde{Z} \subseteq \mathbb{R}^n$$

such that (on $\widetilde{W}$)

$$H \circ F \circ G^{-1}(y_1, \ldots, y_m) = (y_1, \ldots, y_m, 0, \ldots, 0).$$

An immersion is *locally* – but not necessarily globally – one-to-one. For instance, the standard parametrization of the unit circle is an immersion which is clearly *not* one-to-one. Two more instructive examples are given below.

3.3.6 Example. Consider the mapping

$$F : \mathbb{R} \to \mathbb{R}^2, \quad t \mapsto \left(2\cos\left(t - \frac{\pi}{2}\right), \sin 2\left(t - \frac{\pi}{2}\right)\right).$$

It is easy to check that F is an immersion which is not one-to-one. The image of F is a "figure eight" (a self-intersecting geometric curve) with the image point making a complete circuit starting at the origin as t goes from 0 to 2π.

3.3.7 Example. The mapping

$$G : \mathbb{R} \to \mathbb{R}^2, \quad t \mapsto F(g(t)) = \left(2\cos\left(g(t) - \frac{\pi}{2}\right), \sin 2\left(g(t) - \frac{\pi}{2}\right)\right)$$

where $g(t) = \pi + 2\arctan t$, is again an immersion. The image is the "eight figure" as in the previous example, but with an important difference : the image point passes through the origin only once, when $t = 0$; for $t \to -\infty$ and $t \to \infty$ it only approaches the origin as *limit*. Hence G is an one-to-one immersion.

◇ **Exercise 154** Is the mapping

$$F : \mathbb{R} \to \mathbb{R}^2, \quad t \mapsto (t^2, t^3)$$

an immersion ? What about the restriction $F|_U$ of F to $U = \mathbb{R} \setminus \{0\}$? Investigate for injectivity this restriction.

NOTE : An immersion $F : U \subseteq \mathbb{R}^\ell \to \mathbb{R}^m$ is said to be an **embedding** if, in addition,

- F is *injective*. (Observe that the induced mapping $F : U \to \mathbb{F}(U)$ is bijective.)
- $F^{-1} : F(U) \to U$ is *continuous*.

In particular, the mapping $F : U \to F(U)$ is bijective, continuous, and possesses a continuous inverse; hence, is is a *homeomorphism*. Accordingly, an embedding is an immersion which is also a homeomorphism onto its image.

3.3.8 EXAMPLE. The mapping

$$F : \mathbb{R} \to \mathbb{R}^2, \quad t \mapsto (\cos t, \sin t)$$

is a smooth immersion (see **Exercise 153**). Its image set is the unit circle

$$\mathbb{S}^1 = \left\{ x \in \mathbb{R}^2 \,|\, \|x\| = 1 \right\}.$$

We can see that F is *not* one-to-one. However, we can make it so by restricting F to the open interval $J_0 = (0, 2\pi)$ (or, more generally, to an interval of the form $J_a = (a, a + 2\pi)$ with $a \in \mathbb{R}$). The image of this interval under F is a circle with one point left out (a *punctured circle*) :

$$F(J_0) = \mathbb{S}^1 \setminus \{(1, 0)\}.$$

The maping

$$F^{-1} : F(J_0) \to J_0$$

is *continuous*. Consequently, $F : J_0 \to \mathbb{R}^2$ is a smooth *embedding*.

3.3.9 EXAMPLE. The mapping

$$\widetilde{F} : \mathbb{R} \to \mathbb{R}^2, \quad t \mapsto (t^2 - 1, t^3 - t)$$

is a smooth immersion. One has

$$\widetilde{F}(s) = \widetilde{F}(t) \iff t = s \text{ or } s, t \in \{-1, 1\}.$$

This makes the restriction $F := \widetilde{F}\Big|_{(-\infty,1)}$ one-to-one. But it does *not* make F an embedding.

◇ **Exercise 155** Show that the mapping

$$F^{-1} : F((-\infty, 1)) \to (-\infty, 1)$$

is *not* continuous at the point $(0, 0)$.

◇ **Exercise 156** Let $F : U \subseteq \mathbb{R}^m \to V \subseteq \mathbb{R}^n$ be a C^r mapping (between open sets), and $m \geq n$. Show that F is a *submersion* if and only if the derivative $DF(x)$ is *onto* at every point $x \in U$.

When $n = 1$, the mapping $F = f : U \subseteq \mathbb{R}^m \to \mathbb{R}$ is a (differentiable) function defined on the open set U. To verify that f is a submersion it is necessary to check that the Jacobian matrix of f has rank 1 (i.e., one of the partial derivatives of f differs from zero for every $t \in U$).

◇ **Exercise 157** Verify that the following functions are submersions.

(a) $f_1 : \mathbb{R}^m \to \mathbb{R}, \quad x \mapsto a_1 x_1 + \cdots + a_m x_m + c \qquad (a_1^2 + \cdots + a_m^2 \neq 0)$.
(The inverse image of the origin under f_1 is a hyperplane.)

(b) $f_2 : \mathbb{R}^m \setminus \{0\} \to \mathbb{R}, \quad x \mapsto x_1^2 + \cdots + x_m^2 - 1$.
(The inverse image of the origin under f_2 is the unit sphere $\mathbb{S}^{m-1}$.)

3.3.10 COROLLARY. (SUBMERSION THEOREM) *Let $F : U \to V$ be a C^r submersion. Then there are open neighborhoods W of p in U and $Z \supseteq F(W)$ of $F(p)$ in V, respectively, together with C^r diffeomorphisms*

$$G : W \to \widetilde{W} \subseteq \mathbb{R}^m \quad \textit{and} \quad H : Z \to \widetilde{Z} \subseteq \mathbb{R}^n$$

such that (on $\widetilde{W}$)

$$H \circ F \circ G^{-1}(y_1, \ldots, y_m) = (y_1, \ldots, y_n).$$

3.3.11 EXAMPLE. Let $\mathsf{GL}(n,R)$ denote the set (group) of all invertible (i.e., nonsingular) $n \times n$ matrices with real entries. (It can be shown that $\mathsf{GL}(n,\mathbb{R})$ may be viewed as an open subset of Euclidean space $\mathbb{R}^{n^2}$.) The map

$$\det : \mathsf{GL}(n,\mathbb{R}) \to \mathbb{R}^{\times}, \quad A \mapsto \det(A)$$

is differentiable (in fact, smooth) and its derivative is given by

$$D\det(A) \cdot B = (\det A)\,\mathrm{tr}\,\left(A^{-1}B\right).$$

The differentiability of det is clear from its formula in terms of matrix elements. Now

$$\det(I_n + \lambda C) = 1 + \lambda \mathrm{tr}\, C + \cdots + \lambda^n \det C$$

implies

$$\frac{d}{d\lambda}\det(I_n + \lambda C)\Big|_{\lambda=0} = \mathrm{tr}\, C$$

and hence

$$\begin{aligned} D\det(A) \cdot B &= \frac{d}{d\lambda}\det(A + \lambda B)\Big|_{\lambda=0} \\ &= \frac{d}{d\lambda}\left[(\det A)\det(I_n + \lambda A^{-1}B)\right]_{\lambda=0} \\ &= (\det A)\,(\mathrm{tr}\,(A^{-1}B)). \end{aligned}$$

In particular (for $A = I_n$),

$$D\det(I_n) \cdot B = \mathrm{tr}\, B.$$

The map tr is onto, and so the function det is a (smooth) *submersion.*

◇ **Exercise 158** Let $\mathsf{Sym}(n)$ denote the set (vector space) of all symmetric $n \times n$ matrices with real entries, and consider the mapping

$$\Psi : \mathsf{GL}(n,\mathbb{R}) \to \mathsf{Sym}(n), \quad A \mapsto AA^T.$$

Show that Ψ is differentiable (in fact, smooth) and its derivative is given by

$$D\Psi(A) \cdot B = AB^T + BA^T.$$

Hence derive that Ψ is a (smooth) *submersion.*

The Implicit Mapping Theorem

The following result follows easily from the INVERSE MAPPING THEOREM.

3.3.12 PROPOSITION. *Let $U \subseteq \mathbb{R}^k \times \mathbb{R}^{m-k}$ be an open set and let $F : U \to \mathbb{R}^{m-k}$ be of class C^r $(1 \leq r \leq \infty)$. Let $(p,q) \in U$ and suppose that $F(p,q) = 0$ and the matrix $\frac{\partial F}{\partial y}(p,q) \in \mathbb{R}^{(m-k)\times(m-k)}$ is nonsingular. Then there exist an open neighborhood $W \subseteq \mathbb{R}^k$ of p, an open neighborhood $W' \subseteq \mathbb{R}^{m-k}$ of q and a unique C^r mapping $\Phi : W \to W'$ such that $\Phi(p) = q$, and for all $x \in W$, $(x, \Phi(x)) \in U$ and*

$$F(x, \Phi(x)) = 0.$$

NOTE : This result is the so-called IMPLICIT MAPPING THEOREM. It gives sufficient conditions for *local solvability* of a system of equations of the form

$$\begin{aligned} f_1(x_1, \ldots, x_k, y_1, \ldots, y_{m-k}) &= 0 \\ f_2(x_1, \ldots, x_k, y_1, \ldots, y_{m-k}) &= 0 \\ &\vdots \\ f_{m-k}(x_1, \ldots, x_k, y_1, \ldots, y_{m-k}) &= 0 \end{aligned}$$

where the functions f_i are differentiable. (We want to solve for these $m-k$ unknown $y_1, \ldots, y_{m-k}$ in the $m-k$ equations in terms of $x_1, \ldots, x_k$.)

PROOF : Define the mapping $\widetilde{F} : U \to \mathbb{R}^k \times \mathbb{R}^{m-k} = \mathbb{R}^m$ by

$$\widetilde{F}(x,y) := (x, F(x,y))$$

and observe that $\widetilde{F}$ satisfies the hypotheses of the INVERSE MAPPING THEOREM: $\widetilde{F} \in C^r(U, \mathbb{R}^m)$ and $J_{\widetilde{F}}(p,q) = \left|\frac{\partial F}{\partial y}(p,q)\right| \neq 0$. Thus there is an open neighborhood $\widetilde{W} = W_0' \times W'$ of (p,q) and an open neighborhood $W \times W_0$ of $\widetilde{F}(p,q) = (p,0)$ such that $\widetilde{F} : W_0' \times W' \to W \times W_0$ has a C^r inverse $\widetilde{F}^{-1} : W \times W_0 \to W_0' \times W'$; clearly, $\widetilde{F}^{-1}$ is of the form $\widetilde{F}^{-1}(x,y) = (x, H(x,y))$. Now define

$$\Phi : W \to W', \quad \Phi(x) := H(x,0).$$

Then $\Phi \in C^r(W, \mathbb{R}^{m-k})$ and

$$(p, \Phi(p)) = (p, H(p,0) = \widetilde{F}^{-1}(p,0) = (p,q)$$

which implies $\Phi(p) = q$. For $x \in W$, $(x, \Phi(x)) \in U$ and

$$F(x, \Phi(x)) = \left(F \circ \widetilde{F}^{-1}\right)(x, 0) = \left(\mathrm{pr}_2 \circ \widetilde{F} \circ \widetilde{F}^{-1}\right)(x, 0) = \mathrm{pr}_2(x, 0) = 0.$$

□

◇ **Exercise 159** Show that (in PROPOSITION 3.3.10) when $m - k = 1$ we get

$$\frac{\partial \Phi}{\partial x_j} = -\frac{\dfrac{\partial F}{\partial x_j}}{\dfrac{\partial F}{\partial y}} \qquad (j = 1, 2, \ldots, k).$$

NOTE : More generally, the partial derivatives $\dfrac{\partial \Phi_i}{\partial x_j}$ are given by

$$\begin{bmatrix} \frac{\partial \Phi_1}{\partial x_1} & \cdots & \frac{\partial \Phi_1}{\partial x_k} \\ \vdots & & \vdots \\ \frac{\partial \Phi_{m-k}}{\partial x_1} & \cdots & \frac{\partial \Phi_{m-k}}{\partial x_k} \end{bmatrix} = - \begin{bmatrix} \frac{\partial f_1}{\partial y_1} & \cdots & \frac{\partial f_1}{\partial y_{m-k}} \\ \vdots & & \vdots \\ \frac{\partial f_{m-k}}{\partial y_1} & \cdots & \frac{\partial f_{m-k}}{\partial y_{m-k}} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_k} \\ \vdots & & \vdots \\ \frac{\partial f_{m-k}}{\partial x_1} & \cdots & \frac{\partial f_{m-k}}{\partial x_k} \end{bmatrix}.$$

◇ **Exercise 160** Show that the equations

$$\begin{aligned} x + y + t &= 0 \\ xyt + \sin(xyt) &= 0 \end{aligned}$$

define x and y implicitly as functions of t in an open neighborhood of the point $(t, x, y) = (-1, 0, 1)$. Calculate the derivatives $x'(-1)$ and $y'(-1)$.

3.4 Smooth Submanifolds

Linear submanifolds (of some Euclidean space $\mathbb{R}^m$) are a generalization of the notion of line; they are higher-dimensional geometrical objects (subsets) which can be studied rather easily because of their simple algebraic structure : linear submanifolds are "linear" ! The natural "non-linear" analogues are the *smooth submanifolds*; smooth submanifolds are a significant generalization of the notion of smooth curve.

NOTE : All the results proven so far are valid for C^r mappings (or functions). However, the class C^r is not strong enough for some purposes. For this reason, and

since it is very convenient to know that we do not lose differentiability as a result of taking derivatives (the derivatives of a smooth mapping are also smooth). C^∞ is the preferred differentiability class in much of (differentiable) manifold theory. Henceforth we will be concerned almost exclusively with *smooth* mappings (or functions).

We make the following definition.

3.4.1 DEFINITION. A (nonempty) subset S of $\mathbb{R}^m$ is said to be a **smooth submanifold** if, for every $x \in S$, there exist an open neighborhood U of x in $\mathbb{R}^m$ and a smooth *diffeomorphism* $\phi : U \to \widetilde{U} \subseteq \mathbb{R}^m$ such that

$$\phi(S \cap U) = \widetilde{U} \cap \mathbb{R}^\ell,$$

where $0 \leq \ell \leq m$.

We say that S is a smooth submanifold of *dimension* ℓ (or, simply, an ℓ-**submanifold**). The *codimension* of S is $m - \ell$.

NOTE : Roughly speaking, the condition

$$S \cap U = \phi^{-1}(\widetilde{U} \cap \mathbb{R}^\ell)$$

says that the set S looks like $\mathbb{R}^\ell$ and is "flat" in $\mathbb{R}^m$. We may assume, without any loss of generality, that $\phi(x) = o$ (the origin).

3.4.2 THEOREM. *Let $\emptyset \neq S$ be a subset of $\mathbb{R}^m$ and suppose $0 \leq \ell \leq m$. The following statements are equivalent.*

(i) S is an ℓ-submanifold of $\mathbb{R}^m$.

(ii) For every $x \in S$ there exist an open neighborhood U of x in $\mathbb{R}^m$ and smooth functions $f_i : U \to \mathbb{R}$, $i = 1, 2, \ldots, m - \ell$ such that the linear functionals $Df_i(x)$ are linearly independent and

$$S \cap U = \bigcap_{i=1}^{m-\ell} f_i^{-1}(0).$$

(iii) For every $x \in S$ there exist an open neighborhood U of x in $\mathbb{R}^m$ and a smooth submersion $F : U \to \mathbb{R}^{m-\ell}$ such that

$$S \cap U = F^{-1}(0).$$

(iv) *For every* $x \in S$ *there exist an open neighborhood* U *of* $x = (x_1, \ldots, x_m)$ *in* $\mathbb{R}^m$*, an open neighborhood* U' *of* $x' = (x_1, \ldots, x_\ell)$ *in* $\mathbb{R}^\ell$ *and smooth functions* $h_i : U' \to \mathbb{R}$, $i = 1, 2, \ldots, m - \ell$ *such that, possibly after a permutation of coordinates, the intersection* $S \cap U$ *is the graph of the mapping* $H := (h_1, \ldots, h_{m-\ell}) : U' \to \mathbb{R}^{m-\ell}$ *(under the canonical isomorphism):*

$$S \cap U = \operatorname{graph}(H).$$

(v) *For every* $x \in S$ *there exist an open neighborhood* U *of* x *in* $\mathbb{R}^m$*, an open neighborhood* V *of* 0 *in* $\mathbb{R}^\ell$ *and a smooth embedding* $\Phi : V \to \mathbb{R}^m$ *such that* $\Phi(0) = x$ *and*

$$S \cap U = \operatorname{im} \Phi := \{\Phi(y) \,|\, y \in V\}.$$

NOTE : In (ii) we think of a smooth submanifold as an *intersection of hypersurfaces* (i.e., codimension-1 smooth submanifolds) defined by local equations, in (iii) as the *zero-set* of a smooth submersion, in (iv) as a *graph*, and in (v) as the *image set* of a smooth embedding (i.e., a parametrized set). All these are *local* descriptions. (In (v) it is sufficient to assume that the smooth mapping Φ is an embedding only at the origin because if $DG(0)$ is injective, so is $DG(x)$ for x close enough to o.)

PROOF : We shall show that

$$(iii) \Rightarrow (i) \Rightarrow (v) \Rightarrow (iv) \Rightarrow (ii) \Rightarrow (iii).$$

$(iii) \Rightarrow (i)$. This is just the SUBMERSION THEOREM.

$(i) \Rightarrow (v)$. We may assume (by using a translation, if necessary) that $\phi(x) = 0$. Take $V = \phi(S \cap U)$ and $\Phi = \phi^{-1} \circ i$, where $i : \mathbb{R}^\ell \to \mathbb{R}^m$ is the canonical inclusion.

$(v) \Rightarrow (iv)$. After permuting indices, if necessary, we may assume that $D\Phi(0)(\mathbb{R}^\ell) \cap \mathbb{R}^{m-\ell} = 0$. Let $\operatorname{pr}_1 : \mathbb{R}^m = \mathbb{R}^\ell \times \mathbb{R}^{m-\ell} \to \mathbb{R}^\ell$ be the projection on the first factor. From $D\Phi(0)(\mathbb{R}^\ell) \cap \mathbb{R}^{m-\ell} = 0$ we deduce that

$$D(\operatorname{pr}_1 \circ \Phi)(0)(\mathbb{R}^\ell) = \mathbb{R}^\ell.$$

In other words, the mapping $\mathrm{pr}_1 \circ \Phi$ is regular at 0. By the INVERSE MAPPING THEOREM, there exists an open neighborhood V' of 0 such that $\mathrm{pr}_1 \circ \Phi$ is a (smooth) diffeomorphism between V' and $U' = \mathrm{pr}_1(\Phi(V')) \subseteq \mathbb{R}^\ell$. Thus (iv) is satisfied if we take this U' and $h_1, \ldots, h_{m-\ell}$ equal to the $m-\ell$ last component functions of the mapping $H = \Phi \circ (\mathrm{pr}_1 \circ \Phi)^{-1} \in C^\infty(U', \mathbb{R}^m)$. In fact, $H(U') = \Phi(V')$ by assumption, and so there exists an open set $U'' \subseteq \mathbb{R}^m$ (containing U) such that

$$\Phi(V') = H(U') = U'' \cap V.$$

Thus $U'' \cap V$ is the graph of $(h_1, \ldots, h_{m-\ell}) = H$.

$(iv) \Rightarrow (ii)$. Just set

$$f_i(x_1, \ldots, x_m) = h_i(x_1, \ldots, x_\ell) - x_{i+\ell}$$

for $i = 1, 2, \ldots, m-\ell$.

$(ii) \Rightarrow (iii)$. The mapping $F : U \to \mathbb{R}^{m-\ell}$ with component functions $f_1, \ldots, f_{m-\ell}$ is a smooth submersion at x, and remains a submersion on an open neighborhood of x, since the determinant is a continuous function.

□

The following result follows easily from the CONSTANT RANK THEOREM.

3.4.3 PROPOSITION. *Let $U \subseteq \mathbb{R}^m$ and $V \subseteq \mathbb{R}^n$ be open sets and let $F : U \to V$ be a smooth mapping of constant rank k. Let $q \in F(U) \subseteq V$. Then $F^{-1}(q)$ is a smooth submanifold of U of dimension $m-k$.*

PROOF : Let $x \in F^{-1}(q)$. Choose a neighborhood of x as in the CONSTANT RANK THEOREM. Without loss of generality, we can replace W with $\widetilde{W}$ and $F|_W$ with $H \circ F \circ G^{-1}$ on $\widetilde{W}$, all as in that theorem. That is, on W, we assume that

$$F(x_1, \ldots, x_m) = (x_1, \ldots, x_m, 0, \ldots, 0).$$

Thus $q = (a_1, \ldots, a_k, 0, \ldots, 0)$ and $W \cap F^{-1}(q)$ is the set of all points in W of the form

$$(a_1, \ldots, a_k, x_{k+1}, \ldots, x_m).$$

The desired diffeomorphism $\phi : W \to \phi(W) \subseteq \mathbb{R}^m$ will be

$$\phi(x_1, \ldots, x_m) = (x_{k+1}, \ldots, x_m, x_1 - a_1, \ldots, x_k - a_k).$$

□

Examples of smooth submanifolds

3.4.4 EXAMPLE. 0-submanifolds of $\mathbb{R}^m$ are exactly *sets of isolated points*.

◇ **Exercise 161** Show that linear submanifolds are smooth submanifolds.

3.4.5 EXAMPLE. A *parametrized curve* in $\mathbb{R}^m$ is a smooth mapping

$$\alpha : J \to \mathbb{R}^m,$$

where $J \subseteq \mathbb{R}$ is an open interval. If the mapping α is an immersion (i.e., $\dot{\alpha}(t) \neq 0$ for all $t \in J$), we say that the curve is *regular*. In this case, one can show that every $t \in J$ has a neighborhood W such that $\alpha(W) \subseteq \mathbb{R}^m$ is a 1-submanifold of $\mathbb{R}^m$.

NOTE : In general, the trace $\alpha(J)$ of a regular curve is *not* a submanifold, even if the mapping α is one-to-one. For instance, neither the "figure eight" (see EXAMPLE 3.3.6) nor its variation, without self-intersection (see EXAMPLE 3.3.7) are submanifolds of $\mathbb{R}^2$. Both these geometric curves are images of a smooth submanifold – the open interval J – under some smooth immersion.

We have just seen that, in general, the image of a submanifold under an immersion (even a one-to-one immersion) is *not* a submanifold. However, *the inverse image of a point* (i.e., a connected 0-dimensional submanifold) *under a submersion is either the empty set or a submanifold.* (This is a special case of PROPOSITION 3.4.3.)

3.4.6 EXAMPLE. The *sphere*

$$\mathbb{S}^{m-1} := \{x \in \mathbb{R}^m \mid \|x\| = 1\}$$

is a compact, $(m-1)$-submanifold of $\mathbb{R}^m$. ($\mathbb{S}^1$ is the unit *circle*; $\mathbb{S}^0$ is equal to two points.)

To see this, write

$$\mathbb{S}^{m-1} = \{x = (x_1, \ldots, x_m) \,|\, x_1^2 + \cdots + x_m^2 = 1\}.$$

Thus the sphere $\mathbb{S}^{m-1}$ is the zero-set of the smooth function

$$f : \mathbb{R}^m \to \mathbb{R}, \quad (x_1, \ldots, x_m) \mapsto x_1^2 + \cdots + x_m^2 - 1.$$

That is, $\mathbb{S}^{m-1} = f^{-1}(0)$. Since the function f is a smooth submersion, the result follows.

3.4.7 EXAMPLE. A smooth submanifold of codimension one is usually referred to as a (smooth) **hypersurface**. Hyperplanes and spheres are simple examples of hypersurfaces. More generally, (nonempty) subsets of the form

$$S = \{x = (x_1, \ldots, x_m) \in \mathbb{R}^m \,|\, f(x_1, \ldots, x_m) = 0\},$$

where $f : \mathbb{R}^m \to \mathbb{R}$ is a smooth submersion, are hypersurfaces (of $\mathbb{R}^m$).

Another simple way of constructing smooth submanifolds is given now.

3.4.8 PROPOSITION. *Let S_1 be an ℓ_1-submanifold of $\mathbb{R}^m$ and S_2 an ℓ_2-submanifold of $\mathbb{R}^n$. Then $S_1 \times S_2$ is an $(\ell_1 + \ell_2)$-submanifold of $\mathbb{R}^{m+n}$.*

PROOF : THEOREM 3.4.2, applied to $x \in S_1$, and $y \in S_2$, gives $n + m - (\ell_1 + \ell_2)$ (smooth) functions f_i defined on an open neighborhood $U = U_1 \times U_2 \subseteq \mathbb{R}^{m+n}$ of (x, y) and satisfying condition (ii) for $S_1 \times S_2$. □

3.4.9 EXAMPLE. The *k-torus*

$$\mathbb{T}^k := \mathbb{S}^1 \times \cdots \times \mathbb{S}^1 \subseteq \mathbb{R}^2 \times \cdots \times \mathbb{R}^2 = \mathbb{R}^{2k}$$

is a compact, k-submanifold of $\mathbb{R}^{2k}$.

3.4.10 EXAMPLE. m-submanifolds of $\mathbb{R}^m$ are exactly *open subsets* of $\mathbb{R}^m$. We shall see that the set (group) $\mathsf{GL}(n, \mathbb{R})$ of all invertible $n \times n$ matrices with real entries - the so-called (real) *general linear group* - is an open subset of Euclidean space $\mathbb{R}^{n^2}$. Hence *the general linear group* $\mathsf{GL}(n, \mathbb{R})$ *is a smooth submanifold* (of $\mathbb{R}^{n^2}$).

NOTE : Any *closed* subgroup of $\mathsf{GL}(n,\mathbb{R})$ turns out to be a smooth submanifold (of $\mathbb{R}^{n^2}$). This result (by no means obvious) will be proved in the chapter devoted to (abstract) Lie groups.

◇ **Exercise 162** Prove that

(a) each of the following sets is a smooth submanifold of $\mathbb{R}^2$ (of dimension 1):

i. $\{x \in \mathbb{R}^2 \,|\, x_2 = x_1^3\}$;

ii. $\{x \in \mathbb{R}^2 \,|\, x_1 = x_2^3\}$;

iii. $\{x \in \mathbb{R}^2 \,|\, x_1 x_2 = 1\}$.

(b) none of the following sets is a smooth submanifold of $\mathbb{R}^2$:

i. $\{x \in \mathbb{R}^2 \,|\, x_2 = |x_1|\}$;

ii. $\{x \in \mathbb{R}^2 \,|\, (x_1 x_2 - 1)(x_1^2 + x_2^2 - 2) = 0\}$;

iii. $\{x \in \mathbb{R}^2 \,|\, x_2 = -x_1^2 \text{ for } x_1 \leq 0;\ x_2 = x_1^2 \text{ for } x_1 \geq 0\}$.

◇ **Exercise 163** Why is that

$$\{x \in \mathbb{R}^2 \,|\, \|x\| < 1\} \quad \text{and} \quad \{x \in \mathbb{R}^2 \,|\, |x_1| < 1,\ |x_2| < 1\}$$

are submanifolds of $\mathbb{R}^2$, but not

$$\{x \in \mathbb{R}^2 \,|\, \|x\| \leq 1\}\,?$$

◇ **Exercise 164** Which of the following sets are smooth submanifolds (of some appropriate Euclidean space $\mathbb{R}^m$) ?

(a) $\big\{(t^2, t^3) \,|\, t \in \mathbb{R}\big\}$;

(b) $\big\{(x_1, x_2) \in \mathbb{R}^2 \,|\, x_1 = 0 \ \text{ or } \ x_2 = 0\big\}$;

(c) $\big\{(t, t^2) \,|\, t < 0\big\} \cup \big\{(t, -t^2) \,|\, t > 0\big\}$;

(d) $\{(\cos t, \sin t, t) \,|\, t \in \mathbb{R}\}$;

(e) $\big\{(x_1, x_2, x_3) \in \mathbb{R}^3 \,|\, x_1^3 + x_2^3 + x_3^3 - 3x_1 x_2 x_3 = 1\big\}$;

(f) $\big\{(x_1, x_2, x_3) \in \mathbb{R}^3 \,|\, x_1^2 + x_2^2 + x_3^2 = 1 \ \text{ and } \ x_1 + x_2 - x_3 = 0\big\}$.

◇ **Exercise 165** Define

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad x \mapsto x_1^3 - x_2^3.$$

(a) Prove that f is a *surjective* smooth function.

(b) Prove that f is a smooth *submersion* at every point $x \in \mathbb{R}^2 \setminus \{0\}$.

(c) Prove that for all $c \in \mathbb{R}$ the set

$$\{x \in \mathbb{R}^2 \,|\, f(x) = c\}$$

is a *submanifold* of $\mathbb{R}^2$ of dimension 1.

◇ **Exercise 166** Define

$$g : \mathbb{R}^3 \to \mathbb{R}, \quad x \mapsto x_1^2 + x_2^2 - x_2^2.$$

(a) Prove that g is a *surjective* smooth function.

(b) Prove that g is a smooth *submersion* at every point $x \in \mathbb{R}^3 \setminus \{0\}$.

(c) Prove that the two sheets of the cone

$$g^{-1}(0) \setminus \{0\} = \{x \in \mathbb{R}^3 \setminus \{0\} \,|\, x_1^2 + x_2^2 = x_3^2\}$$

form a *submanifold* of $\mathbb{R}^3$ of dimension 2.

◇ **Exercise 167** A *nondegenerate quadric* in $\mathbb{R}^m$ is a set of the form

$$\begin{aligned} Q &:= \{x \in \mathbb{R}^m \,|\, (Ax) \bullet x + b \bullet x + c = 0\} \\ &= \{x \in \mathbb{R}^m \,|\, x^\top Ax + b^\top x + c = 0\}, \end{aligned}$$

where A is a symmetric (i.e., $A^\top = A$) invertible $m \times m$ matrix with real entries, b is a column m-matrix with real entries, and $c \in \mathbb{R}$. Introduce the *discriminant* $\Delta := b^\top A^{-1} b - 4c \in \mathbb{R}$.

(a) Show that Q is a smooth *hypersurface* (i.e., a smooth submanifold of dimension $m-1$) of $\mathbb{R}^m$.

(b) Suppose $\Delta = 0$. Verify that $p := -\frac{1}{2} A^{-1} b \in Q$ and then show that $S = Q \setminus \{p\}$ is also a smooth hypersurface of $\mathbb{R}^m$.

Tangent spaces

Let S be an ℓ-submanifold of Euclidean space $\mathbb{R}^m$ and let $p \in S$. We want to define the (geometric) *tangent space* to S at the point p; this is, locally at p, the "best" approximation of S by a linear ℓ-submanifold.

We shall base our definition of tangent space on the concept of (geometric) tangent vector to a curve in $\mathbb{R}^m$.

Let $\gamma : J \to \mathbb{R}^m$ be a parametrized curve in $\mathbb{R}^m$. (This means that J is an open interval of $\mathbb{R}$ and γ is a smooth mapping. Also, recall that the image set $\gamma(J) \subseteq \mathbb{R}^m$, the so-called trace of γ, is generally *not* a submanifold of $\mathbb{R}^m$.) The (geometric) *tangent vector* to γ at (the point) $\gamma(t)$ is the element

$$\dot{\gamma}(t) = \left(\frac{d\gamma_1}{dt}(t), \cdots, \frac{d\gamma_m}{dt}(t) \right) \in \mathbb{R}^m = T_{\gamma(t)}\mathbb{R}^m,$$

where $\gamma_i : J \to \mathbb{R}$, $i = 1, 2, \ldots, m$ are the component functions of γ.

3.4.11 DEFINITION. Let S be an ℓ-submanifold of $\mathbb{R}^m$ and let $p \in S$. A tangent vector $v \in \mathbb{R}^m = T_p\mathbb{R}^m$ is said to be a **geometric tangent vector** of S at p if there exist a parametrized curve $\gamma : J \to \mathbb{R}^m$ and $t_0 \in J$ such that

(GTV1) $\gamma(t) \in S$ for all $t \in J$;

(GTV2) $\gamma(t_0) = p$;

(GTV3) $\dot{\gamma}(t_0) = v$.

NOTE : We are dealing with two kinds of tangent vectors : those that are "tangent" to the whole space (i.e., the Euclidean space $\mathbb{R}^m$) and those that are tangent to a specific submanifold; the latter will be referred to as *geometric* tangent vectors in order to avoid ambiguity.

The set of all geometric tangent vectors of S at p is denoted by T_pS and is called the **tangent space** to S at p.

NOTE : By definition, T_pS is a subset of (the vector space) $T_p\mathbb{R}^m = \mathbb{R}^m$. It turns out that it is, in fact, a vector subspace of the tangent space $T_p\mathbb{R}^m$. When regarded as a subset of (Euclidean space) $\mathbb{R}^m$, the tangent space T_pS is better viewed as a linear submanifold which is *tangent* to (i.e., has a contact of order one with) the smooth submanifold S at the point p. It is common to refer to $p + T_pS$, as the **geometric tangent space** of S at p. (Obviously, the point p plays an important role in this linear submanifold. By choosing the point p as the origin, one obtains the identification of the geometric tangent space with the vector space T_pS.) Experience

shows that is convenient to regard the tangent space to S at p as a vector space (i.e., to identify the linear submanifold $p + T_pS$ with its direction space T_pS).

Let $v \in T_pS$ and $\lambda \in \mathbb{R}$. Then $\lambda v \in T_pS$. Indeed, we may assume that $\gamma(t) \in S$ for al $t \in J$ with $\gamma(0) = p$ and $\dot{\gamma}(0) = v$. Consider the parametrized curve $\gamma_\lambda : t \mapsto \gamma(\lambda t)$. One has, for sufficiently small t, $\gamma_\lambda(t) \in S$. Also

$$\gamma_\lambda(0) = p \quad \text{and} \quad \dot{\gamma}_\lambda(0) = \lambda v.$$

Hence $\lambda v \in T_pS$. It is less obvious that if $v, w \in T_pS$, then $v + w \in T_pS$.

3.4.12 THEOREM. *Let S be an ℓ-submanifold of $\mathbb{R}^m$ and let $p \in S$. Assume that, locally at p, S is described as in* THEOREM 3.4.2. *Then*

$$\begin{aligned} T_pS &= \ker\left(DF(p)\right) \\ &= \operatorname{graph}\left(DH(z)\right) \\ &= \operatorname{im}\left(D\Phi(0)\right). \end{aligned}$$

In particular, T_pS is an ℓ-dimensional vector subspace of $\mathbb{R}^m = T_p\mathbb{R}^m$.

PROOF : Since S is an ℓ-submanifold of $\mathbb{R}^m$ and $p \in S$, there exists an open neighborhood U of p in $\mathbb{R}^m$ such that we can write

- $S \cap U = F^{-1}(0)$, where $F : U \to \mathbb{R}^{m-\ell}$ is a smooth submersion;
- $S \cap U = \operatorname{graph}(H)$, where $H : W \subseteq \mathbb{R}^\ell \to \mathbb{R}^{m-\ell}$ is a smooth mapping;
- $S \cap U = \operatorname{im}(\Phi)$, where $\Phi : V \subseteq \mathbb{R}^\ell \to \mathbb{R}^m$ is a smooth embedding.

In particular, we assume that

$$\begin{aligned} p &= \left(z, H(z)\right), \quad z \in W \subseteq \mathbb{R}^\ell \\ &= \Phi(y), \quad y \in V \subseteq \mathbb{R}^\ell \end{aligned}$$

and

$$F(p) = 0 \in \mathbb{R}^{m-\ell}.$$

Let $h \in \mathbb{R}^\ell$. Then there exists an $\varepsilon > 0$ such that $z + th \in W$ for all $|t| < \varepsilon$. Consequently,

$$\gamma : t \mapsto \left(z + th, H(z + th)\right), \quad |t| < \varepsilon$$

is a smooth curve in $\mathbb{R}^m$ such that

– $\gamma(t) \in V$;
– $\gamma(0) = (z, H(z)) = p$;
– $\dot{\gamma}(0) = (h, DH(z) \cdot h)$.

This implies

$$\text{graph}\,(H) \subset T_pS.$$

It is equally true that

$$\text{im}\ (D\Phi(y)) \subset T_pS.$$

Hence

$$\text{graph}\,(H) \cup \text{im}\ (D\Phi(y)) \subset T_pS. \qquad (*)$$

Now let $v \in T_pS$ and assume $v = \dot{\gamma}(t_0)$. Then we have $(F \circ \gamma)(t) = 0,\ t \in J$ and hence (by differentiation)

$$0 = D(F \circ \gamma)(t_0) = DF(p) \circ \dot{\gamma}(t_0) = DF(p) \cdot v.$$

Therefore

$$T_pS \subset \ker\,(DF(p))\,. \qquad (**)$$

Since the linear mappings $h \mapsto (h, DH(z) \cdot h)$ and $D\Phi(p)$ are injective and surjective, respectively, from $(*)$ and $(*)$ it follows that

$$\dim \text{graph}\,(DH(z)) = \dim \text{im}\,(D\Phi(y)) = \dim \ker(DF(p)) = \ell.$$

This proves the result. □

◇ **Exercise 168** Let S be an ℓ-submanifold of $\mathbb{R}^m$ and let $p \in S$. Assume that, locally at p, S is described as in THEOREM 3.4.2 (i). Prove that

$$T_pS = (D\phi^{-1}(0))(\mathbb{R}^\ell).$$

3.4.13 EXAMPLE. Let $H = (h_1, h_2) : dom\,(H) \subseteq \mathbb{R} \to \mathbb{R}^2$ be a smooth mapping. The submanifold

$$S = \{(t, h_1(t), h_2(t)) \in \mathbb{R}^3 \,|\, t \in dom\,(H)\}$$

is the (geometric) *curve* in $\mathbb{R}^3$ given as the graph of H.

Then

$$\operatorname{graph}(DH(t)) = \mathbb{R}\,(1, \dot{h}_1(t), \dot{h}_2(t)).$$

A parametric representation of the *geometric tangent line* of S at $(t, H(t))$ (with $t \in R$ fixed) is

$$x = (t, h_1(t), h_2(t)) + \lambda \left(1, \dot{h}_1(t), \dot{h}_2(t)\right), \quad \lambda \in \mathbb{R}.$$

3.4.14 EXAMPLE. Let $S \subset \mathbb{R}^3$ be the (geometric) *helix* such that

$$S \subset \{x \in \mathbb{R}^3 \,|\, x_1^2 + x_2^2 = 1\} \quad \text{and}$$
$$S \cap \{x \in \mathbb{R}^3 \,|\, x_3 = 2k\pi\} = \{(1, 0, 2k\pi)\}, \quad k \in \mathbb{Z}.$$

Then S is the graph of the smooth mapping

$$H : \mathbb{R} \to \mathbb{R}^2, \quad t \mapsto (\cos t, \sin t).$$

That is,

$$S = \{(\cos t, \sin t, t) \,|\, t \in \mathbb{R}\}.$$

It follows that S is a smooth submanifold of $\mathbb{R}^3$ of dimension 1. Moreover, S is a zero-set. Indeed, we have

$$x \in S \iff F(x) = (x_1 - \cos x_3,\ x_2 - \sin x_3) = 0.$$

For $x = (H(t), t)$ we obtain

$$\begin{aligned} T_x S = \operatorname{graph}(DH(t)) &= \mathbb{R}\,(-\sin t, \cos t, 1) \\ &= \mathbb{R}\,(-x_2, x_1, 1), \end{aligned}$$

$$DF(x) = \begin{bmatrix} 1 & 0 & \sin x_3 \\ 0 & 1 & -\cos x_3 \end{bmatrix}.$$

The parametric representation of the geometric tangent line of S at $x = (H(t), t)$ is

$$(\cos t, \sin t, t) + \lambda(-\sin t, \cos t, 1), \quad \lambda \in \mathbb{R}.$$

3.4.15 EXAMPLE. The submanifold $S \subset \mathbb{R}^3$ is given by a smooth embedding $\Phi : dom\,(\Phi) \subseteq \mathbb{R}^2 \to \mathbb{R}^3$. That is,

$$S = \left\{ \Phi(y) \in \mathbb{R}^3 \,|\, y \in dom\,(\Phi) \right\}.$$

Then

$$\begin{aligned} D\Phi(y) &= \left[\frac{\partial \Phi}{\partial y_1}(y) \quad \frac{\partial \Phi}{\partial y_2}(y) \right] \\ &= \begin{bmatrix} \frac{\partial \Phi_1}{\partial y_1}(y) & \frac{\partial \Phi_1}{\partial y_2}(y) \\ \frac{\partial \Phi_2}{\partial y_1}(y) & \frac{\partial \Phi_2}{\partial y_2}(y) \\ \frac{\partial \Phi_3}{\partial y_1}(y) & \frac{\partial \Phi_3}{\partial y_2}(y) \end{bmatrix}. \end{aligned}$$

The tangent space T_xS, with $x = \Phi(y)$, is spanned by the (tangent) vectors

$$\frac{\partial \Phi}{\partial y_1}(y) \quad \text{and} \quad \frac{\partial \Phi}{\partial y_2}(y)\cdot$$

Therefore, a parametric representation of the geometric tangent plane of S at $\Phi(y)$ is

$$\begin{aligned} u_1 &= \Phi_1(y) + \lambda_1 \frac{\partial \Phi_1}{\partial y_1}(y) + \lambda_2 \frac{\partial \Phi_1}{\partial y_2}(y) \\ u_2 &= \Phi_2(y) + \lambda_1 \frac{\partial \Phi_2}{\partial y_1}(y) + \lambda_2 \frac{\partial \Phi_2}{\partial y_2}(y) \\ u_3 &= \Phi_3(y) + \lambda_1 \frac{\partial \Phi_3}{\partial y_1}(y) + \lambda_2 \frac{\partial \Phi_3}{\partial y_2}(y), \quad \lambda \in \mathbb{R}^2. \end{aligned}$$

It turns out that

$$T_xS = \left\{ h \in \mathbb{R}^3 \,|\, h \bullet \frac{\partial \Phi}{\partial y_1}(y) \times \frac{\partial \Phi}{\partial y_2}(y) = 0 \right\}.$$

3.4.16 EXAMPLE. The submanifold $S \subset \mathbb{R}^3$ is the (geometric) *curve* in $\mathbb{R}^3$ given as a zero-set of a smooth submersion $F : dom\,(F) \subseteq \mathbb{R}^3 \to \mathbb{R}^2$. That is,

$$x \in S \iff F(x) = (f_1(x), f_2(x)) = 0.$$

Then

$$DF(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x}(x) \\ \frac{\partial f_2}{\partial x}(x) \end{bmatrix}$$

and thus

$$\ker\left(DF(x)\right) = \left\{h \in \mathbb{R}^3 \,|\, \operatorname{grad} f_1(x) \bullet h = \operatorname{grad} f_2(x) \bullet h = 0\right\}.$$

The tangent space T_xS is seen to be the line in $\mathbb{R}^3$ through the origin, formed by intersection of two planes

$$\{h \in \mathbb{R}^3 \,|\, \operatorname{grad} f_1(x) \bullet h = 0\} \quad \text{and} \quad \{h \in \mathbb{R}^3 \,|\, \operatorname{grad} f_2(x) \bullet h = 0\}.$$

◇ **Exercise 169** Let S be the *hyperboloid of two sheets*

$$S = \{(a\sinh y_1 \cos y_2,\, b\sinh y_1 \sin y_2,\, c\cosh y_1) \,|\, y = (y_1, y_2) \in \mathbb{R}^2\}, \quad a, b, c > 0.$$

(a) Show that S is a smooth *submanifold* of $\mathbb{R}^3$ of dimension 2.

(b) Determine the *geometric tangent space* of S at an arbitrary point p of S in three ways, by successively considering S as a zero-set, a parametrized set and a graph.

◇ **Exercise 170** Let $Q \subset \mathbb{R}^m$ be a *nondegenerate quadric* given by

$$Q = \{x \in \mathbb{R}^m \,|\, x^\top A x + b^\top x + c = 0\}.$$

Let $x \in Q \setminus \left\{-\frac{1}{2}A^{-1}b\right\}$.

(a) Prove that

$$T_xQ = \{h \in \mathbb{R}^m \,|\, (2Ax + b) \bullet h = 0\}.$$

(b) Prove that

$$\begin{aligned} x + T_xQ &= \{h \in \mathbb{R}^m \,|\, (2Ax + b) \bullet (x - h) = 0\} \\ &= \{h \in \mathbb{R}^m \,|\, (Ax) \bullet h + \frac{1}{2}b \bullet (x + h) + c = 0\}. \end{aligned}$$

. . . . . .

Chapter 4

Matrix Groups

Topics :

1. Real and Complex Matrix Groups
2. Examples of Matrix Groups
3. The Exponential Mapping
4. Lie Algebras for Matrix Groups
5. More Properties of the Exponential Mapping
6. Examples of Lie Algebras of Matrix Groups

4.1 Real and Complex Matrix Groups

Throughout, we shall denote by $\mathbb{k}$ either the *field* $\mathbb{R}$ of real numbers or the *field* $\mathbb{C}$ of complex numbers.

The algebra of $n \times n$ matrices over $\mathbb{k}$

Let $\mathbb{k}^m$ be the set of all m-tuples of elements of $\mathbb{k}$. Under the usual addition and scalar multiplication, $\mathbb{k}^m$ is a *vector space* over $\mathbb{k}$. The set $\mathsf{Hom}\,(\mathbb{k}^n, \mathbb{k}^m)$ of all linear mappings from $\mathbb{k}^n$ to $\mathbb{k}^m$ (i.e., mappings $L : \mathbb{k}^n \to \mathbb{k}^m$ such that $L(\lambda x + \mu y) = \lambda L(x) + \mu L(y)$ for every $x, y \in \mathbb{k}^n$ and $\lambda, \mu \in \mathbb{k}$) is also a *vector space* over $\mathbb{k}$.

◇ **Exercise 171** Determine the *dimension* of the vector space $\mathsf{Hom}\,(\mathbb{k}^n, \mathbb{k}^m)$.

Let $\mathbb{k}^{m\times n}$ be the set of all $m \times n$ matrices with elements (entries) from $\mathbb{k}$. It is convenient to *identify*

the m-tuple $(a_1, a_2, \ldots, a_m) \in \mathbb{k}^m$ with the column m-matrix $\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \in \mathbb{k}^{m\times 1}$.

◇ **Exercise 172** Give reasons why the *identification* of $\mathbb{k}^m$ with $\mathbb{k}^{m\times 1}$ is legitimate.

Under the usual matrix addition and multiplication, $\mathbb{k}^{m\times n}$ is a *vector space* over $\mathbb{k}$. There is a natural one-to-one correspondence

$$A \mapsto L_A\,(: x \mapsto Ax)$$

between the $m \times n$ matrices with elements from $\mathbb{k}$ and the linear mappings from $\mathbb{k}^n$ to $\mathbb{k}^m$.

◇ **Exercise 173** Show that the vector spaces $\mathbb{k}^{m\times n}$ and $\mathsf{Hom}\,(\mathbb{k}^n, \mathbb{k}^m)$ are *isomorphic*. Observe that, in particular, the vector spaces $\mathbb{k}^{1\times n}$ and $\mathsf{Hom}\,(\mathbb{k}^n, \mathbb{k}) = (\mathbb{k}^n)^*$ (the *dual* of $\mathbb{k}^n$) are isomorphic.

NOTE : We *do not* identify the row n-matrix $\begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}$ with the n-tuple $(a_1, a_2, \dots, a_n)$ but rather with the linear mapping (functional)

$$(x_1, x_2, \dots, x_n) \mapsto a_1x_1 + a_2x_2 + \cdots + a_nx_n.$$

Any matrix $A \in \mathbb{k}^{m\times n}$ can be *interpreted* as a linear mapping $L_A \in \mathsf{Hom}\,(\mathbb{k}^n, \mathbb{k}^m)$, whereas any linear mapping $L \in \mathsf{Hom}\,(\mathbb{k}^n, \mathbb{k}^m)$ can be *realized* as a matrix $A \in \mathbb{k}^{m\times n}$. Henceforth we shall not distinguish notationwise between a matrix A and its corresponding linear mapping $x \mapsto Ax$.

NOTE : A matrix (or linear mapping, if one prefers) $A \in \mathbb{k}^{n\times n}$ can be viewed as a *vector field* (on $\mathbb{k}^n$) : A associates to each point p in $\mathbb{k}^n$ the tangent vector $A(p) = Ap \in \mathbb{k}^n$. We may think of a *fluid* in motion, so that the velocity of the fluid particles passing through p is always $A(p)$. The vector field is then the current of the *flow* and the paths of the fluid particles are the trajectories. This kind of flow is, of course, very special : $A(p)$ is independent of time, and depends linearly on p.

Notice that $\mathbb{k}^{n\times n}$ is not just a vector space. It also has a multiplication which is associative and distributes over addition (on either side). In other words, under the usual addition and multiplication, $\mathbb{k}^{n\times n}$ is a *ring* (in general not commutative), with identity I_n. Moreover, for all $A, B \in \mathbb{k}^{n\times n}$ and $\lambda \in \mathbb{k}$,

$$\lambda(AB) = (\lambda A)B = A(\lambda B).$$

Such a structure is called an (associative) *algebra* over $\mathbb{k}$.

◇ **Exercise 174** Give the definition of an *algebra* over (the field) $\mathbb{k}$. Write down *all* the axioms.

The topology of $\mathbb{k}^{n\times n}$

For $x \in \mathbb{k}^n$ $\left(= \mathbb{k}^{n\times 1}\right)$, let

$$\|x\|_2 := \sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2}$$

be the 2-norm (or *Euclidean norm*) on $\mathbb{k}^n$.

NOTE : For $r \geq 1$, the *r-norm* of $x \in \mathbb{k}^n$ is defined as

$$\|x\|_r := \left(|x_1|^r + |x_2|^r + \cdots + |x_n|^r\right)^{1/r}.$$

The following properties hold (for $x, y \in \mathbb{k}^n$ and $\lambda \in \mathbb{k}$) :

$$\|x\|_r \geq 0, \quad \text{and} \quad \|x\|_r = 0 \iff x = 0 \ ;$$
$$\|\lambda x\|_r = |\lambda| \, \|x\|_r \ ;$$
$$\|x + y\|_r \leq \|x\|_r + \|y\|_r.$$

In practice, only three of the r-norms are used, and they are :

$$\begin{aligned} \|x\|_1 &= |x_1| + |x_2| + \cdots + |x_n| \quad \text{(the grid norm);} \\ \|x\|_2 &= \sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2} \quad \text{(the Euclidean norm);} \\ \|x\|_\infty = \lim_{r \to \infty} \|x\|_r &= \max\{|x_1|, |x_2|, \ldots, |x_n|\} \quad \text{(the max norm).} \end{aligned}$$

For $x \in \Bbbk^n$, we have

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \cdot \|x\|_2 \leq n \cdot \|x\|_\infty$$

and so any two of these norms are *equivalent* (i.e., the associated metric topologies are identical). In fact, *all norms on a finite dimensional vector space (over $\Bbbk$) are equivalent.*

The metric topology *induced* by (the Euclidean distance)

$$(x, y) \mapsto \|x - y\|_2$$

is the *natural topology* on the set (vector space) $\Bbbk^n$.

◇ **Exercise 175** Show that, for $x, y \in \Bbbk^n$,

$$| \, \|x\|_2 - \|y\|_2 \, | \leq \|x - y\|_2.$$

Hence deduce that the function

$$\| \cdot \|_2 : \Bbbk^n \to \mathbb{R}, \quad x \mapsto \|x\|_2$$

is *continuous* (with respect to the natural topologies on $\Bbbk^n$ and $\mathbb{R}$).

◇ **Exercise 176** Given $A \in \Bbbk^{n \times n}$, show that the linear mapping (on $\Bbbk^n$) $x \mapsto Ax$ is *continuous* (with respect to the natural topology on $\Bbbk^n$).

Let $A \in \Bbbk^{n \times n}$. The 2-norm $\| \cdot \|_2$ on $\Bbbk^{n \times 1}$ induces a (matrix) norm on $\Bbbk^{n \times n}$ by setting

$$\|A\| := \max_{\|x\|_2 = 1} \|Ax\|_2.$$

The subset $K = \{x \in \mathbb{k}^n \,|\, \|x\|_2 = 1\} \subset \mathbb{k}^n$ is closed and bounded, and so is *compact*. [A subset of the metric space $\mathbb{k}^n$ is compact if and only if it is closed and bounded.] On the other hand, the function $f : K \to \mathbb{R}, \quad x \mapsto \|Ax\|_2$ is *continuous*. [The composition of two continuous maps is a continuous map.] Hence the maximum value $\max_{x\in K} \|Ax\|_2$ *must* exist.

NOTE : The following topological result holds : *If $K \subset \mathbb{k}^n$ is a (nonempty) compact set, then any continuous function $f : K \to \mathbb{R}$ is bounded; that is, the image set $f(K) = \{f(x) \,|\, x \in K\} \subseteq \mathbb{R}$ is bounded.* Moreover, f has a global maximum (and a global minimum).

◇ **Exercise 177** Show that the induced norm $\|\cdot\|$ is *compatible* with its underlying norm $\|\cdot\|_2$; that is (for $A \in \mathbb{k}^{n\times n}$ and $x \in \mathbb{k}^n$),

$$\|Ax\|_2 \leq \|A\| \, \|x\|_2.$$

$\|\cdot\|$ is a *matrix norm* on $\mathbb{k}^{n\times n}$, called the **operator norm**; that is, it has the following four properties (for $A, B \in \mathbb{k}^{n\times n}$ and $\lambda \in \mathbb{k}$) :

(MN1) $\|A\| \geq 0, \quad \text{and} \quad \|A\| = 0 \iff A = 0$;

(MN2) $\|\lambda A\| = |\lambda| \, \|A\|$;

(MN3) $\|A + B\| \leq \|A\| + \|B\|$;

(MN4) $\|AB\| \leq \|A\| \, \|B\|$.

NOTE : There is a simple procedure (well-known in numerical linear algebra) for calculating the operator norm of an $n \times n$ matrix A. This is

$$\|A\| = \sqrt{\lambda_{\max}},$$

where $\lambda_{\max}$ is the largest eigenvalue of the matrix A^*A. Here A^* denotes the *Hermitian conjugate* (i.e., the conjugate transpose) matrix of A; in the case $\mathbb{k} = \mathbb{R}$, $A^* = A^\top$.

We define a *metric* ρ on (the algebra) $\mathbb{k}^{n\times n}$ by

$$\rho(A, B) := \|A - B\|.$$

Associated to this metric is a natural *topology* on $\mathbb{k}^{n\times n}$. Hence fundamental topological concepts, like *open sets, closed sets, compactness, connectedness,* as well as *continuity*, can be introduced. In particular, we can speak of continuous functions from $\mathbb{k}^{n\times n}$ into $\mathbb{k}$.

◇ **Exercise 178** For $1 \le i, j \le n$, show that the coordinate function

$$\mathrm{coord}_{ij} : \mathbb{k}^{n\times n} \to \mathbb{k}, \quad A \mapsto a_{ij}$$

is *continuous*. [HINT : Show first that $|a_{ij}| \le \|A\|$ and then verify the defining condition for continuity.]

It follows immediately that *if* $f : \mathbb{k}^{n^2} \to \mathbb{k}$ *is continuous, then the associated function*

$$\widetilde{f} = f \circ (\mathrm{coord}_{ij}) : \mathbb{k}^{n\times n} \to \mathbb{k}, \quad A \mapsto f((a_{ij}))$$

is also continuous. Here $(a_{ij}) = (a_{11}, a_{12}, \ldots, a_{1n}, a_{21}, \ldots, a_{nn}) \in \mathbb{k}^{n^2}$.

◇ **Exercise 179** Show that the *determinant function*

$$\det : \mathbb{k}^{n\times n} \to \mathbb{k}, \quad A \mapsto \det A := \sum_{\sigma\in S_n} (-1)^{|\sigma|} a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)}$$

and the *trace function*

$$\mathrm{tr} : \mathbb{k}^{n\times n} \to \mathbb{k}, \quad A \mapsto \mathrm{tr}\, A := \sum_{i=1}^{n} a_{ii}$$

are *continuous.*

The metric space $(\mathbb{k}^{n\times n}, \rho)$ is *complete.* This means that *every Cauchy sequence* $(A_r)_{r\ge 0}$ *in* $\mathbb{k}^{n\times n}$ *has a unique limit* $\lim_{r\to\infty} A_r$. Furthermore,

$$\Big(\lim_{r\to\infty} A_r\Big)_{ij} = \lim_{r\to\infty} (A_r)_{ij}.$$

Indeed, the limit on the RHS exists, so it is sufficient to check that the required matrix limit is the matrix A with $a_{ij} = \lim_{r\to\infty} (A_r)_{ij}$. The sequence $(A_r - A)_{r\ge 0}$ satisfies

$$\|A_r - A\| \le \sum_{i,j=1}^{n} |(A_r)_{ij} - a_{ij}| \to 0 \quad \text{as } r \to \infty$$

and so $A_r \to A$.

Groups of matrices

Let $\mathsf{GL}(n,\mathbb{k})$ be the set of all invertible $n \times n$ matrices over $\mathbb{k}$ (or, equivalently, the set of all linear transformations on $\mathbb{k}^n$). So

$$\mathsf{GL}(n,\mathbb{k}) := \{A \in \mathbb{k}^{n\times n} \,|\, \det A \neq 0\}.$$

$\diamond$ **Exercise 180** Verify that the set $\mathsf{GL}(n,\mathbb{k})$ is a *group* under matrix multiplication.

$\mathsf{GL}(n,\mathbb{k})$ is called the **general linear group** over $\mathbb{k}$. We will refer to $\mathsf{GL}(n,\mathbb{R})$ and $\mathsf{GL}(n,\mathbb{C})$ as the *real* and *complex* general linear group, respectively.

A 1×1 matrix over $\mathbb{k}$ is just an element of $\mathbb{k}$ and matrix multiplication of two such elements is just multiplication in $\mathbb{k}$. So we see that

$$\mathsf{GL}(1,\mathbb{k}) = \mathbb{k}^\times \quad \text{(the multiplicative group of } \mathbb{k} \setminus \{0\}).$$

4.1.1 Proposition. $\mathsf{GL}(n,\mathbb{k})$ *is an open subset of* $\mathbb{k}^{n\times n}$.

Proof : We have seen that the function $\det : \mathbb{k}^{n\times n} \to \mathbb{k}$ is continuous (see **Exercise 179**). Then observe that

$$\mathsf{GL}(n,\mathbb{k}) = \mathbb{k}^{n\times n} \setminus \det{}^{-1}(0).$$

Since the set $\{0\}$ is closed (in $\mathbb{k}$), it follows that $\det^{-1}(0) = \det^{-1}(\{0\}) \subset \mathbb{k}^{n\times n}$ is also closed. [The preimage of a closed set under a continuous map is a closed set.] Hence $\mathsf{GL}(n,\mathbb{k})$ is open. [The complement of a closed set is an open set.] $\square$

Let G be a subgroup of the general linear group $\mathsf{GL}(n,\mathbb{k})$. If G is also a closed subspace of $\mathsf{GL}(n,\mathbb{k})$, we say that G is a *closed* subgroup.

4.1.2 Definition. A closed subgroup of $\mathsf{GL}(n,\mathbb{k})$ is called a **matrix group** over $\mathbb{k}$ (or a **matrix subgroup** of $\mathsf{GL}(n,\mathbb{k})$).

Matrix groups are also known as *linear* groups or even as *matrix Lie* groups. This latter terminology emphasizes the remarkable fact that *every matrix group is a Lie group.*

NOTE : The condition that the group of matrices $G \subseteq \mathsf{GL}(n,\Bbbk)$ is a closed subset of (the metric space) $\mathsf{GL}(n,\Bbbk)$ means that the following condition is satisfied : *if $(A_r)_{r\geq 0}$ is any sequence of matrices in G and $A_r \to A$, then either $A \in G$ or A is not invertible (i.e. $A \notin \mathsf{GL}(n,\Bbbk)$).*

The condition that G be a *closed* subgroup, as opposed to merely a subgroup, should be regarded as a "technicality" since most of the *interesting* subgroups of $\mathsf{GL}(n,\Bbbk)$ have this property. Almost all of the matrix groups we will consider have the stronger property that if $(A_r)_{r\geq 0}$ is any sequence of matrices in G converging to some matrix A, then $A \in G$.

We will often use the notation $G \leq \mathsf{GL}(n,\Bbbk)$ to indicate that G is a (matrix) subgroup of $\mathsf{GL}(n,\Bbbk)$.

4.1.3 EXAMPLE. The general linear group $\mathsf{GL}(n,\Bbbk)$ is a matrix group (over $\Bbbk$).

4.1.4 EXAMPLE. An example of a group of matrices which is *not* a matrix group is the set of all $n \times n$ invertible matrices all of whose entries are rational numbers. This is in fact a subgroup of $\mathsf{GL}(n,\mathbb{C})$ but not a closed subgroup; that is, one can (easily) have a sequence of invertible matrices with rational entries converging to an invertible matrix with some irrational entries.

◇ **Exercise 181** * Let $a \in \mathbb{R} \setminus \mathbb{Q}$. Show that

$$G = \left\{ \begin{bmatrix} e^{it} & 0 \\ 0 & e^{iat} \end{bmatrix} \mid t \in \mathbb{R} \right\}$$

is a subgroup of $\mathsf{GL}(2,\mathbb{C})$, and then find a sequence of matrices in G which converges to $-I_2 \notin G$. This means that G is *not* a matrix group. [HINT : By taking $t = (2n+1)\pi$ for a suitably chosen $n \in \mathbb{Z}$, we can make ta *arbitrarily close* to an odd integer multiple of π, $(2m+1)\pi$ say. It is sufficient to show that for any positive integer N, there exist $n, m \in \mathbb{Z}$ such that $|(2n+1)a - (2m+1)| < \frac{1}{N}$.]

NOTE : The *closure* of G (in $\mathsf{GL}(2,\mathbb{C})$) can be thought of as (the direct product) $\mathbb{S}^1 \times \mathbb{S}^1$ and so *is* a matrix group (see **Exercise 195**).

4.1.5 PROPOSITION. *Let G be a matrix group over $\Bbbk$ and H a closed subgroup of G. Then H is a matrix group over $\Bbbk$.*

PROOF : Every sequence $(A_r)_{r\geq 0}$ in H with a limit in $\mathsf{GL}(n,\Bbbk)$ actually has its limit in G since each $A_r \in H \subseteq G$ and G is closed in $\mathsf{GL}(n,\Bbbk)$. Since H is closed in G, this means that $(A_r)_{r\geq 0}$ has a limit in H. So H is closed in $\mathsf{GL}(n,\Bbbk)$, showing it is a matrix group over $\Bbbk$. □

◇ **Exercise 182** Prove that any *intersection* of matrix groups (over $\Bbbk$) is a matrix group.

4.1.6 EXAMPLE. Denote by $\mathsf{SL}(n,\Bbbk)$ the set of all $n\times n$ matrices over $\Bbbk$, having determinant one. So

$$\mathsf{SL}(n,\Bbbk) := \{A \in \Bbbk^{n\times n} \mid \det A = 1\} \subseteq \mathsf{GL}(n,\Bbbk).$$

◇ **Exercise 183** Show that $\mathsf{SL}(n,\Bbbk)$ is a closed subgroup of $\mathsf{GL}(n,\Bbbk)$ and hence is a matrix group over $\Bbbk$.

$\mathsf{SL}(n,\Bbbk)$ is called the **special linear group** over $\Bbbk$. We will refer to $\mathsf{SL}(n,\mathbb{R})$ and $\mathsf{SL}(n,\mathbb{C})$ as the *real* and *complex* special linear groups, respectively.

4.1.7 DEFINITION. A closed subgroup of a matrix group G is called a **matrix subgroup** of G.

4.1.8 EXAMPLE. We can consider $\mathsf{GL}(n,\Bbbk)$ as a subgroup of $\mathsf{GL}(n+1,\Bbbk)$ by *identifying* the $n\times n$ matrix $A = \left[a_{ij}\right]$ with

$$\begin{bmatrix} 1 & 0 \\ 0 & A \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & a_{11} & \dots & a_{1n} \\ 0 & a_{21} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n1} & \dots & a_{nn} \end{bmatrix}.$$

It is easy to verify that $\mathsf{GL}(n,\Bbbk)$ is closed in $\mathsf{GL}(n+1,\Bbbk)$ and hence $\mathsf{GL}(n,\Bbbk)$ is a matrix subgroup of $\mathsf{GL}(n+1,\Bbbk)$.

◇ **Exercise 184** Show that $\mathsf{SL}(n,\Bbbk)$ is a matrix subgroup of $\mathsf{SL}(n+1,\Bbbk)$.

4.2 Examples of Matrix Groups

The vector space $\mathbb{k}^{n\times n}$ over $\mathbb{k}$ can be considered to be a *real* vector space, of dimension n^2 or $2n^2$, respectively. Explicitly, $\mathbb{R}^{n\times n}$ is (isomorphic to) $\mathbb{R}^{n^2}$, and $\mathbb{C}^{n\times n}$ is (isomorphic to) $\mathbb{C}^{n^2} = \mathbb{R}^{2n^2}$. Hence we may assume, without any loss of generality, that $\mathbb{k}^{n\times n}$ is some Euclidean space $\mathbb{R}^m$.

The real general linear group $\mathsf{GL}(n,\mathbb{R})$

We showed that $\mathsf{GL}(n,\mathbb{R})$ is a matrix group and that it is an *open* subset of the vector space $\mathbb{R}^{n\times n}$ $\left(=\mathbb{R}^{n^2}\right)$. Since the set $\mathsf{GL}(n,\mathbb{R})$ is not closed, it is *not* compact. [Any compact set is a closed set.]

The determinant function $\det : \mathsf{GL}(n,\mathbb{R}) \to \mathbb{R}$ is continuous (in fact, *smooth*) and maps $\mathsf{GL}(n,\mathbb{R})$ onto the two *components* of $\mathbb{R}^\times$. Thus $\mathsf{GL}(n,\mathbb{R})$ is *not* connected. [The image of a connected set under a continuous map is a connected set.]

NOTE : A matrix group G is said to be **connected** if given any two matrices $A, B \in G$, there exists a continuous *path* $\gamma : [a,b] \to G$ with $\gamma(a) = A$ and $\gamma(b) = B$. This property is what is called **path-connectedness** in topology, which is not (in general) the same as connectedness. However, it is a fact (not particularly obvious at the moment) that *a matrix group is connected if and only if it is path-connected.* So in a slight abuse of terminology we shall continue to refer to the above property as *connectedness.*

A matrix group G which is not connected can be decomposed (uniquely) as a union of several pieces, called *components*, such that two elements of the same component can be joined by a continuous path, but two elements of different components cannot. The component of G containing the identity is a closed subgroup of G (and hence a connected matrix group).

Consider the sets

$$\begin{aligned}\mathsf{GL}^+(n,\mathbb{R}) &:= \{A \in \mathsf{GL}(n,\mathbb{R}) \,|\, \det A > 0\}\\ \mathsf{GL}^-(n,\mathbb{R}) &:= \{B \in \mathsf{GL}(n,\mathbb{R}) \,|\, \det B < 0\}.\end{aligned}$$

These two disjoint subsets of $\mathsf{GL}(n,\mathbb{R})$ are *open* and such that

$$\mathsf{GL}^+(n,\mathbb{R}) \cup \mathsf{GL}^-(n,\mathbb{R}) = \mathsf{GL}(n,\mathbb{R}).$$

[The preimage of an open set under a continuous map is an open set.]

◇ **Exercise 185** Show that $\mathsf{GL}^+(n,\mathbb{R})$ is a matrix subgroup of $\mathsf{GL}(n,\mathbb{R})$ but $\mathsf{GL}^-(n,\mathbb{R})$ is not.

The mapping

$$A \in \mathsf{GL}^+(n,\mathbb{R}) \mapsto SA \in \mathsf{GL}^-(n,\mathbb{R})$$

where $S = \operatorname{diag}(1,1,\ldots,1,-1)$, is a bijection (in fact, a *diffeomorphism*). The transformation $x \mapsto Sx$ may be thought of as a *reflection* in the *hyperplane* $\mathbb{R}^{n-1} = \mathbb{R}^{n-1} \times \{0\} \subset \mathbb{R}^n$.

NOTE : The group $\mathsf{GL}^+(n,\mathbb{R})$ is *connected*, which proves that $\mathsf{GL}^+(n,\mathbb{R})$ *is the connected component of the identity in* $\mathsf{GL}(n,\mathbb{R})$ and that $\mathsf{GL}(n,\mathbb{R})$ has two (connected) components.

The real special linear group $\mathsf{SL}(n,\mathbb{R})$

Recall that

$$\mathsf{SL}(n,\mathbb{R}) := \{A \in \mathsf{GL}(n,\mathbb{R}) \,|\, \det A = 1\} = \det^{-1}(1).$$

It follows that $\mathsf{SL}(n,\mathbb{R})$ is a closed subgroup of $\mathsf{GL}(n,\mathbb{R})$ and hence is a matrix group. [The preimage of a closed set under a continuous map is a closed set.] We introduce a new matrix norm on $\mathbb{R}^{n\times n}$, called the *Frobenius norm*, as follows :

$$\|A\|_F := \sqrt{\operatorname{tr}(A^\top A)} = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2}.$$

NOTE : The Frobenius norm coincides with the Euclidean norm on $\mathbb{R}^{n^2}$, and is much easier to compute than the operator norm. However, *all* matrix norms on $\mathbb{R}^{n\times n}$ are *equivalent* (i.e., they generate the same metric topology).

We shall use this (matrix) norm to show that $\mathsf{SL}(n,\mathbb{R})$ is *not* compact. Indeed, all matrices of the form

$$\begin{bmatrix} 1 & 0 & \dots & t \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

are elements of $\mathsf{SL}(n,\mathbb{R})$ whose norm equals $\sqrt{n+t^2}$ for any $t \in \mathbb{R}$. Thus $\mathsf{SL}(n,\mathbb{R})$ is *not* a bounded subset of $\mathbb{R}^{n\times n}$ and hence is *not* compact. [In a metric space, any compact set is bounded.]

NOTE : The special linear group $\mathsf{SL}(n,\mathbb{R})$ is *connected.*

More on $\mathsf{SL}(2,\mathbb{R})$. .....

The orthogonal and special orthogonal groups $\mathsf{O}(n)$ and $\mathsf{SO}(n)$

The set

$$\mathsf{O}(n) := \{A \in \mathbb{R}^{n\times n} \mid A^\top A = I_n\}$$

is the **orthogonal group**. Clearly, every *orthogonal matrix* $A \in \mathsf{O}(n)$ has an inverse, namely $A^\top$. Hence $\mathsf{O}(n) \subseteq \mathsf{GL}(n,\mathbb{R})$.

◇ **Exercise 186** Verify that $\mathsf{O}(n)$ is a *subgroup* of the general linear group $\mathsf{GL}(n,\mathbb{R})$.

The single matrix equation $A^\top A = I_n$ is equivalent to n^2 equations for the n^2 real numbers a_{ij}, $i,j = 1,2,\dots,n$:

$$\sum_{k=1}^{n} a_{ki}a_{kj} = \delta_{ij}.$$

This means that $\mathsf{O}(n)$ is a *closed* subset of $\mathbb{R}^{n\times n}$ and hence of $\mathsf{GL}(n,\mathbb{R})$.

◇ **Exercise 187** Prove that $\mathsf{O}(n)$ is a closed subset of $\mathbb{R}^{n^2}$.

Thus $\mathsf{O}(n)$ is a matrix group. The group $\mathsf{O}(n)$ is also *bounded* in $\mathbb{R}^{n\times n}$. Indeed, the (Frobenius) norm of $A \in \mathsf{O}(n)$ is

$$\|A\|_F = \sqrt{\operatorname{tr}(A^\top A)} = \sqrt{\operatorname{tr} I_n} = \sqrt{n}.$$

Hence the group $\mathsf{O}(n)$ is *compact*. [A subset of $\mathbb{R}^{n\times n}$ is compact if and only if it is closed and bounded.]

Let us consider the determinant function (restricted to $\mathsf{O}(n)$), $\det : \mathsf{O}(n) \to \mathbb{R}^\times$. Then for $A \in \mathsf{O}(n)$

$$\det I_n = \det(A^\top A) = \det A^\top \cdot \det A = (\det A)^2.$$

Hence $\det A = \pm 1$. So we have

$$\mathsf{O}(n) = \mathsf{O}^+(n) \cup \mathsf{O}^-(n)$$

where

$$\mathsf{O}^+(n) := \{A \in \mathsf{O}(n) \,|\, \det A = 1\} \quad \text{and} \quad \mathsf{O}^-(n) := \{A \in \mathsf{O}(n) \,|\, \det A = -1\}.$$

NOTE : The group $\mathsf{O}^+(n)$ is *connected*, which proves that $\mathsf{O}^+(n)$ *is the connected component of the identity in* $\mathsf{O}(n)$.

The **special orthogonal group** is defined as

$$\mathsf{SO}(n) := \mathsf{O}(n) \cap \mathsf{SL}(n, \mathbb{R}).$$

That is,

$$\mathsf{SO}(n) = \{A \in \mathsf{O}(n) \,|\, \det A = 1\} = \mathsf{O}^+(n).$$

It follows that $\mathsf{SO}(n)$ is a closed subset of $\mathsf{O}(n)$ and hence is *compact*. [A closed subset of a compact set is compact.]

NOTE : One of the main reasons for the study of these groups $\mathsf{O}(n), \mathsf{SO}(n)$ is their relationship with *isometries* (i.e., distance-preserving transformations on the Euclidean space $\mathbb{R}^n$). If such an isometry fixes the origin, then it is actually a linear transformation and so – with respect to the standard basis – corresponds to a matrix A. The isometry condition is equivalent to the fact that (for all $x, y \in \mathbb{R}^n$)

$$Ax \bullet Ay = x \bullet y,$$

which in turn is equivalent to the condition that $A^\top A = I_n$ (i.e., A is orthogonal). Elements of $\mathsf{SO}(n)$ are (identified with) *rotations* (or direct isometries); elements of $\mathsf{O}^-(n)$ are sometimes referred to as indirect isometries.

The Lorentz group $\mathsf{Lor}(1,n)$

Consider the inner-product (i.e., nondegenerate symmetric bilinear form) $\odot$ on (the vector space) $\mathbb{R}^{n+1}$ given by (for $x, y \in \mathbb{R}^{n+1}$)

$$x \odot y := -x_1 y_1 + \sum_{i=2}^{n+1} x_i y_i$$

(the so-called *Minkowski product*). It is standard to denote this inner-product space by $\mathbb{R}^{1,n}$.

◇ **Exercise 188** Show that the group of all linear isometries (i.e., linear transformations on $\mathbb{R}^{1,n}$ that preserve the Minkowski product) is *isomorphic* to the matrix group

$$\mathsf{O}(1,n) := \left\{ A \in \mathsf{GL}(n+1,\mathbb{R}) \,|\, A^\top S A = S \right\}$$

where

$$S = \operatorname{diag}(-1,1,1,\ldots,1) = \begin{bmatrix} -1 & 0 \\ 0 & I_n \end{bmatrix} \in \mathsf{GL}(n+1,\mathbb{R}).$$

In a similar fashion, one can define more general matrix groups

$$\mathsf{O}(k,\ell) \leq \mathsf{GL}(k+\ell,\mathbb{R}) \quad \text{and} \quad \mathsf{SO}(k,\ell) \leq \mathsf{SL}(k+\ell,\mathbb{R})$$

usually called "pseudo-orthogonal" groups.

◇ **Exercise 189** Define the inner-product $\langle \cdot,\cdot \rangle_{k,\ell}$ on $\mathbb{R}^{k+\ell}$ by the formula

$$\langle x,y \rangle_{k,\ell} := -x_1 y_1 - \cdots - x_k y_k + x_{k+1} y_{k+1} + \cdots + x_{k+\ell} y_{k+\ell}.$$

The *pseudo-orthogonal group* $\mathsf{O}(k,\ell)$ consists of all matrices $A \in \mathsf{GL}(k+\ell,\mathbb{R})$ which preserve this inner-product (i.e., such that $\langle Ax, Ay \rangle_{k,\ell} = \langle x,y \rangle_{k,\ell}$ for all $x, y \in \mathbb{R}^{k+\ell}$).

(a) Verify that $\mathsf{O}(k,\ell)$ is a *matrix subgroup* of $\mathsf{GL}(k+\ell,\mathbb{R})$.

(b) Let

$$Q = \operatorname{diag}(-1, \ldots, -1, 1, \ldots, 1) = \begin{bmatrix} -I_k & 0 \\ 0 & I_\ell \end{bmatrix}.$$

Prove that a matrix $A \in \mathsf{GL}(k+\ell, \mathbb{R})$ is in $\mathsf{O}(k, \ell)$ if and only if $A^\top Q A = Q$. Hence deduce that $\det A = \pm 1$.

(c) Verify that $\mathsf{SO}(k, \ell) := \mathsf{O}(k, \ell) \cap \mathsf{SL}(k + \ell, \mathbb{R})$ is a *matrix subgroup* of $\mathsf{SL}(k + \ell, \mathbb{R})$.

NOTE : Since $\mathsf{O}(k, \ell)$ and $\mathsf{O}(\ell, k)$ are essentially the same group, we may assume (without any loss of generality) that $1 \leq k \leq \ell$. *The pseudo-orthogonal groups are neither compact nor connected.* The groups $\mathsf{O}(k, \ell)$ have four (connected) components, whereas $\mathsf{SO}(k, \ell)$ have two components.

For each positive number $\rho > 0$, the *hyperboloid*

$$\mathcal{H}_{1,n}(\rho) := \left\{ x \in \mathbb{R}^{1,n} \mid \langle x, x \rangle = -\rho \right\}$$

has two (connected) components

$$\mathcal{H}^+_{1,n}(\rho) = \{x \in \mathcal{H}_{1,n}(\rho) \mid x_1 > 0\} \quad \text{and} \quad \mathcal{H}^-_{1,n}(\rho) = \{x \in \mathcal{H}_{1,n}(\rho) \mid x_1 < 0\}.$$

We define the **Lorentz group** $\mathsf{Lor}(1, n)$ to be the (closed) subgroup of $\mathsf{SO}(1, n)$ preserving each of the connected sets $\mathcal{H}^\pm_{1,n}(1)$. Thus

$$\mathsf{Lor}(1, n) := \left\{ A \in \mathsf{SO}(1, n) \mid A\mathcal{H}^\pm_{1,n}(1) = \mathcal{H}^\pm_{1,n}(1) \right\} \leq \mathsf{SO}(1, n).$$

It turns out that $A \in \mathsf{Lor}(1, n)$ if and only if it preserves the hyperboloids $\mathcal{H}^\pm_{1,n}(\rho)$, $\rho > 0$ and the "light cones" $\mathcal{H}^\pm_{1,n}(0)$.

NOTE : The Lorentz group $\mathsf{Lor}(1, n)$ is *connected*.

Of particular interest in physics is the Lorentz group $\mathsf{Lor} = \mathsf{Lor}(1, 3)$. That is,

$$\mathsf{Lor} = \left\{ L \in \mathsf{SO}(1, 3) \mid L\mathcal{H}^\pm_{1,3}(\rho) = \mathcal{H}^\pm_{1,3}(\rho),\ \rho \geq 0 \right\} \leq \mathsf{SO}(1, 3).$$

◇ **Exercise 190** Show that

(a) The matrix $A = \begin{bmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{bmatrix}$ is in $\mathsf{SO}(1, 1)$.

(b) For every $s, t \in \mathbb{R}$

$$\begin{bmatrix} \cosh s & \sinh s \\ \sinh s & \cosh s \end{bmatrix} \begin{bmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{bmatrix} = \begin{bmatrix} \cosh(s+t) & \sinh(s+t) \\ \sinh(s+t) & \cosh(s+t) \end{bmatrix}.$$

(c) Every element (matrix) of $\mathsf{O}(1,1)$ can be written in one of the four forms

$$\begin{bmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{bmatrix}, \begin{bmatrix} -\cosh t & \sinh t \\ \sinh t & -\cosh t \end{bmatrix}, \begin{bmatrix} \cosh t & -\sinh t \\ \sinh t & -\cosh t \end{bmatrix}, \begin{bmatrix} -\cosh t & -\sinh t \\ \sinh t & \cosh t \end{bmatrix}.$$

(Since $\cosh t$ is always positive, there is no overlap among the four cases. Matrices of the first two forms have determinant one; matrices of the last two forms have determinant minus one.)

NOTE : We can write

$$\begin{aligned} \mathsf{SO}(1,1) &= \mathsf{Lor}(1,1) \cup \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \mathsf{Lor}(1,1) \\ \mathsf{O}(1,1) &= \mathsf{SO}(1,1) \cup \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \mathsf{SO}(1,1). \end{aligned}$$

More on $\mathsf{Lor}(1,2)$. . . . . .

The real symplectic group $\mathsf{Sp}(2n, \mathbb{R})$

Let

$$\mathbb{J} := \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix} \in \mathsf{SL}(2n, \mathbb{R}).$$

A matrix $A \in \mathbb{R}^{2n \times 2n}$ is called *symplectic* if

$$A^\top \mathbb{J} A = \mathbb{J}.$$

NOTE : The word *symplectic* was invented by HERMANN WEYL (1885-1955), who substituted Greek for Latin roots in the word *complex* to obtain a term which would describe a group (related to "line complexes" but which would not be confused with complex numbers).

Let $\mathsf{Sp}(2n,\mathbb{R})$ be the set of all $2n \times 2n$ symplectic matrices. Taking determinants of the condition $A^\top \mathbb{J} A = \mathbb{J}$ gives

$$1 = \det \mathbb{J} = (\det A^\top) \cdot (\det \mathbb{J}) \cdot (\det A) = (\det A)^2.$$

Hence $\det A = \pm 1$, and so $A \in \mathsf{GL}(2n,\mathbb{R})$. Furthermore, if $A, B \in \mathsf{Sp}(2n,\mathbb{R})$, then

$$(AB)^\top \mathbb{J} (AB) = B^\top A^\top \mathbb{J} AB = \mathbb{J}.$$

Hence $AB \in \mathsf{Sp}(2n,\mathbb{R})$. Now, if $A^\top \mathbb{J} A = \mathbb{J}$, then

$$\mathbb{J} A = (A^\top)^{-1} \mathbb{J} = (A^{-1})^\top \mathbb{J}$$

so

$$\mathbb{J} = (A^{-1})^\top \mathbb{J} A^{-1}.$$

It follows that $A^{-1} \in \mathsf{Sp}(2n,\mathbb{R})$ and hence $\mathsf{Sp}(2n,\mathbb{R})$ is a group. In fact, it is a *closed* subgroup of $\mathsf{GL}(2n,\mathbb{R})$, and thus a matrix group.

NOTE : The symplectic group $\mathsf{Sp}(2n,\mathbb{R})$ is *connected*. (It turns out that the determinant of a symplectic matrix must be positive; this fact is by no means obvious.

◇ **Exercise 191** Check that $\mathsf{Sp}(2,\mathbb{R}) = \mathsf{SL}(2,\mathbb{R})$. (In general, it is *not* true that $\mathsf{Sp}(2n,\mathbb{R}) = \mathsf{SL}(2n,\mathbb{R})$.)

◇ **Exercise 192** Given $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathsf{GL}(2n,\mathbb{R})$, show that $A \in \mathsf{Sp}(2n,\mathbb{R})$ if and only if $a^\top c$ and $b^\top d$ are symmetric and $a^\top d - c^\top b = I_n$.

All matrices of the form

$$\begin{bmatrix} I_n & 0 \\ tI_n & I_n \end{bmatrix}$$

are symplectic. However, the (Frobenius) norm of such a matrix is equal to $\sqrt{2n + t^2 n}$, which is *unbounded* if $t \in \mathbb{R}$. Therefore, $\mathsf{Sp}(2n,\mathbb{R})$ is not a bounded subset of $\mathbb{R}^{2n \times 2n}$ and hence is *not* compact.

◇ **Exercise 193** Consider the skew-symmetric bilinear form on (the vector space) $\mathbb{R}^{2n}$ defined by

$$\Omega(x,y) := \sum_{i=1}^{n} (x_i y_{n+i} - x_{n+i} y_i)$$

(the standard *symplectic form* or the "canonical" symplectic structure). Show that a linear transformation (on $\mathbb{R}^{2n}$) $x \mapsto Ax$ preserves the symplectic form Ω if and only if $A^\top \mathbb{J} A = \mathbb{J}$ (i.e., the matrix A is symplectic). Such a structure-preserving transformation is called a *symplectic transformation*.

The group of all symplectic transformations on $\mathbb{R}^{2n}$ (equipped with the symplectic form Ω) is *isomorphic* to (the matrix group) $\mathsf{Sp}(2n, \mathbb{R})$.

NOTE : The symplectic group is related to *classical mechanics*. Consider a particle of mass m moving in a *potential field* V. Newton's second law states that the particle moves along a curve $t \mapsto x(t)$ in in Cartesian 3-space $\mathbb{R}^3$) in such a way that $m\ddot{x} = -\operatorname{grad} V(x)$. Introduce the conjugate *momenta* $p_i = m\dot{x}_i$, $i = 1, 2, 3$ and the *energy* (Hamiltonian)

$$H(x, p) := \frac{1}{2m} \sum_{i=1}^{3} p_i^2 + V(x).$$

Then

$$\frac{\partial H}{\partial x_i} = \frac{\partial V}{\partial x_i} = -m\ddot{x}_i = -\dot{p}_i \quad \text{and} \quad \frac{\partial H}{\partial p_i} = \frac{1}{m} p_i = \dot{x}_i$$

and hence *Newton's law* $\mathbf{F} = m\,a$ is equivalent to *Hamilton's equations*

$$\dot{x}_i = \frac{\partial H}{\partial p_i} \quad \text{and} \quad \dot{p}_i = -\frac{\partial H}{\partial x_i} \qquad (i = 1, 2, 3).$$

Writing $z = (x, p)$,

$$\mathbb{J} \cdot \operatorname{grad} H(z) = \begin{bmatrix} 0 & I_3 \\ -I_3 & 0 \end{bmatrix} \begin{bmatrix} \frac{\partial H}{\partial x} \\ \\ \frac{\partial H}{\partial p} \end{bmatrix} = (\dot{x}, \dot{p}) = \dot{z}$$

so Hamilton equations read $\dot{z} = \mathbb{J} \cdot \operatorname{grad} H(z)$. Now let

$$F : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3 \times \mathbb{R}^3$$

and write $w(t) = F(z(t))$. If $z(t)$ satisfies Hamilton's equations

$$\dot{z} = \mathbb{J} \cdot \operatorname{grad} H(z)$$

then $w(t) = F(z(t))$ satisfies $\dot{w} = A^\top \dot{z}$, where $A^\top = [\partial w^i / \partial z^j]$ is the Jacobian matrix of f. By the chain rule,

$$\dot{w} = A^\top \mathbb{J} \operatorname{grad}_z H(z) = A^\top \mathbb{J} A \operatorname{grad}_w H(z(w)).$$

Thus, the equations for $w(t)$ have the form of Hamilton's equations with energy $K(w) = H(z(w))$ if and only if $A^\top \mathbb{J} A = \mathbb{J}$; that is, if and only if A is symplectic. A nonlinear transformation F is *canonical* if and only if its Jacobian matrix is symplectic (or, if one prefers, its tangent mapping is a symplectic transformation).

As a special case, consider a (linear transformation) $A \in \mathsf{Sp}(2n, \mathbb{R})$ and let $w = Az$. Suppose H is quadratic (i.e., of the form $H(z) = \frac{1}{2} z^\top B z$ where B is a symmetric matrix). Then $\operatorname{grad} H(z) = Bz$ and thus the equations of motion become the linear equations $\dot{z} = \mathbb{J} B z$. Now

$$\dot{w} = A\dot{z} = A\mathbb{J}Bz = \mathbb{J}(A^\top)^{-1}Bz = \mathbb{J}(A^\top)^{-1}BA^{-1}Az = \mathbb{J}B'w$$

where $B' = (A^\top)^{-1}BA^{-1}$ is symmetric. For the new Hamiltonian we get

$$\begin{aligned} H'(w) &= \frac{1}{2} w^\top (A^\top)^{-1} B A^{-1} w = \frac{1}{2} (A^{-1}w)^\top B A^{-1} w \\ &= H(A^{-1}w) = H(z). \end{aligned}$$

Thus $\mathsf{Sp}(2n, \mathbb{R})$ *is the linear invariance group of classical mechanics.*

The complex general linear group $\mathsf{GL}(n, \mathbb{C})$

Many important matrix groups involve *complex* matrices. As in the real case,

$$\mathsf{GL}(n, \mathbb{C}) := \{A \in \mathbb{C}^{n \times n} \,|\, \det A \neq 0\}$$

is an *open* subset of $\mathbb{C}^{n \times n}$, and hence is *not* compact. Clearly $\mathsf{GL}(n, \mathbb{C})$ is a group under matrix multiplication.

NOTE : The general linear group $\mathsf{GL}(n, \mathbb{C})$ is *connected*. This is in contrast with the fact that $\mathsf{GL}(n, \mathbb{R})$ has two components.

The complex special linear group $\mathsf{SL}(n, \mathbb{C})$

This group is defined by

$$\mathsf{SL}(n, \mathbb{C}) := \{A \in \mathsf{GL}(n, \mathbb{C}) \,|\, \det A = 1\}$$

and is treated as in the real case. The matrix group $\mathsf{SL}(n, \mathbb{C})$ is *not* compact but *connected*.

The unitary and special unitary groups $\mathsf{U}(n)$ and $\mathsf{SU}(n)$

For $A = \left[a_{ij}\right] \in \mathbb{C}^{n\times n}$,

$$A^* := \bar{A}^\top = \overline{A^\top}$$

is the *Hermitian conjugate* (i.e., the conjugate transpose) matrix of A; thus, $(A^*)_{ij} = \bar{a}_{ji}$. The **unitary group** is defined as

$$\mathsf{U}(n) := \{A \in \mathsf{GL}(n,\mathbb{C}) \mid A^*A = I_n\}.$$

◇ **Exercise 194** Verify that $\mathsf{U}(n)$ is a *subgroup* of the general linear group $\mathsf{GL}(n,\mathbb{C})$.

The unitary condition amounts to n^2 equations for the n^2 complex numbers a_{ij}, $i,j = 1,2,\ldots,n$

$$\sum_{k=1}^{n} \bar{a}_{ki}a_{kj} = \delta_{ij}.$$

By taking real and imaginary parts, these equations actually give $2n^2$ equations in the $2n^2$ real and imaginary parts of the a_{ij} (although there is some redundancy). This means that $\mathsf{U}(n)$ is a *closed* subset of $\mathbb{C}^{n\times n} = \mathbb{R}^{2n^2}$ and hence of $\mathsf{GL}(n,\mathbb{C})$. Thus $\mathsf{U}(n)$ is a complex matrix group.

NOTE : The unitary group $U(n)$ is *compact* and *connected*.

Let $A \in \mathsf{U}(n)$. From $|\det A| = 1$, we see that the determinant function $\det : \mathsf{GL}(n,\mathbb{C}) \to \mathbb{C}$ maps $\mathsf{U}(n)$ onto the unit circle $\mathbb{S}^1 = \{z \in \mathbb{C} \mid |z| = 1\}$.

NOTE : In the special case $n = 1$, a complex linear mapping $\phi : \mathbb{C} \to \mathbb{C}$ is multiplication by some complex number z, and ϕ is an *isometry* if and only if $|z| = 1$. In this way, the unitary group $\mathsf{U}(1)$ is *identified* with the unit circle $\mathbb{S}^1$. The group $\mathsf{U}(1)$ is more commonly known as the *circle group* or the 1-dimensional *torus*, and is also denoted by $\mathbb{T}^1$.

The dot product on $\mathbb{R}^n$ can be extended to $\mathbb{C}^n$ by setting (for $x, y \in \mathbb{C}^{n\times 1}$)

$$x \bullet y := x^*y = \bar{x}_1y_1 + \bar{x}_2y_2 + \cdots + \bar{x}_ny_n.$$

NOTE : This is *not* $\mathbb{C}$-linear but satisfies (for $x, y \in \mathbb{C}^{n\times 1}$ and $u, v \in \mathbb{C}$)

$$(ux) \bullet (vy) = \bar{u}v(x \bullet y).$$

This dot product allows us to define the *length* (or norm) of a complex vector $x \in \mathbb{C}^{n\times 1}$ by

$$\|x\| := \sqrt{x \bullet x}.$$

Then a matrix $A \in \mathbb{C}^{n\times n}$ is *unitary* if and only if (for $x, y \in \mathbb{C}^n$)

$$Ax \bullet Ay = x \bullet y.$$

$\diamond$ **Exercise 195** If $G_i \leq \mathsf{GL}(n_i, \Bbbk)$, $i = 1, 2$ are matrix groups, show that their (direct) *product* $G_1 \times G_2$ is also a matrix group (in $\mathsf{GL}(n_1 + n_2, \Bbbk)$). Observe, in particular, that the k-dimensional *torus*

$$\mathbb{T}^k := \mathbb{T}^1 \times \mathbb{T}^1 \times \cdots \times \mathbb{T}^1$$

is a matrix group (in $\mathsf{GL}(k, \mathbb{C})$). These groups are compact connected Abelian matrix groups. In fact, they are the *only* matrix groups with these properties.

The **special unitary group**

$$\mathsf{SU}(n) := \{A \in \mathsf{U}(n) \,|\, \det A = 1\}$$

is a closed subgroup of $\mathsf{U}(n)$ and hence a complex matrix group.

NOTE : The matrix group $\mathsf{SU}(n)$ is *compact* and *connected*. In the special case $n = 2$, $\mathsf{SU}(2)$ is *diffeomorphic* to the unit sphere $\mathbb{S}^3$ in $\mathbb{C}^2$ (or $\mathbb{R}^4$). The group $\mathsf{SU}(2)$ is used in the construction of the gauge group for the Young-Mills equations in *elementary particle physics*. Also, there is a 2 to 1 surjection (in fact, a surjective *submersion*)

$$\pi : \mathsf{SU}(2) \to \mathsf{SO}(3)$$

which is of crucial importance in *computational mechanics* (it is related to the quaternionic representation of rotations in Euclidean 3-space).

The complex orthogonal groups $\mathsf{O}(n, \mathbb{C})$ and $\mathsf{SO}(n, \mathbb{C})$

Consider the bilinear form on (the vector space) $\mathbb{C}^n$ defined by (for $x, y \in \mathbb{C}^n$)

$$(x, y) := x_1y_1 + x_2y_2 + \cdots + x_ny_n.$$

This form is *not* an inner product because of the lack of complex conjugation in the definition. The set of all complex $n \times n$ matrices which preserve this

form (i.e., such that $(Ax, Ay) = (x, y)$ for all $x, y \in \mathbb{C}^n$) is the **complex orthogonal group** $\mathsf{O}(n, \mathbb{C})$. Thus

$$\mathsf{O}(n, \mathbb{C}) := \left\{ A \in \mathsf{GL}(n, \mathbb{C}) \,|\, A^\top A = I_n \right\} \subseteq \mathsf{GL}(n, \mathbb{C}).$$

It is easy to show that O *(n,*$\mathbb{C}$*) is a matrix group*, and that $\det A = \pm 1$ for all $\mathsf{O}(n, \mathbb{C})$.

NOTE : The matrix group $\mathsf{O}(n, \mathbb{C})$ is *not* the same as the unitary group $\mathsf{U}(n)$.

The **complex special orthogonal group**

$$\mathsf{SO}(n, \mathbb{C}) := \{A \in \mathsf{O}(n, \mathbb{C}) \,|\, \det A = 1\}$$

is also a matrix group.

The unipotent group $\mathsf{UT}^u(n, \Bbbk)$

A matrix $A = \left[a_{ij}\right] \in \Bbbk^{n \times n}$ is *upper triangular* if all the entries bellow the main diagonal are equal to 0. Let $\mathsf{UT}(n, \Bbbk)$ denote the set of all $n \times n$ *invertible* upper triangular matrices (over $\Bbbk$). Thus

$$\mathsf{UT}(n, \Bbbk) := \{A \in \mathsf{GL}(n, \Bbbk) \,|\, a_{ij} = 0 \text{ for } i > j\}.$$

◇ **Exercise 196** Show that $\mathsf{UT}(n, \Bbbk)$ is a *closed* subgroup of the general linear group $\mathsf{GL}(n, \Bbbk)$ (and hence a matrix group).

The group $\mathsf{UT}(n, \Bbbk)$ is called the (real or complex) **upper triangular group**. This group is *not* compact.

NOTE : Likewise, one can define the *lower triangular group*

$$\mathsf{LT}(n, \Bbbk) := \{A \in \mathsf{GL}(n, \Bbbk) \,|\, a_{ij} = 0 \text{ for } i < j\}.$$

Clearly, $A \in \mathsf{LT}(n, \Bbbk)$ if and only if $A^\top \in \mathsf{UT}(n, \Bbbk)$. The matrix groups $\mathsf{UT}(n, \Bbbk)$ and $\mathsf{LT}(n, \Bbbk)$ are *isomorphic* and there is no need to distinguish between them.

◇ **Exercise 197** Show that the **diagonal group**

$$\mathsf{D}(n, \Bbbk) := \{A \in \mathsf{GL}(n, \Bbbk) \,|\, a_{ij} = 0 \text{ for } i \neq j\}$$

is a closed subgroup of $\mathsf{UT}(n, \Bbbk)$ (and hence a *matrix group*).

◇ **Exercise 198** For $k \leq n$, let $\mathsf{P}(k)$ denote the group of all linear transformations (i.e., invertible linear mappings) on $\mathbb{R}^n$ that preserve the subspace $\mathbb{R}^k = \mathbb{R}^k \times \{0\} \subseteq \mathbb{R}^n$. Show that $\mathsf{P}(k)$ is (*isomorphic* to) the matrix group

$$\left\{ \begin{bmatrix} A & X \\ 0 & B \end{bmatrix} \mid A \in \mathsf{GL}(k, \mathbb{R}),\ B \in \mathsf{GL}(n-k, \mathbb{R}),\ X \in \mathbb{R}^{k \times (n-k)} \right\}.$$

An upper triangular matrix $A = \left[a_{ij}\right]$ is *unipotent* if it has all diagonal entries equal to 1. The (real or complex) **unipotent group** is (the subgroup of $\mathsf{GL}(n, \Bbbk)$)

$$\mathsf{UT}^u(n, \Bbbk) := \{A \in \mathsf{GL}(n, \Bbbk) \mid a_{ij} = 0 \ \text{ for } i > j \ \text{ and } a_{ii} = 1\}$$

(see also **Exercise 194**). It is easy to see that the unipotent group $\mathsf{UT}^u(n, \Bbbk)$ is a closed subgroup of $\mathsf{GL}(n, \Bbbk)$ and hence a *matrix group*.

NOTE : $\mathsf{UT}^u(n, \Bbbk)$ is a closed subgroup of $\mathsf{UT}(n, \Bbbk)$.

For the case

$$\mathsf{UT}^u(2, \Bbbk) = \left\{ \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \in \mathsf{GL}(n, \Bbbk) \mid t \in \Bbbk \right\}$$

the mapping

$$\theta : \Bbbk \to \mathsf{UT}^u(2, \Bbbk), \quad t \mapsto \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}$$

is a *continuous* group homomorphism which is an isomorphism with continuous inverse. This allows us to *view* $\Bbbk$ *as a matrix group*.

NOTE : Given two matrix groups G and H, a group homomorphism $\theta : G \to H$ is a *continuous homomorphism* if it is continuous and its image $\theta(G) \leq H$ is a closed subset of H. For instance,

$$\theta : \mathsf{UT}^u(2, \mathbb{R}) \to \mathsf{U}(1), \quad \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \mapsto e^{2\pi t i}$$

is a continuous homomorphism of matrix groups, but (for $a \in \mathbb{R} \setminus \mathbb{Q}$)

$$\theta' : G = \left\{ \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix} \in \mathsf{SUT}(2, \mathbb{R}) \mid k \in \mathbb{Z} \right\} \to \mathsf{U}(1), \quad \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix} \mapsto e^{2\pi k a i}$$

is *not* (since its image is a *dense* proper subset of $\mathsf{U}(1)$). Whenever we have a continuous homomorphism of matrix groups $\theta : G \to H$ which is a *homeomorphism* (i.e., a continuous bijection with continuous inverse) we say that θ is a *continuous isomorphism* and regard G and H as "identical" (as matrix groups).

The unipotent group $\mathsf{UT}^u(3,\mathbb{R})$ is the **Heisenberg group**

$$\mathsf{Heis} := \left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} \mid a,b,c \in \mathbb{R} \right\}$$

which is particularly important in *quantum physics*; the *Lie algebra* of Heis gives a realization of the *Heisenberg commutation relations* of quantum mechanics.

◇ **Exercise 199** Verify that the 4×4 unipotent matrices A of the form

$$A = \begin{bmatrix} 1 & a_2 & a_3 & a_4 \\ 0 & 1 & a_1 & \frac{a_1^2}{2} \\ 0 & 0 & 1 & a_1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

form a closed subgroup of $\mathsf{UT}^u(4,\mathbb{R})$ (and hence a matrix group). Generalize.

Several other matrix groups are of great interest. We describe briefly some of them.

The general affine group $\mathsf{GA}(n,\Bbbk)$

The **general affine group** (over $\Bbbk$) is the group

$$\mathsf{GA}(n,\Bbbk) := \left\{ \begin{bmatrix} 1 & 0 \\ c & A \end{bmatrix} \in \mathsf{GL}(n+1,\Bbbk) \mid c \in \Bbbk^{n\times 1} \text{ and } A \in \mathsf{GL}(n,\Bbbk) \right\}.$$

This is clearly a closed subgroup of the general linear group $\mathsf{GL}(n+1,\Bbbk)$ (and hence a *matrix group*). The general affine group $\mathsf{GA}(n,\Bbbk)$ is *not* compact. Likewise the case of the general linear group, the matrix group $\mathsf{GA}(n,\mathbb{C})$ is connected but $\mathsf{GA}(n,\mathbb{R})$ is *not*.

NOTE : If we identify the element $x \in \mathbb{k}^n$ with $\begin{bmatrix} 1 \\ x \end{bmatrix} \in \mathbb{k}^{(n+1)\times 1}$, then since

$$\begin{bmatrix} 1 & 0 \\ c & A \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \begin{bmatrix} 1 \\ Ax + c \end{bmatrix}$$

we obtain an *action* of the group $\mathsf{GA}(n, \mathbb{k})$ on (the vector space) $\mathbb{k}^n$. Transformations on $\mathbb{k}^n$ having the form $x \mapsto Ax + c$ (with A invertible) are called *affine transformations* and they preserve *lines* (i.e., translates of 1-dimensional subspaces of the vector space $\mathbb{k}^n$). The associated geometry is *affine geometry* that has $\mathsf{GA}(n, \mathbb{k})$ as its symmetry group.

The (additive group of the) vector space $\mathbb{k}^n$ (in fact, $\mathbb{k}^{n\times 1}$) can be viewed as (and identified with) the *translation subgroup* of $\mathsf{GA}(n, \mathbb{k})$

$$\left\{ \begin{bmatrix} 1 & 0 \\ c & I_n \end{bmatrix} \in \mathsf{GL}(n+1, \mathbb{k}) \,|\, c \in \mathbb{k}^{n\times 1} \right\} \leq \mathsf{GA}(n, \mathbb{k})$$

and this is a closed subgroup.

The identity component of the (real) general affine group $\mathsf{GA}(n, \mathbb{R})$ is (the matrix group)

$$\mathsf{GA}^+(n, \mathbb{R}) = \left\{ \begin{bmatrix} 1 & 0 \\ c & A \end{bmatrix} \,|\, c \in \mathbb{k}^{n\times 1} \text{ and } A \in \mathsf{GL}^+(n, \mathbb{R}) \right\}.$$

In particular,

$$\mathsf{GA}^+(1, \mathbb{R}) = \left\{ \begin{bmatrix} 1 & 0 \\ c & e^a \end{bmatrix} \,|\, a, c \in \mathbb{R} \right\}$$

is a *connected* matrix group (of "dimension" 2). Its elements are (in fact, can be identified with) transformations on (the real line) $\mathbb{R}$ having the form $x \mapsto bx + c$ (with $b, c \in \mathbb{R}$ and $b > 0$).

The Euclidean group $\mathsf{E}(n)$

This is the matrix group

$$\mathsf{E}(n) := \left\{ \begin{bmatrix} 1 & 0 \\ c & A \end{bmatrix} \in \mathsf{GL}(n+1, \mathbb{R}) \,|\, c \in \mathbb{R}^{n\times 1} \text{ and } A \in \mathsf{O}(n) \right\}.$$

The Euclidean group $\mathsf{E}(n)$ is a closed subgroup of the general affine group $\mathsf{GA}(n,\mathbb{R})$ and also is neither compact nor connected. It can be viewed as (and thus identified with) the group of all *isometries* (i.e., *rigid motions*) on the Euclidean n-space $\mathbb{R}^n$.

The special Euclidean group $\mathsf{SE}(n)$

The **special Euclidean group** $\mathsf{SE}(n)$ is (the matrix group) defined by

$$\mathsf{SE}(n) := \left\{ \begin{bmatrix} 1 & 0 \\ c & R \end{bmatrix} \in \mathsf{GL}(n+1,\mathbb{R}) \,|\, c \in \mathbb{R}^{n\times 1} \text{ and } R \in \mathsf{SO}(n) \right\}.$$

This group is *isomorphic* to the group of all *orientation-preserving isometries* (i.e., *proper rigid motions*) on the Euclidean n-space $\mathbb{R}^n$. It is *not* compact but *connected*.

Some other groups

Several important groups which are not naturally groups of matrices can be viewed as matrix groups. We have seen that the multiplicative groups $\mathbb{R}^\times$ and $\mathbb{C}^\times$ (of non-zero real numbers and complex numbers, respectively) are *isomorphic* to the matrix groups $\mathsf{GL}(1,\mathbb{R})$ and $\mathsf{GL}(1,\mathbb{C})$, respectively. Also, the *circle group* $\mathbb{S}^1$ (of complex numbers with absolute value one) is *isomorphic* to $\mathsf{U}(1)$. The *n-torus* (the direct product of n copies of $\mathbb{S}^1$)

$$\mathbb{T}^n = \mathbb{S}^1 \times \cdots \times \mathbb{S}^1 \leq \mathsf{GL}(n,\mathbb{C})$$

is *isomorphic* to the matrix group of $n \times n$ diagonal matrices with complex entries of modulus one. ($\mathbb{T}^n$ can also be realized as the *quotient group* $\mathbb{R}^n/\mathbb{Z}^n$: an element $(\theta_1,\ldots,\theta_n) \mod \mathbb{Z}^n$ of $\mathbb{R}^n/\mathbb{Z}^n$ can be identified with the diagonal matrix $\operatorname{diag}\left(e^{2\pi i\theta_1},\ldots,e^{2\pi i\theta_n}\right)$.)

NOTE : If $\theta : G \to H$ is a continuous homomorphism of matrix groups, then its *kernel* $\ker\theta \leq G$ is a matrix group. Moreover, the *quotient group* $G/\ker\theta$ can be identified with the matrix group $\theta(G)$ by the usual quotient isomorphism $\widetilde{\theta} : G/\ker\theta \to \theta(G)$.

However, it is important to realize that *not every normal matrix subgroup N of the matrix group G gives rise to a matrix group G/N*; there are examples for which

G/N is a *Lie group* but not a matrix group. (We shall see later that every matrix group is a Lie group.)

Recall that the additive groups $\mathbb{R}$ and $\mathbb{C}$ are *isomorphic* to the unipotent groups $\mathsf{UT}^u(2,\mathbb{R})$ and $\mathsf{UT}^u(2,\mathbb{C})$, respectively.

◇ **Exercise 200** Verify that the map

$$x \in \mathbb{R} \mapsto [e^x] \in \mathsf{GL}^+(1,\mathbb{R})$$

is a *continuous isomorphism* of matrix groups, and then show that the additive group $\mathbb{R}^n$ is *isomorphic* to the matrix group of all $n \times n$ diagonal matrices with positive entries.

◇ **Exercise 201** Let $\mathbb{Z}^n \leq \mathbb{R}^n$ be the *discrete* subgroup of vectors with integer entries and set

$$\mathsf{GL}(n,\mathbb{Z}) := \{A \in \mathsf{GL}(n,\mathbb{R}) \,|\, A(\mathbb{Z}^n) = \mathbb{Z}^n\}.$$

Show that $\mathsf{GL}(n,\mathbb{Z})$ is a matrix group. (This matrix group consists of $n \times n$ matrices over (the ring) $\mathbb{Z}$ with determinant ± 1.)

The *symmetric group* S_n of all permutations on n elements may be considered as well as a matrix group. Indeed, we can make S_n to *act* (from the right) on $\Bbbk^n$ by linear transformations :

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \cdot \sigma = \begin{bmatrix} x_{\sigma^{-1}(1)} \\ x_{\sigma^{-1}(2)} \\ \vdots \\ x_{\sigma^{-1}(n)} \end{bmatrix}.$$

Thus (for the standard unit vectors $e_1, e_2, \ldots, e_n$) $e_i \cdot \sigma = e_{\sigma(i)}, \quad i = 1, 2, \ldots, n.$

The matrix $[\sigma]$ of the linear transformation induced by $\sigma \in S_n$ (with respect to the standard basis) has all its entries 0 or 1, with exactly one 1 in each row and column. Such a matrix is usually called a *permutation matrix*.

◇ **Exercise 202** Write down the permutations matrices induces by the elements (permutations) of S_3.

When $\Bbbk = \mathbb{R}$ each of these permutation matrices is orthogonal, while when $\Bbbk = \mathbb{C}$ it is unitary. So, for a given $n \geq 1$, the symmetric group S_n is (isomorphic to) a closed subgroup of $\mathsf{O}(n)$ or $\mathsf{U}(n)$.

NOTE : Any *finite* group is (isomorphic to) a matrix subgroup of some orthogonal group $\mathsf{O}(n)$.

The following table lists some interesting matrix groups, indicates whether or not the group is compact and/or connected, and gives the number of (connected) components.

Group	**Compact ?**	**Connected ?**	**Components**
$\mathsf{GL}(n, \mathbb{C})$	no	yes	one
$\mathsf{SL}(n, \mathbb{C})$	no	yes	one
$\mathsf{GL}(n, \mathbb{R})$	no	no	two
$\mathsf{GL}^+(n, \mathbb{R})$	no	yes	one
$\mathsf{SL}(n, \mathbb{R})$	no	yes	one
$\mathsf{U}(n)$	yes	yes	one
$\mathsf{SU}(n)$	yes	yes	one
$\mathsf{O}(n)$	yes	no	two
$\mathsf{SO}(n)$	yes	yes	one
$\mathsf{O}(1, n)$	no	no	four
$\mathsf{SO}(1, n)$	no	no	two
$\mathsf{Lor}(1, n)$	no	yes	one
$\mathsf{Sp}(2n, \mathbb{R})$	no	yes	one
$\mathsf{UT}^u(n, \Bbbk)$	no	yes	one
$\mathsf{GA}(n, \Bbbk)$	no	no	two
$\mathsf{GA}^+(n, \Bbbk)$	no	yes	one
$\mathsf{E}(n)$	no	no	two
$\mathsf{SE}(n)$	no	yes	one
$\mathbb{R}^n$	no	yes	one
$\mathbb{T}^n$	yes	yes	one

NOTE : There are more interesting matrix groups, e.g., the *quaternionic matrix groups* (in particular, the quaternionic symplectic group $\mathsf{Sp}(n)$), associated with the

division algebra $\mathbb{H}$ of quaternions, as well as the *spinor groups* $\mathsf{Spin}\,(n)$ and the *pinor groups* $\mathsf{Pin}\,(n)$, associated with (real) Clifford algebras.

Complex matrix groups as real matrix groups

Recall that the (complex) vector space $\mathbb{C}$ can be viewed as a *real* 2-dimensional vector space (with basis $\{1, i\}$, for example).

◇ **Exercise 203** Show that the mapping

$$\rho : \mathbb{C} \to \mathbb{R}^{2\times 2}, \quad z = x + iy \mapsto \begin{bmatrix} x & -y \\ y & x \end{bmatrix}$$

is an injective *ring homomorphism* (i.e., a one-to-one mapping such that, for $z, z' \in \mathbb{C}$,

$$\rho(z + z') = \rho(z) + \rho(z') \quad \text{and} \quad \rho(zz') = \rho(z)\rho(z').)$$

We can view $\mathbb{C}$ as a *subring* of $\mathbb{R}^{2\times 2}$. In other words, we can *identify* the complex number $z = x + iy$ with the 2×2 real matrix $\rho(z)$.

NOTE : This can also be expressed as

$$\rho(x + iy) = xI_2 - yJ_2, \quad \text{where} \quad J_2 := \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Also, for $z \in \mathbb{C}$,

$$\rho(\bar{z}) = \rho(z)^T$$

(complex conjugation corresponds to transposition).

More generally, given $Z = \left[z_{rs}\right] \in \mathbb{C}^{n\times n}$ with $z_{rs} = x_{rs} + iy_{rs}$, we can write

$$Z = X + iY,$$

where $X = \left[x_{rs}\right]$, $Y = \left[y_{rs}\right] \in \mathbb{R}^{n\times n}$.

◇ **Exercise 204** Show that the mapping

$$\rho_n : \mathbb{C}^{n\times n} \to \mathbb{R}^{2n\times 2n}, \quad Z = X + iY \mapsto \begin{bmatrix} X & -Y \\ Y & X \end{bmatrix}$$

is an injective *ring homomorphism.*

Hence we can *identify* the complex matrix $Z = X + iY$ with the $2n \times 2n$ real matrix $\rho_n(Z)$. Let

$$\mathbb{J} = \mathbb{J}_{2n} := \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix} \in \mathsf{SL}\,(2n, \mathbb{R}).$$

Then we can write

$$\rho_n(Z) = \rho_n(X + iY) = \begin{bmatrix} X & 0 \\ 0 & X \end{bmatrix} - \begin{bmatrix} Y & 0 \\ 0 & Y \end{bmatrix} \mathbb{J}.$$

◇ **Exercise 205** First verify that

$$\mathbb{J}^2 = -I_{2n} \quad \text{and} \quad \mathbb{J}^\top = -\mathbb{J}$$

and then show that, for $Z \in \mathbb{C}^{n\times n}$,

$$\rho_n(\bar{Z}) = \rho_n(Z)^\top \iff X = X^\top \quad \text{and} \quad Y = Y^\top.$$

We see that $\rho_n(\mathsf{GL}\,(n, \mathbb{C}))$ is a closed subgroup of $\mathsf{GL}\,(2n, \mathbb{R})$, so *any matrix subgroup G of $\mathsf{GL}\,(n, \mathbb{C})$ can be viewed as a matrix subgroup of $\mathsf{GL}\,(2n, \mathbb{R})$ (by identifying it with its image $\rho_n(G)$ under ρ_n).* The following characterizations are sometimes useful :

$$\begin{aligned} \rho_n(\mathbb{C}^{n\times n}) &= \left\{A \in \mathbb{R}^{n\times n} \,|\, A\mathbb{J} = \mathbb{J}A\right\} \\ \rho_n(\mathsf{GL}\,(n, \mathbb{C})) &= \left\{A \in \mathsf{GL}\,(2n, \mathbb{R}) \,|\, A\mathbb{J} = \mathbb{J}A\right\}. \end{aligned}$$

◇ **Exercise 206** Verify the folowing set of equalities :

$$\begin{aligned} \rho_n(\mathsf{U}\,(n)) &= \mathsf{O}\,(n) \cap \rho_n(\mathsf{GL}\,(n, \mathbb{C})) \\ &= \mathsf{O}\,(n) \cap \mathsf{Sp}\,(2n, \mathbb{R}) \\ &= \rho_n(\mathsf{GL}\,(n, \mathbb{C})) \cap \mathsf{Sp}\,(2n, \mathbb{R}). \end{aligned}$$

NOTE : In a slight abuse of notation, the real symplectic group $\mathsf{Sp}\,(2n, \mathbb{R})$ is related to the unitary group $\mathsf{U}\,(n)$ by

$$\mathsf{Sp}\,(2n, \mathbb{R}) \cap \mathsf{O}\,(2n) = \mathsf{U}\,(n).$$

4.3 The Exponential Mapping

Let $A \in \mathbb{k}^{n\times n}$ and consider the *matrix series*

$$\sum_{k\geq 0} \frac{1}{k!}A^k = I_n + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \cdots$$

NOTE : This matrix series is a series in the complete normed vector space (in fact, algebra) $(\mathbb{k}^{n\times n}, \|\cdot\|)$, where $\|\cdot\|$ is the *operator norm* (induced by the Euclidean norm on $\mathbb{k}^n$). In a complete normed vector space, an absolutely convergent series $\sum_{k\geq 0} a_k$ (i.e., such that the series $\sum_{k\geq 0} \|a_k\|$ is convergent) is convergent, and

$$\left\| \sum_{k=0}^{\infty} a_k \right\| \leq \sum_{k=0}^{\infty} \|a_k\|.$$

(The converse is not true.) Also, every *rearrangement* of an absolutely convergent series is absolutely convergent, with same sum. Given two absolutely convergent series $\sum_{k\geq 0} a_k$ and $\sum_{k\geq 0} b_k$ (in a complete normed algebra), their *Cauchy product* $\sum_{k\geq 0} c_k$, where $c_k = \sum_{i+j=k} a_i b_j = a_0 b_k + a_1 b_{k-1} + \cdots + a_k b_0$ is also absolutely convergent, and

$$\sum_{k=0}^{\infty} c_k = \left(\sum_{k=0}^{\infty} a_k \right) \left(\sum_{k=0}^{\infty} b_k \right).$$

◇ **Exercise 207** Show that the matrix series $\sum_{k\geq 0} \frac{1}{k!}A^k$ is *absolutely convergent* (and hence *convergent*).

Let $\sum_{k=0}^{\infty} \frac{1}{k!}A^k$ denote the *sum* of the (absolutely) convergent matrix series $\sum_{k\geq 0} \frac{1}{k!}A^k$. We set

$$e^A = \exp(A) := \sum_{k=0}^{\infty} \frac{1}{k!}A^k.$$

This matrix is called the **matrix exponential** of A. It follows that

$$\|\exp(A)\| \leq \|I_n\| + \|A\| + \frac{1}{2!}\|A\|^2 + \cdots = e^{\|A\|}$$

and also $\|\exp(A) - I_n\| \leq e^{\|A\|} - 1$.

◇ **Exercise 208** Show that (for $\lambda, \mu \in \mathbb{k}$)

$$\exp((\lambda + \mu)A) = \exp(\lambda A)\exp(\mu A).$$

[HINT : These series are absolutely convergent. Think of the Cauchy product.]

It follows that

$$I_n = \exp(O) = \exp((1 + (-1))A) = \exp(A)\exp(-A)$$

and hence $\exp(A)$ is *invertible* with inverse $\exp(-A)$. So $\exp(A) \in \mathsf{GL}(n, \mathbb{k})$.

NOTE : The "group property" $\exp((\lambda+\mu)A) = \exp(\lambda A)\exp(\mu A)$ may be rephrased by saying that, for fixed $A \in \mathbb{k}^{n\times n}$, the mapping $\lambda \mapsto \exp(\lambda A)$ is a (continuous) *homomorphism* from the additive group of scalars $\mathbb{k}$ into the general linear group $\mathsf{GL}(n, \mathbb{k})$.

4.3.1 DEFINITION. The mapping

$$\exp : \mathbb{k}^{n\times n} \to \mathsf{GL}(n, \mathbb{k}), \quad A \mapsto \exp(A)$$

is called the **exponential mapping**.

4.3.2 PROPOSITION. *If $A, B \in \mathbb{k}^{n\times n}$ commute, then*

$$\exp(A + B) = \exp(A)\exp(B).$$

PROOF : We expand the series and perform a sequence of manipulations that

are legitimate since these series are absolutely convergent :

$$\begin{aligned}
\exp(A)\exp(B) &= \left(\sum_{r=0}^{\infty}\frac{1}{r!}A^r\right)\left(\sum_{s=0}^{\infty}\frac{1}{s!}B^s\right) \\
&= \sum_{r,s=0}^{\infty}\frac{1}{r!s!}A^rB^s \\
&= \sum_{k=0}^{\infty}\left(\sum_{r=0}^{k}\frac{1}{r!(k-r)!}A^rB^{k-r}\right) \\
&= \sum_{k=0}^{\infty}\frac{1}{k!}\left(\sum_{r=0}^{k}\binom{k}{r}A^rB^{k-r}\right) \\
&= \sum_{k=0}^{\infty}\frac{1}{k!}(A+B)^k \\
&= \exp(A+B).
\end{aligned}$$

□

NOTE : We have made crucial use of the *commutativity* of A and B in the identity

$$\sum_{r=0}^{k}\binom{k}{r}A^rB^{k-r} = (A+B)^k.$$

In particular, for the (commuting) matrices λA and μA, we reobtain the property $\exp((\lambda+\mu)A) = \exp(\lambda A)\exp(\mu A)$. It is important to realize that, in fact, the following statements are equivalent (for $A, B \in \mathbb{k}^{n\times n}$) :

(i) $AB = BA$.

(ii) $\exp(\lambda A)\exp(\mu B) = \exp(\mu B)\exp(\lambda A)$ for all $\lambda, \mu \in \mathbb{k}$.

(iii) $\exp(\lambda A + \mu B) = \exp(\lambda A)\exp(\mu B)$ for all $\lambda, \mu \in \mathbb{k}$.

◇ **Exercise 209** Compute (for $a, b \in \mathbb{R}$)

$$\exp\left(\begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}\right), \quad \exp\left(\begin{bmatrix} a & -b \\ b & a \end{bmatrix}\right), \quad \exp\left(\begin{bmatrix} a & b \\ b & a \end{bmatrix}\right), \quad \exp\left(\begin{bmatrix} a & b \\ 0 & a \end{bmatrix}\right).$$

NOTE : Every real 2×2 matrix is *conjugate* to exactly one of the following types (with $a, b \in \mathbb{R},\ b \neq 0$) :

- $a\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$ (scalar).
- $a\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} + b\begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix}$ (elliptic).
- $a\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} + b\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$ (hyperbolic).
- $a\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} + b\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}$ (parabolic).

Hermitian and skew-Hermitian matrices. .....

◇ **Exercise 210**

(a) Show that if $A \in \mathbb{R}^{n\times n}$ is *skew-symmetric*, then $\exp(A)$ is orthogonal.

(b) Show that if $A \in \mathbb{C}^{n\times n}$ is *skew-Hermitian*, then $\exp(A)$ is unitary.

◇ **Exercise 211** Let $A \in \Bbbk^{n\times n}$ and $B \in \mathsf{GL}(n, \Bbbk)$. Show that

$$\exp(BAB^{-1}) = B\exp(A)B^{-1}.$$

Deduce that if $B^{-1}AB = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, then

$$\exp(A) = B\operatorname{diag}\left(e^{\lambda_1}, e^{\lambda_2}, \ldots, e^{\lambda_n}\right)B^{-1}.$$

◇ **Exercise 212** A matrix $A \in \Bbbk^{n\times n}$ is **nilpotent** if $A^k = O$ for some $k \geq 1$.

(a) Prove that a nilpotent matrix is singular.

(b) Prove that a *strictly upper triangular* matrix $A = \left[a_{ij}\right]$ (i.e. with $a_{ij} = 0$ whenever $i \geq j$) is nilpotent.

(c) Find two nilpotent matrices whose product is *not* nilpotent.

◇ **Exercise 213** Suppose that $A \in \Bbbk^{n\times n}$ and $\|A\| < 1$.

(a) Show that the matrix series

$$\sum_{k\geq 0} A^k = I_n + A + A^2 + A^3 + \cdots$$

converges (in $\Bbbk^{n\times n}$).

(b) Show that the matrix $I_n - A$ is *invertible* and find a formula for $(I_n - A)^{-1}$.

(c) If A is *nilpotent*, determine $(I_n - A)^{-1}$ and $\exp(A)$.

◇ **Exercise 214** Show (for $\lambda \in \mathbb{R}$)

$$\exp\left(\begin{bmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & \lambda \end{bmatrix}\right) = \begin{bmatrix} e^{\lambda} & e^{\lambda} & \frac{1}{2!}e^{\lambda} & \dots & \frac{1}{(n-1)!}e^{\lambda} \\ 0 & e^{\lambda} & e^{\lambda} & \dots & \frac{1}{(n-2)!}e^{\lambda} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & e^{\lambda} \end{bmatrix}.$$

NOTE : When the matrix $A \in \Bbbk^{n\times n}$ is *diagonalizable* over $\mathbb{C}$ (i.e., $A = C\,\mathrm{diag}\,(\lambda_1, \lambda_2, \dots, \lambda_n)\,C^{-1}$ for some $C \in \mathsf{GL}\,(n, \mathbb{C})$), we have

$$\exp(A) = C\,\mathrm{diag}\,\left(e^{\lambda_1}, e^{\lambda_2}, \dots, e^{\lambda_n}\right)\,C^{-1}.$$

This means that the problem of calculating the exponential of a diagonalizable matrix is solved once an explicit diagonalization is found. Many important types of matrices are indeed diagonalizable (over $\mathbb{C}$), including skew-symmetric, skew-Hermitian, orthogonal, and unitary matrices. However, there are also many non-diagonalizable matrices. If $A^k = O$ for some positive integer k, then $A^{\ell} = O$ for all $\ell \geq k$. In this case the matrix series which defines $\exp(A)$ terminates after the first k terms, and so can be computed explicitly. A general matrix A may be neither nilpotent nor diagonalizable. This situation is best discussed in terms of the *Jordan canonical form*.

For $\lambda \in \mathbb{C}$ and $r \geq 1$, we have the *Jordan block matrix*

$$J(\lambda, r) := \begin{bmatrix} \lambda & 1 & 0 & \dots & 0 & 0 \\ 0 & \lambda & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda & 1 \\ 0 & 0 & 0 & \dots & 0 & \lambda \end{bmatrix} \in \mathbb{C}^{r\times r}.$$

The characteristic polynomial of $J(\lambda, r)$ is

$$\mathrm{char}_{J(\lambda,r)}(s) := \det\,(sI_r - J(\lambda, r)) = (s - \lambda)^r$$

and by the *Cayley-Hamilton Theorem*, $(J(\lambda, r) - \lambda I_r)^r = O$, which implies that $(J(\lambda, r) - \lambda I_r)^{r-1} \neq O$ (and hence $\mathrm{char}_{J(\lambda,r)}(s) = \min_{J(\lambda,r)}(s) \in \mathbb{C}[s]$). The main result on Jordan form is the following : *Given $A \in \mathbb{C}^{n\times n}$, there exists a matrix*

$P \in \mathsf{GL}(n, \mathbb{C})$ *such that*

$$P^{-1}AP = \begin{bmatrix} J(\lambda_1, r_1) & O & \dots & O \\ O & J(\lambda_2, r_2) & \dots & O \\ \vdots & \vdots & & \vdots \\ O & O & \dots & J(\lambda_m, r_m) \end{bmatrix} \in \mathbb{C}^{n \times n}.$$

This form is unique except for the order in which the Jordan blocks $J(\lambda_i, r_i) \in \mathbb{C}^{r_i \times r_i}$ *occur.* (The elements $\lambda_1, \lambda_2, \dots, \lambda_m$ are the eigenvalues of A and in fact $\operatorname{char}_A(s) = (s - \lambda_1)^{r_1}(s - \lambda_2)^{r_2} \cdots (s - \lambda_m)^{r_m}$.)

Using the Jordan canonical form we can see that every matrix $A \in \mathbb{C}^{n \times n}$ can be written as $A = S + N$, where S is diagonalizable (over $\mathbb{C}$), N is nilpotent, and $SN = NS$.

◇ **Exercise 215** Let $A \in \Bbbk^{n \times n}$.

(a) Prove that A is *nilpotent* if and only if all its eigenvalues are equal to zero.

(b) The matrix A is called **unipotent** if $I_n - A$ is nilpotent (i.e. $(I_n - A)^k = O$ for some $k \geq 1$). Prove that A is *unipotent* if and only if all its eigenvalues are equal to 1.

(c) If A is a strictly upper triangular matrix, show that $\exp(A)$ is unipotent.

◇ **Exercise 216** Compute

$$\exp\left(\begin{bmatrix} \lambda & a & b \\ 0 & \lambda & c \\ 0 & 0 & \lambda \end{bmatrix}\right).$$

◇ **Exercise 217** The power series

$$\sum_{k \geq 1} (-1)^{k+1} \frac{(z-1)^k}{k} = z - 1 - \frac{(z-1)^2}{2} + \frac{(z-1)^3}{3} - \frac{(z-1)^4}{4} + \cdots, \quad z \in \mathbb{C}$$

has *radius of convergence* 1 and hence defines a complex analytic function

$$z \mapsto \log z := \sum_{k=1}^{\infty} (-1)^{k+1} \frac{(z-1)^k}{k}$$

on the set $\{z \,|\, |z - 1| < 1\}$. (This function coincides with the usual logarithm for real z on the interval $(0, 2)$.) Show that

(a) For all z with $|z-1|<1$,

$$e^{\log z} = z.$$

(b) For all w with $|w|<\ln 2$, $|e^w-1|<1$ and

$$\log\left(e^w\right) = w.$$

Let $A \in \mathbb{k}^{n\times n}$. The matrix series

$$\sum_{k\geq 1} \frac{(-1)^{k+1}}{k} A^k = A - \frac{1}{2}A + \frac{1}{3}A^3 - \frac{1}{4}A^4 + \cdots$$

converges (absolutely) for $\|A\|<1$. We define the **logarithm mapping**

$$\log : \mathcal{B}_{\mathbb{k}^{n\times n}}(I_n,1) \to \mathbb{k}^{n\times n}, \quad A \mapsto \log(A) := \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}(A-I_n)^k.$$

(The notation $\mathcal{B}_{\mathbb{k}^{n\times n}}(A,\rho)$ stands for the *open ball* of radius ρ around A in the metric space $\mathbb{k}^{n\times n}$; that is,

$$\mathcal{B}_{\mathbb{k}^{n\times n}}(A,\rho) := \left\{A' \in \mathbb{k}^{n\times n} \mid \|A'-A\|<\rho\right\}.)$$

NOTE : Defining a logarithm for matrices turns out to be at least as difficult as defining a logarithm for complex numbers, and so we cannot hope to define the matrix logarithm for all matrices, or even for all invertible matrices. We content ourselves with defining the logarithm in a *neighborhood* of the identity matrix. The logarithm mapping is *continuous* (on the set of all $n\times n$ matrices A with $\|A-I_n\|<1$) and $\log(A)$ is real if A is real.

◇ **Exercise 218** Show that

(a) For all A with $\|A-I_n\|<1$,

$$\exp\left(\log(A)\right) = A.$$

(b) For all B with $\|B\|<\ln 2$, $\|\exp(B)-I_n\|<1$ and

$$\log\left(\exp(B)\right) = B.$$

The exponential and logarithm mappings

$$\exp : \mathbb{k}^{n\times n} \to \mathsf{GL}\,(n,\mathbb{k}), \quad \text{and} \quad \log : \mathcal{B}_{\mathbb{k}^{n\times n}}(I_n,1) \to \mathbb{k}^{n\times n}$$

are *continuous* (in fact, infinitely differentiable). Indeed, since any power A^k is a continuous mapping of A, the sequence of partial sums $\left(\sum_{k=0}^{r} \frac{1}{k!}A^k\right)_{r\geq 0}$ consists of continuous mappings. But (it is easy to see that) the matrix series (defining the exponential matrix) converges *uniformly* on each set of the form $\{A \,|\, \|A\| \leq \rho\}$, and so the sum (i.e., the limit of its sequence of partial sums) is again continuous. (A similar argument works in the case of the logarithm mapping.)

By continuity (of the exponential mapping at the origin O), there is a number $\delta > 0$ such that

$$\mathcal{B}_{\mathbb{k}^{n\times n}}(O,\delta) \subseteq \exp^{-1}\left(\mathcal{B}_{\mathsf{GL}\,(n,\mathbb{k})}(I_n,1)\right).$$

In fact we can actually take $\delta = \ln 2$ since

$$\exp\left(\mathcal{B}_{\mathbb{k}^{n\times n}}(O,\delta)\right) \subseteq \mathcal{B}_{\mathbb{k}^{n\times n}}\left(I_n, e^{\delta}-1\right).$$

Hence we have the following result

4.3.3 Proposition. *The exponential mapping* exp *is injective when restricted to the open subset* $\mathcal{B}_{\mathbb{k}^{n\times n}}(O,\ln 2)$. *(Hence it is locally a diffeomorphism at the origin* O *with local inverse* log.*)*

Let $A \in \mathbb{k}^{n\times n}$. For every $t \in \mathbb{R}$, the matrix series $\displaystyle\sum_{k\geq 0} \frac{t^k}{k!}A^k$ is (absolutely) *convergent* and we have

$$\sum_{k=0}^{\infty} \frac{t^k}{k!}A^k = \sum_{k=0}^{\infty} \frac{1}{k!}(tA)^k = \exp(tA).$$

So the mapping

$$\alpha : \mathbb{R} \to \mathbb{k}^{n\times n}, \quad t \mapsto \exp(tA)$$

is defined and *differentiable* with

$$\dot{\alpha}(t) = \sum_{k=1}^{\infty} \frac{t^{k-1}}{(k-1)!}A^k = \exp(tA)A = A\exp(tA).$$

NOTE : This mapping can be viewed as a *curve* in $\mathbb{k}^{n\times n}$. The curve is in fact *smooth* (i.e., infinitely differentiable) and satisfies the differential equation (in matrices) $\dot{\alpha}(t) = \alpha(t)A$ with initial condition $\alpha(0) = I_n$. Also (for $t, s \in \mathbb{R}$),

$$\alpha(t+s) = \alpha(t)\alpha(s).$$

In particular, this shows that $\alpha(t)$ is always invertible with $\alpha(t)^{-1} = \alpha(-t)$.

◇ **Exercise 219** Let $A, C \in \mathbb{k}^{n\times n}$. Show that the differential equation (in matrices) $\dot{\alpha} = \alpha A$ has a *unique* differentiable solution $\alpha : \mathbb{R} \to \mathbb{k}^{n\times n}$ for which $\alpha(0) = C$. (This solution is $\alpha(t) = C \exp(tA)$.) Furthermore, if C is invertible, then so is $\alpha(t)$ for $t \in \mathbb{R}$, hence $\alpha : \mathbb{R} \to \mathsf{GL}(n, \mathbb{k})$.

◇ **Exercise 220** Let $A \in \mathbb{k}^{n\times n}$. Show that the functional equation (in matrices) $\alpha(t+s) = \alpha(t)\alpha(s)$ has a *unique* differentiable solution $\alpha : \mathbb{R} \to \mathbb{k}^{n\times n}$ for which $\alpha(0) = I_n$ and $\dot{\alpha}(0) = A$. (This solution is $\alpha(t) = \exp(tA)$.)

◇ **Exercise 221** If $A, B \in \mathbb{k}^{n\times n}$ commute, show that

$$\frac{d}{dt}\exp(A+tB)\bigg|_{t=0} = \exp(A)B = B\exp(A).$$

(This is a formula for the *derivative* of the exponential mapping exp at an arbitrary A, evaluated only at those B such that $AB = BA$. The general situation is more complicated.)

4.4 Lie Algebras for Matrix Groups

One-parameter subgroups

Let $G \leq \mathsf{GL}(n, \mathbb{k})$ be a matrix group and let I denote the identity matrix.

NOTE : The matrix I is the *neutral element* of the group G. When $\mathbb{k} = \mathbb{R}$, then $I = I_n$ whereas when $\mathbb{k} = \mathbb{C}$ and $G \leq \mathsf{GL}(2n, \mathbb{R})$, then $I = I_{2n} = \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix}$.

4.4.1 DEFINITION. A **one-parameter subgroup** of G is a continuous mapping

$$\gamma : \mathbb{R} \to G$$

which is *differentiable* at 0 and satisfies (for $t, s \in \mathbb{R}$)

$$\gamma(s+t) = \gamma(s)\gamma(t).$$

We refer to the last condition as the *homomorphism property.*

NOTE : Recall that $\mathbb{R}$ and G can be viewed as matrix groups (isomorphic to $\mathsf{UT}^u(2,\mathbb{R})$ and to a subgroup of either $\mathsf{GL}(n,\mathbb{R})$ or $\mathsf{GL}(2n,\mathbb{R})$, respectively). Hence, γ is a *continuous homomorphism* of matrix groups.

It suffices to know γ on some *open neighborhood* $(-\varepsilon,\varepsilon)$ of 0 in $\mathbb{R}$. Indeed, let $t \in \mathbb{R}$. Then for some (large enough) natural number m, $\frac{t}{m} \in (-\varepsilon,\varepsilon)$. Hence

$$\gamma\left(\frac{t}{m}\right),\ \left(\gamma\left(\frac{t}{m}\right)\right)^m \in G.$$

◇ **Exercise 222** Show that (for $m, n \in \mathbb{N}$ such that $\frac{t}{m}, \frac{t}{n} \in (-\varepsilon,\varepsilon)$)

$$\left(\gamma\left(\frac{t}{n}\right)\right)^n = \left(\gamma\left(\frac{t}{m}\right)\right)^m.$$

The element $\left(\gamma\left(\frac{t}{m}\right)\right)^m \in G$ is well defined (for every $t \in \mathbb{R}$), and so

$$\gamma(t) = \gamma\left(\frac{t}{m} + \frac{t}{m} + \cdots + \frac{t}{m}\right) = \left(\gamma\left(\frac{t}{m}\right)\right)^m.$$

NOTE : A one-parameter subgroup $\gamma : \mathbb{R} \to G$ can be viewed as a *collection* $(\gamma(t))_{t\in\mathbb{R}}$ of linear transformations on $\Bbbk^n$ such that (for $t, s \in \mathbb{R}$)

- $\gamma(0) = id_{\Bbbk^n}\ (= I)$.
- $\gamma(s+t) = \gamma(s)\gamma(t)$.
- $\gamma(t) \in G$ depends continuously on t.

Moreover, the *curve* $\gamma : \mathbb{R} \to G$ in $G \subseteq \Bbbk^{n\times n}$ has a *tangent vector* $\dot\gamma(0)$ (at $\gamma(0) = I$).

4.4.2 PROPOSITION. *Let $\gamma : \mathbb{R} \to G$ be a one-parameter subgroup of G. Then γ is differentiable at every $t \in \mathbb{R}$ and*

$$\dot\gamma(t) = \dot\gamma(0)\gamma(t) = \gamma(t)\dot\gamma(0).$$

PROOF : We have (for $t, h \in \mathbb{R}$)

$$\begin{aligned}\dot{\gamma}(t) &= \lim_{h\to 0} \frac{1}{h}\left(\gamma(t+h) - \gamma(t)\right) \\ &= \lim_{h\to 0} \frac{1}{h}\left(\gamma(h)\gamma(t) - \gamma(t)\right) \\ &= \left(\lim_{h\to 0} \frac{1}{h}\left(\gamma(h) - I\right)\right)\gamma(t) \\ &= \dot{\gamma}(0)\gamma(t)\end{aligned}$$

and similarly

$$\dot{\gamma}(t) = \gamma(t)\dot{\gamma}(0).$$

□

We can now determine the form of all one-parameter subgroups of G.

4.4.3 THEOREM. *Let $\gamma : \mathbb{R} \to G$ be a one-parameter subgroup of G. Then it has the form*

$$\gamma(t) = \exp(tA)$$

for some $A \in \Bbbk^{n\times n}$.

PROOF : Let $A = \dot{\gamma}(0)$. This means that γ satisfies (the differential equation)

$$\dot{\gamma}(t) = A\gamma(t)$$

and is subject to (the initial condition)

$$\gamma(0) = I.$$

This initial value problem (IVP) has the unique solution $\gamma(t) = \exp(tA)$. □

We cannot yet reverse this process and decide for which $A \in \Bbbk^{n\times n}$ the one-parameter subgroup

$$\gamma : \mathbb{R} \to \mathsf{GL}\,(n, \Bbbk), \quad t \mapsto \exp(tA)$$

actually takes values in G. The answer involves the *Lie algebra* of G.

NOTE : We have a curious phenomenon in the fact that although the definition of a one-parameter group only involves first order differentiability, the general form

$\exp(tA)$ is always infinitely differentiable (and indeed *analytic*) as a function of t. This is an important characteristic of much of the *Lie theory*, namely that conditions of first order diferentiability (and even continuity) often lead to much stronger conditions.

Lie algebras

Let $G \leq \mathsf{GL}(n, \mathbb{k})$ be a matrix group. Recall that $\mathbb{k}^{n\times n}$ may be considered to be some Euclidean space $\mathbb{R}^m$.

4.4.4 Definition. A **curve** in G is a *differentiable* mapping

$$\gamma : (a, b) \subseteq \mathbb{R} \to \mathbb{k}^{n\times n}$$

such that (for $t \in (a, b)$)

$$\gamma(t) \in G.$$

The derivative

$$\dot{\gamma}(t) := \lim_{h\to 0} \frac{1}{h} (\gamma(t + h) - \gamma(t)) \in \mathbb{k}^{n\times n}$$

is called the **tangent vector** to γ at $\gamma(t)$. We will usually assume that $a < 0 < b$.

◇ **Exercise 223** Given two curves $\gamma, \sigma : (a, b) \to G$, we define a new curve, the *product curve*, by

$$(\gamma\sigma)(t) := \gamma(t)\sigma(t).$$

Show that (for $t \in (a, b)$)

$$(\gamma\sigma)^{\cdot}(t) = \gamma(t)\dot{\sigma}(t) + \dot{\gamma}(t)\sigma(t).$$

◇ **Exercise 224**

(a) Let $\gamma : (-1, 1) \to \mathbb{R}^{3\times 3}$ be given by

$$\gamma(t) := \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos t & \sin t \\ 0 & -\sin t & \cos t \end{bmatrix}.$$

Show that γ is a curve in $\mathsf{SO}(3)$ and find $\dot{\gamma}(0)$. Show that

$$(\gamma^2)^{\cdot}(0) = 2\dot{\gamma}(0).$$

(b) Let $\sigma : (-1, 1) \to \mathbb{R}^{3\times 3}$ be given by

$$\sigma(t) := \begin{bmatrix} 0 & 0 & 0 \\ 0 & \cos t & \sin t \\ 0 & -\sin t & \cos t \end{bmatrix}.$$

Calculate $\dot{\sigma}(0)$. Write the matrix $\gamma(t)\sigma(t)$ and verify that

$$(\gamma\sigma)^{\cdot}(0) = \dot{\gamma}(0) + \dot{\sigma}(0).$$

◇ **Exercise 225** Let $\alpha : (-1, 1) \to \mathbb{C}^{n\times n}$ be given by

$$\alpha(t) := \begin{bmatrix} e^{i\pi t} & 0 & \\ 0 & e^{i\frac{\pi t}{2}} & 0 \\ 0 & 0 & e^{i\frac{\pi t}{2}} \end{bmatrix}.$$

Show that α is a curve in $\mathsf{U}(3)$. Calculate $\dot{\alpha}(0)$.

4.4.5 Definition. The **tangent space** to (the matrix group) G at $A \in G$ is the set

$$T_AG := \{\dot{\gamma}(0) \in \mathbb{k}^{n\times n} \,|\, \gamma \text{ is a curve in } G \text{ with } \gamma(0) = A\}.$$

4.4.6 Proposition. *The set T_AG is a real vector subspace of $\mathbb{k}^{n\times n}$.*

Proof : Let $\alpha, \beta : (a, b) \to \mathbb{k}^{n\times n}$ be two curves in G through A (i.e., $\alpha(0) = \beta(0) = A$). Then

$$\gamma : (a, b) \to \mathbb{k}^{n\times n}, \quad t \mapsto \alpha(t)A^{-1}\beta(t)$$

is also a curve in G with $\gamma(0) = A$. We have

$$\dot{\gamma}(t) = \dot{\alpha}(t)A^{-1}\beta(t) + \alpha(t)A^{-1}\dot{\beta}(t)$$

and hence

$$\dot{\gamma}(0) = \dot{\alpha}(0)A^{-1}\beta(0) + \alpha(0)A^{-1}\dot{\beta}(0) = \dot{\alpha}(0) + \dot{\beta}(0)$$

which shows that T_AG is closed under (vector) addition.

Similarly, if $\lambda \in \mathbb{R}$ and $\alpha : (a, b) \to \mathbb{k}^{n\times n}$ is a curve in G with $\alpha(0) = A$, then

$$\eta : (a, b) \to \mathbb{k}^{n\times n}, \quad t \mapsto \alpha(\lambda t)$$

is another such curve. Since

$$\dot{\eta}(0) = \lambda\dot{\alpha}(0)$$

we see that T_AG is closed under (real) scalar multiplication. So T_AG is a (real) vector subspace of $\mathbb{k}^{n\times n}$. □

NOTE : Since the vector space $\mathbb{k}^{n\times n}$ is *finite dimensional*, so is (the tangent space) T_AG.

4.4.7 DEFINITION. If $G \leq GL(n,\mathbb{k})$ is a matrix group, its **dimension** is the dimension of the (real) vector space T_IG (I is the identity matrix). So

$$\dim G := \dim_{\mathbb{R}} T_IG.$$

NOTE : If the matrix group G is complex, then its *complex dimension* is

$$\dim_{\mathbb{C}} G := \dim_{\mathbb{C}} T_IG.$$

◇ **Exercise 226** Show that the matrix group $\mathsf{U}(1)$ has dimension 1.

NOTE : The *only* connected matrix groups (up to isomorphism) of dimension 1 are $\mathbb{T}^1 = \mathsf{U}(1)$ and $\mathbb{R}$, and of dimension 2 are $\mathbb{R}^2, \mathbb{T}^1\times\mathbb{R}, \mathbb{T}^2$, and $\mathsf{GA}^+(1,\mathbb{R})$.

4.4.8 EXAMPLE. The real general linear group $\mathsf{GL}(n,\mathbb{R})$ has dimension n^2.

The determinant function $\det : \mathbb{R}^{n\times n} \to \mathbb{R}$ is *continuous* and $\det(I) = 1$. So there is some ϵ-*ball* about I in $\mathbb{R}^{n\times n}$ such that for each A in this ball $\det(A) \neq 0$ (i.e., $A \in \mathsf{GL}(n,\mathbb{R})$). If $B \in \mathbb{R}^{n\times n}$, then define a curve σ in $\mathbb{R}^{n\times n}$ by

$$\sigma(t) := tB + I.$$

Then $\sigma(0) = I$ and $\dot{\sigma}(0) = B$, and (for *small* t) $\sigma(t) \in \mathsf{GL}(n,\mathbb{R})$. Hence the tangent space $T_I\mathsf{GL}(n,\mathbb{R})$ is all of $\mathbb{R}^{n\times n}$ which has dimension n^2. So

$$\dim \mathsf{GL}(n,\mathbb{R}) = n^2.$$

◇ **Exercise 227** Show that the dimension of the *complex* general linear group $\mathsf{GL}(n,\mathbb{C})$ is $2n^2$.

4.4.9 PROPOSITION. *Let* $\mathsf{Sk\text{-}sym}\,(n)$ *denote the set of all skew-symmetric matrices in* $\mathbb{R}^{n\times n}$. *Then* $\mathsf{Sk\text{-}sym}\,(n)$ *is a linear subspace of* $\mathbb{R}^{n\times n}$ *and its dimension is* $\frac{n(n-1)}{2}$.

PROOF : If $A, B \in \mathsf{Sk\text{-}sym}\,(n)$, then

$$(A+B)^\top + (A+B) = A^\top + A + B^\top + B = 0$$

so that $\mathsf{Sk\text{-}sym}\,(n)$ is closed under (vector) addition.

It is also closed under scalar multiplication, for if $A \in \mathsf{Sk\text{-}sym}\,(n)$ and $\lambda \in \mathbb{R}$, then $(\lambda A)^\top = \lambda A^\top$ so that

$$(\lambda A)^\top + \lambda A = \lambda(A^\top + A) = 0.$$

To check the dimension of $\mathsf{Sk\text{-}sym}\,(n)$ we construct a basis. Let E_{ij} denote the matrix whose entries are all zero except the ij-entry, which is 1, and the ji-entry, which is -1. If we define these E_{ij} *only* for $i < j$, we can see that they form a *basis* for $\mathsf{Sk\text{-}sym}\,(n)$.

It is easy to compute that there are

$$(n-1) + (n-2) + \cdots + 2 + 1 = \frac{n(n-1)}{2}$$

of them. □

◇ **Exercise 228** Show that if σ is a curve through the identity (i.e., $\sigma(0) = I$) in the orthogonal group $\mathsf{O}\,(n)$, then $\dot\sigma(0)$ is skew-symmetric.

NOTE : It follows that $\dim \mathsf{O}\,(n) \leq \frac{n(n-1)}{2}$. Later we will show that this evaluation is an equality.

◇ **Exercise 229** A matrix $A \in \mathbb{C}^{n\times n}$ is called *skew-Hermitian* if $A^* + A = 0$.

(a) Show that the diagonal terms of a skew-Hermitian matrix are purely imaginary and hence deduce that the set $\mathsf{Sk\text{-}Herm}\,(n)$ of all skew-Hermitian matrices in $\mathbb{C}^{n\times n}$ is *not* a vector space over $\mathbb{C}$.

(b) Prove that $\mathsf{Sk\text{-}Herm}\,(n)$ is a *real* vector space of dimension

$$n + 2\,\frac{n(n-1)}{2} = n^2.$$

(c) If σ is a curve through the identity in $\mathsf{U}(n)$, show that $\dot{\sigma}(0)$ is skew-Hermitian an hence

$$\dim \mathsf{U}(n) \leq n^2.$$

We will adopt the notation $\mathfrak{g} := T_I G$ for this real vector subspace of $\mathbb{k}^{n\times n}$. In fact, $\mathfrak{g}$ has a more interesting algebraic structure, namely that of a *Lie algebra*.

NOTE : It is customary to use *lower case Gothic* (Fraktur) characters (such as $\mathfrak{a}, \mathfrak{g}$ and $\mathfrak{h}$) to refer to Lie algebras.

4.4.10 DEFINITION. A (real) **Lie algebra** $\mathfrak{a}$ is a real vector space equipped with a product

$$[\cdot,\cdot] : \mathfrak{a} \times \mathfrak{a} \to \mathfrak{a}, \quad (x, y) \mapsto [x, y]$$

such that (for $\lambda, \mu \in \mathbb{R}$ and $x, y, z \in \mathfrak{a}$)

(LA1) $[x, y] = -[y, x]$.

(LA2) $[\lambda x + \mu y, z] = \lambda[x, z] + \mu[y, z]$.

(LA3) $[x, [y, z]] + [y, [z, x]] + [z, [x, y]] = 0$.

The product $[\cdot,\cdot]$ is called the *Lie bracket* of the Lie algebra $\mathfrak{a}$.

NOTE : (1) Condition (LA3) is called the *Jacobi identity*. So the Lie bracket $[\cdot,\cdot]$ of (the Lie algebra) $\mathfrak{a}$ is a *skew-symmetric bilinear* mapping (on $\mathfrak{a}$) which satisfies the Jacobi identity. Hence Lie algebras are *nonassociative* algebras. The Lie bracket plays for Lie algebras the same role that the associative law plays for associative algebras.

(2) While we can define *complex* Lie algebras (or, more generally, Lie algebras over any field), we shall only consider Lie algebras over the real field $\mathbb{R}$.

4.4.11 EXAMPLE. Let $\mathfrak{a} = \mathbb{R}^n$ and set (for all $x, y \in \mathbb{R}^n$)

$$[x, y] := 0.$$

The *trivial product* is a skew-symmetric bilinear multiplication (on $\mathbb{R}^n$) which satisfies the Jacobi identity and hence is a Lie bracket. $\mathbb{R}^n$ equipped with this product (Lie bracket) is a Lie algebra. Such a Lie algebra is called an *Abelian Lie algebra*.

◇ **Exercise 230** Show that the *only* Lie algebra structure on (the vector space) $\mathbb{R}$ is the trivial one.

4.4.12 EXAMPLE. Let $\mathfrak{a} = \mathbb{R}^3$ and set (for $x, y \in \mathbb{R}^3$)

$$[x, y] := x \times y \quad \text{(the cross product)}.$$

For the *standard unit vectors* e_1, e_2, e_3 we have

$$[e_1, e_2] = -[e_2, e_1] = e_3, \quad [e_2, e_3] = -[e_3, e_2] = e_1, \quad [e_3, e_1] = -[e_1, e_3] = e_2.$$

Then $\mathbb{R}^3$ equipped with this bracket operation is a Lie algebra. In fact, as we will see later, this is the Lie algebra of (the matrix group) $\mathsf{SO}(3)$ and also of $\mathsf{SU}(2)$ in disguise.

Given two matrices $A, B \in \mathbb{k}^{n\times n}$, their **commutator** is

$$[A, B] := AB - BA.$$

A and B *commute* (i.e., $AB = BA$) if and only if $[A, B] = 0$. The commutator $[\cdot, \cdot]$ is a product on $\mathbb{k}^{n\times n}$ satisfying conditions (LA1)-(LA3).

◇ **Exercise 231** Verify the *Jacobi identity* for the commutator $[\cdot, \cdot]$.

The *real* vector space $\mathbb{k}^{n\times n}$ equipped with the commutator $[\cdot, \cdot]$ is a Lie algebra.

NOTE : The procedure to give $\mathbb{k}^{n\times n}$ a Lie algebra structure can be extended to any *associative* algebra. A Lie product (bracket) can be defined in any associative algebra by the comutator $[x, y] = xy - yx$, making it a Lie algebra. Here the skew-symmetry condition (axiom) is clearly satisfied, and one can check easily that in this case the Jacobi identity for the commutator follows from the associtivity law for the ordinary product.

There is another way in which Lie algebras arise in the study of algebras. A *derivation* d of a *nonassociative* algebra $\mathcal{A}$ (i.e., a vector space endowed with a bilinear mapping $\mathcal{A} \times \mathcal{A} \to \mathcal{A}$) is a linear mapping $\mathcal{A} \to \mathcal{A}$ satisfying the formal analogue of the Leibniz rule for differentiating a product (for all $x, y \in \mathcal{A}$)

$$d(xy) = (dx)y + x(dy).$$

(The concept of a derivation is an abstraction of the idea of a *first order differential operator*.) The set of all derivations on $\mathcal{A}$ is clearly a vector subspace of the algebra $\mathrm{End}\,(\mathcal{A})$ of all linear mappings $\mathcal{A} \to \mathcal{A}$. Although the product of derivations is in general *not* a derivation, the commutator $d_1 d_2 - d_2 d_1$ of two derivations is again a derivation. Thus the set of all derivations of a nonassociative algebra is a Lie algebra, called the *derivation algebra* of the given nonassociative algebra.

Suppose that $\mathfrak{a}$ is a vector subspace of the Lie algebra $\mathbb{k}^{n\times n}$. Then $\mathfrak{a}$ is a *Lie subalgebra* of $\mathbb{k}^{n\times n}$ if it is *closed* under taking commutators of pairs of alements in $\mathfrak{a}$; that is,

$$A, B \in \mathfrak{a} \;\Rightarrow\; [A, B] \in \mathfrak{a}.$$

Of course, $\mathbb{k}^{n\times n}$ is a Lie subalgebra of itself.

4.4.13 THEOREM. *If $G \leq \mathsf{GL}\,(n, \mathbb{k})$ is a matrix group, then the tangent space $\mathfrak{g} = T_I G$ (at the identity) is a Lie subalgebra of $\mathbb{k}^{n\times n}$.*

PROOF : We will show that two curves α, β in G with $\alpha(0) = \beta(0) = I$, there is such a curve γ with $\dot\gamma(0) = [\dot\alpha(0), \dot\beta(0)]$.

Consider the mapping

$$F : (s, t) \mapsto F(s, t) := \alpha(s)\beta(t)\alpha(s)^{-1}.$$

This is clearly (continuous and) differentiable with respect to each of the variables s, t. For each s (in the domain of α), $F(s, \cdot)$ is a curve in G with $F(s, 0) = I$. Differentiating gives

$$\frac{d}{dt} F(s, t)\bigg|_{t=0} = \alpha(s)\dot\beta(0)\alpha(s)^{-1}$$

and so

$$\alpha(s)\dot\beta(0)\alpha(s)^{-1} \in \mathfrak{g}.$$

Since $\mathfrak{g}$ is a *closed subspace* of $\mathbb{k}^{n\times n}$ (Any vector subspace is an intersection of hyperplanes), whenever this limit exists we also have

$$\lim_{s\to 0} \frac{1}{s} \Big(\alpha(s)\dot\beta(0)\alpha(s)^{-1} - \dot\beta(0)\Big) \in \mathfrak{g}.$$

◇ **Exercise 232** Verify the following (matrix version of the) usual rule for differentiating an inverse :

$$\frac{d}{dt}\left(\alpha(t)^{-1}\right) = -\alpha(t)^{-1}\dot{\alpha}(t)\alpha(t)^{-1}.$$

We have

$$\begin{aligned}
\lim_{s\to 0}\frac{1}{s}\left(\alpha(s)\dot{\beta}(0)\alpha(s)^{-1} - \dot{\beta}(0)\right) &= \frac{d}{ds}\alpha(s)\dot{\beta}(0)\alpha(s)^{-1}\Big|_{s=0} \\
&= \dot{\alpha}(0)\dot{\beta}(0)\alpha(0) - \alpha(0)\dot{\beta}(0)\alpha(0)^{-1}\dot{\alpha}(0)\alpha(0)^{-1} \\
&= \dot{\alpha}(0)\dot{\beta}(0)\alpha(0) - \alpha(0)\dot{\beta}(0)\dot{\alpha}(0) \\
&= \dot{\alpha}(0)\dot{\beta}(0) - \dot{\beta}(0)\dot{\alpha}(0) \\
&= [\dot{\alpha}(0), \dot{\beta}(0)].
\end{aligned}$$

This shows that $[\dot{\alpha}(0), \dot{\beta}(0)] \in \mathfrak{g}$, hence it must be of the form $\dot{\gamma}(0)$ for some curve.

□

So for each matrix group G there is a Lie algebra $\mathfrak{g} = T_I G$. We call $\mathfrak{g}$ the **Lie algebra** of G.

NOTE : The essential phenomenon behind Lie theory is that *one may associate in a natural way to a matrix group G its Lie algebra $\mathfrak{g}$*. The Lie algebra is first of all a (real) vector space and secondly is endowed with a skew-symmetric bilinear product (called the Lie bracket or commutator). Amazingly, *the group G is almost completely determined by $\mathfrak{g}$ and its Lie bracket*. Thus for many purposes one can replace G with $\mathfrak{g}$. Since G is a complicated nonlinear object and $\mathfrak{g}$ is just a vector space, it is usually vastly simpler to work with $\mathfrak{g}$. Otherwise intractable computations may become straightforward linear algebra. This is one source of the power of Lie theory.

Homomorphisms of Lie algebras

A suitable type of homomorphism $G \to H$ between matrix groups gives rise to a linear mapping $\mathfrak{g} \to \mathfrak{h}$ respecting the Lie algebra structures.

4.4.14 DEFINITION. Let $G \leq \mathsf{GL}(n, \mathbb{k})$, $H \leq \mathsf{GL}(m, \mathbb{k})$ be matrix groups and let $\varphi : G \to H$ be a continuous mapping. Then φ is said to be **differentiable** if for every (differentiable) curve $\gamma : (a, b) \to G$, the composite mapping $\varphi \circ \gamma : (a, b) \to H$ is a (differentiable) curve with derivative

$$(\varphi \circ \gamma)^{\cdot}(t) = \frac{d}{dt}\varphi(\gamma(t))$$

and if whenever two (differentiable) curves $\alpha, \beta : (a, b) \to G$ both satisfy the conditions

$$\alpha(0) = \beta(0) \quad \text{and} \quad \dot{\alpha}(0) = \dot{\beta}(0)$$

then

$$(\varphi \circ \alpha)^{\cdot}(0) = (\varphi \circ \beta)^{\cdot}(0).$$

Such a mapping φ is a *differentiable homomorphism* if it is also a group homomorphism. A continuous homomorphism of matrix groups that is also *differentiable* is called a **Lie homomorphism**.

NOTE : The "technical restriction" in the definition of a Lie homomorphism is in fact *unnecessary.*

If $\varphi : G \to H$ is the restriction of a differentiable mapping $\Phi : \mathsf{GL}(n, \mathbb{k}) \to \mathsf{GL}(m, \mathbb{k})$, then φ is also a differentiable mapping.

4.4.15 PROPOSITION. *Let G, H, K be matrix groups and $\varphi : G \to H, \psi : H \to K$ be differentiable homomorphisms.*

(a) For each $A \in G$ there is a linear mapping $d\varphi_A : T_A G \to T_{\varphi(A)} H$ given by

$$d\varphi_A(\dot{\gamma}(0)) = (\varphi \circ \gamma)^{\cdot}(0).$$

(b) We have

$$d\psi_{\varphi(A)} \circ d\varphi_A = d(\psi \circ \varphi)_A.$$

(c) For the identity mapping $id_G : G \to G$ and $A \in G$,

$$d\, id_G = id_{T_A G}.$$

PROOF : (a) The definition of $d\varphi_A$ makes sense since (by the definition of differentiability), given $X \in T_AG$, for any curve γ with

$$\gamma(0) = A \quad \text{and} \quad \dot{\gamma}(0) = X$$

$(\varphi \circ \gamma)^{\cdot}(0)$ depends *only* on X and *not* on γ.

◇ **Exercise 233** Verify that the maping $d\varphi_A : T_AG \to T_{\varphi(A)}H$ is *linear*.

The identities (b) and (c) are straightforward to verify. □

If $\varphi : G \to H$ is a differentiable homomorphism, then (since $\varphi(I) = I$) $d\varphi_I : T_IG \to T_IH$ is a linear mapping, called the **derivative** of φ and usually denoted by

$$d\varphi : \mathfrak{g} \to \mathfrak{h}.$$

4.4.16 DEFINITION. Let $\mathfrak{g}, \mathfrak{h}$ be Lie algebras. A linear mapping $\Phi : \mathfrak{g} \to \mathfrak{h}$ is a **homomorphism of Lie algebras** if (for $x, y \in \mathfrak{g}$)

$$\Phi([x, y]) = [\Phi(x), \Phi(y)].$$

4.4.17 THEOREM. *Let G, H be matrix groups and $\varphi : G \to H$ be a Lie homomorphism. Then the derivative $d\varphi : \mathfrak{g} \to \mathfrak{h}$ is a homomorphism of Lie algebras.*

PROOF : Following ideas and notation in the proof of THEOREM 4.4.13, for curves α, β in G with $\alpha(0) = \beta(0) = I$, we can use the composite mapping

$$\varphi \circ F : (s, t) \mapsto \varphi(F(s, t)) = \varphi(\alpha(s))\varphi(\beta(t))\varphi(\alpha(s))^{-1}$$

to deduce

$$d\varphi([\dot{\alpha}(0), \dot{\beta}(0)]) = [d\varphi(\dot{\alpha}(0)), d\varphi(\dot{\beta}(0))].$$

□

4.4.18 COROLLARY. *Let G, H be matrix groups and $\varphi : G \to H$ be an isomorphism of matrix groups. Then the derivative $d\varphi : \mathfrak{g} \to \mathfrak{h}$ is an isomorphism of Lie algebras.*

PROOF : $\varphi^{-1} \circ \varphi$ is the identity, so

$$d\varphi^{-1} \circ d\varphi : T_I G \to T_I G$$

is the identity. Thus $d\varphi^{-1}$ is surjective and $d\varphi$ is injective.

Likewise, $\varphi \circ \varphi^{-1}$ is the identity, so $d\varphi \circ d\varphi^{-1}$ is the identity. Thus $d\varphi^{-1}$ is injective, and $d\varphi$ is surjective. The result now follows. □

NOTE : Isomorphic matrix groups have isomorphic Lie algebras. The converse (i.e., matrix groups with isomorphic Lie algebras are isomorphic) is *false.* For example, the rotation group $\mathsf{SO}(2)$ and the diagonal group

$$D_1 = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & e^a \end{bmatrix} \mid a \in \mathbb{R} \right\} \leq \mathsf{GA}^+(1, \mathbb{R})$$

have both Lie algebras isomorphic to $\mathbb{R}$ (the only Lie algebra structure on $\mathbb{R}$), but $\mathsf{SO}(2)$ is homeomorphic to a circle, while D_1 is homeomorphic to $\mathbb{R}$, so they are certainly *not* isomorphic.

However, the converse is – in a sense – almost true, so that the bracket operation on $\mathfrak{g}$ almost determine G as a group. After the existence of the Lie algebra, this fact is the most remarkable in Lie theory. Its precise formulation is known as *Lie's Third Theorem.*

4.5 More Properties of the Exponential Mapping

The following formula can be considered as another definition of the matrix exponential.

4.5.1 PROPOSITION. *Let* $A \in \mathbb{k}^{n \times n}$. *Then*

$$\exp(A) = \lim_{r \to \infty} \left(I + \frac{1}{r} A \right)^r .$$

PROOF : Consider the difference

$$\exp(A) - \left(I + \frac{1}{r} A \right)^r = \sum_{k=0}^{\infty} \left(\frac{1}{k!} - \frac{1}{r^k} \binom{r}{k} \right) A^k .$$

This matrix series converges since the series for the matrix exponential $\exp(A)$ converges and $\left(I + \frac{1}{r}A\right)^r$ is a polynomial. The coefficients in the rhs are nonnegative since

$$\frac{1}{k!} \geq \frac{r(r-1)\cdots(r-k+1)}{r \cdot r \cdots r} \frac{1}{k!}.$$

Therefore, setting $\|A\| = a$, we get

$$\left\| \exp(A) - \left(I + \frac{1}{r}A^r\right)^r \right\| \leq \sum_{k=0}^{\infty} \left(\frac{1}{k!} - \frac{1}{r^k}\binom{r}{k} \right) a^k = e^a - \left(1 + \frac{a}{r}\right)^r$$

where the expression on the right approaches zero (as $r \to \infty$). The result now follows. □

4.5.2 Proposition. *Let $A \in \mathbb{k}^{n\times n}$ and $\epsilon \in \mathbb{R}$. Then*

$$\det(I + \epsilon A) = 1 + \epsilon \operatorname{tr} A + O(\epsilon^2) \quad (as\ \epsilon \to 0).$$

Proof : The determinant of $I + \epsilon A$ equals the product of the eigenvalues of the matrix. But the eigenvalues of $I + \epsilon A$ (with due regard for multiplicity) equal $1 + \epsilon \lambda_i$, where the λ_i are the eigenvalues of A. It follows that

$$\begin{aligned} \det(I + \epsilon A) &= (1 + \epsilon \lambda_1)(1 + \epsilon \lambda_2) \cdots (1 + \epsilon \lambda_n) \\ &= 1 + \epsilon(\lambda_1 + \lambda_2 + \cdots + \lambda_n) + O(\epsilon^2) \\ &= 1 + \epsilon \operatorname{tr} A + O(\epsilon^2). \end{aligned}$$

□

Note : Whenever we have a mapping Z from some (open) interval (a, b), $a < 0 < b$ into a finite-dimensional normed vector space (e.g. $\mathbb{k}^{n\times n}$), then Z will often be denoted by $O(t^k)$ if $t \mapsto \frac{1}{t^k}Z(t)$ is bounded in an (open) neighborhood of the origin 0 (i.e. there are constants C_1 and C_2 such that

$$\|Z(t)\| \leq C_1|t^k| \quad \text{for } |t| < C_2.)$$

Thus $O(t^k)$ may denote different mappings at different times. The big-O notation was first introduced in 1892 by Paul G.H. Bachmann (1837-1920) in a book on number theory, and is currently used in several areas of mathematics and computer science (including mathematical analysis and the theory of algorithms).

The following result is useful.

4.5.3 LEMMA. *Let* $\alpha : (a, b) \to \mathbb{k}^{n\times n}$ *be a curve. Then*

$$\frac{d}{dt} \det\, \alpha(t)\bigg|_{t=0} = \operatorname{tr} \dot{\alpha}(0).$$

PROOF : The operation $\partial := \frac{d}{dt}\big|_{t=0}$ has the *derivation property*

$$\partial(\gamma_1\gamma_2) = (\partial\gamma_1)\gamma_2(0) + \gamma_1(0)\partial\gamma_2.$$

Put $\alpha(t) = \Big[a_{ij}\Big]$ and notice that (when $t = 0$) $a_{ij} = \delta_{ij}$. Write C_{ij} for the *cofactor matrix* obtained from $\alpha(t)$ by deleting the i^{th} row and the j^{th} column. By expanding along the n^{th} row we get

$$\det\, \alpha(t) = \sum_{j=1}^{n} (-1)^{n+j} a_{nj} \det\, C_{nj}.$$

For $t = 0$ (since $\alpha(0) = I$) we have

$$\det\, C_{nj} = \delta_{nj}.$$

Then

$$\begin{aligned}
\partial \det\, \alpha(t) &= \sum_{j=1}^{n} (-1)^{n+j} ((\partial a_{nj}) \det\, C_{nj} + a_{nj}(\partial \det\, C_{nj})) \\
&= \sum_{j=1}^{n} (-1)^{n+j} ((\partial a_{nj}) \det\, C_{nj}) + (\partial \det\, C_{nn}) \\
&= \partial a_{nn} + \partial \det\, C_{nn}.
\end{aligned}$$

We can repeat this calculation with the $(n-1)\times(n-1)$ matrix C_{nn} and so on. This gives

$$\begin{aligned}
\partial \det\, \alpha(t) &= \partial a_{nn} + \partial a_{n-1,n-1} + \partial \det\, C_{n-1,n-1} \\
&\vdots \\
&= \partial a_{nn} + \partial a_{n-1,n-1} + \cdots + \partial a_{11} \\
&= \operatorname{tr} \dot{\alpha}(0).
\end{aligned}$$

□

We can now prove a remarkable (and very useful) result, known as *Liouville's Formula*. Three different proofs will be given.

4.5.4 THEOREM. (LIOUVILLE'S FORMULA) *For* $A \in \mathbb{k}^{n\times n}$ *we have*

$$\det \exp(A) = e^{\operatorname{tr} A}.$$

FIRST SOLUTION (using the second definition of the exponential) : We have

$$\det \exp(A) = \det \lim_{r\to\infty} \left(I + \frac{1}{r}A\right)^r = \lim_{r\to\infty} \det \left(I + \frac{1}{r}A\right)^r$$

since the determinant function $\det : \mathbb{k}^{n\times n} \to \mathbb{k}$ is *continuous*. Moreover, by PROPOSITION 4.5.2,

$$\det \left(I + \frac{1}{r}A\right)^r = \left[\det \left(I + \frac{1}{r}A\right)\right]^r = \left[1 + \frac{1}{r}\operatorname{tr} A + O\left(\frac{1}{r^2}\right)\right]^r \text{ (as } r \to \infty).$$

It only remains to note that (for any $a \in \mathbb{k}$)

$$\lim_{r\to\infty} \left[1 + \frac{a}{r} + O\left(\frac{1}{r^2}\right)\right]^r = e^a.$$

In particular, for $a = \operatorname{tr} A$, we get the desired result. □

SECOND SOLUTION (using differential equations) : Consider the curve

$$\gamma : \mathbb{R} \to \mathsf{GL}(1, \mathbb{k}) = \mathbb{k}^\times, \quad t \mapsto \det \exp(tA).$$

Then (by LEMMA 4.5.3 applied to the curve γ)

$$\begin{aligned}
\dot\gamma(t) &= \lim_{h\to 0} \frac{1}{h} [\det \exp((t+h)A) - \det \exp(tA)] \\
&= \det \exp(tA) \lim_{h\to 0} \frac{1}{h} [\det \exp(hA) - 1] \\
&= \det \exp(tA) \operatorname{tr} A \\
&= \gamma(t) \operatorname{tr} A.
\end{aligned}$$

So γ satisfies the same differential equation and initial condition as the curve $t \mapsto e^{t \operatorname{tr} A}$. By the uniqueness of the solution (see **Exercise 219**), it follows that

$$\gamma(t) = \det \exp(tA) = e^{t \operatorname{tr} A}.$$

In particular, for $t = 1$, we get the desired result. □

THIRD SOLUTION (using Jordan canonical form) : If $B \in \mathsf{GL}(n, \Bbbk)$, then (see **Exercise 221**)

$$\begin{aligned} \det \exp(BAB^{-1}) &= \det(B \exp(A) B^{-1}) \\ &= \det B \cdot \det \exp(A) \cdot \det B^{-1} \\ &= \det \exp(A) \end{aligned}$$

and

$$e^{\operatorname{tr}(BAB^{-1})} = e^{\operatorname{tr} A}.$$

So it suffices to prove the identity for BAB^{-1} for a *suitably chosen* invertible matrix B. Using for example the theory of *Jordan canonical forms*, there is a suitable choice of such a B for which

$$BAB^{-1} = D + N$$

with D diagonal and N *strictly upper triangular* (i.e., $N_{ij} = 0$ for $i \geq j$). Then N is *nilpotent* (i.e., $N^k = O$ for some $k \geq 1$). We have

$$\begin{aligned} \exp(BAB^{-1}) &= \sum_{k=0}^{\infty} \frac{1}{k!} (D+N)^k \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} D^k + \sum_{k=0}^{\infty} \frac{1}{(k+1)!} \left((D+N)^{k+1} - D^{k+1} \right) \\ &= \exp(D) + \sum_{k=0}^{\infty} \frac{1}{(k+1)!} N(D^k + D^{k-1}N + \cdots + N^k). \end{aligned}$$

The matrix

$$N(D^k + D^{k-1}N + \cdots + N^k)$$

is strictly upper triangular, and so

$$\exp(BAB^{-1}) = \exp(D) + N'$$

where N' is strictly upper triangular. Now, if $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, we

have

$$\begin{aligned}
\det \exp(A) &= \det \exp(BAB^{-1}) \\
&= \det \exp(D) \\
&= \det \operatorname{diag}(e^{\lambda_1}, e^{\lambda_2}, \ldots, e^{\lambda_n}) \\
&= e^{\lambda_1} e^{\lambda_2} \cdots e^{\lambda_n} \\
&= e^{\lambda_1+\lambda_2+\cdots+\lambda_n} \\
&= e^{\operatorname{tr} D} \\
&= e^{\operatorname{tr}(BAB^{-1})} \\
&= e^{\operatorname{tr} A}.
\end{aligned}$$

□

The exponential mapping

$$\exp : \mathbb{k}^{n\times n} \to \mathsf{GL}(n, \mathbb{k})$$

is a basic link between the linear structure on $\mathbb{k}^{n\times n}$ and the multiplicative structure on $\mathsf{GL}(n, \mathbb{k})$. Let G be a matrix subgroup of $\mathsf{GL}(n, \mathbb{k})$. Applying PROPOSITION 4.3.3, we may choose $\rho \in \mathbb{R}$ so that $0 < \rho \leq \frac{1}{2}$ and if $A, B \in \mathcal{B}_{\mathbb{k}^{n\times n}}(O, \rho)$, then $\exp(A)\exp(B) \in \exp\left(\mathcal{B}_{\mathbb{k}^{n\times n}}(O, \frac{1}{2})\right)$. Since exp is one-to-one on $\mathcal{B}_{\mathbb{k}^{n\times n}}(O, \rho)$, *there is a unique matrix* $C \in \mathbb{k}^{n\times n}$ *for which*

$$\exp(A)\exp(B) = \exp(C).$$

NOTE : There is a beautiful formula, the *Baker-Campbell-Hausdorff formula* which expresses C as a universal power series in A and B. To develop this completely would take too long. Specifically, (one form of) the B-C-H formula says that *if* X *and* Y *are sufficiently small, then*

$$\exp(X)\exp(Y) = \exp(Z) \quad \text{with}$$

$$Z = X + Y + \frac{1}{2}[X, Y] + \frac{1}{12}[X, [X, Y]] - \frac{1}{12}[Y, [X, Y]] + \cdots$$

It is not supposed to be evident at the moment what "..." refers to. The only important point is that all the terms (in the expansion of Z) are given in terms of X

and Y, Lie brackets of X and Y, Lie brackets of Lie brackets involving X and Y, etc. Then it follows that the mapping $\phi : G \to \mathsf{GL}(n, \mathbb{R})$ "defined" by the relation

$$\phi(\exp(X)) = \exp(\phi(X))$$

is such that on elements of the form $\exp(X)$, with X sufficiently small, is a *group homomorphism*. Hence the B-C-H formula shows that *all the information about the group product, a least near the identity, is "encoded" in the Lie algebra.*

An interesting special case is the following : If $X, Y \in \mathbb{C}^{n\times n}$ and X, Y commute with their commutator (i.e., $[X,[X,Y]] = [Y,[X,Y]]$), then

$$\exp(X)\exp(Y) = \exp\left(X + Y + \frac{1}{2}[X,Y]\right).$$

◇ **Exercise 234** Show by direct computation that for

$$X, Y \in \mathfrak{heis} = \left\{ \begin{bmatrix} 0 & a & b \\ 0 & 0 & c \\ 0 & 0 & 0 \end{bmatrix} \mid a, b, c \in \mathbb{R} \right\}$$

(the Lie algebra of the Heisenberg group Heis)

$$\exp(X)\exp(Y) = \exp\left(X + Y + \frac{1}{2}[X,Y]\right).$$

We set

$$R = C - A - B \in \Bbbk^{n\times n}.$$

For $X \in \Bbbk^{n\times n}$, we have

$$\exp(X) = I + X + R_1(X),$$

where the remainder term $R_1(X)$ is given by

$$R_1(X) = \sum_{k=2}^{\infty} \frac{1}{k!} X^k.$$

Hence

$$\|R_1(X)\| \leq \|X\|^2 \sum_{k=2}^{\infty} \frac{1}{k!} \|X\|^{k-2}$$

and therefore if $\|X\| < 1$, then

$$\|R_1(X)\| \leq \|X\|^2 \sum_{k=2}^{\infty} \frac{1}{k!} = \|X\|^2 (e-2) < \|X\|^2.$$

Now for $X = C \in \mathcal{B}_{\mathbb{k}^{n\times n}}(O, \frac{1}{2})$, we have

$$\exp(C) = I + C + R_1(C)$$

with

$$\|R_1(C)\| < \|C\|^2.$$

Similar considerations lead to

$$\exp(C) = \exp(A)\exp(B) = I + A + B + R_1(A, B),$$

where

$$R_1(A, B) = \sum_{k=2}^{\infty} \frac{1}{k!} \left(\sum_{r=0}^{k} \binom{k}{r} A^r B^{k-r} \right).$$

This gives

$$\begin{aligned} \|R_1(A, B)\| &\leq \sum_{k=2}^{\infty} \frac{1}{k!} \left(\sum_{r=0}^{k} \binom{k}{r} \|A\|^r \|B\|^{k-r} \right) \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} (\|A\| + \|B\|)^k \\ &= (\|A\| + \|B\|)^2 \sum_{k=2}^{\infty} \frac{1}{k!} (\|A\| + \|B\|)^{k-2} \\ &\leq (\|A\| + \|B\|)^2 \end{aligned}$$

since $\|A\| + \|B\| < 1$.

Combining the two ways of writing $\exp(C)$ from above, we have

$$C = A + B + R_1(C) - R_1(A, B)$$

and so

$$\begin{aligned} \|C\| &\leq \|A\| + \|B\| + \|R_1(A, B)\| + \|R_1(C)\| \\ &< \|A\| + \|B\| + (\|A\| + \|B\|)^2 + \|C\|^2 \\ &\leq 2(\|A\| + \|B\|) + \frac{1}{2}\|C\| \end{aligned}$$

since $\|A\|, \|B\|, \|C\| \leq \frac{1}{2}$. Finally this gives

$$\|C\| \leq 4(\|A\| + \|B\|).$$

We also have

$$\begin{aligned} \|R\| = \|C - A - B\| &\leq \|R_1(A,B)\| + \|R_1(C)\| \\ &\leq (\|A\| + \|B\|)^2 + (4(\|A\| + \|B\|))^2 \\ &= 17\,(\|A\| + \|B\|)^2\,. \end{aligned}$$

We have proved the following result.

4.5.5 Proposition. *Let $A, B, C \in \mathcal{B}_{\mathbb{k}^{n\times n}}(O, \frac{1}{2})$ such that $\exp(A)\exp(B) = \exp(C)$. Then $C = A + B + R$, where the remainder term R satisfies*

$$\|R\| \leq 17\,(\|A\| + \|B\|)^2\,.$$

We can refine this estimate (to second order). We only point out the essential steps (details will be omitted). Set

$$S = C - A - B - \frac{1}{2}[A,B] \in \mathbb{k}^{n\times n}$$

and write

$$\exp(C) = I + C + \frac{1}{2}C^2 + R_2(C)$$

with

$$\|R_2(C)\| \leq \frac{1}{3}\|C\|^3.$$

Then

$$\begin{aligned} \exp(C) &= I + A + B + \frac{1}{2}[A,B] + S + \frac{1}{2}C^2 + R_2(C) \\ &= I + A + B + \frac{1}{2}(A^2 + 2AB + B^2) + T, \end{aligned}$$

where

$$T = S + \frac{1}{2}(C^2 - (A+B)^2) + R_2(C).$$

Also

$$\exp(A)\exp(B) = I + A + B + \frac{1}{2}(A^2 + 2AB + B^2) + R_2(A,B)$$

with

$$\|R_2(A,B)\| \leq \frac{1}{3}\,(\|A\| + \|B\|)^3\,.$$

We see that

$$S = R_2(A,B) + \frac{1}{2}((A+B)^2 - C^2) - R_2(C)$$

and by taking norms we get

$$\begin{aligned}\|S\| &\leq \|R_2(A,B)\| + \frac{1}{2}\|(A+B)(A+B-C) + (A+B-C)C\| + \|R_2(C)\| \\ &\leq \frac{1}{3}(\|A\| + \|B\|)^3 + \frac{1}{2}(\|A\| + \|B\| + \|C\|)\|A+B-C\| + \frac{1}{3}\|C\|^3 \\ &\leq 65(\|A\| + \|B\|)^3 .\end{aligned}$$

The following estimation holds.

4.5.6 Proposition. *Let* $A, B, C \in \mathcal{B}_{\mathbb{k}^{n\times n}}(O, \frac{1}{2})$ *such that* $\exp(A)\exp(B) = \exp(C)$. *Then* $C = A+B+\frac{1}{2}[A,B]+S$, *where the remainder term* S *satisfies*

$$\|S\| \leq 65(\|A\| + \|B\|)^3 .$$

We will derive two main consequences of Proposition 4.5.5 and Proposition 4.5.6. These relate group operations in $\mathsf{GL}(n,\mathbb{k})$ to the linear operations in $\mathbb{k}^{n\times n}$ and are crucial ingredients in the proof that *every matrix group is a Lie group.*

4.5.7 Theorem. (Lie-Trotter Product Formula) *For* $U, V \in \mathbb{k}^{n\times n}$ *we have*

$$\exp(U+V) = \lim_{r\to\infty}\left(\exp\left(\frac{1}{r}U\right)\exp\left(\frac{1}{r}V\right)\right)^r .$$

(This formula relates addition in $\mathbb{k}^{n\times n}$ to multiplication in $\mathsf{GL}(n,\mathbb{k})$.)

Proof : For *large* r we may take $A = \frac{1}{r}U$ and $B = \frac{1}{r}V$ and apply Proposition 4.5.5 to give

$$\exp\left(\frac{1}{r}U\right)\exp\left(\frac{1}{r}V\right) = \exp(C_r)$$

with

$$\left\|C_r - \frac{1}{r}(U+V)\right\| \leq \frac{17(\|U\| + \|V\|)^2}{r^2}.$$

As $r \to \infty$,

$$\|rC_r - (U+V)\| \leq \frac{17(\|U\| + \|V\|)^2}{r} \to 0$$

and hence

$$rC_r \to U + V.$$

Since $\exp(rC_r) = \exp(C_r)^r$, the *Lie-Trotter product formula* follows by continuity of the exponential mapping. □

4.5.8 THEOREM. (COMMUTATOR FORMULA) *For $U, V \in \mathbb{k}^{n\times n}$ we have*

$$\exp([U,V]) = \lim_{r\to\infty} \left(\exp\left(\frac{1}{r}U\right)\exp\left(\frac{1}{r}V\right)\exp\left(-\frac{1}{r}U\right)\exp\left(-\frac{1}{r}V\right)\right)^{r^2}.$$

(This formula relates the Lie bracket - or commutator - in $\mathbb{k}^{n\times n}$ to the *group commutator* in $\mathsf{GL}(n,\mathbb{k})$.)

PROOF : For *large* r (as in the proof of THEOREM 4.5.7) we have

$$\exp\left(\frac{1}{r}U\right)\exp\left(\frac{1}{r}V\right) = \exp(C_r)$$

with (as $r \to \infty$)

$$rC_r \to U + V.$$

We also have

$$C_r = \frac{1}{r}(U+V) + \frac{1}{2r^2}[U,V] + S_r,$$

where

$$\|S_r\| \leq 65\frac{(\|U\| + \|V\|)^3}{r^3}.$$

Similarly (replacing U, V with $-U, -V$) we obtain :

$$\exp\left(-\frac{1}{r}U\right)\exp\left(-\frac{1}{r}V\right) = \exp(C'_r),$$

where

$$C'_r = -\frac{1}{r}(U+V) + \frac{1}{2r^2}[U,V] + S'_r$$

and

$$\|S'_r\| \leq 65\frac{(\|U\| + \|V\|)^3}{r^3}.$$

Combining these we get

$$\begin{aligned}\exp\left(\frac{1}{r}U\right)\exp\left(\frac{1}{r}V\right)\exp\left(-\frac{1}{r}U\right)\exp\left(-\frac{1}{r}V\right) &= \exp(C_r)\exp(C'_r)\\ &= \exp(E_r),\end{aligned}$$

where

$$\begin{aligned} E_r &= C_r + C_r' + \frac{1}{2}[C_r, C_r'] + T_r \\ &= \frac{1}{r^2}[U, V] + \frac{1}{2}[C_r, C_r'] + S_r + S_r' + T_r. \end{aligned}$$

◇ **Exercise 235** Verify that

$$\begin{aligned} [C_r, C_r'] &= \frac{1}{r^3}\,[U+V,[U,V]] + \frac{1}{r}\,[U+V, S_r + S_r'] \\ &\quad + \frac{1}{2r^2}\,[[U,V], S_r' - S_r] + [S_r, S_r']. \end{aligned}$$

All four of these terms have norm bounded by an expression of the form $\frac{\text{constant}}{r^3}$ so the same is true of $[C_r, C_r']$. Also S_r, S_r', T_r have similarly bounded norms. Setting

$$Q_r := r^2 E_r - [U, V]$$

we obtain (as $r \to \infty$)

$$\|Q_r\| = r^2 \|E_r - \frac{1}{r^2}[U, V]\| \leq \frac{\text{constant}}{r} \to 0$$

and hence

$$\exp(E_r)^{r^2} = \exp\left([U, V] + Q_r\right) \to \exp([U, V]).$$

The *commutator formula* now follows using continuity of the exponential mapping.

□

NOTE : If g, h are elements of a group, then the expression $ghg^{-1}h^{-1}$ is called the *group commutator* of g and h.

There is one further concept involving the exponential mapping that is basic in Lie theory. It involves *conjugation*, which is generally referred to as the "adjoint action". For $g \in \mathsf{GL}(n, \Bbbk)$ and $A \in \Bbbk^{n \times n}$, we can form the conjugate

$$\mathrm{Ad}_g(A) := gAg^{-1}.$$

◇ **Exercise 236** Let $A, B \in \Bbbk^{n \times n}$ and $g, h \in \mathsf{GL}(n, \Bbbk)$. Show that (for $\lambda, \mu \in \Bbbk$)

(a) $\mathrm{Ad}_g(\lambda A + \mu B) = \lambda \mathrm{Ad}_g(A) + \mu \mathrm{Ad}_g(B)$.

(b) $\mathrm{Ad}_g([A,B]) = [\mathrm{Ad}_g(A), \mathrm{Ad}_g(B)]$.

(c) $\mathrm{Ad}_{gh}(A) = \mathrm{Ad}_g(\mathrm{Ad}_h(A))$.

In particular, $\mathrm{Ad}_g^{-1} = \mathrm{Ad}_{g^{-1}}$.

Formulas (a) an (b) say that Ad_g is an *automorphism* of the Lie algebra $\mathbb{k}^{n\times n}$, and formula (c) says the mapping

$$\mathrm{Ad} : \mathsf{GL}\,(n,\mathbb{k}) \to \mathsf{Aut}\,(\mathbb{k}^{n\times n}), \quad g \mapsto \mathrm{Ad}_g$$

is a *group homomorphism*. The mapping Ad is called the **adjoint representation** of (the matrix group) $\mathsf{GL}\,(n,\mathbb{k})$.

Formula (c) implies in particular that if $t \mapsto \exp(tA)$ is a one-parameter subgroup of $\mathsf{GL}\,(n,\mathbb{k})$, then $\mathrm{Ad}_{\exp(tA)}$ is a one-parameter group (of linear transformations) in $\mathbb{k}^{n\times n}$. Observe that we can identify $\mathsf{Aut}\,(\mathbb{k}^{n\times n})$ with $\mathsf{GL}\,(n^2,\mathbb{k})$ (and thus view $\mathsf{Aut}\,(\mathbb{k}^{n\times n})$ as a matrix group). Then (by THEOREM 4.4.3)

$$\mathrm{Ad}_{\exp(tA)} = \exp(t\mathcal{A})$$

for some $\mathcal{A} \in \mathbb{k}^{n^2\times n^2} = \mathsf{End}\,(\mathbb{k}^{n\times n})$. Since

$$\begin{aligned}\mathcal{A}(B) &= \frac{d}{dt}\mathrm{Ad}_{\exp(tA)}(B)\Big|_{t=0}\\ &= \frac{d}{dt}\exp(tA)B\exp(-tA)\Big|_{t=0}\\ &= [A,B]\end{aligned}$$

by setting (for $A, B \in \mathbb{k}^{n\times n}$)

$$\mathrm{ad}\,A(B) := [A,B]$$

we have the following formula

$$\mathrm{Ad}_{\exp(tA)} = \exp(t\,\mathrm{ad}\,A).$$

Explicitly, the formula says that

$$\exp(tA)B\exp(-tA) = \sum_{k=0}^{\infty}\frac{t^k}{k!}(\mathrm{ad}\,A)^k\,B.$$

(Here $(\operatorname{ad} A)^0 = A$ and $(\operatorname{ad} A)^k = \operatorname{ad}(\operatorname{ad} A)^{k-1}$ for $k \geq 1$.)

NOTE : The mapping

$$\operatorname{ad} : \mathbb{k}^{n\times n} \to \mathsf{End}\,(\mathbb{k}^{n\times n}), \quad X \mapsto \operatorname{ad} X$$

is called the *adjoint representation* of (the Lie algebra) $\mathbb{k}^{n\times n}$. From the Jacobi identity for Lie algebras, we have

$$\operatorname{ad} X([Y,Z]) = [\operatorname{ad} X(Y), Z] + [Y, \operatorname{ad} X(Z)].$$

That is, $\operatorname{ad} X$ is a *derivation* of the Lie algebra $\mathbb{k}^{n\times n}$. The formula above gives the relation between the automorphism $\operatorname{Ad}_{\exp(tX)}$ of the Lie algebra $\mathbb{k}^{n\times n}$ and the derivation $\operatorname{ad} X$ of $\mathbb{k}^{n\times n}$. One also has

$$\exp(t\operatorname{Ad}_g(X)) = g\exp(tX)g^{-1}.$$

Using this formula, we can see that $[X,Y] = 0$ if and only if $\exp(tX)$ and $\exp(sY)$ commute for arbitrary $s,t \in \mathbb{R}$.

◇ **Exercise 237** Let $A, B \in \mathbb{k}^{n\times n}$.

(a) Verify that

$$\operatorname{ad}[A,B] = \operatorname{ad} A \operatorname{ad} B - \operatorname{ad} B \operatorname{ad} A = [\operatorname{ad} A, \operatorname{ad} B].$$

(This means that $\operatorname{ad} : \mathbb{k}^{n\times n} \to \mathsf{End}\,(\mathbb{k}^{n\times n})$ is a *Lie algebra homomorphism.*)

(b) Show by induction that

$$(\operatorname{ad} A)^n(B) = \sum_{k=0}^{n} \binom{n}{k} A^k B (-A)^{n-k}.$$

(c) Show by direct computation that

$$\exp(\operatorname{ad} A)(B) = \operatorname{Ad}_{\exp(A)}(B) = \exp(A)B\exp(-A).$$

4.6 Examples of Lie Algebras of Matrix Groups

The Lie algebras of $\mathsf{GL}(n,\mathbb{R})$ and $\mathsf{GL}(n,\mathbb{C})$

Let us start with the *real* general linear group $\mathsf{GL}(n,\mathbb{R}) \subseteq \mathbb{R}^{n\times n}$. We have shown (see EXAMPLE 4.4.8) that $T_I\mathsf{GL}(n,\mathbb{R}) = \mathbb{R}^{n\times n}$. Hence *the Lie algebra* $\mathfrak{gl}(n,\mathbb{R})$ *of* $\mathsf{GL}(n,\mathbb{R})$ *consists of all* $n \times n$ *matrices (with real entries), with the commutator as the Lie bracket.* Thus

$$\mathfrak{gl}(n,\mathbb{R}) = \mathbb{R}^{n\times n}.$$

It follows that

$$\dim \mathsf{GL}(n,\mathbb{R}) = \dim \mathfrak{gl}(n,\mathbb{R}) = n^2.$$

Similarly, *the Lie algebra of the* complex *general linear group* $\mathsf{GL}(n,\mathbb{C})$ *is*

$$\mathfrak{gl}(n,\mathbb{C}) = \mathbb{C}^{n\times n}$$

and

$$\dim \mathsf{GL}(n,\mathbb{C}) = \dim_{\mathbb{R}} \mathfrak{gl}(n,\mathbb{C}) = 2n^2.$$

The Lie algebras of $\mathsf{SL}(n,\mathbb{R})$ and $\mathsf{SL}(n,\mathbb{C})$

For $\mathsf{SL}(n,\mathbb{R}) \leq \mathsf{GL}(n,\mathbb{R})$, suppose that

$$\alpha : (a,b) \to \mathsf{SL}(n,\mathbb{R})$$

is a curve in $\mathsf{SL}(n,\mathbb{R})$ with $\alpha(0) = I$. For $t \in (a,b)$ we have $\det \alpha(t) = 1$ and so

$$\frac{d}{dt} \det \alpha(t) = 0.$$

Using LEMMA 4.5.3, it follows that

$$\operatorname{tr} \dot{\alpha}(0) = 0$$

and thus

$$T_I\mathsf{SL}(n,\mathbb{R}) \subseteq \ker \operatorname{tr} := \left\{ A \in \mathbb{R}^{n\times n} \,|\, \operatorname{tr} A = 0 \right\}.$$

If $A \in \ker \operatorname{tr} \subseteq \mathbb{R}^{n\times n}$, the curve

$$\alpha : (a,b) \to \mathbb{R}^{n\times n}, \quad t \mapsto \exp(tA)$$

satisfies (the boundary conditions)

$$\alpha(0) = I \quad \text{and} \quad \dot{\alpha}(0) = A.$$

Moreover, using *Liouville's Formula*, we get

$$\det \alpha(t) = \det \exp(tA) = e^{t \operatorname{tr} A} = 1.$$

Hence *the Lie algebra* $\mathfrak{sl}(n, \mathbb{R})$ *of* $\mathsf{SL}(n, \mathbb{R})$ *consists of all* $n \times n$ *matrices (with real entries) having trace zero, with the commutator as the Lie bracket.* Thus

$$\mathfrak{sl}(n, \mathbb{R}) = T_I \mathsf{SL}(n, \mathbb{R}) = \{A \in \mathfrak{gl}(n, \mathbb{R}) \mid \operatorname{tr} A = 0\}.$$

Since $\operatorname{tr} A = 0$ imposes one condition on A, it follows that

$$\dim \mathsf{SL}(n, \mathbb{R}) = \dim_{\mathbb{R}} \mathfrak{sl}(n, \mathbb{R}) = n^2 - 1.$$

Similarly, *the Lie algebra of the complex special linear group* $\mathsf{SL}(n, \mathbb{C})$ *is*

$$\mathfrak{sl}(n, \mathbb{C}) = T_I \mathsf{SL}(n, \mathbb{C}) = \{A \in \mathfrak{gl}(n, \mathbb{C}) \mid \operatorname{tr} A = 0\}$$

and

$$\dim \mathsf{SL}(n, \mathbb{C}) = \dim_{\mathbb{R}} \mathfrak{sl}(n, \mathbb{C}) = 2(n^2 - 1).$$

The Lie algebras of $\mathsf{O}(n)$ and $\mathsf{SO}(n)$

First, consider the orthogonal group $\mathsf{O}(n)$; that is,

$$\mathsf{O}(n) = \left\{A \in \mathsf{GL}(n, \mathbb{R}) \mid A^\top A = I\right\} \leq \mathsf{GL}(n, \mathbb{R}).$$

Given a curve $\alpha : (a, b) \to \mathsf{O}(n)$ with $\alpha(0) = I$, we have

$$\frac{d}{dt}\alpha(t)^T \alpha(t) = 0$$

and so

$$\dot{\alpha}(t)^T \alpha(t) + \alpha(t)^T \dot{\alpha}(t) = 0$$

which implies

$$\dot{\alpha}(0)^T + \dot{\alpha}(0) = 0.$$

Thus we must have $\dot{\alpha}(0) \in \mathbb{R}^{n\times n}$ is *skew-symmetric*. So

$$T_I\mathsf{O}(n) \subseteq \mathsf{Sk\text{-}sym}(n) = \left\{A \in \mathbb{R}^{n\times n} \mid A^\top + A = 0\right\}$$

(the set of all $n \times n$ skew-symmetric matrices in $\mathbb{R}^{n\times n}$).

On the other hand, if $A \in \mathsf{Sk\text{-}sym}(n) \subseteq \mathbb{R}^{n\times n}$, we consider the curve

$$\alpha : (a,b) \to \mathsf{GL}(n,\mathbb{R}), \quad t \mapsto \exp(tA).$$

Then

$$\begin{aligned}\alpha(t)^\top\alpha(t) &= \exp(tA)^\top \exp(tA)\\ &= \exp(tA^\top)\exp(tA)\\ &= \exp(-tA)\exp(tA)\\ &= I.\end{aligned}$$

Hence we can view α as a curve in $\mathsf{O}(n)$. Since $\dot{\alpha}(0) = A$, this shows that

$$\mathsf{Sk\text{-}sym}(n) \subseteq T_I\mathsf{O}(n)$$

and hence *the Lie algebra* $\mathfrak{o}(n)$ *of the orthogonal group* $\mathsf{O}(n)$ *consists of all* $n\times n$ *skew-symmetric matrices, with the usual commutator as the Lie bracket.* Thus

$$\mathfrak{o}(n) = T_I\mathsf{O}(n) = \mathsf{Sk\text{-}sym}(n) = \left\{A \in \mathbb{R}^{n\times n} \mid A^\top + A = 0\right\}.$$

It follows that (see PROPOSITION 4.4.9)

$$\dim \mathsf{O}(n) = \dim \mathfrak{o}(n) = \frac{n(n-1)}{2}.$$

◇ **Exercise 238** Show that if $A \in \mathsf{Sk\text{-}sym}(n)$, then $\operatorname{tr} A = 0$.

By *Liouville's Formula*, we have

$$\det \alpha(t) = \det \exp(tA) = 1$$

and hence $\alpha : (a,b) \to \mathsf{SO}(n)$, where $\mathsf{SO}(n)$ is the *special orthogonal group.* We have actually shown that *the Lie algebra of the special orthogonal group* $\mathsf{SO}(n)$ *is*

$$\mathfrak{so}(n) = \mathfrak{o}(n) = \left\{A \in \mathbb{R}^{n\times n} \mid A^\top + A = 0\right\}.$$

The Lie algebra of $\mathsf{SO}(3)$

We will discuss the Lie algebra $\mathfrak{so}(3)$ of the *rotation group* $\mathsf{SO}(3)$ in some detail.

◇ **Exercise 239** Show that

$$\mathfrak{so}(3) = \left\{ \begin{bmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{bmatrix} \in \mathbb{R}^{3\times 3} \mid a, b, c \in \mathbb{R} \right\}.$$

The Lie algebra $\mathfrak{so}(3)$ is a 3-dimensional real vector space. Consider the *rotations*

$$R_1(t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos t & -\sin t \\ 0 & \sin t & \cos t \end{bmatrix}, \; R_2(t) = \begin{bmatrix} \cos t & 0 & \sin t \\ 0 & 1 & 0 \\ -\sin t & 0 & \cos t \end{bmatrix}, \; R_3 = \begin{bmatrix} \cos t & -\sin t & 0 \\ \sin t & \cos t & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Then the mappings

$$\rho_i : t \mapsto R_i(t), \quad i = 1, 2, 3$$

are curves in $\mathsf{SO}(3)$ and clearly $\rho_i(0) = I$. It follows that

$$\dot{\rho}_i(0) := A_i \in \mathfrak{so}(3), \quad i = 1, 2, 3.$$

These elements (matrices) are

$$A_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

◇ **Exercise 240** Verify that the matrices A_1, A_2, A_3 form a basis for $\mathfrak{so}(3)$. We shall refer to this basis as the *standard basis*.

◇ **Exercise 241** Compute all the Lie brackets (commutators) $[A_i, A_j]$, $i, j = 1, 2, 3$ and then determine the coefficients c_{ij}^k defined by

$$[A_i, A_j] = c_{ij}^1 A_1 + c_{ij}^2 A_2 + c_{ij}^3 A_3, \quad i, j = 1, 2, 3.$$

These coefficients are called the *structure constants* of the Lie algebra. They determine completely the Lie bracket $[\cdot, \cdot]$.

The Lie algebra $\mathfrak{so}(3)$ may be *identified* with (the Lie algebra) $\mathbb{R}^3$ as follows. We define the mapping

$$\widehat{\ }: \mathbb{R}^3 \to \mathfrak{so}(3), \quad x = (x_1, x_2, x_3) \mapsto \widehat{x} := \begin{bmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{bmatrix}.$$

This mapping is called the *hat mapping*.

◇ **Exercise 242** Show that the hat mapping $\widehat{\ }: \mathbb{R}^3 \to \mathfrak{so}(3)$ is an *isomorphism* of vector spaces.

◇ **Exercise 243** Show that (for $x, y \in \mathbb{R}^3$)

(a) $x \times y = \widehat{x}\, y$.

(b) $\widehat{x \times y} = [\widehat{x}, \widehat{y}]$.

(c) $x \bullet y = -\frac{1}{2}\mathrm{tr}\,(\widehat{x}\,\widehat{y})$.

Formula (b) says that the hat mapping is in fact an isomorphism of Lie algebras and so *we can identify the Lie algebra* $\mathfrak{so}(3)$ *with (the Lie algebra)* $\mathbb{R}^3$.

NOTE : For $x \in \mathbb{R}^3$ and $t \in \mathbb{R}$, the matrix exponential $\exp(t\,\widehat{x})$ is a *rotation* about (the axis) x through the angle $t\|x\|$. The following explicit formula for $\exp(\widehat{x})$ is known as *Rodrigues' Formula* :

$$\exp(\widehat{x}) = I + \frac{\sin\|x\|}{\|x\|}\,\widehat{x} + \frac{1}{2}\left[\frac{\sin\left(\frac{\|x\|}{2}\right)}{\frac{\|x\|}{2}}\right]^2 \widehat{x}^2.$$

This result says that *the exponential mapping*

$$\exp : \mathfrak{so}(3) \to \mathsf{SO}(3)$$

is onto. Rodrigues' Formula is useful in *computational solid mechanics*, along with its quaternionic counterpart.

The Lie algebras of $\mathsf{U}(n)$ and $\mathsf{SU}(n)$

Consider the unitary group $\mathsf{U}(n)$; that is,

$$\mathsf{U}(n) = \{A \in \mathsf{GL}(n, \mathbb{C}) \mid A^*A = I\}.$$

For a curve α in $\mathsf{U}(n)$ with $\alpha(0) = I$, we obtain

$$\dot{\alpha}(0)^* + \dot{\alpha}(0) = 0$$

and so $\dot{\alpha}(0) \in \mathbb{C}^{n\times n}$ is skew-Hermitian. So

$$T_I\mathsf{U}(n) \subseteq \mathsf{Sk\text{-}Herm}(n) = \left\{A \in \mathbb{C}^{n\times n} \mid A^* + A = 0\right\}$$

(the set of all $n \times n$ skew-Hermitian matrices in $\mathbb{C}^{n\times n}$).

If $H \in \mathsf{Sk\text{-}Herm}(n)$, then the curve

$$\alpha : (a, b) \to \mathsf{GL}(n, \mathbb{C}), \quad t \mapsto \exp(tH)$$

satisfies

$$\begin{aligned} \alpha(t)^*\alpha(t) &= \exp(tH)^* \exp(tH) \\ &= \exp(tH^*) \exp(tH) \\ &= \exp(-tH) \exp(tH) \\ &= I. \end{aligned}$$

Hence we can view α as a curve in $\mathsf{U}(n)$. Since $\dot{\alpha}(0) = H$, this shows that

$$\mathsf{Sk\text{-}Herm}(n) \subseteq T_I\mathsf{U}(n)$$

and hence *the Lie algebra* $\mathfrak{u}(n)$ *of the unitary group* $\mathsf{U}(n)$ *consists of all* $n\times n$ *skew-Hermitian matrices, with the usual commutator as the Lie bracket.* Thus

$$\mathfrak{u}(n) = T_I\mathsf{U}(n) = \mathsf{Sk\text{-}Herm}(n) = \left\{H \in \mathbb{C}^{n\times n} \mid H^* + H = 0\right\}.$$

It follows that (see **Exercise 229**)

$$\dim \mathsf{U}(n) = \dim_{\mathbb{R}} \mathfrak{u}(n) = n^2.$$

The special unitary group $\mathsf{SU}(n)$ can be handled in a similar way. Again we have

$$\mathfrak{su}(n) = T_I\mathsf{SU}(n) \subseteq \mathsf{Sk\text{-}Herm}(n).$$

But also if $\alpha : (a,b) \to \mathsf{SU}(n)$ is a curve with $\alpha(0) = I$ then, as in the analysis for $\mathsf{SL}(n,\mathbb{R})$,

$$\operatorname{tr}\dot{\alpha}(0) = 0.$$

Writing

$$\mathsf{Sk\text{-}Herm}^0(n) := \{H \in \mathsf{Sk\text{-}Herm}(n) \mid \operatorname{tr} H = 0\}$$

this gives $\mathfrak{su}(n) \subseteq \mathsf{Sk\text{-}Herm}^0(n)$. On the other hand, if $H \in \mathsf{Sk\text{-}Herm}^0(n)$ then the curve

$$\alpha : (a,b) \to \mathsf{U}(n), \quad t \mapsto \exp(tH)$$

takes values in $\mathsf{SU}(n)$ and $\dot{\alpha}(0) = H$. Hence

$$\mathfrak{su}(n) = T_I\mathsf{SU}(n) = \mathsf{Sk\text{-}Herm}^0(n) = \left\{H \in \mathbb{C}^{n\times n} \mid H^* + H = 0 \text{ and } \operatorname{tr} H = 0\right\}.$$

NOTE : For a matrix group $G \leq \mathsf{GL}(n,\mathbb{R})$ (with Lie algebra $\mathfrak{g}$), the following are true (and can be used in determining Lie algebras of matrix groups).

- The mapping

  $$\exp_G : \mathfrak{g} \to \mathsf{GL}(n,\mathbb{R}), \quad X \mapsto \exp(X)$$

  has image contained in G, $\exp_G(\mathfrak{g}) \subseteq G$. We will normally write $\exp_G : \mathfrak{g} \to G$ for the exponential mapping on G (and sometimes even just exp). In general, the exponential mapping $\exp_G$ is neither one-to-one nor onto.

- If G is compact and connected, then $\exp_G$ is onto.

- The mapping $\exp_G$ maps a neighborhood of 0 in $\mathfrak{g}$ bijectively onto a neighborhood of I in G.

◇ **Exercise 244** Verify that the exponential mapping

$$\exp_{\mathsf{U}(1)} : \mathbb{R} \to \mathsf{U}(1) = \mathbb{S}^1, \quad t \mapsto e^{it}$$

is onto but not one-to-one.

4.6.1 EXAMPLE. The exponential mapping

$$\exp_{\mathsf{SL}\,(2,\mathbb{R})} : \mathfrak{sl}(2,\mathbb{R}) \to \mathsf{SL}\,(2,\mathbb{R})$$

is not onto. Let

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \frac{1}{\lambda} \end{bmatrix} \quad \text{with } \lambda < -1.$$

We see that $A \in \mathsf{SL}\,(2,\mathbb{R})$ and we shall show that A is *not* of the form $\exp(X)$ with $X \in \mathfrak{sl}(2,\mathbb{R})$. If $A = \exp(X)$, then the eigenvalues of A are of the form e^a and e^b, where a and b are the eigenvalues of X. Suppose $\lambda = e^a$ and $\frac{1}{\lambda} = e^b$. Then

$$a = -b + 2k\pi\, i, \quad k \in \mathbb{Z}.$$

However, since λ is negative, a is actually complex and therefore its conjugate is also an eigenvalue; that is, $b = \bar{a}$. This gives a as pure imaginary. Then

$$1 = |e^a| = |\lambda| = -\lambda$$

which contradicts the assumption that $\lambda < -1$.

The Lie algebra of $\mathsf{SU}\,(2)$

We will discuss the Lie algebra $\mathfrak{su}(2)$ in some detail.

◇ **Exercise 245** Show that

$$\mathfrak{su}(2) = \left\{ \begin{bmatrix} ci & -b+ai \\ b+ai & -ci \end{bmatrix} \in \mathbb{C}^{2\times 2} \mid a,b,c \in \mathbb{R} \right\}.$$

The Lie algebra $\mathfrak{su}(2)$ is a 3-dimensional real vector space. Consider the matrices

$$H_1 = \frac{1}{2}\begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}, \quad H_2 = \frac{1}{2}\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad H_3 = \frac{1}{2}\begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}.$$

Clearly,

$$H_i \in \mathfrak{su}(2), \quad i = 1,2,3.$$

◇ **Exercise 246** Verify that the matrices H_1, H_2, H_3 form a basis for $\mathfrak{su}(2)$.

◇ **Exercise 247** Compute all the Lie brackets (commutators) $[H_i, H_j]$, $i, j = 1, 2, 3$ and then determine the structure constants of (the Lie algebra) $\mathfrak{su}(2)$.

Consider the mapping

$$\phi : \mathbb{R}^3 \to \mathfrak{su}(2), \quad x = (x_1, x_2, x_3) \mapsto x_1 H_1 + x_2 H_2 + x_3 H_3.$$

◇ **Exercise 248** Show that the mapping $\phi : \mathbb{R}^3 \to \mathfrak{su}(2)$ is an *isomorphism* of Lie algebras.

Thus *we can identify the Lie algebra* $\mathfrak{su}(2)$ *with (the Lie algebra)* $\mathbb{R}^3$.

NOTE : The Lie algebras $\mathfrak{su}(2)$ and $\mathfrak{so}(3)$ look the same algebraically (they are isomorphic). An explicit isomorphism (of Lie algebras) is given by

$$\psi : x_1 H_1 + x_2 H_2 + x_3 H_3 \mapsto x_1 A_1 + x_2 A_2 + x_3 A_3.$$

This suggests that there might be a close relationship between the matrix groups themselves. Indeed there is a (surjective) *Lie homomorphism* $\mathsf{SU}(2) \to \mathsf{SO}(3)$ whose derivative (at I) is ψ. Recall the adjoint representation

$$\mathrm{Ad} : \mathsf{SU}(2) \to \mathsf{Aut}(\mathfrak{su}(2)), \quad A \mapsto \mathrm{Ad}_A \, (: U \mapsto AUA^*).$$

Each Ad_A is a *linear isomorphism* of $\mathfrak{su}(2)$. Ad_A is actually an *orthogonal transformation* on $\mathfrak{su}(2)$ (the mapping $(X, Y) \mapsto -\mathrm{tr}\,(XY)$ is an inner product on $\mathfrak{su}(2)$) and so Ad_A corresponds to an element of $\mathsf{O}(3)$ (in fact, $\mathsf{SO}(3)$). The mapping

$$\overline{\mathrm{Ad}} : \mathsf{SU}(2) \to \mathsf{SO}(3), \quad A \mapsto \mathrm{Ad}_A$$

turns out to be a continuous homomorphism of matrix groups that is differentiable (i.e., a Lie homomorphism) and such that its derivative $d\,\overline{\mathrm{Ad}} : \mathfrak{su}(2) \to \mathfrak{so}(3)$ is ψ.

The Lie algebras of $\mathsf{UT}(n, \Bbbk)$ and $\mathsf{UT}^u(n, \Bbbk)$

Let $\alpha : (a, b) \to \mathsf{UT}(n, \Bbbk)$ be a curve in $\mathsf{UT}(n, \Bbbk)$ with $\alpha(0) = I$. Then $\dot{\alpha}(0)$ is upper triangular. Moreover, using the argument for $\mathsf{GL}(n, \Bbbk)$ we see that given any upper triangular matrix $A \in \Bbbk^{n \times n}$, there is a curve

$$\sigma : (-\epsilon, \epsilon) \to \Bbbk^{n \times n}, \quad t \mapsto tA + I$$

such that $\sigma(t) \in \mathsf{UT}(n,\Bbbk)$ and $\dot{\sigma}(0) = A$. Hence *the Lie algebra* $\mathfrak{ut}(n,\Bbbk)$ *of* $\mathsf{UT}(n,\Bbbk)$ *consists of all* $n \times n$ *upper triangular matrices, with the usual commutator as the Lie bracket.* Thus

$$\mathfrak{ut}(n,\Bbbk) = T_I\mathsf{UT}(n,\Bbbk) = \left\{A \in \Bbbk^{n\times n} \,|\, a_{ij} = 0 \ \text{ for } i > j\right\}.$$

It follows that

$$\dim \mathsf{UT}(n,\Bbbk) = \dim_{\mathbb{R}} \mathfrak{ut}(n,\Bbbk) = \frac{n(n+1)}{2} \dim_{\mathbb{R}} \Bbbk.$$

An upper triangular matrix $A \in \Bbbk^{n\times n}$ is *strictly upper triangular* if all its diagonal entries are 0. Then *the Lie algebra of the unipotent group* $\mathsf{UT}^u(n,\Bbbk)$ *consists of all* $n \times n$ *strictly upper triangular matrices, with the usual commutator as the Lie bracket.* So

$$\mathfrak{sut}(n,\Bbbk) = T_I\mathsf{UT}^u(n,\Bbbk) = \left\{A \in \Bbbk^{n\times n} \,|\, a_{ij} = 0 \ \text{ for } i \geq j\right\}.$$

◇ **Exercise 249** Find $\dim_{\mathbb{R}} \mathfrak{sut}(n,\Bbbk)$.

◇ **Exercise 250** For each of the following matrix group G, determine its Lie algebra $\mathfrak{g}$ and hence its dimension.

(a) $G = \left\{A \in \mathsf{GL}(2,\mathbb{R}) \,|\, AQA^\top = Q\right\}$, where $Q = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$.

(b) $G = \left\{A \in \mathsf{GL}(2,\mathbb{R}) \,|\, AQA^\top = Q\right\}$, where $Q = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$.

(c) $G = \mathsf{GA}(3,\mathbb{R})$.

(d) $G = \mathsf{Heis}$.

(e) $G = G_4 \leq \mathsf{UT}^u(4,\mathbb{R})$ from **Exercise 199**.

(f) $G = \mathsf{E}(n)$.

(g) $G = \mathsf{SE}(n)$.

◇ **Exercise 251**

(a) Show that the Lie algebra of the symplectic group $\mathsf{Sp}(2n,\mathbb{R})$ is

$$\mathfrak{sp}(2n,\mathbb{R}) = \left\{A \in \mathbb{R}^{2n\times 2n} \,|\, A^\top\mathbb{J} + \mathbb{J}A = 0\right\}.$$

(b) If

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathfrak{sl}(2n, \mathbb{R})$$

show that $A \in \mathfrak{sp}(2n, \mathbb{R})$ if and only if

$$d = -a^\top, \quad c = c^\top, \quad \text{and} \quad b = b^\top.$$

(c) Calculate the dimension of $\mathfrak{sp}(2n, \mathbb{R})$.

◇ **Exercise 252** Show that the Lie algebra of the Lorentz group Lor is

$$\mathfrak{lor} = \left\{ A \in \mathbb{R}^{4\times 4} \,|\, SA + A^\top S = 0 \right\} = \left\{ \begin{bmatrix} 0 & a_1 & a_2 & a_3 \\ -a_1 & 0 & a_4 & a_5 \\ -a_2 & -a_4 & 0 & a_6 \\ a_3 & a_5 & a_6 & 0 \end{bmatrix} \,|\, a_1, a_2, a_3, a_4, a_5, a_6 \in \mathbb{R} \right\}.$$

◇ **Exercise 253** Consider the matrix group $\mathbb{k}^\times = \mathsf{GL}(1, \mathbb{k})$. (Its Lie algebra is clearly $\mathbb{k}$.)

(a) Show that the *determinant* function

$$\det : \mathsf{GL}(n, \mathbb{k}) \to \mathbb{k}^\times$$

is a *Lie homomorphism* (i.e. a continuous homomorphism of matrix groups that is also *differentiable*; cf. Definition 4.4.14).

(b) Show that the induced *homomorphism* of Lie algebras (i.e. the derivative of det) is the *trace* function

$$\mathrm{tr} : \mathbb{k}^{n\times n} \to \mathbb{k}.$$

(c) Derive from (*b*) that (for $A, B \in \mathbb{k}^{n\times n}$)

$$\mathrm{tr}(AB) = \mathrm{tr}(BA).$$

Chapter 5

Manifolds

Topics :

1. Manifolds: Definition and Examples
2. Smooth Functions and Mappings
3. The Tangent and Cotangent Spaces
4. Smooth Submanifolds
5. Vector Fields
6. Differential Forms

5.1 Manifolds: Definition and Examples

Submanifolds (in fact, immersed submanifolds) of Euclidean space $\mathbb{R}^m$ are a generalization of the concept of regular curve in the Euclidean 3-space $\mathbb{R}^3$. The major defect of the definition of a submanifold is its dependence of $\mathbb{R}^m$. Indeed, the natural idea of an ℓ-dimensional smooth submanifold is of *a set which is ℓ-dimensional* (in a certain sense) *and to which the differential calculus of $\mathbb{R}^m$ can be applied*; the unnecessary presence of $\mathbb{R}^m$ is simply an imposition of our physical nature.

NOTE : In his monograph on surface theory, published in 1827, CARL F. GAUSS (1777-1855) developed the geometry on a surface (based on its fundamental form); the necessity of an *abstract* idea of surface – that is, without involving the ambient space – was already clear to him. This was generalized by BERNHARD RIEMANN (1826-1866) to m-dimensions in his inaugural lecture (Habilitationschrift) at Göttingen, "On the Hypotheses which lie at the Foundation of Geometry" (1854), marking the birth of modern (differential) geometry. However, it was nearly a century before such an idea attained the definite form that we shall present here.

The concept of *manifold* is one of the most sophisticated basic concepts in mathematics.

Definition (of a manifold) and examples

Let $\mathbb{R}^m$ denote the Euclidean m-space *in the broad sense* (i.e., the vector space $\mathbb{R}^m$ equipped with its *canonical topology* and natural *differentiable structure*).

Let M be a set.

5.1.1 DEFINITION. A (coordinate) **chart** on M is a pair (U, ϕ), where $U \subseteq M$ and $\phi : U \to \mathbb{R}^m$ is a one-to-one mapping *onto* an open subset $\phi(U)$ of $\mathbb{R}^m$.

One often writes $\phi(p) = (\phi_1(p), \ldots, \phi_n(p))$, viewing this as the coordinate m-tuple of the point $p \in U$. The functions $\phi_i : U \to \mathbb{R}$, $i = 1, 2, \ldots, m$ are called the *coordinate functions* associated with the chart (U, ϕ).

NOTE : A chart is also called a (local) *coordinate system* (on M).

Relative to such a *coordinatization*, one can do calculus in the region U of M. The problem is that the point p will generally belong to infinitely many different coordinate charts and calculus in one of these coordinatizations about p might not agree with calculus in another. One needs the coordinate systems to be *smoothly compatible* in the following sense.

5.1.2 DEFINITION. Two charts (U, ϕ) and (V, ψ) on M are said to be C^∞-*related* if either $U \cap V = \emptyset$ or

$$\psi \circ \phi^{-1} : \phi(U \cap V) \to \psi(U \cap V)$$

is a smooth diffeomeorphism (between open subsets in $\mathbb{R}^m$).

We think of $\psi \circ \phi^{-1}$ as a smooth change of coordinates (on $\phi(U \cap V)$). Thus, on $U \cap V$, functions are smooth relative to one coordinate system if and only if they are smooth relative to the other. Indeed, differential calculus carried out in $U \cap V$ via the coordinates of $\phi(U \cap V)$ is equivalent to the calculus carried out via the coordinates of $\psi(U \cap V)$. (The explicit formulas will, of course, change from the one coordinate system to the other.) Furthermore, piecing together these local calculi produces a global calculus on M. The concept that allows us to make these remarks precise is that of a *smooth atlas*.

5.1.3 DEFINITION. A (smooth) **atlas** on M is a family $\mathcal{A} = \{(U_\alpha, \phi_\alpha)\}_{\alpha \in \mathfrak{A}}$ of charts (on M) such that

(AT1) $M = \bigcup_{\alpha \in \mathfrak{A}} U_\alpha$;

(AT2) (U_α, ϕ_α) is C^∞-related to (U_β, ϕ_β) for every $\alpha, \beta \in \mathfrak{A}$.

Two atlases $\mathcal{A}$ and $\mathcal{A}'$ on M are *compatible* provided their union $\mathcal{A} \cup \mathcal{A}'$ is also an atlas on M. Compatibility is an equivalence relation (on the set of all atlases on M). *Each atlas on M is equivalent to a unique maximal atlas on M.* Thus we arrive at the definition of a manifold.

5.1.4 DEFINITION. A maximal atlas $\mathcal{A}$ on M is called a **smooth structure** on M (also called a *differentiable structure* or a C^∞ structure). An n-**dimensional smooth** (or differentiable or C^∞) **manifold** is a pair $(M, \mathcal{A})$ (i.e. a set equipped with a smooth structure).

By a typical abuse of notation, we usually write M for the smooth manifold, the presence of the differentiable structure $\mathcal{A}$ being understood. An **admissible chart** on (the smooth manifold) M is any chart belonging to any (smooth) atlas in the differentiable structure of M.

NOTE : (1) We often refer to m-dimensional smooth manifolds simply as *m-manifolds*.

(2) In practice one defines a manifold M by means of a single (smooth) atlas (not necessarily maximal) on M which completely determines the differentiable structure.

We will now define on a manifold M a *canonical topology*, one that only depends on the differentiable structure.

NOTE : One could also have started from a topological space M and required that the domains U_α of the charts be open sets in M and that the mappings $\phi_\alpha : U_\alpha \to \phi_\alpha(U_\alpha)$ be homeomorphisms.

5.1.5 PROPOSITION. *Let M be a (smooth) m-manifold. The collection of unions of domains of admissible charts on M forms a topology (called the canonical topology) on M.*

PROOF : Let $\mathcal{O}$ be the set thus defined. Clearly, $M \in \mathcal{O}$ and we have to show that $\mathcal{O}$ satisfies the two axioms for a topology :

(O1) Every union of elements of $\mathcal{O}$ is an element of $\mathcal{O}$.

(O2) Every finite intersection of elements of $\mathcal{O}$ is an element of $\mathcal{O}$.

Clearly (O1) is satisfied, since a set is in $\mathcal{O}$ if and only if it is a union of domains of charts. To show (O2), we just have to consider the intersection of two elements of $\mathcal{O}$. Let them be $A = \cup_{\alpha\in\mathfrak{A}_1} U_\alpha$ and $B = \cup_{\beta\in\mathfrak{A}_2} U_\beta$; then

$$A \cap B = \bigcup_{(\alpha,\beta)\in\mathfrak{A}_1\times\mathfrak{A}_2} (U_\alpha \cap U_\beta).$$

We have to show that each intersection $U_\alpha \cap U_\beta$ can be taken as the domain of a chart *compatible* with the differentiable structure (i.e. an admissible chart on M). Let (U_α, ϕ_α) be an admissible chart on M and set $\psi := \phi_\alpha|_{U_\alpha \cap U_\beta}$; we claim that $(U_\alpha \cap U_\beta, \psi)$ is the desired admissible chart. Clearly $\psi(U_\alpha \cap U_\beta) = \phi_\alpha(U_\alpha \cap U_\beta)$ is open in $\mathbb{R}^m$. If (U, ϕ) is any admissible chart, the composition $\phi \circ \phi_\alpha^{-1}$ is a (smooth) diffeomorphism between (the open sets) $\phi_\alpha(U \cap U_\alpha)$ and $\phi(U \cap U_\alpha)$, so

$$\phi \circ \psi^{-1} = \phi \circ \phi_\alpha^{-1}\big|_{\phi_\alpha(U_\alpha \cap U_\beta \cap U)}$$

is a (smooth) diffeomorphism between $\psi(U \cap (U_\alpha \cap U_\beta))$ and $\phi(U \cap (U_\alpha \cap U_\beta))$. Similarly, $\psi \circ \phi^{-1}$ is a (smooth) diffeomorphism between $\phi(U \cap (U_\alpha \cap U_\beta))$ and $\psi(U \cap (U_\alpha \cap U_\beta))$. This proves compatibility.

□

NOTE : Sometimes it is desirable to characterize the open sets in the canonical topology of M in terms of a single atlas. One can prove that *given an atlas* $\mathcal{A} = \{(U_\alpha, \phi_\alpha)\}_{\alpha \in \mathfrak{A}}$ *on an* m*-manifold* M*, a subset* $U \subseteq M$ *is open if and only if the set* $\phi_\alpha(U \cap U_\alpha) \subseteq \mathbb{R}^m$ *is open for every chart* $(U_\alpha, \phi_\alpha) \in \mathcal{A}$. This result provides another way of defining the (canonical) topology of a manifold : for every chart (U, ϕ) on an m-manifold M, considered with its canonical topology, the mapping $\phi : U \to \phi(U) \subseteq \mathbb{R}^m$ is a homeomorphism.

The canonical topology of a manifold can be quite strange. In particular, it can happen that one (or both) of the following conditions (axioms) *not* be satisfied :

(A) *Hausdorff Axiom :* Given two distinct points of M, there exist (open) neighborhoods of these points that do not intersect.

(B) *Countable Basis Axiom :* M can be covered by a countable number of coordinate neighborhoods (i.e. domains of admissible charts on M). We say then that M *has a countable basis* (or that M is second countable).

NOTE : Axiom (A) is essential for the uniqueness of limits of convergent sequences whereas Axiom (B) is essential for the existence of a (smooth) *partition of unity*, an

almost indispensabil tool for the study of certain questions on manifolds. A topological space which is locally compact (each point has at least one compact neighborhood), Hausdorff, and has a countable basis (of open sets) is *paracompact*, and hence admits a partition of the unity. For example, a partition of unity is required for piecing together global functions and structures out of local ones, and conversely for representing global structures as locally finite sums of local ones. The following result holds : A (smooth) manifold M has a (smooth) partition of unity if and only if every connected component of M is Hausdorff and has a countable basis.

For all practical purposes, we shall be interested in *only* (smooth) manifolds that satisfy Axiom (A) and Axiom (B). Henceforth, we shall refer to such objects, simply, as *manifolds*.

NOTE : (1) Manifolds are locally Euclidean spaces (A Hausdorff topological space is said to be *locally Euclidean* of dimension m if each point p has an open neighborhood homeomorphic to an open set of $\mathbb{R}^m$). Second countable locally Euclidean spaces are known as *topological manifolds*. A topological manifold is *smoothable* provides it can be given a smooth structure. For $m = 1, 2, 3$ it is known that all topological m-manifolds are smoothable. The first dimension in which there exist nonsmoothable manifolds is $m = 4$.

(2) Manifolds are paracompact spaces (A Hausdorff space is called *paracompact* if every open cover has a locally finite subcover). Moreover, manifolds are metrizable spaces (A topological space is called *metrizable* if there exists a metric such that its associated metric topology coincides with the space topology; any metrizable space is paracompact).

(3) Any m-manifold admits a *finite* atlas consisting of $m + 1$ (not necessarily connected) charts. This is a consequence of *topological dimension theory*.

(4) A manifold is connected if and only if it is path-connected. (A path-connected topological space is connected, but the converse is *not* true in general.)

(5) A natural question in the theory of (differentiable) manifolds is to know whether a given manifold can be *immersed* (or even *embedded*) into some Euclidean space. A fundamental result in this direction is the famous theorem of HASSLER WHITNEY (1907-1989) which states the following : *Any m-manifold can be immersed in $\mathbb{R}^{2m}$ and embedded in $\mathbb{R}^{2m+1}$* (in fact, the theorem can be improved, for $m \geq 2$, to $\mathbb{R}^{2m-1}$

and $\mathbb{R}^{2m}$, respectively).

(6) A set M may have more than one inequivalent smooth structure. For instance, the spheres from dimension 7 on have finitely many. A most surprising result is that *on* $\mathbb{R}^4$ *there are uncountably many pairwise inequivalent (exotic) smooth structures.*

We give now some preliminary examples of manifolds.

5.1.6 EXAMPLE. (*Euclidean space*) The standard smooth structure on the Euclidean m-space $\mathbb{E}^m$ is obtained by taking the atlas consisting of a single (global) chart $(\mathbb{E}^m, \iota)$, where $\iota : \mathbb{E}^m \to \mathbb{R}^m$ is the identity mapping. (Many examples will make it abundantly clear that manifolds in general can *not* be covered by a single coordinate system nor are there preferred coordinates.)

NOTE : It is common practice to identify $\mathbb{E}^m$ and $\mathbb{R}^m$; however, we DO NOT follow this custom. It is often better in thinking of the Euclidean space $\mathbb{E}^m$ as a "flat" Riemannian manifold (i.e. a "geometrical" model for classical geometry, *without* coordinates; a Riemannian manifold is a manifold equipped with an additional "geometrical" structure, called a *Riemannian metric*) and of the Cartesian space $\mathbb{R}^m$ as a normed vector space (i.e. an "algebraic" model for classical geometry, *with* coordinates). The additive group of $\mathbb{R}^m$, also denoted by $\mathbb{R}^m$, is a matrix group. This group is isomorphic to (and customarily identified with) the group of all translations on the Euclidean space $\mathbb{E}^m$.

5.1.7 EXAMPLE. Let V be an m-dimensional vector space (over $\mathbb{R}$). Then V has a natural manifold structure. Indeed, if $\{v_1, \ldots, v_m\}$ is a basis in V, then the correspondence

$$\phi : p = p_1 v_1 + \cdots + p_m v_m \mapsto (p_1, \ldots, p_m)$$

is a bijection (between V and the open set $\mathbb{R}^m$). The pair (V, ϕ) is a (global) chart on V and hence uniquely determines a smooth structure on V. This smooth structure is independent of the choice of the basis, since different bases give C^∞-related charts. (In fact, the change of coordinates is given simply by an $m \times m$ invertible matrix.)

5.1.8 EXAMPLE. (*Open submanifolds*) An open subset U of a manifold M is itself a manifold. Indeed, if $\{(U_\alpha, \phi_\alpha)\}_{\alpha\in\mathfrak{A}}$ is the (maximal) atlas of admissible charts on M, then the family of charts (atlas)

$$\mathcal{A}_U = \{(U \cap U_\alpha, \phi_\alpha | \, U \cap U_\alpha) \,|\, (U_\alpha, \phi_\alpha) \in \mathcal{A}\}$$

defines a smooth structure on U. Unless otherwise stated, open subsets of manifolds will always be given this natural (induced) smooth structure.

More generally, any ℓ-dimensional smooth submanifold of some Euclidean space $\mathbb{E}^m$ is a (smooth) ℓ-manifold.

◇ **Exercise 254** Let S be a (non-empty) subset of the Euclidean space $\mathbb{E}^m$ and assume that S satisfies the *ℓ-submanifold property* (i.e. S is an *ell*-dimensional submanifold of $\mathbb{E}^m$). Show that S is naturally endowed with a smooth structure, hence it *is* an *ell*-manifold.

5.1.9 EXAMPLE. (*The general linear group*) The general linear group $\mathsf{GL}(n, \mathbb{R})$ is an open subset of the manifold $\mathbb{E}^{n^2}$ (we may identify $\mathbb{R}^{n\times n}$ with the Cartesian space $\mathbb{R}^{n^2}$). Hence $\mathsf{GL}(n, \mathbb{R})$ is a manifold.

5.1.10 EXAMPLE. (*The sphere*) The *n-sphere* is the set

$$\mathbb{S}^n := \left\{x \in \mathbb{E}^{n+1} \,|\, x_1^2 + \cdots + x_{n+1}^2 = 1\right\}.$$

(We have seen that $\mathbb{S}^n$ is an n-dimensional smooth submanifold of $\mathbb{E}^{n+1}$.) Let $p_N = (0, \ldots, 0, 1)$ be the north pole and $p_S = (0, \ldots, 0, -1)$ the south pole of $\mathbb{S}^n$. Define the mapping $\phi_1 : U_1 := \mathbb{S}^n \setminus \{p_N\} \to \mathbb{R}^n$ that takes the point $p = (x_1, \ldots, x_{n+1})$ in U_1 into the intersection of the hyperplane $x_{n+1} = 0$ with the line that passes through p and p_N. This mapping is the so-called *stereographic projection* from the north pole. In a similar manner one defines the stereographic projection $\phi_{-1} : U_{-1} := \mathbb{S}^n \setminus \{p_S\} \to \mathbb{R}^n$ from the south pole.

◇ **Exercise 255** Show that the stereographic projections (ϕ_1 and ϕ_{-1}) are given by

$$\phi_{\pm 1}(x_1, \ldots, x_{n+1}) = \left(\frac{x_1}{1 \mp x_{n+1}}, \cdots, \frac{x_n}{1 \mp x_{n+1}}\right).$$

Clearly, the stereographic projections are one-to-one and hence the pairs (U_1, ϕ_1) and (U_{-1}, ϕ_{-1}) are charts on $\mathbb{S}^n$. The domains (coordinate neighborhoods) of these two charts cover $\mathbb{S}^n$ and is not difficult to see that they are C^∞-related (and hence form a smooth atlas on the sphere). Indeed, the change of coordinates

$$y_i = \frac{x_i}{1 - x_{n+1}} \longleftrightarrow y'_i = \frac{x_i}{1 + x_{n+1}} \quad (i = 1, 2, \ldots, n)$$

is given by

$$y'_i = \frac{y_i}{y_1^2 + \cdots + y_n^2}$$

(here we use the fact that $x_1^2 + \cdots + x_{n+1}^2 = 1$). Therefore, the n-sphere $\mathbb{S}^n$ is an n-manifold.

5.1.11 EXAMPLE. (*Product manifolds*) Let M and N be manifolds (of dimension m and n, respectively). Suppose that $\mathcal{A} = \{(U_\alpha, \phi_\alpha)\}_{\alpha \in \mathfrak{A}}$ and $\mathcal{B} = \{(V_\beta, \psi_\beta)\}_{\beta \in \mathfrak{B}}$ are the maximal atlases on M and N, respectively.

◇ **Exercise 256** Show that the family (of charts)

$$\{(U_\alpha \times V_\beta, \phi_\alpha \times \psi_\beta) \,|\, (U_\alpha, \phi_\alpha) \in \mathcal{A},\ (V_\beta, \psi_\beta) \in \mathcal{B}\}$$

where $\phi_\alpha \times \psi_\beta(p, q) := (\phi_\alpha(p), \psi_\beta(q)) \in \mathbb{R}^m \times \mathbb{R}^n$, is a smooth atlas on $M \times N$ (which determines a smooth structure).

With this smooth structure $M \times N$ is an $(m + n)$-manifold, called the *product manifold* of M and N. An important example is the *torus* $\mathbb{T}^2 = \mathbb{S}^1 \times \mathbb{S}^1$, the product of two *circles*. More generally, the k-dimensional torus $\mathbb{T}^k = \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$ is a k-manifold obtained as a Cartesian product.

5.2 Smooth Functions and Mappings

On a topological space the concept of continuity has meaning; in an analogous way, on a manifold we may define the concept of *smooth* (also called differentiable or C^∞) function. Let M be an m-manifold.

5.2.1 DEFINITION. A function $f : M \to \mathbb{R}$ is said to be **smooth** if for any point $p \in M$ there is an admissible chart (U, ϕ) on M such that $p \in U$ and the composite function

$$f \circ \phi^{-1} : \phi(U) \subseteq \mathbb{R}^m \to \mathbb{R}$$

is smooth.

Clearly, a smooth function is continuous. The set of all smooth functions on M will be denoted by $C^\infty(M)$. It is a consequence of the definition that if $f \in C^\infty(M)$ and $W \subseteq M$ is an open set, then $f|_W$ is smooth (on the manifold W).

NOTE : The definition only requires us to be able to find *some* chart about each point $p \in M$, but the following result assures us that *all* admissible charts will then work : *The function $f : M \to \mathbb{R}$ is smooth if and only if $f \circ \phi^{-1}$ is smooth for every admissible chart (U, ϕ) on M.*

We think of $f \circ \phi^{-1}$ as a *formula* for $f|_U$ relative to the coordinate system (U, ϕ). For $x \in U$, with coordinates $\phi(x) = (x_1, \ldots, x_m)$, we can write

$$\begin{aligned} y &= f(x) \\ &= f \circ \phi^{-1}(\phi(x)) \\ &= f \circ \phi^{-1}(x_1, \ldots, x_m). \end{aligned}$$

We shall refer to $f \circ \phi^{-1}$ as the *local representation* of f with respect to (U, ϕ).

5.2.2 EXAMPLE. Among the smooth functions on M are the coordinate functions of an admissible chart (U, ϕ). Indeed, for each $i = 1, 2, \ldots, m$, the local representation of $\phi_i = \mathrm{pr}_i \circ \phi$ is given by

$$\begin{aligned} y &= \phi_i(x) \\ &= \phi_i \circ \phi^{-1}(x_1, \ldots, x_m) \\ &= \mathrm{pr}_i \circ \phi \circ \phi^{-1}(x_1, \ldots, x_m) \\ &= \mathrm{pr}_i(x_1, \ldots, x_m) \\ &= x_i \end{aligned}$$

which is clearly smooth (see also **Exercise 120**).

Just as in the case of (the manifold) $\mathbb{E}^n$ we proceed from definition of smooth function to definition of smooth mapping. Suppose that M and N are manifolds.

5.2.3 DEFINITION. A mapping $F : M \to N$ is said to be **smooth** if for any point $p \in M$ there is an admissible chart (U, ϕ) on M with $p \in U$ and an admissible chart (V, ψ) on N with $F(p) \in V$ such that $F(U) \subseteq V$ and the composite mapping

$$\psi \circ F \circ \phi^{-1} : \phi(U) \to \psi(V)$$

is smooth.

Smooth mappings are continuous; their restrictions to open subsets are also smooth. The set of all smooth mappings from M into N will be denoted by $C^\infty(M, N)$.

NOTE : A smooth mapping is a more general notion than smooth function, the latter being a mapping (from a manifold M) into $N = \mathbb{R}$, which is, of course, the same as (the manifold) $\mathbb{E}^1$.

The local representation of F with respect to (U, ϕ) and (V, ψ) is given by

$$y_i = \psi \circ F \circ \phi^{-1}(x_1, \ldots, x_m), \qquad i = 1, 2, \ldots, n.$$

◇ **Exercise 257** Prove that a mapping $F : M \to N$ is smooth if and only if for any smooth function $f : N \to \mathbb{R}$, the function $f \circ F$ is smooth (on M). (We write F^*f for the function $f \circ F$, and shall refer to F^*f as the *pull-back* of f under F.)

An open interval J of $\mathbb{R}$ is an open submanifold of $\mathbb{R}$ (in fact, the Euclidean 1-space $\mathbb{E}^1$) and hence is a manifold. Then *a curve $\sigma : J \to N$ is smooth if and only if for any smooth function f on N, (the pull-back of f under σ) $\sigma^*f : J \to \mathbb{R}$ is a smooth function.*

◇ **Exercise 258** Let M and N be manifolds. Prove that the *canonical projections*

$$\mathrm{pr}_M : M \times N \to M \quad \text{and} \quad \mathrm{pr}_N : M \times N \to N$$

are smooth mappings (between manifolds).

◇ **Exercise 259** Let M, N, and P be manifolds. Prove that if $F : M \to N$ and $G : N \to P$ are smooth mappings, then $G \circ F : M \to P$ is also smooth.

◇ **Exercise 260** Let M be a manifold. Show that the set $C^\infty(M)$ of all smooth functions on M is an *algebra* (over $\mathbb{R}$) under the natural operations of addition, scalar multiplication, and product.

5.3 The Tangent and Cotangent Spaces

The tangent space

There are several alternative ways in which we can define *tangent vectors* (and hence tangent spaces) to a manifold, independent of any embedding in some Euclidean space.

NOTE : The whole reason for introducing tangent vectors is to produce linear approximations to nonlinear problems.

An intuitive (and very useful) way to define tangent vectors is as equivalence classes of curves. (Roughly speaking, two curves are equivalent if they have the same velocity vector at some point.)

Let M be an m-manifold and let $\mathbf{C}(p)$ denote the set of all smooth curves $\sigma : (-\varepsilon, \varepsilon) \to M$ such that $\sigma(0) = p$. Elements (curves) α and β in $\mathbf{C}(p)$ are said to be *infinitesimally equivalent* at p and we write $\alpha \sim_p \beta$ if

$$\frac{d}{dt}\phi(\alpha(t))\bigg|_{t=0} = \frac{d}{dt}\phi(\beta(t))\bigg|_{t=0}$$

for any admissible chart (U, ϕ) on M.

◇ **Exercise 261** Show that if (U, ϕ) and (V, ψ) are two admissible charts at p (i.e. such that $p \in U \cap V$), then

$$\frac{d}{dt}\phi \circ \alpha(t)\bigg|_{t=0} = \frac{d}{dt}\phi \circ \beta(t)\bigg|_{t=0}$$

if and only if

$$\frac{d}{dt}\psi \circ \alpha(t)\bigg|_{t=0} = \frac{d}{dt}\psi \circ \beta(t)\bigg|_{t=0}.$$

(The infinitesimal equivalence is well defined.)

It is easy to check that $\sim_p$ is an equivalence relation on the set $\mathbf{C}(p)$. The infinitesimal equivalence class of α in $\mathbf{C}(p)$ is denoted by $[\alpha]_p$ and is called an *infinitesimal curve* at p. An infinitesimal curve at p is also called a **tangent vector** to M at p.

5.3.1 DEFINITION. The (quotient) set $T_pM := \mathbf{C}(p)/_{\sim_p}$ of all infinitesimal curves at p is called the **tangent space** to M at p.

Let (U, ϕ) be any admissible chart on M such that $p \in U$. The mapping

$$\bar{\phi} : T_pM \to \mathbb{R}^m, \quad [\alpha]_p \mapsto \left.\frac{d}{dt}\phi(\alpha(t))\right|_{t=0}$$

is one-to-one and onto $\mathbb{R}^m$. In fact, for any $v \in \mathbb{R}^m$, $\alpha(t) := \phi^{-1}(\phi(p) + tv)$ is a curve such that $\bar{\phi}([\alpha]_p) = v$. We define the vector structure on T_pM so that $\bar{\phi}$ becomes a linear isomorphism. That is, for $[\alpha]_p, [\beta]_p \in T_pM$ and $a \in \mathbb{R}$,

$$\begin{aligned} [\alpha]_p + [\beta]_p &:= \bar{\phi}^{-1}\left(\bar{\phi}([\alpha]_p) + \bar{\phi}([\beta]_p)\right) \\ a[\alpha]_p &:= \bar{\phi}^{-1}\left(a\bar{\phi}([\alpha]_p)\right). \end{aligned}$$

Under the forgoing addition and scalar multiplication, *the tangent space T_pM is an m-dimensional vector space over $\mathbb{R}$.*

NOTE : The linear structure of T_pM is *canonical* in the sense that it is independent of the choice of (local) coordinates. Indeed, let (U, ϕ) and (V, ψ) be two admissible charts at p. Let $\bar{\phi}([\alpha]_p) = v$ and let $\bar{\psi}([\alpha]_p) = w$. It follows that

$$w = \left.\frac{d}{dt}\psi \circ \alpha(t)\right|_{t=0} = \left.\frac{d}{dt}\psi \circ \phi^{-1} \circ \phi \circ \alpha(t)\right|_{t=0}.$$

Therefore the coordinates of v and w transform according to the following formula :

$$w_i = \frac{\partial y_i}{\partial x_1}v_1 + \cdots + \frac{\partial y_i}{\partial x_m}v_m$$

where $y_i = y_i(x_1, \ldots, x_m)$, $i = 1, 2, \ldots, m$ denote the coordinate functions of the mapping $\psi \circ \phi^{-1}$. Hence the vector structure on T_pM is independent of the particular chart (used to define it).

The cotangent space

Let M be an m-dimensional manifold and let $\mathbf{F}(p)$ denote the set of all smooth functions f, defined in some (open) neighborhood of $p \in M$, that satisfy $f(p) = 0$. $\mathbf{F}(p)$ *will have a natural vector space structure* (in fact, associative algebra with unity) *provided that functions that agree on a common domain are regarded as equal.* (The domains of elements of $\mathbf{F}(p)$ need not be the same.)

NOTE : Actually, an element of (the algebra) $\mathbf{F}(p)$ is a certain set (equivalence class) of smooth functions, commonly referred to as a *function germ* at p, which is conveniently identified with any one of its representatives.

Elements (function germs) f and g in $\mathbf{F}(p)$ are said to be *equivalent* (at p) and we write $f \approx_p g$ if

$$D\left(f \circ \phi^{-1}\right)(\phi(p)) = D\left(g \circ \phi^{-1}\right)(\phi(p))$$

for any admissible chart (U, ϕ) on M.

NOTE : We shall write, by a slight abuse of notation,

$$f \circ \phi^{-1}(x_1, \ldots, x_m) = f(x_1, \ldots, x_m) \quad \text{and} \quad D(f \circ \phi^{-1}) = \frac{\partial f}{\partial x} := \left[\frac{\partial f}{\partial x_1} \quad \cdots \quad \frac{\partial f}{\partial x_m}\right].$$

Again, it is easy to check that $\approx_p$ is an equivalence relation on the set $\mathbf{F}(p)$. The equivalence class of f in $\mathbf{F}(p)$ is denoted by $[f]_p$ and is called a **tangent covector** to M at p.

5.3.2 DEFINITION. The (quotient) set $T_p^*M := \mathbf{F}(p)/_{\approx_p}$ is called the **cotangent space** to M at p.

◇ **Exercise 262** Let $f, \bar{f}, g, \bar{g} \in \mathbf{F}(p)$ and $a \in \mathbb{R}$. Show that

(a) If $f \approx_p \bar{f}$ and $g \approx_p \bar{g}$, then $f + g \approx_p \bar{f} + \bar{g}$.

(b) If $f \approx_p \bar{f}$, then $af \approx_p a\bar{f}$.

That is, for $[f]_p, [g]_p \in T_p^*M$ and $a \in \mathbb{R}$, the following operations

$$\begin{aligned} [f]_p + [g]_p &:= [f + g]_p \\ a[f]_p &:= [af]_p \end{aligned}$$

are well-defined. Under the foregoing addition and scalar multiplication, *the cotangent space* T_p^*M *is a real vector space.*

For each admissible chart (U, ϕ) on M such that $p \in U$, the mapping

$$\underline{\phi} : T_p^*M \to (\mathbb{R}^m)^*, \quad [f]_p \mapsto D\left(f \circ \phi^{-1}\right) \in \mathbb{R}^{1\times m}$$

is a linear *isomorphism.* For each i, the (smooth) function

$$f_i : U \to \mathbb{R}, \quad x \mapsto f_i(x) := \phi_i(x) - \phi_i(p)$$

is an element of $\mathbf{F}(p)$ and $\underline{\phi}\,([f_i]_p) = \begin{bmatrix}\delta_{i1} & \cdots & \delta_{im}\end{bmatrix} \in \mathbb{R}^{1\times m}$. So $[f_1]_p, \cdots, [f_m]_p$ form a basis for (the vector space) T_p^*M.

NOTE : The linear structure of T_p^*M is *canonical.* Indeed, let (U, ϕ) and (V, ψ) be two admissible charts at p which produce their own bases $[f_1]_p, \ldots, [f_m]_p$ and $[g_1]_p, \ldots, [g_m]_p$, respectively. Let $[f]_p$ be an arbitrary element of T_p^*M. Then

$$\begin{aligned}[f]_p &= v_1[f_1]_p + \cdots + v_m[f_m]_p \\ &= w_1[g_1]_p + \cdots + w_m[g_m]_p.\end{aligned}$$

It follows that the coordinates $(w_1, \ldots, w_m)$ are related to the coordinates $(v_1, \ldots, v_m)$ via the following formula

$$v_i = \frac{\partial y_1}{\partial x_i} w_1 + \cdots + \frac{\partial y_m}{\partial x_i} w_m$$

where $y_i = y_i(x_1, \ldots, x_m)$, $i = 1, 2, \ldots, m$ denote the coordinate functions of the mapping $\psi \circ \phi^{-1}$. Hence the vector structure of T_p^*M is independent of the particular choice of admissible chart.

We shall show now the *duality* between the elements of T_pM and those of T_p^*M. For any $f \in \mathbf{F}(p)$ and any $\sigma \in \mathbf{C}(p)$, consider the *pairing*

$$\langle [f]_p,\, [\sigma]_p \rangle := \frac{d}{dt} f \circ \sigma \Big|_{t=0}.$$

Because $f \circ \sigma = f \circ \phi^{-1} \circ \phi \circ \sigma$, it follows that the foregoing pairing is well defined and is *bilinear.* More explicitly,

$$\langle [f]_p,\, [\sigma]_p \rangle = \frac{\partial f}{\partial x_1} \frac{d\sigma_1}{dt} + \cdots + \frac{\partial f}{\partial x_m} \frac{d\sigma_m}{dt}$$

with

$$D\left(f\circ\phi^{-1}\right)=\left[\begin{array}{ccc}\frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_m}\end{array}\right] \quad \text{and} \quad \left.\frac{d}{dt}\phi\circ\sigma(t)\right|_{t=0}=\left[\begin{array}{c}\frac{d\sigma_1}{dt}\\ \vdots \\ \frac{d\sigma_m}{dt}\end{array}\right].$$

Therefore, *each element of* T_p^*M *is a linear functional on* T_pM, and hence

$$T_p^*M=(T_pM)^*\,.$$

NOTE : It is useful to think of tangent vectors as objects that act (linearly) on functions and produce *directional derivatives*. Let M be a smooth manifold and let $\mathbf{F}(p)$ be the *algebra* of function germs at $p\in M$. A *linear functional* $X_p:\mathbf{F}(p)\to\mathbb{R}$ is called a **derivation** at p if (for every $f,g\in\mathbf{F}(p)$)

$$X_p(f\cdot g)=f(p)\cdot X_p(g)+g(p)\cdot X_p(f) \qquad \text{(Leibniz rule).}$$

If $f=1$ (i.e. $f(x)=1$ for all $x\in M$), then $X_p(f)=2X_p(f)$, and therefore $X_p(f)=0$. Thus any derivation of a constant function is zero. It is easy to check that the set of all derivations at p is in fact a vector space (over $\mathbb{R}$). Moreover, this vector space is *isomorphic* to (the tangent space) T_pM. (In general, for manifolds that are *not* smooth, the space of derivations is an infinite dimensional vector space and so cannot be isomorphic to T_pM.)

For each $[\alpha]_p\in T_pM$ and $f\in\mathbf{F}(p)$, let

$$\langle f,\,[\alpha]_p\rangle=\left.\frac{d}{dt}f\circ\alpha(t)\right|_{t=0}.$$

Such action is well defined, for if $\alpha\sim_p\bar{\alpha}$, then

$$\begin{aligned}\left.\frac{d}{dt}f\circ\bar{\alpha}(t)\right|_{t=0} &= \left.\frac{d}{dt}f\circ\phi^{-1}\circ\phi\circ\bar{\alpha}(t)\right|_{t=0}\\ &= \left.\frac{d}{dt}(f\circ\phi^{-1})\circ\phi\circ\alpha(t)\right|_{t=0}\\ &= \left.\frac{d}{dt}f\circ\alpha(t)\right|_{t=0}.\end{aligned}$$

$[\alpha]_p$ acts linearly on $\mathbf{F}(p)$ and it follows that such an operation is a *derivation*. Let $D_{[\alpha]_p}$ denote the derivation (at p) induced by the foregoing pairing. It can be shown that *for each derivation* X_p *at* p*, there exists an element (infinitesimal curve)* $[\alpha]_p$

in T_pM *such that* $X_p = D_{[\alpha]_p}$. (Given a fixed admissible chart (U, ϕ) at p, consider the curves

$$\alpha_i : t \mapsto \alpha_i(t) := \phi^{-1}\left(\phi(p) + te_i\right), \qquad i = 1, 2, \ldots, m.$$

Then $[\alpha_1], \cdots, [\alpha_m]_p$ form a basis for T_pM and $X_p = a_1 D_{[\alpha_1]_p} + \cdots + a_m D_{[\alpha_m]_p}$ for some numbers $a_1, \ldots, a_m$.)

Following the usual practice, we shall write $\left.\frac{\partial}{\partial x_i}\right|_p$ for $D_{[\alpha_i]_p}$. Then $\left.\frac{\partial}{\partial x_1}\right|_p, \cdots, \left.\frac{\partial}{\partial x_m}\right|_p$ is a basis for the (vector space of) derivations at p, and each derivation is an expression of the form

$$a_1 \left.\frac{\partial}{\partial x_1}\right|_p + \cdots + a_m \left.\frac{\partial}{\partial x_m}\right|_p .$$

We shall find it convenient to use two notations for the tangent vectors at p, each of which is suggestive in its own way. If we think of T_pM as the set (vector space) of equivalence classes of curves at p, then we shall denote its elements by

$$\left.\frac{d\alpha}{dt}\right|_{t=0}$$

and if we think of T_pM as the (vector) space of derivations at p, then we shall denote its elements as

$$a_1 \left.\frac{\partial}{\partial x_1}\right|_p + \cdots + a_m \left.\frac{\partial}{\partial x_m}\right|_p$$

the meaning being that

$$\left.\frac{d\alpha}{dt}\right|_{t=0} = a_1 \left.\frac{\partial}{\partial x_1}\right|_p + \cdots + a_m \left.\frac{\partial}{\partial x_m}\right|_p \quad \Longleftrightarrow \quad D_{[\alpha]_p} = a_1 D_{[\alpha_1]_p} + \cdots + a_m D_{[\alpha_m]_p}.$$

We shall adopt a similar convention with the elements of (the cotangent space) T_p^*M: $(df)_p$ is the equivalence class of f in T_p^*M, with the understanding that

$$(df)_p \cdot \left.\frac{d\alpha}{dt}\right|_{t=0} = \langle [f]_p, [\alpha]_p \rangle = \left.\frac{d}{dt} f \circ \alpha(t)\right|_{t=0} .$$

In particular, then $(dx_1)_p, \cdots, (dx_m)_p$ denotes the *dual basis* of $\left.\frac{\partial}{\partial x_1}\right|_p, \cdots, \left.\frac{\partial}{\partial x_m}\right|_p$.

NOTE : The definition of the tangent space T_pM uses only (the algebra) $\mathbf{F}(p)$, not all M; thus if U is any open subset of M containing p, then T_pU and T_pM are

naturally identified. Also, recall that $T_p\mathbb{E}^m = \{p\} \times \mathbb{E}^m$ is commonly identified with (the vector space) $\mathbb{R}^m$. We can write, for $U \subseteq \mathbb{E}^m$ (open),

$$T_pU = T_p\mathbb{E}^m = \{p\} \times \mathbb{E}^m = \mathbb{R}^m.$$

◇ **Exercise 263** Let $U \subseteq \mathbb{E}^m$ be open and let $f : U \to \mathbb{R}$ be a smooth function. Compare $Df(p)$ and $(df)_p$ for $p \in U$.

Tangent mappings (differentials)

For every smooth mapping $F : \mathbb{E}^m \to \mathbb{E}^n$ between Euclidean spaces and any point $p \in \mathbb{E}^m$, the derivative of F at p is a linear mapping $DF(p) : T_p\mathbb{E}^m = \mathbb{R}^m \to T_{F(p)}\mathbb{E}^n = \mathbb{R}^n$. Now that we have tangent spaces to manifolds, we are ready to associate analogous (linear) mappings (between tangent spaces) to smooth mappings (between manifolds).

Let M and N be smooth manifolds, and $\Phi : M \to N$ a smooth mapping. We have already mentioned that Φ *pulls back* smooth functions on N into smooth functions on M. However, for smooth curves the situation is different : for any smooth curve σ on M, $\Phi \circ \sigma$ is a smooth curve on N. Thus Φ *pushes forward* curves on M into curves on N. We shall write $\Phi_*\sigma$ for the curve $\Phi \circ \sigma$. Both the push-forward Φ_* and the pull-back Φ^* induce linear mappings between tangent spaces and cotangent spaces, respectively.

5.3.3 Definition. Suppose $\Phi : M \to N$ is a smooth mapping between manifolds and $p \in M$. The **tangent mapping** $\Phi_{*,p} : T_pM \to T_{\Phi(p)}N$ (of Φ at p) is defined by

$$\Phi_{*,p} : [\alpha]_p \mapsto [\Phi_*\alpha]_{\Phi(p)}.$$

◇ **Exercise 264** Show that the tangent mapping $\Phi_{*,p}$ is well-defined and is linear.

It is immediate that if $\Phi : M \to M$ is the identity, then $\Phi_{*,p} : T_pM \to T_pM$ is the identity isomorphism.

◇ **Exercise 265** Suppose that $\Phi : M \to N$ and $\Psi : N \to P$ are smooth mappings between manifolds and $p \in M$. Verify that

$$(\Psi \circ \Phi)_{*,p} = \Psi_{*,\Phi(p)} \circ \Phi_{*,p}.$$

NOTE : The linear mapping $\Phi_{*,p} : T_pM \to T_{\Phi(p)}N$ is often called the *differential* of Φ at p. One frequently sees other notations for $\Phi_{*,p}$, for example $(d\Phi)_p, \Phi'(p)$, or $T_p\Phi$. The $*$ is a subscript since the mapping is in the same "direction" as Φ (i.e. from M to N).

Recall from linear algebra that every linear mapping $\Phi = \Phi_* : V \to W$ between vector spaces induces a *dual* (linear) mapping $\Phi^* : W^* \to V^*$ by the prescription

$$\begin{aligned} (\Phi^*\lambda)(v) &= \lambda(\Phi_*(v)) \\ &= \lambda \circ \Phi(v) \quad \text{for } v \in V \text{ and } \lambda \in W^* \end{aligned}$$

(or, if one prefers, $\langle \Phi^*(\lambda), v\rangle = \langle \lambda, \Phi_*(v)\rangle$).

NOTE : The definition of Φ^* does *not* require the choice of a basis; therefore Φ^* is *naturally* (or canonically) determined by Φ_*. The vector spaces V and V^* have the same dimension, thus they must be isomorphic. There is no natural isomorphism; however, we do have the following property : *There is a natural isomorphism between V and $(V^*)^*$ given by $v \mapsto \langle \cdot, v\rangle$ (i.e. v is mapped to the linear functional on V^* whose value on any $\lambda \in V^*$ is $\lambda(v) = \langle \lambda, v\rangle$).* Observe that the mapping $(v, \lambda) \mapsto \langle \lambda, v\rangle$ is *bilinear* (i.e. linear in each variable separately). This shows that the dual of V^* is V itself, accounts for the name "dual" space, and validates the use of the symmetric notation $\langle \lambda, v\rangle$ in preference to the functional notation $\lambda(v)$.

We make the following definition.

5.3.4 DEFINITION. Suppose $\Phi : M \to N$ is a smooth mapping between manifolds and $p \in M$. The **cotangent mapping** $\Phi_p^* : T^*_{\Phi(p)}N \to T^*_pM$ (of Φ at p) is the dual of the tangent mapping $\Phi_{*,p} : T_pM \to T_{\Phi(p)}N$ (i.e. $\Phi_p^* = (\Phi_{*,p})^*$).

The cotangent mapping $\Phi_p^* : T^*_{\Phi(p)}N \to T^*_pM$ is defined by

$$\Phi_p^* : [f]_{\Phi(p)} \mapsto [\Phi^* f]_p.$$

NOTE : The foregoing mapping (between cotangent spaces) is well-defined and acts like the dual of the tangent mapping (between tangent spaces).

◇ **Exercise 266** Suppose that $\Phi : M \to N$ and $\Psi : N \to P$ are smooth mappings between manifolds and $p \in M$. Verify that

$$(\Psi \circ \Phi)^*_p = \Psi^*_{\Phi(p)} \circ \Phi^*_p.$$

In terms of the admissible charts (U, ϕ) at $p \in M$ and (V, ψ) at $\Phi(p) \in N$, we have the following formulas. Let

$$v = \frac{d}{dt}\phi \circ \alpha(t)\Big|_{t=0}, \qquad w = \frac{d}{dt}\psi \circ \Phi \circ \alpha(t)\Big|_{t=0}$$

and

$$\phi_i(x_1, \ldots, x_m) = \psi_i \circ \Phi \circ \phi^{-1}(x_1, \ldots, x_m), \qquad i = 1, 2, \ldots, n.$$

Then the *local representation* of the tangent mapping $\Phi_{*,p}$ is

$$w_i = \frac{\partial \Phi_i}{\partial x_1} v_1 + \cdots + \frac{\partial \Phi_i}{\partial x_m} v_m, \qquad i = 1, 2, \ldots, n.$$

We can write (in compact form)

$$\begin{aligned} w = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} &= \begin{bmatrix} \frac{\partial \Phi}{\partial x_1} & \cdots & \frac{\partial \Phi}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial \Phi}{\partial x_1} & \cdots & \frac{\partial \Phi}{\partial x_m} \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} \\ &= \frac{\partial \Phi}{\partial x} v. \end{aligned}$$

In order to get an analogous expression for the cotangent mapping, let f be a smooth function on N at $\Phi(p)$, and g its pull-back $\Phi^* f$. Denote

$$g(x_1, \ldots, x_m) = \Phi^* f \circ \phi^{-1}(x_1, \ldots, x_m) \quad \text{and} \quad f(y_1, \ldots, y_n) = f \circ \psi^{-1}(y_1, \ldots, y_n).$$

Then $g(x_1, \ldots, x_m) = f(\Phi_1(x), \ldots, \Phi_n(x))$, and hence

$$\frac{\partial g}{\partial x_i} = \frac{\partial \Phi_1}{\partial x_i}\frac{\partial f}{\partial y_1} + \cdots + \frac{\partial \Phi_n}{\partial x_i}\frac{\partial f}{\partial y_n}, \qquad i = 1, 2, \ldots, m.$$

Likewise, we can write (in compact form)

$$\frac{\partial g}{\partial x} = \begin{bmatrix} \frac{\partial g}{\partial x_1} & \cdots & \frac{\partial g}{\partial x_m} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial y_1} & \cdots & \frac{\partial f}{\partial y_n} \end{bmatrix} \begin{bmatrix} \frac{\partial \Phi}{\partial x_1} & \cdots & \frac{\partial \Phi}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial \Phi}{\partial x_1} & \cdots & \frac{\partial \Phi}{\partial x_m} \end{bmatrix}$$
$$= \frac{\partial f}{\partial y} \frac{\partial \Phi}{\partial x}.$$

. . . . . .

The tangent bundle and the cotangent bundle

It is natural to assemble all tangent spaces of a (smooth) manifold together into a new structure – and conceivably, this set should again have a natural manifold structure. (We will omit some of the more technical details of this structure.) As discussed earlier, it is desirable to distinguish between tangent vectors at different points.

Let M be a smooth n-manifold, and consider the *set*

$$TM := \left\{ (p, X_p) \in M \times \bigcup_{p \in M} T_pM \,|\, X_p \in T_pM \right\}$$

which is the (disjoint) union of all tangent spaces to M at all points $p \in M$. Let

$$\pi : TM \to M, \qquad (p, X_p) \mapsto p$$

be the *projection* onto M. The *fibre* over $\in M$ is the preimage $\pi^{-1}(p) = \{p\} \times T_pM$. (Occasionally it is convenient to identify the fibre $\pi^{-1}(p)$ with the tangent space T_pM – technically, this includes a tacit projection onto the second factor.) We call TM the **tangent bundle** of M.

NOTE : To illustrate the natural manifold structure of tangent bundles, consider the example of $M = \mathbb{S}^1$. The *naive* collection of all *tangent lines* to the (embedded) circle $\mathbb{S}^1 \subseteq \mathbb{E}^2$ is full of intersections. More suitable for our purposes is to embed the circle in $\mathbb{E}^3$ as $\{x \in \mathbb{E}^3 \,|\, x_1^2 + x_2^2 = 1,\ x_3 = 0\}$ and *attach* at every point $p \in \mathbb{S}^1$ a *vertical* line, yielding a cylinder. As a set, this cylinder is in bijection with the (disjoint) collection of all *tangent lines* to the circle (embedded in the plane). It is clear that one can

consistently choose an orientation of the lines (and even more a consistent scaling). Intuitively, identify the *naive tangent vector* $((\cos\theta, \sin\theta), (-L\sin\theta, L\cos\theta))$ with the point $(\cos\theta, \sin\theta, L) \in \mathbb{E}^3$.

In complete analogy, we may intuitively think of the tangent bundle $T\mathbb{R}$ of the real line $\mathbb{R}$ as $\mathbb{R}^2$. However, for dimensional reasons it is clear that these two examples are the only tangent bundles amenable to such immediate visualization. How quickly things get complicated becomes clear if one tries to think of $T\mathbb{S}^2$ as a sphere with a (different) planes attached to each of its points. A *vector field* on the sphere simply selects one point on each plane. However, from algebraic topology it is known that *there does not exist any continuous vector field on the sphere that vanish nowhere.* In our picture this means that it is impossible to continuously select one point on each *tangent plane* avoiding the origin (zero vector) in each $T_p\mathbb{S}^2$. Intuitively, $T\mathbb{S}^2$ must be nontrivially twisted (when compared to e.g. $T\mathbb{S}^1$ which is the very tame cylinder) and hence must be very different from the *trivial* Cartesian product $\mathbb{S}^2 \times \mathbb{R}^2$.

The set TM has a canonical (smooth) manifold structure of dimension $2n$.

NOTE : The key idea is that locally, above an admissible chart (U, ϕ), the tangent bundle "looks like" $\mathbb{R}^m \times \mathbb{R}^m = \mathbb{R}^{2m}$. This observation is captured in the concept of *local triviality* (compare the later short note on vector bundles). Thus the topology and geometry of M are captured, in the global structure of the tangent bundle, by how the trivial bundles are pieced together with *twists*.

Starting with a (smooth) atlas on M, we shall find it easy to obtain a candidate (smooth) atlas on TM. This can be done as follows. Let $(U_\alpha, \phi_\alpha) \in \{(U_\alpha, \phi_\alpha)\}_{\alpha\in\mathfrak{A}}$ be an admissible chart on M with $p \in U_\alpha$, and consider the set

$$\begin{aligned} TU_\alpha &:= \pi^{-1}(U_\alpha) \\ &= \text{the (disjoint) union of all } T_xM \text{ with } x \in U_\alpha. \end{aligned}$$

To any element $(p, v) \in TU_\alpha \subseteq TM$, where $v = X_p \in T_pM$, we associate the point

$$\big(\phi_\alpha(p), \bar{\phi}_\alpha(v)\big) \in \phi_\alpha(U_a) \times \mathbb{R}^m \subseteq \mathbb{R}^{2m}$$

where $\bar{\phi}_\alpha : T_pM \to \mathbb{R}^m$ is the linear isomorphism asoociated with (U_α, ϕ_α) at p. The mapping

$$T\phi_\alpha : TU_\alpha \to \mathbb{R}^{2m}, \qquad (p, v) \mapsto \big(\phi_\alpha(p), \bar{\phi}_\alpha(v)\big)$$

is one-to-one and onto an open subset $\phi_\alpha(U_\alpha) \times \mathbb{R}^m$ of $\mathbb{R}^{2m}$. We claim that *the family (of charts)* $\{(TU_\alpha, T\phi_\alpha)\}_{\alpha \in \mathfrak{A}}$ *is a smooth atlas on* TM, determining a smooth structure.

◇ **Exercise 267** Verify the preceding statement.

The induced canonical topology on TM is such that all the coordinate mappings $T\phi_\alpha : TU_\alpha \to T\phi_\alpha(TU_\alpha) \subseteq \mathbb{R}^{2m}$ are homeomorphisms (in fact, it is the weakest topology on TM with this property).

NOTE : Alternatively, the canonical topology on the tangent bundle TM can be characterized as the strongest topology under which the projection mapping $\pi : TM \to M$ is continuous.

Recall that, in order for TM to qualify as a smooth manifold, we still need that the (canonical) topology is reasonably *nice* – Hausdorff and second countable. (It is easy to see that the topology is Hausdorff; however, the second condition is rather tricky, and we shall skip the details.)

◇ **Exercise 268** Show that (as a mapping between smooth manifolds) the projection mapping $\pi : TM \to M$ is smooth.

5.3.5 EXAMPLE. If an m-dimensional vector space V is regarded as a (smooth) manifold (see EXAMPLE 4.2.7), then the tangent bundle TV is *isomorphic* to $V \times V$.

NOTE : It is often convenient to replace $\phi_\alpha(U_\alpha) \times \mathbb{R}^m$ with $U_\alpha \times \mathbb{R}^m$, identifying $T\phi_\alpha$ with the mapping $v \mapsto (p, \bar{\phi}_\alpha(v))$. (This minor abuse of notation turns out to be a major convenience.) For each $\alpha \in \mathfrak{A}$, we get a commutative diagram

$$\begin{array}{ccc} TU_\alpha & \xrightarrow{T\phi_\alpha} & U_\alpha \times \mathbb{R}^m \\ {\scriptstyle \pi}\downarrow & & \downarrow{\scriptstyle pr_1} \\ U_\alpha & \xrightarrow[id]{} & U_\alpha \end{array}$$

where pr_1 denotes projection on the first factor and $T\phi_\alpha$ is a diffeomorphism that restricts to be a linear isomorphism $T_pM \to \{p\} \times \mathbb{R}^m$ for every $p \in U_\alpha$. Thus, TM *is "locally" a Cartesian product of* M *and* $\mathbb{R}^m$*, the projection* π *being "locally" the projection of the Cartesian product onto the first factor, and the fiber* $\pi^{-1}(p) = T_pM$ *has a canonical vector space structure, for every* $p \in M$.

Tangent bundles are examples of vector bundles. (Vector bundles play a very important role in manifold theory.)

NOTE : Let M be a smooth m-manifold, E a smooth $(m+k)$-dimensional manifold, and $\pi : E \to M$ a smooth mapping. The triple (E, M, π) is called a **vector bundle** over M (of *fibre dimension* k) if the following properties hold.

(VB1) For each $p \in M$, the *fibre* $E_p := \pi^{-1}(p)$ has the structure of a (real) k-dimensional vector space.

(VB2) For each $p \in M$, there exist an open neighborhood W and a (smooth) diffeomorphism $\zeta : \pi^{-1}(W) \to W \times \mathbb{R}^k$ such the following diagram commutes

$$\begin{array}{ccc} \pi^{-1}(W) & \xrightarrow{\zeta} & W \times \mathbb{R}^k \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle pr_1} \\ W & \xrightarrow[id]{} & W \end{array}$$

(Any such pair $(\pi^{-1}(W), \zeta)$ is called a (vector) *bundle chart* on (E, M, π).)

(VB3) For each $p \in W$, the restriction

$$\zeta_p = \zeta|_{E_p} : E_p \to \{p\} \times \mathbb{R}^k$$

is a linear isomorphism.

We call E the *total space*, M the *base space*, and π the *bundle projection*. We shall denote a vector bundle (over M), simply, $\pi : E \to M$. An obvious example of a vector bundle is given by $pr_1 : M \times \mathbb{R}^k \to M$. Here $(M \times \mathbb{R}^k, id)$ is a global bundle chart and the vector bundle is said to be *trivial*.

Given two vector bundles $\pi_1 : E_1 \to M$ and $\pi_2 : E_2 \to M$ over M, a (vector) *bundle isomorphism* is a commutative diagram

$$\begin{array}{ccc} E_1 & \xrightarrow{\varphi} & E_2 \\ {\scriptstyle\pi_1}\downarrow & & \downarrow{\scriptstyle\pi_2} \\ M & \xrightarrow[id]{} & M \end{array}$$

such that φ is a (smooth) diffeomorphism, and carries E_{1p} isomorphically (as a vector space) onto E_{2p}, for every $p \in M$.

So far, the only real eaxamples of vector bundles that we have seen are the tangent bundles and trivial bundles. The following is the least complicated example of a nontrivial vector bundle. We give an example of a "line" bundle (i.e. of fibre dimension 1) over the circle, known as the *Möbius bundle*. On $\mathbb{R} \times \mathbb{R}$, define the equivalence relation $(s,t) \sim (s+n,(-1)^n t)$, $n \in \mathbb{Z}$. Observe that $t \mapsto (-1)^n t$ is a linear automorphism of $\mathbb{R}$. The projection $(s,t) \mapsto s$ passes to a well defined mapping $\pi : (\mathbb{R}\times\mathbb{R})/_\sim \to \mathbb{R}/\mathbb{Z} = \mathbb{S}^1$. It should be clear, intuitively, that this is a vector bundle over $\mathbb{S}^1$ of fibre dimension 1, but a rigorous proof of this involves checking many details.

Beyond vector bundles are *fibre bundles* in which the fibres need not necessarily be vector spaces. Arguably the most important such fibre bundle is the *principal bundle* in which each fibre is a copy of the general linear group $\mathsf{GL}(k,\mathbb{R})$. (Differential geometry may be described as *the study of a connection on a principal bundle.*)

In complete analogy to the tangent bundle we assemble all cotangent spaces T_p^*M into the **cotangent bundle**, denoted T^*M. It is a vector bundle (of fibre dimension m) over the (smooth) m-manifold M with bundle projection again denoted by π. The set (total space)

$$T^*M := \left\{ (p,\omega_p) \in M \times \bigcup_{\in M} T_p^*M \,|\, \omega_p \in T_p^*M \right\}$$

has a (smooth) manifold structure of dimension $2m$, given by the (smooth) atlas $\{(T^*U_\alpha, T^*\phi_\alpha)\}_{\alpha\in\mathfrak{A}}$ where $T^*U_\alpha = \pi^{-1}(U_\alpha) \subseteq T^*M$ and

$$T^*\phi_\alpha : (p,\omega) \mapsto (\phi_\alpha(p), \underline{\phi}_\alpha(\omega)) \in \phi_\alpha(U_\alpha) \times (\mathbb{R}^m)^* \subseteq \mathbb{R}^{2m}$$

($\{(U_\alpha,\phi_\alpha)\}_{\alpha\in\mathfrak{A}}$ is an atlas on M).

5.3.6 Example. If an m-dimensional vector space V is regarded as a (smooth) manifold, then the cotangent bundle T^*V is *isomorphic* to $V \times V^*$.

Note : One can show that the tangent and cotangent bundles are isomorphic, but not canonically. We *do not* identify these (vector) bundles.

5.4 Smooth Submanifolds

. . . . . .

5.5 Vector Fields

Vector fields and flows

Let M be a manifold and let TM be the corresponding tangent bundle.

5.5.1 DEFINITION. A **vector field** X (on M) is a mapping from M into TM such that for each $p \in M$ the natural projection $\pi : TM \to M$ projects $X(p)$ to p (i.e. the compositon $\pi \circ X$ is the identity on M).

Rather than considering arbitrary such mappings, our interest is primarily in those that vary smoothly. (In topological considerations continuity may suffice.) Since a vector field is defined as a mapping between (smooth) manifolds M and TM, we already have a notion of smoothness : We say that X is a *smooth* vector field provided $X : M \to TM$ is a smooth mapping. We shall write $\mathfrak{X}(M)$ for the set of all smooth vector fields on M.

NOTE : A *section* of the vector bundle $\pi : E \to M$ is a smooth mapping $s : M \to E$ such that $\pi \circ s = id_M$. The set of all such sections is denoted by $\Gamma^\infty(E)$. It is easy to verify that *the set* $\Gamma^\infty(E)$ *is a* $C^\infty(M)$*-module* under the natural (pointwise) operations of addition and (function) multiplication.

Thus, $\mathfrak{X}(M) = \Gamma^\infty(TM)$. In complete analogy (to the tangent bundle), the $C^\infty(M)$-module of all smooth *covector fields* on M is

$$A^1(M) := \Gamma^\infty(T^*M).$$

Covector fields are also (and more commonly) called **differential 1-forms**. If $\omega \in A^1(M)$, then $\omega : M \to T^*M$ is writen as

$$p \mapsto \omega_p \in T_p^*M.$$

NOTE : Having defined these objects in an "intrinsic" way, let us now examine their meaning in a more intuitive way. It is well known in physics that the position

of a particle is a scalar-like quantity, and its velocity is a vector quantity. Therefore, if $t \mapsto p(t)$ is a curve that describes the position, then the velocity $\frac{dp}{dt}$ is a different object, since it is a vector. These two objects "live" in different spaces. The manifolds formalism clarifies this issue, and it provides a natural point of view from which differential equations (systems) should be studied.

If $\dot{x} = F(x)$ is a differential equation in $\mathbb{R}^m$, then F *cannot* be viewed as a mapping from $\mathbb{R}^m$ into $\mathbb{R}^m$. Rather, it must be viewed as a mapping form $\mathbb{R}^m$ (in fact, $\mathbb{E}^m$) into the tangent bundle of $\mathbb{R}^m$, since $F(x(t))$ is equal to the tangent vector of the curve $x(\cdot)$ at $x(t)$. For equations in $\mathbb{R}^m$, it is easy to confuse mappings and vector fields (in much the same way as it is to confuse vectors with their duals). It is only on arbitrary manifolds that the genuine differences of these objects become apparent.

Each vector field $X \in \mathfrak{X}(M)$ in some admissible chart (U, ϕ) becomes an expression of the form

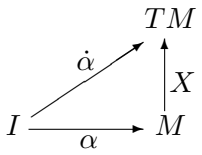
$$X_1(x_1, \ldots, x_m)\frac{\partial}{\partial x_1} + X_2(x_1, \ldots, x_m)\frac{\partial}{\partial x_2} + \cdots + X_m(x_1, \ldots, x_m)\frac{\partial}{\partial x_m}.$$

The (smooth) functions $X_1, \ldots, X_m$ are called the *coordinate functions* of the vector field X. (Strictly speaking, X should be expressed in terms of $2m$ coordinates; however, because the first m coordinates contain redundant information, they are suppressed.)

Let $\alpha : J = (a, b) \to M$ be a smooth curve on the manifold M. Then the *tangent vector* to α at $t \in I$ is given by

$$\dot{\alpha}(t) := \alpha_{*,t}\left(\frac{d}{dt}\right) \in T_{\alpha(t)}M.$$

(Thus $\dot{\alpha} : J \to TM$ is a smooth curve in TM, commonly referred to as the *lift* of α.) Let X be a smooth vector field on M. A smooth curve $\alpha : J \to M$ is an *integral curve* of X provided the tangent vector to α at each $t \in J$ equals the value of X at $\alpha(t)$ (i.e. $\dot{\alpha}(t) = X(\alpha(t))$ for each $t \in J$). Thus the accompanying diagram



is commutative. (The lift $\dot{\alpha}$ of α coincides with $X \circ \alpha$.)

Let (U, ϕ) be an admissible chart on M and let $\alpha : J \to U \subseteq M$ be a smooth curve as before.

◇ **Exercise 269** Verify that (for $t \in J$)

$$\dot{\alpha}(t) = \frac{dx_1}{dt}(t) \left.\frac{\partial}{\partial x_1}\right|_{\alpha(t)} + \cdots + \frac{dx_m}{dt}(t) \left.\frac{\partial}{\partial x_m}\right|_{\alpha(t)}$$

where $x_i = \phi_i \circ \alpha, \quad i = 1, 2, \ldots, m.$

Then $\dot{\alpha} = X \circ \alpha$ yields

$$\frac{dx_i}{dt} = X_i(x_1, \ldots, x_m), \qquad i = 1, 2, \ldots, m.$$

This system of differential equations admits solution curves in the open set $\phi(U)$. That is, through each point x_0 in $\phi(U)$ there exists a solution curve $x(\cdot) : J_0 \to \phi(U) \subseteq \mathbb{R}^m$ that passes through x_0 at $t = 0$ (i.e. $x(0) = x_0$). Any two such solutions curves agree for values of t for which they are both defined. It follows from the theory of differential equations that *for each* x_0 *there exist a maximum open interval* J_{max} *(that contains* 0*) and a unique solution curve* $x(\cdot) : J_{max} \to \mathbb{R}^m$ *such that* $x(0) = x_0$. We shall refer to such a solution curve as *the* solution curve through x_0.

Any solution curve $x(\cdot)$ in $\phi(U)$ defines an integral curve

$$t \mapsto p(t) = \phi^{-1}(x_1(t), \ldots, x_m(t))$$

on M.

NOTE : Consider another admissible chart (V, ψ) on M such that $p(t_0) \in U \cap V$ for some t_0. We denote by $(y_1, \ldots, y_m)$ the coordinates on V, and by $Y_1, \ldots, Y_m$ the coordinate functions of X relative to (V, ψ). The curve $t \mapsto y(t) = \psi \circ p(t)$ is a (smooth) curve in $\psi(V)$ defined in some neighborhood of t_0. Furthermore, $y(t) = \psi \circ \phi^{-1}(x(t))$ and

$$\begin{aligned} \frac{dy_i}{dt} &= \frac{\partial y_i}{\partial x_1}(x(t))\frac{dx_1}{dt} + \cdots + \frac{\partial y_i}{\partial x_m}(x(t))\frac{dx_m}{dt} \\ &= \frac{\partial y_i}{\partial x_1}(x(t))X_1(x(t)) + \cdots + \frac{\partial y_i}{\partial x_m}(x(t))X_m(x(t)). \end{aligned}$$

Because $(Y_1, \ldots, Y_m)$ and $(X_1, \ldots, X_m)$ are the coordinates of the same tangent vector $X(p)$, they are related through

$$Y_i(y) = \frac{\partial y_i}{\partial x_1} X_1(x) + \cdots + \frac{\partial y_i}{\partial x_m} X_m(x), \quad i = 1, 2, \ldots, m.$$

Therefore, $y(\cdot)$ is a solution curve of the system of differential equations

$$\frac{dy_i}{dt} = Y_i(y_1, \ldots, y_m), \qquad i = 1, 2, \ldots, m.$$

Let $\bar{y}(\cdot)$ be the solution curve of this differential system in $\psi(V)$ that passes through $y_0 = \psi \circ p(t_0)$ at $t = t_0$, and denote $\bar{p}(t) = \psi^{-1} \circ \bar{y}(t)$. It then follows that the two integral curves $p(\cdot)$ and $\bar{p}(\cdot)$ on M agree at all values of t for which they are both defined.

5.5.2 DEFINITION. We say that an integral curve $\gamma = \gamma_p$ of $X \in \mathfrak{X}(M)$ is *the* **integral curve** through $p \in M$ provided $\gamma_p(0) = p$ and the domain $J_p \subseteq \mathbb{R}$ of γ_p is maximal.

That is, if α is any integral curve of X that satisfies $\alpha(0) = p$, then its domain can be extended to J_p so that $\alpha(t) = \gamma_p(t)$ for all t.

A (smooth) vector field X is called *complete* if the integral curves γ_p through each point $p \in M$ are defined for *all* values of $t \in \mathbb{R}$. In such case, X is said to define a *flow* $\Phi = \Phi^X$ on M.

NOTE : A *flow* on M is a smooth mapping $\Phi : \mathbb{R} \times M \to M$ such that (for all $t_1, t_2 \in \mathbb{R}$ and all $p \in M$)

(FL1) $\Phi(0, p) = p$.

(FL2) $\Phi(t_1 + t_2, p) = \Phi(t_1, \Phi(t_2, p))$.

(If we fix p and let t vary, we get a smooth curve $\Phi(\cdot, p)$ in M; thus as t varies each point of M moves smoothly inside M, and various points move in a coherent fashion, so that we can form a mental picture of them "flowing" through M, each point along its individual path.) For each $t \in \mathbb{R}$, the (smooth) mapping

$$\varphi_t : M \to M, \qquad p \mapsto \Phi(t, p)$$

is a smooth diffeomorphism of M. We have $\varphi_0 = id_M$ and (for all $t_1, t_2 \in \mathbb{R}$)

$$\varphi_{t_1 + t_2} = \varphi_{t_1} \circ \varphi_{t_2}.$$

Hence the collection $\{\varphi_t \,|\, t \in \mathbb{R}\}$ forms a group under the composition of mappings. Such a group is called a *one-parameter group of diffeomorphisms* of M (or a smooth *action* of $\mathbb{R}$ on M) and is denoted by $\{\varphi_t\}$ or, simply, by φ_t.

The flow Φ^X (generated by the complete vector field X) is defined by

$$\Phi^X(t,p) := \gamma_p(t).$$

We shall also use $\exp tX$ to denote the mapping (diffeomorphism) $\varphi_t = \varphi_t^X$. (Each notation is fairly standard, and each has different merits, depending on the context.)

Each (smooth) flow Φ on M is generated by a vector field X, called the *infinitesimal generator* of Φ. The relation between X and Φ is given by

$$X(p) = \left.\frac{d}{dt}\Phi(t,p)\right|_{t=0}.$$

($X(p) \in T_pM$ is the value of the lift of $\Phi(\cdot,p) : \mathbb{R} \to M$ at $t = 0$.) Therefore, *there is a one-to-one correspondence between complete vector fields and flows.*

NOTE : The *support* of a vector field X is the closure of the set $\{p \in M \,|\, X(p) \neq 0\}$. It can be shown that *every vector field with compact support on M is complete.* So on a compact manifold M, each vector field is complete. If M is not compact and of dimension ≥ 2 the set of complete vector fields is not even a vector space as the following example (on $\mathbb{E}^2$) shows : the vector fields

$$X = x_2\frac{\partial}{\partial x_1} \quad \text{and} \quad Y = \frac{x_1^2}{2}\frac{\partial}{\partial x_2}$$

are complete, but $X + Y$ is not.

◇ **Exercise 270** Show that the (smooth) vector field

$$X = -x_2\frac{\partial}{\partial x_1} + x_1\frac{\partial}{\partial x_2}$$

is complete (on $\mathbb{E}^2$). Is the vector field

$$Y = e^{-x_1}\frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_2}$$

complete ?

◇ **Exercise 271** Consider the (smooth) vector field on $\mathbb{E}^3$ defined by

$$X = x_2\frac{\partial}{\partial x_1} + x_3\frac{\partial}{\partial x_2} + x_1\frac{\partial}{\partial x_3}\cdot$$

Find the integral curve γ of X so that $\gamma(0) = (-1, 1, 1)$.

We have seen that not every vector field is complete. If this is the case, then $X \in \mathfrak{X}(M)$ generates (only) a *local flow* on M.

5.5.3 EXAMPLE. Let $M = \mathbb{E}^2$ and let (the flow) $\Phi : \mathbb{R} \times M \to M$ be defined by

$$(t, (x_1, x_2)) \mapsto (x_1 + t, x_2).$$

Then the infinitesimal generator is $X = \frac{\partial}{\partial x_1}\cdot$ Suppose now that we remove the origin $(0, 0)$ from $\mathbb{E}^2$; let $M_0 = \mathbb{E}^2 \setminus \{(0, 0)\}$. For most points (the diffeomorphism) φ_t is defined as before; however, we cannot obtain an action of $\mathbb{R}$ on M_0 by restriction of Φ to $\mathbb{R} \times M_0$ since points of the (closed) set

$$\{(t, (x_1, 0)) \,|\, x_1 + t = 0\} = \Phi^{-1}((0, 0)) \subseteq \mathbb{R} \times M$$

are mapped by Φ to the origin. On the other hand, let $W \subseteq \mathbb{R} \times M_0$ be the open set defined by

$$W = \left(\bigcup_{x_2 \neq 0} \mathbb{R} \times \{(x_1, x_2)\}\right) \cup \{(t, (x_1, 0)) \,|\, x_1(x_1 + t) > 0\}.$$

Then $\Phi = \Phi|_W$ maps W onto M_0 and preserves many of the features of Φ which we have used. For example, let $p = (x_1, x_2) \in M_0$. Then

- $(0, p) \in W$ and $\Phi(0, p) = p$
- $\Phi(t_1, \Phi(t_2, p)) = \Phi(t_1 + t_2, p)$

if all terms are defined, and the infinitesimal generator is again $X = \frac{\partial}{\partial x_1}\cdot$ Finally, we have *orbits* $t \mapsto \Phi(t, p)$, which are the lines $x_2 =$ constant (as before) when $p = (x_1, x_2)$, $x_2 \neq 0$, and for $p = (x_1, 0)$ the portion of the x_1-axis minus the origin which contains p. This curve is not defined for all values of t in the case of the orbit of a point on the x_1-axis.

NOTE : In order to define the local flow of a vector field at $p \in M$, it is first necessary to define the escape times of the integral curve γ_p of X through p. The *positive escape time* $e^+(p)$ is defined to be the supremum of t such that an integral curve passing through p can be defined at t. The *negative escape time* $e^-(p)$ is defined similarly. Let $W := \{(t,p) \,|\, e^-(p) < t, e^+(p)\}$. Then W is an open subset of $\mathbb{R} \times M$ and a neighborhood of $\{0\} \times M$. The *local flow* Φ of X is defined on W and it satisfies the following :

- The mapping $\Phi : W \subseteq \mathbb{R} \times M \to M$ is smooth
- $\Phi(0,p) = p$ for all $p \in W$.
- $\Phi(t_1 + t_2, p) = \Phi(t_1, \Phi(t_2, p))$ whenever each of (t_1, p) and $(t_1, \Phi(t_2, p))$ is contained in W.
- $\frac{d\Phi}{dt}(t,p) = X \circ \Phi(t,p)$.

5.5.4 EXAMPLE. Let $M = \mathbb{E}^m$, and let

$$X : x \mapsto a = \begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix} \in \mathbb{R}^m \ \left(= \mathbb{R}^{m \times 1}\right)$$

be a *constant* (or parallel) vector field on M. Then (in the "derivation notation")

$$X(x) = a_1 \left.\frac{\partial}{\partial x_1}\right|_x + \cdots + a_m \left.\frac{\partial}{\partial x_m}\right|_x .$$

The integral curves of X are parallel lines, all in the direction of a. For each t, the mapping (diffeomorphism) $\varphi_t : t \mapsto \Phi(t,x)$ is a translation of x by ta. Hence $\{\varphi_t\}$ is a one-parameter group of translations on $\mathbb{E}^m$.

5.5.5 EXAMPLE. Let $M = \mathbb{E}^m$ and $A \in \mathbb{R}^{m \times m}$. Let

$$X : x = (x_1, \ldots, x_m) \mapsto Ax := \begin{bmatrix} a_{11}x_1 + \cdots + a_{1m}x_m \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mm}x_m \end{bmatrix} \in \mathbb{R}^m \ \left(= \mathbb{R}^{m \times 1}\right)$$

be a *linear* vector field on M. So

$$X(x_1, \ldots, x_m) = X_1(x_1, \ldots, x_m) \left.\frac{\partial}{\partial x_1}\right|_x + \cdots + X_m(x_1, \ldots, x_m) \left.\frac{\partial}{\partial x_m}\right|_x$$

with coordinate functions given by

$$X_i(x_1, \ldots, x_m) = a_{i1}x_1 + \cdots + a_{im}x_m, \qquad i = 1, 2, \ldots, m.$$

Each integral curve of X is of the form $t \mapsto \exp(tA)x$, where $\exp(tA) = \sum_{k=0}^{\infty} \frac{t^k}{k!}A^k$ (the matrix exponential of tA). Thus $\varphi_t(x) = \exp(tA)x$, and therefore (the one-parameter group of diffeomorphisms) $\{\varphi_t\}$ is a subgroup of the group of all linear transformations on $\mathbb{R}^n$ (i.e. a matrix group). Here are two familiar cases (for $n = 2$):

- $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, $\exp(tA) = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}$. (The one-parameter group $\{\varphi_t\}$ is the rotation group $\mathsf{SO}(2)$, and the integral curves are concentric circles centered at the origin.)
- $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $\exp(tA) = \begin{bmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{bmatrix}$. (The one-parameter group $\{\varphi_t\}$ is a subgroup of $\mathsf{SL}(2, \mathbb{R})$, and the integral curves are hyperbolas.)

5.5.6 Example. Let $M = \mathbb{E}^3$ and consider the vector field (on M)

$$X : x \mapsto X(x) := Ax + a$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad a = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Then

$$\begin{aligned} \Phi(t, x) = \varphi_t(x) &= \exp(tA)x + ta \\ &= \begin{bmatrix} \cos t & \sin t & 0 \\ -\sin t & \cos t & 0 \\ 0 & 0 & 1 \end{bmatrix} x + t \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \end{aligned}$$

Integral curves are helices (with centers along the x_3-axis).

5.5.7 EXAMPLE. Let $M = \mathsf{GL}^+(2,\mathbb{R})$. For $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}$, let X be the vector field on M defined by $p \mapsto Ap$. Then

$$\Phi(t,p) = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} p, \quad p \in M.$$

$\varphi_t(p)$ is the matrix multiplication of $p \in M$ by $\exp(tA)$ from the left for each $t \in \mathbb{R}$.

Vector fields as differential operators

Let M be a manifold and let TM be the corresponding tangent bundle. The algebra of smooth functions on M is denoted by $C^\infty(M)$ (see **Execise 224**).

Recall that tangent vectors act on smooth functions and produce directional derivatives. Specifically, if $X_p = \frac{d\alpha}{dt}\big|_{t=0} \in T_pM$ and $f \in C^\infty(M)$, then

$$X_p f = \frac{d}{dt} f \circ \alpha(t)\Big|_{t=0} \in \mathbb{R}$$

is the directional derivative of f along X_p.

◇ **Exercise 272** Given a mapping $X : M \to TM$, show that the following statements are logically equivalent :

(a) X is smooth (as a mapping between manifolds). In other words, X is a smooth vector field on M.

(b) For each admissible chart (U,ϕ) on M, the coordinate functions $X_i : U \to \mathbb{R}$ of X are smooth.

(c) For each smooth function $f : M \to \mathbb{R}$, the function $x \mapsto X(x)f$ is also smooth.

Smooth vector fields act as derivations on the space of smooth functions. Indeed, let $X \in \mathfrak{X}(M)$ and $f \in C^\infty(M)$. Then Xf will denote the *smooth* function

$$x \mapsto (Xf)(x) := X(x)f.$$

The function Xf is often known as the **Lie derivative** of the function f along the vector field X, and is then denoted $\mathfrak{L}_X f$. In local coordinates, if X is of the form

$$X = X_1 \frac{\partial}{\partial x_1} + \cdots + X_m \frac{\partial}{\partial x_m}$$

then

$$\begin{aligned} \mathfrak{L}_X f &= \frac{\partial f}{\partial x_1} X_1 + \cdots + \frac{\partial f}{\partial x_m} X_m \\ &= \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_m} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix} \\ &= \frac{\partial f}{\partial x} X. \end{aligned}$$

NOTE : One can also define the *Lie derivative* of a function by the formula

$$\mathfrak{L}_X f := \lim_{t \to 0} \frac{\varphi_t^* f - f}{t}$$

where φ_t is the (local) flow of X. (It is then easy to see that $\mathfrak{L}_X f = Xf$.)

◇ **Exercise 273** Given $X \in \mathfrak{X}(M), f, g \in C^\infty(M)$ and $\lambda \in \mathbb{R}$, verify that

(D1) $X(f+g) = Xf + Xg$;

(D2) $X(\lambda f) = \lambda X f$;

(D3) $X(f \cdot g) = f \cdot Xg + g \cdot Xf$.

This shows that the mapping $f \mapsto Xf$ (i.e. the Lie derivative $\mathfrak{L}_X : C^\infty(M) \to C^\infty(M)$) is linear and satisfies the Leibniz rule, hence is a *derivation* of (the ring) $C^\infty(M)$.

NOTE : Derivations of $C^\infty(M)$ are also called *first order differential operators*. The set $\mathfrak{D}(M)$ of all such derivations is a vector space (over $\mathbb{R}$).

We have a natural *inclusion* ($X \mapsto \mathfrak{L}_X$)

$$\mathfrak{X}(M) \subseteq \mathfrak{D}(M)$$

(every smooth vector field is a derivation). One can prove that *all derivations of* $C^\infty(M)$ *are smooth vector fields on* M (i.e. the reverse inclusion $\mathfrak{D}(M) \subseteq \mathfrak{X}(M)$ holds).

NOTE : For this, we need to show that a derivation of $C^\infty(M)$ can be *localized* to a derivation of the algebra $C^\infty(p)$ of function germs at each $p \in M$. (Caution : For $f \in C^\infty(p)$ we do *not* require that $f(p) = 0$. Such elements form a subalgebra $\mathbf{F}(p)$ of $C^\infty(p)$.) This is by no means evident. The "tricky" part is to show that (for $\Delta \in \mathfrak{D}(M)$ and $p \in M$) the mapping

$$\Delta_p : C^\infty(p) \to \mathbb{R}, \quad f \mapsto \Delta(f)(p)$$

is *well-defined* (i.e. depends only on Δ and the function germ $f = \langle f \rangle_p$). Then it follows that

$$\widetilde{\Delta} : p \mapsto \Delta_p \in T_pM$$

is a smooth section of (the tangent bundle) TM, hence a smooth vector field on M.

Henceforth, we shall regard a smooth vector field (on a given manifold) either as a smooth section of the tangent bundle of the manifold or as a derivation of the algebra of smooth functions on that manifold.

The Lie algebra of vector fields

Given a manifold M, the set of all smooth vector fields on M is denoted by $\mathfrak{X}(M)$. It is itself a *vector space* (over $\mathbb{R}$) since any linear combination (with constant coefficients) of two smooth vector fields is also a smooth vector field. More precisely, if $X, Y \in \mathfrak{X}(M)$ and $\lambda, \mu \in \mathbb{R}$, then (for $f \in C^\infty(M)$)

$$\lambda X + \mu Y : f \mapsto (\lambda X + \mu Y)f := \lambda Xf + \mu Yf$$

is a derivation of $C^\infty(M)$, hence a smooth vector field on M.

NOTE : As a vector space, $\mathfrak{X}(M)$ is *not* finite-dimensional. In fact, $\mathfrak{X}(M)$ is more than just a vector space; it is a Lie algebra as we shall see.

Let $X, Y \in \mathfrak{X}(M)$ (viewed as derivations of $C^\infty(M)$). Then, in general, neither YX nor XY is a derivation. However, oddly enough, the operator $YX - XY$ is a derivation (of $C^\infty(M)$).

◇ **Exercise 274** Given $X, Y \in \mathfrak{X}(M)$, verify that the operator $YX - XY : C^\infty(M) \to C^\infty(M)$ is a derivation, hence is (identified with) a smooth vector field on M.

We make the following definition.

5.5.8 DEFINITION. The smooth vector field $[X, Y] \in \mathfrak{X}(M)$, defined by

$$[X, Y]f := Y(Xf) - X(Yf)$$

is called the **Lie bracket** of X and Y.

It is easy to check that the Lie bracket $[\cdot, \cdot] : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$ has the following properties (for $\lambda, \mu \in \mathbb{R}$ and $X, Y, Z \in \mathfrak{X}(M)$) :

(LA1) $[X, Y] = -[Y, X]$;

(LA2) $[X, \lambda Y + \mu Z] = \lambda[X, Y] + \mu[X, Z]$;

(LA3) $[X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] = 0.$

This means that the real vector space $\mathfrak{X}(M)$ equipped with the Lie bracket $[\cdot, \cdot]$ is a *Lie algebra*.

We may now derive the expression in local coordinates for $[X, Y]$. Let

$$X = X_1 \frac{\partial}{\partial x_1} + \cdots + X_m \frac{\partial}{\partial x_m} \quad \text{and} \quad Y = Y_1 \frac{\partial}{\partial x_1} + \cdots + Y_m \frac{\partial}{\partial x_m}$$

be local representations of X and Y, respectively (in an admissible chart (U, ϕ) of M). Then

$$\begin{aligned} [X, Y]f &= Y(Xf) - X(Yf) \\ &= \sum_{i,j=1}^{m} Y_i \frac{\partial X_j}{\partial x_i} \frac{\partial f}{\partial x_j} - \sum_{i,j=1}^{m} X_i \frac{\partial Y_j}{\partial x_i} \frac{\partial f}{\partial x_j} \\ &= \sum_{j=1}^{m} \left(\sum_{i=1}^{m} Y_i \frac{\partial X_j}{\partial x_i} - X_i \frac{\partial Y_j}{\partial x_i} \right) \frac{\partial f}{\partial x_j}. \end{aligned}$$

Thus

$$\begin{aligned} [X, Y] &= \sum_{j=1}^{m} \left(\sum_{i=1}^{m} Y_i \frac{\partial X_j}{\partial x_i} - X_i \frac{\partial Y_j}{\partial x_i} \right) \frac{\partial}{\partial x_j} \\ &= \begin{bmatrix} \frac{\partial X_1}{\partial x_1} & \cdots & \frac{\partial X_1}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial X_m}{\partial x_1} & \cdots & \frac{\partial X_m}{\partial x_m} \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_m \end{bmatrix} - \begin{bmatrix} \frac{\partial Y_1}{\partial x_1} & \cdots & \frac{\partial Y_1}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial Y_m}{\partial x_1} & \cdots & \frac{\partial Y_m}{\partial x_m} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix} \\ &= \frac{\partial X}{\partial x} Y - \frac{\partial Y}{\partial x} X. \end{aligned}$$

5.5.9 EXAMPLE. For *constant* (or parallel) vector fields

$$X : x \mapsto a = \begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix} \quad \text{and} \quad Y : x \mapsto b = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

on $M = \mathbb{E}^m$, we have $[X, Y] = 0$.

5.5.10 EXAMPLE. Let $X, x \mapsto Ax$ be a *linear* vector field and $Y, x \mapsto b$ be a *constant* vector field on $M = \mathbb{E}^m$. Then

$$\begin{aligned} X &= (a_{11}x_1 + \cdots + a_{1m}x_m)\frac{\partial}{\partial x_1} + \cdots + (a_{m1}x_1 + \cdots + a_{mm}x_m)\frac{\partial}{\partial x_m} \\ Y &= b_1\frac{\partial}{\partial x_1} + \cdots + b_m\frac{\partial}{\partial x_m} \end{aligned}$$

and so

$$[X, Y] = \frac{\partial X}{\partial x}Y - \frac{\partial Y}{\partial x}X = Ab - 0 = Ab.$$

Therefore $[X, Y]$ is a constant vector field $x \mapsto c$, with $c = Ab$.

5.5.11 EXAMPLE. If $X, x \mapsto Ax$ and $Y, x \mapsto Bx$ are both *linear* vector fields (on $M = \mathbb{E}^m$), then

$$[X, Y] = \frac{\partial X}{\partial x}Y - \frac{\partial Y}{\partial x}X = ABx - BAx = (AB - BA)x.$$

Therefore $[X, Y]$ is also a linear vector field $x \mapsto Cx$, with $C = [A, B]$ (the *commutator* of the matrices A and B).

We have seen that the set $\mathfrak{X}(M)$ (of all smooth vector fields on M) has a natural structure of Lie algebra. In addition to this structure, $\mathfrak{X}(M)$ admits another algebraic structure : for any $f \in C^\infty(M)$ and any $X \in \mathfrak{X}(M)$,

$$fX : p \mapsto (fX)(p) := f(p)X(p) \in T_pM$$

is a smooth vector field on M. (Caution : do not confuse Xf and fX.) With this operation, $\mathfrak{X}(M)$ becomes a *module* over the ring $C^\infty(M)$.

NOTE : The Lie bracket $[\cdot, \cdot] : C^\infty(M) \times C^\infty(M) \to C^\infty(M)$ is *not* $C^\infty(M)$-bilinear. In fact (for $g \in C^\infty(M)$),

$$[X, gY] = g[X, Y] - (Xg)Y.$$

◇ **Exercise 275** Let $X, Y \in \mathfrak{X}(M)$ and $f, g \in C^\infty(M)$. Show that

$$[fX, gY] = fg[X, Y] - f(Xg)Y + g(Yf)X.$$

Use this formula to derive the formula for the components of $[X, Y]$ in local coordinates.

Commutativity of vector fields

. . . . . .

Orbits of vector fields

. . . . . .

5.6 Differential Forms

. . . . . .

Chapter 6

Lie Groups

Topics :

1. Lie Groups: Definition and Examples
2. Invariant Vector Fields
3. The Exponential Mapping
4. Matrix Groups as Lie Groups
5. Hamiltonian Vector Fields
6. Lie-Poisson Reduction

6.1 Lie Groups: Definition and Examples

Lie groups form an important class of smooth (in fact, *analytic*) manifolds. (Their prototype is any finite-dimensional group of linear transformations on a vector space.) The key idea of a Lie group is that it is a group in the usual sense, but with the additional property that it is also a smooth manifold, and in such a way that the group operations are smooth. A good example is the circle $\mathbb{S}^1 = \{z \in \mathbb{C} \,|\, |z| = 1\}$.

Lie groups (and their Lie algebras) play a central role in geometry, topology, and analysis, as well as in modern theoretical physics. The precise definition is given below.

6.1.1 DEFINITION. A (real) **Lie group** is a smooth manifold G which is also a group such that the operations

$$G \times G \to G, \quad (g_1, g_2) \mapsto g_1 g_2 \qquad \text{and} \qquad G \to G, \quad g \mapsto g^{-1}$$

are smooth mapings.

6.1.2 EXAMPLE. The vector space $\mathbb{R}^m$, when equipped with its natural smooth structure (i.e., viewed as the Euclidean space $\mathbb{R}^m$ in the broad sense), is an m-dimensional (Abelian) Lie group.

6.1.3 EXAMPLE. The *general linear group* $\mathsf{GL}(n, \mathbb{R})$ is evidently a Lie group. It is an open subset of (the vector space) $\mathbb{R}^{n \times n}$ (and hence a smooth submanifold of $\mathbb{R}^{n^2}$) and the group operations are given by rational functions of the coordinates.

NOTE : Let V be an n-dimensional vector space (over $\mathbb{R}$). Then the group $\mathsf{GL}(V)$ of all linear transformations on V is an n^2-manifold. Any choice of a basis in V induces a linear isomorphism from $\mathsf{GL}(V)$ onto $\mathsf{GL}(n, \mathbb{R}) \subseteq \mathbb{R}^{n^2}$ (an hence a global chart on $\mathsf{GL}(V)$). The coordinates of any product (composition) ST of elements in $\mathsf{GL}(V)$ are polynomial expressions of the coordinates of S and T, and the coordinates of S^{-1} are rational functions of the coordinates of S. It therefore follows that both group operations $(S, T) \mapsto ST$ and $S \mapsto S^{-1}$ are smooth (in fact, real analytic) mappings from $\mathsf{GL}(V) \times \mathsf{GL}(V)$ and $\mathsf{GL}(V)$, respectively, onto $\mathsf{GL}(V)$.

6.1.4 EXAMPLE. The *special linear group* $\mathsf{SL}(n,\mathbb{R})$ and the *orthogonal group* $\mathsf{O}(n)$ are clearly Lie groups. Both subgroups $\mathsf{SL}(n,\mathbb{R})$ and $\mathsf{O}(n)$ are smooth submanifolds of (the Lie group) $\mathsf{GL}(n,\mathbb{R})$, hence smoothness of the group operations on $\mathsf{GL}(n,\mathbb{R})$ implies smoothness of their restrictions to $\mathsf{SL}(n,\mathbb{R})$ and $\mathsf{O}(n)$.

6.1.5 EXAMPLE. The *complex general linear group* $\mathsf{GL}(n,\mathbb{C}) \subseteq \mathbb{R}^{2n^2}$ is a (real) Lie group. In particular, $\mathbb{C}^\times = \mathsf{GL}(1,\mathbb{C})$ is a Lie group. The unit circle $\mathbb{S}^1 \subseteq \mathbb{C}^\times$ is a subgroup and a (smoothly embedded) submanifold, hence also a Lie group.

6.1.6 EXAMPLE. If G_1 and G_2 are Lie groups, then $G_1 \times G_2$ is a Lie group under the usual Cartesian group operations and the smooth product structure. In particular, the m-dimensional *torus*

$$\mathbb{T}^m = \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$$

is a Lie group.

6.1.7 EXAMPLE. Let $\mathbb{H}$ denote the division algebra of *quaternions*. The nonzero quaternions $\mathbb{H}^\times$ form a multiplicative group and a (smooth) manifold diffeomorphic to $\mathbb{R}^4 \setminus \{0\}$. It is clear that the group operations are smooth, so $\mathbb{H}^\times$ is a Lie group. The 3-*sphere* $\mathbb{S}^3 \subseteq \mathbb{H}^\times$ consists of the unit length quaternions, hence it is closed under multiplication and passing to inverses. This gives a Lie group structure on $\mathbb{S}^3$.

Usually, the *identity element* of a Lie group will be denoted by e. (For matrix groups, however, the customary symbol for the identity is I.)

NOTE : In most of the literature, Lie groups are defined to be *real analytic*. That is, G is a manifold with a C^ω (real analytic) atlas and the group operations are real analytic. In fact, no generality is lost by this more restrictive definition. *Smooth Lie groups always support an analytic group structure*, and something even stronger is true. HILBERT'S FIFTH PROBLEM was to show that if G is only assumed to be a topological manifold with continuous group operations, then it is, in fact, a real analytic Lie group. This was finally proven by the combined work of A. GLEASON, D. MONTGOMERY, and L. ZIPPIN (195?).

6.2 Invariant Vector Fields

One of the most important features of a Lie group is the existence of an associated Lie algebra that encodes many of the properties of the group. The crucial property of a Lie group that enables this to occur is the existence of the left and right translations on the group.

Let G be a Lie group. For any $g \in G$, the mappings

$$L_g : G \to G, \quad x \mapsto gx \qquad \text{and} \qquad R_g : G \to G, \quad x \mapsto xg$$

are called the **left** and **right translation** (by g), respectively. For each $g \in G$, both L_g and R_g are smooth mappings on G.

◇ **Exercise 276** Verify that (for every $g_1, g_2, g, h \in G$)

(a) $L_{g_1} \circ L_{g_2} = L_{g_1 g_2}$.

(b) $R_{g_1} \circ R_{g_2} = R_{g_2 g_1}$.

(c) $L_e = R_e = id_G$ ($e \in G$ denotes the identity element).

(d) $(L_g)^{-1} = L_{g^{-1}}$ and $(R_g)^{-1} = R_{g^{-1}}$. (Hence L_g and R_g are diffeomorphisms.)

(e) $L_g \circ R_h = R_h \circ L_g$.

NOTE : Given any admissible chart on G, one can construct an entire atlas on the Lie group G by use of left (or right) translations. Suppose, for example, that (U, ϕ) is an admissible chart with $e \in U$. Define a chart (U_g, ϕ_g) with $g \in U_g$ by letting

$$U_g := L_g(U) = \{L_g(x) \,|\, x \in U\}$$

and defining

$$\phi_g := \phi \circ L_{g^{-1}} : U_g \to \phi(U), \quad x \mapsto \phi(g^{-1}x).$$

The collection of charts $\{(U_g, \phi_g)\}_{g \in G}$ forms a (smooth) atlas provided one can show that the transition mappings

$$\phi_{g_2} \circ \phi_{g_1}^{-1} = \phi \circ L_{g_2^{-1} g_1} \circ \phi^{-1} : \phi_{g_1}(U_{g_1} \cap U_{g_2}) \to \phi_{g_2}(U_{g_1} \cap U_{g_2})$$

is smooth. But this follows from the smoothness of group multiplication and passing to inverse.

By the chain rule,

$$\left(L_{g^{-1}}\right)_{*,gh} \circ (L_g)_{*,h} = \left(L_{g^{-1}} \circ L_g\right)_{*,h} = id_G.$$

Thus the tangent mapping $(L_g)_{*,h}$ is invertible and so, in particular,

$$(L_g)_* = (L_g)_{*,e} : T_eG \to T_gG$$

is a linear isomorphism. Likewise, $(R_g)_{*,h}$ is invertible.

6.2.1 DEFINITION. A vector field X on G is called

- **left-invariant** if for every $g \in G$

$$(L_g)_* X(e) = X(g).$$

- **right-invariant** if for every $g \in G$

$$(R_g)_* X(e) = X(g).$$

It follows that a vector field (on G) that is either left- or right-invariant is determined by its value at the identity.

NOTE : Recall that smooth vector fields act as derivations on the space of smooth functions. (If X is a smooth vector field and f is a smooth function on M, then Xf denotes the (smooth) function $x \mapsto X(x)f$.) For any smooth vector fields X and Y, their *Lie bracket* $[X, Y]$ defined by

$$[X, Y]f = Y(Xf) - X(Yf)$$

is also a smooth vector field. The (vector) space $\mathfrak{X}(M)$ of all smooth vector space on M has the structure of a (real) Lie algebra, with the product given by the Lie bracket.

The set of all left-invariant (respectively, right-invariant) vector fields on a Lie group G is denoted $\mathfrak{X}_L(G)$ (respectively, $\mathfrak{X}_R(G)$). Clearly, both $\mathfrak{X}_L(G)$ and $\mathfrak{X}_R(G)$ are (real) vector spaces (under the pointwise vector addition and scalar multiplication).

NOTE : We defined the push forward $\Phi_{*,p} : T_pM \to T_{\Phi(p)}N$ induced by the (smooth) mapping $\Phi : M \to N$ (the so-called tangent mapping of Φ at $p \in M$).

This is a linear mapping between the vector spaces T_pM and $T_{\Phi(p)}N$, and the question arises of whether it is similarly possible to define an induced mapping between the (vector) spaces of smooth vector fields $\mathfrak{X}(M)$ and $\mathfrak{X}(N)$. Given a vector field $X \in \mathfrak{X}(M)$ and a smooth mapping $\Phi : M \to N$, a natural choice for an induced vector field $\Phi_*X \in \mathfrak{X}(N)$ might appear to be

$$\Phi_*X(\Phi(p)) = \Phi_{*,p}(X(p))$$

but this may fail to be well-defined for two reasons :

- If there are points $p_1, p_2 \in M$ such that $\Phi(p_1) = \Phi(p_2)$ (i.e. the mapping Φ is not one-to-one), then the "definition" above will be ambiguous when $\Phi_*X(p_1) \neq \Phi_*X(p_2)$.
- If Φ is not onto, then the defining equation does not specify the induced vector field outside the range of Φ.

Observe that if Φ is a diffeomorphism from M to N, then neither of these objections apply and an induced vector field Φ_*X *can* be defined via the above equation. However, it is possible that in certain cases the idea will work, even if Φ is not a diffeomeorphism, and this motivates the following definition : vector fields $X \in \mathfrak{X}(M)$ and $Y \in \mathfrak{X}(N)$ are said to be Φ-*related* provided $\Phi_*X(p) = Y(\Phi(p))$ for all $p \in M$. We then write $\Phi_*X = Y$. It is not difficult to see that if $\Phi_*X_1 = Y_1$ and $\Phi_*X_2 = Y_2$, then $[X_1, X_2]$ is Φ-related to $[Y_1, Y_2]$ with

$$\Phi_*[X_1, X_2] = [\Phi_*X_1, \Phi_*X_2].$$

6.2.2 PROPOSITION. *Let X and Y be any left-invariant (respectively, right-invariant) vector fields. Then $[X, Y]$ is a left-invariant (respectively, right-invariant) vector field.*

PROOF : Let $X, Y \in \mathfrak{X}_L(G)$ and $g \in G$. Then (and only then) $(L_g)_*X = X$ and $(L_g)_*Y = Y$. Hence

$$(L_g)_*[X, Y] = [(L_g)_*X, (L_g)_*Y] = [X, Y]$$

and so $[X, Y] \in \mathfrak{X}_L(G)$. The case of right-invariant vector fields is similar. □

Therefore, both $\mathfrak{X}_L(G)$ and $\mathfrak{X}_R(G)$ are Lie subalgebras of the (infinite dimensional) Lie algebra $\mathfrak{X}(G)$ of all smooth vector fields on G.

For each $A \in T_eG$, we define a (smooth) vector field X_A on G by letting

$$X_A(g) := (L_g)_{*,e}A.$$

Then

$$\begin{aligned}(L_g)_*X_A(e) &= (L_g)_*\left((L_e)_*A\right)\\ &= (L_g)_* \circ (L_e)_*A\\ &= (L_{ge})_{*,e}A\\ &= (L_g)_{*,e}A\\ &= X_A(g)\end{aligned}$$

which shows that X_A is left-invariant. Consider the mappings

$$\zeta_1 : \mathfrak{X}_L(G) \to T_eG, \qquad X \mapsto X(e)$$

and

$$\zeta_2 : T_eG \to \mathfrak{X}_L(G), \qquad A \mapsto X_A.$$

◇ **Exercise 277** Verify that ζ_1 and ζ_2 are linear mappings that satisfy

$$\zeta_1 \circ \zeta_2 = id_{T_e(G)} \qquad \text{and} \qquad \zeta_2 \circ \zeta_1 = id_{\mathfrak{X}_L(G)}.$$

(It is clear that ζ_2 is the inverse of ζ_1, and hence for a left-invariant vector field X

$$(L_g)_*X(e) = X(g) \qquad \text{and} \qquad (L_{g^{-1}})_*X_A(g) = A.)$$

Therefore, $\mathfrak{X}_L(G)$ *and* T_eG *are isomorphic (as vector spaces).* It follows that the dimension of the vector space $\mathfrak{X}_L(G)$ is equal to $\dim T_eG = \dim G$.

NOTE : Since, by assumption, G is a (finite-dimensional) manifold it follows that $\mathfrak{X}_L(G)$ is a finite-dimensional, nontrivial subalgebra of the Lie algebra of all (smoth) vector fields on G.

For any $A, B \in T_eG$, we define their *Lie product* (bracket) $[A, B]$ by

$$[A, B] := [X_A, X_B](e)$$

where $[X_A, X_B]$ is the Lie bracket of vector fields. This makes T_eG into a Lie algebra. We say that this defines a Lie product in T_eG via *left extension.*

NOTE : By construction,

$$[X_A, X_B] = X_{[A,B]}$$

for all $A, B \in T_eG$.

6.2.3 DEFINITION. The vector space T_eG with this Lie algebra structure is called the **Lie algebra** of G and is denoted by $\mathfrak{g}$.

◇ **Exercise 278** Let $\varphi : G \to H$ be a smooth homomorphism between the Lie groups G and H. Show that the induced mapping

$$d\varphi = \varphi_{*,e} : T_eG = \mathfrak{g} \to T_eH = \mathfrak{h}$$

is a homomorphism between the Lie algebras of the groups.

A similar construction to the above can be carried out with the Lie algebra $\mathfrak{X}_R(G)$ of right-invariant vector fields on G. In this case, for each $A \in T_eG$, the corresponding right-invariant vector field is defined by

$$Y_A(g) := (R_g)_{*,e}A.$$

We have (for $A, B \in T_eG$)

$$[Y_A, Y_B](e) = -[X_A, X_B](e).$$

Therefore, the Lie product $[\cdot,\cdot]^R$ in $\mathfrak{g}$ defined by *right extension* of elements of $\mathfrak{g}$:

$$[A, B]^R := [Y_A, Y_B](e)$$

is the *negative* of the one defined by left extension; that is,

$$[A, B]^R = -[A, B].$$

NOTE : There is a natural isomorphism between the (Lie algebras) $\mathfrak{X}_L(G)$ and $\mathfrak{X}_R(G)$. It is equal to the tangent mapping of $\Phi : G \to G, \quad x \mapsto x^{-1}$. In particular, we have (for $A \in \mathfrak{g} = T_eG$)

$$\Phi_* X_A = -Y_A.$$

Orbits of invariant vector fields

. . . . . .

6.3 The Exponential Mapping

. . . . . .

6.4 Matrix Groups as Lie Groups

We have seen that the matrix groups $\mathsf{GL}(n,\mathbb{k})$, $\mathsf{SL}(n,\mathbb{k})$, and $\mathsf{O}(n)$ are all Lie groups. These examples are typical of what happens for any matrix group that is a Lie subgroup of $\mathsf{GL}(n,\mathbb{R})$. The following important result holds.

6.4.1 THEOREM. *Let $G \leq \mathsf{GL}(n,\mathbb{R})$ be a matrix group. Then G is a Lie subgroup of $\mathsf{GL}(n,\mathbb{R})$.*

NOTE : In fact, a more general result also holds (but we will not give a proof) : *Every closed subgroup of a Lie group is a Lie subgroup*.

Our aim in this section is to prove THEOREM 4.5.1.

Let $G \leq \mathsf{GL}(n,\mathbb{R})$ be a matrix group, and let $\mathfrak{g} = T_I G$ denote its Lie algebra.

6.4.2 PROPOSITION. *Let*

$$\widetilde{\mathfrak{g}} := \{A \in \mathbb{R}^{n\times n} \mid \exp(tA) \in G \text{ for all } t\}.$$

Then $\widetilde{\mathfrak{g}}$ is a Lie subalgebra of $\mathbb{R}^{n\times n}$.

PROOF : By definition, $\widetilde{\mathfrak{g}}$ is closed under (real) scalar multiplication. If $U, V \in \widetilde{\mathfrak{g}}$ and $r \geq 1$, then the following are in G :

$$\exp\left(\frac{1}{r}U\right)\exp\left(\frac{1}{r}V\right), \quad \left(\exp\left(\frac{1}{r}U\right)\exp\left(\frac{1}{r}V\right)\right)^r,$$
$$\exp\left(\frac{1}{r}U\right)\exp\left(\frac{1}{r}V\right)\exp\left(-\frac{1}{r}U\right)\exp\left(-\frac{1}{r}V\right),$$
$$\left(\exp\left(\frac{1}{r}U\right)\exp\left(\frac{1}{r}V\right)\exp\left(-\frac{1}{r}U\right)\exp\left(-\frac{1}{r}V\right)\right)^{r^2}.$$

For $t \in \mathbb{R}$, by the LIE-TROTTER PRODUCT FORMULA we have

$$\exp(tU + tV) = \lim_{r\to\infty}\left(\exp\left(\frac{1}{r}tU\right)\exp\left(\frac{1}{r}tV\right)\right)^r$$

and by the COMMUTATOR FORMULA

$$\begin{aligned}\exp(t[U,V]) &= \exp([tU,V]) \\ &= \lim_{r\to\infty}\left(\exp\left(\frac{1}{r}tU\right)\exp\left(\frac{1}{r}V\right)\exp\left(-\frac{1}{r}tU\right)\exp\left(-\frac{1}{r}V\right)\right)^{r^2}.\end{aligned}$$

As these are both limits of elements of the closed subgroup $G \leq \mathsf{GL}(n,\mathbb{R})$, they are also in G. This shows that $\widetilde{\mathfrak{g}}$ is a Lie subalgebra of $\mathfrak{gl}(n,\mathbb{R}) = \mathbb{R}^{n\times n}$. □

6.4.3 COROLLARY. *$\widetilde{\mathfrak{g}}$ is a Lie subalgebra of $\mathfrak{g}$.*

PROOF : Let $U \in \widetilde{\mathfrak{g}}$. Then the curve

$$\gamma : \mathbb{R} \to G, \quad t \mapsto \exp(tU)$$

has $\gamma(0) = I$ and $\dot{\gamma}(0) = U$, hence $U \in \mathfrak{g}$. □

NOTE : Eventually we will see that $\widetilde{\mathfrak{g}} = \mathfrak{g}$.

We will require a technical result.

6.4.4 LEMMA. *Let $(A_r)_{r\geq 1}$ and $(\lambda_r)_{r\geq 1}$ be sequences in $\exp^{-1}(G)$ and $\mathbb{R}$, respectively. If $\|A_r\| \to 0$ and $\lambda_r A_r \to A \in \mathbb{R}^{n\times n}$ as $r \to \infty$, then $A \in \widetilde{\mathfrak{g}}$.*

PROOF : Let $t \in \mathbb{R}$. For each r, choose an integer $m_r \in \mathbb{Z}$ so that $|t\lambda_r - m_r| \leq 1$. Then

$$\begin{aligned}\|m_r A_r - tA\| &\leq \|(m_r - t\lambda_r)A_r\| + \|t\lambda_r A_r - tA\| \\ &= |m_r - t\lambda_r|\|A_r\| + \|t\lambda_r A_r - tA\| \\ &\leq \|A_r\| + |t|\|\lambda_r A_r - A\| \to 0\end{aligned}$$

as $r \to \infty$, showing that $m_r A_r \to tA$. Since $\exp(m_r A_r) = \exp(A_r)^{m_r} \in G$ and G is closed in $\mathsf{GL}(n,\mathbb{R})$, we have

$$\exp(tA) = \lim_{r\to\infty}\exp(m_r A_r) \in G.$$

Thus every scalar multiple tA is in $\exp^{-1}(G)$, showing that $A \in \widetilde{\mathfrak{g}}$. □

PROOF OF THEOREM 4.5.1 : Choose a complementary $\mathbb{R}$-subspace $\mathfrak{w}$ to $\widetilde{\mathfrak{g}}$ in $\mathbb{R}^{n\times n}$; that is, any vector subspace such that

$$\begin{aligned} \widetilde{\mathfrak{g}} + \mathfrak{w} &= \mathbb{R}^{n\times n} \\ \dim \widetilde{\mathfrak{g}} + \dim \mathfrak{w} &= \dim \mathbb{R}^{n\times n} = n^2. \end{aligned}$$

(The second of these conditions is equivalent to $\widetilde{\mathfrak{g}} \cap \mathfrak{w} = 0$.) This gives a a *direct sum decomposition* of $\mathbb{R}^{n\times n}$, so every element $X \in \mathbb{R}^{n\times n}$ has a unique decomposition of the form

$$X = U + V \qquad (U \in \widetilde{\mathfrak{g}},\, V \in \mathfrak{w}).$$

Consider the mapping

$$\Phi : \mathbb{R}^{n\times n} \to \mathsf{GL}\,(n, \mathbb{R}), \quad U + V \mapsto \exp(U)\exp(V).$$

Φ is a smooth mapping which maps O to I. Observe that the factor $\exp(U)$ is in G. Consider the derivative (at O)

$$D\Phi(O) : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}.$$

To determine $D\Phi(O)\cdot(A+B)$, where $A \in \widetilde{\mathfrak{g}}$ and $B \in \mathfrak{h}$, we differentiate the curve $t \mapsto \Phi(t(A+B))$ at $t = 0$. Assuming that A and B small enough, for small $t \in \mathbb{R}$, there is a unique matrix $C(t)$ (depending on t) for which

$$\Phi(t(A+B)) = \exp(C(t)).$$

Then (by using the estimate in PROPOSITION 3.5.6)

$$\|C(t) - tA - tB - \frac{t^2}{2}[A,B]\| \le 65|t|^3 \left(\|A\| + \|B\|\right)^3.$$

From this we obtain

$$\begin{aligned} \|C(t) - tA - tB\| &\le \frac{t^2}{2}\|[A,B]\| + 65|t|^3 \left(\|A\| + \|B\|\right)^3 \\ &= \frac{t^2}{2}\left(\|[A,B]\| + 130|t|\left(\|A\| + \|B\|\right)^3\right) \end{aligned}$$

and so

$$\begin{aligned} D\Phi(O)\cdot(A+B) &= \frac{d}{dt}\Phi(t(A+B))\Big|_{t=0} \\ &= \frac{d}{dt}\exp(C(t))\Big|_{t=0} \\ &= A+B. \end{aligned}$$

Hence $D\Phi(O)$ is the identity mapping on $\mathbb{R}^{n\times n}$, and by the Inverse Mapping Theorem, there exists an open neighborhood (and we may take this to be an open ball) $\mathcal{B}_{\mathbb{R}^{n\times n}}(O,\delta)$ of O such that the restriction

$$\Phi_1 := \Phi|_{\mathcal{B}(O,\delta)} : \mathcal{B}(O,\delta) \to \Phi\left(\mathcal{B}(O,\delta)\right)$$

is a smooth diffeomeorphism.

Now we must show that Φ maps some open subset (which we may assume to be an open ball) of $\mathcal{B}_{\mathbb{R}^{n\times n}}(O,\delta)\cap\widetilde{\mathfrak{g}}$ onto an open neighborhood of I in G. Suppose not. Then there is a sequence of elements $(U_r)_{r\geq 1}$ in G with $U_r \to I$ as $r\to\infty$ but $U_r \notin \Phi(\widetilde{\mathfrak{g}})$. For large enough r, $U_r \in \Phi(\mathcal{B}(O,\delta))$, hence there are unique elements $A_r \in \widetilde{\mathfrak{g}}$ and $B_r \in \mathfrak{w}$ with $\Phi(A_r+B_r)=U_r$. Notice that $B_r \neq O$ since otherwise $U_r \in \Phi(\widetilde{\mathfrak{g}})$. As Φ_1 is a diffeomorphism, $A_r+B_r \to O$ and this implies that $A_r \to O$ and $B_r \to O$. By definition of Φ,

$$\exp(B_r) = \exp(A_r)^{-1}U_r \in G.$$

Hence $B_r \in \exp^{-1}(G)$. Consider the elements $\bar{B}_r = \frac{1}{\|B_r\|}B_r$ of unit norm. Each $\bar{B}_r$ is in the unit sphere in $\mathbb{R}^{n\times n}$, which is compact hence there is a convergent subsequence of $(\bar{B}_r)_{r\geq 1}$. By renumbering this subsequence, we can assume that $\bar{B}_r \to B$, where $\|B\|=1$. Applying Lemma 4.5.4 to the sequences $(B_r)_{r\geq 1}$ and $\left(\frac{1}{\|B_r\|}\right)_{r\geq 1}$, we find that $B\in\widetilde{g}$. But each B_r (and hence $\bar{B}_r$) is in $\mathfrak{w}$, so B must be too. Thus $B\in\widetilde{g}\cap\mathfrak{w}$, which contradicts the fact that $B\neq O$.

So there must be an open ball

$$\mathcal{B}_{\widetilde{\mathfrak{g}}}(O,\delta_1) = \mathcal{B}_{\mathbb{R}^{n\times n}}(O,\delta_1)\cap\widetilde{\mathfrak{g}}$$

which is mapped by Φ onto an open neighborhood of I in G. So the restriction of Φ to this open ball is a local diffeomorphism at O. The inverse

mapping gives a local chart for G at I (and moreover $\mathcal{B}_{\widetilde{\mathfrak{g}}}(O,\delta_1)$ is then a smooth submanifold of $\mathbb{R}^{n\times n}$). We can use left translation to move this local chart to a new chart at any other point $U \in G$ (by considering $L_U \circ \Phi$).

So we have shown that $G \leq \mathsf{GL}(n,\mathbb{R})$ is a smooth submanifold. The matrix product $(A,B) \mapsto AB$ is clearly a smooth (in fact, analytic) function of the entries of A and B, and (in light of Cramer's rule) $A \mapsto A^{-1}$ is a smooth (in fact, analytic) function of the entries of A. Hence G is a Lie subgroup, proving THEOREM 4.5.1.

□

This is a fundamental result that can be usefully reformulated as follows : *A subgroup of* $\mathsf{GL}(n,\mathbb{R})$ *is a closed Lie subgroup if and only if it is a matrix subgroup.* (More generally, a subgroup of a Lie group G is a closed Lie subgroup if and only if is a closed subgroup.)

NOTE : Recall that the dimension of a matrix group G (as a manifold) is $\dim \widetilde{\mathfrak{g}}$. By COROLLARY 4.5.3, $\widetilde{\mathfrak{g}} \subseteq \mathfrak{g}$ and so $\dim \widetilde{\mathfrak{g}} \leq \dim \mathfrak{g}$. By definition of $\mathfrak{g} = T_I G$, these dimensions are in fact equal, giving

$$\widetilde{\mathfrak{g}} = \mathfrak{g}.$$

Combining with PROPOSITION 3.3.3, this gives the following result : *For a matrix group* $G \leq \mathsf{GL}(n,\mathbb{R})$*, the exponential mapping*

$$\exp : \mathfrak{g} \to \mathbb{R}^{n\times n}$$

has image in G*. Moreover,* $\exp_G$ *is a local diffeomorphism at the origin (mapping some open neighborhood of* 0 *onto an open neighborhood of* I *in* G*).*

It is a remarkable fact that most of the important examples of Lie groups are (or can easily be represented as) matrix groups. However, *not all Lie groups are matrix groups.* For the sake of completeness, we shall describe the simplest example of a Lie group which is *not* a matrix group.

Consider the matrix group (of *unipotent* 3×3 matrices)

$$\mathsf{H}(1) = \left\{ \gamma(x,y,t) = \begin{bmatrix} 1 & x & t \\ 0 & 1 & y \\ 0 & 0 & 1 \end{bmatrix} \mid x,y,t \in \mathbb{R} \right\} \leq \mathsf{GL}(3,\mathbb{R})$$

commonly referred to as the *Heisenberg group*. $\mathsf{H}(1)$ is a 3-dimensional Lie group.

NOTE : More generally, the *Heisenberg group* $\mathsf{H}(n)$ is defined by

$$\mathsf{H}(n) = \left\{ \gamma(x,y,t) = \begin{bmatrix} 1 & x^T & t \\ 0 & I_n & y \\ 0 & 0 & 1 \end{bmatrix} \mid (x,y) \in \mathbb{R}^{2n}, t \in \mathbb{R} \right\} \leq \mathsf{GL}(n+2, \mathbb{R}).$$

This (matrix) group is *isomorphic* to either one of the following groups :

- $\mathbb{R}^{2n+1}$ equipped with the group multiplication

$$(x,y,t) * (x',y',t') = (x+x', y+y', t+t'+x \bullet y').$$

- $\mathbb{R}^{2n+1}$ equipped with the group multiplication

$$(x,y,t)(x',y',t') = \left(x+x', y+y', t+t' + \frac{1}{2}(\Omega((x,y),(x',y'))) \right)$$

  where $\Omega((x,y),(x',y')) = x \bullet y' - x' \bullet y$ is the standard *symplectic form* on $\mathbb{R}^{2n}$.

The Lie algebra $\mathfrak{h}(n)$ of $\mathsf{H}(n)$ is given by

$$\mathfrak{h}(n) = \left\{ \Gamma(x,y,t) = \begin{bmatrix} 0 & x^T & t \\ 0 & O_n & y \\ 0 & 0 & 0 \end{bmatrix} \mid (x,y) \in \mathbb{R}^{2n}, t \in \mathbb{R} \right\}.$$

(The Lie algebra $\mathfrak{h}(1)$, which occurs throughout quantum physics, is essentially the same as the Lie algebra of operators on differentiable functions $f : \mathbb{R} \to \mathbb{R}$ spanned by the three operators $\mathbf{1}, \mathbf{p}, \mathbf{q}$ defined by

$$\mathbf{1}f(x) := f(x), \quad \mathbf{p}f(x) := \frac{d}{dx}f(x), \quad \mathbf{q}f(x) := xf(x).$$

The non-trivial commutator involving these three operators is given by the *canonical commutation relation* $[\mathbf{p}, \mathbf{q}] = \mathbf{pq} - \mathbf{qp} = \mathbf{1}$.)

◇ **Exercise 279** Determine the (group) commutator in $\mathsf{H}(1)$ (i.e. the product $\gamma\gamma'\gamma^{-1}\gamma'^{-1}$ for $\gamma, \gamma' \in \mathsf{H}(1)$) and hence deduce that the *centre* $Z(\mathsf{H}(1))$ of $\mathsf{H}(1)$ is

$$Z(\mathsf{H}(1)) = \{\gamma(0,0,t) \mid t \in \mathbb{R}\}.$$

Clearly, there is an isomorphism (of Lie groups) between $\mathbb{R}$ and $Z(\mathsf{H}(1))$, under which the subgroup $\mathbb{Z}$ of integers corresponds to the subgroup $\mathcal{Z}$ of $Z(\mathsf{H}(1))$. Thus

$$\mathcal{Z} = \{\gamma(0,0,t) \,|\, t \in \mathbb{Z}\}\,.$$

The subgroup $\mathcal{Z}$ is *discrete* and also *normal*.

NOTE : (1) By a *discrete* group Γ is meant a group with a countable number of elements and the discrete topology (every point is an open set). A discrete group is a 0-dimensional Lie group. Closed 0-dimensional Lie subgroups of a Lie group are usually called *discrete subgroups*. The following remarkable result holds : *If* Γ *is a discrete subgroup of a Lie group* G*, then the space of right (or left) cosets* G/Γ *is a smooth manifold* (and the natural projection $G \to G/\Gamma$ is a smooth mapping).
(2) A subgroup N of G is *normal* if for any $n \in N$ and $g \in G$ we have $gng^{-1} \in N$. A kernel of a homomorphism is normal. Conversely, if N is normal, we can define the *quotient group* G/N whose elements are equivalence classes $[g]$ of elements in G, and two elements g, h are equivalent if and only if $g = hn$ for some $n \in N$. The multiplication is given by $[g][h] = [gh]$ and the fact that N is normal says that this is well-defined. Thus *normal subgroups are exactly kernels of homomorphisms.*

Hence we can form the *quotient group*

$$\mathsf{H}\,(1)/\mathcal{Z}$$

which is in fact a (3-dimensional) Lie group. (Its Lie algebra is $\mathfrak{h}\,(1)$.)

The following result (which we will not prove) tells that *the Lie group* $\mathsf{H}\,(1)/\mathcal{Z}$ *cannot be realized as a matrix group.*

6.4.5 PROPOSITION. *There are no continuous homomorphisms* $\varphi : \mathsf{H}\,(1)/\mathcal{Z} \to \mathsf{GL}\,(n,\mathbb{C})$ *with trivial kernel.*

6.5 Hamiltonian Vector Fields

. . . . . .

6.6 Lie-Poisson Reduction

. . . . . .

Bibliography

[AM78] R. Abraham and J.E. Marsden – *Foundations of Mechanics* (Second Edition), Benjamin/Cummings, 1978.

[AMR88] R. Abraham, J.E. Marsden, and T. Ratiu – *Manifolds, Tensor Analysis, and Applications* (Second Edition), Springer-Verlag, 1988.

[Arm] M.A. Armstrong – *Groups and Symmetry*, Springer-Verlag, 1988.

[Arn73] V.I. Arnold – *Ordinary Differential Equations*, The MIT Press, 1973.

[Arn78] V.I. Arnold – *Mathematical Methods of Classical Mechanics* (Second Edition), Springer-Verlag, 1989.

[Arv03] A. Arvanitoyeorgos – *An Introduction to Lie Groups and the Geometry of Homogeneous Spaces*, Amer. Math. Soc., 2003.

[Aus67] L. Auslander – *Differential Geometry*, Harper and Row, 1967.

[Bak02] A. Baker – *Matrix Groups. An Introduction to Lie Group Theory*, Springer-Verlag, 2002.

[Bea05] A.F. Beardon – *Algebra and Geometry*, Cambridge University Press, 2005.

[BK89] J.G.F. Belinfante and B. Kolman – *A Survey of Lie Groups and Lie Algebras with Applications and Computational Methods*, SIAM, 1989.

[Bel97] R. Bellman – *Introduction to Matrix Analysis* (Second Edition), SIAM, 1997.

[BG88] M. Berger and B. Gostiaux – *Differential Geometry : Manifolds, Curves, and Surfaces*, Springer-Verlag, 1988.

[Boo03] W.M. Boothby – *An Introduction to Differentiable Manifolds and Riemannian Geometry* (Revised Second Edition), Academic Press, 2003.

[Bre97] O. Bretscher – *Linear Algebra with Applications*, Prentice Hall, 1997.

[Brow96] A. Browder – *Mathematical Analysis. An introduction*, Springer-Verlag, 1996.

[Bur91] R.P. Burn – *Groups. A Path to Geometry*, Cambridge University Press, 1991.

[Con01] L. Conlon – *Differentiable Manifolds* (Second Edition), Birkhäuser, 2001.

[Cur84] M.L. Curtis – *Matrix Groups* (Second Edition), Springer-Verlag, 1984.

[Die69] J. Dieudonné – *Foundations of Modern Analysis*, Academic Press, 1969.

[DoC76] M.P. do Carmo – *Differential Geometry of Curves and Surfaces*, Prentice Hall, 1976.

[DK00] J.J. Duistermaat and J.A.C. Kolk – *Lie Groups*, Springer-Verlag, 2000.

[DK04] J.J. Duistermaat and J.A.C. Kolk – *Multidimensional Real Analysis I. Differentiation*, Cambridge University Press, 2004.

[Fra99] T. Frankel – *The Geometry of Physics. An Introduction*, Cambridge University Press, 1999.

[Gib01] C.G. Gibson – *Elementary Geometry of Differentiable Curves*, Cambridge University Press, 2001.

[Gre80] M.J. Greenberg – *Euclidean and Non-Euclidean Geometries. Development and History*, W.H. Freeman, 1980.

[Hal03] B.C. Hall – *Lie Groups, Lie Algebras, and Representations*, Springer-Verlag, 2003.

[Hel78] S. Helgason – *Differential Geometry, Lie Groups, and Symmetric Spaces*, Academic Press, 1978.

[Hen01] M. Henle – *Modern Geometries. Non-Euclidean, Projective, and Discrete* (Second Edition), Prentice Hall, 2001.

[How83] R. Howe – Very basic Lie theory, *Amer. Math. Monthly* **90**(9)(1983), 600-623.

[Hsi97] C-C. Hsiung – *A First Course in Differential Geometry*, International Press, 1997.

[Jen94] G.A. Jennings – *Modern Geometry with Applications*, Springer-Verlag, 1994.

[Kuh02] W. Kühnel – *Differential Geometry. Curves - Surfaces - Manifolds*, Amer. Math. Soc., 2002.

[Lee03] J.M. Lee – *Introduction to Smooth Manifolds*, Springer-Verlag, 2003.

[MR94] J.E. Marsden and T.S. Ratiu – *Introduction to Mechanics and Symmetry* (Seconfd Edition), Springer-Verlag, 1999.

[Mat72] Y. Matsushima – *Differentiable Manifolds*, Marcel Dekker, 1972.

[McC94] J.McCleary – *Geometry from a Differentiable Viewpoint*, Cambridge University Press, 1994.

[Mey00] C.D. Meyer – *Matrix Analysis and Applied Linear Algebra*, SIAM, 2000.

[Mil77] R.S. Millman – Kleinian transformation geometry, *Amer. Math. Monthly* **84**(5)(1977), 338-349.

[MS73] R.S. Millmann and A.K. Stehney – The geometry of connections, *Amer. Math. Monthly* **80**(5)(1973), 475-500.

[Nab92] G.L. Naber – *The Geometry of Minkowski Spacetime*, Springer-Verlag, 1992.

[ONe97] B. O'Neill – *Elementary Differential Geometry* (Second Edition), Academic Press, 1997.

[Opr97] J. Oprea – *Differential Geometry and its Applications*, Prentice Hall, 1997.

[Pri77] J.F. Price – *Lie Groups and Compact Groups*, Cambridge University Press, 1977.

[Roe93] J. Roe – *Elementary Geometry*, Oxford University Press, 1993.

[Ros02] W. Rossmann – *Lie Groups. An Introduction Through Linear Groups*, Oxford University Press, 2002.

[Rud76] W. Rudin – *Principles of Mathematical Analysis* (Third Edition), McGraw-Hill, 1976.

[SW73] A.A. Sagle and R.E. Walde – *Introduction to Lie Groups and Lie Algebras*, Academic Press, 1973.

[Sha97] R.W. Sharpe – *Differential Geometry. Cartan's Generalization of Klein's Erlangen Program*, Springer-Verlag, 1997.

[Sib98] T.Q. Sibley – *The Geometric Viewpoint. A Survey of Geometries*, Addison-Wesley, 1998.

[Spi65] M. Spivak – *Calculus on Manifolds*, W.A. Benjamin, 1965.

[Spi99] M. Spivak – *A Comprehensive Introduction to Differential Geometry* (Third Edition), vol.I-V, Publish or Perish, 1999.

[Tap05] K. Tapp – *Matrix Groups for Undergraduates*, Amer. Math. Soc., 2005.

[Ton69] P. Tondeur – *Introduction to Lie Groups and Transformation Groups* (Second Edition), Springer-Verlag, 1969.

[War83] F.W. Warner – *Foundations of Differentiable Manifolds and Lie Groups*, Springer-Verlag, 1983.

[Yag88] I.M. Yaglom – *Felix Klein and Sophus Lie. Evolution of the Idea of Symmetry in the Nineteenth Century*, Birkhäuser, 1988.