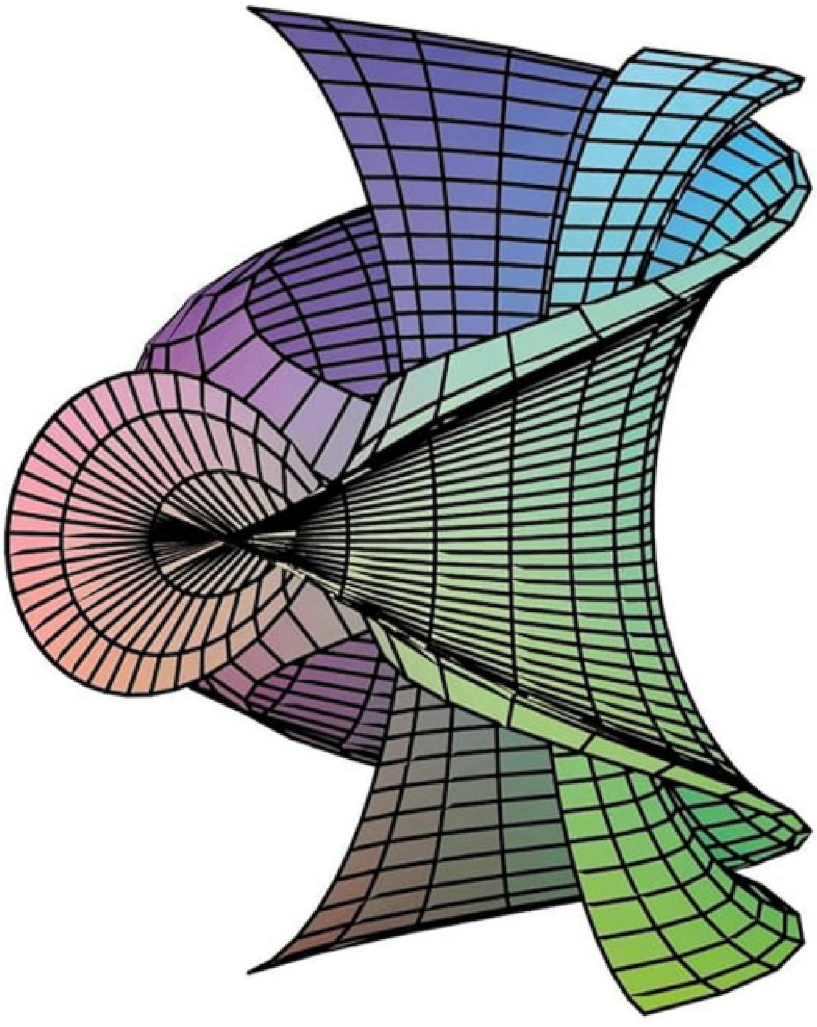
DE GRUYTER GRADUATE

Gerard Walschap

MULTIVARIABLE CALCULUS AND DIFFERENTIAL GEOMETRY



DE G

Gerard Walschap

Multivariable Calculus and Differential Geometry

Gerard Walschap

Multivariable Calculus and Differential Geometry

DE GRUYTER

ISBN 978-3-11-036949-6

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Typesetting: PTP-Berlin, Protago-TEX-Production GmbH, Berlin
Printing and binding: CPI books GmbH, Leck
♾ Printed on acid-free paper
Printed in Germany

www.degruyter.com

Preface

The purpose of this text is to introduce the reader to the basic concepts of differential geometry with a minimum of prerequisites. The only absolute requirement is a solid background in single variable calculus, although some mathematical maturity would also be helpful. All the material from multivariable calculus, linear algebra, and basic analysis that is needed in the study of elementary differential geometry has been included. Many students who have completed a standard course in multivariable calculus still struggle to grasp manifold theory. There are several factors contributing to this; chief among them, at least in the author's view, is the lack of integration of linear algebra with calculus. Although most texts on calculus of several variables introduce the reader to matrices, few emphasize concepts such as linear transformations. The latter are admittedly of little use to someone who is only interested in computing, say, partial derivatives of a map, but they are a cornerstone of both calculus and differential geometry. These areas are concerned with retrieving information about a smooth map between Euclidean spaces from its derivative, and the derivative of such a map at a point is essentially synonymous with the notion of linear transformation. We hope that even the reader who is familiar with linear algebra will find it useful to have access to this material should it prove necessary. It constitutes the core of the first chapter. We emphasize this is not a course in linear algebra: only those topics that will be needed in studying geometry are tackled. Furthermore, the presentation and proofs of the material are carried out through the prism of calculus rather than in full generality. For example, several properties that hold in general vector spaces are only established for inner product spaces if their proof is shorter or easier in that context. The remaining part of the chapter is devoted to metric properties of Euclidean spaces, and to the notions of limits and continuity of maps between those spaces.

The second chapter introduces the reader to differentiation of maps between Euclidean spaces, as well as to further concepts from linear algebra that are relevant to inner product spaces. Special emphasis is given to vector fields and their Lie brackets.

The third chapter discusses the spaces that are studied in differential geometry, namely differentiable manifolds. Instead of defining them abstractly, we only consider submanifolds of Euclidean space, which, as geometers know, does not constitute – at least in principle – a restriction. The reason for this is twofold: on the one hand, it is much more intuitive to consider a surface in 3-space and make the leap of imagination to higher dimensions than to study an abstract topological space; on the other, the abstract approach requires concepts from general topology, such as paracompactness, which lie outside the scope of this text. We next proceed to calculus on manifolds, and introduce the basic concepts of covariant derivatives, geodesics, and curvature.

Chapter 4 discusses integration of functions of several variables. We first look at functions with domain in some Euclidean space, determine conditions for integrability, and highlight some of the main tools available, such as Fubini's theorem and the

change of variable formula. The special cases of cylindrical and spherical change of coordinates is discussed in detail. Several applications of multiple integrals to concepts from physics are also examined.

Chapter 5 briefly introduces tensors and then discusses differential forms on manifolds and their integration. This leads us to the modern formulation of Stokes' theorem, and how it unifies the classical versions of the fundamental theorem of Calculus, Green's theorem, and Stokes' theorem. As a practical example, we discuss how Green's theorem explains the principle behind the polar planimeter, a device that calculates the area of a plane region by merely tracing out its boundary.

The next chapter examines manifolds as spaces where a distance between any two points can be defined; when the space is complete, this distance is given by the length of the shortest curve joining the points. Further properties of these metric spaces are discussed, and the chapter ends with an illustration of how curvature affects the shape of space.

The last chapter examines all these concepts in the special case of hypersurfaces; i.e., manifolds whose dimension is one less than that of the ambient Euclidean space. In this setting, there are additional tools such as the Gauss map that provide further insight into the structure of the space. Many features, in particular geodesics, convexity, and curvature become more transparent in this context. We also briefly discuss the geometry of some classical surfaces.

Contents

1 Euclidean Space

We begin by exploring some of the properties of Euclidean space that will be needed to study Calculus. They roughly fall into two categories, algebraic and topological. The former are closely related to the concept of vector spaces.

1.1 Vector spaces

As a set, k-dimensional Euclidean space consists of all k-tuples $(a_1, \ldots, a_k)$ of real numbers a_i, $i = 1, \ldots, k$ (the notation Euclidean space usually requires some additional structure as we shall see in section 4; we will, for now, ignore this out of laziness and for lack of a better name). Terms such as "dimension" and "Euclidean" will be examined in more detail later in the chapter. The number a_i is called the *i-th coordinate* of the point, and the map $u^i : \mathbb{R}^k \to \mathbb{R}$ which assigns to a point its i-th coordinate is called the *i-th projection*. When k equals 2 or 3, $\mathbb{R}^k$ can be visualized geometrically as the plane or 3-space, with (a_1, a_2, a_3) representing the point in 3-space whose orthogonal projection onto the x, y, and z-axes equals a_1, a_2, and a_3 respectively (the plane may be identified with those points in $\mathbb{R}^3$ with zero third coordinate). The motivation behind the notation $\mathbb{R}^k$ is that it is a kind of product of $\mathbb{R}$ with itself k times: the *cartesian product* $A \times B$ of two sets A and B is defined to be the set of all pairs (a, b) where a and b range over A and B respectively, and $\mathbb{R}^k$ is then the product $\mathbb{R} \times \cdots \times \mathbb{R}$ of $\mathbb{R}$ with itself k times.

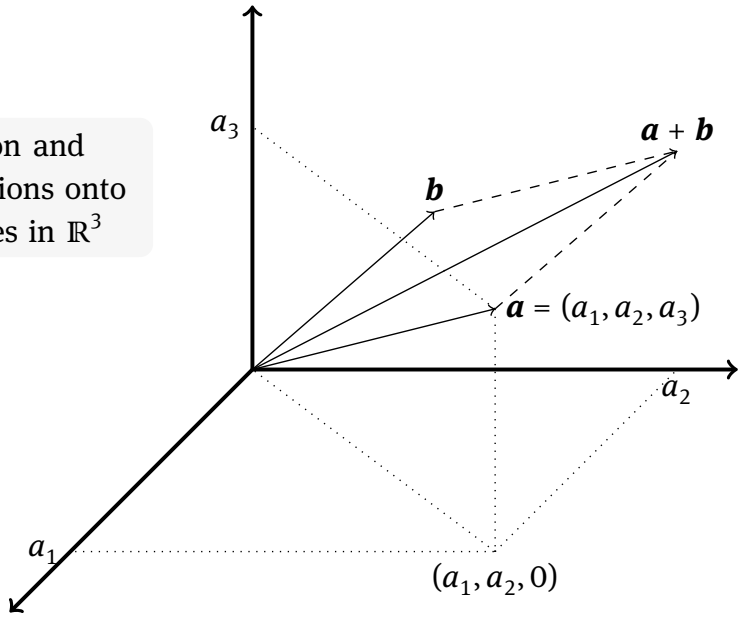
$\mathbb{R}^k$ comes with two operations, *addition* $+ \colon \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^k$, given by

$$(a_1, \ldots, a_k) + (b_1, \ldots, b_k) = (a_1 + b_1, \ldots, a_k + b_k), \qquad a_i \in \mathbb{R},$$

and *scalar multiplication* $\cdot : \mathbb{R} \times \mathbb{R}^k \to \mathbb{R}^k$,

$$c \cdot (a_1, \ldots, a_k) = (ca_1, \ldots, ca_k), \qquad c, a_i \in \mathbb{R}.$$

We will mostly dispense with the dot in the scalar multiplication, and often abbreviate $(a_1, \ldots, a_k)$ by $\boldsymbol{a}$. If one visualizes a point $\boldsymbol{a}$ in 3-space as a directed line segment from the origin $\mathbf{0}$ to $\boldsymbol{a}$, then the sum of $\boldsymbol{a}$ and $\boldsymbol{b}$ is the diagonal of the parallelogram determined by $\boldsymbol{a}$ and $\boldsymbol{b}$ (unless $\boldsymbol{a}$ and $\boldsymbol{b}$ are parallel, in which case $\boldsymbol{a} + \boldsymbol{b}$ is also parallel).



Addition and projections onto the axes in $\mathbb{R}^3$

Because of the properties of scalar addition and product in $\mathbb{R}$, one easily verifies that $\mathbb{R}^k$ with these operations has the following structure:

Definition 1.1.1. A (real) *vector space* is a set V together with two operations $+\colon V\times V \to V$ and $\cdot\colon \mathbb{R}\times V \to V$ that satisfy for all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$, $a, b \in \mathbb{R}$:

(1) $\boldsymbol{u}+\boldsymbol{v}=\boldsymbol{v}+\boldsymbol{u}$;
(2) $(\boldsymbol{u}+\boldsymbol{v})+\boldsymbol{w}=\boldsymbol{u}+(\boldsymbol{v}+\boldsymbol{w})$;
(3) There exists an element $\boldsymbol{0} \in V$, called the *zero vector*, satisfying $\boldsymbol{u}+\boldsymbol{0}=\boldsymbol{u}$ for all $\boldsymbol{u}$;
(4) Any $\boldsymbol{u}$ in V has an additive inverse; i.e., an element $-\boldsymbol{u}$ such that $\boldsymbol{u}+(-\boldsymbol{u})=\boldsymbol{0}$.
(5) $(a+b)\cdot\boldsymbol{u}=a\cdot\boldsymbol{u}+b\cdot\boldsymbol{u}$;
(6) $a\cdot(\boldsymbol{u}+\boldsymbol{v})=a\cdot\boldsymbol{u}+b\cdot\boldsymbol{v}$;
(7) $1\cdot\boldsymbol{u}=\boldsymbol{u}$;
(8) $a\cdot(b\cdot\boldsymbol{u})=(ab)\cdot\boldsymbol{u}$

Elements of V are called *vectors*. As in $\mathbb{R}^k$, we will for the most part omit the dot in scalar multiplication.

Examples 1.1.1. (i) An $m\times n$ *matrix* A is a rectangular array of mn real numbers organized in m rows and n columns:

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

The element a_{ij} which lies on the i-th row and j-th column is referred to as the (i,j)-th entry of A. The sum of two such matrices A and B is defined to be the matrix of the same size whose (i,j)-th entry is $a_{ij}+b_{ij}$. Given $c \in \mathbb{R}$, cA is the matrix whose

(i,j)-th entry equals ca_{ij}. The collection $M_{m,n}$ of all $m \times n$ matrices is a vector space under these operations. In fact, when $m = 1$, we recover $\mathbb{R}^n$. Much of the time, though, we will identify $\mathbb{R}^n$ with column matrices (i.e., with $M_{n,1}$) rather than with row matrices.

(ii) The collection $\mathcal{F}(\mathbb{R})$ of all functions $f : \mathbb{R} \to \mathbb{R}$ with addition and scalar multiplication defined in the usual way,

$$(f + g)(x) = f(x) + g(x), \quad (cf)(x) = cf(x), \qquad c, x \in \mathbb{R}, \quad f, g \in \mathcal{F}(\mathbb{R}),$$

is a vector space.

(iii) The collection $\mathcal{C}(\mathbb{R})$ of continuous functions $f : \mathbb{R} \to \mathbb{R}$ is a vector space with the operations from (ii).

Notice that a zero vector in a vector space V is necessarily unique: for if $\boldsymbol{0}$ and $\boldsymbol{0}'$ are two zero vectors, then $\boldsymbol{0} + \boldsymbol{0}' = \boldsymbol{0}$ because $\boldsymbol{0}'$ is a zero vector, and $\boldsymbol{0} + \boldsymbol{0}' = \boldsymbol{0}'$ because $\boldsymbol{0}$ is also one. A similar argument shows that a vector has a unique inverse. Furthermore, for any $\boldsymbol{v} \in V$, $0\boldsymbol{v} = \boldsymbol{0}$, since $0\boldsymbol{v} = (0 + 0)\boldsymbol{v} = 0\boldsymbol{v} + 0\boldsymbol{v}$, and the conclusion follows by adding the inverse of $0\boldsymbol{v}$ to both sides.

Example (iii) above is a subset of (ii). This is a quite common occurrence: A nonempty subset of a vector space V is called a *subspace* of V if it is a vector space with the operations inherited from V. Clearly, a necessary condition for $W \subset V$ to be a subspace is that it be closed under addition and scalar multiplication; i.e., that the sum of vectors in W is again in W, and similarly for scalar multiplication. It turns out this is also a sufficient condition:

Proposition 1.1.1. *A nonempty subset W of a vector space V is a subspace of V if*

(1) $\boldsymbol{v} + \boldsymbol{w} \in W$ *for any* $\boldsymbol{v}, \boldsymbol{w} \in W$, *and*

(2) $c\boldsymbol{v} \in W$ *for any* $c \in \mathbb{R}$, $\boldsymbol{v} \in W$.

Proof. All the axioms for vector space, except possibly the existence of zero vector and inverses, are satisfied because they hold for elements of V and therefore also for elements of W. The remaining two axioms follow from closure under scalar multiplication: pick any $\boldsymbol{v} \in W$; then $\boldsymbol{0} = 0\boldsymbol{v} \in W$. For the other axiom, observe that $\boldsymbol{0} = 0\boldsymbol{v} = (1 - 1)\boldsymbol{v} = \boldsymbol{v} + (-1)\boldsymbol{v}$, and by uniqueness of inverses, $-\boldsymbol{v} = (-1)\boldsymbol{v}$. This shows that $-\boldsymbol{v} \in W$. □

Examples 1.1.2. (i) The transpose A^T of an $m \times n$ matrix A is the $n \times m$ matrix obtained by interchanging the rows and columns of A; i.e., the (i,j)-th element of A^T is a_{ji}. An $n \times n$ matrix A is said to be symmetric (respectively skew-symmetric) if $A = A^T$ (resp. $A = -A^T$). The easily checked identities

$$(A + B)^T = A^T + B^T, \qquad (cA)^T = cA^T,$$

imply that the set of symmetric matrices and that of skew-symmetric ones are both subspaces of $M_{n,n}$.

(ii) The *trace* of an $n \times n$ matrix A is the sum $\operatorname{tr} A = \sum_{i=1}^{n} a_{ii}$ of the elements on the diagonal. The collection of matrices with trace equal to 0 is a subspace of $M_{n,n}$, but those with trace any other number is not.

(iii) A *linear combination* of vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in V$ is a vector of the form $c_1\boldsymbol{v}_1 + \cdots + c_k\boldsymbol{v}_k$, where $c_i \in \mathbb{R}$. The collection $\operatorname{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ of all such linear combinations is called the *span* of $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$. It follows immediately from Proposition 1.1.1 that the span of a subset $S \subset V$ is a subspace of V. For instance, if $V = \mathbb{R}^3$, then the span of $\{(1,2,0), (-1,1,2)\}$ is the space

$$\{s(1,2,0) + t(-1,1,2) \mid s,t \in \mathbb{R}\} = \{(s-t, 2s+t, 2t) \in \mathbb{R}^3 \mid s,t \in \mathbb{R}\}.$$

Definition 1.1.2. Vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in V$ are said to be *linearly dependent* if there exist scalars $a_1, \ldots, a_k$, *not all zero*, such that $a_1\boldsymbol{v}_1 + \cdots + a_k\boldsymbol{v}_k = \boldsymbol{0}$. Otherwise, they are said to be *linearly independent*.

For vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$, the equation $x_1\boldsymbol{v}_1 + \cdots + x_k\boldsymbol{v}_k = \boldsymbol{0}$ in the variables x_i always has the trivial solution $x_1 = \cdots = x_k = 0$. In fact, the collection of solutions $(x_1, \ldots, x_k)$ is a subspace of $\mathbb{R}^k$. The $\boldsymbol{v}_i$'s are linearly independent iff the trivial solution is the only one.

Another way to characterize linear dependence is by means of linear combinations (see Example 1.1.2 (iii)): A set $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ is linearly dependent if and only if one of the vectors is a linear combination of the others. Indeed, if $\boldsymbol{v}_i = \sum_{j \neq i} a_j\boldsymbol{v}_j$, then $\sum_l x_l\boldsymbol{v}_l = \boldsymbol{0}$ for $x_l = a_l$ when $l \neq i$, and $x_i = -1 \neq 0$. Conversely, if the set is linearly dependent, then $\sum a_i\boldsymbol{v}_i = \boldsymbol{0}$ for some scalars a_i, with at least one of them, say a_j, different from zero. Then $\boldsymbol{v}_j = -\sum_{i \neq j}(a_i/a_j)\boldsymbol{v}_i$ is a linear combination of the others.

Notice that if a spanning set for a vector space is linearly dependent, then any vector that is a linear combination of the others can be discarded, and the remaining ones still span the space. The result is a minimal spanning set:

Definition 1.1.3. A *basis* of a vector space V is a linearly independent spanning set.

Examples 1.1.3. (i) Let $\boldsymbol{e}_i$, $1 \leq i \leq n$, denote the point in $\mathbb{R}^n$ with $u^j(\boldsymbol{e}_i) = 1$ if $i = j$ and 0 otherwise (recall that u^j is the j-th projection). The identity $(a_1, \ldots, a_n) = \sum_i a_i\boldsymbol{e}_i$ implies that the $\boldsymbol{e}_i$ are both independent and span $\mathbb{R}^n$. This is the so-called *standard basis* of $\mathbb{R}^n$.

(ii) The collection of points $(x,y,z) \in \mathbb{R}^3$ such that $x + y + z = 0$ is a subspace V of $\mathbb{R}^3$. The vectors $\boldsymbol{u} = (1,0,-1)$ and $\boldsymbol{v} = (0,1,-1)$ belong to V, and are linearly independent: the equation $a\boldsymbol{u} + b\boldsymbol{v} = \boldsymbol{0}$ yields

$$a(1,0,-1) + b(0,1,-1) = (a, b, -a-b) = (0,0,0),$$

so that $a = b = 0$. They also span V, since $(x,y,z) \in V$ iff $z = -x - y$; i.e., if and only if

$$(x,y,z) = (x,y,x-y) = (x,0,x) + (0,y,-y) = x\boldsymbol{u} + y\boldsymbol{v}.$$

Thus, they form a basis of V.

Theorem 1.1.1. *If $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ is a basis of V, then any subset of V that contains more than n vectors is linearly dependent.*

Proof. Let $S = \{\boldsymbol{u}_1, \dots, \boldsymbol{u}_k\}$ be a subset with $k > n$ elements. If one of the $\boldsymbol{u}_i$ is the zero vector, then S is certainly linearly dependent (see Exercise 1.4), so we may assume this is not the case. Since the $\boldsymbol{v}_i$ span V, we may express $\boldsymbol{u}_1$ as a linear combination $\boldsymbol{u}_1 = \sum_i a_i \boldsymbol{v}_i$, where at least one of the scalar coefficients is nonzero. By renumbering the basis elements if necessary, we may assume $a_1 \neq 0$, so that the above equation can be solved for $\boldsymbol{v}_1$:

$$\boldsymbol{v}_1 = (1/a_1)(\boldsymbol{u}_1 - a_2\boldsymbol{v}_2 - \cdots - a_n\boldsymbol{v}_n).$$

This implies that V is spanned by $\{\boldsymbol{u}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n\}$. Continuing in this way, we can replace the $\boldsymbol{v}$'s one by one with $\boldsymbol{u}$'s. More precisely, suppose V has been shown to be spanned by $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_l, \boldsymbol{v}_{l+1}, \dots, \boldsymbol{v}_n\}$. Then $\boldsymbol{u}_{l+1}$ can be expressed as a linear combination

$$\boldsymbol{u}_{l+1} = \sum_{i \leq l} b_i \boldsymbol{u}_i + \sum_{i > l} c_i \boldsymbol{v}_i.$$

where at least one of the scalars is nonzero. We may assume one of the c_i is nonzero, for if they are all zero, then $\boldsymbol{u}_{l+1}$ is a linear combination of the $\boldsymbol{u}$'s and S is then linearly dependent. Renumbering the vectors if necessary, we may suppose that $c_{l+1} \neq 0$, and we can write $\boldsymbol{v}_{l+1}$ as a linear combination of $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_{l+1}, \boldsymbol{v}_{l+2}, \dots, \boldsymbol{v}_n\}$. This means the latter set spans V. By induction, we conclude that V is spanned by $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$. But then any $\boldsymbol{u}_{n+i}$ is a linear combination of $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$, and the result follows. □

Corollary 1.1.1. *If $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_k\}$ and $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ are two bases of V, then $k = n$.*

Proof. Since the $\boldsymbol{v}_i$ are linearly independent, Theorem 1.1.1 implies that $n \leq k$. By symmetry, $k \leq n$. □

Definition 1.1.4. If V has a basis consisting of n elements, the *dimension* $\dim V$ of V is defined to be n.

A trivial vector space (one that consists of a single element, which is then necessarily the zero vector) has no nonempty linearly independent spanning set. We therefore say it has the empty set as basis, and its dimension is zero. A vector space with dimension n for some n is called a *finite-dimensional* space. Unless specified otherwise, we will concern ourselves here only with finite-dimensional vector spaces, even though it should be noted that many spaces have infinite bases: for example, one such is the space of all polynomials with the usual addition of polynomials and scalar multiplication. The reader is invited to check that the collection $\{p_k(x) \mid k = 0, 1, \dots\}$, where $p_k(x) = x^k$, is a basis of this space.

One particularly useful feature of bases is that they can be used to uniquely identify vectors:

Proposition 1.1.2. *If $\mathcal{B} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ is a basis of V, then any $\boldsymbol{v} \in V$ can be written in one and only one way as a linear combination of the basis elements.*

Proof. That $\boldsymbol{v}$ can be written as a linear combination follows from the fact that the $\boldsymbol{v}_i$ span V. That the expression is unique follows from linear independence: if $\sum a_i\boldsymbol{v}_i = \sum b_i\boldsymbol{v}_i$, then $\sum(a_i - b_i)\boldsymbol{v}_i = \boldsymbol{0}$, so that each $a_i - b_i$ must be zero. □

The uniqueness part of the above proposition enables us to make the following:

Definition 1.1.5. Let $\mathcal{B} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ be an ordered basis of V. The *coordinate vector* of $\boldsymbol{v} = a_1\boldsymbol{v}_1 + \cdots + a_n\boldsymbol{v}_n \in V$ with respect to $\mathcal{B}$ is the element $[\boldsymbol{v}]_{\mathcal{B}} = [a_1, \ldots, a_n]^T \in M_{n,1}$.

For example, if $V = \mathbb{R}^n$, then the coordinate vector of $\boldsymbol{v}$ with respect to the standard basis is just $\boldsymbol{v}$ itself, written as a column rather than a row. A more interesting example is the plane V with the ordered basis $(\boldsymbol{u}, \boldsymbol{v})$ from Examples 1.1.3 (ii): The vector $(1, -1, 0)$ belongs to V, and since it equals $\boldsymbol{u} - \boldsymbol{v}$, its coordinate vector in that basis is $[1, -1]^T$.

Notice that the order in which the basis elements are listed is crucial. We will, when referring to an ordered basis, dispense with braces when listing the elements, and either use parentheses as in the above definition or no symbols at all.

1.2 Linear transformations

Definition 1.2.1. A map $L : V \to W$ between vector spaces V and W is said to be *linear* if it preserves the vector space operations; i.e., if
(1) $L(\boldsymbol{u} + \boldsymbol{v}) = L(\boldsymbol{u}) + L(\boldsymbol{v})$ for all $\boldsymbol{u}, \boldsymbol{v} \in V$;
(2) $L(c\boldsymbol{u}) = cL(\boldsymbol{u})$ for all $c \in \mathbb{R}$, $\boldsymbol{u} \in V$.

When there is a single vector $\boldsymbol{u}$ in the argument, it is customary to write $L\boldsymbol{u}$ instead of $L(\boldsymbol{u})$. Notice that the + on the right of the equality sign in (1) denotes the addition in W, whereas the one on the left refers to addition in V. A similar observation involving scalar multiplication holds for (2). If L is invertible, i.e., if there exists a map $L^{-1} : W \to V$ such that $L \circ L^{-1} = 1_W$ and $L^{-1} \circ L = 1_V$ (1_V is the identity map on V), then L is said to an *isomorphism*; in this case V and W are said to be *isomorphic*. Observe that the inverse of an isomorphism is again linear: given $\boldsymbol{w}_i \in W$, $i = 1, 2$, there exist unique $\boldsymbol{v}_i \in V$ such that $L\boldsymbol{v}_i = \boldsymbol{w}_i$. Then

$$\begin{aligned} L^{-1}(\boldsymbol{w}_1 + \boldsymbol{w}_2) &= L^{-1}(L\boldsymbol{v}_1 + L\boldsymbol{v}_2) = L^{-1}(L(\boldsymbol{v}_1 + \boldsymbol{v}_2)) = \boldsymbol{v}_1 + \boldsymbol{v}_2 \\ &= L^{-1}\boldsymbol{w}_1 + L^{-1}\boldsymbol{w}_2, \end{aligned}$$

and a similar argument shows that L^{-1} preserves scalar multiplication. Isomorphic vector spaces are identical as far as their algebraic structure is concerned, and can thus be identified.

Examples 1.2.1. (i) The map $[]_{\mathcal{S}} : \mathbb{R}^n \to M_{n,1}$ that assigns to a point its coordinate vector with respect to the standard basis (see Definition 1.1.5) is an isomorphism. This formally justifies why we can use either row or column matrices to denote elements of $\mathbb{R}^n$.

(ii) More generally, the map $M_{m,n} \to M_{n,m}$ that sends a matrix to its transpose is an isomorphism.

(iii) Let V be any n-dimensional vector space with basis $\mathcal{B}$. As in (i), we see that the map $[]_{\mathcal{B}} : V \to M_{n,1}$ that assigns to a vector in V its coordinate vector with respect to $\mathcal{B}$ is an isomorphism. Together with (i), this shows that any n-dimensional vector space is isomorphic to $\mathbb{R}^n$. This isomorphism is not canonical, however, in the sense that it depends on the chosen basis.
More generally, any two vector spaces of the same dimension are isomorphic: if $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ is a basis of V, and $\{\boldsymbol{w}_1, \dots, \boldsymbol{w}_n\}$ is a basis of W, define $L : V \to W$ as follows: given $\boldsymbol{v} \in V$, it can be uniquely written as $\boldsymbol{v} = \sum a_i \boldsymbol{v}_i$ for some choices of $a_i \in \mathbb{R}$. Set $L\boldsymbol{v} = \sum a_i \boldsymbol{w}_i$. It is straightforward to check that L is a well-defined linear transformation, and that it is entirely determined by the fact that $L\boldsymbol{v}_i = \boldsymbol{w}_i$, $i = 1, \dots, n$. In fact, if W is a vector space of any dimension, then for any vectors $\boldsymbol{w}_i \in W$, $i = 1, \dots, n$, there exists a unique linear map $L : V \to W$ such that $L(\boldsymbol{v}_i) = \boldsymbol{w}_i$ for all $i = 1, \dots, n$, see Exercise 1.10. L will not, in general, be an isomorphism, though.

(iv) Let $\mathcal{C}^\infty(\mathbb{R})$ denote the space of functions from $\mathbb{R}$ to itself that have derivatives of any order. The map $D : \mathcal{C}^\infty(\mathbb{R}) \to \mathcal{C}^\infty(\mathbb{R})$, $Df = f'$, is linear. Why is it not an isomorphism?

Linear maps between finite-dimensional Euclidean spaces are most conveniently expressed in terms of matrices. If $A = [a_{ij}]$ is an $m \times n$ matrix, and $B = [b_{ij}]$ is an $n \times k$ matrix, the *product* matrix is the $m \times k$ matrix AB whose (i,j)-th entry is $\sum_{l=1}^{n} a_{il} b_{lj}$. For example,

$$\begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} -2 & 4 \end{bmatrix}, \qquad \begin{bmatrix} 1 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}.$$

One easily verifies that matrix multiplication has the following properties:

Proposition 1.2.1. *For appropriately sized matrices A, B, C, and $c \in \mathbb{R}$,*

(1) $A(BC) = (AB)C$;

(2) $(A + B)C = AC + BC$, *and* $A(B + C) = AB + AC$;

(3) $c(AB) = (cA)B = A(cB)$.

The expression "appropriately sized" means that the matrices are assumed to have the correct size for addition and multiplication; for instance, in the first identity, the number of columns of A equals the number of rows of B, and the number of columns of B equals the number of rows of C. Unlike scalar multiplication, though, $AB \neq BA$ in general (even when both products make sense). There is nevertheless a matrix that

plays the same role that 1 does in scalar multiplication: the $n \times n$ *identity matrix*

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

is the matrix whose (i,j)-th entry is 1 if $i = j$ and 0 otherwise. It follows from the definition of multiplication that if A is $m \times n$, then $AI_n = A$ and if B is $n \times m$, then $I_n B = B$. A (necessarily square) $n \times n$ matrix A is said to be *invertible* if there exists a matrix B such that $AB = BA = I_n$. If such a B exists (and in general it need not), then it is unique: indeed, if B and B' are two such, then $B' = B'I_n = B'(AB) = (B'A)B = I_n B = B$. We then call B the *inverse* of A, and denote it A^{-1}.

Examples 1.2.2. (i) The reader is invited to verify that the matrices

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \text{ and } B = \begin{bmatrix} -1 & 1 \\ 2 & -1 \end{bmatrix}$$

satisfy $AB = BA = I_2$, so that $B = A^{-1}$.

(ii) The matrix

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

is not invertible. One way of seeing this is by noting that AA equals the zero matrix $\mathbf{0}$ whose entries are all 0. If A did have an inverse, then we would have that $A = I_2 A = (A^{-1}A)A = A^{-1}(AA) = A^{-1}\mathbf{0} = \mathbf{0}$, which is not true.

Matrix multiplication provides many examples of linear transformations:

Example 1.2.3. Let A be an $m \times n$ matrix. For any $\boldsymbol{u} \in \mathbb{R}^n$, $A\boldsymbol{u}^T \in M_{1,m}$, so that $(A\boldsymbol{u}^T)^T \in \mathbb{R}^m$. By Proposition 1.2.1 (2), (3), and the fact that transposing is linear, *left multiplication L_A by A*,

$$\begin{aligned} L_A : \mathbb{R}^n &\to \mathbb{R}^m, \\ \boldsymbol{u} &\mapsto (A\boldsymbol{u}^T)^T \end{aligned}$$

is a linear transformation. We will shortly see that *every* linear transformation between Euclidean spaces is of this form.

A word about terminology: left multiplication by A may sound like a poor choice of words, since $L_A\boldsymbol{u} = \boldsymbol{u}A^T$, so that L_A is really right multiplication by the transpose of A. Nevertheless, as observed earlier, it is customary to identify elements of $\mathbb{R}^n$ with column rather than row matrices, and when doing so, we may write $L_A\boldsymbol{u} = A\boldsymbol{u}$. We will be deliberately vague as to which representation we use, but a general rule of thumb is that an element of $\mathbb{R}^n$ considered as a point is represented by a row matrix, whereas an element considered as a vector is represented as a column.

Definition 1.2.2. Let $L : V \to W$ be a linear transformation, and suppose $\mathcal{B} = (\boldsymbol{u}_1, \dots, \boldsymbol{u}_n)$ and $\mathcal{C} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_m)$ are ordered bases of V and W respectively. The *matrix of L with respect to $\mathcal{B}$ and $\mathcal{C}$* is the $m \times n$ matrix

$$[L]_{\mathcal{B},\mathcal{C}} = \begin{bmatrix} [L\boldsymbol{u}_1]_{\mathcal{C}} & \vdots & \cdots & \vdots & [L\boldsymbol{u}_n]_{\mathcal{C}} \end{bmatrix}$$

whose i-th column is the coordinate vector of $L\boldsymbol{u}_i$ with respect to $\mathcal{C}$, cf. Examples 1.2.1 (iii).

Example 1.2.4. Let P_n denote the $(n+1)$-dimensional space of polynomials of degree $\le n$ with its standard basis $\mathcal{S} = \{1, x, \dots, x^n\}$, and consider the derivative operator $D : P_2 \to P_1$. If $\mathcal{B}$ is the basis $\{1, 1+x\}$ of P_1, then $D1 = 0$, $Dx = 1$, and $Dx^2 = 2x = -2 \cdot 1 + 2(1+x)$. Thus,

$$[D]_{\mathcal{S},\mathcal{B}} = \begin{bmatrix} 0 & 1 & -2 \\ 0 & 0 & 2 \end{bmatrix}.$$

The following result implies that any linear transformation is entirely determined by its matrix with respect to any given bases; in fact, it says that once bases are fixed, the vector space of all linear transformations from V to W is isomorphic to the space of $m \times n$ matrices, where $\dim V = n$ and $\dim W = m$, see Exercise 1.16:

Theorem 1.2.1. *Let $L : V \to W$ be a linear transformation, and suppose $\mathcal{B} = \{\boldsymbol{u}_1, \dots, \boldsymbol{u}_n\}$ and $\mathcal{C} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_m\}$ are bases of V and W respectively. Then for any $\boldsymbol{u} \in V$,*

$$[L\boldsymbol{u}]_{\mathcal{C}} = [L]_{\mathcal{B},\mathcal{C}}[\boldsymbol{u}]_{\mathcal{B}}.$$

Proof. We begin by observing that if A is an $m \times n$ matrix with columns A_i, and $\boldsymbol{u} = [x_1 \dots x_n]^T \in \mathbb{R}^n$, then by definition of matrix multiplication, $A\boldsymbol{u} = x_1A_1 + \cdots + x_nA_n$. Now, if $\boldsymbol{v} = \sum a_i\boldsymbol{v}_i \in V$, then $[\boldsymbol{v}]_{\mathcal{B}} = [a_1 \dots a_n]^T$, and together with Examples 1.2.1 (iii),

$$[L\boldsymbol{v}]_{\mathcal{C}} = \Big[L\Big(\sum a_i\boldsymbol{v}_i\Big)\Big]_{\mathcal{C}} = \Big[\sum a_iL\boldsymbol{v}_i\Big]_{\mathcal{C}} = \sum a_i[L\boldsymbol{v}_i]_{\mathcal{C}} = [L]_{\mathcal{B},\mathcal{C}}[\boldsymbol{v}]_{\mathcal{B}},$$

as claimed. □

Corollary 1.2.1. (1) *If A is the matrix of a linear transformation $L : \mathbb{R}^n \to \mathbb{R}^m$ with respect to the standard bases, then $L = L_A$ (see Example 1.2.3).*

(2) *Let $L : V_1 \to V_2$, $T : V_2 \to V_3$ be linear, and $\mathcal{B}_i$ be a basis of V_i, $1 \le i \le 3$. Then $[T \circ L]_{\mathcal{B}_1,\mathcal{B}_3} = [T]_{\mathcal{B}_2,\mathcal{B}_3}[L]_{\mathcal{B}_1,\mathcal{B}_2}$. In particular, if $L : V_1 \to V_2$ is an isomorphism, then $[L^{-1}]_{\mathcal{B}_2,\mathcal{B}_1} = [L]^{-1}_{\mathcal{B}_1,\mathcal{B}_2}$.*

(3) (Change of basis) *If $\mathcal{B}$ and $\mathcal{C}$ are bases of V, then $[\boldsymbol{v}]_{\mathcal{C}} = [1_V]_{\mathcal{B},\mathcal{C}}[\boldsymbol{v}]_{\mathcal{B}}$ for any $\boldsymbol{v} \in V$.*

Proof. (1) We denote by the same letter $\mathcal{S}$ the standard bases in both Euclidean spaces. Recalling that a vector in Euclidean space equals its coordinate vector in the standard basis, Theorem 1.2.1 implies that for any $\boldsymbol{v} \in \mathbb{R}^n$,

$$L\boldsymbol{v} = [L\boldsymbol{v}]_{\mathcal{S}} = [L]_{\mathcal{S},\mathcal{S}}[\boldsymbol{v}]_{\mathcal{S}} = A\boldsymbol{v} = L_A\boldsymbol{v}.$$

(2) Given $\boldsymbol{u} \in V_1$,

$$[(T \circ L)\boldsymbol{u}]_{\mathcal{B}_3} = [T(L\boldsymbol{u})]_{\mathcal{B}_3} = [T]_{\mathcal{B}_2,\mathcal{B}_3}[L\boldsymbol{u}]_{\mathcal{B}_2} = [T]_{\mathcal{B}_2,\mathcal{B}_3}[L]_{\mathcal{B}_1,\mathcal{B}_2}[\boldsymbol{u}]_{\mathcal{B}_1}.$$

The second statement follows by taking $T = L^{-1}$, observing that $L^{-1} \circ L$ is the identity map, and the matrix $[1_{V_1}]_{\mathcal{B},\mathcal{B}}$ of the identity map with respect to any basis $\mathcal{B}$ is the identity matrix.
(3) is an immediate consequence of Theorem 1.2.1. □

In light of Theorem 1.2.1 (3), we refer to $[1_V]_{\mathcal{B},\mathcal{C}}$ as the *change of basis matrix* from the basis $\mathcal{B}$ to $\mathcal{C}$.

Examples 1.2.5. (i) Theorem 1.2.1 has many other applications that can be proved by arguments similar to the ones used above. Suppose, for example, that $L : V \to V$ is a linear transformation, and $\mathcal{B}$, $\mathcal{C}$ are two bases of V. How are the matrices of L with respect to the two bases related? To answer this, let $\boldsymbol{v} \in V$, and observe that

$$[L\boldsymbol{v}]_{\mathcal{C}} = [(1_V \circ L \circ 1_V)\boldsymbol{v}]_{\mathcal{C}} = [1_V]_{\mathcal{C},\mathcal{B}}[L]_{\mathcal{C},\mathcal{C}}[1_V]_{\mathcal{B},\mathcal{C}}[\boldsymbol{v}]_{\mathcal{B}},$$

so that if P denotes the change of basis matrix $[1_V]_{\mathcal{B},\mathcal{C}}$ from $\mathcal{B}$ to $\mathcal{C}$, then

$$[L]_{\mathcal{C},\mathcal{C}} = P^{-1}[L]_{\mathcal{B},\mathcal{B}}P. \tag{1.2.1}$$

For the sake of brevity, the matrix $[L]_{\mathcal{B},\mathcal{B}}$ of L in a given basis $\mathcal{B}$ will often be denoted by $[L]_{\mathcal{B}}$, so that (1.2.1) reads

$$[L]_{\mathcal{C}} = P^{-1}[L]_{\mathcal{B}}P.$$

Two $n \times n$ matrices A and B are said to be *similar* if there exists an invertible matrix P such that $A = P^{-1}BP$. This is easily seen to be an equivalence relation for the class of all $n \times n$ matrices (meaning (1) any A is similar to itself, (2) if A is similar to B, then B is similar to A, and (3) if A is similar to B and B is similar to C, then A is similar to C). (1.2.1) says that the matrices of a linear transformation $L : V \to V$ with respect to two different bases are similar.

(ii) The set $\mathcal{B} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3)$ where

$$\boldsymbol{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}, \quad \boldsymbol{v}_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix},$$

is an ordered basis of $\mathbb{R}^3$. Furthermore, $\boldsymbol{e}_1 = (1/2)\boldsymbol{v}_1 + (1/4)\boldsymbol{v}_2 - (1/4)\boldsymbol{v}_3$, $\boldsymbol{e}_2 = (-1/2)\boldsymbol{v}_1 + (1/4)\boldsymbol{v}_2 - (1/4)\boldsymbol{v}_3$, and $e_3 = (1/2)\boldsymbol{v}_1 - (1/4)\boldsymbol{v}_2 + (1/4)\boldsymbol{v}_3$, so that the change of basis matrix from the standard basis $\mathcal{S}$ to $\mathcal{B}$ is

$$\left[1_{\mathbb{R}^3}\right]_{\mathcal{S},\mathcal{B}} = \begin{bmatrix} [\boldsymbol{e}_1]_{\mathcal{B}} & [\boldsymbol{e}_2]_{\mathcal{B}} & [\boldsymbol{e}_3]_{\mathcal{B}} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & \frac{1}{4} \end{bmatrix}.$$

This means for example that the coordinate vector of $\begin{bmatrix}1 & 2 & 1\end{bmatrix}^T$ in the basis $\mathcal{B}$ equals

$$\begin{bmatrix}1\\2\\1\end{bmatrix}_{\mathcal{B}} = \left[1_{\mathbb{R}^3}\right]_{\mathcal{B},\mathcal{S}} \begin{bmatrix}1\\2\\1\end{bmatrix}_{\mathcal{S}} = \begin{bmatrix}\frac{1}{2} & -\frac{1}{2} & \frac{1}{2}\\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{4}\\ -\frac{1}{4} & \frac{3}{4} & \frac{1}{4}\end{bmatrix}\begin{bmatrix}1\\2\\1\end{bmatrix} = \begin{bmatrix}0\\\frac{1}{2}\\\frac{3}{2}\end{bmatrix},$$

which reflects the fact that $\begin{bmatrix}1 & 2 & 1\end{bmatrix}^T = (1/2)\boldsymbol{v}_2 + (3/2)\boldsymbol{v}_3$.

Definition 1.2.3. The *kernel* or *nullspace* $\ker L$ of a linear map $L : V \to W$ is the set of all $\boldsymbol{v} \in V$ that are mapped to $\boldsymbol{0} \in W$ by L. The *image* $\operatorname{Im} L$ of L is the collection of all elements of W that can be expressed as $L\boldsymbol{v}$ for some $\boldsymbol{v} \in V$. It is often denoted $L(V)$.

It is easily seen that kernel and image are subspaces of V and W respectively.

Theorem 1.2.2. *If $L : V \to W$ is linear, then* $\dim V = \dim \ker L + \dim \operatorname{Im} L$.

Proof. Let $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_k\}$ denote a basis of $\ker L$, and $\{\boldsymbol{w}_1, \dots, \boldsymbol{w}_r\}$ one of $\operatorname{Im} L$. For each $i = 1, \dots, r$, choose some $\boldsymbol{v}_i \in V$ such that $L\boldsymbol{v}_i = \boldsymbol{w}_i$. The claim follows once we establish that the set $S = \{\boldsymbol{u}_1, \dots, \boldsymbol{u}_k, \boldsymbol{v}_1, \dots, \boldsymbol{v}_r\}$ is a basis of V. To see that it spans V, take an arbitrary element $\boldsymbol{v} \in V$. By assumption, $L\boldsymbol{v} = \sum_i a_i \boldsymbol{w}_i$ for some scalars a_i. Since $L(\boldsymbol{v} - \sum_i a_i \boldsymbol{v}_i) = L\boldsymbol{v} - \sum_i a_i \boldsymbol{w}_i = \boldsymbol{0}$, the vector $\boldsymbol{v} - \sum_i a_i \boldsymbol{v}_i$ lies in the kernel of L and can therefore be expressed as a linear combination of the $\boldsymbol{u}_i$. In other words, $\boldsymbol{v}$ lies in the span of S. For linear independence, consider the equation

$$\sum_{i=1}^{k} a_i \boldsymbol{u}_i + \sum_{j=1}^{r} b_j \boldsymbol{v}_j = \boldsymbol{0}.$$

Applying L to both sides, and recalling that the first summand lies in the kernel of L, we obtain $\sum b_j \boldsymbol{w}_j = \boldsymbol{0}$. By linear independence of the $\boldsymbol{w}_j$, each b_j must vanish. But then only the first summand remains in the above equation, and by linear independence of the $\boldsymbol{u}_i$, each $a_i = 0$. □

The dimension of the kernel of L is called the *nullity* of L, that of its image the *rank* of L.

Example 1.2.6. Let $L : \mathbb{R}^3 \to \mathbb{R}^3$ be given by

$$L\begin{bmatrix}x\\y\\z\end{bmatrix} = \begin{bmatrix}x + z\\ y + 2z\\ 2x + y + 4z\end{bmatrix}.$$

L is linear, since it is left multiplication by the matrix

$$A = \begin{bmatrix}1 & 0 & 1\\ 0 & 1 & 2\\ 2 & 1 & 4\end{bmatrix}.$$

The columns of A are the image via L of the standard coordinate vectors, and they are linearly dependent: in fact, $L\boldsymbol{e}_3 = L\boldsymbol{e}_1 + 2L\boldsymbol{e}_2$. Thus, the image of L is spanned by the

first 2 columns of A. Since they are linearly independent, L has rank 2. On the other hand, $[\,x\,y\,z\,]^T$ belongs to the kernel of L iff $x + z = 0$, $y + 2z = 0$, and $2x + y + 4z = 0$. The last equation is just the sum of the second one with two times the first one, and can be discarded. The first two equations enable us to express x and y in terms of z, so that the kernel consists of all vectors of the form

$$\begin{bmatrix} -z \\ -2z \\ z \end{bmatrix} = z \begin{bmatrix} -1 \\ -2 \\ 1 \end{bmatrix}, \qquad z \in \mathbb{R},$$

and is therefore one-dimensional, as predicted by Theorem 1.2.2.

1.3 Determinants

Determinants play a crucial role in Linear Algebra, and by extension, in Calculus and Differential Geometry. Their properties are closely related to those of permutations: let $J_n = \{1, \ldots, n\}$. A *permutation* of J_n is a bijection $\sigma : J_n \to J_n$. It is often represented by

$$\begin{bmatrix} \sigma(1) & \sigma(2) & \ldots & \sigma(n) \end{bmatrix}.$$

For example,

$$\sigma = \begin{bmatrix} 2 & 4 & 1 & 3 \end{bmatrix}$$

is the permutation of J_4 such that $\sigma(1) = 2$, $\sigma(2) = 4$, $\sigma(3) = 1$, and $\sigma(4) = 3$. The composition $\sigma \circ \tau$ of two permutations σ and τ is again a permutation, denoted $\sigma\tau$ for brevity, and the collection of all permutations of J_n is denoted S_n. It is easy to see that S_n consists of $n!$ elements, where

$$n! = n(n-1)(n-2)\cdots 2,$$

since there are n possible choices in assigning the value of σ at, say, 1, then only $n - 1$ remaining choices for $\sigma(2)$, and so on.

A *transposition* is a permutation that interchanges two elements and leaves all the others fixed. We denote by (i, j) the transposition that interchanges i and j.

Lemma 1.3.1. *Every permutation is a product of transpositions.*

Proof. The term product in the statement actually means composition; it is suggested by the notation in use. The argument will be by induction on n. For $n = 1$, there is nothing to prove. So assume that every permutation of J_{n-1} is a product of transpositions, and consider $\sigma \in S_n$. Let $k = \sigma(n)$. Define τ to be the transposition (k, n) if $k \neq n$, and the identity if $k = n$. Then $\tau\sigma(n) = \tau(k) = n$, so $\tau\sigma \in S_{n-1}$ and by the induction hypothesis, there exist transpositions $\tau_1, \ldots, \tau_j$ such that $\tau\sigma = \tau_1 \cdots \tau_j$. Since τ equals its own inverse,

$$\sigma = \tau\tau_1 \cdots \tau_j$$

as claimed. □

The argument explicitly yields an algorithm for expressing a permutation as a product of transpositions. For example, the permutation

$$\sigma = \begin{bmatrix} 2 & 4 & 1 & 3 \end{bmatrix}$$

considered earlier can be written as $(3,4)(3,1)(1,2)$. There is no unique representation of a permutation as a product of transpositions, but the number of transpositions that appear is either always even or always odd. To see this, consider the polynomial p in n variables

$$p(x_1, \ldots, x_n) = \prod_{1 \le i < j \le n} (x_i - x_j).$$

For example, when $n = 3$,

$$p(x_1, x_2, x_3) = (x_1 - x_2)(x_1 - x_3)(x_2 - x_3).$$

Define a map

$$\begin{aligned} F : S_n \times \{\pm p\} &\to \{\pm p\}, \\ (\sigma, \pm p) &\mapsto \pm \prod_{i<j} (x_{\sigma(i)} - x_{\sigma(j)}). \end{aligned}$$

It is easily checked that $F(\mathrm{id}, \pm p) = \pm p$, and $F(\sigma\tau, \pm p) = F(\sigma, F(\tau, \pm p))$ for any $\tau, \sigma \in S_n$ (such a map is called an *action* of S_n on the set $\{\pm p\}$). Observe that for a transposition τ, $F(\tau, p) = -p$. Thus, if $\sigma \in S_n$ can be represented as the product of an even number of transpositions, then $F(\sigma, p) = p$, and consequently any other representation of σ also involves an even number of transpositions.

Definition 1.3.1. A permutation is said to be *even* if it can be written as a product of an even number of transpositions. A permutation that is not even is said to be *odd*. The *sign* $\varepsilon(\sigma)$ of σ is defined to be $+1$ if σ is even, and -1 if it is odd.

A map $M : V_1 \times \cdots \times V_k \to V$ from a Cartesian product of vector spaces V_i to a vector space V is said to be *multilinear* if it is linear in each component; i.e., if for any $1 \le i \le k$ and $\boldsymbol{v}_j \in V_j$, $j \ne i$, the map

$$\begin{aligned} V_i &\to V \\ \boldsymbol{v} &\mapsto M(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}, \boldsymbol{v}, \boldsymbol{v}_{i+1}, \ldots, \boldsymbol{v}_k) \end{aligned}$$

is linear. A multilinear map M is said to be *alternating* if

$$M(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_j, \ldots, \boldsymbol{v}_k) = -M(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_j, \ldots, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_k),$$

for all $1 \le i < j \le k$, $\boldsymbol{v}_l \in V_l$, $l = 1, \ldots, k$; equivalently, M is alternating if and only if $M(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_i \ldots, \boldsymbol{v}_j, \ldots, \boldsymbol{v}_k) = 0$ whenever $\boldsymbol{v}_i = \boldsymbol{v}_j$ for some $1 \le i < j \le k$: clearly, if M is alternating, then M evaluates to zero when applied to a list of vectors at least one of which is repeated. Conversely, if M evaluates to zero under these conditions, then

$$M(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_i + \boldsymbol{v}_j, \ldots, \boldsymbol{v}_i + \boldsymbol{v}_j, \ldots, \boldsymbol{v}_k) = 0.$$

By multilinearity, the expression on the left is a sum of four terms: $M(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_k)$, which by assumption vanishes; a similar term with $\boldsymbol{v}_j$ replacing $\boldsymbol{v}_i$ which also vanishes; and

$$M(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_j, \ldots, \boldsymbol{v}_k) + M(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_j, \ldots, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_k).$$

Since this must equal zero, the claim follows.

With these preliminaries out of the way, we are able to prove the following:

Theorem 1.3.1. *There exists a unique map* $\det : (\mathbb{R}^n)^n = \mathbb{R}^n \times \cdots \times \mathbb{R}^n \to \mathbb{R}$ *satisfying the following:*
(1) det *is multilinear;*
(2) det *is alternating;*
(3) $\det(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n) = 1$.

Proof. We begin with uniqueness. Let $\boldsymbol{v}_i = \sum_k a_{ki}\boldsymbol{e}_k$, $i = 1, \ldots, n$. Since det is multilinear,

$$\begin{aligned}\det(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) &= \det(a_{11}\boldsymbol{e}_1 + \cdots + a_{n1}\boldsymbol{e}_n, \ldots, a_{1n}\boldsymbol{e}_1 + \cdots + a_{nn}\boldsymbol{e}_n) \\ &= \sum_\sigma a_{\sigma(1)1} \cdots a_{\sigma(n)n} \det(\boldsymbol{e}_{\sigma(1)}, \ldots, \boldsymbol{e}_{\sigma(n)}),\end{aligned}$$

where the sum runs over all maps $\sigma : J_n \to J_n$. Now, if σ is not one-to-one, i.e., if $\sigma(i) = \sigma(j)$ for some $i \neq j$, then $\det(\boldsymbol{e}_{\sigma(1)}, \ldots, \boldsymbol{e}_{\sigma(n)}) = 0$ because det is alternating. Thus, the sum actually runs over all permutations of J_n. Finally, by properties (2), (3), and the definition of the sign of a permutation, $\det(\boldsymbol{e}_{\sigma(1)}, \ldots, \boldsymbol{e}_{\sigma(n)}) = \varepsilon(\sigma)$. Summarizing, if det is to satisfy the three properties listed, then

$$\det\left(\begin{bmatrix} a_{11} \\ \vdots \\ a_{n1} \end{bmatrix}, \ldots, \begin{bmatrix} a_{1n} \\ \vdots \\ a_{nn} \end{bmatrix}\right) = \sum_{\sigma \in S_n} \varepsilon(\sigma) a_{\sigma(1)1} \cdots a_{\sigma(n)n}, \tag{1.3.1}$$

which establishes uniqueness. To determine existence, it is enough to show that the above equation defines a map satisfying the three stated properties. For the first one, we compute

$$\begin{aligned}&\det\left(\begin{bmatrix} a_{11} \\ \vdots \\ a_{n1} \end{bmatrix}, \ldots, \begin{bmatrix} a_{1i} \\ \vdots \\ a_{ni} \end{bmatrix} + c\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}, \ldots, \begin{bmatrix} a_{1n} \\ \vdots \\ a_{nn} \end{bmatrix}\right) \\ &= \det\left(\begin{bmatrix} a_{11} \\ \vdots \\ a_{n1} \end{bmatrix}, \ldots, \begin{bmatrix} a_{1i} + cb_1 \\ \vdots \\ a_{ni} + cb_n \end{bmatrix} + \cdots \begin{bmatrix} a_{1n} \\ \vdots \\ a_{nn} \end{bmatrix}\right) \\ &= \sum_{\sigma \in S_n} \varepsilon(\sigma) a_{\sigma(1)1} \cdots (a_{\sigma(i)i} + cb_{\sigma(i)}) \cdots a_{\sigma(n)n}\end{aligned}$$

$$= \sum_{\sigma\in S_n} \varepsilon(\sigma) a_{\sigma(1)1}\cdots a_{\sigma(i)i}\cdots a_{\sigma(n)n} + c\sum_{\sigma\in S_n} \varepsilon(\sigma) a_{\sigma(1)1}\cdots b_{\sigma(i)}\cdots a_{\sigma(n)n}$$

$$= \det\left(\begin{bmatrix} a_{11}\\ \vdots\\ a_{n1}\end{bmatrix},\ldots,\begin{bmatrix} a_{1i}\\ \vdots\\ a_{ni}\end{bmatrix},\ldots\begin{bmatrix} a_{1n}\\ \vdots\\ a_{nn}\end{bmatrix}\right)$$

$$+ c\det\left(\begin{bmatrix} a_{11}\\ \vdots\\ a_{n1}\end{bmatrix},\ldots,\begin{bmatrix} b_{1}\\ \vdots\\ b_{n}\end{bmatrix},\ldots,\begin{bmatrix} a_{1n}\\ \vdots\\ a_{nn}\end{bmatrix}\right),$$

which shows that det is multilinear. For the second property, let $\tau \in S_n$, and $\boldsymbol{v}_i = \begin{bmatrix} a_{1i} & \ldots & a_{ni}\end{bmatrix}^T$. Then, with notation as above, and noting that $\varepsilon^2(\tau) = 1$, we have

$$\begin{aligned}\det(\boldsymbol{v}_{\tau(1)},\ldots,\boldsymbol{v}_{\tau(n)}) &= \sum_{\sigma\in S_n} \varepsilon(\sigma) a_{\sigma\circ\tau(1)1}\cdots a_{\sigma\circ\tau(n)n}\\ &= \varepsilon(\tau)\sum_{\sigma\in S_n} \varepsilon(\sigma)\varepsilon(\tau) a_{\sigma\circ\tau(1)1}\cdots a_{\sigma\circ\tau(n)n}\\ &= \varepsilon(\tau)\sum_{\sigma\circ\tau\in S_n} \varepsilon(\sigma\circ\tau) a_{\sigma\circ\tau(1)1}\cdots a_{\sigma\circ\tau(n)n}\\ &= \varepsilon(\tau)\det(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n).\end{aligned}$$

Notice that for the third equality, we used the fact that $S_n = \{\sigma\circ\tau \mid \sigma\in S_n\}$. Taking τ to be a transposition now shows that det is alternating. The last property is immediate. □

If A is an $n\times n$ matrix, we define its *determinant* to be the number $\det A = \det(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n)$, where $\boldsymbol{a}_i$ is the i-th column of A. Thus, by (1.3.1),

$$\det\begin{bmatrix} a_{11} & \ldots & a_{1_n}\\ \vdots & \ldots & \vdots\\ a_{n1} & \ldots & a_{nn}\end{bmatrix} = \sum_{\sigma\in S_n} \varepsilon(\sigma) a_{\sigma(1)1}\cdots a_{\sigma(n)n}. \tag{1.3.2}$$

For example, in the case of a 2×2 matrix, S_2 consists only of the identity and the transposition $(1,2)$, so that

$$\det\begin{bmatrix} a & b\\ c & d\end{bmatrix} = ad - bc.$$

The definition given here is not very convenient for computing determinants of larger matrices. In order to describe a different approach for $n > 1$, let us denote by A_{ij} the $(n-1)\times(n-1)$ matrix obtained by deleting row i and column j from A.

Theorem 1.3.2 (Expansion along the i-th row).

$$\det A = (-1)^{i+1} a_{i1}\det A_{i1} + \cdots + (-1)^{i+n} a_{in}\det A_{in}.$$

Proof. According to Theorem 1.3.1, it suffices to show that the function defined by the right side of the above identity satisfies the three properties stated in the theorem. To check linearity in the k-th column of A, let $\tilde{A}$ denote the matrix obtained by adding $\boldsymbol{b} \in \mathbb{R}^n$ to the k-th column of A, and $\bar{A}$ the one obtained by replacing the k-th column of A by $\boldsymbol{b}$. We must show that

$$\sum_j (-1)^{i+j} \tilde{a}_{ij} \det \tilde{A}_{ij} = \sum_j (-1)^{i+j} a_{ij} \det A_{ij} + \sum_j (-1)^{i+j} \bar{a}_{ij} \det \bar{A}_{ij}. \tag{1.3.3}$$

Now, if $j \neq k$, then $\tilde{A}_{ij}$ is obtained by adding $(b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_n)^T$ to a column of A_{ij}, whereas $\bar{A}_{ij}$ is obtained by replacing that column by the same vector. Thus, $\det \tilde{A}_{ij} = \det A_{ij} + \det \bar{A}_{ij}$; furthermore, $a_{ij} = \tilde{a}_{ij} = \bar{a}_{ij}$, so that

$$\tilde{a}_{ij} \det \tilde{A}_{ij} = a_{ij} \det A_{ij} + \bar{a}_{ij} \det \bar{A}_{ij}, \qquad j \neq k. \tag{1.3.4}$$

When $j = k$ on the other hand, $A_{ik} = \tilde{A}_{ik} = \bar{A}_{ik}$, and $\tilde{a}_{ik} = a_{ik} + b_k = a_{ik} + \bar{a}_{ik}$, so that

$$\tilde{a}_{ik} \det \tilde{A}_{ik} = a_{ik} \det A_{ik} + \bar{a}_{ik} \det \bar{A}_{ik}. \tag{1.3.5}$$

Identities (1.3.4) and (1.3.5) then imply (1.3.3). A similar argument shows that if B denotes the matrix obtained by multiplying the k-th column of A by c, then

$$\sum_j (-1)^{i+j} b_{ij} \det B_{ij} = c \sum_j (-1)^{i+j} a_{ij} \det A_{ij}.$$

This shows that the expansion along the i-th row is multilinear.
To see that the expansion is alternating, consider first the matrix B obtained by interchanging two adjacent columns from A, say columns k and $k+1$. Then for $j \neq k, k+1$, B_{ij} is obtained from A_{ij} by interchanging two columns, so that $\det B_{ij} = -\det A_{ij}$, and

$$b_{ij} \det B_{ij} = a_{ij} \det B_{ij} = -a_{ij} \det A_{ij}.$$

For the remaining two cases, observe that $A_{i(k+1)} = B_{ik}$, $B_{i(k+1)} = A_{ik}$, and similar identities hold for lower case letters. Thus,

$$b_{ik} \det B_{ik} = a_{i(k+1)} \det A_{i(k+1)}, \quad a_{ik} \det A_{ik} = b_{i(k+1)} \det B_{i(k+1)}.$$

However, the terms have opposite sign in the corresponding sums, since one is in column k, and the other in column $k+1$. Thus, $\det B = -\det A$. Finally, if B is obtained by interchanging two non-adjacent columns i and j from A, it is also obtained by interchanging adjacent columns an odd number of times: assuming, without loss of generality, that $i < j$, interchange column i with the column to its right $j - i$ times in succession, then interchange column j with the column to its left $j - i - 1$ times. This shows that the sum is alternating.
It remains to check that the right side of the identity equals 1 when A is the $n \times n$ identity matrix I_n. Since $a_{ij} = 0$ unless $i = j$, the only remaining term in the sum is $(-1)^{2i} a_{ii} \det A_{ii} = \det I_{n-1}$. Applying this repeatedly yields $\det I_n = \det I_{n-1} = \cdots = \det I_1 = 1$. □

Theorem 1.3.3. *If A is a square matrix, then* $\det A = \det(A^T)$.

Proof. Recall that for an $n \times n$ matrix A,

$$\det A = \sum_{\sigma \in S_n} \varepsilon(\sigma) a_{\sigma(1)1} \dots a_{\sigma(n)n}.$$

Now, if $\sigma(i) = j$, then $i = \sigma^{-1}(j)$, and $a_{\sigma(i)i} = a_{j\sigma^{-1}(j)}$. Furthermore, $a_{j\sigma^{-1}(j)}$ occurs exactly once for each integer j, so that

$$a_{\sigma(1)1} \dots a_{\sigma(n)n} = a_{1\sigma^{-1}(1)} \cdots a_{n\sigma^{-1}(n)}.$$

Finally, $\varepsilon(\sigma^{-1}) = \varepsilon(\sigma)$ (since $\varepsilon(\sigma)\varepsilon(\sigma^{-1}) = \varepsilon(\sigma \circ \sigma^{-1}) = \varepsilon(\mathrm{id}) = 1$), and the map that assigns to each $\sigma \in S_n$ its inverse is a bijection. Thus,

$$\begin{aligned} \det A &= \sum_{\sigma^{-1} \in S_n} \varepsilon(\sigma^{-1}) a_{1\sigma^{-1}(1)} \cdots a_{n\sigma^{-1}(n)} = \sum_{\sigma \in S_n} \varepsilon(\sigma) a_{1\sigma(1)} \cdots a_{n\sigma(n)} \\ &= \det(A^T). \end{aligned}$$

□

Corollary 1.3.1 (Expansion along j-th column).

$$\det A = (-1)^{1+j} a_{1j} \det(A_{1j}) + \cdots + (-1)^{n+j} a_{nj} \det(A_{nj}).$$

Proof. Let B denote the transpose of A. Thus, B_{ji} is the matrix obtained by deleting row j and column i from A^T; i.e., it is the transpose of the matrix obtained by deleting row i and column j from A. By Theorem 1.3.3, $\det B_{ji} = \det A_{ij}$. Using Theorem 1.3.3 once again, and expanding along the j-th row, we have

$$\begin{aligned} \det A = \det B &= (-1)^{1+j} b_{j1} \det(B_{j1}) + \cdots + (-1)^{n+j} b_{jn} \det(B_{nj}) \\ &= (-1)^{1+j} a_{1j} \det(A_{1j}) + \cdots + (-1)^{n+j} a_{nj} \det(A_{nj}). \end{aligned}$$

□

For example, to compute the determinant of the three by three matrix

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 1 & 2 \\ 1 & 4 & 1 \end{bmatrix},$$

we can expand along the second row (which has the only zero) to obtain

$$1 \det \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} - 2 \det \begin{bmatrix} 1 & 2 \\ 1 & 4 \end{bmatrix} = 2 - 2 \cdot 2 = -2.$$

Better yet, the multilinear and alternating properties imply that adding a multiple of a row to another row (or a multiple of a column to another column) does not change the determinant. In the previous example, we may therefore subtract row 1 from row 3, and expand along the first column to get

$$\det A = \det \begin{bmatrix} 1 & 2 & -1 \\ 0 & 1 & 2 \\ 0 & 2 & 2 \end{bmatrix} = 1 \det \begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix} = -2.$$

Even though the determinant of a sum is not the sum of the determinants, products are another matter:

Theorem 1.3.4. $\det(AB) = \det A \det B$.

Proof. Suppose both matrices are $n \times n$. As observed in the proof of Theorem 1.2.1, the k-th column of AB equals

$$A\boldsymbol{b}_k = b_{1k}\boldsymbol{a}_1 + \cdots + b_{nk}\boldsymbol{a}_n,$$

where $\boldsymbol{b}_k$ denotes the k-th column of B, and similarly, $\boldsymbol{a}_i$ is the i-th column of A. Since the determinant is multilinear and alternating, we obtain, just as in the proof of Theorem 1.3.1,

$$\begin{aligned}\det(AB) &= \det(b_{11}\boldsymbol{a}_1 + \cdots + b_{n1}\boldsymbol{a}_n, \ldots, b_{1n}\boldsymbol{a}_1 + \cdots + b_{nn}\boldsymbol{a}_n)\\ &= \sum_\sigma b_{\sigma(1)1}\cdots b_{\sigma(n)n}\det(\boldsymbol{a}_{\sigma(1)}, \ldots, \boldsymbol{a}_{\sigma(n)})\\ &= \sum_\sigma \varepsilon(\sigma) b_{\sigma(1)1}\cdots b_{\sigma(n)n}\det(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)\\ &= \det B \det A.\end{aligned}$$

□

As a special case, if A is an invertible matrix, then $\det(A^{-1}) = 1/\det A$, since $AA^{-1} = I_n$ and the identity matrix has determinant 1. Another important consequence is that similar matrices (see Example 1.2.5) have the same determinant. In particular, the determinant of the matrix $[L]_{\mathcal{B},\mathcal{B}}$ of a linear transformation $L : V \to V$ in a basis $\mathcal{B}$ equals that of the matrix of L in any other basis. This justifies the following:

Definition 1.3.2. The *determinant* of a linear transformation $L : V \to V$ is the determinant of the matrix $[L]_{\mathcal{B},\mathcal{B}}$ of L with respect to any given basis $\mathcal{B}$ of V.

Determinants are often useful for, well, determining whether a linear map is an isomorphism:

Theorem 1.3.5. *Let V be a finite-dimensional vector space. A linear transformation $L : V \to V$ is an isomorphism if and only if* $\det L \neq 0$.

Proof. If L is an isomorphism, then it has an inverse L^{-1}, and the product $\det L \det(L^{-1}) = \det(L \circ L^{-1}) = \det(1_V) = 1$, so L cannot have vanishing determinant. If, on the other hand, L is not an isomorphism, then by Theorem 1.2.2, there exists a nonzero $\boldsymbol{v} \in V$ such that $L\boldsymbol{v} = \boldsymbol{0}$. Let $\mathcal{B} = \boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ denote an ordered basis of V, and write $\boldsymbol{v}$ as a linear combination $\sum_i c_i\boldsymbol{v}_i$ of the basis elements. Notice that not all coefficients vanish. Then

$$\boldsymbol{0} = [L\boldsymbol{v}]_{\mathcal{B}} = [L(\sum_i c_i\boldsymbol{v}_i)]_{\mathcal{B}} = \sum_i c_i[L\boldsymbol{v}_i]_{\mathcal{B}}.$$

This says that the columns $[L\boldsymbol{v}_i]_{\mathcal{B}}$ of the matrix $[L]_{\mathcal{B},\mathcal{B}}$ of L in the basis $\mathcal{B}$ are linearly dependent. But any square matrix with linearly dependent columns (or rows) has vanishing determinant: Indeed, suppose A has linearly dependent columns $\boldsymbol{a}_i$ (for rows,

consider the transpose of A instead). Then, one of the columns, say, $\boldsymbol{a}_j$, can be written as a linear combination

$$\boldsymbol{a}_j = \sum_{i \neq j} \alpha_i \boldsymbol{a}_i$$

of the others. By linearity of det in each column,

$$\det A = \det(\boldsymbol{a}_1, \dots, \boldsymbol{a}_j, \dots, \boldsymbol{a}_n) = \sum_{i \neq j} \alpha_i \det(\boldsymbol{a}_1, \dots, \boldsymbol{a}_i, \dots, \boldsymbol{a}_n).$$

Each determinant in this sum is the determinant of a matrix with two equal columns (the i-th and the j-th). Such a determinant must vanish by the alternating property (interchanging columns i and j changes the sign of the determinant, but the matrix remains the same). This establishes the result. □

The proof simplifies considerably if one uses the fact that given any nonzero $\boldsymbol{v} \in V$, there is a basis of V containing $\boldsymbol{v}$ (this will be proved in the next section). In fact, if L is not an isomorphism, consider any nonzero vector in the kernel of L, and extend it to a basis. The matrix of L in that basis then has a zero column and therefore also zero determinant.

Theorem 1.3.6. *Let V be a finite-dimensional vector space. For any linear transformation $L : V \to V$, there exists a linear map $\tilde{L} : V \to V$ such that*

$$L \circ \tilde{L} = (\det L) 1_V.$$

In particular, if L is invertible, then $L^{-1} = (1/\det L)\tilde{L}$, and if not, then $L \circ \tilde{L} = 0$.

Proof. Consider any ordered basis $\mathcal{B}$ of V and denote by $A = (a_{ij})$ the matrix of L with respect to $\mathcal{B}$. For each i and j between 1 and $n = \dim V$, let A_{ij} be the $(n-1) \times (n-1)$ matrix obtained by deleting row i and column j from A, and define an $n \times n$ matrix $\tilde{A} = (\tilde{a}_{ij})$ by setting

$$\tilde{a}_{ij} = (-1)^{i+j} \det A_{ji}.$$

A straightforward computation yields

$$A\tilde{A} = (\det A) I_n. \tag{1.3.6}$$

The result now follows if we let $\tilde{L}$ be the linear map whose matrix in the basis $\mathcal{B}$ equals $\tilde{A}$. □

The map $\tilde{L}$ constructed in the above proof will be called the *linear map adjugate to L*. It can be constructed without using a basis, but this requires extra work.

1.4 Euclidean spaces

An elementary concept that is used throughout Calculus is the distance $|a - b|$ between two points a and b in the real line. We now generalize this concept to $\mathbb{R}^n$. The

(standard) *inner product* of $\boldsymbol{a}$ and $\boldsymbol{b} \in \mathbb{R}^n$ is the number

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{i=1}^{n} u^i(\boldsymbol{a}) u^i(\boldsymbol{b}).$$

This is nothing but the entry of the 1×1 matrix $\boldsymbol{a}^T \cdot \boldsymbol{b}$. The vector space $\mathbb{R}^n$ together with this inner product is called *n-dimensional Euclidean space*. Some authors use this terminology for a larger class of spaces, which we call inner product spaces, to be introduced shortly. The *norm* of $\boldsymbol{a}$ is

$$|\boldsymbol{a}| = \langle \boldsymbol{a}, \boldsymbol{a} \rangle^{1/2}.$$

When $n = 1$, the norm of $\boldsymbol{a}$ is just its absolute value.

Theorem 1.4.1. *If $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^n$, and $a \in \mathbb{R}$, then*
(1) $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \langle \boldsymbol{b}, \boldsymbol{a} \rangle$*;*
(2) $\langle a\boldsymbol{a} + \boldsymbol{b}, \boldsymbol{c} \rangle = a\langle \boldsymbol{a}, \boldsymbol{b} \rangle + \langle \boldsymbol{b}, \boldsymbol{c} \rangle$*;*
(3) $|\boldsymbol{a}| \geq 0$*, and* $|\boldsymbol{a}| = 0$ *if and only if* $\boldsymbol{a} = \boldsymbol{0}$*;*
(4) $|a\boldsymbol{a}| = |a||\boldsymbol{a}|$*;*
(5) (Cauchy-Schwarz inequality) $|\langle \boldsymbol{a}, \boldsymbol{b} \rangle| \leq |\boldsymbol{a}||\boldsymbol{b}|$*;*
(6) (Triangle inequality) $|\boldsymbol{a} + \boldsymbol{b}| \leq |\boldsymbol{a}| + |\boldsymbol{b}|$.

Proof. The first four statements are obvious. For the Cauchy-Schwarz inequality, let $a_i = u^i(\boldsymbol{a})$, $b_i = u^i(\boldsymbol{b})$. We may assume that $\boldsymbol{b} \neq \boldsymbol{0}$, since otherwise the conclusion is trivial. Now,

$$\begin{aligned} 0 &\leq \sum_i \left(|\boldsymbol{b}|^2 a_i - \langle \boldsymbol{a}, \boldsymbol{b} \rangle b_i \right)^2 \\ &= |\boldsymbol{b}|^4 \sum a_i^2 + \langle \boldsymbol{a}, \boldsymbol{b} \rangle^2 \sum b_i^2 - 2|\boldsymbol{b}|^2 \langle \boldsymbol{a}, \boldsymbol{b} \rangle \sum a_i b_i \\ &= |\boldsymbol{b}|^4 |\boldsymbol{a}|^2 + |\boldsymbol{b}|^2 \langle \boldsymbol{a}, \boldsymbol{b} \rangle^2 - 2|\boldsymbol{b}|^2 \langle \boldsymbol{a}, \boldsymbol{b} \rangle^2 \\ &= |\boldsymbol{b}|^2 \left(|\boldsymbol{a}|^2 |\boldsymbol{b}|^2 - \langle \boldsymbol{a}, \boldsymbol{b} \rangle^2 \right). \end{aligned}$$

Since $|\boldsymbol{b}|^2 > 0$, $|\boldsymbol{a}|^2 |\boldsymbol{b}|^2 - \langle \boldsymbol{a}, \boldsymbol{b} \rangle^2 \geq 0$, which is the desired inequality.
The triangle inequality is a direct consequence of the Cauchy-Schwarz inequality:

$$\begin{aligned} |\boldsymbol{a} + \boldsymbol{b}|^2 &= \langle \boldsymbol{a} + \boldsymbol{b}, \boldsymbol{a} + \boldsymbol{b} \rangle = |\boldsymbol{a}|^2 + |\boldsymbol{b}|^2 + 2\langle \boldsymbol{a}, \boldsymbol{b} \rangle \leq |\boldsymbol{a}|^2 + |\boldsymbol{b}|^2 + 2|\boldsymbol{a}||\boldsymbol{b}| \\ &= (|\boldsymbol{a}| + |\boldsymbol{b}|)^2. \end{aligned}$$

□

Corollary 1.4.1. $||\boldsymbol{a}| - |\boldsymbol{b}|| \leq |\boldsymbol{a} - \boldsymbol{b}|$ *for* $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$.

Proof. It is enough to show that $|\boldsymbol{a}| - |\boldsymbol{b}| \leq |\boldsymbol{a} - \boldsymbol{b}|$ (why?). But this follows from the triangle inequality, since

$$|\boldsymbol{a}| = |\boldsymbol{a} - \boldsymbol{b} + \boldsymbol{b}| \leq |\boldsymbol{a} - \boldsymbol{b}| + |\boldsymbol{b}|.$$

□

Definition 1.4.1. The *operator norm* of a linear transformation $L : \mathbb{R}^n \to \mathbb{R}^m$ is

$$|L| = \sup\{|L\boldsymbol{u}| \mid |\boldsymbol{u}| = 1\},$$

where sup denotes the supremum or least upper bound of a set as introduced in Section 1.7.

Notice that the norm of L is always a finite number: If $\boldsymbol{u} = \sum_i a_i \boldsymbol{e}_i$ has norm 1, then each $|a_i| \leq 1$, so that

$$|L\boldsymbol{u}| = |\sum_i a_i L\boldsymbol{e}_i| \leq \sum_i |a_i||L\boldsymbol{e}_i| \leq \sum_i |L\boldsymbol{e}_i|,$$

and consequently $|L| \leq \sum_i |L\boldsymbol{e}_i|$.

It also follows from the definition that for all $\boldsymbol{u} \in \mathbb{R}^n$,

$$|L\boldsymbol{u}| \leq |L||\boldsymbol{u}|. \tag{1.4.1}$$

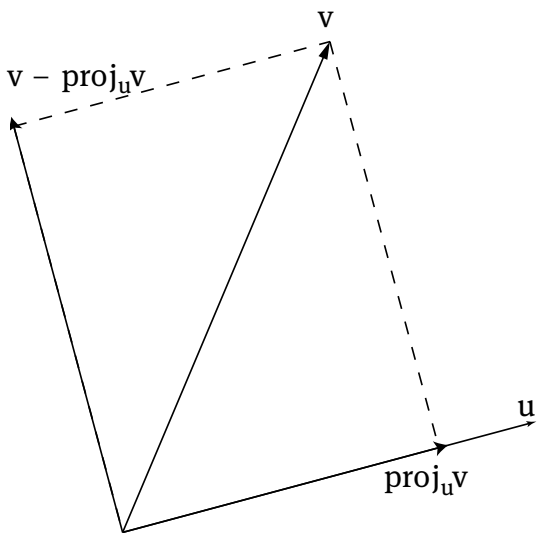
More generally, an *inner product* on a vector space V is a map $\langle , \rangle : V \times V \to \mathbb{R}$ that satisfies (1), (2), and (3) in Theorem 1.4.1. In this case, $(V, \langle , \rangle)$ is called an *inner product space*. $\boldsymbol{v}, \boldsymbol{w} \in V$ are said to be *orthogonal* if $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = 0$. Given a nonzero vector $\boldsymbol{u} \in V$, the *projection of* $\boldsymbol{v} \in V$ *along* $\boldsymbol{u}$ is the vector

$$\text{proj}_{\boldsymbol{u}} \boldsymbol{v} = \frac{\langle \boldsymbol{u}, \boldsymbol{v} \rangle}{\langle \boldsymbol{u}, \boldsymbol{u} \rangle} \boldsymbol{u}.$$

Any $\boldsymbol{v} \in V$ decomposes uniquely as the sum

$$\boldsymbol{v} = \text{proj}_{\boldsymbol{u}} \boldsymbol{v} + (\boldsymbol{v} - \text{proj}_{\boldsymbol{u}} \boldsymbol{v})$$

of a vector parallel to $\boldsymbol{u}$ (namely its projection along $\boldsymbol{u}$) and a vector orthogonal to $\boldsymbol{u}$.



A basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ of an inner product space is said to be *orthonormal* if $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \delta_{ij}$(the *Kronecker delta* δ_{ij} is the symbol that equals 1 if $i = j$ and 0 otherwise); in

other words, a basis is orthonormal if it consists of unit vectors any two of which are orthogonal.

The following theorem provides an algorithm for obtaining an orthogonal basis from an arbitrary one. To get an orthonormal one, it only remains to divide each basis element by its length.

Theorem 1.4.2 (Gram-Schmidt orthogonalization). *Suppose* $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_n\}$ *denotes a basis of an inner product space* V. *Then* $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$, *where*

$$\begin{aligned}
\boldsymbol{v}_1 &= \boldsymbol{u}_1, \\
\boldsymbol{v}_2 &= \boldsymbol{u}_2 - \operatorname{proj}_{\boldsymbol{v}_1} \boldsymbol{u}_2, \\
\boldsymbol{v}_3 &= \boldsymbol{u}_3 - \operatorname{proj}_{\boldsymbol{v}_1} \boldsymbol{u}_3 - \operatorname{proj}_{\boldsymbol{v}_2} \boldsymbol{u}_3, \\
&\vdots \\
\boldsymbol{v}_n &= \boldsymbol{u}_n - \sum_{i=1}^{n-1} \operatorname{proj}_{\boldsymbol{v}_i} \boldsymbol{u}_n,
\end{aligned}$$

is an orthogonal basis of V.

Proof. By construction, each $\boldsymbol{v}_i$ is orthogonal to $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i-1}$, so that any two of them are orthogonal. Moreover,

$$\boldsymbol{u}_i = \boldsymbol{v}_i + \sum_{l=1}^{i-1} \operatorname{proj}_{\boldsymbol{v}_l} \boldsymbol{u}_i \in \operatorname{span}\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_i\},$$

so that $\mathcal{B} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ spans V. But then $\mathcal{B}$ is linearly independent, for otherwise a strict subset of $\mathcal{B}$ would span V, contradicting the fact that $\dim V = n$. Thus, $\mathcal{B}$ is an orthogonal basis of V. □

Example 1.4.1. Let $\boldsymbol{u}_1 = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}^T$, $\boldsymbol{u}_2 = \begin{bmatrix} 2 & 1 & 0 \end{bmatrix}^T$, and V the subspace of $\mathbb{R}^3$ spanned by $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. Then $\operatorname{proj}_{\boldsymbol{u}_1} \boldsymbol{u}_2 = (\langle \boldsymbol{u}_1, \boldsymbol{u}_2 \rangle / |\boldsymbol{u}_1|^2)\boldsymbol{u}_1 = \boldsymbol{u}_1$, and $\boldsymbol{v}_2 = \boldsymbol{u}_2 - \boldsymbol{u}_1 = \begin{bmatrix} 1 & 1 & -1 \end{bmatrix}^T$ is a vector in V orthogonal to $\boldsymbol{u}_1$. Thus, the vectors

$$\boldsymbol{w}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \qquad \boldsymbol{w}_2 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$$

form an orthonormal basis of V. If we wish to extend this basis to an orthonormal basis of $\mathbb{R}^3$, we need only find a vector that does not belong to V, say, $\boldsymbol{e}_3$, and apply the Gram-Schmidt process to it:

$$\boldsymbol{v}_3 = \boldsymbol{e}_3 - \operatorname{proj}_{\boldsymbol{w}_1} \boldsymbol{e}_3 - \operatorname{proj}_{\boldsymbol{w}_2} \boldsymbol{e}_3 = \boldsymbol{e}_3 - \boldsymbol{w}_1 + \boldsymbol{w}_2 = \begin{bmatrix} -1/6 \\ 1/3 \\ 1/6 \end{bmatrix}.$$

$\{\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{v}_3/|\boldsymbol{v}_3|\}$ is then an orthonormal basis of $\mathbb{R}^3$.

One reason orthonormal bases are useful is that it is quite easy to compute coordinate vectors in them:

Theorem 1.4.3. *If $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ is an orthonormal basis of V, then for any $\boldsymbol{v} \in V$,*

$$\boldsymbol{v} = \langle \boldsymbol{v}, \boldsymbol{v}_1 \rangle \boldsymbol{v}_1 + \cdots + \langle \boldsymbol{v}, \boldsymbol{v}_n \rangle \boldsymbol{v}_n.$$

Thus, the coordinate vector of $\boldsymbol{v}$ in the ordered basis $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ has $\langle \boldsymbol{v}, \boldsymbol{v}_i \rangle$ as i-th entry.

Proof. By assumption $\boldsymbol{v} = \sum a_i \boldsymbol{v}_i$ for unique scalars $a_1, \dots, a_n$. Taking the inner product on both sides with $\boldsymbol{v}_j$ yields

$$\langle \boldsymbol{v}, \boldsymbol{v}_j \rangle = \sum_{i=1}^{n} a_i \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \sum_{i=1}^{n} a_i \, \delta_{ij} = a_j. \qquad \square$$

It is now easy to see that the Cauchy-Schwarz inequality is valid in any inner product space: just replace a_i in the proof by $\langle \boldsymbol{a}, \boldsymbol{v}_i \rangle$, where $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ is an orthonormal basis, and similarly for b_i.

Definition 1.4.2. Let A be a nonempty subset of an inner product space V. The *orthogonal complement* of A is the set

$$A^{\perp} = \{\boldsymbol{v} \in V \mid \langle \boldsymbol{v}, \boldsymbol{a} \rangle = 0 \text{ for every } \boldsymbol{a} \in A\}.$$

Linearity of the inner product in each entry implies that $A^{\perp}$ is a subspace of V. For the same reason, $(\operatorname{span} A)^{\perp} = A^{\perp}$.

In general, if W_1, W_2 are subspaces of a vector space V, then the set $W_1 + W_2 = \{\boldsymbol{v}_1 + \boldsymbol{v}_2 \mid \boldsymbol{v}_i \in W_i\}$ is easily seen to be a subspace of V, called the *sum of W_1 and W_2*. If $W_1 \cap W_2 = \{0\}$, this sum is called a *direct sum*, and is denoted $W_1 \oplus W_2$. In this case, any $\boldsymbol{v} \in W_1 \oplus W_2$ can be written in one and only one way as a sum of a vector in W_1 and a vector in W_2: indeed if $\boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2 = \boldsymbol{v}_1' + \boldsymbol{v}_2'$ with $\boldsymbol{v}_i, \boldsymbol{v}_i' \in W_i$, then $\boldsymbol{v}_1 - \boldsymbol{v}_1' = \boldsymbol{v}_2 - \boldsymbol{v}_2'$. Since the left side is in W_1 and the right side in W_2, both terms vanish, and $\boldsymbol{v}_i = \boldsymbol{v}_i'$ as claimed.

Proposition 1.4.1. *If A is a nonempty subset of an inner product space V, then $V = (\operatorname{span} A) \oplus A^{\perp}$.*

Proof. If $\boldsymbol{v} \in (\operatorname{span} A) \cap A^{\perp}$, then $\boldsymbol{v}$ is orthogonal to itself, and must then equal $\mathbf{0}$. It therefore remains to show that $V \subset (\operatorname{span} A) + A^{\perp}$. Given $\boldsymbol{v} \in V$, and an orthononormal basis $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\}$ of span A, write

$$\boldsymbol{v} = \boldsymbol{u}_1 + \boldsymbol{u}_2, \qquad \boldsymbol{u}_1 = \sum_{i=1}^{k} \operatorname{proj}_{\boldsymbol{v}_i} \boldsymbol{v}, \quad \boldsymbol{u}_2 = \boldsymbol{v} - \boldsymbol{u}_1. \tag{1.4.2}$$

By definition $\boldsymbol{u}_1 \in \operatorname{span} A$. For any $j = 1, \dots, k$,

$$\langle \boldsymbol{u}_2, \boldsymbol{v}_j \rangle = \langle \boldsymbol{v} - \sum_{i=1}^{k} \operatorname{proj}_{\boldsymbol{v}_i} \boldsymbol{v}, \boldsymbol{v}_j \rangle = \langle \boldsymbol{v}, \boldsymbol{v}_j \rangle - \langle \operatorname{proj}_{\boldsymbol{v}_j} \boldsymbol{v}, \boldsymbol{v}_j \rangle = 0,$$

so that $\boldsymbol{u}_2$ is orthogonal to every $\boldsymbol{v}_j$ and therefore to every element in span A; i.e., $\boldsymbol{u}_2 \in A^\perp$. □

The (necessarily unique) vector $\boldsymbol{u}_1$ in (1.4.2) is called the *orthogonal projection of* $\boldsymbol{v}$ *onto* span A. When A consists of a single vector $\boldsymbol{u}$, it coincides with what we defined earlier as the projection of $\boldsymbol{v}$ along $\boldsymbol{u}$.

Incidentally, we have in essence established the following:

Proposition 1.4.2. *If A is a linearly independent subset of $\mathbb{R}^n$, then there exists a basis of $\mathbb{R}^n$ that contains A.*

Proof. Let $\mathcal{B}$ denote a basis of $A^\perp$. Then $A \cup \mathcal{B}$ is a basis as in the statement. □

It should be noted that this is true in any vector space, since such a space is isomorphic to some Euclidean space. In fact, any two spaces with the same dimension are isomorphic, as observed in Examples 1.2.1 (iii). Such an isomorphism depends of course on the choice of bases. More to the point, the above proposition can be proved without introducing an inner product. The argument given here is, however, sufficient for our purposes.

The following concept too does not require the existence of an inner product:

Definition 1.4.3. The *dual space* of a vector space V is the collection V^* of all linear maps $V \to \mathbb{R}$.

An element of V^* is also called a *one-form on* V. These maps can be added in the usual way, as well as multiplied by scalars, thereby inducing a vector space structure on V^*. If $\mathcal{B} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ denotes a basis of V, then the map

$$\begin{aligned} V^* &\to M_{n,1}, \\ \alpha &\mapsto \begin{bmatrix} \alpha(\boldsymbol{v}_1) & \cdots & \alpha(\boldsymbol{v}_n) \end{bmatrix} \end{aligned}$$

is by definition linear, surjective, and has zero kernel. In particular, V^* has the same dimension as V and is therefore isomorphic to V. Although, as noted earlier, isomorphisms between spaces of the same dimension depend in general on the choice of bases, in this case there is a canonical one, provided V is an inner product space:

Theorem 1.4.4. *Let V be a finite-dimensional inner product space. The map $\flat : V \to V^*$, given by $\boldsymbol{u}^\flat(\boldsymbol{v}) = \langle \boldsymbol{u}, \boldsymbol{v} \rangle$, is an isomorphism.*

Proof. The map is clearly linear, and since the spaces have the same dimension, it suffices to check its kernel is zero. But if $\boldsymbol{u} \in \ker \flat$, then $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$ for all $\boldsymbol{v} \in V$, and in particular, $|\boldsymbol{u}| = \langle \boldsymbol{u}, \boldsymbol{u} \rangle^{1/2} = 0$, so that $\boldsymbol{u} = \boldsymbol{0}$. □

The expression $\boldsymbol{u}^\flat$ reads "$\boldsymbol{u}$ flat". $\flat$ and its inverse $\sharp$ (the existence of which is guaranteed by Theorem 1.4.4) are called the *musical isomorphisms* associated to the inner product. An important consequence of Theorem 1.4.4 is the following:

Corollary 1.4.2. *If V is a finite-dimensional inner product space, then for any $\alpha \in V^*$, there exists a unique $\boldsymbol{u} \in V$ such that*

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \alpha(\boldsymbol{v}), \qquad \boldsymbol{v} \in V.$$

Proof. $\boldsymbol{u} = \alpha^\sharp$. □

Examples 1.4.2. (i) If $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is a basis of V, then there exists, for each $j = 1, \ldots, n$, a unique element $\alpha^j \in V^*$ satisfying $\alpha^j(\sum_{i=1}^n a_i \boldsymbol{v}_i) = a_j$. In fact, the matrix of α^j in the basis $\mathcal{B}$ is $\boldsymbol{e}_j^T$, which also shows that the α^j are linearly independent. By dimension considerations, they form a basis of V^*, called the *basis dual* to $\mathcal{B}$. (It is also easy to see directly that they span V^*, since any $\alpha \in V^*$ may be written as $\alpha = \sum_{i=1}^n \alpha(\boldsymbol{v}_i)\alpha^i$.) For example, if $V = \mathbb{R}^n$, then the basis of V^* dual to the standard basis is $u^1, \ldots, u^n$, because $u^j(\sum_i a_i \boldsymbol{e}_i) = a_j$. Superscripts are traditionally used with dual elements.

(ii) A simple yet illustrative example of a musical isomorphism is the one corresponding to the standard inner product of $\mathbb{R}^n$. If α is a one-form, and $[\alpha]$ denotes its $n \times 1$ matrix in the standard basis, then, recalling that any vector is its own coordinate vector in that basis, we have

$$\alpha(\boldsymbol{v}) = [\alpha]\boldsymbol{v} = \langle [\alpha]^T, \boldsymbol{v} \rangle, \qquad \boldsymbol{v} \in \mathbb{R}^n.$$

Thus, $\alpha^\sharp$ is the transpose of the matrix of α in the standard basis. For example, if $\alpha : \mathbb{R}^3 \to \mathbb{R}$ is given by $\alpha(x, y, z) = 2x - 3y + z$, then

$$\alpha^\sharp = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}.$$

It should be noted that the musical isomorphism $\flat$ maps the standard basis of $\mathbb{R}^n$ to its dual basis in the sense of (i). This is because the standard basis is orthonormal, and the above claim holds in any inner product space: If $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ is an orthonormal basis of an inner product space V, then

$$\boldsymbol{v}_j^\flat(\sum_{i=1}^n a_i \boldsymbol{v}_i) = \langle \boldsymbol{v}_j, \sum_i a_i \boldsymbol{v}_i \rangle = a_j.$$

1.5 Subspaces of Euclidean space

We wish to take a closer look at subspaces of dimension 1 or $n - 1$ of $\mathbb{R}^n$, since these represent all proper subspaces in the important special case when $n = 3$.

If V is a subspace of dimension $n - 1$, then its orthogonal complement $V^\perp$ is one-dimensional, and an element of $V^\perp$ is called a *normal vector* of V. If $\boldsymbol{n} \neq \boldsymbol{0}$ is one such, then any other is a multiple of it, and V consists of the set of all $\boldsymbol{v} \in \mathbb{R}^n$ such

that $\langle \boldsymbol{v}, \boldsymbol{n} \rangle = 0$. Equivalently, V is the kernel of left multiplication $L_{\boldsymbol{n}^T}$ by $\boldsymbol{n}^T$. If $\boldsymbol{n}^T = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix}$, then V consist of all $\boldsymbol{v} = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}^T$ such that $a_1 x_1 + \dots + a_n x_n = 0$. V is called a *hyperplane* (or plane when $n = 3$) through the origin.

An *affine hyperplane* of $\mathbb{R}^n$ is a set of the form $\boldsymbol{u} + V$, where V is a hyperplane and $\boldsymbol{u} \in \mathbb{R}^n$. In other words, it it the subspace V "parallel translated" by $\boldsymbol{u}$. If $\boldsymbol{u} = \begin{bmatrix} u_1 & \dots & u_n \end{bmatrix}^T$ and V has equation $a_1 x_1 + \dots + a_n x_n = 0$ as above, then the equation of the affine hyperplane is

$$\{(x_1, \dots, x_n) \mid a_1 x_1 + \dots + a_n x_n = b\}, \qquad \text{where } b = a_1 u_1 + \dots + a_n u_n.$$

Indeed, the affine hyperplane consists of all points $\boldsymbol{u} + \boldsymbol{v}$, where $\langle \boldsymbol{v}, \boldsymbol{n} \rangle = 0$. Letting $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{v}$, this is equivalent to the set of all $\boldsymbol{x}$ such that $\langle \boldsymbol{x} - \boldsymbol{u}, \boldsymbol{n} \rangle = 0$; setting $x_i = u^i(\boldsymbol{x})$, this correspondents to all $\begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}^T$ such that $\sum a_i(x_i - u_i) = 0$, as claimed.

Even though we are primarily interested in the cases when $\dim V = 1$ or $n - 1$, a similar approach can be adopted to describe a subspace V of dimension $n - k$ for any k: If $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_k\}$ is a basis of $V^\perp$, then $\boldsymbol{a} \in \mathbb{R}^n$ belongs to V if and only if $\langle \boldsymbol{a}, \boldsymbol{u}_i \rangle = 0$ for $i = 1, \dots, k$. Each of these equations represents a hyperplane, and V is then the intersection of these k hyperplanes. As above, we obtain an affine subspace by translating V away from the origin.

A one-dimensional subspace V of $\mathbb{R}^n$ is called a *line* through the origin. We could describe it as an intersection of $n - 1$ hyperplanes, but when n is large, it is usually more convenient to resort to a vector $\boldsymbol{a}$ that spans V. $\boldsymbol{v} \in \mathbb{R}^n$ belongs to V if and only if $\boldsymbol{v} = t\boldsymbol{a}$ for some $t \in \mathbb{R}$. Letting $x_i = u^i(\boldsymbol{v})$, $a_i = u^i(\boldsymbol{a})$, we obtain so-called *parametric equations* of the line

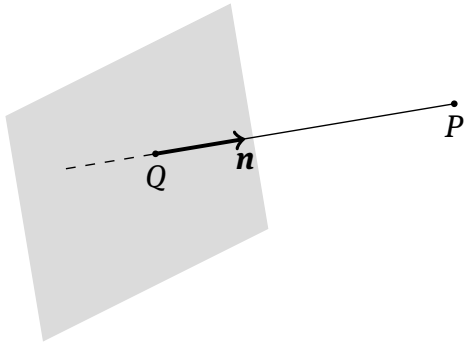
$$\{(x_1, \dots, x_n) \mid x_1 = ta_1, \dots, x_n = ta_n,\ t \in \mathbb{R}\}. \tag{1.5.1}$$

When all $a_i \neq 0$, one can eliminate the parameter t in (1.5.1) and describe the line by the equation $x_1/a_1 = \dots = x_n/a_n$.

As in the case of hyperplanes, one can parallel translate one-dimensional subspaces to describe all possible lines. Thus, the line parallel to the one in (1.5.1) that passes through $\boldsymbol{p} = (p_1, \dots, p_n)$ has parametric equations

$$x_1 = p_1 + ta_1, \dots, x_n = p_n + ta_n, \qquad t \in \mathbb{R},$$

which can be written $(x_1 - p_1)/a_1 = \dots = (x_n - p_n)/a_n$ if every $a_i \neq 0$.



Example 1.5.1. Suppose we are asked to find the distance from the point $P = (1, 2, -1)$ to the plane $x + y - z = 1$ in $\mathbb{R}^3$. This distance equals the length $|PQ|$ of the line segment PQ, where Q is the point of intersection of the plane with the line passing through P perpendicular to the plane. Since a vector normal to the plane is $\boldsymbol{n} = \begin{bmatrix} 1 & 1 & -1 \end{bmatrix}^T$, the line in question has vector equation

$$\boldsymbol{v} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} + t \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}.$$

Thus, a point on the line has coordinates $(1 + t, 2 + t, -1 - t)$, and the intersection with the plane occurs for the value of t for which

$$x + y - z = 1 + t + 2 + t + 1 + t = 1,$$

or $t = -1$. This means that $Q = (0, 1, 0)$. Elementary geometry implies that the distance between vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ in $\mathbb{R}^3$ is $|\boldsymbol{q} - \boldsymbol{p}|$, so that

$$|PQ| = \left((0-1)^2 + (1-2)^2 + (0-(-1))^2\right)^{1/2} = \sqrt{3}.$$

We will later find another solution to this problem, one that uses Calculus.

1.6 Determinants as volume

Any two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathbb{R}^3$ determine a linear map

$$\begin{aligned} \mathbb{R}^3 &\to \mathbb{R}, \\ \boldsymbol{u} &\mapsto \det \begin{bmatrix} \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{u} \end{bmatrix}, \end{aligned}$$

and thus an element of the dual space of $\mathbb{R}^3$. By Corollary 1.4.2, there is a unique vector $\boldsymbol{a} \times \boldsymbol{b} \in \mathbb{R}^3$, called the *cross product* of $\boldsymbol{a}$ and $\boldsymbol{b}$, such that

$$\langle \boldsymbol{a} \times \boldsymbol{b}, \boldsymbol{u} \rangle = \det \begin{bmatrix} \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{u} \end{bmatrix}, \qquad \boldsymbol{u} \in \mathbb{R}^3.$$

The following theorem is an immediate consequence of properties of the determinant:

Theorem 1.6.1. *Given $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^3$, $t \in \mathbb{R}$,*

- $\boldsymbol{b} \times \boldsymbol{a} = -\boldsymbol{a} \times \boldsymbol{b}$;
- $\boldsymbol{a} \times \boldsymbol{b}$ *is orthogonal to* $\boldsymbol{a}$ *and* $\boldsymbol{b}$;
- $(t\boldsymbol{a}) \times \boldsymbol{b} = \boldsymbol{a} \times (t\boldsymbol{b}) = t(\boldsymbol{a} \times \boldsymbol{b})$;
- $\boldsymbol{a} \times (\boldsymbol{b} + \boldsymbol{c}) = \boldsymbol{a} \times \boldsymbol{b} + \boldsymbol{a} \times \boldsymbol{c}$.

It is straightforward to compute the components of the vector $\boldsymbol{a} \times \boldsymbol{b}$: for example, if $a_i = u^i(\boldsymbol{a})$ and $b_i = u^i(\boldsymbol{b})$, then the first component equals

$$\begin{aligned}\langle \boldsymbol{a} \times \boldsymbol{b}, \boldsymbol{e}_1 \rangle &= \det \begin{bmatrix} \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{e}_1 \end{bmatrix} = \det \begin{bmatrix} \boldsymbol{a}^T \\ \boldsymbol{b}^T \\ \boldsymbol{e}_1^T \end{bmatrix} = \det \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ 1 & 0 & 0 \end{bmatrix} \\ &= \det \begin{bmatrix} a_2 & a_3 \\ b_2 & b_3 \end{bmatrix},\end{aligned}$$

and similar formulas hold for the other components, replacing $\boldsymbol{e}_1$ by $\boldsymbol{e}_2$ and $\boldsymbol{e}_3$. A common way of writing these formulas is

$$\boldsymbol{a} \times \boldsymbol{b} = \det \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \boldsymbol{e}_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{bmatrix},$$

where the meaningless right side is meant to be "expanded" along the first row; i.e.,

$$\boldsymbol{a} \times \boldsymbol{b} = \det \begin{bmatrix} a_2 & a_3 \\ b_2 & b_3 \end{bmatrix} \boldsymbol{e}_1 - \det \begin{bmatrix} a_1 & a_3 \\ b_1 & b_3 \end{bmatrix} \boldsymbol{e}_2 + \det \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix} \boldsymbol{e}_3. \tag{1.6.1}$$

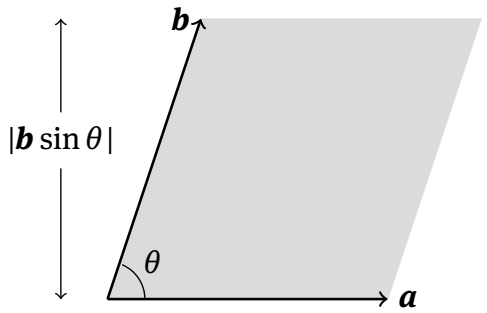
Using the latter formula, it is straightforward to compute that the norm squared of a cross product equals

$$\begin{aligned}|\boldsymbol{a} \times \boldsymbol{b}|^2 &= (a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2) - (a_1 b_1 + a_2 b_2 + a_3 b_3)^2 \\ &= |\boldsymbol{a}|^2 |\boldsymbol{b}|^2 - \langle \boldsymbol{a}, \boldsymbol{b} \rangle^2 = |\boldsymbol{a}|^2 |\boldsymbol{b}|^2 - |\boldsymbol{a}|^2 |\boldsymbol{b}|^2 \cos^2 \theta \\ &= |\boldsymbol{a}|^2 |\boldsymbol{b}|^2 \sin^2 \theta,\end{aligned}$$

with θ denoting the angle between $\boldsymbol{a}$ and $\boldsymbol{b}$. Thus,

$$|\boldsymbol{a} \times \boldsymbol{b}| = |\boldsymbol{a}||\boldsymbol{b}|| \sin \theta |. \tag{1.6.2}$$

This means that $|\boldsymbol{a} \times \boldsymbol{b}|$ represents the area of the parallelogram spanned by $\boldsymbol{a}$ and $\boldsymbol{b}$: notice that the area equals that of a rectangle with same base length $|\boldsymbol{a}|$ and height $|\boldsymbol{b}|| \sin \theta |$.



If, on the other hand, we take the two vectors to lie in the *xy*-plane, so that their third component is zero, then by (1.6.1),

$$\boldsymbol{a} \times \boldsymbol{b} = \det \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix} \boldsymbol{e}_3.$$

Combining this with the previous remark, we conclude: *The absolute value of a* 2×2 *determinant equals the area of the parallelogram spanned by the columns (or the rows).*

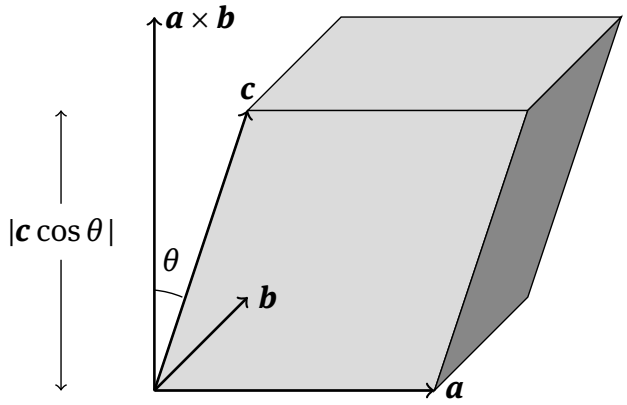
A similar interpretation exists for 3×3 determinants. If $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c} \in \mathbb{R}^3$, then the volume of the parallelepiped

$$\{a\boldsymbol{a} + b\boldsymbol{b} + c\boldsymbol{c} \mid 0 \le a, b, c \le 1\}$$

spanned by the three vectors equals the area of the base times the height. The former is $|\boldsymbol{a} \times \boldsymbol{b}|$, and the latter is $|\boldsymbol{c}||\cos\theta|$, where θ is the angle between $\boldsymbol{a} \times \boldsymbol{b}$ and $\boldsymbol{c}$. Thus, the volume equals

$$|\langle \boldsymbol{a} \times \boldsymbol{b}, \boldsymbol{c} \rangle| = |\det \begin{bmatrix} \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{c} \end{bmatrix}|,$$

so that *the absolute value of a* 3×3 *determinant equals the volume of the parallelepiped spanned by the three columns (or rows).* Incidentally, we define this to be also the volume of the "open" (see the following section) parallelepiped $\{a\boldsymbol{a} + b\boldsymbol{b} + c\boldsymbol{c} \mid 0 < a, b, c < 1\}$. This is because the open solid may be expressed as an increasing union of "closed" ones together with continuity of volume: for example, $(0, 1) = \cup_{i=1}^{\infty} [1/n, 1 - 1/n]$, so that the volume or length of $(0, 1)$ equals $\lim_{n\to\infty} 1 - 1/n - 1/n = 1$.



Later, we will discuss volume in n-dimensional space. For now, the previous observations provide some justification for the following:

Definition 1.6.1. The *volume* of the (open or closed) parallelepiped in $\mathbb{R}^n$ spanned by the vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{R}^n$ is defined to be the absolute value of

$$\det \begin{bmatrix} \boldsymbol{a}_1 \cdots \boldsymbol{a}_n \end{bmatrix}.$$

1.7 Elementary topology of Euclidean spaces

A fundamental application of the inner product introduced in Section 1.3 is the concept of distance: The *distance* between $\boldsymbol{a}$ and $\boldsymbol{b} \in \mathbb{R}^n$ is $\mathrm{d}(\boldsymbol{a}, \boldsymbol{b}) = |\boldsymbol{a} - \boldsymbol{b}|$. The distance function d satisfies the following properties for any $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^n$:

- $\mathrm{d}(\boldsymbol{a}, \boldsymbol{b}) = \mathrm{d}(\boldsymbol{b}, \boldsymbol{a})$;
- $\mathrm{d}(\boldsymbol{a}, \boldsymbol{b}) \geq 0$, and $\mathrm{d}(\boldsymbol{a}, \boldsymbol{b}) = 0$ if and only if $\boldsymbol{a} = \boldsymbol{b}$;
- $\mathrm{d}(\boldsymbol{a}, \boldsymbol{b}) \leq \mathrm{d}(\boldsymbol{a}, \boldsymbol{c}) + \mathrm{d}(\boldsymbol{c}, \boldsymbol{b})$.

All these properties follow from corresponding properties of the norm; for example, the third one is a consequence of the triangle inequality (and is actually responsible for that name, since it says that in a triangle, the length of one side cannot exceed the sum of the other two):

$$\mathrm{d}(\boldsymbol{a}, \boldsymbol{b}) = |\boldsymbol{a} - \boldsymbol{b}| = |(\boldsymbol{a} - \boldsymbol{c}) + (\boldsymbol{c} - \boldsymbol{b})| \leq |(\boldsymbol{a} - \boldsymbol{c})| + |(\boldsymbol{c} - \boldsymbol{b})| = \mathrm{d}(\boldsymbol{a}, \boldsymbol{c}) + \mathrm{d}(\boldsymbol{c}, \boldsymbol{b}).$$

More generally, a set X together with a function $\mathrm{d} : X \times X \to \mathbb{R}$ that satisfies the above three properties is called a *metric space*. Thus, every inner product space is a metric space. Notice that the inner product itself is not essential, but rather the norm associated to it.

Definition 1.7.1. A *norm* on a vector space E is a map $|| : E \to \mathbb{R}$ such that for all $\boldsymbol{a}, \boldsymbol{b} \in E$ and $\alpha \in \mathbb{R}$,

(1) $|\boldsymbol{a}| \geq 0$, and $|\boldsymbol{a}| = 0$ if and only if $\boldsymbol{a} = \boldsymbol{0}$;
(2) $|\boldsymbol{a} + \boldsymbol{b}| \leq |\boldsymbol{a} + \boldsymbol{b}|$;
(3) $|\alpha \boldsymbol{a}| = |\alpha||\boldsymbol{a}|$.

A *normed vector space* is a vector space together with a norm. It follows that any normed space is a metric space if one defines the distance by $\mathrm{d}(\boldsymbol{a}, \boldsymbol{b}) = |\boldsymbol{a} - \boldsymbol{b}|$. One important example is the space of all linear transformations $L : \mathbb{R}^n \to \mathbb{R}^m$ with the operator norm from Definition 1.4.1.

Definition 1.7.2. Let (X, d) be a metric space.

(1) The *open ball of radius* $r > 0$ *around* $\boldsymbol{a}$ is the set $B_r(\boldsymbol{a})$ of all points at distance less than r from $\boldsymbol{a}$;
(2) A set $U \subset X$ is said to be a *neighborhood* of $\boldsymbol{a} \in X$ if U contains some open ball around $\boldsymbol{a}$; in this case, we say $\boldsymbol{a}$ is an *interior point* of U;
(3) A set $V \subset X$ is said to be *open* if it is a neighborhood of each and every one of its elements.
(4) A set $C \subset X$ is said to be *closed* if its complement $X \setminus C$ is open.

It follows from the definition that both X and the empty set are open. Being complements of each other, they are also closed. Although we will mostly deal with the case $X = \mathbb{R}^n$, it is useful, whenever possible, to state properties in terms of abstract

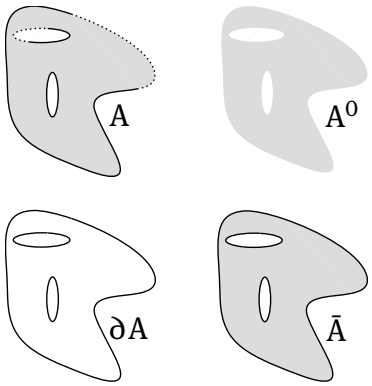
metric spaces. Observe for example that any subset of $\mathbb{R}^n$ becomes a metric space when endowed with the restriction of the metric on $\mathbb{R}^n$.

The triangle inequality guarantees that open balls are indeed open: for if $\boldsymbol{b} \in B_r(\boldsymbol{a})$, then $\varepsilon := r - \mathrm{d}(\boldsymbol{a}, \boldsymbol{b}) > 0$, and the ball of radius ε around y is contained inside $B_r(\boldsymbol{a})$. In fact, given $\boldsymbol{c} \in B_\varepsilon(\boldsymbol{b})$, $\mathrm{d}(\boldsymbol{a}, \boldsymbol{c}) \leq \mathrm{d}(\boldsymbol{a}, \boldsymbol{b}) + \mathrm{d}(\boldsymbol{b}, \boldsymbol{c}) < \mathrm{d}(\boldsymbol{a}, \boldsymbol{b}) + \varepsilon = r$.

Notice that a union of open sets is again open. Since the complement of a union is the intersection of the complements, an arbitrary intersection of closed sets is closed. The intersection of a finite number of open sets is open (and therefore a finite union of closed sets is closed): if $\boldsymbol{a} \in U_i$ for $i = 1, \ldots, k$ and each U_i is open, then there exists $r_i > 0$ such that the ball of radius r_i around $\boldsymbol{a}$ is contained in U_i. Consequently the ball of radius r around $\boldsymbol{a}$, where $r = \min\{r_1, \ldots, r_k\}$, is contained in every U_i, and therefore in their intersection. On the other hand, arbitrary intersections of open sets need not be open: for example, the intersection of all open balls of radius $1/k$, $k \in \mathbb{N}$, about a point $\boldsymbol{p} \in \mathbb{R}^n$, consists of the single point $\boldsymbol{p}$, and one-point sets are closed.

Given real numbers $a < b$, the interval (a, b) is open in $\mathbb{R}$. Viewing the real line as a subset of $\mathbb{R}^2$ (i.e., identifying $\mathbb{R}$ with $\mathbb{R} \times \{0\} \subset \mathbb{R}^2$) $(a, b) \times \{0\}$ is no longer open. In order to be able to say that it is open as a subset of $\mathbb{R} \times \{0\}$, we introduce the following:

Definition 1.7.3. Let A be a subset of a metric space X. A subset B of A is said to be *open* (resp. *closed) in* A or *relative to* A if $B = U \cap A$, where U is open (resp. closed) in X.



Definition 1.7.4. Let $A \subset X$.

(1) The *interior* A^0, also denoted int A, of A is the set of all interior points of A.
(2) A point is said to be a *boundary point of* A if every neighborhood of that point intersects both A and the complement of A. The *boundary* ∂A of A is the collection of all boundary points of A.
(3) The *closure* of A is the set $\overline{A} = A \cup \partial A$.
(4) A is said to be *bounded* if it is contained in some metric ball (of finite radius).

It is clear that A^0 is open, ∂A and $\overline{A}$ are closed, and $A^0 \subset A \subset \overline{A}$. In fact, A^0 is the largest open set that is contained in A, and $\overline{A}$ is the smallest closed set that contains A, see Exercise 1.20.

Definition 1.7.5. An *open cover* $\mathcal{O}$ of $A \subset X$ is a collection of open sets whose union contains A. A subcollection of $\mathcal{O}$ is called a *subcover* if it is also a cover of A. A is said to be *compact* if any open cover of A contains a finite subcover.

Any finite set is of course compact, but finiteness is not a requisite: If $\{a_n\}$ is any sequence of real numbers that converges to some a, then the set $K = \{a\} \cup \{a_n \mid n \in \mathbb{N}\}$ is compact, since any open interval containing a will contain all but finitely many elements of K.

Theorem 1.7.1. (1) *Any compact set is closed;*
(2) *Any closed subset of a compact set is compact.*

Proof. For (1), suppose K is compact. We will show that the complement of K is open. Given $\boldsymbol{a} \notin K$, choose for every $\boldsymbol{b} \in K$ open balls $U_{\boldsymbol{b}}$ and $V_{\boldsymbol{b}}$ centered at $\boldsymbol{b}$ and $\boldsymbol{a}$ respectively of small enough radius that they don't intersect. By compactness of K, finitely many of these, say, $U_{\boldsymbol{b}_1}, \ldots, U_{\boldsymbol{b}_k}$ cover K. Then $V_{\boldsymbol{b}_1} \cap \cdots \cap V_{\boldsymbol{b}_k}$ is an open ball around $\boldsymbol{a}$ that does not intersect any $U_{\boldsymbol{b}_j}$, $j = 1, \ldots, k$, and therefore does not intersect K.
For (2), suppose $C \subset K$, where C is closed and K is compact, and consider an open cover $\mathcal{O}$ of C. Then $\mathcal{O}$ together with the complement of C is an open cover of K and we may extract a finite subcover. If $X \setminus C$ is a member of this subcover, remove it. What remains is a finite subcollection of $\mathcal{O}$ that still covers C. □

Compact sets are important enough to warrant a more concrete characterization. This alternative description does not, however, hold in arbitrary metric spaces. Recall that $\alpha \in \mathbb{R}$ is said to be an *upper bound* of a set A if $\alpha \geq x$ for any $x \in A$, and the *least upper bound* of A if it no larger than any other upper bound; i.e., if $\alpha \leq \beta$ for any upper bound β of A. In this case, we write $\alpha = \sup A$. Notice that even though α itself need not belong to A, any neighborhood of α must contain an element of A: for if, say, $(\alpha - \varepsilon, \alpha + \varepsilon)$ did not intersect A for some $\varepsilon > 0$, then $\alpha - \varepsilon$ would be an upper bound of A smaller than α. A key property of the real numbers is that any nonempty set of reals that is bounded above has a least upper bound. Appendix A explores these concepts in further detail.

A *box* or *rectangle* in $\mathbb{R}^n$ is a cartesian product of n intervals. In the definition of an open set, we could have replaced open metric balls by open boxes: the ball $B_r(\boldsymbol{p})$ contains the open box

$$\left(p_1 - \frac{r}{\sqrt{n}}, p_1 + \frac{r}{\sqrt{n}}\right) \times \cdots \times \left(p_n - \frac{r}{\sqrt{n}}, p_n + \frac{r}{\sqrt{n}}\right)$$

(with $p_i = u^i(\boldsymbol{p})$) centered at $\boldsymbol{p}$, and this box in turn contains the ball $B_{r/\sqrt{n}}(\boldsymbol{p})$. In particular, if U is any open set containing a point $\boldsymbol{p}$, then there exists an open box R centered at $\boldsymbol{p}$ which lies inside U. Conversely, if U is a set with the property that any point of U admits an open box centered at that point which is contained in U, then any $\boldsymbol{p}$ in U also admits an open ball centered at $\boldsymbol{p}$ which is contained in U, so that U is

open. Our first goal is to show that a closed and bounded box is compact. We begin in dimension one:

Theorem 1.7.2 (Heine-Borel). *For any $a < b$, the interval $[a, b]$ is compact.*

Proof. Given an open cover $\mathcal{O}$ of $[a, b]$, denote by A the set of all $x \in [a, b]$ such that $[a, x]$ is covered by finitely many sets in $\mathcal{O}$. Then $a \in A$, A is bounded above by b, and so A has a least upper bound $\alpha \in [a, b]$. We first observe that $\alpha \in A$: indeed, α belongs to some element U of $\mathcal{O}$. Since α is the least upper bound of A, U must contain some element $x \in A$. By assumption, $[a, x]$ is then covered by finitely many sets in $\mathcal{O}$, and $[x, \alpha]$ by one, namely U. Thus $[a, \alpha] = [a, x] \cup [x, \alpha]$ is also covered by finitely many sets, and $\alpha \in A$. We now conclude the proof by showing that $\alpha = b$. If $\alpha < b$, then the set U above must contain some $c \in (\alpha, b)$, and the above argument shows that $c \in A$, contradicting the fact that α is an upper bound. □

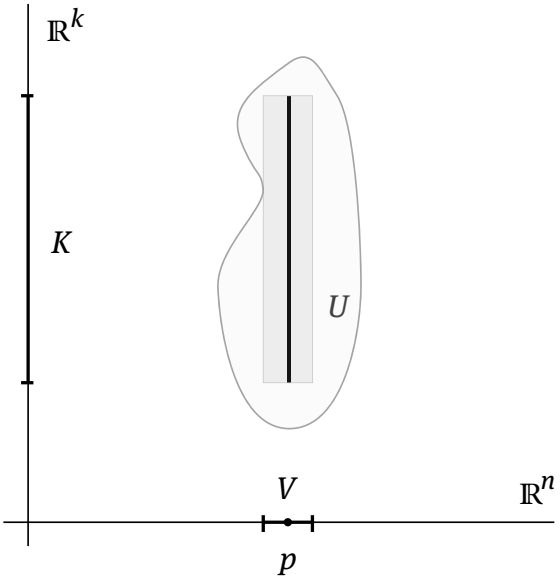


Fig. 1.1: A tube about $\{\boldsymbol{p}\} \times K$ contained in U

In order to show that a Cartesian product of compact sets is compact, we will need the following:

Lemma 1.7.1 (the tube lemma). *Suppose $\boldsymbol{p} \in \mathbb{R}^n$, and K is a compact subset of $\mathbb{R}^k$. If U is an open set in $\mathbb{R}^{n+k}$ that contains $\{\boldsymbol{p}\} \times K$, then U contains $V \times K$ for some open neighborhood V of $\boldsymbol{p}$ in $\mathbb{R}^n$.*

Proof. For any $\boldsymbol{a} \in K$, there exists an open box inside U that contains the point $(\boldsymbol{p}, \boldsymbol{a})$. Such a box is a product $V_{\boldsymbol{a}} \times W_{\boldsymbol{a}}$ of open boxes in $\mathbb{R}^n$ and $\mathbb{R}^k$ containing $\boldsymbol{p}$ and $\boldsymbol{a}$ respectively. Now, K is compact, and is therefore covered by finitely many $W_{\boldsymbol{a}_1}, \ldots, W_{\boldsymbol{a}_l}$.

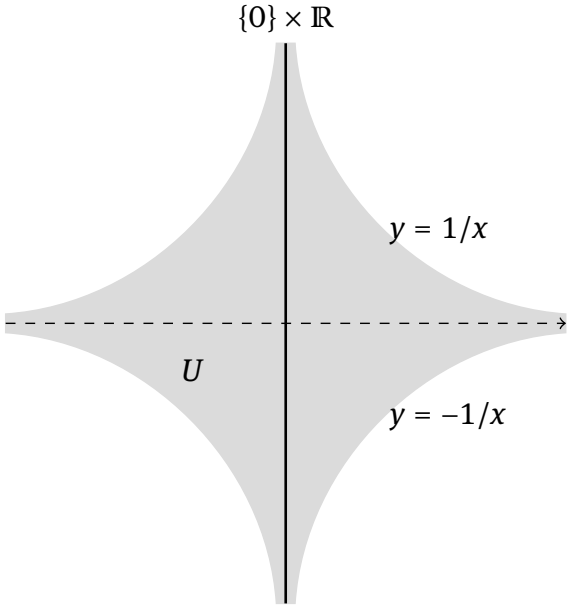
If $V = V_{\boldsymbol{a}_1} \cap \cdots \cap V_{\boldsymbol{a}_l}$, then V is an open neighborhood of p, and

$$V \times K \subset V \times \left(\bigcup_{i=1}^{l} W_{\boldsymbol{a}_i} \right) = \bigcup_{i=1}^{l} (V \times W_{\boldsymbol{a}_i}) \subset \bigcup_{i=1}^{l} (V_{\boldsymbol{a}_i} \times W_{\boldsymbol{a}_i}) \subset U,$$

as claimed. $V \times K$ is called a *tube* about $\{\boldsymbol{p}\} \times K$. □

Notice that the set $\{\boldsymbol{p}\} \times K$ in the above lemma is compact if K is: indeed, if $\pi : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^k$ denotes the projection onto the second factor, then π is an *open map*; i.e., π maps open sets to open sets – in fact, π maps balls of a given radius onto balls of the same radius. Thus, if $\mathcal{O}$ is an open cover of $\{\boldsymbol{p}\} \times K$, then the sets $\pi(U)$, $U \in \mathcal{O}$, form an open cover of K, and a finite subcover $\pi(U_1), \ldots, \pi(U_l)$ of K may be chosen. This means that $U_1 \cup \cdots \cup U_l \supset \{\boldsymbol{p}\} \times K$.

It is easily seen that compactness of K is essential in the tube lemma: if, for example $n = k = 1$, $\boldsymbol{p} = 0$, and $K = \mathbb{R}$, then $U = \{(x, y) \mid |xy| < 1\}$ is an open subset of $\mathbb{R}^2$ containing $\{0\} \times \mathbb{R}$, but it cannot contain any tube around it.



Theorem 1.7.3. *A Cartesian product of compact sets is compact. In particular, if $a_i < b_i$, $i = 1, \ldots, n$, then $[a_1, b_1] \times \cdots \times [a_n, b_n]$ is compact.*

Proof. The second statement follows from the first one together with Theorem 1.7.2. For the first one, it is enough to show that $A \times B$ is compact whenever A and B are. Let $\mathcal{O}$ denote an open cover of $A \times B$. For each $\boldsymbol{p} \in A$, the set $\{\boldsymbol{p}\} \times B$, being compact, can be covered by a finite subcollection $\mathcal{O}_{\boldsymbol{p}}$ of $\mathcal{O}$. By the tube lemma, the union of the sets in $\mathcal{O}_{\boldsymbol{p}}$ contains $U_{\boldsymbol{p}} \times B$ for some neighborhood $U_{\boldsymbol{p}}$ of $\boldsymbol{p}$, and A can be covered by finitely many of these, say, $U_{\boldsymbol{p}_1}, \ldots, U_{\boldsymbol{p}_l}$. But then $\mathcal{O}_{\boldsymbol{p}_1} \cup \cdots \cup \mathcal{O}_{\boldsymbol{p}_l}$ is a finite cover of $A \times B$. □

Theorem 1.7.4. *A subset K of $\mathbb{R}^n$ is compact if and only if it is closed and bounded.*

Proof. If K is compact, then it is closed by Theorem 1.7.1. It is also bounded, for if $\boldsymbol{p} \in K$, then the collection of balls $B_k(\boldsymbol{p})$, $k \in \mathbb{N}$, is an open cover of K, and admits a finite subcover. The ball of largest radius in that subcover contains K. Conversely, suppose K is closed and bounded. Being bounded, it is contained inside some closed rectangle. The latter is compact by Theorem 1.7.3. Since K is closed, it is compact by Theorem 1.7.1. □

By definition, if E is not compact, then there exists some open cover of E with no finite subcover. However, if E lies in Euclidean space, there will always be a countable subcover:

Theorem 1.7.5. *If $E \subset \mathbb{R}^n$, then any open cover $\{U_\alpha\}_{\alpha \in J}$ of E has a countable subcover; i.e., there exists a subset A of natural numbers, and a map $f : A \to J$ such that*

$$E \subset \bigcup_{k \in A} U_{f(k)}.$$

Proof. The argument uses the notion of countable set and dense set, both of which are examined in Appendix A. For each $\boldsymbol{a} \in E$, choose some U_α that contains it. Since this set is open, there exists some $r > 0$ such that the ball $B_r(\boldsymbol{a}) \subset U_\alpha$. The fact that rationals are dense in $\mathbb{R}$ (see Appendix A) is easily extended to $\mathbb{Q}^n$ being dense in $\mathbb{R}^n$. Thus, there exists a ball $B_{\boldsymbol{a}}$ with center $\boldsymbol{a} \in \mathbb{Q}^n$ and rational radius such that $\boldsymbol{a} \in B_{\boldsymbol{a}} \subset U_\alpha$. But $\mathbb{Q}^n \times \mathbb{Q}$ is countable, and therefore so is the collection

$$\{B_{\boldsymbol{a}} \mid \boldsymbol{a} \in E\} = \{V_1, V_2, \dots\}.$$

By assumption, each V_k is contained in some U_α. Choosing some such α and setting $\alpha = f(k)$, we conclude that $E \subset \bigcup_k V_k \subset \bigcup_k U_{f(k)}$. □

The least upper bound of a set A of real numbers is also called the *supremum* of A, and is denoted $\sup A$. A similar notion can be introduced for lower bounds: the *infimum* or *greatest lower bound* $\alpha = \inf A$ of A is a lower bound of A (i.e., $\alpha \leq x$ for any $x \in A$) that is greater than or equal to any other lower bound of A. Notice that α is a lower bound of A if and only if $-\alpha$ is an upper bound of $-A = \{-a \mid a \in A\}$. It easily follows that $\inf A = -\sup(-A)$, in the sense that if one of the two exists, then so does the other, and they are equal. In particular, any nonempty set of real numbers that is bounded below has a greatest lower bound.

A useful observation is that if A is closed, nonempty, and bounded above (respectively below), then A contains its least upper bound (resp. greatest lower bound): indeed, by definition, given any $\varepsilon > 0$, there must be a point $x \in A$ such that

$$\sup A - \varepsilon < x \leq \sup A,$$

since otherwise $\sup A - \varepsilon$ would be an upper bound of A. Thus, any open interval of radius ε around $\sup A$ contains points of A (as was just pointed out) and also points

outside A (any $y \in (\sup A, \sup A + \varepsilon)$), so that $\sup A$ is a boundary point of A. But closed sets contain their boundary points. The argument for infimum is similar.

The concepts of infimum and supremum also enable us to prove a property of intervals that will be quite useful in the sequel:

Proposition 1.7.1. *If I is an interval of real numbers, then any subset that is both open and closed in I is either empty or equals all of I.*

Proof. We argue by contradiction: suppose A is a nonempty subset of I that is both open and closed, but there exists some $c \in I$ that does not belong to A. Consider any $t_0 \in A$. Then $t_0 < c$ or $t_0 > c$. The argument is similar in both cases, so we only consider the former. Let

$$B = \{t \in I \mid [t_0, t] \subset A\}.$$

Then B is nonempty (since it contains t_0), bounded above (by c), and so admits a supremum α. By definition, α is a boundary point of A. Since A is closed in I, A equals the intersection of I with some closed set C, and $\alpha \in C$: indeed, if some neighborhood of α is contained in C, then certainly $\alpha \in C$. Otherwise, every neighborhood of α contains points outside C. It must also contain points in C since it contains points of $A \subset C$. This means α is a boundary point of C, and since C is closed, $\alpha \in C$ as claimed. But α also belongs to I, because the latter is an interval that contains a point to the left of α (e.g., t_0) and one to the right of it (e.g., c). Thus, $\alpha \in A = C \cap I$, and so $\alpha < c$. A is open in I, however, so there exists some $\varepsilon > 0$ such that $(\alpha, \alpha + \varepsilon) \cap I \subset A$. Thus, if β is any point in the nonempty set $(\alpha, \alpha + \varepsilon) \cap I$, then $[t_0, \beta] \subset A$, contradicting the assumption that α is an upper bound of B. □

1.8 Sequences

The reader is already familiar with sequences of real numbers from Calculus. The generalization to metric spaces is straightforward: A *sequence* in a metric space (X, d) is a map from the set $\mathbb{N}$ of natural numbers into X. The value of the map at $k \in \mathbb{N}$ is usually denoted $\boldsymbol{a}_k$ (or some other letter such as $\boldsymbol{p}_k$) when $X = \mathbb{R}^n$, $n > 1$, and the sequence itself by $\{\boldsymbol{a}_k\}$ – although we will often use $\boldsymbol{a}_k$ to denote either one when there is no risk of confusion. When $n = 1$ or X is not Euclidean space, we use the regular font a_k instead of the bold one.

Definition 1.8.1. A sequence $\{\boldsymbol{a}_k\}$ is said to *converge* to $\boldsymbol{a}$ if for any $\varepsilon > 0$, there is a positive integer N such that $\mathrm{d}(\boldsymbol{a}_k, \boldsymbol{a}) < \varepsilon$ whenever $k \geq N$. In this case, $\boldsymbol{a}$ is called the *limit* of the sequence, and we write $\boldsymbol{a}_k \to \boldsymbol{a}$ or $\lim_{k\to\infty} \boldsymbol{a}_k = \boldsymbol{a}$. If no such $\boldsymbol{a}$ exists, the sequence is said to *diverge*. The sequence is said to be *bounded* if the set $\{\boldsymbol{a}_k \mid k \in \mathbb{N}\}$ of values is bounded; i.e., if there exists some $R > 0$ such that $\mathrm{d}(\boldsymbol{a}_1, \boldsymbol{a}_k) \leq R$ for all k.

Another way to describe convergence $\boldsymbol{a}_k \to \boldsymbol{a}$ is to say that any neighborhood of $\boldsymbol{a}$ contains $\boldsymbol{a}_k$ except perhaps for finitely many values of k. The one in the above definition is of course just the open ball of radius ε about $\boldsymbol{a}$. This also justifies using the terminology "the limit", for a sequence can have at most one limit: if $\boldsymbol{a} \neq \boldsymbol{b}$, then these two points admit disjoint neighborhoods (such as the open metric balls of radius half the distance between them), and then both neighborhoods cannot contain $\boldsymbol{a}_k$ for all sufficiently large k. It also shows that any convergent sequence is bounded: there are at most finitely many points of the sequence outside the ball of radius, say, 1 about the limit, and a finite set is alway bounded: specifically, if $\boldsymbol{a}_k \in B_1(\boldsymbol{a})$ for all $k \geq N$, let $R = \max\{1, \mathrm{d}(\boldsymbol{a}_1, \boldsymbol{a}), \ldots, \mathrm{d}(\boldsymbol{a}_{N-1}, \boldsymbol{a})$. Then $\mathrm{d}(\boldsymbol{a}_k, \boldsymbol{a}) \leq R$ for all k. It is also clear from the above discussion that changing finitely many terms in a sequence does not affect convergence; i.e., if $\boldsymbol{a}_k$ and $\boldsymbol{b}_k$ are two sequences such that $\boldsymbol{a}_k = \boldsymbol{b}_k$ for all sufficiently large k, then they either both converge to the same limit, or they both diverge.

We first consider convergence of sequences of real numbers, since it is very much related to that of sequences in higher-dimensional Euclidean spaces. Even though the reader is probably familiar with the contents of the following theorem from a previous Calculus course, it has been included because its proof is often not covered there.

Theorem 1.8.1. *Let $\{x_k\}$ and $\{y_k\}$ be sequences of real numbers that converge to x and y respectively. Then*

(1) $x_k + y_k \to x + y$;
(2) $cx_k \to cx$ *for any* $c \in \mathbb{R}$;
(3) $x_k y_k \to xy$;
(4) *If* $x_k, x \neq 0$, *then* $1/x_k \to 1/x$;
(5) *If* $y_k, y \neq 0$, *then* $(x_k/y_k) \to (x/y)$.

Proof. (1) Let $\varepsilon > 0$ be given. By assumption, there exist positive integers N_1 and N_2 such that $|x_k - x| < \varepsilon/2$ for all $k \geq N_1$, and $|y_k - y| < \varepsilon/2$ for all $k \geq N_2$. So, if N is the largest of the two numbers N_1 and N_2, then for any $k \geq N$,

$$|(x_k + y_k) - (x + y)| = |(x_k - x) + (y_k - y)| \leq |x_k - x| + |y_k - y| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

(2) This is an immediate consequence of (3), taking $y_k = c$ for all k.
(3) Since $\{x_k\}$ converges, there exists some $M > 0$ such that $|x_k| \leq M$ for all k. Choose N_1 so that $|y_k - y| < \varepsilon/(2M)$ for $k \geq N_1$, and N_2 so that $|x_k - x| < \varepsilon/(2|y|)$ if $k \geq N_2$ (unless $y = 0$, in which case take $N_2 = 1$). If N is the larger of N_1 and N_2, then for $k \geq N$,

$$|x_k y_k - xy| = |x_k(y_k - y) + y(x_k - x)| \leq |x_k||y_k - y| + |y||x_k - x|.$$

The first term on the right side of the above inequality is smaller than $\varepsilon/2$. The second term is either zero (if $y = 0$) or less than $\varepsilon/2$ (if $y \neq 0$). In either case, the left side is less than ε, which establishes the claim.
(4) Take $\varepsilon = |x|/2$ in the definition of convergence of $\{x_k\}$ to conclude that there exists N_1 such that $|x_k| > |x|/2$ whenever $n \geq N_1$. Next, for a given $\varepsilon > 0$, there exists N_2 such

that $|x_k - x| < \varepsilon|x|^2/2$ whenever $k \geq N_2$. Let $N = \max\{N_1, N_2\}$. If $k \geq N$, then

$$\left|\frac{1}{x_k} - \frac{1}{x}\right| = \left|\frac{x_k - x}{x_k x}\right| < \frac{2|x_k - x|}{|x|^2} < \varepsilon.$$

(5) is an easy consequence of (3) and (4). □

Examples and Remarks 1.8.1. (i) A sequence $\{a_k\}$ of real numbers is said to be *increasing* (resp. *decreasing*) if $a_k \leq a_{k+1}$ (resp. $a_k \geq a_{k+1}$) for all k. An increasing sequence that is bounded above converges: to see this, let α denote the supremum of $\{a_k \mid k \in \mathbb{N}\}$. Given $\varepsilon > 0$, there exists some $N \in \mathbb{N}$ such that $0 \leq \alpha - a_N < \varepsilon$. Since the sequence is increasing, $0 \leq \alpha - a_k \leq \alpha - a_N < \varepsilon$ for all $k \geq N$, and $a_k \to \alpha$. This also implies that a decreasing sequence that is bounded below converges to the infimum of its set of values, because if $\{a_k\}$ is decreasing bounded below, then $\{-a_k\}$ is increasing and bounded above.

(ii) Another useful tool for proving convergence of a real sequence is the so-called squeeze theorem: if $a_k \leq b_k \leq c_k$, and both $\{a_k\}$ and $\{c_k\}$ converge to the same limit L, then so does the middle sequence $\{b_k\}$. To see this, let $\varepsilon > 0$, and choose N_1 such that $|a_k - L| < \varepsilon$ whenever $k \geq N_1$. Similarly, let N_2 be such that $|c_k - L| < \varepsilon$ for $k \geq N_2$. If k is larger than $N = \max\{N_1, N_2\}$, then $|b_k - L| < \varepsilon$, because

$$-\varepsilon < a_k - L \leq b_k - L \leq c_k - L < \varepsilon.$$

(iii) One of the simplest applications of (i) and (ii) is the so-called *geometric sequence* r^k, where $0 < r < 1$. It is a decreasing sequence, bounded below by zero, and therefore converges. This does not quite tell us what the limit is, but if $r^k \to L$, then $a_k \to L$, where $a_k = r^{k+1}$: indeed, given $\varepsilon > 0$, choose N such that $|r^k - L| < \varepsilon$ for $k > N$; then $|a_k - L| = |r^{k+1} - L|$ is also less than ε for $k > N$ because $k + 1 > k > N$. On the other hand, by Theorem 1.8.1 (2), $a_k \to rL$. Thus, $L = rL$, and since $r \neq 1$, $L = 0$. In conclusion, $r^k \to 0$.
We have, in fact, that $r^k \to 0$ for any $r \in (-1, 1)$; indeed, we know that $|r|^k \to 0$. By Theorem 1.8.1 (2), $-|r|^k \to 0$. Since $-|r|^k \leq r^k \leq |r|^k$, the claim follows from the squeeze theorem in (ii).

For the sake of convenience, we will often abbreviate the sentence "There exists $N \in \mathbb{N}$ such that the statement $P(k)$ is true whenever $k \geq N$" by "$P(k)$ is true for sufficiently large k". Now that we have some experience with real-valued sequences, let us look at sequences in higher-dimensional Euclidean space.

Theorem 1.8.2. *A sequence $\{\boldsymbol{a}_k\}$ in $\mathbb{R}^n$ converges to $\boldsymbol{a}$ iff $u^i(\boldsymbol{a}_k) \to u^i(\boldsymbol{a})$ in $\mathbb{R}$ for $i = 1, \ldots, n$.*

Proof. If $\boldsymbol{a}_k \to \boldsymbol{a}$, then $u^i(\boldsymbol{a}_k) \to u^i(\boldsymbol{a})$ since $|u^i(\boldsymbol{a}_k) - u^i(\boldsymbol{a})| \leq |\boldsymbol{a}_k - \boldsymbol{a}|$. Conversely, if $u^i(\boldsymbol{a}_k) \to u^i(\boldsymbol{a})$ for each $i = 1, \ldots, n$, then given $\varepsilon > 0$, $|u^i(\boldsymbol{a}_k) - u^i(\boldsymbol{a})| < \varepsilon/\sqrt{n}$ for all i

and sufficiently large k. But then

$$|\boldsymbol{a}_k - \boldsymbol{a}| = \left(\sum_{i=1}^{n} |u^i(\boldsymbol{a}_k) - u^i(\boldsymbol{a})|^2 \right)^{\frac{1}{2}} < \varepsilon$$

for k large enough, which shows that $\boldsymbol{a}_k \to \boldsymbol{a}$. □

Corollary 1.8.1. *If $\{\boldsymbol{a}_k\}$ and $\{\boldsymbol{b}_k\}$ are sequences in $\mathbb{R}^n$ that converge to $\boldsymbol{a}$ and $\boldsymbol{b}$ respectively, then*

(1) $\boldsymbol{a}_k + \boldsymbol{b}_k \to \boldsymbol{a} + \boldsymbol{b}$;
(2) *If $c_k \to c \in \mathbb{R}$, then $c_k \boldsymbol{a}_k \to c\boldsymbol{a}$;*
(3) $\langle \boldsymbol{a}_k, \boldsymbol{b}_k \rangle \to \langle \boldsymbol{a}, \boldsymbol{b} \rangle$;
(4) $|\boldsymbol{a}_k| \to |\boldsymbol{a}|$.

Proof. The first three statements follow immediately from the previous two theorems. The last one is a consequence of the third. □

Sequences in compact sets have additional properties. Before describing one such, we need some terminology: If $\{\boldsymbol{a}_k\}$ is a sequence, and $k_1 < k_2 < \dots$ is a strictly increasing sequence of positive integers, then the sequence $\boldsymbol{a}_{k_1}, \boldsymbol{a}_{k_2}, \dots$ is called a *subsequence* of $\boldsymbol{a}_k$ (more formally, if $f : \mathbb{N} \to \mathbb{R}^n$, $f(k) = \boldsymbol{a}_k$, is the function defining the sequence, then a subsequence of f is a function $f \circ g$, where $g : \mathbb{N} \to \mathbb{N}$ is strictly increasing). Elementary examples are the subsequences $\{a_{2k}\}$ of even terms and $\{a_{2k-1}\}$ of odd terms. It is an easy exercise to show that if a sequence converges, then any subsequence converges to the same limit. In fact, we proved this directly for the subsequence $\{r^{k+1}\}$ of $\{r^k\}$ in Examples and Remarks 1.8.1 (iii). The general case is similar.

Theorem 1.8.3. *If $K \subset \mathbb{R}^n$ is compact, then any sequence in K contains a convergent subsequence.*

Proof. Let $A = \{\boldsymbol{a}_k \mid k \in \mathbb{N}\}$. If A is a finite set, then there must be some $\boldsymbol{a}$ that equals $\boldsymbol{a}_k$ for infinitely many k. This yields a constant subsequence. We may therefore assume that A is infinite, and claim that there exists some $\boldsymbol{a} \in K$ with the following property: every neighborhood of $\boldsymbol{a}$ intersects $A \setminus \{\boldsymbol{a}\}$ (a point with this property is called a *limit point* of A; notice that any neighborhood of a limit point $\boldsymbol{a}$ of A must contain infinitely many elements of A: if it contained only finitely many, then we could find $\boldsymbol{b} \in A \setminus \{\boldsymbol{a}\}$ closest to $\boldsymbol{a}$, and the open ball around $\boldsymbol{a}$ with radius $|\boldsymbol{a} - \boldsymbol{b}|$ would no longer intersect $A \setminus \{\boldsymbol{a}\}$). To establish the claim, we argue by contradiction: suppose that every $\boldsymbol{a}$ in K has a neighborhood that does not intersect $A \setminus \{\boldsymbol{a}\}$. Apply this first to points in the complement of A: every point outside A (be it in K or outside of K) has a neighborhood disjoint from A, so that the complement of A is open, and A is closed. By Theorem 1.7.1, A is compact. Next, apply it to points inside A: every $\boldsymbol{a} \in A$ has an open neighborhood $U_{\boldsymbol{a}}$ such that $U_{\boldsymbol{a}} \cap A = \{\boldsymbol{a}\}$. Since A is infinite, this means that $\{U_{\boldsymbol{a}} \mid \boldsymbol{a} \in A\}$ is an open cover of A with no finite subcover, contradicting compactness.

Now that we have established the existence of a limit point $\boldsymbol{a}$, we can construct inductively a subsequence that converges to it. By assumption, there is an integer k_1 such that $\boldsymbol{a}_{k_1} \in B_1(\boldsymbol{a})$. If k_j is a given integer, then the ball $B_{1/(j+1)}(\boldsymbol{a})$ contains infinitely many elements of A, and we may therefore choose an integer $k_{j+1} > k_j$ such that $\boldsymbol{a}_{k_{j+1}}$ belongs to it. The subsequence $\{\boldsymbol{a}_{k_j}\}$ thus constructed satisfies $|\boldsymbol{a}_{k_j} - \boldsymbol{a}| < 1/j$ for all j, and therefore converges to $\boldsymbol{a}$. □

Corollary 1.8.2. *A bounded sequence in $\mathbb{R}^k$ has a convergent subsequence.*

Proof. By assumption, the sequence lies inside some ball, and this ball has compact closure, so Theorem 1.8.3 applies. □

Example 1.8.1. Let $r > 0$, $\alpha \in \mathbb{R}$, and $\boldsymbol{a}_k = \begin{bmatrix} r^k \cos(k\alpha) & r^k \sin(k\alpha) \end{bmatrix}^T \in \mathbb{R}^2$. If $r > 1$, then $\{\boldsymbol{a}_k\}$ diverges, because the sequence is unbounded: $|\boldsymbol{a}_k| = r^k$. Suppose next that $r < 1$. Since

$$0 \leq |u^1(\boldsymbol{a}_k)|, |u^2(\boldsymbol{a}_k)| \leq r^k,$$

and $r^k \to 0$, $|u^i(\boldsymbol{a}_k)| \to 0$ for $i = 1, 2$ by the squeeze theorem. But then $u^i(\boldsymbol{a}_k) \to 0$ because

$$-|u^i(\boldsymbol{a}_k)| \leq u^i(\boldsymbol{a}_k) \leq |u^i(\boldsymbol{a}_k).$$

Finally, when $r = 1$, the sequence will, in general, diverge, but must admit a convergent subsequence by Corollary 1.8.2 because $|\boldsymbol{a}_k| = 1$.

There is an alternative characterization of convergent sequences, one that is often useful because it does not explicitly involve a limit:

Definition 1.8.2. $\{\boldsymbol{a}_k\}$ is said to be a *Cauchy sequence* if for any $\varepsilon > 0$, there exists an integer N such that $|\boldsymbol{a}_k - \boldsymbol{a}_l| < \varepsilon$ for all $k, l \geq N$.

If $\boldsymbol{a}_k \to \boldsymbol{a}$, then $\{\boldsymbol{a}_k\}$ is Cauchy: given $\varepsilon > 0$ choose N so that $|\boldsymbol{a}_k - \boldsymbol{a}| < \varepsilon/2$ whenever $k \geq N$. If $k, l \geq N$, then

$$|\boldsymbol{a}_k - \boldsymbol{a}_l| = |(\boldsymbol{a}_k - \boldsymbol{a}) + (\boldsymbol{a} - \boldsymbol{a}_l)| \leq |\boldsymbol{a}_k - \boldsymbol{a}| + |\boldsymbol{a}_l - \boldsymbol{a}| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

On the other hand, if $\{\boldsymbol{a}_k\}$ is Cauchy, then it is bounded: there exists a positive integer N such that $|\boldsymbol{a}_k - \boldsymbol{a}_l| < 1$ whenever $k, l \geq N$. But then the sequence is contained inside the bounded set

$$B_1(\boldsymbol{a}_N) \cup \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{N-1}\}.$$

This, together with Corollary 1.8.2, implies:

Theorem 1.8.4. *A sequence in $\mathbb{R}^n$ converges if and only if it is a Cauchy sequence.*

Proof. We've already established that a convergent sequence is Cauchy, and that conversely, a Cauchy sequence $\{\boldsymbol{a}_k\}$ in $\mathbb{R}^n$ has a convergent subsequence, say, $\boldsymbol{a}_{k_j} \to \boldsymbol{a}$. But then the sequence itself must converge to $\boldsymbol{a}$: given $\varepsilon > 0$, choose N_1 such that

$|\boldsymbol{a}_k - \boldsymbol{a}_l| < \varepsilon/2$ whenever $k, l \geq N_1$, and choose N_2 such that $|\boldsymbol{a} - \boldsymbol{a}_{k_j}| < \varepsilon/2$ if $j \geq N_2$. Let N denote the larger of N_1 and N_2. Observing that $k_N \geq N$, we have

$$|\boldsymbol{a} - \boldsymbol{a}_k| \leq |\boldsymbol{a} - \boldsymbol{a}_{k_N}| + |\boldsymbol{a}_{k_N} - \boldsymbol{a}_k| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \text{ if } k \geq N,$$

which establishes the claim. □

There are metric spaces in which not every Cauchy sequence converges: consider for example the set $\mathbb{Q}$ of rational numbers with the distance function inherited as a subset of the real numbers. If x_n is the number representing the first n digits in the decimal expansion of $\sqrt{2}$, then $x_n \to \sqrt{2}$ in $\mathbb{R}$, and is therefore Cauchy. Each x_n is rational, and the sequence does not converge in $\mathbb{Q}$. A metric space in which every Cauchy sequence converges is said to be *complete*. Notice that every closed subset of a complete metric space is complete, and $\mathbb{Q}$ is of course not closed in $\mathbb{R}$.

Given a sequence $\{\boldsymbol{a}_m\}$ in $\mathbb{R}^n$, consider the sequence $\{\boldsymbol{s}_k\}$ of *partial sums*, where

$$\boldsymbol{s}_k = \boldsymbol{a}_1 + \cdots + \boldsymbol{a}_k.$$

We say the *infinite series* $\sum_{m=1}^{\infty} \boldsymbol{a}_m$ converges to L (and write $\sum_{m=1}^{\infty} \boldsymbol{a}_m = L$) if $\lim_{k\to\infty} \boldsymbol{s}_k = L$. The series is said to *converge absolutely* if the series of real numbers $\sum_m |\boldsymbol{a}_m|$ converges.

Completeness implies the following:

Theorem 1.8.5. *An absolutely convergent series converges.*

Proof. Let $\boldsymbol{s}_k = \boldsymbol{a}_1 + \cdots + \boldsymbol{a}_k$, $s_k = |\boldsymbol{a}_1| + \cdots + |\boldsymbol{a}_k|$. The claim will follow once we show that $\{\boldsymbol{s}_k\}$ is a Cauchy sequence. So let $\varepsilon > 0$. Since $\{s_k\}$ converges, it is Cauchy, and there exists an integer N such that $|s_m - s_l| < \varepsilon$ whenever $m > l \geq N$. By the triangle inequality, if $m > l \geq N$,

$$|\boldsymbol{s}_m - \boldsymbol{s}_l| = |\boldsymbol{a}_{l+1} + \cdots + \boldsymbol{a}_m| \leq |\boldsymbol{a}_{l+1}| + \cdots + |\boldsymbol{a}_m| = |s_m - s_l| < \varepsilon.$$ □

1.9 Limits and continuity

Recall from Section 1.8 that $\boldsymbol{a}$ is said to be a *limit point* of a set $A \subset \mathbb{R}^n$ if every neighborhood of $\boldsymbol{a}$ intersects $A \setminus \{\boldsymbol{a}\}$. Maps from A to $\mathbb{R}^m$ will be denoted in bold font when $m > 1$ (provided they are not linear). If $\boldsymbol{f} : A \to \mathbb{R}^m$ is a map, its *component functions* are the maps $f^i = u^i \circ \boldsymbol{f} : A \to \mathbb{R}$, $i = 1, \ldots, m$. The *graph* of $\boldsymbol{f}$ is the subset of $\mathbb{R}^{n+m}$ that consists of all $(\boldsymbol{a}, \boldsymbol{f}(\boldsymbol{a}))$ as $\boldsymbol{a}$ ranges over A. If $B \subset A$, define the *image* of B as $\boldsymbol{f}(B) = \{\boldsymbol{f}(\boldsymbol{b}) \mid \boldsymbol{b} \in B\}$. Finally, if $C \subset \mathbb{R}^m$, the *pre-image* of C is $\boldsymbol{f}^{-1}(C) = \{\boldsymbol{a} \in A \mid \boldsymbol{f}(\boldsymbol{a}) \in C\}$.

Definition 1.9.1. Let $A \subset \mathbb{R}^n$, $\boldsymbol{f} : A \to \mathbb{R}^m$ a map from A to $\mathbb{R}^m$, and $\boldsymbol{a}$ a limit point of A. We say the *limit of* $\boldsymbol{f}$ *at* $\boldsymbol{a}$ *equals* $\boldsymbol{b} \in \mathbb{R}^m$, and write

$$\lim_{\boldsymbol{p}\to\boldsymbol{a}} \boldsymbol{f}(\boldsymbol{p}) = \boldsymbol{b}, \quad \text{or} \quad \boldsymbol{f}(\boldsymbol{p}) \to \boldsymbol{b} \text{ as } \boldsymbol{p} \to \boldsymbol{a},$$

if for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|\boldsymbol{f}(\boldsymbol{p}) - \boldsymbol{b}| < \varepsilon \quad \text{whenever} \quad 0 < |\boldsymbol{p} - \boldsymbol{a}| < \delta, \quad \boldsymbol{p} \in A.$$

Notice that $\boldsymbol{a}$ need not belong to the domain A of $\boldsymbol{f}$. An alternative characterization is that for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\boldsymbol{f}(B_\delta(\boldsymbol{a}) \setminus \{\boldsymbol{a}\}) \subset B_\varepsilon(\boldsymbol{b}).$$

A useful way of determining limits is by means of the following:

Theorem 1.9.1. *With notation as in Definition 1.9.1, the limit of $\boldsymbol{f}$ at $\boldsymbol{a}$ equals $\boldsymbol{b}$ if and only if for every sequence $\{\boldsymbol{a}_k\}$ in A, $\boldsymbol{a}_k \neq \boldsymbol{a}$, that converges to $\boldsymbol{a}$, the sequence $\{\boldsymbol{f}(\boldsymbol{a}_k)\}$ converges to $\boldsymbol{b}$.*

Proof. Suppose $\lim_{\boldsymbol{p}\to\boldsymbol{a}} \boldsymbol{f}(\boldsymbol{p}) = \boldsymbol{b}$, and consider a sequence $\boldsymbol{a}_k \to \boldsymbol{a}$, $\boldsymbol{a}_k \neq \boldsymbol{a}$. Given $\varepsilon > 0$, we must show that there exists some integer N such that $|\boldsymbol{f}(\boldsymbol{a}_k) - \boldsymbol{b}| < \varepsilon$ for $k \geq N$. By hypothesis, for this ε there exists $\delta > 0$ such that $|\boldsymbol{f}(\boldsymbol{p}) - \boldsymbol{b}| < \varepsilon$ whenever $0 < |\boldsymbol{p} - \boldsymbol{a}| < \delta$, $\boldsymbol{p} \in A$. Since $\boldsymbol{a}_k \to \boldsymbol{a}$, there exists a positive integer N such that $0 < |\boldsymbol{a}_k - \boldsymbol{a}| < \delta$ when $k \geq N$. Thus, for $k \geq N$, $|\boldsymbol{f}(\boldsymbol{a}_k) - \boldsymbol{b}| < \varepsilon$, and $\boldsymbol{f}(\boldsymbol{a}_k) \to \boldsymbol{b}$.
Conversely, if the limit of $\boldsymbol{f}$ at $\boldsymbol{a}$ does not equal $\boldsymbol{b}$, then there exists an $\varepsilon > 0$ such that for every $\delta > 0$, there exists some $\boldsymbol{p} \in A$ within distance less than δ from $\boldsymbol{a}$ such that $|\boldsymbol{f}(\boldsymbol{p}) - \boldsymbol{b}| \geq \varepsilon$. Choose some such point $\boldsymbol{a}_k$ for each $\delta = 1/k$, $k \in \mathbb{N}$. Then $\{\boldsymbol{a}_k\}$ is a sequence that converges to $\boldsymbol{a}$, and $\lim_{k\to\infty} \boldsymbol{f}(\boldsymbol{a}_k)) \neq \boldsymbol{b}$. □

Since sequences have at most one limit, the limit of a function, if it exists, is also unique. Furthermore, the properties of sequences imply the following:

Theorem 1.9.2. *Let $A \subset \mathbb{R}^n$, $f, g, h : A \to \mathbb{R}$, and $\boldsymbol{a}$ a limit point of A.*

(a) *If $\lim_{\boldsymbol{p}\to\boldsymbol{a}} f(\boldsymbol{p}) = L$, and $\lim_{\boldsymbol{p}\to\boldsymbol{a}} g(\boldsymbol{p}) = M$, then $\lim_{\boldsymbol{p}\to\boldsymbol{a}}(f + g)(\boldsymbol{p}) = L + M$. A similar property holds for fg and for f/g (the latter provided $M \neq 0$).*

(b) *Let U be a neighborhood of $\boldsymbol{a}$. If $f(\boldsymbol{p}) \leq g(\boldsymbol{p}) \leq h(\boldsymbol{p})$ for all $\boldsymbol{p}$ in $U \setminus \{\boldsymbol{a}\}$, and if $\lim_{\boldsymbol{p}\to\boldsymbol{a}} f(\boldsymbol{p}) = \lim_{\boldsymbol{p}\to\boldsymbol{a}} h(\boldsymbol{p}) = L$, then $\lim_{\boldsymbol{p}\to\boldsymbol{a}} g(\boldsymbol{p}) = L$.*

Examples 1.9.1. (i) A *polynomial of degree k* on $\mathbb{R}^n$ is a function f of the form

$$f = a_0 + \sum_{j=1}^{k} \sum_{1\leq i_1,\ldots,i_j\leq n} a_{i_1\ldots i_j} u^{i_1} \cdots u^{i_j}, \qquad a_0, a_{i_1\ldots i_j} \in \mathbb{R}.$$

By Theorem 1.9.2, if f is a polynomial, then $\lim_{\boldsymbol{p}\to\boldsymbol{a}} f(\boldsymbol{p}) = f(\boldsymbol{a})$ for any $\boldsymbol{a}$.

(ii) Let $f : \mathbb{R}^2 \setminus \{\boldsymbol{0}\} \to \mathbb{R}$ be given by $f(x, y) = xy/(x^2 + y^2)$. Then $\boldsymbol{0}$ is a limit point of the domain of f. If $\boldsymbol{a}_n = (1/n, 0)$ then $\boldsymbol{a}_n \to \boldsymbol{0}$ and $f(\boldsymbol{a}_n) = 0$. If $\boldsymbol{b}_n = (1/n, 1/n)$, then $\boldsymbol{b}_n \to \boldsymbol{0}$ but $f(\boldsymbol{b}_n) = 1/2$ for all n. Thus, $\boldsymbol{f}$ has no limit at $\boldsymbol{0}$. The graph of f in $\mathbb{R}^3$ intersects each plane $ax + by = 0$ containing the z-axis in two half-lines at constant height $-ab/(a^2 + b^2)$ above the plane $\mathbb{R}^2 \times \{0\}$.

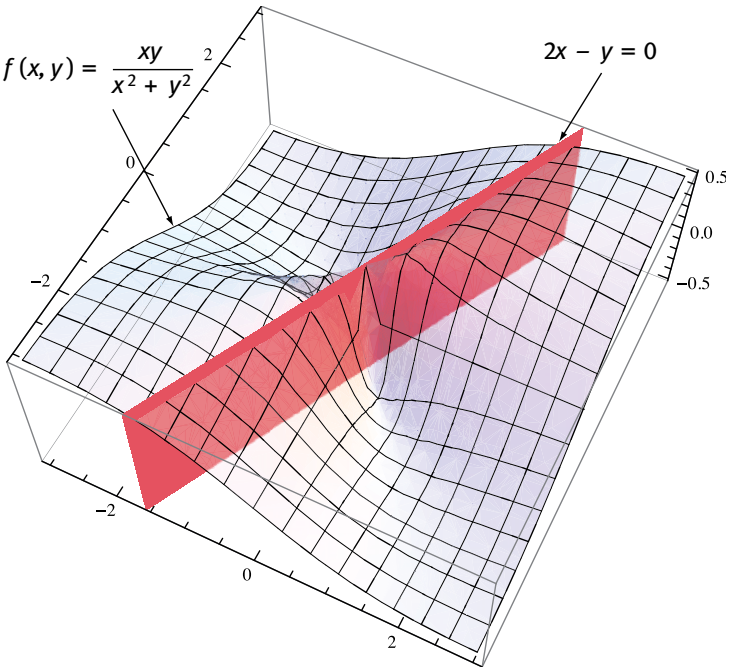


Fig. 1.2: The graph of the function from Examples 1.9.1 (ii)

Recall that the *composition* $\boldsymbol{f} \circ \boldsymbol{g}$ of maps $\boldsymbol{f}$ and $\boldsymbol{g}$ with appropriate domain and range is defined by $(\boldsymbol{f} \circ \boldsymbol{g})(\boldsymbol{a}) = \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{a}))$.

Corollary 1.9.1. *Suppose* $\boldsymbol{a}$ *is a limit point of* $A \subset \mathbb{R}^n$, $\boldsymbol{f}, \boldsymbol{g} : A \to \mathbb{R}^m$. *If* $\lim_{\boldsymbol{p}\to\boldsymbol{a}} \boldsymbol{f}(\boldsymbol{p}) = \boldsymbol{u}$ *and* $\lim_{\boldsymbol{p}\to\boldsymbol{a}} \boldsymbol{g}(\boldsymbol{p}) = \boldsymbol{v}$, *then for any* $c \in \mathbb{R}$,

$$\lim_{\boldsymbol{p}\to\boldsymbol{a}} (\boldsymbol{f} + \boldsymbol{g})(\boldsymbol{p}) = \boldsymbol{u} + \boldsymbol{v}, \quad \lim_{\boldsymbol{p}\to\boldsymbol{a}} c\boldsymbol{f}(\boldsymbol{p}) = c\,\boldsymbol{u}, \quad \lim_{\boldsymbol{p}\to\boldsymbol{a}} \langle \boldsymbol{f}(\boldsymbol{p}), \boldsymbol{g}(\boldsymbol{p}) \rangle = \langle \boldsymbol{u}, \boldsymbol{v} \rangle.$$

Furthermore, if $\boldsymbol{h} : B \subset \mathbb{R}^m \to \mathbb{R}^l$ *has limit* $\boldsymbol{r}$ *at* $\boldsymbol{u} \in B$, *then*

$$\lim_{\boldsymbol{p}\to\boldsymbol{a}} (\boldsymbol{h} \circ \boldsymbol{f}) = \boldsymbol{r}.$$

Proof. The first statement follows from Theorem 1.9.1 and Corollary 1.8.1. The second one is immediate from the definition of limit: Let $\boldsymbol{a}_i \to \boldsymbol{a}$. Since $\lim_{\boldsymbol{p}\to\boldsymbol{a}} \boldsymbol{f}(\boldsymbol{p}) = \boldsymbol{u}$, $\boldsymbol{f}(\boldsymbol{a}_i) \to \boldsymbol{u}$. But $\lim_{\boldsymbol{p}\to\boldsymbol{u}} \boldsymbol{h}(\boldsymbol{p}) = \boldsymbol{r}$, so $(\boldsymbol{h} \circ \boldsymbol{f})(\boldsymbol{a}_i) \to \boldsymbol{r}$, which implies the claimed limit. □

The following fundamental concept is defined here in the context of Euclidean spaces, since these remain our main focus. The reader should be aware, though, that the same definition can be, and is used in arbitrary metric spaces. Furthermore, all the results mentioned below that do not rely on the algebraic or order structures of $\mathbb{R}^n$ hold in these metric spaces.

Definition 1.9.2. Let $\boldsymbol{a} \in A \subset \mathbb{R}^n$. A map $\boldsymbol{f} : A \to \mathbb{R}^m$ is said to be *continuous at* $\boldsymbol{a}$ if for every $\varepsilon > 0$, there exists some $\delta > 0$ such that

$$\boldsymbol{f}(B_\delta(\boldsymbol{a}) \cap A) \subset B_\varepsilon(\boldsymbol{f}(\boldsymbol{a})).$$

$\boldsymbol{f}$ is said to be *continuous on* A if it continuous at every $\boldsymbol{a} \in A$.

If $\boldsymbol{a}$ is a limit point of A, then $\boldsymbol{f}$ is continuous at $\boldsymbol{a}$ if and only if $\lim_{\boldsymbol{p}\to\boldsymbol{a}} \boldsymbol{f}(\boldsymbol{p}) = \boldsymbol{f}(\boldsymbol{a})$. As an application, any polynomial is continuous on all Euclidean space by Examples 1.9.1 (i). If $\boldsymbol{a}$ is not a limit point of A, then $\boldsymbol{f}$ is automatically continuous at $\boldsymbol{a}$. For example, every sequence $\boldsymbol{f} : \mathbb{N} \subset \mathbb{R} \to \mathbb{R}^n$ is continuous on its domain: no matter what $\boldsymbol{f}(k)$ is, $B_{1/2}(k) \cap \mathbb{N} = \{k\}$, and so its image is contained in $B_\varepsilon(\boldsymbol{f}(k))$ for any ε.

Proposition 1.9.1. *$\boldsymbol{f} : A \to \mathbb{R}^m$ is continuous at $\boldsymbol{a} \in A$ if and only if $\boldsymbol{f}(\boldsymbol{a}_k) \to \boldsymbol{f}(\boldsymbol{a})$ for every sequence $\boldsymbol{a}_k \to \boldsymbol{a}$.*

Proof. If $\boldsymbol{a}$ is a limit point of A, the claim follows from Theorem 1.9.1. Otherwise, we have seen that the map is necessarily continuous at $\boldsymbol{a}$; on the other hand, if $\boldsymbol{a}_k \to \boldsymbol{a}$, then $\boldsymbol{a}_k$ must equal $\boldsymbol{a}$ for k sufficiently large, so certainly $\boldsymbol{f}(\boldsymbol{a}_k) \to \boldsymbol{f}(\boldsymbol{a})$. □

The above proposition, together with previous results, implies that continuity is well behaved with respect to common operations on maps.

Theorem 1.9.3. *Suppose $\boldsymbol{f}, \boldsymbol{g} : A \subset \mathbb{R}^n \to \mathbb{R}^m$, $\boldsymbol{h} : B \subset \mathbb{R}^m \to \mathbb{R}^l$, $c \in \mathbb{R}$. If $\boldsymbol{f}$ and $\boldsymbol{g}$ are continuous at $\boldsymbol{a} \in A$, then so are $\boldsymbol{f} + \boldsymbol{g}$, $c\boldsymbol{f}$, and $\langle \boldsymbol{f}, \boldsymbol{g} \rangle$. If $\boldsymbol{h}$ is continuous at $\boldsymbol{f}(\boldsymbol{a})$, then $\boldsymbol{h} \circ \boldsymbol{f}$ is continuous at $\boldsymbol{a}$.*

Proof. This follows, as noted above, from corresponding properties of sequences. Alternatively, one can use Corollary 1.9.1 when $\boldsymbol{a}$ is a limit point of A. If $\boldsymbol{a}$ is not a limit point, then there is nothing to prove. □

Examples 1.9.2. (i) A *rational function* is a quotient of two polynomials. Such a function is continuous on its domain.

(ii) Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x,y) = \begin{cases} \frac{xy^2}{x^2+y^2}, & \text{if } (x,y) \neq \mathbf{0}, \\ 0, & \text{if } (x,y) = \mathbf{0}. \end{cases}$$

By (i), f is continuous everywhere except perhaps at the origin. But $|x| \leq (x^2 + y^2)^{1/2}$ and $y^2 \leq x^2 + y^2$, so that $0 \leq f(x,y) \leq (x^2 + y^2)^{1/2}$. Since the term on the right goes to 0 as $(x,y) \to \mathbf{0}$, so does f, and f is continuous everywhere.

(iii) Since $\det : M_{n,n} \cong \mathbb{R}^{n^2} \to \mathbb{R}$ is a polynomial, it is continuous everywhere. It follows that the *general linear group* $GL(n)$, which by definition is the collection of all invertible $n \times n$ matrices, is open; this is because $GL(n) = \det^{-1}(\mathbb{R} \setminus \{0\})$.

Recall that for $\boldsymbol{f} : A \subset \mathbb{R}^n \to \mathbb{R}^m$, the *pre-image* $\boldsymbol{f}^{-1}(U)$ of a set $U \subset \mathbb{R}^m$ is the collection of all points $\boldsymbol{a}$ in A such that $\boldsymbol{f}(\boldsymbol{a}) \in U$. One often useful characterization of continuity is the following:

Theorem 1.9.4. *A map $\boldsymbol{f} : A \subset \mathbb{R}^n \to \mathbb{R}^m$ is continuous if and only if the preimage $\boldsymbol{f}^{-1}(U)$ of any open set U in $\mathbb{R}^m$ is open in A.*

Proof. Suppose $\boldsymbol{f}$ is continuous, and U is open in $\mathbb{R}^m$. It must be shown that if $\boldsymbol{a} \in \boldsymbol{f}^{-1}(U)$, then there exists an open set V in $\mathbb{R}^n$ such that $\boldsymbol{a} \in V \cap A \subset \boldsymbol{f}^{-1}(U)$. Since U is

open, there exists an $\varepsilon > 0$ such that the open ball of radius ε about $\boldsymbol{f}(\boldsymbol{a})$ is contained in U. Continuity of $\boldsymbol{f}$ means that there exists $\delta > 0$ such that $\boldsymbol{f}(B_\delta(\boldsymbol{a}) \cap A) \subset B_\varepsilon(\boldsymbol{f}(\boldsymbol{a}))$. In other words, $B_\delta(\boldsymbol{a}) \cap A \subset \boldsymbol{f}^{-1}(B_\varepsilon(\boldsymbol{f}(\boldsymbol{a}))) \subset \boldsymbol{f}^{-1}(U)$, so that $\boldsymbol{f}^{-1}(U)$ is open in A.
Conversely, suppose open sets have open pre-images in A. Given $\boldsymbol{a} \in A$ and $\varepsilon > 0$, $\boldsymbol{f}^{-1}(B_\varepsilon(\boldsymbol{f}(\boldsymbol{a})))$ is then open in A, and thus equals $U \cap A$ for some open set U in $\mathbb{R}^n$. Since $\boldsymbol{a} \in U$, there exists $\delta > 0$ such that the ball of radius δ around $\boldsymbol{a}$ is contained inside U. But then $B_\delta(\boldsymbol{a}) \cap A \subset U \cap A = \boldsymbol{f}^{-1}(B_\varepsilon(\boldsymbol{f}(a)))$, and $\boldsymbol{f}(B_\delta(\boldsymbol{a}) \cap A) \subset B_\varepsilon(\boldsymbol{f}(a))$. This shows that $\boldsymbol{f}$ is continuous at every $\boldsymbol{a} \in A$. □

Theorem 1.9.4 provides an easy proof of the fact that continuous maps send compact sets to compact sets:

Theorem 1.9.5. *If $\boldsymbol{f} : A \subset \mathbb{R}^n \to \mathbb{R}^m$ is continuous, then $\boldsymbol{f}(K)$ is compact for every compact set K contained in A.*

Proof. Let $\{U_\alpha\}$ be an open cover of $\boldsymbol{f}(K)$. By Theorem 1.9.4, each $\boldsymbol{f}^{-1}(U_\alpha)$ equals $V_\alpha \cap A$ for some open set V_α in $\mathbb{R}^n$. Since K is compact, there exist finitely many indices $\alpha_1, \ldots, \alpha_k$ such that $K \subset V_{\alpha_1} \cup \cdots \cup V_{\alpha_k}$. But then, $\boldsymbol{f}(K) \subset U_{\alpha_1} \cup \cdots \cup U_{\alpha_k}$, and $\boldsymbol{f}(K)$ is compact. Notice that we have used the fact, easily verified, that for any sets U and V, $\boldsymbol{f}(\boldsymbol{f}^{-1}(U) \cup \boldsymbol{f}^{-1}(V)) \subset U \cup V$. □

Theorem 1.9.5 has the following immediate application, usually referred to as the *extreme value theorem*:

Theorem 1.9.6. *Let $f : K \subset \mathbb{R}^n \to \mathbb{R}$ be continuous. If K is compact, then there exist $\boldsymbol{a}$, $\boldsymbol{b} \in K$ such that $f(\boldsymbol{a}) \leq f(\boldsymbol{p}) \leq f(\boldsymbol{b})$ for every $\boldsymbol{p} \in K$. (The numbers $f(\boldsymbol{a})$, $f(\boldsymbol{b})$ are called the* minimum and maximum values*, respectively, of f on K.)*

Proof. Since $f(K)$ is compact by Theorem 1.9.5, it is closed and bounded by Theorem 1.7.4. Being bounded, it has a least upper bound α, and being closed, it contains α, so that $\alpha = f(\boldsymbol{b})$ for some $\boldsymbol{b} \in K$. This implies the second inequality in the theorem. The first one is proved in a similar way. Alternatively, one can apply the above argument to $-f$, since the maximum value of $-f$ is the negative of the minimum value of f. □

There is a stronger version of continuity that plays an important role in Calculus:

Definition 1.9.3. A map $\boldsymbol{f} : A \subset \mathbb{R}^n \to \mathbb{R}^m$ is said to be *uniformly continuous* if for any $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$\boldsymbol{f}(B_\delta(\boldsymbol{a}) \cap A) \subset B_\varepsilon(\boldsymbol{f}(\boldsymbol{a})) \tag{1.9.1}$$

for any $\boldsymbol{a} \in A$. Alternatively, for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|\boldsymbol{f}(\boldsymbol{a}) - \boldsymbol{f}(\boldsymbol{b})| < \varepsilon, \text{ whenever } |\boldsymbol{a} - \boldsymbol{b}| < \delta, \quad \boldsymbol{a}, \boldsymbol{b} \in A.$$

For $\boldsymbol{f}$ to be continuous, one needs to find for each $\varepsilon > 0$ *and each* $\boldsymbol{a} \in A$ some $\delta > 0$ (which usually depends on *both* ε and $\boldsymbol{a}$) such (1.9.1) holds. When $\boldsymbol{f}$ is uniformly continuous, for a given $\varepsilon > 0$ there exists some $\delta > 0$ which works for *all* $\boldsymbol{a} \in A$. In particular,

a uniformly continuous map is continuous at every point. The converse is not true: for example, the function $f : (0,\infty) \to (0,\infty)$, where $f(x) = 1/x$, is continuous, but not uniformly continuous. Intuitively, given a fixed ε, the corresponding δ becomes smaller and smaller as the point a approaches 0. Specifically, let $\varepsilon = 1$. We claim there is no $\delta > 0$ such that $|1/a - 1/b| < 1$ for all a, b at distance less than δ from each other. In fact, if $\delta \geq 1$, take $a = 1/4$, $b = 1/2$. a and b are less than δ apart, but $1/a - 1/b = 2$. If $\delta < 1$, let $a = \delta/4$, $b = \delta/2$. Then $|a - b| < \delta$, but $1/a - 1/b = 2/\delta > 2$.

The obstruction in the above example is due to the fact that the domain of f is not compact:

Theorem 1.9.7. *If $K \subset \mathbb{R}^n$ is compact, then any continuous map $\boldsymbol{f} : K \to \mathbb{R}^m$ is uniformly continuous.*

Proof. Given $\varepsilon > 0$, there exists, for each $\boldsymbol{a} \in K$, some $\delta(\boldsymbol{a}) > 0$ such that

$$\boldsymbol{f}(B_{\delta(\boldsymbol{a})}(\boldsymbol{a})) \subset B_{\varepsilon/2}(\boldsymbol{f}(\boldsymbol{a})),$$

by continuity of $\boldsymbol{f}$. The collection of all $U_{\boldsymbol{a}} := B_{\delta(\boldsymbol{a})/2}(\boldsymbol{a})$, as $\boldsymbol{a}$ ranges over K, is an open cover of K, and therefore admits a finite subcover $U_{\boldsymbol{a}_1}, \ldots, U_{\boldsymbol{a}_k}$. Set $\delta := \min\{\delta(\boldsymbol{a}_1)/2, \ldots, \delta(\boldsymbol{a}_k)/2\} > 0$. We claim that if $\boldsymbol{a}$ and $\boldsymbol{b}$ are points in K at distance less than δ from each other, then $|\boldsymbol{f}(\boldsymbol{a}) - \boldsymbol{f}(\boldsymbol{b})| < \varepsilon$. To see this, observe that by assumption $\boldsymbol{a}$ belongs to some $U_{\boldsymbol{a}_i}$, and therefore

$$|\boldsymbol{f}(\boldsymbol{a}) - \boldsymbol{f}(\boldsymbol{a}_i)| < \frac{\varepsilon}{2}. \tag{1.9.2}$$

On the other hand,

$$|\boldsymbol{b} - \boldsymbol{a}_i| \leq |\boldsymbol{b} - \boldsymbol{a}| + |\boldsymbol{a} - \boldsymbol{a}_i| < \delta + \frac{\delta(\boldsymbol{a}_i)}{2} \leq \delta(\boldsymbol{a}_i),$$

so that

$$|\boldsymbol{f}(\boldsymbol{b}) - \boldsymbol{f}(\boldsymbol{a}_i)| < \frac{\varepsilon}{2}. \tag{1.9.3}$$

(1.9.2) and (1.9.3) together with the triangle inequality then yield the claim. □

Examples and Remarks 1.9.1. (i) Let N be a norm on $\mathbb{R}^n$, not necessarily the standard one. We claim N is uniformly continuous. To see this, observe that if $M = \max\{N(\boldsymbol{e}_i) \mid 1 \leq i \leq n\}$, then

$$\begin{aligned} N(\boldsymbol{a}) = N(\sum_i a_i \boldsymbol{e}_i) &\leq \sum_i |a_i| N(\boldsymbol{e}_i) \leq M(\sum_i |a_i|) \leq nM \max_i\{|a_i|\} \\ &\leq nM|\boldsymbol{a}|. \end{aligned}$$

By the triangle inequality,

$$|N(\boldsymbol{a}) - N(\boldsymbol{b})| \leq N(\boldsymbol{a} - \boldsymbol{b}) \leq nM|\boldsymbol{a} - \boldsymbol{b}|,$$

which establishes the claim. One consequence is the following equivalence of norms property for Euclidean space: given any norm N on $\mathbb{R}^n$, there exist $\alpha, \beta > 0$

such that

$$\alpha|\boldsymbol{a}| \le N(\boldsymbol{a}) \le \beta|\boldsymbol{a}|, \qquad \boldsymbol{a} \in \mathbb{R}^n. \tag{1.9.4}$$

The second inequality has already been proved. For the first one, we may assume that $\boldsymbol{a} \neq \boldsymbol{0}$. If α denotes the minimum value of N on the compact unit sphere, then $N(\boldsymbol{a}/|\boldsymbol{a}|) \ge \alpha$, so $N(\boldsymbol{a})$ is indeed no less than $\alpha|\boldsymbol{a}|$.
(1.9.4) in turn has the following important consequence: Euclidean space is complete with respect to any norm: the first inequality in (1.9.4) implies that a sequence which is Cauchy in the N-norm is Cauchy in the Euclidean one, hence converges in the Euclidean one. The second inequality then guarantees convergence in the N-norm.

(ii) A linear transformation from a vector space to itself is also called an *operator*. Consider the space of operators on $\mathbb{R}^n$, which we also identify with $M_{n,n}$ in the usual way. Let us denote the norm operator by $N : M_{n,n} \to \mathbb{R}$ for the moment, to distinguish it from the Euclidean norm on $M_{n,n} = \mathbb{R}^{n^2}$. It was shown in (i) that the space of operators on $\mathbb{R}^n$ is complete in the operator norm. We now revert to the old notation; i.e., $|L|$ will denote the operator norm of L. By the proof of Theorem 1.8.5, any absolutely convergent series converges. We may use this property to introduce the exponential of a linear $L : \mathbb{R}^n \to \mathbb{R}^n$. This concept will be useful in the next chapter, as it provides further insight into a large class of vector fields. The reader is assumed to be familiar with the fact that for a real number α, $e^\alpha = \sum_{k=0}^{\infty} \alpha^k/k$. The discussion will mostly be framed in the context of $n \times n$ matrices in view of the isomorphism $L_A \leftrightarrow A$ between the two spaces together with the fact that $L_A \circ L_B = L_{AB}$ for $A, B \in M_{n,n}$. First of all, notice that $|AB| \le |A||B|$, where $|A|$ is defined to be $|L_A|$: indeed, if $\boldsymbol{u} \in \mathbb{R}^n$ has unit norm, then by (1.4.1)

$$|(AB)\boldsymbol{u}| = |A(B\boldsymbol{u})| \le |A||B\boldsymbol{u}| \le |A||B|.$$

Induction now implies that $|A^k| \le |A|^k$. In particular, the *exponential series*

$$\exp(A) = e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!}, \qquad A \in M_{n,n},$$

converges absolutely (to the real number $e^{|A|}$), and therefore also converges. We call e^A the *exponential* of A. When $n = 1$, this is the usual exponential.
Absolute convergence also implies that

$$|e^A| \le e^{|A|}.$$

We emphasize again that if L^k denotes the composition of the operator L with itself k times, then, as noted earlier, the series

$$\exp(L) = e^L = \sum_{k=0}^{\infty} \frac{L^k}{k!}$$

converges absolutely; if L has matrix A in the standard basis, then e^L has matrix e^A.

(iii) As a concrete example, let $b \in \mathbb{R}$, and

$$B = \begin{bmatrix} 0 & -b \\ b & 0 \end{bmatrix}.$$

We assume the reader is familiar with the series

$$\cos x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!}, \qquad \sin x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!},$$

which converge for any real number x. A direct computation yields $B^2 = -b^2 I_2$, and an easy induction implies that $B^{2k} = (-1)^k b^{2k} I_2$. Thus,

$$B^{2k} = \begin{bmatrix} (-1)^k b^{2k} & 0 \\ 0 & (-1^k) b^{2k} \end{bmatrix},$$

and multiplying the above expression by B yields

$$B^{2k+1} = (-1)^k b^{2k} B = \begin{bmatrix} 0 & -(-1)^k b^{2k+1} \\ (-1)^k b^{2k+1} & \end{bmatrix}.$$

This means that

$$\sum_{i=0}^{k} \frac{B^i}{i!} = \begin{bmatrix} s_k & -t_k \\ t_k & s_k \end{bmatrix},$$

where s_k and t_k are the k-th partial sums for the above $\sin b$ and $\cos b$ series respectively. In other words,

$$e^B = \begin{bmatrix} \cos b & -\sin b \\ \sin b & \cos b \end{bmatrix}.$$

1.10 Exercises

1.1. Let a_i, $i = 1, \ldots, n$ be real numbers. Determine which, if any, of the following sets are subspaces of $\mathbb{R}^n$:
(i) $\{(x_1, \ldots, x_n) \mid \sum_{i=1}^n a_i x_i = 0\}$;
(ii) $\{(x_1, \ldots, x_n) \mid \sum_{i=1}^n a_i x_i^2 = 0\}$;
(iii) $\{(x_1, \ldots, x_n) \mid \sum_{i=1}^n a_i x_i = 1\}$;
(iv) $\{(x_1, \ldots, x_n) \mid (\sum_{i=1}^n a_i^2) x_1 + a_n x_2 = 0\}$.

1.2. (a) Show that the collection P_n of all polynomials (with real coefficients) of degree no larger than n is a subspace of the vector space of all real-valued functions.
(b) Prove that P_n has $\{1, x, x^2, \ldots, x^n\}$ as basis, and therefore has dimension $n + 1$.
(c) Show that $\{1, 1 + x, 1 + x + x^2, \ldots, 1 + x + \cdots x^n\}$ is also a basis of P_n. Find the coordinate vector of $a + bx + cx^2$ with respect to this basis.

1.3. Prove that the vector space of all polynomials is infinite-dimensional, and exhibit a basis for it.

1.4. Prove that any subset of a vector space that contains the zero vector is linearly dependent.

1.5. Let V denote an n-dimensional vector space, and $E \subset V$ a subset consisting of n elements. Prove that E is linearly independent if and only if E spans V, and is in this case a basis of V.

1.6. Determine whether the following sets form a basis of $\mathbb{R}^3$:
(i) $\{(1,0,1),(2,1,0),(4,1,2)\}$;
(ii) $\{(1,0,1),(2,1,0)\}$;
(iii) $\{(1,0,1),(2,1,0),(4,1,2),(1,1,1)\}$;
(iv) $\{(1,0,1),(2,1,0),(5,1,2)\}$.

1.7. Let $L : \mathbb{R}^3 \to \mathbb{R}^3$ be given by

$$L\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x + 2y + 4z \\ y + z \\ x + 2z \end{bmatrix}.$$

Find bases for the kernel and for the image of L.

1.8. Let P_n denote the space of polynomials of degree $\leq n$ with its standard basis $\{1, x, \ldots, x^n\}$. Find the matrix of the derivative operator $D : P_n \to P_n$, $Dp(x) := p'(x)$, with respect to this basis.

1.9. A set $A \subset \mathbb{R}^n$ is said to be *convex* if it contains the line segment joining any two points of A; i.e., $\boldsymbol{a} + t(\boldsymbol{b} - \boldsymbol{a}) \in A$ for all $\boldsymbol{a}, \boldsymbol{b} \in A$ and $t \in [0,1]$. Show that if A is convex and $L : \mathbb{R}^n \to \mathbb{R}^m$ is linear, then $L(A)$ is also convex.

1.10. Let V be a vector space with basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$. Show that for any n elements $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ in some vector space W, there exists one and only one linear transformation $L : V \to W$ such that $L\boldsymbol{v}_i = \boldsymbol{w}_i$ for $i = 1, \ldots, n$. Prove that if the set $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$ is linearly independent, then L is one-to-one, and if it is a basis, then L is an isomorphism.

1.11. (a) Show that if $L : V \to W$ is a linear transformation, then the kernel of L is a subspace of V and the image of L a subspace of W.
(b) Prove that L is one-to-one if and only if its kernel consists of the zero vector only.

1.12. Suppose $L : V \to W$ is a linear transformation between vector spaces of the same dimension. Use Theorem 1.2.2 to show that the following statements are equivalent:
(i) L is one-to-one;
(ii) L is onto;
(iii) L is an isomorphism.

1.13. A linear operator L on an inner product space V is said to be a *linear isometry* or an *orthogonal transformation* if it preserves the inner product; i.e., if $\langle L\boldsymbol{v}, L\boldsymbol{w}\rangle = \langle \boldsymbol{v}, \boldsymbol{w}\rangle$ for all $\boldsymbol{v}, \boldsymbol{w} \in V$.

(a) Show that L is a linear isometry if and only if L preserves norms; i.e., iff $|L\boldsymbol{v}| = |\boldsymbol{v}|$ for all $\boldsymbol{v} \in V$. *Hint:* $\langle \boldsymbol{v}, \boldsymbol{w}\rangle = \frac{1}{2}(|\boldsymbol{v} + \boldsymbol{w}|^2 - |\boldsymbol{v}|^2 - |\boldsymbol{w}|^2)$.

(b) Prove that a linear isometry is an isomorphism.

(c) Show that a linear operator on an n-dimensional inner product space V is a linear isometry if and only if its matrix A with respect to any orthonormal basis satisfies $AA^T = I_n$.

1.14. Recall that the *trace* of an $n \times n$ matrix A is the sum $\operatorname{tr} A = \sum_i a_{ii}$ of its diagonal elements.

(a) Show that $\operatorname{tr}(AB) = \operatorname{tr}(BA)$. *Hint:* Write the (i, i)-th element of AB as $\sum_k a_{ik}b_{ki}$. Next write the (k, k)-th element of BA as $\sum_i b_{ki}a_{ik}$.

(b) Use Example 1.2.5 to show that similar matrices have the same trace. One may therefore define the trace of a linear transformation $L : V \to V$ to be the trace of its matrix with respect to any basis of V.

1.15. Let V be a (finite-dimensional) inner product space, and A a nonempty subset of V. Prove that $(A^\perp)^\perp = \operatorname{span} A$.

1.16. The collection $\mathcal{L}(V, W)$ of all linear transformations $L : V \to W$ is a vector space with the operations $(L_1 + L_2)\boldsymbol{v} = L_1\boldsymbol{v} + L_2\boldsymbol{v}$, $(aL)\boldsymbol{v} = aL\boldsymbol{v}$ for $L_i \in \mathcal{L}(V, W)$, $a \in \mathbb{R}$, $\boldsymbol{v} \in V$. Show that if V is n-dimensional with basis $\mathcal{B}$ and W m-dimensional with basis $\mathcal{C}$, then the map

$$\begin{aligned} \mathcal{L}(V, W) &\to M_{m,n}, \\ L &\mapsto [L]_{[\mathcal{B}],[\mathcal{C}]} \end{aligned}$$

which assigns to each linear transformation its matrix with respect to the given bases is an isomorphism. This shows in particular that $\mathcal{L}(V, W)$ has dimension $(\dim V)(\dim W)$.

1.17. Find the determinant of the matrix

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix}$$

whose elements below the diagonal are all zero.

1.18. Prove that adding a multiple of a column to another column (or a multiple of a row to another row) of a matrix does not change its determinant. This allows us to replace the matrix by one that has only one nonzero entry in some row or column and

expand along that row or column. Use this to compute

$$\det \begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 4 \\ 2 & 1 & 2 \end{bmatrix}.$$

1.19. Prove (1.3.6). Show that the set $U \subset \mathbb{R}^{n^2}$ of invertible matrices is open, and that the map $U \to U$ which sends a matrix to its inverse is continuous.

1.20. Show that equality holds in the Cauchy-Schwarz inequality from Theorem 1.4.1 if and only if the vectors are linearly dependent.

1.21. Let V_1, V_2 be subspaces of V. Prove that $\dim(V_1 + V_2) = \dim(V_1) + \dim(V_2) - \dim(V_1 \cap V_2)$. *Hint:* Start out with a basis of $V_1 \cap V_2$, and extend it first to a basis of V_1, then to a basis of V_2.

1.22. (a) Show that the interior A^0 of a subset A of a metric space is equal to the union of all open sets that are contained in A. In other words, A^0 is the largest open set contained in A.

(b) Prove that the closure $\bar{A}$ of A is equal to the intersection of all closed sets that contain A; i.e., $\bar{A}$ is the smallest closed set that contains A.

1.23. Let $A, B \subset \mathbb{R}^n$.

(a) Show that $(A \cap B)^0 = A^0 \cap B^0$, but that in general, $(A \cup B)^0 \neq A^0 \cup B^0$.

(b) Prove that $\overline{A \cup B} = \overline{A} \cup \overline{B}$, but in general $\overline{A \cap B} \neq \overline{A} \cap \overline{B}$.

1.24. Determine the interior, boundary, and closure of the following subsets of $\mathbb{R}^2$:

(a) $\{(x, y) \mid y \geq x^2\}$;

(b) $\mathbb{R} \times \{0\}$;

(c) $\mathbb{R} \times \mathbb{Q}$;

(d) $\mathbb{Q} \times \mathbb{Q}$.

(You may use the fact that any open interval in $\mathbb{R}$ contains a rational number).

1.25. (a) Show that if $\boldsymbol{a}$ is a boundary point of a set A, then there exists a sequence in A that converges to $\boldsymbol{a}$.

(b) Prove that a set C is closed if and only if for any convergent sequence contained in C, the limit of the sequence belongs to C.

1.26. Prove that a set $A \subset \mathbb{R}^n$ is open if and only if $A \cap \partial A = \emptyset$, and closed if and only if $\partial A \subset A$.

1.27. Show that a subset C of A is closed in A if and only if for any sequence in C that converges in A, the limit of the sequence belongs to C.

1.28. For a subset A of a metric space X, denote by A^c the complement $X \setminus A$ of A. Prove that $\partial A = \overline{A} \cap \overline{A^c}$.

1.29. Let U be open in $\mathbb{R}^n$, $\boldsymbol{f} : U \to \mathbb{R}^n$ a continuous map, and A a set whose closure lies in U.
(a) Show that $\boldsymbol{f}(\overline{A}) \subset \overline{f(A)}$.
(b) Show that if $\boldsymbol{f}$ is a *homeomorphism* (i.e., $\boldsymbol{f}$ is one-to-one and its inverse is also continuous), then $\boldsymbol{f}(\partial A) = \partial \boldsymbol{f}(A)$.

1.30. Let A be a nonempty subset of $\mathbb{R}^n$. Prove or disprove:
(i) Any boundary point of A is a limit point of A.
(ii) Any limit point of A is a boundary point of A.
(iii) Any boundary point of A that does not belong to A is a limit point of A.
(iv) Any limit point of A that does not belong to A is a boundary point of A.
(v) If A is closed, then it contains all its limit points.
(vi) If A contains all its limit points, then it is closed.

1.31. Give one or more examples of a set $A \subset \mathbb{R}^n$ that
(i) is neither open nor closed;
(ii) is both open and closed;
(iii) has empty boundary;
(iv) has all of $\mathbb{R}^n$ as boundary.

1.32. (a) Let $f : [a, b] \to \mathbb{R}$ be a continuous function. Show that the graph $\{(x, f(x)) \mid a \le x \le b\}$ is a closed subset of $\mathbb{R}^2$. Give an example of a noncontinuous function $f : [a, b] \to \mathbb{R}$ whose graph is not closed.
(b) Is the set $A = \{(x, \sin(1/x) \mid 0 < x < 1\} \subset \mathbb{R}^2$ closed? If not, find its closure. ($\overline{A}$ is called the *topologist's sine curve*).

1.33. Suppose $\{\boldsymbol{a}_k\}$ is a convergent sequence in $\mathbb{R}^n$. Prove that any subsequence converges to the same limit.

1.34. Determine whether the sequence

$$\sqrt{2}, \sqrt{2 + \sqrt{2}}, \sqrt{2 + \sqrt{2 + \sqrt{2}}}, \ldots$$

converges. If it does converge, determine its limit.

1.35. Suppose $\{\boldsymbol{a}_k\}$ is a sequence in $\mathbb{R}^n$ that satisfies

$$|\boldsymbol{a}_{k+1} - \boldsymbol{a}_k| \le \frac{1}{2}|\boldsymbol{a}_k - \boldsymbol{a}_{k-1}|, \qquad k > 1.$$

Prove that the sequence converges.

1.36. Suppose $C_k \subset \mathbb{R}^n$, $k = 1, 2, \ldots$, is a sequence of nonempty compact sets with $C_k \supset C_{k+1}$ for all k.
(a) Show that $\cap_{k=1}^{\infty} C_k \neq \emptyset$.
(b) Prove, by means of an example, that the statement is in general false if the C_k are not compact.

1.37. The *diameter* of a bounded set $A \subset \mathbb{R}^n$ is

$$\operatorname{diam} A = \sup\{|\boldsymbol{a} - \boldsymbol{b}| \mid \boldsymbol{a}, \boldsymbol{b} \in A\}.$$

Prove that if A is compact, then there exist $\boldsymbol{a}, \boldsymbol{b} \in A$ such that $|\boldsymbol{a} - \boldsymbol{b}| = \operatorname{diam} A$.

1.38. Given nonempty subsets A, B of $\mathbb{R}^n$, define the *distance* between A and B to be the number $\mathrm{d}(A, B) = \inf\{|\boldsymbol{a} - \boldsymbol{b}| \mid \boldsymbol{a} \in A, \boldsymbol{b} \in B\}$.
(a) Give an example of two disjoint closed sets A and B with $\mathrm{d}(A, B) = 0$.
(b) Prove that if A is compact and B is closed, then there exist $\boldsymbol{a} \in A$ and $\boldsymbol{b} \in B$ such that $\mathrm{d}(A, B) = |\boldsymbol{a} - \boldsymbol{b}|$.

1.39. Let $U \subset \mathbb{R}^n$, $f : U \to \mathbb{R}$ be uniformly continuous. Show that if U is bounded, then so is f; i.e., $|f(\boldsymbol{a})| \leq M$ for some $M > 0$.

1.40. Suppose $U \subset \mathbb{R}^n$, and $\boldsymbol{f} : U \to \mathbb{R}^m$ is uniformly continuous. Show that if $\{\boldsymbol{a}_k\}$ is a Cauchy sequence in U, then $\{\boldsymbol{f}(\boldsymbol{a}_k)\}$ is also Cauchy. Show by means of an example that the conclusion is not necessarily true if $\boldsymbol{f}$ is merely continuous.

1.41. A subset $A \subset \mathbb{R}^n$ is said to be *complete* if any Cauchy sequence in A converges to some point in A. Prove that if A is closed, then it is complete. Give examples that show the above conclusion is not necessarily true if A is not closed.

1.42. Let U denote an open set in $\mathbb{R}^2$, and $f : U \to \mathbb{R}$ a continuous function. Prove that f cannot be one-to-one.

1.43. Determine whether or not the following maps are continuous:
(i) $f : \mathbb{R}^2 \to \mathbb{R}$, where

$$f(x, y) = \begin{cases} \frac{x^2 y^{1/3}}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0); \\ 0 & \text{if } (x, y) = (0, 0). \end{cases}$$

(ii) $f : \mathbb{R}^2 \to \mathbb{R}$, where $f(x, y) = \max\{x, y\}$.
(iii) $f : \mathbb{R}^n \to \mathbb{R}$, where $f(\boldsymbol{a}) = |\boldsymbol{a}|$.
(iv) $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$, where

$$\boldsymbol{f}(\boldsymbol{a}) = \begin{cases} \boldsymbol{a}/\sqrt{|\boldsymbol{a}|} & \text{if } \boldsymbol{a} \neq \boldsymbol{0}; \\ \boldsymbol{0} & \text{if } \boldsymbol{a} = \boldsymbol{0}. \end{cases}$$

1.44. Show that any linear transformation $L : \mathbb{R}^n \to \mathbb{R}^m$ is absolutely continuous on $\mathbb{R}^n$.

1.45. A *separation* of a metric space X is a pair U, V of nonempty disjoint open sets whose union equals X. X is said to be *connected* if there is no separation of X.
(a) Prove that X is connected if and only if the only subsets of X that are both open and closed are $\emptyset$ and X.
(b) Show that a pair U, V of nonempty subsets of X whose union equals X form a separation of X if and only if $\overline{U} \cap V = U \cap \overline{V} = \emptyset$.

1.46. (a) With the terminology from Exercise 1.45, suppose U and V form a separation of X. Show that any connected subset of X is entirely contained in either U or V.
(b) Prove that if a subset of X is connected, then so is its closure.

1.47. (a) Let $A \subset \mathbb{R}^n$, $\boldsymbol{f} : A \to \mathbb{R}^m$ a continuous map. Prove that $\boldsymbol{f}(A)$ is connected if A is.
(b) $A \subset \mathbb{R}^n$ is said to be *path connected* if given any $\boldsymbol{p}, \boldsymbol{q} \in A$, there exists a continuous map $\boldsymbol{c} : [a, b] \to A$ such that $\boldsymbol{c}(a) = \boldsymbol{p}$ and $\boldsymbol{c}(b) = \boldsymbol{q}$. Show that if A is path connected, then it is connected.
(c) Prove that the topologist's sine curve E from Exercise 1.32 is connected but not path connected. *Hint:* for connectivity, use part (a) and Exercise 1.46 (b). For path connectivity, consider a point $\boldsymbol{p}$ in E that lies on the y-axis. Show that the image of any continuous $\boldsymbol{c} : [a, b] \to E$ with $\boldsymbol{c}(a) = \boldsymbol{p}$ is contained inside the y-axis.

1.48. Let $A \subset \mathbb{R}^n$ be connected, $\boldsymbol{f} : A \to \mathbb{R}^m$ a continuous map. Prove the intermediate value theorem: If $\boldsymbol{a}, \boldsymbol{b} \in A$ and c is a number between $f(\boldsymbol{a})$ and $f(\boldsymbol{b})$, then there exists some $\boldsymbol{x} \in A$ such that $f(\boldsymbol{x}) = c$.

1.49. Suppose that $\sum_{k=0}^{\infty} A_k = A$, $\sum_{k=0}^{\infty} B_k = B$ are absolutely convergent series of $n \times n$ matrices (or operators on $\mathbb{R}^n$). Define

$$C_k = \sum_{i=0}^{k} A_i B_{k-i}.$$

The goal of this exercise is to establish that $\sum_{k=0}^{\infty} C_k = AB$.
(a) Let a_k, b_k, and c_k denote the k-th partial sums of the series $\sum A_i$, $\sum B_i$, and $\sum C_i$ respectively. Show that $c_{2k} - a_k b_k = s_k + t_k$, where

$$s_k = \sum_{\substack{0 \le i \le k \\ k+1 \le j \le 2k \\ i+j \le 2k}} A_i B_j, \qquad t_k = \sum_{\substack{0 \le i \le k \\ k+1 \le j \le 2k \\ i+j \le 2k}} A_j B_i.$$

(b) Prove that

$$|s_k| \le \left(\sum_{i=0}^{\infty} |A_i| \right) \left(\sum_{j=k+1}^{2k} |B_j| \right),$$

and deduce that $|s_k| \to 0$.
(c) Prove a similar result for $|t_k|$, and conclude that $\sum_{k=0}^{\infty} C_k = AB$.

1.50. (a) Use Exercise 1.49 to prove if two $n \times n$ matrices A and B commute (i.e., $AB = BA$), then $e^{A+B} = e^A e^B = e^B e^A$.
(b) Show that for $A, B \in M_n$, if B is invertible, then $Be^A B^{-1} = e^{BAB^{-1}}$.

1.51. (a) Prove that for any $n \times n$ matrix A, e^A is invertible, and $(e^A)^{-1} = e^{-A}$.
(b) Show that

$$\exp \begin{bmatrix} a & -b \\ b & a \end{bmatrix} = e^a \begin{bmatrix} \cos b & -\sin b \\ \sin b & \cos b \end{bmatrix}.$$

1.52. An $n \times n$ matrix A is said to be *skew-symmetric* if $A + A^T = \mathbf{0}$, and *orthogonal* if $AA^T = I_n$, see also Exercise 1.13. Show that the exponential of a skew-symmetric matrix is orthogonal.

1.53. Let $A \subset \mathbb{R}^n$, $f : A \to \mathbb{R}$ a bounded function. Given $B \subset A$, define

$$m_B(f) = \inf\{f(\boldsymbol{b}) \mid \boldsymbol{b} \in B\}, \qquad M_B(f) = \sup\{f(\boldsymbol{b}) \mid \boldsymbol{b} \in B\}.$$

Denote by D the set of points in A where f is discontinuous.
(a) Show that for any $\varepsilon > 0$, the set

$$D_\varepsilon = \{\boldsymbol{a} \in A \mid M_U(f) - m_U(f) \geq \varepsilon \text{ for any neighborhood } U \text{ of } A\}$$

is contained inside D.
(b) Prove that $D = \cup_{k=1}^{\infty} D_{1/k}$.

2 Differentiation

In calculus of one variable, the derivative of a function $f : A \subset \mathbb{R} \to \mathbb{R}$ at an interior point a of A is defined to be the number

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h},$$

provided the limit exists. It is geometrically interpreted as the slope of the line that best approximates the graph of f at $(a, f(a))$. This line, when parallel translated so that it passes through the origin is the graph of the linear transformation $L : \mathbb{R} \to \mathbb{R}$ given by $Lh = f'(a)h$. The above equation can be rewritten in terms of L:

$$\lim_{h \to 0} \frac{f(a+h) - f(a) - Lh}{h} = 0.$$

If we now replace numerator and denominator by their absolute values or norms, then this expression makes sense even when f is a map between Euclidean spaces of dimension higher than one. This is the approach we will adopt.

2.1 The derivative

Definition 2.1.1. Let $\boldsymbol{f} : A \subset \mathbb{R}^n \to \mathbb{R}^m$ be a map. $\boldsymbol{f}$ is said to be *differentiable* at an interior point $\boldsymbol{a}$ of A if there exists a linear transformation $L : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\lim_{\boldsymbol{h} \to \boldsymbol{0}} \frac{|\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a}) - L\boldsymbol{h}|}{|\boldsymbol{h}|} = 0. \tag{2.1.1}$$

It is worth remarking that if such an L exists, it is unique: Indeed, if M is another linear transformation with the same property, then for any unit vector $\boldsymbol{u} \in \mathbb{R}^n$, and $0 \neq t \in \mathbb{R}$,

$$\begin{aligned} |(L-M)\boldsymbol{u}| &= \frac{|(L-M)(t\boldsymbol{u})|}{|t\boldsymbol{u}|} \\ &= \frac{|(L(t\boldsymbol{u}) - \boldsymbol{f}(\boldsymbol{a}+t\boldsymbol{u}) + \boldsymbol{f}(\boldsymbol{a})) + (\boldsymbol{f}(\boldsymbol{a}+t\boldsymbol{u}) - \boldsymbol{f}(\boldsymbol{a}) - M(t\boldsymbol{u}|))|}{|t\boldsymbol{u}|} \\ &\leq \frac{|\boldsymbol{f}(\boldsymbol{a}+t\boldsymbol{u}) - \boldsymbol{f}(\boldsymbol{a}) - Lt(\boldsymbol{u})|}{|t\boldsymbol{u}|} + \frac{|\boldsymbol{f}(\boldsymbol{a}+t\boldsymbol{u}) - \boldsymbol{f}(\boldsymbol{a}) - M(t\boldsymbol{u}|)}{|t\boldsymbol{u}|}. \end{aligned}$$

Letting $t \to 0$, we see that $|(L-M)\boldsymbol{u}|$ is less than any positive number, so that $L\boldsymbol{u} = M\boldsymbol{u}$. Since this is true for arbitrary unit $\boldsymbol{u}$, $L = M$ (recall that a linear transformation is entirely determined by the image of basis vectors).

Definition 2.1.2. If $\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$, the *derivative* $D\boldsymbol{f}(\boldsymbol{a})$ *of* $\boldsymbol{f}$ *at* $\boldsymbol{a}$ is defined to be the linear transformation L from (2.1.1).

Examples 2.1.1. (i) If $L : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation, then L is differentiable at any $\boldsymbol{a} \in \mathbb{R}^n$, and $DL(\boldsymbol{a}) = L$. This is clear, since when substituting in (2.1.1), the numerator $L(\boldsymbol{a} + \boldsymbol{h}) - L(\boldsymbol{a}) - L(\boldsymbol{h})$ is identically zero.

(ii) A *curve* in $\mathbb{R}^n$ is a continuous map $\boldsymbol{c} : I \to \mathbb{R}^n$, where I is an interval in the real line. Let $c^i = u^i \circ \boldsymbol{c}$ denote the i-th component function of $\boldsymbol{c}$. It is a real-valued function of one variable. Now, $\boldsymbol{c}$ is differentiable at an interior point t of I if and only if there exists a linear transformation $D\boldsymbol{c}(t) : \mathbb{R} \to \mathbb{R}^n$ such that

$$\lim_{h\to 0} \frac{|\boldsymbol{c}(t+h) - \boldsymbol{c}(t) - D\boldsymbol{c}(t)h|}{h} = 0.$$

This is equivalent to

$$\lim_{h\to 0} \frac{c^i(t+h) - c^i(t) - (u^i \circ D\boldsymbol{c}(t))h}{h} = 0, \qquad i = 1, \dots, n.$$

But c^i is an ordinary function, so that $\boldsymbol{c}$ is differentiable if and only if each component function is differentiable in the usual sense, and in this case, the matrix of $D\boldsymbol{c}(t)$ is commonly referred to as the *velocity vector*

$$\boldsymbol{c}'(t) = \begin{bmatrix} c^{1\prime}(t) \\ \vdots \\ c^{n\prime}(t) \end{bmatrix}$$

at t of the curve. Its norm is called the *speed* of the curve.

(iii) Let $f : \mathbb{R}^n \to \mathbb{R}$, $f(\boldsymbol{u}) = |\boldsymbol{u}|^2$, denote the square of the distance from the origin function. For any $\boldsymbol{a} \in \mathbb{R}^n$, the map $L_{\boldsymbol{a}} : \mathbb{R}^n \to \mathbb{R}$, where $L_{\boldsymbol{a}}\boldsymbol{h} = 2\langle \boldsymbol{a}, \boldsymbol{h}\rangle$, is linear. We claim that $Df(\boldsymbol{a}) = L_{\boldsymbol{a}}$. In fact,

$$\frac{1}{|\boldsymbol{h}|}(f(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a}) - L_{\boldsymbol{a}}\boldsymbol{h}) = \frac{1}{|\boldsymbol{h}|}(|\boldsymbol{a}+\boldsymbol{h}|^2 - |\boldsymbol{a}|^2 - 2\langle \boldsymbol{a}, \boldsymbol{h}\rangle) = |\boldsymbol{h}|$$

clearly goes to zero as $\boldsymbol{h} \to \boldsymbol{0}$. This establishes the claim.

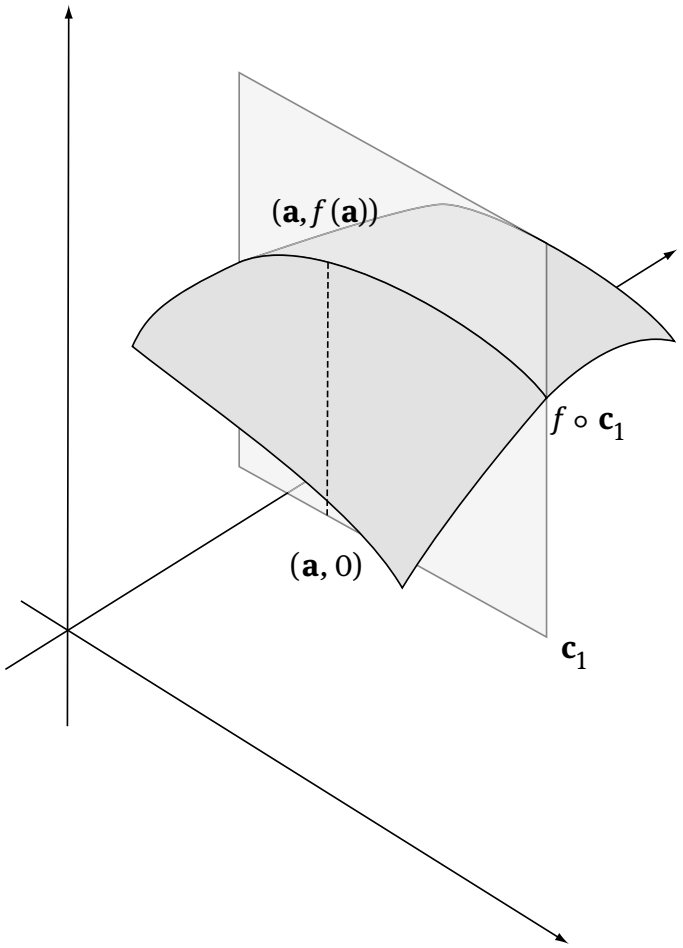
In order to see what the matrix of the linear transformation $D\boldsymbol{f}(a)$ looks like in general, we need some terminology:

Definition 2.1.3. Let $f : A \subset \mathbb{R}^n \to \mathbb{R}$ be a real-valued function. The *i-th partial derivative* $D_i f(\boldsymbol{a})$ *of* f at an interior point $\boldsymbol{a}$ of A is defined to be the limit

$$D_i f(\boldsymbol{a}) = \lim_{t\to 0} \frac{f(\boldsymbol{a} + t\boldsymbol{e}_i) - f(\boldsymbol{a})}{t}, \qquad i = 1, \dots, n,$$

if it exists.

Notice that $D_i f(\boldsymbol{a})$ is just the derivative at 0 of the real-valued function of one variable $f \circ \boldsymbol{c}_i$, where $t \mapsto \boldsymbol{c}_i(t) = \boldsymbol{a} + t\boldsymbol{e}_i$ is a parametrization of the line though $\boldsymbol{a}$ parallel to the i-th axis. The graph of $f \circ \boldsymbol{c}_i$ is the intersection in $\mathbb{R}^{n+1}$ of the graph of f with the the plane through $(\boldsymbol{a}, 0)$ that is parallel to the i-th and the $(n+1)$-th coordinate axes. We will sometimes use the classical notation $\partial f/\partial x^i$ for $D_i f$.



The matrix of $D\boldsymbol{f}(\boldsymbol{a})$ with respect to the standard bases is called the *Jacobian matrix of* $\boldsymbol{f}$ *at* $\boldsymbol{a}$, and will be denoted $[D\boldsymbol{f}(\boldsymbol{a})]$.

Theorem 2.1.1. *Suppose* $\boldsymbol{f} : A \subset \mathbb{R}^n \to \mathbb{R}^m$ *is differentiable at* $\boldsymbol{a}$. *If* $f^i = u^i \circ \boldsymbol{f}$, $i = 1, \dots, m$, *denote the component functions of* $\boldsymbol{f}$, *then the partial derivatives* $D_j f^i(\boldsymbol{a})$ *exist, and the Jacobian matrix of* $\boldsymbol{f}$ *at* $\boldsymbol{a}$ *is*

$$[D\boldsymbol{f}(\boldsymbol{a})] = \begin{bmatrix} D_1 f^1(\boldsymbol{a}) & D_2 f^1(\boldsymbol{a}) & \dots & D_n f^1(\boldsymbol{a}) \\ D_1 f^2(\boldsymbol{a}) & D_2 f^2(\boldsymbol{a}) & \dots & D_n f^2(\boldsymbol{a}) \\ \vdots & \vdots & \ddots & \vdots \\ D_1 f^m(\boldsymbol{a}) & D_2 f^m(\boldsymbol{a}) & \dots & D_n f^m(\boldsymbol{a}) \end{bmatrix}.$$

Thus, the i*-th row of* $[D\boldsymbol{f}(\boldsymbol{a})]$ *is the Jacobian matrix of* f^i *at* $\boldsymbol{a}$.

Proof. For $j = 1, \dots, n$, define a curve $\boldsymbol{r}_j$ in a neighborhood of 0 by

$$\boldsymbol{r}_j(t) = \boldsymbol{f}(\boldsymbol{a} + t\boldsymbol{e}_j) - \boldsymbol{f}(\boldsymbol{a}) - D\boldsymbol{f}(\boldsymbol{a})(t\boldsymbol{e}_j).$$

Then $|\boldsymbol{r}_j(t)|/t \to 0$ as $t \to 0$, and by linearity of $D\boldsymbol{f}(\boldsymbol{a})$,

$$\frac{\boldsymbol{f}(\boldsymbol{a} + t\boldsymbol{e}_j) - \boldsymbol{f}(\boldsymbol{a})}{t} = D\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{e}_j) + \frac{\boldsymbol{r}_j(t)}{t}.$$

Taking the i-th component on both sides and letting $t \to 0$, we obtain

$$D_j f^i(\boldsymbol{a}) = u^i \circ D\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{e}_j).$$

The right side of this identity is the (i,j)-th entry of $[D\boldsymbol{f}(\boldsymbol{a})]$ by Definition 1.2.2 and Theorem 1.2.1. This establishes the claim. □

It is worth noting that the matrix of partial derivatives may exist without $\boldsymbol{f}$ being differentiable: Consider the function $f(x,y) = xy/(x^2+y^2)$ from Example 1.9.1, and define it to equal zero at the origin. Since $f(x,0) \equiv 0$, $D_1 f(0,0) = 0$, and similarly $D_2 f(0,0) = 0$. f, however, is not differentiable at $\mathbf{0}$. This follows from the fact that it has no limit at that point (and is in particular discontinuous there), together with:

Theorem 2.1.2. *If $\boldsymbol{f} : A \subset \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at an interior point $\boldsymbol{a}$ of A, then it continuous at that point.*

Proof. Define a map $\boldsymbol{r}$ from a neighborhood of $\mathbf{0} \in \mathbb{R}^n$ to $\mathbb{R}^m$ by

$$\boldsymbol{r}(\boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a}) - D\boldsymbol{f}(\boldsymbol{a})\boldsymbol{h}.$$

Then $|\boldsymbol{r}(\boldsymbol{h})|/|\boldsymbol{h}| \to 0$ as $\boldsymbol{h} \to \mathbf{0}$, and

$$|\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a})| = |D\boldsymbol{f}(\boldsymbol{a})\boldsymbol{h} + \boldsymbol{r}(\boldsymbol{h})| \le |D\boldsymbol{f}(\boldsymbol{a})\boldsymbol{h}| + |\boldsymbol{r}(\boldsymbol{h})| \le |D\boldsymbol{f}(\boldsymbol{a})||\boldsymbol{h}| + |\boldsymbol{r}(\boldsymbol{h})|.$$

Since the last term on the right approaches 0 as $\boldsymbol{h} \to \mathbf{0}$, $\boldsymbol{f}$ is continuous at $\boldsymbol{a}$. □

It turns out that if the partial derivatives exist and are continuous, then $\boldsymbol{f}$ is differentiable. Before arguing this, we introduce some terminology:

Definition 2.1.4. (1) $\boldsymbol{f} : A \subset \mathbb{R}^n \to \mathbb{R}^m$ is said to be *differentiable on A* if there exists an open set U containing A and a map $\boldsymbol{g} : U \to \mathbb{R}^m$ that is differentiable at every point of U and whose restriction $\boldsymbol{g}_{|A}$ to A equals $\boldsymbol{f}$.

(2) Suppose $\boldsymbol{f}$ is differentiable on an open set U. If for every $\boldsymbol{u} \in U$ and every $\varepsilon > 0$, there exists $\delta > 0$ such that $|D\boldsymbol{f}(\boldsymbol{p}) - D\boldsymbol{f}(\boldsymbol{u})| < \varepsilon$ whenever $\boldsymbol{p} \in U$ and $|\boldsymbol{p} - \boldsymbol{u}| < \delta$, then $\boldsymbol{f}$ is said to be *continuously differentiable* on U, and we write $\boldsymbol{f} \in \mathcal{C}^1(U)$.

For example, according to the first part of the above definition, the absolute value function $f : \mathbb{R} \to \mathbb{R}$, $f(t) = |t|$, is differentiable on $[0,\infty)$. The following fact sheds some light on the meaning of the second part:

Lemma 2.1.1. *Suppose $\boldsymbol{f}$ is differentiable on an open set U. Then $\boldsymbol{f}$ is continuously differentiable on U if and only if all partial derivatives $D_j f^i$ are continuous on U.*

Proof. The statement easily follows from the string of inequalities:

$$\begin{aligned} |D_j f^i(\boldsymbol{p}) - D_j f^i(\boldsymbol{u})| &\le |D\boldsymbol{f}(\boldsymbol{p}) - D\boldsymbol{f}(\boldsymbol{u})| \\ &\le \sum_i \left(\sum_j (D_j f^i(\boldsymbol{p}) - D_j f^i(\boldsymbol{u}))^2 \right)^{\frac{1}{2}}. \end{aligned}$$

The first inequality holds because

$$\begin{aligned}|D_j f^i(\boldsymbol{p}) - D_j f^i(\boldsymbol{u})| &= |\langle (D\boldsymbol{f}(\boldsymbol{p}) - D\boldsymbol{f}(\boldsymbol{u}))\boldsymbol{e}_j, \boldsymbol{e}_i\rangle| \le |(D\boldsymbol{f}(\boldsymbol{p}) - D\boldsymbol{f}(\boldsymbol{u}))\boldsymbol{e}_j| \\ &\le |D\boldsymbol{f}(\boldsymbol{p}) - D\boldsymbol{f}(\boldsymbol{u})|.\end{aligned}$$

For the second inequality, recall that if L is any linear transformation, then its norm $|L|$ satisfies $|L| \le \sum_i |L\boldsymbol{e}_i|$, cf. the paragraph following Definition 1.4.1. Thus,

$$\begin{aligned}|D\boldsymbol{f}(\boldsymbol{p}) - D\boldsymbol{f}(\boldsymbol{u})| &\le \sum_i |(D\boldsymbol{f}(\boldsymbol{p}) - D\boldsymbol{f}(\boldsymbol{u}))\boldsymbol{e}_i| \\ &= \sum_i \left(\sum_j \langle (D\boldsymbol{f}(\boldsymbol{p}) - D\boldsymbol{f}(\boldsymbol{u}))\boldsymbol{e}_i, \boldsymbol{e}_j\rangle^2 \right)^{\frac{1}{2}} \\ &= \sum_i \left(\sum_j (D_j f^i(\boldsymbol{p}) - D_j f^i(\boldsymbol{u}))^2 \right)^{\frac{1}{2}}.\end{aligned}$$

□

Next, we drop the assumption of differentiability in the lemma:

Theorem 2.1.3. *Let $\boldsymbol{f}$ map an open set $U \subset \mathbb{R}^n$ to $\mathbb{R}^m$. Then $\boldsymbol{f}$ is continuously differentiable on U if and only if the partial derivatives $D_j f^i$ exist and are continuous on U.*

Proof. In light of Lemma 2.1.1, it only remains to show that if the partial derivatives exist and are continuous on U, then $\boldsymbol{f}$ is differentiable on U. We begin by considering the case when $\boldsymbol{f}$ is a real-valued function f; i.e., when $m = 1$. So let $\boldsymbol{a} \in U$, $r > 0$ small enough that $B_r(\boldsymbol{a}) \subset U$. For $\boldsymbol{h} \in \mathbb{R}^n$ with $|\boldsymbol{h}| < r$, set $h_i = u^i(\boldsymbol{h})$, so that $\boldsymbol{h} = \sum h_i \boldsymbol{e}_i$. Define $\boldsymbol{u}_0 = \boldsymbol{0}$, and $\boldsymbol{u}_j = h_1\boldsymbol{e}_1 + \cdots + h_j\boldsymbol{e}_j$ for $j = 1, \ldots, n$. Then

$$f(\boldsymbol{a} + \boldsymbol{h}) - f(\boldsymbol{a}) = \sum_{i=1}^n (f(\boldsymbol{a} + \boldsymbol{u}_i) - f(\boldsymbol{a} + \boldsymbol{u}_{i-1})) = \sum_{i=1}^n (g_i(h_i) - g_i(0)),$$

where $g_i(t) = f(\boldsymbol{a} + \boldsymbol{u}_{i-1} + t\boldsymbol{e}_i)$. Notice that the line segment connecting $\boldsymbol{a} + \boldsymbol{u}_{i-1}$ with $\boldsymbol{a} + \boldsymbol{u}_i$ lies inside U, since its endpoints lie in $B_r(\boldsymbol{a})$ and the latter ball is convex. We may therefore apply the ordinary mean-value theorem to g_i, and conclude that there exists $t_i \in (0, h_i)$ such that

$$g_i(h_i) - g_i(0) = h_i g_i'(t_i) = h_i D_i f(\boldsymbol{a} + \boldsymbol{u}_{i-1} + t_i\,\boldsymbol{e}_i).$$

Let $L : \mathbb{R}^n \to \mathbb{R}$ denote the linear transformation whose matrix in the standard basis has the partial derivatives of f at $\boldsymbol{a}$ as entries. Then

$$\begin{aligned}\frac{|f(\boldsymbol{a} + \boldsymbol{h}) - f(\boldsymbol{a}) - L\boldsymbol{h}|}{|\boldsymbol{h}|} &\le \sum_{i=1}^n \frac{|D_i f(\boldsymbol{a} + \boldsymbol{u}_{i-1} + t_i\,\boldsymbol{e}_i) - D_i f(\boldsymbol{a})||h_i|}{|\boldsymbol{h}|} \\ &\le \sum_i |D_i f(\boldsymbol{a} + \boldsymbol{u}_{i-1} + t_i\,\boldsymbol{e}_i) - D_i f(\boldsymbol{a})|.\end{aligned}$$

Since the right side of the inequality approaches 0 as $\boldsymbol{h} \to \boldsymbol{0}$ by continuity of the partial derivatives, this proves the claim when $m = 1$. For general m, let L denote the linear

transformation whose matrix in the standard basis has $D_j f^i(\boldsymbol{a})$ as (i,j)-th entry. Then

$$\frac{|\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h})-\boldsymbol{f}(\boldsymbol{a})-L\boldsymbol{h}|}{|\boldsymbol{h}|} = \frac{\left(\sum_{i=1}^n (f^i(\boldsymbol{a}+\boldsymbol{h})-f^i(\boldsymbol{a})-Df^i(\boldsymbol{a})\boldsymbol{h})^2\right)^{\frac{1}{2}}}{|\boldsymbol{h}|}$$

approaches 0 as $\boldsymbol{h} \to \boldsymbol{0}$ by the one-dimensional case. □

Remark 2.1.1. Continuity of the partial derivatives of a function is not a necessary condition for the function to be differentiable. In the next section, it will be shown that sums and compositions of differentiable functions are again differentiable. Assuming this for now, consider first the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(t) = \begin{cases} t^2 \sin \frac{1}{t} & \text{if } t \neq 0, \\ 0 & \text{if } t = 0. \end{cases}$$

f is differentiable everywhere: $f'(0) = \lim_{h\to 0} h\sin(1/h) = 0$, since $0 \leq |h\sin(1/h)| \leq |h|$, and $f'(t) = 2t\sin(1/t) - \cos(1/t)$ if $t \neq 0$. However the limit of f' does not exist at 0. Consider now $g : \mathbb{R}^2 \to \mathbb{R}$, where $g = f \circ u^1 + f \circ u^2$. As a sum of compositions of differentiable functions, g is differentiable everywhere. Nevertheless, $D_1 g = f' \circ u^1$ and $D_2 g = f' \circ u^2$ are not continuous at the origin.

2.2 Basic properties of the derivative

Many of the derivation techniques for real-valued functions of one variable carry over to maps defined on Euclidean spaces. We begin with the following:

Theorem 2.2.1. *Let U be open in $\mathbb{R}^n$, $c \in \mathbb{R}$.*

- *If $\boldsymbol{f}, \boldsymbol{g} : U \to \mathbb{R}^m$ are differentiable at $\boldsymbol{a} \in U$, then so are $\boldsymbol{f}+\boldsymbol{g}$, $c\boldsymbol{f}$, and*

$$D(\boldsymbol{f}+\boldsymbol{g})(\boldsymbol{a}) = D(\boldsymbol{f}(\boldsymbol{a})) + D(\boldsymbol{g}(\boldsymbol{a})), \qquad D(c\boldsymbol{f})(\boldsymbol{a}) = cD(\boldsymbol{f})(\boldsymbol{a}).$$

- *If $f, g : U \to \mathbb{R}$ are differentiable at $\boldsymbol{a} \in U$, then so are $f \cdot g$, f/g (the latter provided $g(\boldsymbol{a}) \neq 0$), and*

$$D(f\cdot g)(\boldsymbol{a}) = f(\boldsymbol{a})Dg(\boldsymbol{a}) + g(\boldsymbol{a})Df(\boldsymbol{a}),$$
$$D(f/g)(\boldsymbol{a}) = \frac{g(\boldsymbol{a})Df(\boldsymbol{a}) - f(\boldsymbol{a})Dg(\boldsymbol{a})}{g^2(\boldsymbol{a})}.$$

Proof. Notice that if the maps are known to be continuously differentiable, then the statements follow immediately from the analogous ones for functions of one variable, since by Theorem 2.1.3, it suffices to look at partial derivatives, which, as noted before, are ordinary derivatives of functions of one variable. We will prove the first statement in each item, and leave the others to the reader.

For $\boldsymbol{f}+\boldsymbol{g}$, it must be shown that

$$\lim_{\boldsymbol{h}\to 0}\frac{|(\boldsymbol{f}+\boldsymbol{g})(\boldsymbol{a}+\boldsymbol{h})-(\boldsymbol{f}+\boldsymbol{g})(\boldsymbol{a})-(D(\boldsymbol{f}(\boldsymbol{a}))+D(\boldsymbol{g}(\boldsymbol{a})))\boldsymbol{h}|}{|\boldsymbol{h}|}=0.$$

But this follows from the fact that the expression inside the limit is nonnegative and bounded above by

$$\frac{|\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h})-\boldsymbol{f}(\boldsymbol{a})-D\boldsymbol{f}(\boldsymbol{a})\boldsymbol{h}|}{|\boldsymbol{h}|}+\frac{|\boldsymbol{g}(\boldsymbol{a}+\boldsymbol{h})-\boldsymbol{g}(\boldsymbol{a})-D\boldsymbol{g}(\boldsymbol{a})\boldsymbol{h}|}{|\boldsymbol{h}|},$$

which goes to 0 as $\boldsymbol{h}\to\boldsymbol{0}$.

For $f\cdot g$, we must show that

$$\lim_{\boldsymbol{h}\to 0}\frac{|fg(\boldsymbol{a}+\boldsymbol{h})-fg(\boldsymbol{a})-(f(\boldsymbol{a})Dg(\boldsymbol{a})+g(\boldsymbol{a})Df(\boldsymbol{a}))\boldsymbol{h}|}{|\boldsymbol{h}|}=0. \tag{2.2.1}$$

Let

$$r_f(\boldsymbol{h})=f(\boldsymbol{a}+\boldsymbol{h})-f(\boldsymbol{a})-Df(\boldsymbol{a})\boldsymbol{h},\quad r_g(\boldsymbol{h})=g(\boldsymbol{a}+\boldsymbol{h})-g(\boldsymbol{a})-Dg(\boldsymbol{a})\boldsymbol{h}.$$

By assumption, $|r_f(\boldsymbol{h})|/|\boldsymbol{h}|, |r_g(\boldsymbol{h})|/|\boldsymbol{h}|\to 0$ as $\boldsymbol{h}\to\boldsymbol{0}$. The expression inside the limit in (2.2.1) is nonnegative and can be rewritten

$$\frac{1}{|\boldsymbol{h}|}|f(\boldsymbol{a}+\boldsymbol{h})r_g(\boldsymbol{h})+(f(\boldsymbol{a}+\boldsymbol{h})-f(\boldsymbol{a}))Dg(\boldsymbol{a})\boldsymbol{h}+g(\boldsymbol{a})r_f(\boldsymbol{h})|,$$

which is no larger than

$$|f(\boldsymbol{a}+\boldsymbol{h})|\frac{|r_g(\boldsymbol{h})|}{|\boldsymbol{h}|}+|f(\boldsymbol{a}+\boldsymbol{h})-f(\boldsymbol{a})||Dg(\boldsymbol{a})|+|g(\boldsymbol{a})|\frac{|r_f(\boldsymbol{h})|}{|\boldsymbol{h}|}.$$

The first and last term in this sum go to 0 as $\boldsymbol{h}\to\boldsymbol{0}$ as noted earlier, and the middle term also by continuity of f at $\boldsymbol{a}$ (notice that we also need continuity at $\boldsymbol{a}$ to ensure that the first term goes to 0). This establishes (2.2.1). □

Theorem 2.2.2 (The chain rule). *Let $\boldsymbol{f}: U\to\mathbb{R}^m$, where U is an open subset of $\mathbb{R}^n$, be differentiable at $\boldsymbol{a}\in U$. Let V be a neighborhood of $\boldsymbol{f}(\boldsymbol{a})$ in $\mathbb{R}^m$, and suppose $\boldsymbol{g}: V\to\mathbb{R}^k$ is differentiable at $\boldsymbol{f}(\boldsymbol{a})$. Then $\boldsymbol{g}\circ\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$, and*

$$D(\boldsymbol{g}\circ\boldsymbol{f})(\boldsymbol{a})=D\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a}))\circ D\boldsymbol{f}(\boldsymbol{a})).$$

Proof. Set $T=D\boldsymbol{f}(\boldsymbol{a})$, $S=D\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a}))$. Begin by rewriting

$$\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}))-\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a}))-(S\circ T)\boldsymbol{h}$$

as

$$S(\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h})-\boldsymbol{f}(\boldsymbol{a})-T\boldsymbol{h})+\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a})+\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h})-\boldsymbol{f}(\boldsymbol{a}))-\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a}))-S(\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h})-\boldsymbol{f}(\boldsymbol{a})), \tag{2.2.2}$$

and let

$$r_f(\boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a}) - T\boldsymbol{h},$$
$$r_g(\boldsymbol{h}) = \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a})+\boldsymbol{h}) - \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a})) - S\boldsymbol{h},$$
$$\rho(\boldsymbol{h}) = \frac{|r_g(\boldsymbol{h})|}{|\boldsymbol{h}|} \text{ if } \boldsymbol{h} \neq 0, \quad \rho(\boldsymbol{0}) = 0.$$

By assumption, $|r_f(\boldsymbol{h})|/|\boldsymbol{h}|, \rho(\boldsymbol{h}) \to 0$ as $\boldsymbol{h} \to \boldsymbol{0}$. Finally, let

$$\boldsymbol{k}(\boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a}) = r_f(\boldsymbol{h}) + T\boldsymbol{h},$$

which also goes to 0 as $\boldsymbol{h} \to \boldsymbol{0}$. Now, by (2.2.2),

$$\frac{|\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h})) - \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a})) - ST\boldsymbol{h}|}{|\boldsymbol{h}|} = \frac{|S(r_f(\boldsymbol{h})) + r_g(\boldsymbol{k}(\boldsymbol{h}))|}{|\boldsymbol{h}|}$$
$$\leq \frac{|S(r_f(\boldsymbol{h}))|}{|\boldsymbol{h}|} + \frac{|r_g(\boldsymbol{k}(\boldsymbol{h}))|}{|\boldsymbol{h}|},$$

so it suffices to check that each of the last two terms goes to 0 as $\boldsymbol{h} \to \boldsymbol{0}$. But

$$\frac{|S(r_f(\boldsymbol{h}))|}{|\boldsymbol{h}|} \leq |S|\frac{|r_f(\boldsymbol{h})|}{|\boldsymbol{h}|} \to 0 \text{ as } \boldsymbol{h} \to \boldsymbol{0},$$

whereas

$$\frac{|r_g(\boldsymbol{k}(\boldsymbol{h}))|}{|\boldsymbol{h}|} = \rho(\boldsymbol{k}(\boldsymbol{h}))\frac{|\boldsymbol{k}(\boldsymbol{h})|}{|\boldsymbol{h}|} = \rho(\boldsymbol{k}(\boldsymbol{h}))\frac{|r_f(\boldsymbol{h}) + T\boldsymbol{h}|}{|\boldsymbol{h}|}$$
$$\leq \rho(\boldsymbol{k}(\boldsymbol{h}))\left(\frac{r_f(\boldsymbol{h})}{|\boldsymbol{h}|} + |T|\right) \to 0 \text{ as } \boldsymbol{h} \to \boldsymbol{0}. \qquad \square$$

Recall from Examples 2.1.1 (i) that the derivative of a linear transformation is the transformation itself. We now generalize this result.

Definition 2.2.1. Let V, V_i be vector spaces, $1 \leq i \leq k$. A map $M : V_1 \times \cdots \times V_k \to V$ is said to be *multilinear* if for every $i = 1, \ldots, k$, and any choice of $\boldsymbol{v}_j \in V_j$, $j \neq i$, the map

$$V_i \to V,$$
$$\boldsymbol{v} \mapsto M(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}, \boldsymbol{v}, \boldsymbol{v}_{i+1}, \ldots, \boldsymbol{v}_k)$$

is linear.

Thus, the map obtained by fixing all but one arbitrary variable is linear.

Theorem 2.2.3. *If $M : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k} \to \mathbb{R}^m$ is multilinear, then M is continuously differentiable on its domain, and*

$$DM(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k)(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k) = \sum_{i=1}^{k} M(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \boldsymbol{b}_i, \boldsymbol{a}_{i+1}, \ldots, \boldsymbol{a}_k).$$

Proof. For each $l = 1, \dots, k$, define

$$\begin{aligned} \imath_l : \mathbb{R}^{n_l} &\to \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}, \\ \boldsymbol{v} &\mapsto (\boldsymbol{a}_1, \dots, \boldsymbol{a}_{l-1}, \boldsymbol{v}, \boldsymbol{a}_{l+1}, \dots, \boldsymbol{a}_k). \end{aligned}$$

We may assume without loss of generality that $m = 1$. If $i \le n_l$, then the $(n_1 + \cdots + n_{l-1} + i)$-th partial derivative of M at $(\boldsymbol{a}_1, \dots, \boldsymbol{a}_k)$ equals the i-th partial derivative of $M \circ \imath_l$ at $\boldsymbol{a}_l$. But $M \circ \imath_l$ is linear, so this derivative equals the i-th entry of (the matrix of) $M \circ \imath_l$. In particular, it is continuous, and M is continuously differentiable by Theorem 2.1.3. Furthermore,

$$DM(\boldsymbol{a}_1, \dots, \boldsymbol{a}_k) = (M \circ \imath_1, \dots, M \circ \imath_k),$$

and

$$\begin{aligned} DM(\boldsymbol{a}_1, \dots, \boldsymbol{a}_k)(\boldsymbol{b}_1, \dots, \boldsymbol{b}_k) &= \sum_{l=1}^{k} (M \circ \imath_l)\boldsymbol{b}_l \\ &= \sum_{l=1}^{k} M(\boldsymbol{a}_1, \dots, \boldsymbol{a}_{l-1}, \boldsymbol{b}_l, \boldsymbol{a}_{l+1} \dots, \boldsymbol{a}_k). \end{aligned}$$

□

Corollary 2.2.1. *The derivative of the inner product map* $\langle , \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *is given by*

$$D\langle , \rangle(\boldsymbol{a}_1, \boldsymbol{a}_2)(\boldsymbol{b}_1, \boldsymbol{b}_2) = \langle \boldsymbol{a}_1, \boldsymbol{b}_2 \rangle + \langle \boldsymbol{b}_1, \boldsymbol{a}_2 \rangle.$$

In particular, if $\boldsymbol{c}_i : I \to \mathbb{R}^n$, $i = 1, 2$, *are differentiable curves defined on a common interval* I, *then the real-valued function* $\langle \boldsymbol{c}_1, \boldsymbol{c}_2 \rangle$ *has derivative*

$$\langle \boldsymbol{c}_1, \boldsymbol{c}_2 \rangle'(t) = \langle \boldsymbol{c}_1'(t), \boldsymbol{c}_2(t) \rangle + \langle \boldsymbol{c}_1(t), \boldsymbol{c}_2'(t) \rangle.$$

Proof. The first identity is an immediate consequence of Theorem 2.2.3. The second one follows from the first together with the chain rule: if $\boldsymbol{c}$ is the curve $(\boldsymbol{c}_1, \boldsymbol{c}_2)$ in $\mathbb{R}^{2n}$, then $\langle \boldsymbol{c}_1, \boldsymbol{c}_2 \rangle = \langle , \rangle \circ \boldsymbol{c}$, so that

$$\langle \boldsymbol{c}_1, \boldsymbol{c}_2 \rangle'(t) = D\langle , \rangle(\boldsymbol{c}_1(t), \boldsymbol{c}_2(t))(\boldsymbol{c}_1'(t), \boldsymbol{c}_2'(t) = \langle \boldsymbol{c}_1'(t), \boldsymbol{c}_2(t) \rangle + \langle \boldsymbol{c}_1(t), \boldsymbol{c}_2'(t) \rangle.$$

□

If $f : U \subset \mathbb{R}^n \to \mathbb{R}$ is differentiable on an open set U, its partial derivatives $D_i f$ are again real-valued functions on U. When the $D_i f$ are themselves differentiable, the *second partial derivatives* of f are defined by

$$D_{ij} f = D_j(D_i f), \qquad 1 \le i, j \le n.$$

This process can of course be iterated. We say f *is of class* $\mathcal{C}^k$, and write $f \in \mathcal{C}^k(U)$, if all partial derivatives of order k of f exist and are continuous on U. If this holds for any k, we say f is *smooth*, and write $f \in \mathcal{C}^\infty(U)$. Our next theorem, which is usually interpreted as saying that if f is of class $\mathcal{C}^2$, then $D_{ij} f = D_{ji} f$, can be stated somewhat more generally:

Theorem 2.2.4. *Suppose f is a real-valued function defined on an open set $U \subset \mathbb{R}^n$. If f, $D_i f$, $D_j f$, and $D_{ij} f$ exist and are continuous on U, then $D_{ji} f$ exists on U, and equals $D_{ij} f$.*

Proof. We may, without loss of generality, assume that $n = 2$: for if $\boldsymbol{a} = (a_1, \dots, a_n) \in U$ and $i < j$, define

$$\begin{aligned} \iota &: \mathbb{R}^2 \to \mathbb{R}^n, \\ &(x, y) \mapsto (a_1, \dots, a_{i-1}, x, a_{i+1}, \dots, a_{j-1}, y, a_{j+1}, \dots, a_n). \end{aligned}$$

Then $D_i f(\boldsymbol{a}) = D_1(f \circ \iota)(a_i, a_j)$, and $D_j(f \circ \iota)(\boldsymbol{a}) = D_2 f(a_i, a_j)$. So let $(a, b) \in U$, and define a real-valued function φ on a neighborhood of $\mathbf{0} \in \mathbb{R}^2$ by

$$\varphi(s, t) = [f(a + s, b + t) - f(a + s, b)] - [f(a, b + t) - f(a, b)].$$

Notice that if $g(x) = f(x, b + t) - f(x, b)$, then $\varphi(s, t) = g(a + s) - g(a)$. Applying the mean value theorem to g on $[a, a + s]$, there is an $\varepsilon_1 \in (0, 1)$ such that

$$\varphi(s, t) = sg'(a + \varepsilon_1 s) = s[D_1 f(a + \varepsilon_1 s, b + t) - D_1 f(a + \varepsilon_1 s, b)].$$

Next, apply the mean value theorem to $y \mapsto D_1 f(a + \varepsilon_1 s, y)$ on $[b, b + t]$ to conclude that

$$\varphi(s, t) = stD_{12} f(a + \varepsilon_1 s, b + \varepsilon_2 t)$$

for some $0 < \varepsilon_1, \varepsilon_2 < 1$. Thus, if $t \neq 0$,

$$\frac{f(a + s, b + t) - f(a + s, b)}{t} - \frac{f(a, b + t) - f(a, b)}{t} = sD_{12} f(a + \varepsilon_1 s, b + \varepsilon_2 t).$$

Taking the limit as $t \to 0$ and using continuity of $D_{12} f$, we obtain

$$D_2 f(a + s, b) - D_2 f(a, b) = sD_{12} f(a + \varepsilon_1 s, b).$$

Dividing this equation by s and letting $s \to 0$ yields $D_{21} f(a, b) = D_{12} f(a, b)$, again by continuity of $D_{12} f$. This establishes the claim. □

We end this section with the construction of a "bump" function that will be needed in later chapters. Given $r > 0$, denote by C_r the open "cube" $(-r, r)^n$ in $\mathbb{R}^n$.

Lemma 2.2.1. *For any $0 < r < R$, there exists a differentiable function $\varphi : \mathbb{R}^n \to \mathbb{R}$ with the following properties:*
(1) $\varphi \equiv 1$ *on* $\bar{C}_r$;
(2) $0 < \varphi < 1$ *on* $C_R - \bar{C}_r$, *and*
(3) $\varphi \equiv 0$ *on* $\mathbb{R}^n \setminus C_R$.

Proof. Let $h : \mathbb{R} \to \mathbb{R}$ be given by

$$h(x) = \begin{cases} e^{-1/x}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

h is C^∞ by l'Hospital's rule. Define

$$f(x) = \frac{h(R+x)h(R-x)}{h(R+x)h(R-x) + h(x-r) + h(-x-r)}.$$

This expression makes sense because $h(x-r) + h(-x-r)$ is nonnegative, and equals 0 only when $|x| \le r$, in which case $h(R+x)h(R-x) > 0$. Furthermore, $f(x) = 1$ if $|x| \le r$, $0 < f(x) < 1$ if $r < |x| < R$, and $f(x) = 0$ if $|x| \ge R$. Now let $\varphi(a_1, \dots, a_n) = \Pi_{i=1}^n f(a_i)$. □

The *support* $\operatorname{supp}\varphi$ of a function $\varphi : U \subset \mathbb{R}^n \to \mathbb{R}$ is defined to be the closure of the set $\{\boldsymbol{p} \in U \mid \varphi(\boldsymbol{p}) \neq 0\}$. The function φ from the Lemma has its support in the closure of C_R, but since R is arbitrary, it is also true that there exists a function satisfying (1) and (2), but with support in C_R itself.

Theorem 2.2.5. *Given any open set U in $\mathbb{R}^n$ and any compact subset K of U, there exists a differentiable function $\varphi : \mathbb{R}^n \to \mathbb{R}$ such that*
(1) $0 \le \varphi \le 1$;
(2) $\varphi \equiv 1$ *on* K, *and*
(3) $\operatorname{supp}\varphi \subset U$.

Proof. For each $\boldsymbol{a} \in K$, choose $R(\boldsymbol{a}) > 0$ such that the cube $C_R(\boldsymbol{a}) := \prod_{i=1}^n (a_i - R, a_i + R)$ of radius R centered at $\boldsymbol{a}$ is contained in U. Since the collection of all cubes of the form $C_{R/2}(\boldsymbol{a})$, with $\boldsymbol{a} \in K$, covers K, there exist $\boldsymbol{a}_1, \dots, \boldsymbol{a}_k \in K$ and $R_1, \dots, R_k > 0$ such that

$$K \subset \bigcup_{i=1}^k C_{R_i/2}(\boldsymbol{a}_i).$$

Lemma 2.2.1 guarantees for any $R > 0$ the existence of a differentiable function φ_R on $\mathbb{R}^n$ that equals 1 on $C_{R/2}$, has support in C_R, and takes values between 0 and 1. For each i between 1 and k, define a function φ_i by

$$\varphi_i(\boldsymbol{p}) = \varphi_{R_i}(\boldsymbol{p} - \boldsymbol{a}_i).$$

Then φ_i is identically 1 on $C_{R_i/2}(\boldsymbol{a}_i)$, has support in $C_R(\boldsymbol{a}_i) \subset U$, and takes values between 0 and 1. Thus,

$$\varphi := \frac{1}{k}\sum_{i=1}^k \varphi_i$$

satisfies the conditions of the theorem. □

We will later see that the compactness assumption on K may be dropped.

2.3 Differentiation of integrals

The sole purpose of this section is to establish a result that will be needed in a later chapter: suppose $f : [a,b] \times [c,d] \to \mathbb{R}$ is a continuous function. Integrating f with respect to the first variable x over $[a,b]$ defines a function φ of y only, $\varphi(y) = \int_a^b f(x,y)\,dx$,

$c \le y \le d$. We wish to determine when φ is differentiable, and what its derivative $\frac{d}{dy}\int_a^b f(x,y)\,dx$ is. It turns out that, loosely speaking,

$$\frac{d}{dy}\left(\int_a^b f(x,y)\,dx\right) = \int_a^b \frac{\partial f}{\partial y}(x,y)\,dx$$

whenever the partial derivative $\partial f/\partial y = D_2 f$ is continuous. The latter condition is actually sufficient but not necessary. Nevertheless, we will adopt it since it is not too restrictive for our purposes.

Theorem 2.3.1. *If f and $D_2 f$ are continuous on $[a,b]\times[c,d]$, then the function*

$$\varphi : [c,d] \to \mathbb{R},$$
$$y \mapsto \int_a^b f(x,y)\,dx$$

is differentiable on (c,d), and

$$\varphi'(y_0) = \int_a^b D_2 f(x,y_0)\,dx, \qquad y_0 \in (c,d).$$

Proof. Fix $y_0 \in (c,d)$, and set

$$g(x,y) = \frac{f(x,y) - f(x,y_0)}{y - y_0}.$$

Then

$$\frac{\varphi(y) - \varphi(y_0)}{y - y_0} = \int_a^b g(x,y)\,dx,$$

so that the theorem will follow once we establish that

$$\lim_{y\to y_0}\int_a^b g(x,y)\,dx = \int_a^b D_2 f(x,y_0)\,dx. \tag{2.3.1}$$

To prove this, let $\varepsilon > 0$. Using the fact that $D_2 f$ is continuous, for each $x_0 \in [a,b]$ there exists a $\delta_{x_0} > 0$ such that

$$|D_2 f(x_0,y) - D_2 f(x_0,y_0)| < \frac{1}{2}\frac{\varepsilon}{b-a}, \text{ provided } |y - y_0| < \delta_{x_0}.$$

Again by continuity, there exists, for each $x_0 \in [a,b]$, a neighborhood U_{x_0} of x_0 such that

$$|D_2 f(x,y) - D_2 f(x,y_0)| < \frac{\varepsilon}{b-a}, \qquad x \in U_{x_0}, \quad |y - y_0| < \delta_{x_0}.$$

Next, choose some finite subcover $\{U_{x_1}, \dots, U_{x_k}\}$ of the cover $\{U_x \mid x \in [a,b]\}$ of $[a,b]$, and set $\delta = \min\{\delta_{x_1}, \dots, \delta_{x_k}\}$. it follows that

$$|D_2 f(x,y) - D_2 f(x,y_0)| < \frac{\varepsilon}{b-a}, \qquad x \in [a,b], \quad |y - y_0| < \delta\,. \tag{2.3.2}$$

By the mean value theorem, for any $(x,y) \in [a,b]\times[c,d]$, there is some s between y and y_0 such that $g(x,y) = D_2 f(x,s)$. (2.3.2) then implies

$$|g(x,y) - D_2 f(x,y_0)| < \frac{\varepsilon}{b-a}, \qquad x \in [a,b], \quad |y - y_0| < \delta\,,$$

and therefore

$$\left| \int_a^b g(x,y)\,\mathrm{d}x - \int_a^b D_2 f(x,y_0)\,\mathrm{d}x \right| < \varepsilon$$

whenever $|y - y_0| < \delta$. This proves (2.3.1). □

2.4 Curves

In Examples 2.1.1 (ii), we introduced curves as continuous maps from an interval to Euclidean space. In geometry, however, many concepts require smoothness. We will therefore restrict ourselves to differentiable curves, or more precisely to the following somewhat larger class, that we still simply call curves for convenience:

Definition 2.4.1. A map $\boldsymbol{c} : [a,b] \to \mathbb{R}^n$ is said to be a *curve* if it is piecewise smooth; i.e., if there exists a partition $P : t_0 = a < t_1 < \cdots < t_k = b$ of $[a,b]$ such that the restriction of $\boldsymbol{c}$ to each subinterval (t_{i-1}, t_i) is smooth.

If $c^i = u^i \circ \boldsymbol{c}$, so that $\boldsymbol{c} = \begin{bmatrix} c^1 & \cdots & c^n \end{bmatrix}^T$, we define

$$\int_a^b \boldsymbol{c} := \begin{bmatrix} \int_a^b c^1 & \cdots & \int_a^b c^n \end{bmatrix}^T.$$

Thus, by the fundamental theorem of Calculus,

$$\int_a^b \boldsymbol{c}' = \sum_{i=1}^n \int_{t_{i-1}}^{t_i} \boldsymbol{c}' = \sum_{i=1}^n \boldsymbol{c}(t_i) - \boldsymbol{c}(t_{i-1}) = \boldsymbol{c}(b) - \boldsymbol{c}(a).$$

Lemma 2.4.1. *If $\boldsymbol{c} : [a,b] \to \mathbb{R}^n$ is a curve, then*

$$\left| \int_a^b \boldsymbol{c}' \right| \le \int_a^b |\boldsymbol{c}'|.$$

Proof. Define $\alpha_i = \int_a^b c^{i\prime}$, $i = 1, \ldots, n$. Then

$$|\int_a^b \boldsymbol{c}'|^2 = \sum_i (\int_a^b c^{i\prime})^2 = \sum \alpha_i \int_a^b c^{i\prime} = \int_a^b \sum \alpha_i c^{i\prime}.$$

By the Cauchy-Schwarz inequality,

$$\sum \alpha_i c^{i\prime} \leq (\sum \alpha_i^2)^{1/2} (\sum c^{i\prime 2})^{1/2} = (|\int_a^b \boldsymbol{c}'|)|\boldsymbol{c}'|.$$

Thus,

$$|\int_a^b \boldsymbol{c}'|^2 \leq |\int_a^b \boldsymbol{c}'| \cdot \int_a^b |\boldsymbol{c}'|,$$

and the claim now follows. □

Our next goal is to define the length of a piecewise smooth curve $\boldsymbol{c} : [a, b] \to \mathbb{R}^n$. To this end, associate to each partition (not necessarily the one from Definition 2.4.1) $P : t_0 = a < t_1 < \cdots < t_k = b$ of $[a, b]$ the number

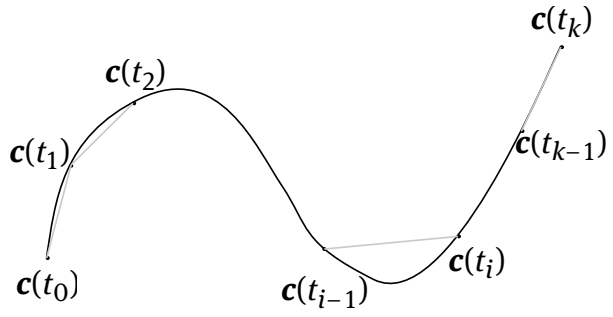
$$\ell(P, \boldsymbol{c}) = \sum_{i=1}^{k} |\boldsymbol{c}(t_i) - \boldsymbol{c}(t_{i-1})|,$$

which represents the sum of the distances between consecutive points $\boldsymbol{c}(t_0)$, $\boldsymbol{c}(t_1), \ldots, \boldsymbol{c}(t_n)$. This number should be no larger than the length of the curve, and approaches it as the partition becomes finer, thereby motivating the following:

Definition 2.4.2. The *length* of a curve $c : [a, b] \to \mathbb{R}^n$ is

$$\ell(\boldsymbol{c}) = \sup\{\ell(P, \boldsymbol{c}) \mid P \text{ is a partition of } [a, b]\},$$

provided this number exists, and ∞ otherwise.



Theorem 2.4.1. *Any curve $\boldsymbol{c} : [a, b] \to \mathbb{R}^n$ has finite length $\ell(\boldsymbol{c}) = \int_a^b |\boldsymbol{c}'|$.*

Proof. Assume first that $\boldsymbol{c}$ is smooth. The result will follow once we establish that (1) $\ell(P,\boldsymbol{c}) \le \int_a^b |\boldsymbol{c}'|$ for any partition P of $[a,b]$, and (2) given any $\varepsilon > 0$, there exists a partition P such that $\int_a^b |\boldsymbol{c}'| \le \ell(P,\boldsymbol{c}) + \varepsilon$.
For the first assertion, consider a partition $P : a = t_0 < t_1 < \cdots < t_k = b$. Then by Lemma 2.4.1,

$$\ell(P,\boldsymbol{c}) = \sum_i |\boldsymbol{c}(t_i) - \boldsymbol{c}(t_{i-1})| = \sum \Big| \int_{t_{i-1}}^{t_i} \boldsymbol{c}' \Big| \le \sum \int_{t_{i-1}}^{t_i} |\boldsymbol{c}'| = \int_a^b |\boldsymbol{c}'|.$$

For the second assertion, observe that $\boldsymbol{c}'$ is uniformly continuous by compactness of $[a,b]$. Given $\varepsilon > 0$, choose $\delta > 0$ so that $|\boldsymbol{c}'(x) - \boldsymbol{c}'(y)| < \varepsilon/2(b-a)$ whenever $|x-y| < \delta$, and consider any partition $P : a = t_0 < t_1 < \cdots < t_k = b$ with $t_i - t_{i-1} < \delta$ for all i. Then

$$|\boldsymbol{c}'(t) - \boldsymbol{c}'(t_i)| < \frac{\varepsilon}{2(b-a)} \text{ and } |\boldsymbol{c}'(t)| < |\boldsymbol{c}'(t_i)| + \frac{\varepsilon}{2(b-a)}$$

for all $t \in [t_{i-1}, t_i]$. It follows that

$$\begin{aligned}
\int_{t_{i-1}}^{t_i} |\boldsymbol{c}'| &< \left(|\boldsymbol{c}'(t_i)| + \frac{\varepsilon}{2(b-a)} \right) (t_i - t_{i-1}) \\
&= \left| \int_{t_{i-1}}^{t_i} \boldsymbol{c}'(t_i) - \boldsymbol{c}' + \boldsymbol{c}' \right| + \frac{\varepsilon}{2(b-a)} (t_i - t_{i-1}) \\
&\le \left| \int_{t_{i-1}}^{t_i} \boldsymbol{c}'(t_i) - \boldsymbol{c}' \right| + \left| \int_{t_{i-1}}^{t_i} \boldsymbol{c}' \right| + \frac{\varepsilon}{2(b-a)} (t_i - t_{i-1}) \\
&\le \left(\int_{t_{i-1}}^{t_i} |\boldsymbol{c}'(t_i) - \boldsymbol{c}'| \right) + |\boldsymbol{c}(t_i) - \boldsymbol{c}(t_{i-1})| + \frac{\varepsilon}{2(b-a)} (t_i - t_{i-1}) \\
&\le |\boldsymbol{c}(t_i) - \boldsymbol{c}(t_{i-1})| + \frac{\varepsilon}{(b-a)} (t_i - t_{i-1}).
\end{aligned}$$

Adding these inequalities for each i then yields the second claim.

If $\boldsymbol{c}$ is only piecewise smooth, then the above argument holds over every sub-interval on which $\boldsymbol{c}$ is smooth, and the result follows from the next lemma. □

Lemma 2.4.2. *If $\boldsymbol{c} : [a,b] \to \mathbb{R}^n$ is a curve, then for any $t_0 \in [a,b]$, $\ell(\boldsymbol{c}) = \ell(\boldsymbol{c}|_{[a,t_0]}) + \ell(\boldsymbol{c}|_{[t_0,b]})$.*

Proof. If P_1 is a partition of the first sub-interval, and P_2 one of the second, then $P := P_1 \cup P_2$ is a partition of $[a,b]$, so that

$$\ell(P_1, \boldsymbol{c}|_{[a,t_0]}) + \ell(P_2, \boldsymbol{c}|_{[t_0,b]}) = \ell(P,\boldsymbol{c}) \le \ell(\boldsymbol{c}).$$

Since this is true for any partitions, $\ell(\boldsymbol{c}) \geq \ell(\boldsymbol{c}|_{[a,t_0]}) + \ell(\boldsymbol{c}|_{[t_0,b]})$. On the other hand, if P is a partition of $[a,b]$, then the partition P' obtained by adding t_0 to P is finer than P, and is a union of a partition P_1 of $[a,t_0]$ and a partition P_2 of $[t_0,b]$. Hence

$$\ell(P,\boldsymbol{c}) \leq \ell(P',\boldsymbol{c}) = \ell(P_1,\boldsymbol{c}|_{[a,t_0]}) + \ell(P_2,\boldsymbol{c}|_{[t_0,b]}) \leq \ell(\boldsymbol{c}|_{[a,t_0]}) + \ell(\boldsymbol{c}|_{[t_0,b]}),$$

and since P was arbitrary, $\ell(\boldsymbol{c}) \leq \ell(\boldsymbol{c}|_{[a,t_0]}) + \ell(\boldsymbol{c}|_{[t_0,b]})$. This establishes the result. □

Definition 2.4.3. Let I and $\tilde{I}$ denote compact intervals in $\mathbb{R}$. A curve $\tilde{\boldsymbol{c}} : \tilde{I} \to \mathbb{R}^n$ is called a *reparametrization* of the curve $\boldsymbol{c} : I \to \mathbb{R}^n$ if $\tilde{\boldsymbol{c}} = \boldsymbol{c} \circ f$, where $f : \tilde{I} \to I$ is a strictly monotone (i.e., either increasing or decreasing) bijection.

The reader can easily verify that the relation $\boldsymbol{c} \sim \tilde{\boldsymbol{c}}$ if $\tilde{\boldsymbol{c}}$ is a reparametrization of $\boldsymbol{c}$ is an equivalence relation: i.e., $\boldsymbol{c} \sim \boldsymbol{c}$ for any $\boldsymbol{c}$, $\boldsymbol{c} \sim \tilde{\boldsymbol{c}}$ implies $\tilde{\boldsymbol{c}} \sim \boldsymbol{c}$, and if $\boldsymbol{c}_1 \sim \boldsymbol{c}_2$, $\boldsymbol{c}_2 \sim \boldsymbol{c}_3$, then $\boldsymbol{c}_1 \sim \boldsymbol{c}_3$. This divides the collection of all such curves into disjoint subsets called equivalence classes, with two curves belonging to the same equivalence class if and only if one is a reparametrization of the other. Two curves in the same class have the same image, and in particular the same length. It turns out that any piecewise-smooth curve admits a smooth reparametrization. In order to show this, we will need the following multi-purpose lemma:

Lemma 2.4.3. *Given $a < b$, there exists a smooth function $f : \mathbb{R} \to \mathbb{R}$ such that:*
(1) *$f(t) = 0$ if $t \leq a$, $0 < f(t) < 1$ if $t \in (a,b)$, and $f(t) = 1$ if $t \geq b$;*
(2) *f is strictly increasing on (a,b).*

Proof. The result is actually an easy consequence of Theorem 2.2.5. Alternatively, define $h : \mathbb{R} \to \mathbb{R}$ by

$$h(t) = \begin{cases} e^{-\frac{1}{t-a}+\frac{1}{t-b}} & \text{if } t \in (a,b), \\ 0 & \text{otherwise.} \end{cases}$$

By l'Hospital's rule, the left and right derivatives of h of any order exist and equal 0 at a and b. Thus, h is differentiable on $\mathbb{R}$ and strictly positive on (a,b). Now let f be given by:

$$f(t) = \frac{\int_a^t h}{\int_a^b h}.$$

□

Proposition 2.4.1. *Any curve $\boldsymbol{c} : [a,b] \to \mathbb{R}^n$ admits a smooth reparametrization $\tilde{\boldsymbol{c}} : [a,b] \to \mathbb{R}^n$.*

Proof. By assumption, there exist numbers $a = t_0 < t_1 < \cdots < t_k = b$ such that the restriction of $\boldsymbol{c}$ to each (t_{i-1},t_i), $i = 1,\ldots,k$, is smooth. By Lemma 2.4.3, there exist functions $f_i : [a,b] \to \mathbb{R}$ such that $f_i(t) = 0$ if $t \leq t_{i-1}$, $f_i(t) = 1$ if $t \geq t_i$, and f_i is strictly increasing on $[t_{i-1},t_i]$. Define a function f on $[a,b]$ by

$$f(t) = t_0 + \sum_{i=1}^{k} (t_i - t_{i-1}) f_i(t).$$

f is a differentiable, strictly increasing function with $f(t_i) = t_i$ for each $i = 1, \dots, k$. The curve $\tilde{\boldsymbol{c}} := \boldsymbol{c} \circ f$ is then a reparametrization of $\boldsymbol{c}$ which is smooth, since $u^i \circ \tilde{\boldsymbol{c}}$ has vanishing left and right derivative at any t_j for $i = 1, \dots, n$, $j = 1, \dots, k$. □

Definition 2.4.4. A curve $\boldsymbol{c} : [a, b] \to \mathbb{R}^n$ is said to be *regular* if $\boldsymbol{c}'(t) \neq \mathbf{0}$ for $t \in [a, b]$. A regular curve is said to be *normal* or *parametrized by arc length* if $|\boldsymbol{c}'(t)| = 1$ for all t.

Since a normal curve has unit speed, its length equals the length of the interval on which it is defined. It follows from the proof of Proposition 2.4.1 that the smooth reparametrization of a piecewise smooth curve is in general not regular. Regularity, though, has one nice advantage:

Theorem 2.4.2. *Every regular curve admits a reparametrization by arc length.*

Proof. Let $\boldsymbol{c} : [a, b] \to \mathbb{R}^n$ be a curve, and denote by L its length $\ell(c)$. Define a function $\ell_{\boldsymbol{c}} : [a, b] \to [0, L]$ by letting $\ell_{\boldsymbol{c}}(t)$ equal to the length $\ell(\boldsymbol{c}|_{[a,t]})$ of the restriction of $\boldsymbol{c}$ to $[a, t]$. Since $\boldsymbol{c}$ is regular, $\ell'_{\boldsymbol{c}}(t) = |\boldsymbol{c}'(t)| > 0$, so that $\ell_{\boldsymbol{c}}$ admits a differentiable strictly increasing inverse $\ell_{\boldsymbol{c}}^{-1}$. We claim that $\tilde{\boldsymbol{c}} := \boldsymbol{c} \circ \ell_{\boldsymbol{c}}^{-1} : [0, L] \to \mathbb{R}^n$ is normal. To see this, notice that if $f : [\tilde{a}, \tilde{b}] \to [a, b]$ is an increasing differentiable function, then

$$\ell_{\boldsymbol{c}\circ f}(t) = \int_{\tilde{a}}^{t} |(\boldsymbol{c} \circ f)'| = \int_{\tilde{a}}^{t} |\boldsymbol{c}' \circ f| f' = \int_{a}^{f(t)} |\boldsymbol{c}'| = \ell_{\boldsymbol{c}} \circ f(t).$$

Thus, $\ell_{\tilde{\boldsymbol{c}}} = \ell_{\boldsymbol{c}\circ\ell_{\boldsymbol{c}}^{-1}} = \ell_{\boldsymbol{c}} \circ \ell_{\boldsymbol{c}}^{-1}$ is the identity, and $|\tilde{\boldsymbol{c}}'(t)| = \ell'_{\tilde{\boldsymbol{c}}}(t) = 1$, as claimed. □

We end this section with a couple of applications of differentiation that involve the use of curves. In order to state them, we need the following:

Definition 2.4.5. A subset E of $\mathbb{R}^n$ is said to be *path connected* if any two points of E can be joined by a curve that lies in E. It is said to be *convex* if it contains the line segment joining any two of its points.

It is easy to see that a path connected set E is also connected in the sense of Exercise 1.45: otherwise there would exist a separation $X = U \cup V$ of X, and for any curve $\boldsymbol{c} : [a, b] \to E$, the connected set $\boldsymbol{c}[a, b]$ would have to lie entirely in U or in V. Thus, there would be no curve joining a point in U to one in V, contradicting the assumption that E is path connected. The converse is not true, however. The topologist's sine curve from Exercise 1.32 is connected but not path connected. For our purposes, however, only path connectivity matters, so from now on, for the sake of brevity, a connected set will mean a path connected one.

Lemma 2.4.4. *Let $\boldsymbol{c} : [0, 1] \to \mathbb{R}^n$ be a curve, and suppose its speed $|\boldsymbol{c}'|$ does not exceed $\alpha > 0$. Then the distance $|\boldsymbol{c}(1) - \boldsymbol{c}(0)|$ between its endpoints is less than or equal to α.*

Proof. This is an easy consequence of Lemma 2.4.1:

$$|\boldsymbol{c}(1) - \boldsymbol{c}(0)| = |\int_0^1 \boldsymbol{c}'| \leq \int_0^1 |\boldsymbol{c}'| \leq \int_0^1 \alpha = \alpha. \qquad \square$$

Theorem 2.4.3. *Let $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ be a differentiable map.*
(1) *If U is connected and $D\boldsymbol{f} = 0$ on U, then $\boldsymbol{f}$ is constant.*
(2) *If U is convex and $|D\boldsymbol{f}| \leq M$ on U, then*

$$|\boldsymbol{f}(\boldsymbol{b}) - \boldsymbol{f}(\boldsymbol{a})| \leq M|\boldsymbol{b} - \boldsymbol{a}| \text{ for all } \boldsymbol{a}, \boldsymbol{b} \in U.$$

Proof. We begin with (1): Given $\boldsymbol{a}, \boldsymbol{b} \in U$, we must show that $\boldsymbol{f}(\boldsymbol{a}) = \boldsymbol{f}(\boldsymbol{b})$. Let $\boldsymbol{c} : [0,1] \to U$ be a curve that joins $\boldsymbol{a}$ to $\boldsymbol{b}$, and assume for now that $\boldsymbol{c}$ is smooth. The curve $\boldsymbol{f} \circ \boldsymbol{c}$ in $\mathbb{R}^m$ has zero velocity everywhere, since by the chain rule, $(\boldsymbol{f} \circ \boldsymbol{c})' = D(\boldsymbol{f} \circ \boldsymbol{c}) = ((D\boldsymbol{f}) \circ \boldsymbol{c}) \circ D\boldsymbol{c}$. But a curve that has zero velocity is constant, since its components are real-valued functions with vanishing derivative on an interval. Thus $\boldsymbol{f}(\boldsymbol{a}) = (\boldsymbol{f} \circ \boldsymbol{c})(0) = (\boldsymbol{f} \circ \boldsymbol{c})(1) = \boldsymbol{f}(\boldsymbol{b})$. If $\boldsymbol{c}$ is merely piecewise-smooth, the same argument shows that $\boldsymbol{f} \circ \boldsymbol{c}$ is constant on each subinterval where $\boldsymbol{c}$ is smooth, and again we conclude that $\boldsymbol{f}(\boldsymbol{a}) = \boldsymbol{f}(\boldsymbol{b})$.
For (2), the curve $\boldsymbol{c} : [0,1] \to \mathbb{R}^n$, where $\boldsymbol{c}(t) = \boldsymbol{a} + t(\boldsymbol{b} - \boldsymbol{a})$, lies in U, since it parametrizes the line segment between its endpoints. Thus, the speed

$$|(\boldsymbol{f} \circ \boldsymbol{c})'| = |((D\boldsymbol{f}) \circ \boldsymbol{c}) \circ D\boldsymbol{c}| \leq |(D\boldsymbol{f}) \circ \boldsymbol{c}||D\boldsymbol{c}| = |(D\boldsymbol{f}) \circ \boldsymbol{c}||\boldsymbol{b} - \boldsymbol{a}|$$

of the curve $\boldsymbol{f} \circ \boldsymbol{c}$ never exceeds $M|\boldsymbol{b} - \boldsymbol{a}|$, and the claim follows from Lemma 2.4.4. For the inequality above, we used the fact that if L, M are linear transformations for which the composition $L \circ M$ is defined, then $|L \circ M| \leq |L||M|$. This is because

$$|(L \circ M)\boldsymbol{u}| = |L(M(\boldsymbol{u}))| \leq |L||M\boldsymbol{u}| \leq |L||M||\boldsymbol{u}|$$

for any $\boldsymbol{u}$ by (1.4.1). $\square$

Finally, we mention a property of connected sets that will often be useful in the future:

Theorem 2.4.4. *Let E be a connected set. If A is a nonempty subset that is both open and closed in E, then $A = E$.*

Proof. This result is actually the object of Exercise 1.45. For the sake of completeness, we provide an independent proof in the context of path connected sets. Fix some $\boldsymbol{a} \in A$, and let $\boldsymbol{b} \in E$ be arbitrary. It must be shown that $\boldsymbol{b} \in A$. Consider a curve $\boldsymbol{c} : [0,1] \to E$ with $\boldsymbol{c}(0) = \boldsymbol{a}$ and $\boldsymbol{c}(1) = \boldsymbol{b}$, and define $I = \{t \in [0,1] \mid \boldsymbol{c}(t) \in A\}$. I is nonempty because it contains 0, and is open by continuity of $\boldsymbol{c}$: in fact, if $t_0 \in I$, then A is a neighborhood of $\boldsymbol{c}(t_0)$, and there exists some $\varepsilon > 0$ such that $\boldsymbol{c}(t) \in A$ for all t satisfying $|t - t_0| < \varepsilon$. We claim that I is also closed in $[0,1]$: suppose t_0 is a boundary point of I. Then for any natural number i, there exists $t_i \in I$ such that $|t_i - t_0| < 1/i$. Thus,

$t_i \to t_0$, and $\boldsymbol{c}(t_i) \to \boldsymbol{c}(t_0)$ by continuity of $\boldsymbol{c}$. This implies that any neighborhood of $\boldsymbol{c}(t_0)$ contains points of A, so that $\boldsymbol{c}(t_0)$ belongs to A or is a boundary point of A. But A, being closed in E, contains all its boundary points that lie in E, and so $\boldsymbol{c}(t_0) \in A$. This shows that I is also closed. By Proposition 1.7.1, $I = [0, 1]$, and $\boldsymbol{b} \in A$. □

2.5 The inverse and implicit function theorems

Definition 2.5.1. A map $\boldsymbol{f} : U \subset \mathbb{R}^n \to \boldsymbol{f}(U) \subset \mathbb{R}^n$ is said to be a *diffeomorphism* of class C^k ($k \geq 1$) if it is of class C^k on U, and admits an inverse $\boldsymbol{f}^{-1}$ of class C^k on $\boldsymbol{f}(U)$.

The following theorem has many applications; it is yet another indication that the behavior of a map is locally similar to that of its derivative:

Theorem 2.5.1 (Inverse function theorem). *Suppose $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable. If $D\boldsymbol{f}(\boldsymbol{a})$ is invertible at some $\boldsymbol{a} \in U$, then there exists a neighborhood V of $\boldsymbol{a}$ such that the restriction of $\boldsymbol{f}$ to V is a diffeomorphism of class C^1. Furthermore, given $\boldsymbol{p} \in V$, and $\boldsymbol{q} = \boldsymbol{f}(\boldsymbol{p})$, $D(\boldsymbol{f}^{-1})(\boldsymbol{q}) = \big(D\boldsymbol{f}(\boldsymbol{p})\big)^{-1}$.*

The argument will use a property of complete metric spaces (recall that a metric space is complete if every Cauchy sequence converges). It will only be applied in a very specific context, namely for a closed subset C of Euclidean space, but its proof is the same in the general setting. A map $f : X \to X$ from a metric space (X, d) into itself is said to be a *contraction* if there exists $C \in [0, 1)$ such that

$$\mathrm{d}(f(p), f(q)) \leq C\,\mathrm{d}(p, q), \qquad p, q \in X.$$

Notice that a contraction is always continuous; it is, in fact, uniformly continuous: given $\varepsilon > 0$, take $\delta = \varepsilon/C$ if $C > 0$ (if $C = 0$, any δ will do).

Lemma 2.5.1. *If $f : X \to X$ is a contraction of a complete metric space, then f has a unique fixed point; i.e., a point p such that $f(p) = p$.*

Proof of lemma. Uniqueness is clear: if p and q are fixed points of f, then $\mathrm{d}(p, q) = \mathrm{d}(f(p), f(q)) \leq C\,\mathrm{d}(p, q)$, which can only happen if $\mathrm{d}(p, q) = 0$; i.e., if $p = q$. For existence, let q be any point of X, and define a sequence recursively by $x_1 = q$, $x_{k+1} = f(x_k)$. Then $\mathrm{d}(x_k, x_{k+1}) \leq C\,\mathrm{d}(x_{k-1}, x_k)$, and using induction,

$$\mathrm{d}(x_k, x_{k+1}) \leq C^{k-1}\,\mathrm{d}(x_1, x_2).$$

It follows that for $m > n$,

$$\begin{aligned}
\mathrm{d}(x_n, x_m) &\leq \mathrm{d}(x_n, x_{n+1}) + \mathrm{d}(x_{n+1}, x_{n+2}) + \cdots + \mathrm{d}(x_{m-1}, x_m) \\
&\leq (C^{n-1} + C^n + \cdots + C^{m-2})\,\mathrm{d}(x_1, x_2) \\
&= C^{n-1}(1 + C + \cdots + C^{m-n-1})\,\mathrm{d}(x_1, x_2) \\
&\leq C^{n-1}\Big(\sum_{k=0}^{\infty} C^k\Big)\,\mathrm{d}(x_1, x_2) = C^{n-1}\,\frac{\mathrm{d}(x_1, x_2)}{1 - C}.
\end{aligned}$$

Since $C < 1$, $C^n \to 0$, and $\{x_n\}$ is a Cauchy sequence, which must therefore converge to some x. But f is continuous, so that

$$f(x) = f(\lim_{n\to\infty} x_n) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = x,$$

which proves that x is a fixed point of f. □

Proof of Theorem 2.5.1. To simplify notation, it may be assumed that $\boldsymbol{a} = \boldsymbol{f}(\boldsymbol{a}) = \boldsymbol{0}$: for the map $\boldsymbol{F}$, where $\boldsymbol{F}(\boldsymbol{p}) = \boldsymbol{f}(\boldsymbol{p} + \boldsymbol{a}) - \boldsymbol{f}(\boldsymbol{a})$, satisfies these conditions, and if $\boldsymbol{F}$ is a diffeomorphism in a neighborhood of $\boldsymbol{0}$, then $\boldsymbol{p} \mapsto \boldsymbol{f}(\boldsymbol{p}) = \boldsymbol{F}(\boldsymbol{p} - \boldsymbol{a}) + \boldsymbol{f}(\boldsymbol{a})$, being a composition of local diffeomorphisms, is one too. Similarly, $D\boldsymbol{f}(\boldsymbol{0})$ may be assumed to be the identity I (replacing $\boldsymbol{f}$, if need be, by $\boldsymbol{p} \mapsto (D\boldsymbol{f}(\boldsymbol{0}))^{-1} \circ \boldsymbol{f}(\boldsymbol{p})$, which by the chain rule has the identity as derivative at $\boldsymbol{0}$). Thus, the associated map

$$\begin{aligned} \boldsymbol{g} : U &\to \mathbb{R}^n, \\ \boldsymbol{p} &\mapsto \boldsymbol{p} - \boldsymbol{f}(\boldsymbol{p}) \end{aligned}$$

satisfies $\boldsymbol{g}(\boldsymbol{0}) = \boldsymbol{0}$ and $D\boldsymbol{g}(\boldsymbol{0}) = 0$. B_r will denote the *closed* ball of radius r centered at $\boldsymbol{0}$. Recalling that $D\boldsymbol{f}$ is continuous and that $D\boldsymbol{f}(\boldsymbol{0}) = I$, there exists $r > 0$ small enough that

$$|D\boldsymbol{g}(\boldsymbol{p})| = |I - D\boldsymbol{f}(\boldsymbol{p})| < \frac{1}{2}, \qquad \boldsymbol{p} \in B_r.$$

Together with Theorem 2.4.3, this means that

$$|\boldsymbol{g}(\boldsymbol{p}) - \boldsymbol{g}(\boldsymbol{q})| \le \frac{1}{2}|\boldsymbol{p} - \boldsymbol{q}|, \qquad \boldsymbol{p}, \boldsymbol{q} \in B_r, \tag{2.5.1}$$

and consequently,

$$|\boldsymbol{f}(\boldsymbol{p}) - \boldsymbol{f}(\boldsymbol{q})| \ge |\boldsymbol{p} - \boldsymbol{q}| - |\boldsymbol{g}(\boldsymbol{p}) - \boldsymbol{g}(\boldsymbol{q})| \ge \frac{1}{2}|\boldsymbol{p} - \boldsymbol{q}|, \qquad \boldsymbol{p}, \boldsymbol{q} \in B_r. \tag{2.5.2}$$

Notice that (2.5.2) already implies that $\boldsymbol{f}$ is one-to-one on B_r. Next, we claim that $\boldsymbol{f}(B_r)$ contains $B_{r/2}$. To see why, consider any $\boldsymbol{q} \in B_{r/2}$. In order to find some $\boldsymbol{p} \in B_r$ that gets mapped to $\boldsymbol{q}$, observe that $\boldsymbol{g}_{\boldsymbol{q}}$, where $\boldsymbol{g}_{\boldsymbol{q}}(\boldsymbol{p}) = \boldsymbol{g}(\boldsymbol{p}) + \boldsymbol{q}$, sends B_r to itself. But B_r, being closed, is complete, and by (2.5.1), $\boldsymbol{g}_{\boldsymbol{q}}$ is a contraction, since

$$|\boldsymbol{g}_{\boldsymbol{q}}(\boldsymbol{p}) - \boldsymbol{g}_{\boldsymbol{q}}(\boldsymbol{r})| = |\boldsymbol{g}(\boldsymbol{p}) - \boldsymbol{g}(\boldsymbol{r})| \le \frac{1}{2}|\boldsymbol{p} - \boldsymbol{r}|.$$

By Lemma 2.5.1, $\boldsymbol{g}_{\boldsymbol{q}}$ has a unique fixed point $\boldsymbol{p}$; i.e., $\boldsymbol{g}_{\boldsymbol{q}}(\boldsymbol{p}) = \boldsymbol{p} - \boldsymbol{f}(\boldsymbol{p}) + \boldsymbol{q} = \boldsymbol{p}$, or $\boldsymbol{q} = \boldsymbol{f}(\boldsymbol{p})$, as claimed.

Set $W = \operatorname{int}(B_{r/2})$, $V = \boldsymbol{f}^{-1}(W)$. Then the restriction $\boldsymbol{f} : V \to W$ admits an inverse $\boldsymbol{f}^{-1} : W \to V$. This inverse is certainly continuous, since by (2.5.2),

$$|\boldsymbol{f}^{-1}(\boldsymbol{p}) - \boldsymbol{f}^{-1}(\boldsymbol{q})| \le 2|\boldsymbol{p} - \boldsymbol{q}|.$$

We claim it is actually differentiable at any $\boldsymbol{q}_0 \in W$. To see this, let $\boldsymbol{p}_0 = \boldsymbol{f}^{-1}(\boldsymbol{q}_0)$, $L = D\boldsymbol{f}(\boldsymbol{p}_0)$, and $r_{\boldsymbol{f}}$ the map defined by

$$r_{\boldsymbol{f}}(\boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{p}_0 + \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{p}_0) - L\boldsymbol{h},$$

so that as usual,

$$\boldsymbol{f}(\boldsymbol{p}_0 + \boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{p}_0) + L\boldsymbol{h} + r_{\boldsymbol{f}}(\boldsymbol{h}), \text{ where } \lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{|r_{\boldsymbol{f}}(\boldsymbol{h})|}{|\boldsymbol{h}|} = 0.$$

Apply L^{-1} to both sides of the above identity to conclude that

$$\boldsymbol{h} = L^{-1}\left(\boldsymbol{f}(\boldsymbol{p}_0 + \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{p}_0)\right) - L^{-1}(r_{\boldsymbol{f}}(\boldsymbol{h})). \tag{2.5.3}$$

Set $\boldsymbol{k} = \boldsymbol{f}(\boldsymbol{p}_0 + \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{p}_0)$. Then

$$\begin{aligned}\boldsymbol{f}^{-1}(\boldsymbol{q}_0 + \boldsymbol{k}) - \boldsymbol{f}^{-1}(\boldsymbol{q}_0) &= \boldsymbol{f}^{-1}(\boldsymbol{q}_0 + \boldsymbol{f}(\boldsymbol{p}_0 + \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{p}_0)) - \boldsymbol{f}^{-1}(\boldsymbol{q}_0)\\ &= \boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{p}_0 + \boldsymbol{h})) - \boldsymbol{p}_0\\ &= \boldsymbol{h},\end{aligned}$$

so that by (2.5.3) and the definition of $\boldsymbol{k}$,

$$\boldsymbol{f}^{-1}(\boldsymbol{q}_0 + \boldsymbol{k}) - \boldsymbol{f}^{-1}(\boldsymbol{q}_0) - L^{-1}\boldsymbol{k} = -L^{-1} \circ r_{\boldsymbol{f}}(\boldsymbol{h}).$$

This means that $\boldsymbol{f}$ is differentiable at $\boldsymbol{q}_0$ with derivative L^{-1}, provided

$$\lim_{\boldsymbol{k}\to\boldsymbol{0}} \frac{|L^{-1} \circ r_{\boldsymbol{f}}(\boldsymbol{h})|}{|\boldsymbol{k}|} = 0. \tag{2.5.4}$$

To establish this identity, notice that the definition of $\boldsymbol{k}$ together with (2.5.2) implies that $|\boldsymbol{k}| \geq |\boldsymbol{h}|/2$, so that

$$\frac{|L^{-1} \circ r_{\boldsymbol{f}}(\boldsymbol{h})|}{|\boldsymbol{k}|} \leq |L^{-1}| \frac{|r_{\boldsymbol{f}}(\boldsymbol{h})|}{|\boldsymbol{h}|} \frac{|\boldsymbol{h}|}{|\boldsymbol{k}|} \leq 2|L^{-1}| \frac{|r_{\boldsymbol{f}}(\boldsymbol{h})|}{|\boldsymbol{h}|}.$$

By continuity of $\boldsymbol{f}^{-1}$, $\boldsymbol{h} \to \boldsymbol{0}$ when $\boldsymbol{k} \to \boldsymbol{0}$. This proves (2.5.4) and the differentiability of $\boldsymbol{f}^{-1}$. It only remains to show that $\boldsymbol{f}^{-1}$ is continuously differentiable. But this is an immediate consequence of the fact that the matrix of $D(\boldsymbol{f}^{-1})$ is the inverse of the matrix of $D\boldsymbol{f}$, and inversion of matrices is a continuous map, see Exercise 1.19. □

Examples 2.5.1. (i) The *polar coordinates* of a point in the plane $\mathbb{R}^2$ are (r, θ), where r equals the distance from the point to the origin, and θ is the angle between the vector representing the point and $\boldsymbol{e}_1$. They are well defined for any point different from the origin; in fact, in terms of Cartesian coordinates,

$$(r, \theta)(x, y) = \boldsymbol{f}(x, y) = (\sqrt{x^2 + y^2}, \arctan \frac{y}{x} + c), \tag{2.5.5}$$

if $x \neq 0$. Here, $c = 0$, π, or 2π depending on which quadrant (x, y) lies, cf. Section 4.6.1. When $x = 0$, $\theta = \pi/2$ if $y > 0$, and $3\pi/2$ if $y < 0$. The Jacobian matrix of $\boldsymbol{f}$ is

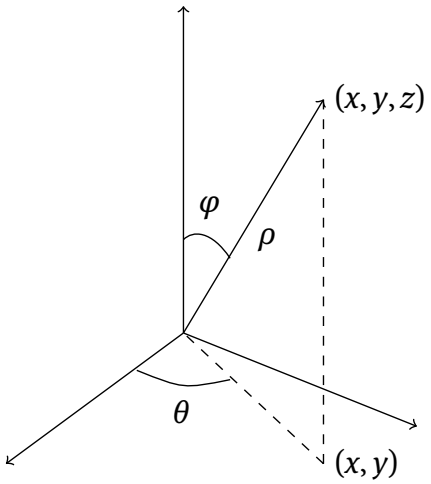
$$D\boldsymbol{f}(x, y) = \begin{bmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} \\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{bmatrix},$$

which has determinant $(x^2 + y^2)^{-1/2}$. By Theorem 2.5.1, $\boldsymbol{f}$ is invertible in a neighborhood of any point $\boldsymbol{p}$ in its domain, with

$$(x, y) = \boldsymbol{f}^{-1}(r, \theta) = (r\cos\theta, r\sin\theta).$$

The domain of the inverse can be taken to be $(0, \infty) \times [a, a + 2\pi)$, $a \in \mathbb{R}$, with the value of a depending on the point $\boldsymbol{p}$. When $\boldsymbol{p}$ does not lie on the positive x-axis, it is customary to take $a = 0$.

(ii) The *spherical coordinates* of a point (x, y, z) in $\mathbb{R}^3$ are (ρ, θ, φ), where ρ is the distance from the point to the origin, θ is the polar coordinate angle of the projection (x, y) of the point in the plane, and φ is the angle between the vector representing the point and $\boldsymbol{e}_3$.



Spherical coordinates (ρ, θ, φ) of the point (x, y, z) in $\mathbb{R}^3$

ρ, θ, and φ are the components of the map $\boldsymbol{f}$ given by

$$\boldsymbol{f}(x, y, z) = \left((x^2 + y^2 + z^2)^{1/2}, \arctan\frac{y}{x} + c, \arccos\frac{z}{(x^2 + y^2 + z^2)^{1/2}}\right)$$

(with c as above), and its inverse is

$$\boldsymbol{f}^{-1}(\rho, \theta, \varphi) = (\rho\sin\varphi\cos\theta, \rho\sin\varphi\sin\theta, \rho\cos\varphi).$$

One easily computes that the Jacobian of $\boldsymbol{f}^{-1}$ has determinant $-\rho^2\sin\varphi$, see Section 4.6.2.

(iii) Although the inverse function theorem asserts that a map from $\mathbb{R}^n$ to itself is one-to-one in a neighborhood of any point where the derivative is one-to-one, the converse is not true already for $n = 1$. For example, $f : \mathbb{R} \to \mathbb{R}$, where $f(x) = x^3$, is globally one-to-one even though its derivative is zero at the origin. Of course, even though a map may still be one-to-one with vanishing derivative at a point, it cannot in that case be a diffeomorphism, since diffeomorphisms always have invertible derivative.

The inverse function theorem, which deals with maps from $\mathbb{R}^n$ to $\mathbb{R}^n$, admits generalizations to maps from $\mathbb{R}^n$ to $\mathbb{R}^k$, when k is larger or smaller than n. They are known as the implicit function theorems. Before stating them, we need some terminology. If $\boldsymbol{f}$ is differentiable, its *rank* at $\boldsymbol{p}$ is defined to be the rank of the linear map $D\boldsymbol{f}(\boldsymbol{p})$; i.e., the dimension of the image $D\boldsymbol{f}(\boldsymbol{p})\mathbb{R}^n$ of the derivative at $\boldsymbol{p}$. Since this image is spanned by the columns $D\boldsymbol{f}(\boldsymbol{p})\boldsymbol{e}_1, \dots, D\boldsymbol{f}(\boldsymbol{p})\boldsymbol{e}_n$ of the Jacobian matrix, the rank equals the number of linearly independent columns of the Jacobian. It turns out that the dimension of the space spanned by the columns of an $m \times n$ matrix A equals that of the space spanned by the rows of A: indeed, apply Theorem 1.2.2 to the linear transformation $L_A : \mathbb{R}^n \to \mathbb{R}^m$ to deduce that the former equals $n - \dim \ker L_A$. The latter, on the other hand, equals the dimension of the space spanned by $\boldsymbol{a}_1, \dots, \boldsymbol{a}_m$, where $\boldsymbol{a}_i$ denotes the transpose (so it can be viewed as a vector) of the i-th row of the matrix. But for $\boldsymbol{p} \in \mathbb{R}^n$,

$$A\boldsymbol{p} = \begin{bmatrix} \langle \boldsymbol{a}_1, \boldsymbol{p} \rangle \\ \vdots \\ \langle \boldsymbol{a}_m, \boldsymbol{p} \rangle \end{bmatrix},$$

so that the kernel of L_A coincides with the orthogonal complement of the row space. By Proposition 1.4.1, the row space also has dimension $n - \dim \ker L_A$, as claimed.

Theorem 2.5.2 (Implicit Function Theorem). *Let U be a neighborhood of $\boldsymbol{0}$ in $\mathbb{R}^n$, $\boldsymbol{f} : U \to \mathbb{R}^k$ a continuously differentiable map with $\boldsymbol{f}(\boldsymbol{0}) = \boldsymbol{0}$. For $n \le k$, let $\imath : \mathbb{R}^n \to \mathbb{R}^k$ denote the inclusion*

$$\imath(a_1, \dots, a_n) = (a_1, \dots, a_n, 0, \dots, 0),$$

and for $n \ge k$, let $\pi : \mathbb{R}^n \to \mathbb{R}^k$ denote the projection

$$\pi(a_1, \dots, a_k, \dots, a_n) = (a_1, \dots, a_k).$$

(1) *If $n \le k$ and $\boldsymbol{f}$ has maximal rank ($= n$) at 0, then there exists a diffeomorphism $\boldsymbol{g}$ of a neighborhood of $\boldsymbol{0}$ in $\mathbb{R}^k$ such that $\boldsymbol{g} \circ \boldsymbol{f} = \imath$ in a neighborhood of $\boldsymbol{0} \in \mathbb{R}^n$.*
(2) *If $n \ge k$ and $\boldsymbol{f}$ has maximal rank ($= k$) at 0, then there exists a diffeomorphism $\boldsymbol{h}$ of a neighborhood of $\boldsymbol{0}$ in $\mathbb{R}^n$ such that $\boldsymbol{f} \circ \boldsymbol{h} = \pi$.*

Proof. As usual, we denote the component functions $u^i \circ \boldsymbol{f}$ of $\boldsymbol{f}$ by f^i. In order to prove (1), observe that the $k \times n$ matrix $[D_j f^i(\boldsymbol{0})]$ has rank n. By rearranging the component functions f^i of $\boldsymbol{f}$ if necessary (which amounts to composing $\boldsymbol{f}$ with an invertible transformation, hence a diffeomorphism of $\mathbb{R}^k$), we may assume that the $n \times n$ submatrix $[D_j f^i(\boldsymbol{0})]_{1 \le i,j \le n}$ is invertible. Define $\boldsymbol{F} : U \times \mathbb{R}^{k-n} \to \mathbb{R}^k$ by

$$\boldsymbol{F}(a_1, \dots, a_n, a_{n+1}, \dots, a_k) := \boldsymbol{f}(a_1, \dots, a_n) + (0, \dots, 0, a_{n+1}, \dots, a_k).$$

Then $\boldsymbol{F} \circ \imath = \boldsymbol{f}$, and the Jacobian matrix of $\boldsymbol{F}$ at $\boldsymbol{0}$ is

$$\begin{bmatrix} [D_j f^i(\boldsymbol{0})]_{1 \le i \le n} & 0 \\ [D_j f^i(\boldsymbol{0})]_{n+1 \le i \le k} & 1_{\mathbb{R}^{k-n}} \end{bmatrix},$$

which has nonzero determinant. Consequently, $\boldsymbol{F}$ has a local inverse $\boldsymbol{g}$, and $\boldsymbol{g} \circ \boldsymbol{f} = \boldsymbol{g} \circ \boldsymbol{F} \circ \iota = \iota$. This establishes (1). Similarly, in (2), we may assume that the $k \times k$ submatrix $[D_j f^i(\mathbf{0})]_{1 \le i,j \le k}$ is invertible. Define $\boldsymbol{F} : U \to \mathbb{R}^n$ by

$$\boldsymbol{F}(a_1, \dots, a_n) := (\boldsymbol{f}(a_1, \dots, a_n), a_{k+1}, \dots, a_n).$$

Then $\boldsymbol{f} = \pi \circ \boldsymbol{F}$, and the Jacobian of $\boldsymbol{F}$ at $\mathbf{0}$ is

$$\begin{bmatrix} [D_j f^i(\mathbf{0})]_{1 \le j \le k} & [D_j f^i(\mathbf{0})]_{k+1 \le j \le n} \\ 0 & 1_{\mathbb{R}^{n-k}} \end{bmatrix},$$

which is invertible. Thus, $\boldsymbol{F}$ has a local inverse $\boldsymbol{h}$, and $\boldsymbol{f} \circ \boldsymbol{h} = \pi \circ \boldsymbol{F} \circ \boldsymbol{h} = \pi$. □

The reason the above is called the implicit function theorem is that under certain circumstances, given a differentiable map $\boldsymbol{f} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, the equation

$$\boldsymbol{f}(\boldsymbol{p}, \boldsymbol{q}) = \mathbf{0}, \qquad \boldsymbol{p} \in \mathbb{R}^n, \quad \boldsymbol{q} \in \mathbb{R}^m,$$

implicitly defines $\boldsymbol{p}$ as a function $\boldsymbol{g}(\boldsymbol{q})$ of $\boldsymbol{q}$; i.e., for each $\boldsymbol{q}$, there exists a unique $\boldsymbol{g}(\boldsymbol{q})$ such that $\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{q}), \boldsymbol{q}) = \mathbf{0}$, and the map $\boldsymbol{g}$ is differentiable; more precisely, we have:

Corollary 2.5.1 (Classical implicit function theorem). *Let $\boldsymbol{f}$ be a continuously differentiable map from an open neighborhood of $(\boldsymbol{0}, \boldsymbol{0}) \in \mathbb{R}^n \times \mathbb{R}^m$ to $\mathbb{R}^n$ such that $\boldsymbol{f}(\boldsymbol{0}, \boldsymbol{0}) = \boldsymbol{0}$.*

If the matrix $[D_j f^i(\boldsymbol{0}, \boldsymbol{0})]_{1 \le i,j \le n}$ is invertible, then there exist open neighborhoods V of $\boldsymbol{0} \in \mathbb{R}^m$, U of $(\boldsymbol{0}, \boldsymbol{0}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that for every $\boldsymbol{q} \in V$, there exists a unique $\boldsymbol{p}$ satisfying $(\boldsymbol{p}, \boldsymbol{q}) \in U$ and $\boldsymbol{f}(\boldsymbol{p}, \boldsymbol{q}) = 0$. Furthermore, the map $\boldsymbol{g} : V \to \mathbb{R}^n$ that assigns to each $\boldsymbol{q}$ the unique $\boldsymbol{p} = \boldsymbol{g}(\boldsymbol{q})$ such that $\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{q}), \boldsymbol{q}) = \boldsymbol{0}$ is continuously differentiable.

Proof. By Theorem 2.5.2 (2), there exists a local diffeomorphism $\boldsymbol{h}$ such that $\boldsymbol{f} \circ \boldsymbol{h}$ equals the projection $\pi_1 : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ onto the first factor. If π_2 is the projection onto the second factor, then the inverse $\boldsymbol{F}$ of $\boldsymbol{h}$ was shown to satisfy $\pi_2 \circ \boldsymbol{F} = \pi_2$. This implies that if $\boldsymbol{f}(\boldsymbol{p}, \boldsymbol{q}) = \mathbf{0}$, then

$$\mathbf{0} = \boldsymbol{f}(\boldsymbol{p}, \boldsymbol{q}) = \boldsymbol{f} \circ \boldsymbol{h} \circ \boldsymbol{F}(\boldsymbol{p}, \boldsymbol{q}) = \pi_1 \circ \boldsymbol{F}(\boldsymbol{p}, \boldsymbol{q}),$$

so that $\boldsymbol{F}(\boldsymbol{p}, \boldsymbol{q}) = (\mathbf{0}, \boldsymbol{q})$; i.e., $\boldsymbol{p} = \pi_1 \circ \boldsymbol{F}^{-1}(\mathbf{0}, \boldsymbol{q}) = \pi_1 \circ \boldsymbol{h}(\mathbf{0}, \boldsymbol{q})$. This shows not only uniqueness of $\boldsymbol{p}$, but also existence and smoothness of $\boldsymbol{g}$: let

$$\begin{aligned} \iota : V &\mapsto \mathbb{R}^n \times \mathbb{R}^m, \\ \boldsymbol{q} &\mapsto (\mathbf{0}, \boldsymbol{q}). \end{aligned}$$

Then the map $\boldsymbol{g} := \pi_1 \circ \boldsymbol{h} \circ \iota$, being a composition of $\mathcal{C}^1$ maps, is continuously differentiable, and

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{q}), \boldsymbol{q}) &= \boldsymbol{f}(\pi_1(\boldsymbol{h}(\mathbf{0}, \boldsymbol{q})), \boldsymbol{q}) = \boldsymbol{f}(\pi_1(\boldsymbol{h}(\mathbf{0}, \boldsymbol{q})), \pi_2(\boldsymbol{h}(\mathbf{0}, \boldsymbol{q})) \\ &= (\boldsymbol{f} \circ \boldsymbol{h})(\mathbf{0}, \boldsymbol{q})) = \pi_1(\mathbf{0}, \boldsymbol{q}) \\ &= \mathbf{0}. \end{aligned}$$

□

Examples and Remarks 2.5.1. (i) Define

$$f : \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R},$$
$$(x,(y,z)) \mapsto (x-r)^2 + y^2 + z^2 - r^2,$$

with $r > 0$. The set of points $(x,y,z) \in \mathbb{R}^3$ satisfying $f(x,y,z) = 0$ is a sphere of radius r that passes through the origin. Since $D_1 f(0,0) = -2r \neq 0$, a part of the sphere containing the origin may be represented by the graph of a function $x = g(y,z)$, even though the whole sphere cannot. In fact, solving $f(x,y,z) = 0$ in terms of x yields $x = g(y,z) = r - \sqrt{r^2 - y^2 - z^2}$.

(ii) Although the classical implicit function theorem was stated at the origin, it is straightforward to obtain a version of it that is valid at any point: specifically, if $\boldsymbol{f}$ is a continuously differentiable map from a neighborhood of $(\boldsymbol{p}_0, \boldsymbol{q}_0) \in \mathbb{R}^n \times \mathbb{R}^m$ to $\mathbb{R}^n$ such that $\boldsymbol{f}(\boldsymbol{p}_0, \boldsymbol{q}_0) = \boldsymbol{r}_0 \in \mathbb{R}^n$, and if the matrix $[D_j f^i(\boldsymbol{p}_0, \boldsymbol{q}_0)]_{1 \le i,j, \le n}$ is invertible, then there exists an open neighborhood V of $\boldsymbol{q}_0 \in \mathbb{R}^m$, and a continuously differentiable map $\boldsymbol{g} : V \to \mathbb{R}^m$ such that $\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{q}), \boldsymbol{q}) = \boldsymbol{r}_0$.
To see this, observe that the map $\boldsymbol{h}$, where $\boldsymbol{h}(\boldsymbol{p}, \boldsymbol{q}) := \boldsymbol{f}(\boldsymbol{p} + \boldsymbol{p}_0, \boldsymbol{q} + \boldsymbol{q}_0) - \boldsymbol{r}_0$, satisfies the hypotheses of the implicit function theorem, so that there exists a continuously differentiable map $\boldsymbol{k} : V_0 \to \mathbb{R}^n$ defined on a neighborhood V_0 of the origin in $\mathbb{R}^m$ such that $\boldsymbol{h}(\boldsymbol{k}(\boldsymbol{q}), \boldsymbol{q}) = \boldsymbol{0}$. Then $V := \{\boldsymbol{q} + \boldsymbol{q}_0 \mid \boldsymbol{q} \in V_0\}$ is a neighborhood of $\boldsymbol{q}_0$, and $\boldsymbol{g} : V \to \mathbb{R}^n$, where $\boldsymbol{g}(\boldsymbol{q}) := \boldsymbol{k}(\boldsymbol{q} - \boldsymbol{q}_0) + \boldsymbol{p}_0$, is the map we are looking for, since

$$\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{q}), \boldsymbol{q}) = \boldsymbol{f}(\boldsymbol{k}(\boldsymbol{q} - \boldsymbol{q}_0) + \boldsymbol{p}_0, \boldsymbol{q}) = \boldsymbol{h}(\boldsymbol{k}(\boldsymbol{q} - \boldsymbol{q}_0), \boldsymbol{q} - \boldsymbol{q}_0) + \boldsymbol{r}_0 = \boldsymbol{r}_0.$$

(iii) To illustrate (ii) above, suppose we are asked to show that there is a neighborhood U of $(1,0)$ in $\mathbb{R}^2$, and a differentiable function $f : U \to \mathbb{R}$ satisfying

$$xf^2(x,y) + yf(x,y) - x^2 f^5(x,y) + y = 0, \qquad f(1,0) = 1.$$

Finding an explicit formula for f seems pretty much hopeless, so let us define $F : \mathbb{R}^3 \to \mathbb{R}$ by $F(a_1, a_2, a_3) = a_2 a_1^2 + a_3 a_1 - a_2^2 a_1^5 + a_3$. Then $F(1,1,0) = 0$, and $D_1 F(1,1,0) = -3 \neq 0$, so there exists a function f on a neighborhood of $(1,0)$ such that

$$F(f(a_2, a_3), a_2, a_3) = a_2 f^2(a_2, a_3) + a_3 f(a_2, a_3) - a_2^2 f^5(a_2, a_3) + a_3 = 0.$$

Substituting x for a_2 and y for a_3 shows this is the function we are looking for.

2.6 The spectral theorem and scalar products

In this section, we take a closer look at symmetric matrices since they pop up frequently in calculus and differential geometry: for instance, to each point of a surface,

one can associate a symmetric matrix that measures how curved the surface is at that point. As another example, in the next section, we will use symmetric matrices to classify extreme values of real-valued functions. The discussion will be couched in terms of linear maps rather than matrices.

Recall that a linear map from a vector space V to itself is also called an *operator* on V. If V is an inner product space, there are several types of operators on V that play a special role; we now introduce two of these:

Definition 2.6.1. A linear operator L on an inner product space V is said to be *self-adjoint* if $\langle L\boldsymbol{u}, \boldsymbol{v}\rangle = \langle \boldsymbol{u}, L\boldsymbol{v}\rangle$ for all $\boldsymbol{u}, \boldsymbol{v} \in V$. L is said to be *skew-adjoint* if $\langle L\boldsymbol{u}, \boldsymbol{v}\rangle = -\langle \boldsymbol{u}, L\boldsymbol{v}\rangle$ for all $\boldsymbol{u}, \boldsymbol{v} \in V$.

Proposition 2.6.1. *Let V be an inner product space with ordered orthonormal basis $\mathcal{B} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$.*

(1) *If L is a linear operator on V, then the (i, j)-th entry of the matrix $[L]_{\mathcal{B},\mathcal{B}}$ of L in the basis $\mathcal{B}$ is $\langle \boldsymbol{u}_i, L\boldsymbol{u}_j\rangle$.*
(2) *L is self-adjoint if and only if $[L]_{\mathcal{B},\mathcal{B}}$ is symmetric.*
(3) *L is skew-adjoint if and only if $[L]_{\mathcal{B},\mathcal{B}}$ is skew-symmetric.*

Proof. For (1), recall that the j-th column of the matrix of L in the basis $\mathcal{B}$ is the coordinate vector $[L\boldsymbol{u}_j]_{\mathcal{B}}$ of $L\boldsymbol{u}_j$ in the basis $\mathcal{B}$. By Theorem 1.4.3, the i-th entry of that vector is $\langle \boldsymbol{u}_i, L\boldsymbol{u}_j\rangle$. For (2), if L is self-adjoint, then the matrix of L in $\mathcal{B}$ is symmetric by (1). Conversely, if the matrix of L is symmetric, then by (1), $\langle L\boldsymbol{u}_i, \boldsymbol{u}_j\rangle = \langle \boldsymbol{u}_i, L\boldsymbol{u}_j\rangle$ for all i and j, and given $\boldsymbol{u}, \boldsymbol{v} \in V$,

$$\begin{aligned}\langle L\boldsymbol{u}, \boldsymbol{v}\rangle &= \langle L\left(\sum_i \langle \boldsymbol{u}, \boldsymbol{u}_i\rangle \boldsymbol{u}_i\right), \sum_j \langle \boldsymbol{v}, \boldsymbol{u}_j\rangle \boldsymbol{u}_j\rangle \\ &= \sum_{i,j} \langle \boldsymbol{u}, \boldsymbol{u}_i\rangle \langle \boldsymbol{v}, \boldsymbol{u}_j\rangle \langle L\boldsymbol{u}_i, \boldsymbol{u}_j\rangle = \sum_{i,j} \langle \boldsymbol{u}, \boldsymbol{u}_i\rangle \langle \boldsymbol{v}, \boldsymbol{u}_j\rangle \langle \boldsymbol{u}_i, L\boldsymbol{u}_j\rangle \\ &= \langle \sum_i \langle \boldsymbol{u}, \boldsymbol{u}_i\rangle \boldsymbol{u}_i, L\left(\sum_j \langle \boldsymbol{v}, \boldsymbol{u}_j\rangle \boldsymbol{u}_j\right)\rangle = \langle \boldsymbol{u}, L\boldsymbol{v}\rangle,\end{aligned}$$

so that L is self-adjoint. The proof of (3) is similar. □

A third important type of operator that occurs in the presence of an inner product can be formulated more generally in terms of a linear transformation between two (possibly) different inner product spaces:

Definition 2.6.2. A linear transformation $L : (V_1, \langle , \rangle_1) \to (V_2, \langle , \rangle_2)$ between two inner product spaces of the same dimension is said to be a *linear isometry* if it preserves the inner product; i.e., if

$$\langle L\boldsymbol{u}, L\boldsymbol{v}\rangle_2 = \langle \boldsymbol{u}, \boldsymbol{v}\rangle_1, \qquad \boldsymbol{u}, \boldsymbol{v} \in V_1.$$

Equivalently, L is a linear isometry if it preserves the norm of vectors. It follows that a linear isometry has trivial kernel (for if $L\boldsymbol{u} = \boldsymbol{0}$, then $|\boldsymbol{u}| = |L\boldsymbol{u}| = 0$, so $\boldsymbol{u} = \boldsymbol{0}$), and is

therefore an isomorphism. Two inner product spaces with a linear isometry between them are said to be *isometric*. Isometric spaces are essentially the same both from a vector space perspective and from an inner product one. For example, it is easy to check that if $T : V \to W$ is a linear isometry and L is a self-adjoint (respectively skew-adjoint) operator on W, then $T^{-1} \circ L \circ T$ is a self-adjoint (resp. skew-adjoint) operator on V.

Proposition 2.6.2. *Let V be an inner product space with ordered orthonormal basis $\mathcal{B} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$, and L an operator on V. Then L is an isometry if and only if its matrix P with respect to $\mathcal{B}$ is* orthogonal, *meaning that $P^{-1} = P^T$.*

Proof. By Exercise 2.16, the isomorphism $V \to \mathbb{R}^n$ that maps a vector to its coordinate vector in the basis $\mathcal{B}$ is a linear isometry, so that

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = [\boldsymbol{u}]_{\mathcal{B}}^T [\boldsymbol{v}]_{\mathcal{B}}, \qquad \boldsymbol{u}, \boldsymbol{v} \in V.$$

Recall that the *Kronecker delta* δ_{ij} is defined to be 1 when $i = j$ and 0 otherwise. If L is an isometry, then

$$\delta_{ij} = \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = \langle L\boldsymbol{u}_i, L\boldsymbol{u}_j \rangle = [L\boldsymbol{u}_i]_{\mathcal{B}}^T [L\boldsymbol{u}_j]_{\mathcal{B}}.$$

The last term in the above identity is the (i, j)-th entry of $P^T P$, which shows that P is orthogonal.
Conversely, if $P^T P = I_n$, then

$$\delta_{ij} = [L\boldsymbol{u}_i]_{\mathcal{B}}^T [L\boldsymbol{u}_j]_{\mathcal{B}} = \langle L\boldsymbol{u}_i, L\boldsymbol{u}_j \rangle,$$

so that $\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = \langle L\boldsymbol{u}_i, L\boldsymbol{u}_j \rangle$. Thus, given $\boldsymbol{v}, \boldsymbol{w} \in V$,

$$\begin{aligned}
\langle L\boldsymbol{v}, L\boldsymbol{w} \rangle &= \langle L\left(\sum_i \langle \boldsymbol{v}, \boldsymbol{u}_i \rangle \boldsymbol{u}_i\right), L\left(\sum_j \langle \boldsymbol{w}, \boldsymbol{u}_j \rangle \boldsymbol{u}_j\right) \rangle \\
&= \sum_{i,j} \langle \boldsymbol{v}, \boldsymbol{u}_i \rangle \langle \boldsymbol{w}, \boldsymbol{u}_j \rangle \langle L\boldsymbol{u}_i, L\boldsymbol{u}_j \rangle = \sum_{i,j} \langle \boldsymbol{v}, \boldsymbol{u}_i \rangle \langle \boldsymbol{w}, \boldsymbol{u}_j \rangle \delta_{ij} \\
&= \sum_i \langle \boldsymbol{v}, \boldsymbol{u}_i \rangle \langle \boldsymbol{w}, \boldsymbol{u}_i \rangle = [\boldsymbol{v}]_{\mathcal{B}}^T [\boldsymbol{w}]_{\mathcal{B}} \\
&= \langle \boldsymbol{v}, \boldsymbol{w} \rangle,
\end{aligned}$$

which shows that L is an isometry. □

For example, the operator on $\mathbb{R}^2$ consisting of counterclockwise rotation by angle θ about the origin is a linear isometry. This can be checked by either noticing that such a rotation leaves the norm of vectors unchanged, or by recalling that the matrix of this operator in the standard basis is

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

which is clearly orthogonal.

Definition 2.6.3. A real number λ is said to be an *eigenvalue* of a linear operator L on V if there exists some $\boldsymbol{v} \neq \mathbf{0}$ in V such that $L\boldsymbol{v} = \lambda \boldsymbol{v}$. In this case, $\boldsymbol{v}$ is called an *eigenvector* of L. When L is the operator on $\mathbb{R}^n$ given by left multiplication L_A by some $n \times n$ matrix A, its eigenvalues and eigenvectors are also often called the eigenvalues and eigenvectors of A.

The reason why one requires $\boldsymbol{v} \neq \mathbf{0}$ in the above definition is that otherwise any λ would always be an eigenvalue of every operator L, since $L\mathbf{0} = \lambda \mathbf{0}$. With this restriction, on the other hand, 0 is an eigenvalue of L if and only if L has nontrivial kernel, or equivalently (at least if V is finite-dimensional) if and only if L is not an isomorphism. Notice also that if λ is an eigenvalue of L, then the collection of corresponding eigenvectors becomes a subspace of V, once we add the vector $\mathbf{0}$ to it. It is called the λ*-eigenspace* of L.

In general, an operator need not admit any eigenvalues. One such is a rotation by, say, $\pi/4$ in the plane. The situation is entirely different when L is self-adjoint, however:

Theorem 2.6.1. *Any self-adjoint operator on an inner product space V admits an eigenvalue.*

Proof. Let us assume for now that V is $\mathbb{R}^n$ with the standard inner product, and let L denote the operator. The function

$$\begin{aligned} f : \mathbb{R}^n &\to \mathbb{R}, \\ \boldsymbol{v} &\mapsto \langle L\boldsymbol{v}, \boldsymbol{v} \rangle \end{aligned}$$

is continuous (in fact, differentiable, being a composition of differentiable maps), and therefore admits a maximum when restricted to the unit sphere S^{n-1} by compactness of the sphere. Denote by $\boldsymbol{u}$ a point of the sphere at which f takes on this maximum value. We claim that $\boldsymbol{u}$ is an eigenvalue of L. To see this, it is enough to show that $L\boldsymbol{u}$ is a multiple of $\boldsymbol{u}$, or equivalently, that $L\boldsymbol{u}$ is orthogonal to $\boldsymbol{v}$ for any $\boldsymbol{v}$ orthogonal to $\boldsymbol{u}$. Notice that we only need to check this for any $\boldsymbol{v}$ of norm 1 orthogonal to $\boldsymbol{u}$. Now, the curve $\boldsymbol{c} : \mathbb{R} \to S^{n-1}$, where $\boldsymbol{c}(t) = \cos t\boldsymbol{u} + \sin t\boldsymbol{v}$, has as image the great circle through $\boldsymbol{u}$ and $\boldsymbol{v}$, and passes through $\boldsymbol{u}$ at 0. Since $f \circ \boldsymbol{c}$ is differentiable and has a maximum at 0, its derivative at 0 vanishes. But by the chain rule together with Corollary 2.2.1 and the fact that the derivative of an operator is the operator itself,

$$\begin{aligned} (f \circ \boldsymbol{c})'(t) = \langle L \circ \boldsymbol{c}, \boldsymbol{c} \rangle'(t) &= \langle (L \circ \boldsymbol{c})'(t), \boldsymbol{c}(t) \rangle + \langle L(\boldsymbol{c}(t)), \boldsymbol{c}'(t) \rangle \\ &= \langle L(\boldsymbol{c}'(t)), \boldsymbol{c}(t) \rangle + \langle L(\boldsymbol{c}(t)), \boldsymbol{c}'(t) \rangle = 2\langle L(\boldsymbol{c}(t)), \boldsymbol{c}'(t) \rangle. \end{aligned}$$

Thus,

$$0 = (f \circ \boldsymbol{c})'(0) = 2\langle L(\boldsymbol{c}(0)), \boldsymbol{c}'(0) \rangle = 2\langle L\boldsymbol{u}, \boldsymbol{v} \rangle,$$

which establishes the theorem in the case when V is Euclidean space. In the general case, consider an orthonormal basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ of V. The map $T : V \to \mathbb{R}^n$ that maps a vector to its coordinate vector in this basis is then a linear isometry, and as remarked

earlier, $T \circ L \circ T^{-1}$ is now a self-adjoint operator on Euclidean space. If $\boldsymbol{u} \in \mathbb{R}^n$ is a λ-eigenvector of $T \circ L \circ T^{-1}$, then $\boldsymbol{v} = T^{-1}\boldsymbol{u}$ is a λ-eigenvector of L, because

$$L\boldsymbol{v} = L \circ T^{-1}\boldsymbol{u} = T^{-1}(T \circ L \circ T^{-1}\boldsymbol{u}) = T^{-1}(\lambda \boldsymbol{u}) = \lambda T^{-1}\boldsymbol{u} = \lambda \boldsymbol{v}. \qquad \square$$

Theorem 2.6.2 (Spectral theorem). *Let L be a self-adjoint operator on an inner product space V. Then V has an orthonormal basis consisting of eigenvectors of L. The matrix of L with respect to this basis is diagonal, with the diagonal entries being the eigenvalues of L.*

Proof. The argument will be by induction on the dimension n of V. If $n = 1$, then any unit vector in V is an eigenvector and forms an orthonormal basis. Assume the statement holds for any space of dimension smaller than n, and consider an n-dimensional inner product space V with self-adjoint operator L. By Theorem 2.6.1, L admits an eigenvalue λ with corresponding eigenvector $\boldsymbol{u}$, which may be assumed to have norm 1. Observe that L restricts to an operator on the subspace $\boldsymbol{u}^\perp$ of V consisting of all vectors orthogonal to $\boldsymbol{u}$; i.e., $L(\boldsymbol{u}^\perp) \subset \boldsymbol{u}^\perp$: indeed, if $\boldsymbol{v} \in \boldsymbol{u}^\perp$, then

$$\langle L\boldsymbol{v}, \boldsymbol{u}\rangle = \langle \boldsymbol{v}, L\boldsymbol{u}\rangle = \langle \boldsymbol{v}, \lambda \boldsymbol{u}\rangle = \lambda \langle \boldsymbol{v}, \boldsymbol{u}\rangle = 0,$$

so that $L\boldsymbol{v} \in \boldsymbol{u}^\perp$ as claimed. Since L remains self-adjoint when restricted to $\boldsymbol{u}^\perp$, the induction hypothesis asserts that this subspace admits an orthonormal basis of eigenvectors. Adding $\boldsymbol{u}$ to this set yields an orthonormal basis of V that consists of eigenvectors.

The last statement of the theorem is clear: if $\mathcal{B} = \boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ is a basis of V with $L\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$, then $[L\boldsymbol{v}_i]_{\mathcal{B}}$ is the vector in $\mathbb{R}^n$ with λ_i in the i-th entry and 0 elsewhere, and by definition, the matrix of L in $\mathcal{B}$ is

$$\begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix}. \qquad \square$$

The above theorem does not provide a constructive way of finding eigenvalues and eigenvectors of a self-adjoint operator L. However, $\boldsymbol{u}$ is a λ-eigenvector of L iff $L\boldsymbol{u} = \lambda \boldsymbol{u}$; this is equivalent to the equation

$$(L - \lambda 1_V)\boldsymbol{u} = \boldsymbol{0}.$$

In other words, to determine the eigenvalues of L, it suffices to determine those values of λ for which the operator $L - \lambda 1_V$ has nontrivial kernel. The nonzero elements in that kernel then correspond to the eigenvectors. Notice that $L - \lambda 1_V$ has nontrivial kernel if and only if its determinant is zero. We illustrate the procedure in the following:

Example 2.6.1. Consider the operator L on $\mathbb{R}^3$, where

$$L\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x+z \\ y+z \\ x+y+z \end{bmatrix}.$$

Thus, L is left multiplication by the matrix

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

and since A is symmetric, L is self-adjoint by Proposition 2.6.1. Now, expanding along the first row, we obtain

$$\begin{aligned} \det(L - \lambda 1_{\mathbb{R}^3}) = \det(A - \lambda I_3) &= \det \begin{bmatrix} 1-\lambda & 0 & 1 \\ 0 & 1-\lambda & 1 \\ 1 & 1 & 1-\lambda \end{bmatrix} \\ &= (1-\lambda)\left((1-\lambda)^2 - 1\right) - (1-\lambda) = (1-\lambda)\left((1-\lambda)^2 - 2\right) \\ &= (1-\lambda)(1-\lambda-\sqrt{2})(1-\lambda+\sqrt{2}). \end{aligned}$$

Setting this determinant equal to zero implies that L has 1, $1-\sqrt{2}$, and $1+\sqrt{2}$ as eigenvalues. To find an eigenvector corresponding to the eigenvalue 1, we must find a nontrivial vector in the kernel of $L - \lambda 1_{\mathbb{R}^3}$ with $\lambda = 1$. The equation $(L - 1_{\mathbb{R}^3})\boldsymbol{u} = \mathbf{0}$ is equivalent to $(A - I_3)\boldsymbol{u} = \mathbf{0}$; i.e., to

$$(A - I_3)\boldsymbol{u} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Thus, $x + y = z = 0$, and the 1-eigenvectors are the nonzero multiples of $\begin{bmatrix} 1 & -1 & 0 \end{bmatrix}^T$.

The same procedure works for $\lambda = 1 - \sqrt{2}$:

$$(A - (1-\sqrt{2})I_3)\boldsymbol{u} = \begin{bmatrix} \sqrt{2} & 0 & 1 \\ 0 & \sqrt{2} & 1 \\ 1 & 1 & \sqrt{2} \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

which yields $\sqrt{2}x + z = 0$, $\sqrt{2}y + z = 0$, and $x + y + \sqrt{2}z = 0$. If one multiplies the first two equations by $1/\sqrt{2}$ and adds them, the resulting equation is the third one, so the latter may be discarded. The solutions to the first two are all multiples of $\begin{bmatrix} 1 & 1 & -\sqrt{2} \end{bmatrix}^T$. Repeating this procedure with $\lambda = 1 + \sqrt{2}$ yields all nontrivial multiples of $\begin{bmatrix} 1 & 1 & \sqrt{2} \end{bmatrix}^T$ as eigenvectors.

The polynomial $\det(A - \lambda I_n)$ whose roots are the eigenvalues of A is called the *characteristic polynomial* of A.

Corollary 2.6.1. *Any symmetric matrix A is conjugate to a diagonal matrix D; i.e., there exists an invertible matrix P such that $D = P^{-1}AP$. In fact, P may be chosen to be orthogonal.*

Proof. As remarked already in the previous example, Proposition 2.6.1 applied to $\mathbb{R}^n$ with its standard basis $\mathcal{S}$ and left multiplication L_A by A implies that the operator L_A is self-adjoint. By the spectral theorem, there exists an orthonormal basis $\mathcal{B}$ of eigenvectors of L_A, and in that basis, the matrix $[L_A]_{\mathcal{B}}$ of L_A is some diagonal matrix D. (1.2.1) now says that the matrices of L_A in the two bases satisfy

$$D = [L_A]_{\mathcal{B}} = P^{-1}[L_A]_{\mathcal{S}}P = P^{-1}AP,$$

where $P = [1_{\mathbb{R}^n}]_{\mathcal{B},\mathcal{S}}$ is the change of basis matrix from $\mathcal{B}$ to $\mathcal{S}$. This means that P has as columns the vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$ of $\mathcal{B}$, and thus, the (i,j)-th entry of P^TP is $\boldsymbol{u}_i^T\boldsymbol{u}_j = \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = \delta_{ij}$. In other words P is orthogonal. □

Example 2.6.2. Consider the matrix A from Example 2.6.1. Normalizing the eigenvectors of L_A so they have length 1, the argument from the above corollary implies that

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1-\sqrt{2} & 0 \\ 0 & 0 & 1+\sqrt{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{-\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} \end{bmatrix} A \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{2} & \frac{1}{2} \\ \frac{-1}{\sqrt{2}} & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{-\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Observe that the order in which the eigenvalues appear in the diagonal matrix is the same as the order in which the eigenvectors appear in the last matrix on the right.

Definition 2.6.4. A *symmetric bilinear form* or *scalar product* on a vector space V is a map $b : V \times V \to \mathbb{R}$ such that $b(\boldsymbol{u}, \boldsymbol{v}) = b(\boldsymbol{v}, \boldsymbol{u})$ and $b(a\boldsymbol{u} + \boldsymbol{v}, \boldsymbol{w}) = a\, b(\boldsymbol{u}, \boldsymbol{w}) + b(\boldsymbol{v}, \boldsymbol{w})$ for all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$ and $a \in \mathbb{R}$.

Notice that an inner product is just a symmetric bilinear form that is *positive definite*; i.e., $b(\boldsymbol{u}, \boldsymbol{u}) > 0$ if $\boldsymbol{u} \neq \boldsymbol{0}$. b is said to be *negative definite* if $-b$ is positive definite. A bilinear form that is either positive definite or negative definite is said to be *definite*.

It is quite easy to construct examples of scalar products:

Example 2.6.3. Let $\mathcal{B} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ denote a basis of a vector space V. Any symmetric matrix A induces a symmetric bilinear form b on V by defining

$$b(\boldsymbol{u}, \boldsymbol{v}) = [\boldsymbol{u}]_{\mathcal{B}}^T A [\boldsymbol{v}]_{\mathcal{B}}, \qquad \boldsymbol{u}, \boldsymbol{v} \in V,$$

where as usual we identify the 1×1 matrix $[\boldsymbol{u}]_{\mathcal{B}}^T A [\boldsymbol{v}]_{\mathcal{B}}$ on the right with its single entry. The fact that b as defined above is bilinear follows immediately from properties of matrix operations. Symmetry in turn follows from the fact that a 1×1 matrix is alway symmetric, so that

$$b(\boldsymbol{u}, \boldsymbol{v}) = [\boldsymbol{u}]_{\mathcal{B}}^T A [\boldsymbol{v}]_{\mathcal{B}} = \left([\boldsymbol{u}]_{\mathcal{B}}^T A [\boldsymbol{v}]_{\mathcal{B}}\right)^T = [\boldsymbol{v}]_{\mathcal{B}}^T A^T [\boldsymbol{u}]_{\mathcal{B}}^{TT} = [\boldsymbol{v}]_{\mathcal{B}}^T A [\boldsymbol{u}]_{\mathcal{B}} = b(\boldsymbol{v}, \boldsymbol{u}).$$

It turns out that the above is essentially the only example possible:

Theorem 2.6.3. *Let b denote a symmetric bilinear form on V. Given any basis $\mathcal{B}$ of V, there exists a unique symmetric matrix A such that $b(\boldsymbol{u},\boldsymbol{v}) = [\boldsymbol{u}]_{\mathcal{B}}^T A[\boldsymbol{v}]_{\mathcal{B}}$ for all $\boldsymbol{u},\boldsymbol{v} \in V$.*

Proof. Uniqueness is clear, because if $\boldsymbol{v}_1,\dots,\boldsymbol{v}_n$ are the (ordered) basis vectors, then the above formula implies that

$$b(\boldsymbol{v}_i,\boldsymbol{v}_j) = [\boldsymbol{v}_i]_{\mathcal{B}}^T A[\boldsymbol{v}_j]_{\mathcal{B}} = \boldsymbol{e}_i^T A\boldsymbol{e}_j = a_{ij},$$

which also shows symmetry of A. To check existence, define the (i,j)-th entry of A to be $b(\boldsymbol{v}_i,\boldsymbol{v}_j)$. By linearity of b in each component, given $\boldsymbol{u} = \sum_i x_i\boldsymbol{v}_i$ and $\boldsymbol{v} = \sum_j y_j\boldsymbol{v}_j$,

$$\begin{aligned} b(\boldsymbol{u},\boldsymbol{v}) &= b\left(\sum_i x_i\boldsymbol{v}_i, \sum_j y_j\boldsymbol{v}_j\right) = \sum_{i,j} x_i b(\boldsymbol{v}_i,\boldsymbol{v}_j)y_j = \sum_{i,j} x_i a_{ij} y_j \\ &= [\boldsymbol{u}]_{\mathcal{B}}^T A[\boldsymbol{v}]_{\mathcal{B}}. \end{aligned}$$

□

The theorem above can also be proved in a more elegant way, by observing that there is a one-to-one correspondence between scalar products and self-adjoint operators on an inner product space: If L is self-adjoint, then the formula

$$b(\boldsymbol{u},\boldsymbol{v}) = \langle L\boldsymbol{u},\boldsymbol{v}\rangle \tag{2.6.1}$$

defines a scalar product. Conversely, given a symmetric bilinear form b on an inner product space V, define $L : V \to V$ as follows: for $\boldsymbol{u} \in V$, the map $\boldsymbol{v} \mapsto b(\boldsymbol{u},\boldsymbol{v})$ defines an element of the dual space V^*, so that by Corollary 1.4.2, there is unique element $L\boldsymbol{u} \in V$ satisfying (2.6.1). The fact that L is linear and self-adjoint is immediate. The theorem now follows by observing that

$$[\boldsymbol{u}]_{\mathcal{B}}^T A[\boldsymbol{v}]_{\mathcal{B}} = \langle \boldsymbol{u}, L\boldsymbol{v}\rangle.$$

The spectral theorem yields a convenient way of expressing scalar products: Select any basis $\mathcal{B}$ of V to obtain a matrix A such that $b(\boldsymbol{u},\boldsymbol{v}) = [\boldsymbol{u}]_{\mathcal{B}}^T A[\boldsymbol{v}]_{\mathcal{B}}$ as in the theorem. Since A is symmetric, there exists, by Corollary 2.6.1, a diagonal matrix D and an orthogonal matrix Q (set Q equal to P^T in the corollary) such that $A = Q^T DQ$. Since Q is invertible, the vectors $\boldsymbol{u}_i = \sum_j q_{ij}\boldsymbol{v}_j$, $i = 1,\dots,n$, form a basis $\mathcal{C}$ of V, and by the change of basis formula,

$$[\boldsymbol{u}]_{\mathcal{C}} = Q[\boldsymbol{u}]_{\mathcal{B}}, \qquad \boldsymbol{u} \in V.$$

Thus, if $\boldsymbol{u} = \sum_i x_i\boldsymbol{u}_i$, $\boldsymbol{v} = \sum_j y_j\boldsymbol{u}_j$, and D is the diagonal matrix with entries $\lambda_1,\dots,\lambda_n$, then

$$\begin{aligned} b(\boldsymbol{u},\boldsymbol{v}) &= [\boldsymbol{u}]_{\mathcal{B}}^T A[\boldsymbol{v}]_{\mathcal{B}} = [\boldsymbol{u}]_{\mathcal{B}}^T Q^T DQ[\boldsymbol{v}]_{\mathcal{B}} = (Q[\boldsymbol{u}]_{\mathcal{B}})^T D(Q[\boldsymbol{v}]_{\mathcal{B}}) = [\boldsymbol{u}]_{\mathcal{C}}^T D[\boldsymbol{v}]_{\mathcal{C}} \\ &= \sum_i \lambda_i x_i y_i. \end{aligned}$$

In particular,

$$b(\boldsymbol{u},\boldsymbol{u}) = \sum_i \lambda_i x_i^2,$$

so that b is positive-(respectively negative-)definite if and only if all the eigenvalues of A are positive (respectively) negative.

Example 2.6.4. Consider the symmetric bilinear form on $\mathbb{R}^2$ given by

$$b(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}) = 3x_1y_1 - x_1y_2 - x_2y_1 + 3x_2y_2.$$

The matrix of b in the standard basis is

$$A = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}.$$

Its characteristic polynomial is

$$\det(A - \lambda I_2) = (3-\lambda)^2 - 1 = (4-\lambda)(2-\lambda),$$

which has roots 4 and 2. Thus, b is positive definite. Using the same procedure as that in the previous example, one easily computes that $\boldsymbol{v}_1 = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}^T$ and $\boldsymbol{v}_2 = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}^T$ are unit eigenvectors corresponding to the eigenvalues 2 and 4 respectively. If $\boldsymbol{u} = x_1\boldsymbol{v}_1 + x_2\boldsymbol{v}_2$ and $\boldsymbol{v} = y_1\boldsymbol{v}_1 + y_2\boldsymbol{v}_2$, then $b(\boldsymbol{u},\boldsymbol{v}) = 2x_1y_1 + 4x_2y_2$.

2.7 Taylor polynomials and extreme values

A point $\boldsymbol{a} \in U \subset \mathbb{R}^n$ is said to be a *critical point* of a map $\boldsymbol{f} : U \to \mathbb{R}^m$ if $D\boldsymbol{f}(\boldsymbol{a}) : \mathbb{R}^n \to \mathbb{R}^m$ is either not onto or does not exist. In this case, $\boldsymbol{f}(\boldsymbol{a})$ is called a *critical value* of $\boldsymbol{f}$. Notice that if $m = 1$, then $\boldsymbol{a}$ is critical if and only if $Df(\boldsymbol{a}) = 0$ or does not exist. If in addition, $n = 1$, we recover the classical notion of critical point from one-variable Calculus. In the case of a real-valued function $f : U \to \mathbb{R}$, f is said to have a local *maximum* (resp. *minimum*) at $\boldsymbol{a}$ if there exists a neighborhood V of $\boldsymbol{a}$ such that $f(\boldsymbol{p}) \leq$ (resp. $\geq$)$f(\boldsymbol{a})$ for all $\boldsymbol{p} \in V$. A point $\boldsymbol{a}$ where f has a local maximum or minimum is called an *extremum*, and $f(\boldsymbol{a})$ is then said to be an *extreme value*.

Theorem 2.7.1. *If $f : U \subset \mathbb{R}^n \to \mathbb{R}$ has an extremum at $\boldsymbol{a}$, then $\boldsymbol{a}$ is a critical point of f.*

Proof. Let $\boldsymbol{a}$ be an extremum of f. We may assume that f is differentiable at $\boldsymbol{a}$, for otherwise the conclusion certainly holds. In particular, f is defined on some open neighborhood of $\boldsymbol{a}$. We wish to show that for any $\boldsymbol{h} \in \mathbb{R}^n$, $Df(\boldsymbol{a})\boldsymbol{h} = \boldsymbol{0}$. Choose a small enough interval I around 0 such that the curve $\boldsymbol{c} : I \to \mathbb{R}^n$, where $\boldsymbol{c}(t) = \boldsymbol{a} + t\boldsymbol{h}$, has its image inside U. Then the function $g := f \circ \boldsymbol{c} : I \to \mathbb{R}$ of one variable has an extremum at 0, and being differentiable there,

$$0 = g'(0) = (f \circ \boldsymbol{c})'(0) = Df(\boldsymbol{c}(0))\boldsymbol{c}'(0) = Df(\boldsymbol{a})\boldsymbol{h}. \qquad \square$$

The converse is already false when $n = 1$, as the function $x \mapsto x^3$ shows: 0 is a critical point of f, but not an extremum. Recall that for functions of one variable, if $f'(a) = 0$, then a is a local minimum (resp. maximum) if $f''(a) > 0$ (resp. < 0). Now, a function defined on $\mathbb{R}^n$ does not have one second derivative if $n > 1$, but it does have an $n \times n$ matrix $[D_{ij}f(\boldsymbol{a})]$ of second partial derivatives, which is symmetric if f is $\mathcal{C}^2$. We will see that if the corresponding symmetric bilinear form is positive (resp. negative) definite, then f has a local minimum (resp. maximum) at a. In order to show this, we first generalize Taylor polynomials to functions defined on Euclidean space.

Definition 2.7.1. Let $f : U \to \mathbb{R}$ be a function of class $\mathcal{C}^k$ on an open set $U \subset \mathbb{R}^n$. The *Taylor polynomial of degree k* of f at $\boldsymbol{a} \in U$ is the function $P_{f,\boldsymbol{a},k} : \mathbb{R}^n \to \mathbb{R}$ given by

$$P_{f,\boldsymbol{a},k} = \sum_{j=0}^{k} \frac{1}{j!} \sum_{1 \le i_1,\dots,i_j \le n} D_{i_1 \cdots i_j} f(\boldsymbol{a}) u^{i_1} u^{i_2} \cdots u^{i_j}.$$

Thus, for $\boldsymbol{p} = (x_1, \dots, x_n) \in \mathbb{R}^n$,

$$P_{f,\boldsymbol{a},k}(\boldsymbol{p}) = \sum_{j=0}^{k} \frac{1}{j!} \sum_{1 \le i_1,\dots,i_j \le n} D_{i_1 \cdots i_j} f(\boldsymbol{a}) x_{i_1} x_{i_2} \cdots x_{i_j},$$

with $D_0 f = f$.

Example 2.7.1. If $f : \mathbb{R}^2 \to \mathbb{R}$ is defined by $f(x, y) = e^{xy}$, then $D_1 f(x, y) = ye^{xy}$, $D_2 f(x, y) = xe^{xy}$, $D_{12} f(x, y) = D_{21} f(x, y) = (1 + xy)e^{xy}$, $D_{11} f(x, y) = y^2 exy$, and $D_{22} f(x, y) = x^2 e^{xy}$. Thus the Taylor polynomial of degree 2 of f at $\mathbf{0}$ is

$$P_{f,\mathbf{0},2}(x, y) = 1 + xy.$$

Theorem 2.7.2 (Taylor's theorem). *Suppose $f : U \to \mathbb{R}$ is of class $\mathcal{C}^{k+1}$ on the open set $U \subset \mathbb{R}^n$, and let $\boldsymbol{a} \in U$, $\varepsilon > 0$ small enough so that the closed ball of radius ε around $\boldsymbol{a}$ is contained in U. Define a function r on this ball by setting*

$$r(\boldsymbol{p}) = f(\boldsymbol{a} + \boldsymbol{p}) - P_{f,\boldsymbol{a},k}(\boldsymbol{p}).$$

Then the remainder *r satisfies*

$$\lim_{\boldsymbol{p} \to \boldsymbol{0}} \frac{r(\boldsymbol{p})}{|\boldsymbol{p}|^k} = 0.$$

Proof. Fix any $\boldsymbol{p} = (x_1, \dots, x_n) \in \mathbb{R}^n$ of norm less than ε. The curve $\boldsymbol{c} : [0, 1] \to \mathbb{R}^n$, $\boldsymbol{c}(t) = \boldsymbol{a} + t\boldsymbol{p}$, has its image in U, so we have a well-defined function $g := f \circ \boldsymbol{c}$ of class $\mathcal{C}^{k+1}$. By the chain rule,

$$g'(t) = Df(\boldsymbol{c}(t))\boldsymbol{c}'(t) = \sum_{i=1}^{n} D_i f(\boldsymbol{c}(t)) x_i,$$

and arguing inductively, we obtain

$$g^{(j)}(t) = \sum_{1 \le i_1,\dots,i_j \le n} D_{i_1 \cdots i_j} f(\boldsymbol{c}(t)) x_{i_1} \cdots x_{i_j}$$

for $j = 1, \dots, k+1$. Taylor's theorem for functions of one variable asserts that there exists some $t \in (0,1)$ such that

$$g(1) = \sum_{j=0}^{k} \frac{g^{(j)}(0)}{j!} + \frac{g^{(k+1)}(t)}{(k+1)!};$$

i.e.,

$$\begin{aligned} f(\boldsymbol{a}+\boldsymbol{p}) &= \sum_{j=0}^{k} \frac{1}{j!} \sum_{1 \le i_1, \dots, i_j \le n} D_{i_1 \cdots i_j} f(\boldsymbol{a}) x_{i_1} \cdots x_{i_j} + \frac{g^{(k+1)}(t)}{(k+1)!} \\ &= P_{f,\boldsymbol{a},k}(\boldsymbol{p}) + r(\boldsymbol{p}), \end{aligned}$$

where

$$r(\boldsymbol{p}) = \frac{1}{(k+1)!} \sum D_{i_1 \cdots i_{k+1}} f(\boldsymbol{a}+t\boldsymbol{p}) x_{i_1} \cdots x_{i_{k+1}},$$

and $t = t(\boldsymbol{p})$ depends on $\boldsymbol{p}$. It only remains to show that $r(\boldsymbol{p})/|\boldsymbol{p}|^k \to 0$ as $\boldsymbol{p} \to \boldsymbol{0}$. But f is of class $\mathcal{C}^{k+1}$, so that each $D_{i_1 \cdots i_{k+1}} f(\boldsymbol{a}+t\boldsymbol{p})$ converges to the value of that function at $\boldsymbol{a}$. Furthermore, each $|x_i| \le |\boldsymbol{p}|$, so that

$$\frac{|x_{i_1} \cdots x_{i_{k+1}}|}{|\boldsymbol{p}|^k} \le \min\{|x_{i_1}|, \dots, |x_{i_{k+1}}|\}, \tag{2.7.1}$$

and the right side goes to 0 as $\boldsymbol{p} \to \boldsymbol{0}$. This proves the assertion. □

Remark 2.7.1. Since the right side of the inequality 2.7.1 is no larger than $|\boldsymbol{p}|$, for each sufficiently small neighborhood V of $\boldsymbol{a}$, there exists a constant α that depends on f and V such that $|r(\boldsymbol{p})| < \alpha |\boldsymbol{p}|^{k+1}$ for all $\boldsymbol{p}$ such that $\boldsymbol{a}+\boldsymbol{p} \in V$. Furthermore, if C is any compact set in U, uniform continuity of the derivatives of order $k+1$ of f on C implies that there exists $\beta > 0$ that depends on C and f such that

$$|r(\boldsymbol{p}, \boldsymbol{h})| := |f(\boldsymbol{p}+\boldsymbol{h}) - P_{f,\boldsymbol{p},k}(\boldsymbol{h})| \le \beta |\boldsymbol{h}|^{k+1} \text{ for all } \boldsymbol{p}, \boldsymbol{p}+\boldsymbol{h} \in C.$$

Definition 2.7.2. Let $f : U \subset \mathbb{R}^n \to \mathbb{R}$ be of class $\mathcal{C}^2$, and $\boldsymbol{a}$ an interior point of U. The *Hessian* $H_f(\boldsymbol{a})$ of f at $\boldsymbol{a}$ is the symmetric operator on $\mathbb{R}^n$ given by left multiplication by the matrix $[D_{ij} f(\boldsymbol{a})]$ of second derivatives of f at $\boldsymbol{a}$. The *Hessian form* $h_f(\boldsymbol{a})$ of f at $\boldsymbol{a}$ is the associated symmetric bilinear form

$$h_f(\boldsymbol{a})(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^T \begin{bmatrix} D_{11} f(\boldsymbol{a}) & \cdots & D_{1n} f(\boldsymbol{a}) \\ \vdots & \dots & \vdots \\ D_{n1} f(\boldsymbol{a}) & \cdots & D_{nn} f(\boldsymbol{a}) \end{bmatrix} \boldsymbol{v}, \qquad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n.$$

Observe that $h_f(\boldsymbol{a})(\boldsymbol{u}, \boldsymbol{u})$ equals two times the term corresponding to $j = 2$ in the Taylor polynomial of f at $\boldsymbol{a}$. In particular, if a $\mathcal{C}^2$ function f has a critical point at $\boldsymbol{a}$, then its Taylor polynomial of order 2 at $\boldsymbol{a}$ is

$$P_{f,\boldsymbol{a},2}(\boldsymbol{u}) = f(\boldsymbol{a}) + \frac{1}{2} h_f(\boldsymbol{a})(\boldsymbol{u}, \boldsymbol{u}).$$

This fact is used to prove the following:

Theorem 2.7.3. *Suppose $f : U \subset \mathbb{R}^n \to \mathbb{R}$ is of class $\mathcal{C}^2$, $\boldsymbol{a} \in U$, and $Df(\boldsymbol{a}) = 0$.*

(1) *If the Hessian form $h_f(\boldsymbol{a})$ of f at $\boldsymbol{a}$ is positive (resp. negative) definite, then f has a local minimum (resp. maximum) at $\boldsymbol{a}$.*

(2) *If the Hessian $H_f(\boldsymbol{a})$ has at least one positive and one negative eigenvalues, then f does not have a local maximum nor a local minimum at $\boldsymbol{a}$. f is then said to have a* saddle point *at $\boldsymbol{a}$.*

Proof. (1): We will consider the case when the Hessian is positive definite. The other case follows by looking at $-f$. By assumption, the function $\boldsymbol{u} \mapsto h_f(\boldsymbol{a})(\boldsymbol{u}, \boldsymbol{u})$ assumes a positive minimum value 2α when restricted to the unit sphere S^{n-1}, since the latter is compact. Let $0 < \varepsilon < \alpha$. By Taylor's theorem, there is $\delta > 0$ such that for $0 < |\boldsymbol{u}| < \delta$,

$$\frac{|r(\boldsymbol{u})|}{|\boldsymbol{u}|^2} = \left| \frac{f(\boldsymbol{a}+\boldsymbol{u}) - f(\boldsymbol{a})}{|\boldsymbol{u}|^2} - \frac{1}{2} h_f(\boldsymbol{a}) \left(\frac{\boldsymbol{u}}{|\boldsymbol{u}|}, \frac{\boldsymbol{u}}{|\boldsymbol{u}|} \right) \right| < \varepsilon. \tag{2.7.2}$$

Thus, for sufficiently small $|\boldsymbol{u}| > 0$,

$$\frac{f(\boldsymbol{a}+\boldsymbol{u}) - f(\boldsymbol{a})}{|\boldsymbol{u}|^2} > \frac{1}{2} h_f(\boldsymbol{a}) \left(\frac{\boldsymbol{u}}{|\boldsymbol{u}|}, \frac{\boldsymbol{u}}{|\boldsymbol{u}|} \right) - \varepsilon \geq \alpha - \varepsilon > 0,$$

and $f(\boldsymbol{a}+\boldsymbol{u}) > f(\boldsymbol{a})$ as claimed.

(2): By assumption, $H_f(\boldsymbol{a})$ has some positive eigenvalue 2α. We will show that if $\boldsymbol{w}$ is a corresponding eigenvector with small enough norm, then $f(\boldsymbol{a}+\boldsymbol{w}) > f(\boldsymbol{a})$. It then follows that f does not have a local maximum at $\boldsymbol{a}$. So choose some $0 < \varepsilon < \alpha$. By Taylor's theorem, there exists $\delta > 0$ such that (2.7.2) holds for any $\boldsymbol{u}$ such that $0 < |\boldsymbol{u}| < \delta$. Now, if $\boldsymbol{w}$ is a 2α-eigenvector of the Hessian, then

$$h_f(\boldsymbol{a})\left(\frac{\boldsymbol{w}}{|\boldsymbol{w}|}, \frac{\boldsymbol{w}}{|\boldsymbol{w}|}\right) = \frac{1}{|\boldsymbol{w}|^2} \boldsymbol{w}^T H_f(\boldsymbol{a}) \boldsymbol{w} = \frac{1}{|\boldsymbol{w}|^2} \boldsymbol{w}^T 2\alpha \boldsymbol{w} = 2\alpha,$$

so that for $0 < |\boldsymbol{w}| < \delta$,

$$\frac{f(\boldsymbol{a}+\boldsymbol{w}) - f(\boldsymbol{a})}{|\boldsymbol{w}|^2} > \frac{1}{2} h_f(\boldsymbol{a}) \left(\frac{\boldsymbol{w}}{|\boldsymbol{w}|}, \frac{\boldsymbol{w}}{|\boldsymbol{w}|} \right) - \varepsilon = \alpha - \varepsilon > 0.$$

The same argument applied to $-f$ and any negative eigenvalue of $H_f(\boldsymbol{a})$ implies that f does not have a local minimum at $\boldsymbol{a}$. □

Examples 2.7.2. (i) Suppose we are asked to find and classify the critical points of $f : \mathbb{R}^2 \to \mathbb{R}, f(x, y) = x^2 + y^2 - 2x^2 y$. The Jacobian matrix of f at (x, y) is

$$[Df](x, y) = \begin{bmatrix} 2x - 4xy & 2y - 2x^2 \end{bmatrix} = 2 \begin{bmatrix} x(1-2y) & y - x^2 \end{bmatrix},$$

and (x, y) is a critical point of f if and only if $x = 0$ or $y = 1/2$, and $y = x^2$. Thus, there are 3 critical points: the origin, and $(\pm 1/\sqrt{2}, 1/2)$. The Hessian of f has as matrix

$$\begin{bmatrix} 2 - 4y & -4x \\ -4x & 2 \end{bmatrix}.$$

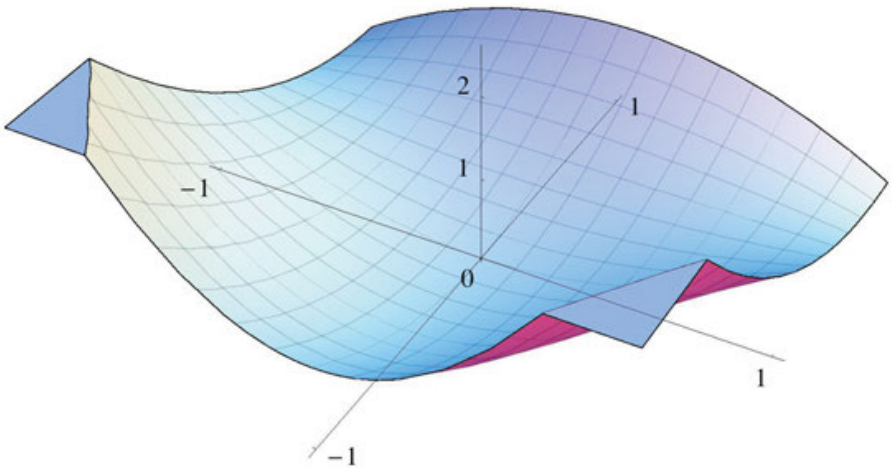


Fig. 2.1: $z = x^2 + y^2 - 2x^2y$

At the origin, the matrix equals $2I_2$, so f has a local minimum there. At the other two points, the Hessian becomes

$$\begin{bmatrix} 0 & \frac{4}{\sqrt{2}} \\ \frac{4}{\sqrt{2}} & 2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & -\frac{4}{\sqrt{2}} \\ -\frac{4}{\sqrt{2}} & 2 \end{bmatrix}.$$

The eigenvalue-finding procedure outlined in the previous section easily yields the same eigenvalues 4 and -2 for both matrices. Thus, both points are saddle points.

(ii) The classification of critical points in the previous example could also have been done without explicitly finding the eigenvalues of the Hessian. This is because for a 2×2 symmetric matrix A, the sign of the eigenvalues is easily determined: since there is a diagonal matrix of eigenvalues that is similar to A, the determinant of A is the product of the eigenvalues, and the trace equals the sum (this is of course true for matrices of any size). Thus, if $\det A < 0$, then the eigenvalues have different signs, corresponding to a saddle point. If the determinant is positive, both eigenvalues have the same sign, which coincides with the sign of the $(1, 1)$ element of the matrix: the reason being that both diagonal terms must have the same sign (otherwise the determinant would not be positive) and their sum equals the sum of the eigenvalues.

Summarizing, if f is a function of two variables which has $\boldsymbol{a}$ as a critical point, and if

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

denotes the Hessian matrix at $\boldsymbol{a}$, then

- f has a saddle point at $\boldsymbol{a}$ if $\det A < 0$;
- f has a local minimum at $\boldsymbol{a}$ if $\det A > 0$ and $a > 0$;
- f has a local maximum at $\boldsymbol{a}$ if $\det A > 0$ and $a < 0$;

Notice that if the determinant is positive, then a cannot vanish. This means that the above covers all but one possibility, namely the determinant being zero. If this is the case, then no conclusion can be made: for example, the function f given by $f(x,y) = x^2y^2$ has a global minimum at the origin but its Hessian there is zero. Taking its negative yields a function with a maximum at the origin with zero Hessian. Finally, f, where $f(x,y) = x^3$, has the whole y-axis as critical points with vanishing Hessian, but none of them are maxima or minima.

2.8 Vector fields

From now on, differentiable maps will be assumed to be smooth (i.e., $\mathcal{C}^\infty$) unless specified otherwise.

One tends to visualize vectors in Euclidean space as arrows whose base point can be assigned arbitrarily. When dealing with vector fields, it is more efficient to have fixed base points. In order to specify a vector with base point at, say, $\boldsymbol{p}$, we use the following concept:

Definition 2.8.1.

The *tangent space* of $\mathbb{R}^n$ at $\boldsymbol{p} \in \mathbb{R}^n$ is the collection

$$\mathbb{R}^n_{\boldsymbol{p}} = \{\boldsymbol{p}\} \times \mathbb{R}^n = \{(\boldsymbol{p}, \boldsymbol{u}) \mid \boldsymbol{u} \in \mathbb{R}^n\}.$$

An element of $\mathbb{R}^n_{\boldsymbol{p}}$ is called a *tangent vector* at $\boldsymbol{p}$. $\mathbb{R}^n_{\boldsymbol{p}}$ is a vector space with the operations

$$(\boldsymbol{p}, \boldsymbol{u}) + (\boldsymbol{p}, \boldsymbol{q}) = (\boldsymbol{p}, \boldsymbol{u} + \boldsymbol{v}), \quad a(\boldsymbol{p}, \boldsymbol{u}) = (\boldsymbol{p}, a\boldsymbol{u}), \qquad a \in \mathbb{R}, \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n.$$

There is a canonical isomorphism $\mathcal{I}_{\boldsymbol{p}} : \mathbb{R}^n \to \mathbb{R}^n_{\boldsymbol{p}}$ that maps $\boldsymbol{u}$ to $(\boldsymbol{p}, \boldsymbol{u})$, $\boldsymbol{u} \in \mathbb{R}^n$. For any $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^n$, the map

$$\begin{aligned} \mathcal{I}_{\boldsymbol{q}} \circ \mathcal{I}_{\boldsymbol{p}}^{-1} : \mathbb{R}^n_{\boldsymbol{p}} &\to \mathbb{R}^n_{\boldsymbol{q}}, \\ (\boldsymbol{p}, \boldsymbol{u}) &\mapsto (\boldsymbol{q}, \boldsymbol{u}) \end{aligned}$$

is an isomorphism between tangent spaces, called *parallel translation*.

Derivatives can be reformulated in terms of tangent spaces:

Definition 2.8.2. If $\boldsymbol{f} : U \to \mathbb{R}^m$ is a differentiable map on an open set $U \in \mathbb{R}^n$, the *derivative of* $\boldsymbol{f}$ *at* $\boldsymbol{p} \in U$ is the linear transformation

$$\begin{aligned} \boldsymbol{f}_{*\boldsymbol{p}} : \mathbb{R}^n_{\boldsymbol{p}} &\to \mathbb{R}^m_{\boldsymbol{f}(\boldsymbol{p})} \\ (\boldsymbol{p}, \boldsymbol{u}) &\mapsto (\boldsymbol{f}(\boldsymbol{p}), D\boldsymbol{f}(\boldsymbol{p})\boldsymbol{u}). \end{aligned} \tag{2.8.1}$$

It is worth emphasizing that the only difference between $\boldsymbol{f}_*$ and $D\boldsymbol{f}$ is one of notation: the former includes the base point of the vector, the latter doesn't. In terms of

the canonical isomorphism between Euclidean space and its tangent space at $\boldsymbol{p}$, the following diagram commutes:

$$\begin{array}{ccc} \mathbb{R}^n_{\boldsymbol{p}} & \xrightarrow{\boldsymbol{f}_{*\boldsymbol{p}}} & R^m_{\boldsymbol{f}(\boldsymbol{p})} \\ {\scriptstyle \mathcal{I}_{\boldsymbol{p}}}\uparrow & & \uparrow{\scriptstyle \mathcal{I}_{\boldsymbol{f}(\boldsymbol{p})}} \\ U & \xrightarrow[D\boldsymbol{f}(\boldsymbol{p})]{} & \mathbb{R}^m. \end{array}$$

In order to avoid cumbersome notation, we will often denote $(\boldsymbol{p}, \boldsymbol{e}_i)$ by $\boldsymbol{D}_i(\boldsymbol{p})$, and $(t, 1) \in \mathbb{R}_t$ by $D(t)$ when $n = 1$. Thus, the tangent vector $(\boldsymbol{p}, \sum a_i \boldsymbol{e}_i)$ can also be written as $\sum a_i \boldsymbol{D}_i(\boldsymbol{p})$. Furthermore, when it is clear that $\boldsymbol{u}$ belongs to the tangent space of, say, $\boldsymbol{p}$, we often write $\boldsymbol{f}_* \boldsymbol{u}$ instead of $\boldsymbol{f}_{*\boldsymbol{p}} \boldsymbol{u}$ for brevity.

Definition 2.8.3. A *vector field* on an open set $U \subset \mathbb{R}^n$ is a map $\boldsymbol{X}$ that assigns to each $\boldsymbol{p} \in U$ a tangent vector $\boldsymbol{X}(\boldsymbol{p}) \in \mathbb{R}^n_{\boldsymbol{p}}$ at $\boldsymbol{p}$.

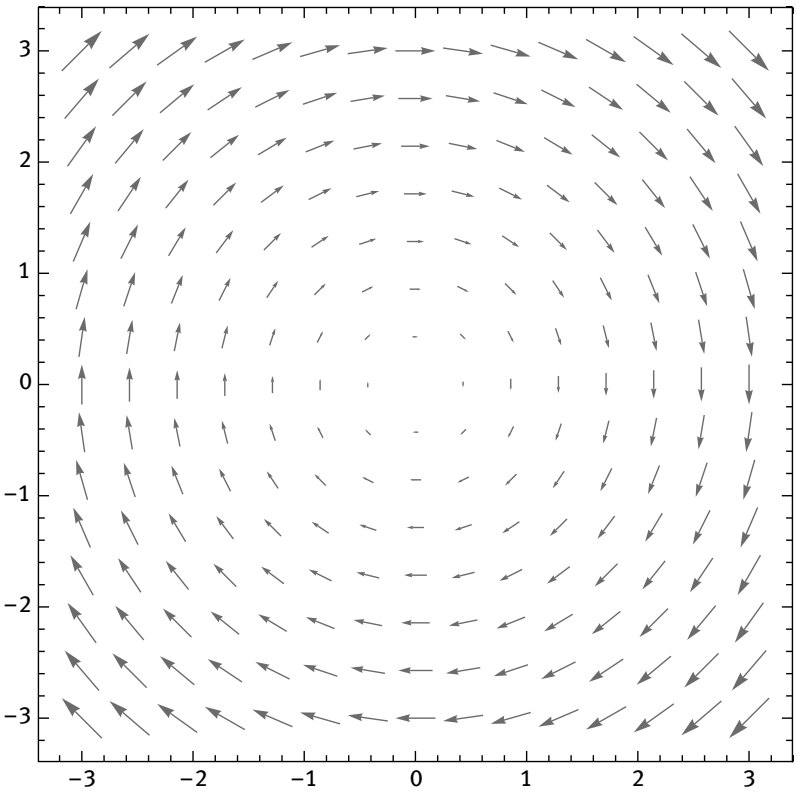


Fig. 2.2: The vector field $\boldsymbol{X} = u^2 \boldsymbol{D}_1 - u^1 \boldsymbol{D}_2$ in the plane

By definition, vector fields on U are in one-one correspondence with maps $\boldsymbol{f} : U \to \mathbb{R}^n$; for each $\boldsymbol{p} \in U$, the value of $\boldsymbol{X}$ at $\boldsymbol{p}$ can be written as $(\boldsymbol{p}, \boldsymbol{f}(\boldsymbol{p}))$ for a unique map $\boldsymbol{f} : U \to \mathbb{R}^n$, and conversely, any such map defines a vector field $\boldsymbol{X}$ by the formula $\boldsymbol{X}(\boldsymbol{p}) = \mathcal{I}_{\boldsymbol{p}} \boldsymbol{f}(\boldsymbol{p})$. We say the vector field is differentiable or smooth if $\boldsymbol{f}$ has that property.

If the vector field $\boldsymbol{X}$ is represented by the map $\boldsymbol{f}$, i.e., if $\boldsymbol{X}(\boldsymbol{p}) = \mathcal{I}_{\boldsymbol{p}}\boldsymbol{f}(\boldsymbol{p})$ for all $\boldsymbol{p}$ in the domain, then

$$\boldsymbol{X} = \sum_i f^i \boldsymbol{D}_i, \qquad f^i = u^i \circ \boldsymbol{f}.$$

For example, the *position vector field* $\boldsymbol{P}$ on $\mathbb{R}^n$, which is defined by $\boldsymbol{P}(\boldsymbol{p}) = \mathcal{I}_{\boldsymbol{p}}\boldsymbol{p}$, can be written $\boldsymbol{P} = \sum_i u^i \boldsymbol{D}_i$.

When $\boldsymbol{f}$ is a constant map, the vector field is said to be *parallel*; equivalently, such a vector field is obtained by parallel translating its value at any one point to every other point.

It is useful to introduce a more general concept:

Definition 2.8.4. Let $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ be a map. A *vector field along* $\boldsymbol{f}$ is a map $\boldsymbol{X}$ that assigns to each $\boldsymbol{p} \in U$ an element $\boldsymbol{X}(\boldsymbol{p}) \in \mathbb{R}^m_{\boldsymbol{f}(\boldsymbol{p})}$ of the tangent space of $\mathbb{R}^m$ at $\boldsymbol{f}(\boldsymbol{p})$. As above, such a field $\boldsymbol{X}$ is represented by a unique map $\boldsymbol{g} : U \to \mathbb{R}^m$, where $\boldsymbol{X}(\boldsymbol{p}) = (\boldsymbol{f}(\boldsymbol{p}), \boldsymbol{g}(\boldsymbol{p}))$, and $\boldsymbol{X}$ is said to be smooth if $\boldsymbol{g}$ is.

A recurring example of such a vector field is the *velocity field* or *tangent field* of a curve $\boldsymbol{c} : I \to \mathbb{R}^n$: by definition it is the vector field $\dot{\boldsymbol{c}}$ along $\boldsymbol{c}$ given by

$$\dot{\boldsymbol{c}}(t) = \mathcal{I}_{\boldsymbol{c}(t)}\boldsymbol{c}'(t), \qquad t \in I.$$

Alternatively,

$$\dot{\boldsymbol{c}}(t) = \boldsymbol{c}_{*t}D(t).$$

For example, if $\boldsymbol{c}(t) = (\cos t, -\sin t)$, then

$$\dot{\boldsymbol{c}}(t) = -\sin t \boldsymbol{D}_1 \circ \boldsymbol{c}(t) - \cos t \boldsymbol{D}_2 \circ \boldsymbol{c}(t) = (u^2\boldsymbol{D}_1 - u^1\boldsymbol{D}_2) \circ \boldsymbol{c}(t).$$

Notice that if $\boldsymbol{X}$ is the vector field from Figure 2.2, then $\dot{\boldsymbol{c}} = \boldsymbol{X} \circ \boldsymbol{c}$; i.e., the "restriction" of $\boldsymbol{X}$ to the image of $\boldsymbol{c}$ is the velocity field of the curve.

Definition 2.8.5. An *integral curve* of a vector field $\boldsymbol{X}$ is a curve $\boldsymbol{c}$ that satisfies $\dot{\boldsymbol{c}} = \boldsymbol{X} \circ \boldsymbol{c}$.

It turns out that for any one point in the domain of a given vector field, there exists an integral curve of that field that passes through the point. This remarkable fact is a consequence of a basic theorem from the theory of ordinary differential equations, a proof of which can be found for example in [9]. The relation between integral curves and differential equations stems from the fact that if $\boldsymbol{X}$ is represented by $\boldsymbol{f} : U \to \mathbb{R}^n$, then $\boldsymbol{c}$ is an integral curve of $\boldsymbol{X}$ if and only if $\boldsymbol{c}' = \boldsymbol{f} \circ \boldsymbol{c}$; in terms of the components $x_i = u^i \circ \boldsymbol{c}$ and $f^i = u^i \circ \boldsymbol{f}$ of $\boldsymbol{c}$ and $\boldsymbol{f}$ respectively, this is equivalent to the system of ordinary differential equations (ODEs)

$$\begin{aligned} x_1'(t) &= f^1(x_1(t), \ldots, x_n(t)) \\ &\vdots \\ x_n'(t) &= f^n(x_1(t), \ldots, x_n(t)) \end{aligned}$$

We state without proof the existence and uniqueness theorems for integral curves of vector fields:

Theorem 2.8.1 (Existence of Solutions). *Let $\boldsymbol{f} : U \to \mathbb{R}^n$ be a differentiable map, where U is open in $\mathbb{R}^n$. For any $\boldsymbol{a} \in U$, there exists a neighborhood W of $\boldsymbol{a}$, an interval I around $0 \in \mathbb{R}$, and a differentiable map $\boldsymbol{\Psi} : I \times W \to U$ such that*
(1) $\boldsymbol{\Psi}(0, \boldsymbol{u}) = \boldsymbol{u}$, *and*
(2) $D\boldsymbol{\Psi}(t, \boldsymbol{u})\boldsymbol{e}_1 = \boldsymbol{f} \circ \Psi(t, \boldsymbol{u})$
for $t \in I$ and $\boldsymbol{u} \in W$.

Recalling that maps $\boldsymbol{f} : U \to \mathbb{R}^n$ are in bijective correspondence with vector fields $\boldsymbol{X}$ defined on U, Theorem 2.8.1 asserts that integral curves $t \mapsto \boldsymbol{c}_{\boldsymbol{u}}(t) := \boldsymbol{\Psi}(t, \boldsymbol{u})$ exist for arbitrary "initial conditions" $\boldsymbol{c}_{\boldsymbol{u}}(0) = \boldsymbol{u}$, that they depend smoothly on the initial conditions, and that at least locally, they can be defined on a fixed common interval. A map $\boldsymbol{\Psi}$ as in the above theorem is called a *local flow* of $\boldsymbol{X}$.

In the same way, uniqueness of solutions of the above system of differential equations implies uniqueness of integral curves:

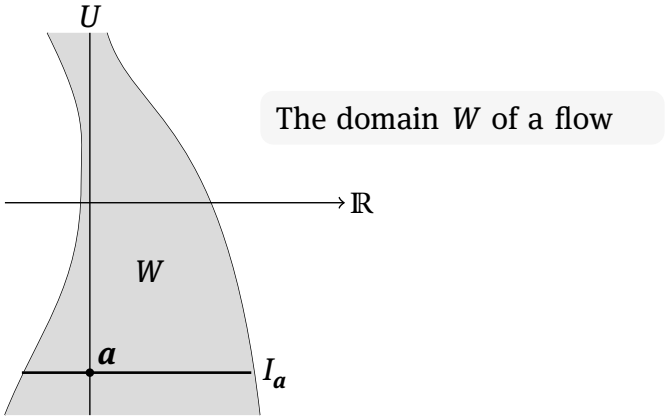
Theorem 2.8.2 (Uniqueness of Solutions). *If $\boldsymbol{c}, \tilde{\boldsymbol{c}} : I \to U$ are two integral curves of a vector field $\boldsymbol{X}$ with $\boldsymbol{c}(t_0) = \tilde{\boldsymbol{c}}(t_0)$ for some $t_0 \in I$, then $\boldsymbol{c} = \tilde{\boldsymbol{c}}$.*

Our next aim is to group all local flows into a single one so that the corresponding integral curves are defined on a maximal interval: For each $\boldsymbol{a} \in U$, let $I_{\boldsymbol{a}}$ denote the maximal open interval around 0 on which the (unique by Theorem 2.8.2) integral curve $\boldsymbol{\Psi}_{\boldsymbol{a}}$ of $\boldsymbol{X}$ that equals $\boldsymbol{a}$ at 0 is defined.

Theorem 2.8.3. *Given any vector field $\boldsymbol{X}$ on $U \subset \mathbb{R}^n$, there exists a unique open set $W \subset \mathbb{R} \times U$ and a unique differentiable map $\boldsymbol{\Psi} : W \to U$ such that*
(1) $I_{\boldsymbol{a}} \times \{\boldsymbol{a}\} = W \cap (\mathbb{R} \times \{\boldsymbol{a}\})$ *for all $\boldsymbol{a} \in U$, and*
(2) $\boldsymbol{\Psi}(t, \boldsymbol{a}) = \boldsymbol{\Psi}_{\boldsymbol{a}}(t)$ *if $(t, \boldsymbol{a}) \in W$.*

$\boldsymbol{\Psi}$ is called the *flow* of $\boldsymbol{X}$. By (2), $\{0\} \times U \subset W$, and (1), (2) of Theorem 2.8.1 are satisfied.

Proof. (1) determines W uniquely, while (2) does the same for $\boldsymbol{\Psi}$. It thus remains to show that W is open, and that $\boldsymbol{\Psi}$ is differentiable.



The domain W of a flow

Fix $\boldsymbol{a} \in U$, and let I denote the set of all $t \in I_{\boldsymbol{a}}$ for which there exists a neighborhood of $(t, \boldsymbol{a})$ contained in W on which $\boldsymbol{\Psi}$ is differentiable. We will establish that I is nonempty, open and closed in $I_{\boldsymbol{a}}$, so that $I = I_{\boldsymbol{a}}$ by Proposition 1.7.1: I is nonempty because $0 \in I$ by Theorem 2.8.1, and is open by definition. To see that it is closed, consider a point t_0 in the closure $\bar{I}$ of I; by 2.8.1, there exists a local flow $\boldsymbol{\Psi}' : I' \times V' \to U$ with $0 \in I'$ and $\boldsymbol{\Psi}_{\boldsymbol{a}}(t_0) \in V'$. Let $t_1 \in I$ be close enough to t_0 that $t_0 - t_1 \in I'$ (recall that t_0 belongs to the closure of I) and $\boldsymbol{\Psi}_{\boldsymbol{a}}(t_1) \in V'$ (by continuity of $\boldsymbol{\Psi}_{\boldsymbol{a}}$). Choose an interval I_0 around t_0 such that $t - t_1 \in I'$ for $t \in I_0$. Finally, by continuity of $\boldsymbol{\Psi}$ at $(t_1, \boldsymbol{a})$, there exists a neighborhood V of $\boldsymbol{a}$ such that $\boldsymbol{\Psi}(t_1 \times V) \subset V'$.
We claim that $\boldsymbol{\Psi}$ is defined and differentiable on $I_0 \times V$, so that $t_0 \in I$: Indeed, if $t \in I_0$ and $\boldsymbol{u} \in V$, then by definition of I_0 and V, $t - t_1 \in I'$ and $\boldsymbol{\Psi}(t_1, \boldsymbol{u}) \in V'$, so that $\boldsymbol{\Psi}'(t - t_1, \boldsymbol{\Psi}(t_1, \boldsymbol{u}))$ is defined. The curve $s \mapsto \boldsymbol{\Psi}'(s - t_1, \boldsymbol{\Psi}(t_1, \boldsymbol{u}))$ is an integral curve of $\boldsymbol{X}$ which equals $\boldsymbol{\Psi}(t_1, \boldsymbol{u})$ at t_1. By uniqueness, $\boldsymbol{\Psi}(t, \boldsymbol{u}) = \boldsymbol{\Psi}'(t - t_1, \boldsymbol{\Psi}(t_1, \boldsymbol{u}))$ is defined, and $\boldsymbol{\Psi}$ is therefore differentiable at $(t, \boldsymbol{u})$. □

A vector field on U is said to be *complete* if its flow has domain $\mathbb{R} \times U$; i.e., if its integral curves are defined for all time. The vector field on $\mathbb{R}^2$ from Figure 2.2 is complete. The vector field X on $\mathbb{R}$, with $X(t) = -t^2 D(t)$, is not: for $a \neq 0$, the maximal integral curve c_a of X with $c_a(0) = a$ is given by $c_a(t) = 1/(t + 1/a)$, since

$$\dot{c}_a(t) = -\left(\frac{1}{t + \frac{1}{a}}\right)^2 D\left(\frac{1}{t + \frac{1}{a}}\right) = X\left(\frac{1}{t + \frac{1}{a}}\right) = (X \circ c_a)(t).$$

Example 2.8.1. Consider a vector field $\boldsymbol{X}$ on $\mathbb{R}^n$ of the form

$$\boldsymbol{X} = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij} u^j \right) \boldsymbol{D}_i, \qquad a_{ij} \in \mathbb{R}.$$

Notice that $\boldsymbol{X}(\boldsymbol{u}) = \mathcal{I}_{\boldsymbol{u}}(A\boldsymbol{u})$, where A is the square matrix (a_{ij}). Thus, $\boldsymbol{c} : I \to \mathbb{R}^n$ is an integral curve of $\boldsymbol{X}$ if and only if $\boldsymbol{c}' = A\boldsymbol{c}$. Writing out this equation in components yields a so-called *linear system of ordinary differential equations with constant coefficients*

$$\begin{aligned} x_1'(t) &= \sum_{i=1}^{n} a_{1i} x_i(t) \\ &\vdots \\ x_n'(t) &= \sum_{i=1}^{n} a_{ni} x_i(t) \end{aligned}$$

where $x_i = u^i \circ \boldsymbol{c}$. We claim that $\boldsymbol{X}$ is complete. In fact, the integral curve of $\boldsymbol{X}$ that passes through $\boldsymbol{u} \in \mathbb{R}^n$ at time 0 is given by

$$\begin{aligned} \boldsymbol{c} : \mathbb{R} &\to \mathbb{R}^n, \\ t &\mapsto e^{tA}\boldsymbol{u}. \end{aligned}$$

Indeed, by Exercise 1.50,

$$\begin{aligned}\boldsymbol{c}'(t) &= \lim_{h\to 0}\frac{\boldsymbol{c}(t+h)-\boldsymbol{c}(t)}{h} = \lim_{h\to 0}\frac{e^{(t+h)A}\boldsymbol{u}-e^{tA}\boldsymbol{u}}{h}\\ &= \left(\lim_{h\to 0}\frac{e^{hA}-I_n}{h}\right)e^{tA}\boldsymbol{u}\\ &= A\cdot\boldsymbol{c}(t),\end{aligned}$$

where the last line follows from the series definition of the exponential, cf. Examples and Remarks 1.9.1 (ii).

For example, if $\boldsymbol{X} = u^2D_1 - u^1D_2$ is the vector field from Figure 2.2, the associated matrix is

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

and by Examples and Remarks 1.9.1 (iii), the integral curve $\boldsymbol{c}$ of $\boldsymbol{X}$ with $\boldsymbol{c}(0) = \boldsymbol{u}$ is given by

$$\boldsymbol{c}(t) = \boldsymbol{\Phi}(t,\boldsymbol{u}) = \begin{bmatrix}\cos t & \sin t\\ -\sin t & \cos t\end{bmatrix}\cdot\boldsymbol{u},$$

where $\boldsymbol{\Phi}$ denotes the flow of $\boldsymbol{X}$.

Definition 2.8.6. Let $\boldsymbol{X}$ denote a vector field along a curve $\boldsymbol{c} : I \to \mathbb{R}^n$, so that $\boldsymbol{X}(t) = \mathcal{I}_{\boldsymbol{c}(t)}\boldsymbol{c}_1(t)$ for some curve $\boldsymbol{c}_1$ with the same domain as $\boldsymbol{c}$. The *covariant derivative* of $\boldsymbol{X}$ at $t \in I$ is the tangent vector

$$\boldsymbol{X}'(t) = \mathcal{I}_{\boldsymbol{c}(t)}\boldsymbol{c}_1'(t) \in \mathbb{R}^n_{\boldsymbol{c}(t)}.$$

Thus, the covariant derivative of a vector field along a curve is again a vector field along the same curve. For example, if $\boldsymbol{P}$ is the position vector field on $\mathbb{R}^n$ and $\boldsymbol{c}$ is any curve, then the velocity field of $\boldsymbol{c}$ can be written as

$$\dot{\boldsymbol{c}} = (\boldsymbol{P}\circ\boldsymbol{c})'. \tag{2.8.2}$$

This is because $(\boldsymbol{P}\circ\boldsymbol{c})(t) = \mathcal{I}_{\boldsymbol{c}(t)}\boldsymbol{c}(t)$, so that

$$(\boldsymbol{P}\circ\boldsymbol{c})'(t) = \mathcal{I}_{\boldsymbol{c}(t)}\boldsymbol{c}'(t) = \dot{\boldsymbol{c}}(t).$$

Definition 2.8.7. (1) Let $\boldsymbol{X}$ denote a vector field on $U \subset \mathbb{R}^n$. For $\boldsymbol{p} \in U$, $\boldsymbol{u} \in \mathbb{R}^n_{\boldsymbol{p}}$, the *covariant derivative* $D_{\boldsymbol{u}}\boldsymbol{X}$ *of* $\boldsymbol{X}$ *with respect to* $\boldsymbol{u}$ is defined to be

$$D_{\boldsymbol{u}}\boldsymbol{X} := (\boldsymbol{X}\circ\boldsymbol{c})'(0),$$

where $\boldsymbol{c}$ is any curve with $\dot{\boldsymbol{c}}(0) = \boldsymbol{u}$.

(2) If f is a function on U, the *derivative of* f *with respect to* $\boldsymbol{u}$ is by definition the number

$$\boldsymbol{u}(f) := (f\circ\boldsymbol{c})'(0),$$

with $\boldsymbol{c}$ as above.

It must of course be checked that this definition is independent of the particular curve chosen. Let $\boldsymbol{f} : U \to \mathbb{R}^n$ denote the map $\mathcal{I}^{-1} \circ \boldsymbol{X}$ representing $\boldsymbol{X}$; i.e., $\boldsymbol{X}(\boldsymbol{p}) = \mathcal{I}_{\boldsymbol{p}}\boldsymbol{f}(\boldsymbol{p})$, $\boldsymbol{p} \in U$. Then for any curve $\boldsymbol{c}$ with $\dot{\boldsymbol{c}}(0) = \boldsymbol{u}$,

$$(\boldsymbol{X} \circ \boldsymbol{c})'(0) = \mathcal{I}_{\boldsymbol{c}(0)}(\boldsymbol{f} \circ \boldsymbol{c})'(0) = \mathcal{I}_{\boldsymbol{c}(0)}D\boldsymbol{f}(\boldsymbol{p})\boldsymbol{c}'(0) = \mathcal{I}_{\boldsymbol{p}}D\boldsymbol{f}(\boldsymbol{p})\mathcal{I}_{\boldsymbol{p}}^{-1}\boldsymbol{u},$$

and the derivative is indeed independent of the curve. The argument for the directional derivative of a function is similar. Recalling that $\mathcal{I}_{\boldsymbol{f}(\boldsymbol{p})} \circ D\boldsymbol{f}(\boldsymbol{p}) = \boldsymbol{f}_{*\boldsymbol{p}} \circ \mathcal{I}_{\boldsymbol{p}}$, the right side of the above equality may be rewritten as $\mathcal{I}_{\boldsymbol{p}}\mathcal{I}_{\boldsymbol{f}(\boldsymbol{p})}^{-1}\boldsymbol{f}_{*\boldsymbol{p}}\boldsymbol{u}$. The identity

$$D_{\boldsymbol{u}}\boldsymbol{X} = \mathcal{I}_{\boldsymbol{p}}\mathcal{I}_{\boldsymbol{f}(\boldsymbol{p})}^{-1}\boldsymbol{f}_{*\boldsymbol{p}}\boldsymbol{u}, \qquad \boldsymbol{f} = \mathcal{I}^{-1} \circ \boldsymbol{X}, \tag{2.8.3}$$

says that the covariant derivative of a vector field represented by a map $\boldsymbol{f}$ with respect to a given tangent vector $\boldsymbol{u}$ at a point $\boldsymbol{p}$ is essentially the ordinary derivative $\boldsymbol{f}_*\boldsymbol{u}$. The latter, however, is a vector based at $\boldsymbol{f}(\boldsymbol{p})$, and must therefore be parallel translated back to $\boldsymbol{p}$. Similarly, if the tangent vector $\boldsymbol{u}$ equals $\mathcal{I}_{\boldsymbol{p}}\boldsymbol{v}$ with $\boldsymbol{v} \in \mathbb{R}^n$, then $\boldsymbol{u}(f) = Df(\boldsymbol{p})\boldsymbol{v}$. The motivation behind the somewhat strange notation $\boldsymbol{u}(f)$ is that when $\boldsymbol{u} = \boldsymbol{D}_i(\boldsymbol{p})$, then $\boldsymbol{u}(f)$ equals the i-th partial derivative of f at $\boldsymbol{p}$; i.e., $\boldsymbol{D}_i(\boldsymbol{p})(f) = D_if(\boldsymbol{p})$.

More generally, one can define covariant derivatives of vector fields along maps by using the same equation as in Definition 2.8.7:

Definition 2.8.8. Let $\boldsymbol{f} : \mathbb{R}^n \supset U \to \mathbb{R}^m$, $\boldsymbol{X}$ a vector field along $\boldsymbol{f}$. Given $\boldsymbol{p} \in U$, $\boldsymbol{u} \in \mathbb{R}^n_{\boldsymbol{p}}$, define the *covariant derivative of $\boldsymbol{X}$ with respect to $\boldsymbol{u}$* to be $D_{\boldsymbol{u}}\boldsymbol{X} = (\boldsymbol{X} \circ \boldsymbol{c})'(0)$, where $\boldsymbol{c}$ is a curve in U with $\dot{\boldsymbol{c}}(0) = \boldsymbol{u}$.

Of course, if $\boldsymbol{f}$ is the identity map, one recovers the original concept from Definition 2.8.7. On the other hand, if $\boldsymbol{X}$ is a vector field on $\mathbb{R}^m$ and $\boldsymbol{f} : \mathbb{R}^n \supset U \to \mathbb{R}^m$, then $\boldsymbol{X} \circ \boldsymbol{f}$ is a vector field along $\boldsymbol{f}$, and by the above definition, $D_{\boldsymbol{u}}(\boldsymbol{X} \circ \boldsymbol{f}) = (\boldsymbol{X} \circ \boldsymbol{f} \circ \boldsymbol{c})'(0)$, where $\boldsymbol{c}$ is a curve with $\dot{\boldsymbol{c}}(0) = \boldsymbol{u}$. But $\boldsymbol{f}\dot{\circ}\boldsymbol{c}(0) = \boldsymbol{f}_*\boldsymbol{u}$, so that

$$D_{\boldsymbol{u}}(\boldsymbol{X} \circ \boldsymbol{f}) = D_{\boldsymbol{f}_*\boldsymbol{u}}\boldsymbol{X}. \tag{2.8.4}$$

Similarly, if $\varphi : \mathbb{R}^m \to \mathbb{R}$, then

$$\boldsymbol{u}(\varphi \circ \boldsymbol{f}) = \boldsymbol{f}_*\boldsymbol{u}(\varphi). \tag{2.8.5}$$

Examples 2.8.2. (i) The "coordinate vector field" $\boldsymbol{D}_i$ is parallel, meaning that $D_{\boldsymbol{u}}\boldsymbol{D}_i = 0$ for any $\boldsymbol{u}$. More generally, a vector field $\boldsymbol{X}$ defined on $U \subset \mathbb{R}^n$ is parallel if and only it is of the form

$$\boldsymbol{X} = a_1\boldsymbol{D}_1 + \cdots + a_n\boldsymbol{D}_n, \qquad a_1, \ldots, a_n \in \mathbb{R}.$$

(ii) The position vector field $\boldsymbol{P}$ is represented by the identity map. Since the derivative of the identity is the identity,

$$D_{\boldsymbol{u}}\boldsymbol{P} = \boldsymbol{u}$$

for any vector $\boldsymbol{u}$. Alternatively, this also follows directly from (2.8.2) together with the definition of covariant derivative.

(iii) The *acceleration* of a curve $\boldsymbol{c} : I \to \mathbb{R}^n$ is the vector field $\dot{\boldsymbol{c}}'$ along $\boldsymbol{c}$. It follows from the definition that $\dot{\boldsymbol{c}}' = \mathcal{I}_{\boldsymbol{c}}\boldsymbol{c}''$. In particular, $\boldsymbol{c}$ has zero acceleration if and only if $\boldsymbol{c}'$ is a constant vector $\boldsymbol{v}$, and in this case $\boldsymbol{c}(t) = \boldsymbol{c}(0) + t\boldsymbol{v}$; the image of $\boldsymbol{c}$ is a straight line parallel to $\boldsymbol{v}$.

(iv) The *gradient* of a function $f : U \subset \mathbb{R}^n \to \mathbb{R}$ is the vector field ∇f on U given by $\nabla f(\boldsymbol{p}) = \mathcal{I}_{\boldsymbol{p}}[Df(\boldsymbol{p})]^T$, $\boldsymbol{p} \in U$. Alternatively,

$$\nabla f = \sum_i (D_i f)\boldsymbol{D}_i.$$

The inner product in $\mathbb{R}^n$ induces one in each tangent space $\mathbb{R}^n_{\boldsymbol{p}}$, by setting

$$\langle \mathcal{I}_{\boldsymbol{p}}\boldsymbol{u}, \mathcal{I}_{\boldsymbol{p}}\boldsymbol{v}\rangle := \langle \boldsymbol{u}, \boldsymbol{v}\rangle, \qquad \boldsymbol{p}, \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n.$$

This collection of inner products is called the *standard flat Riemannian metric* on $\mathbb{R}^n$.

As a first application, we shall derive a geometric interpretation of the gradient vector field of a function f. When $\boldsymbol{u}$ is a *unit vector* (meaning $\boldsymbol{u}$ has norm 1), the derivative $\boldsymbol{u}(f)$ is called the *directional derivative of f in direction $\boldsymbol{u}$*. It measures how fast f is changing along a curve that has $\boldsymbol{u}$ as velocity vector. If $\nabla f(\boldsymbol{p}) \neq 0$, we claim that $(\nabla f/|\nabla f|)(\boldsymbol{p})$ is the direction of greatest increase of f at $\boldsymbol{p}$, and $-(\nabla f/|\nabla f|)(\boldsymbol{p})$ the direction of greatest decrease: indeed, for any unit $\boldsymbol{u} \in \mathbb{R}^n_{\boldsymbol{p}}$, if $\boldsymbol{c}$ is a curve with $\dot{\boldsymbol{c}}(0) = \boldsymbol{u}$,

$$\boldsymbol{u}(f) = (f \circ \boldsymbol{c})'(0) = [Df(\boldsymbol{p})]\boldsymbol{c}'(0) = \langle [Df(\boldsymbol{p})]^T, \boldsymbol{c}'(0)\rangle = \langle \nabla f(\boldsymbol{p}), \boldsymbol{u}\rangle,$$

so that if θ denotes the angle between $\boldsymbol{u}$ and $\nabla f(\boldsymbol{p})$, then

$$\boldsymbol{u}(f) = \langle \nabla f(\boldsymbol{p}), \boldsymbol{u}\rangle = |\nabla f(\boldsymbol{p})||\boldsymbol{u}| \cos\theta = |\nabla f(\boldsymbol{p})| \cos\theta\,.$$

This expression is maximal when $\theta = 0$, i.e., when $\boldsymbol{u}$ points in the direction of ∇f, and minimal when $\theta = \pi$, which corresponds to the opposite direction. The *level sets* of $f : \mathbb{R}^n \supset U \to \mathbb{R}$ are the sets $f^{-1}(c) = \{\boldsymbol{p} \in U \mid f(\boldsymbol{p}) = c\}$, $c \in \mathbb{R}$. It follows from the above that the gradient of f is always orthogonal to these level sets; i.e., if $\boldsymbol{c}$ is a smooth curve contained in a level set, then $\langle (\nabla f) \circ \boldsymbol{c}, \dot{\boldsymbol{c}}\rangle = (f \circ \boldsymbol{c})' \equiv 0$. For example, topographic maps of a region usually feature curves connecting points at the same altitude. These curves are the level curves (also called contour lines) of the altitude function, and traveling in a direction perpendicular to them means you are following the path of steepest ascent or descent.

Notice that for vector fields $\boldsymbol{X}$, $\boldsymbol{Y}$ on U, $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle$ is a function on U, if we let $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle(\boldsymbol{p}) = \langle \boldsymbol{X}(\boldsymbol{p}), \boldsymbol{Y}(\boldsymbol{p})\rangle$. In much the same way, given $f : U \to \mathbb{R}$, define new vector fields $f\boldsymbol{X} + \boldsymbol{Y}$ and $D_{\boldsymbol{X}}\boldsymbol{Y}$ on U by

$$(f\boldsymbol{X} + \boldsymbol{Y})(\boldsymbol{p}) := f(\boldsymbol{p})\boldsymbol{X}(\boldsymbol{p}) + \boldsymbol{Y}(\boldsymbol{p}), \quad D_{\boldsymbol{X}}\boldsymbol{Y}(\boldsymbol{p}) := D_{\boldsymbol{X}(\boldsymbol{p})}\boldsymbol{Y}, \qquad \boldsymbol{p} \in U.$$

Similarly, if $\boldsymbol{Y}$ is a vector field on $\mathbb{R}^m$, and $\boldsymbol{X}$ a vector field along a map $\boldsymbol{f} : \mathbb{R}^n \supset U \to \mathbb{R}^n$, the formula $D_{\boldsymbol{X}}\boldsymbol{Y}(\boldsymbol{p}) := D_{\boldsymbol{X}(\boldsymbol{p})}\boldsymbol{Y}$, for $\boldsymbol{p} \in U$, defines a vector field along $\boldsymbol{f}$.

Theorem 2.8.4. *Let $\boldsymbol{X}$, $\boldsymbol{Y}$ denote vector fields on $U \subset \mathbb{R}^n$, $f : U \to \mathbb{R}$, $a \in \mathbb{R}$. Given $\boldsymbol{p} \in U$, $\boldsymbol{u} \in \mathbb{R}^n_{\boldsymbol{p}}$,*

(1) $D_{\boldsymbol{u}}(a\boldsymbol{X} + \boldsymbol{Y}) = aD_{\boldsymbol{u}}\boldsymbol{X} + D_{\boldsymbol{u}}\boldsymbol{Y}$;

(2) $D_{a\boldsymbol{u}+\boldsymbol{v}}\boldsymbol{X} = aD_{\boldsymbol{u}}\boldsymbol{X} + D_{\boldsymbol{v}}\boldsymbol{X}$;

(3) $D_u(f\boldsymbol{X}) = (\boldsymbol{u}(f))X(\boldsymbol{p}) + f(\boldsymbol{p})D_{\boldsymbol{u}}\boldsymbol{X}$;

(4) $D_{f\boldsymbol{X}}\boldsymbol{Y} = fD_{\boldsymbol{X}}\boldsymbol{Y}$;

(5) *Let $\boldsymbol{X}$, $\boldsymbol{Y}$ denote vector fields along a map $\boldsymbol{f} : \mathbb{R}^n \supset U \to \mathbb{R}^m$, $\boldsymbol{p} \in U$, $\boldsymbol{u} \in \mathbb{R}^n_{\boldsymbol{p}}$. Then*

$$D_{\boldsymbol{u}}\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \langle D_{\boldsymbol{u}}\boldsymbol{X}, \boldsymbol{Y}(\boldsymbol{p})\rangle + \langle \boldsymbol{X}(\boldsymbol{p}), D_{\boldsymbol{u}}\boldsymbol{Y}\rangle.$$

Proof. (1), (2), and (4) are left as easy exercises. For (3), let $\boldsymbol{g} = \mathcal{I}^{-1}\boldsymbol{X}$ (as always, in the sense that $\boldsymbol{X}(\boldsymbol{p}) = \mathcal{I}_{\boldsymbol{p}}\boldsymbol{g}(\boldsymbol{p})$) . If $\boldsymbol{c}$ is a curve with $\dot{\boldsymbol{c}}(0) = \boldsymbol{u}$, then

$$\begin{aligned} D_{\boldsymbol{u}}f\boldsymbol{X} &= ((f \circ \boldsymbol{c})(\boldsymbol{X} \circ \boldsymbol{c}))'(0) = \mathcal{I}_{\boldsymbol{p}}((f \circ \boldsymbol{c})(\boldsymbol{g} \circ \boldsymbol{c}))'(0) \\ &= \mathcal{I}_{\boldsymbol{p}}(f(\boldsymbol{p})(\boldsymbol{g} \circ \boldsymbol{c})'(0) + (f \circ \boldsymbol{c})'(0)\boldsymbol{g}(\boldsymbol{p})) \\ &= f(\boldsymbol{p})\mathcal{I}_{\boldsymbol{p}}(\boldsymbol{g} \circ \boldsymbol{c})'(0) + (f \circ \boldsymbol{c})'(0)\mathcal{I}_{\boldsymbol{p}}\boldsymbol{g}(\boldsymbol{p}) \\ &= f(\boldsymbol{p})D_{\boldsymbol{u}}\boldsymbol{X} + (\boldsymbol{u}(f))\boldsymbol{X}(\boldsymbol{p}). \end{aligned}$$

To prove (5), let $\boldsymbol{c} : I \to U$ denote a curve with $\dot{\boldsymbol{c}}(0) = \boldsymbol{u}$. Then $\boldsymbol{X} \circ \boldsymbol{c} = \mathcal{I}_{\boldsymbol{c}}\boldsymbol{f}$ and $\boldsymbol{Y} \circ \boldsymbol{c} = \mathcal{I}_{\boldsymbol{c}}\boldsymbol{g}$ for some curves $\boldsymbol{f}, \boldsymbol{g} : I \to \mathbb{R}^m$. Since $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle \circ \boldsymbol{c} = \langle \boldsymbol{f}, \boldsymbol{g}\rangle$, Corollary 2.2.1 implies

$$\begin{aligned} D_{\boldsymbol{u}}\langle \boldsymbol{X}, \boldsymbol{Y}\rangle &= \langle \boldsymbol{f}, \boldsymbol{g}\rangle'(0) = \langle \boldsymbol{f}', \boldsymbol{g}\rangle(0) + \langle \boldsymbol{f}, \boldsymbol{g}'\rangle(0) \\ &= \langle \mathcal{I}_{\boldsymbol{c}}\boldsymbol{f}', \mathcal{I}_{\boldsymbol{c}}\boldsymbol{g}\rangle(0) + \langle \mathcal{I}_{\boldsymbol{c}}\boldsymbol{f}, \mathcal{I}_{\boldsymbol{c}}\boldsymbol{g}'\rangle(0) \\ &= \langle D_{\boldsymbol{u}}\boldsymbol{X}, \boldsymbol{Y}(\boldsymbol{p})\rangle + \langle \boldsymbol{X}(\boldsymbol{p}), D_{\boldsymbol{u}}\boldsymbol{Y}\rangle. \end{aligned}$$

□

The following is an immediate application of the above theorem together with Examples 2.8.2 (i):

Corollary 2.8.1. *Let $\boldsymbol{X}$ be a vector field on $U \subset \mathbb{R}^n$, and write*

$$\boldsymbol{X} = \sum_i X^i \boldsymbol{D}_i, \qquad X^i = \langle \boldsymbol{X}, \boldsymbol{D}_i\rangle, \quad i = 1, \ldots, n.$$

Then $D_{\boldsymbol{u}}\boldsymbol{X} = \sum_i (\boldsymbol{u}(X^i))\boldsymbol{D}_i(\boldsymbol{p})$ for any vector $\boldsymbol{u} \in \mathbb{R}^n_{\boldsymbol{p}}$, $\boldsymbol{p} \in U$.

Example 2.8.3. Consider the vector fields $\boldsymbol{X}$ and $\boldsymbol{Y}$ in $\mathbb{R}^2$, where

$$\boldsymbol{X} = u^1 \sin u^2 \boldsymbol{D}_1 + (1 + e^{u^2})\boldsymbol{D}_2, \qquad \boldsymbol{Y} = \frac{u^2}{1 + (u^1)^2}\boldsymbol{D}_1 + 3\boldsymbol{D}_2.$$

By Theorem 2.8.4 and the fact that the $\boldsymbol{D}_i$ are parallel,

$$\begin{aligned} D_{\boldsymbol{X}}\boldsymbol{Y} &= u^1 \sin u^2 D_{\boldsymbol{D}_1}\boldsymbol{Y} + (1 + e^{u^2})D_{\boldsymbol{D}_2}\boldsymbol{Y} \\ &= u^1 \sin u^2 D_1\left(\frac{u^2}{1 + (u^1)^2}\right)\boldsymbol{D}_1 + (1 + e^{u^2})D_2\left(\frac{u^2}{1 + (u^1)^2}\right)\boldsymbol{D}_1 \\ &= \left(-\frac{2(u^1)^2 u^2 \sin u^2}{(1 + (u^1)^2)^2} + \frac{1 + e^{u^2}}{1 + (u^1)^2}\right)\boldsymbol{D}_1. \end{aligned}$$

2.9 Lie brackets

If $\boldsymbol{X}$ is a vector field on an open set U in Euclidean space, one can associate to any function $f : U \to \mathbb{R}$ a new function $\boldsymbol{X}f$ by the formula

$$(\boldsymbol{X}f)(\boldsymbol{p}) = \boldsymbol{X}(\boldsymbol{p})(f), \qquad \boldsymbol{p} \in U.$$

Notice that if $\tilde{\boldsymbol{X}} : U \to \mathbb{R}^n$ denotes the map $\mathcal{I}^{-1} \circ \boldsymbol{X}$ representing the vector field $\boldsymbol{X}$, then for $f : U \to \mathbb{R}$,

$$\boldsymbol{X}f = Df(\tilde{\boldsymbol{X}}). \tag{2.9.1}$$

Indeed, if $\boldsymbol{c}$ denote the maximal integral curve of $\boldsymbol{X}$ with $\boldsymbol{c}(0) = \boldsymbol{p} \in U$, then

$$\boldsymbol{X}(\boldsymbol{p})(f) = (f \circ \boldsymbol{c})'(0) = Df(\boldsymbol{c}(0))\boldsymbol{c}'(0) = Df(\boldsymbol{p})\tilde{\boldsymbol{X}}(\boldsymbol{p}).$$

An important special case of (2.9.1) is that the i-th component X^i of a vector field $\boldsymbol{X} = \sum_i X^i \boldsymbol{D}_i$ on U is given by

$$X^i = \langle \boldsymbol{X}, \boldsymbol{D}_i \rangle = u^i \circ \tilde{\boldsymbol{X}} = Du^i(\tilde{\boldsymbol{X}}) = \boldsymbol{X}u^i, \tag{2.9.2}$$

since u^i, being a linear map, equals Du^i.

Given vector fields $\boldsymbol{X}$, $\boldsymbol{Y}$, and a function f, set $\boldsymbol{XY}f := \boldsymbol{X}(\boldsymbol{Y}f)$.

Lemma 2.9.1. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be vector fields on an open set $U \subset \mathbb{R}^n$. Then there exists a unique vector field $[\boldsymbol{X}, \boldsymbol{Y}]$ on U such that*

$$D_{[\boldsymbol{X},\boldsymbol{Y}]}f = (\boldsymbol{XY} - \boldsymbol{YX})f$$

for any function f on U. $[\boldsymbol{X}, \boldsymbol{Y}]$ is called the Lie bracket *of $\boldsymbol{X}$ and $\boldsymbol{Y}$.*

Proof. Write $\boldsymbol{X} = \sum_i X^i \boldsymbol{D}_i$, and similarly for $\boldsymbol{Y}$. Then

$$\begin{aligned} \boldsymbol{XY}f &= \sum_j \boldsymbol{X}(Y^j D_j f) = \sum_{i,j} X^i D_i (Y^j D_j f) \\ &= \sum_{i,j} X^i Y^j D_{ij} f + X^i D_i Y^j D_j f. \end{aligned}$$

Thus,

$$(\boldsymbol{XY} - \boldsymbol{YX})f = \sum_j \left(\sum_i X^i D_i Y^j - Y^i D_i X^j \right) D_j f.$$

Now, if $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are two vectors in $\mathbb{R}^n_{\boldsymbol{p}}$ such that $\boldsymbol{v}_1 f = \boldsymbol{v}_2 f$ for every function f, then the vectors are equal: indeed, taking $f = u^i$ in (2.9.2) shows that the i-th components of both vectors coincide. This proves uniqueness in the definition of the bracket. Existence is also clear: define

$$[\boldsymbol{X}, \boldsymbol{Y}] = \sum_j \left(\sum_i X^i D_i Y^j - Y^i D_i X^j \right) \boldsymbol{D}_j. \tag{2.9.3}$$

□

Together with Theorem 2.2.4, Lemma 2.9.1 immediately implies that $[\boldsymbol{D}_i, \boldsymbol{D}_j] = \mathbf{0}$. It turns out that the Lie bracket of two vector fields measures the amount by which their flows fail to commute; specifically, we will see that if $\boldsymbol{X}$ has flow $\boldsymbol{\Phi}$ and $\boldsymbol{Y}$ has flow $\boldsymbol{\Psi}$, then $[\boldsymbol{X}, \boldsymbol{Y}] \equiv 0$ if and only if $\boldsymbol{\Phi}_t \circ \boldsymbol{\Psi}_s = \boldsymbol{\Psi}_s \circ \boldsymbol{\Phi}_t$ for all s, t, where $\boldsymbol{\Phi}_t(\boldsymbol{a}) = \boldsymbol{\Phi}(t, \boldsymbol{a})$, etc. But first, we state some general properties of the bracket:

Proposition 2.9.1. $[\boldsymbol{X}, \boldsymbol{Y}] = D_{\boldsymbol{X}}\boldsymbol{Y} - D_{\boldsymbol{Y}}\boldsymbol{X}$.

Proof.

$$D_{\boldsymbol{X}}\boldsymbol{Y} = \sum_j (\boldsymbol{X}Y^j)\boldsymbol{D}_j = \sum_j \left((\sum_i X^i \boldsymbol{D}_i Y^j) \right) \boldsymbol{D}_j = \sum_{i,j} X^i D_i Y^j \boldsymbol{D}_j,$$

and a similar expression holds for $D_{\boldsymbol{Y}}\boldsymbol{X}$. The claim now follows from (2.9.3). □

More generally, we have:

Theorem 2.9.1. *Let $\boldsymbol{f} : \mathbb{R}^n \supset U \to \mathbb{R}^m$. If $\boldsymbol{X}$, $\boldsymbol{Y}$ are vector fields on U, then*

$$\boldsymbol{f}_*[\boldsymbol{X}, \boldsymbol{Y}] = D_{\boldsymbol{X}}\boldsymbol{f}_*\boldsymbol{Y} - D_{\boldsymbol{Y}}\boldsymbol{f}_*\boldsymbol{X}.$$

Proof. Write $\boldsymbol{X} = \sum_i X^i \boldsymbol{D}_i$ as usual, and similarly for $\boldsymbol{Y}$. Then

$$\boldsymbol{f}_*\boldsymbol{Y} = \sum_j Y^j \boldsymbol{f}_* \boldsymbol{D}_j = \sum_{i,j} (Y^j D_j f^i) \boldsymbol{D}_i \circ \boldsymbol{f},$$

so that

$$\begin{aligned} D_{\boldsymbol{X}}\boldsymbol{f}_*\boldsymbol{Y} &= \sum_k X^k D_k \left(\sum_{i,j} (Y^j D_j f^i) \boldsymbol{D}_i \circ \boldsymbol{f} \right) \\ &= \sum_{i,j,k} X^k [(D_k Y^j)(D_j f^i) + Y^j (D_k D_j f^i)] \boldsymbol{D}_i \circ \boldsymbol{f}. \end{aligned}$$

Thus,

$$\begin{aligned} D_{\boldsymbol{X}}\boldsymbol{f}_*\boldsymbol{Y} - D_{\boldsymbol{Y}}\boldsymbol{f}_*\boldsymbol{X} &= \sum_{i,j,k} [X^k (D_k Y^j) - Y^k (D_k X^j)](D_j f^i) \boldsymbol{D}_i \circ \boldsymbol{f} \\ &\quad + \sum_{i,j,k} [X^k Y^j D_{kj} f^i - Y^k X^j D_{kj} f^i] \boldsymbol{D}_i \circ \boldsymbol{f}. \end{aligned}$$

The second sum in the above identity vanishes, because

$$\begin{aligned} \sum_{j,k} X^k Y^j D_{kj} f^i - Y^k X^j D_{kj} f^i &= \sum_{j,k} X^k Y^j D_{kj} f^i - Y^j X^k D_{jk} f^i \\ &= \sum_{j,k} X^k Y^j (D_{kj} f^i - D_{jk} f^i) = 0. \end{aligned}$$

Therefore

$$\begin{aligned} D_{\boldsymbol{X}}\boldsymbol{f}_*\boldsymbol{Y} - D_{\boldsymbol{Y}}\boldsymbol{f}_*\boldsymbol{X} &= \sum_{j,k}[X^k(D_kY^j) - Y^k(D_kX^j)]\boldsymbol{f}_*\boldsymbol{D}_j \\ &= \sum_j(\boldsymbol{X}(Y^j) - \boldsymbol{Y}(X^j))\boldsymbol{f}_*\boldsymbol{D}_j \\ &= \boldsymbol{f}_*\sum_j[D_{\boldsymbol{X}}(Y^j\boldsymbol{D}_j) - D_{\boldsymbol{Y}}(X^j\boldsymbol{D}_j)] = \boldsymbol{f}_*(D_{\boldsymbol{X}}\boldsymbol{Y} - D_{\boldsymbol{Y}}\boldsymbol{X}) \\ &= \boldsymbol{f}_*[\boldsymbol{X},\boldsymbol{Y}]. \end{aligned}$$

□

Proposition 2.9.2. *Let $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ denote vector fields on $U \subset \mathbb{R}^n$, $f, g : U \to \mathbb{R}$, $c \in \mathbb{R}$. Then*

(1) $[c\boldsymbol{X} + \boldsymbol{Y}, \boldsymbol{Z}] = c[\boldsymbol{X},\boldsymbol{Z}] + [\boldsymbol{Y},\boldsymbol{Z}]$;

(2) $[\boldsymbol{Y},\boldsymbol{X}] = -[\boldsymbol{X},\boldsymbol{Y}]$;

(3) $[\boldsymbol{X},[\boldsymbol{Y},\boldsymbol{Z}]] + [\boldsymbol{Y},[\boldsymbol{Z},\boldsymbol{X}]] + [\boldsymbol{Z},[\boldsymbol{X},\boldsymbol{Y}]] = 0$;

(4) $[f\boldsymbol{X},\boldsymbol{Y}]g = f[\boldsymbol{X},\boldsymbol{Y}]g - (\boldsymbol{Y}f)(\boldsymbol{X}g)$.

Proof. The first two properties are immediate. For (3), write

$$[\boldsymbol{X},[\boldsymbol{Y},\boldsymbol{Z}]]f = (\boldsymbol{XYZ} - \boldsymbol{XZY} - \boldsymbol{YZX} + \boldsymbol{ZYX})f.$$

Writing out the other two remaining terms in the same way and summing, one easily sees that all the terms cancel. To establish the last property, we compute

$$\begin{aligned} [f\boldsymbol{X},\boldsymbol{Y}]g &= f\boldsymbol{X}(\boldsymbol{Y}g) - \boldsymbol{Y}((f\boldsymbol{X})g) = f\boldsymbol{XY}g - \boldsymbol{Y}(f(\boldsymbol{X}g)) \\ &= f\boldsymbol{XY}g - (\boldsymbol{Y}f)(\boldsymbol{X}g) - f\boldsymbol{YX}g \\ &= f[\boldsymbol{X},\boldsymbol{Y}]g - (\boldsymbol{Y}f)(\boldsymbol{X}g). \end{aligned}$$

□

A vector space V together with an operation $[,] : V \times V \to \mathbb{R}$ that satisfies the first three properties of Proposition 2.9.2 is called a *Lie algebra*. The third property is known as the *Jacobi identity*.

Let $U \subset \mathbb{R}^n$, $\boldsymbol{X}$ a vector field on U, and $\boldsymbol{f} : U \to \mathbb{R}^n$. For each $\boldsymbol{p} \in U$, $\boldsymbol{f}_*\boldsymbol{X}(\boldsymbol{p})$ is a vector in $\mathbb{R}^n_{\boldsymbol{f}(\boldsymbol{p})}$, but in general, this formula does not define a vector field on $\boldsymbol{f}(U)$, since $\boldsymbol{f}$ may not be injective. If $\boldsymbol{Y}$ is a vector field on $\boldsymbol{f}(U)$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to be $\boldsymbol{f}$*-related* if $\boldsymbol{f}_*\boldsymbol{X} = \boldsymbol{Y} \circ \boldsymbol{f}$.

Theorem 2.9.2. *If $\boldsymbol{X}_i$ is $\boldsymbol{f}$-related to $\boldsymbol{Y}_i$, $i = 1, 2$, then $[\boldsymbol{X}_1, \boldsymbol{X}_2]$ is $\boldsymbol{f}$-related to $[\boldsymbol{Y}_1, \boldsymbol{Y}_2]$.*

Proof. We will show that the i-th components of $\boldsymbol{f}_*[\boldsymbol{X}_1, \boldsymbol{X}_2]$ and $[\boldsymbol{Y}_1, \boldsymbol{Y}_2] \circ \boldsymbol{f}$ coincide. By (2.9.2), that of the latter equals

$$([\boldsymbol{Y}_1, \boldsymbol{Y}_2] \circ \boldsymbol{f})u^i = (\boldsymbol{Y}_1 \circ \boldsymbol{f})\boldsymbol{Y}_2u^i - (\boldsymbol{Y}_2 \circ \boldsymbol{f})\boldsymbol{Y}_1u^i,$$

which by (2.8.5) can be written

$$\begin{aligned} (\boldsymbol{f}_*\boldsymbol{X}_1)\boldsymbol{Y}_2u^i - (\boldsymbol{f}_*\boldsymbol{X}_2)\boldsymbol{Y}_1u^i &= \boldsymbol{X}_1((\boldsymbol{Y}_2u^i) \circ \boldsymbol{f}) - \boldsymbol{X}_2((\boldsymbol{Y}_1u^i) \circ \boldsymbol{f}) \\ &= [\boldsymbol{X}_1, \boldsymbol{X}_2](u^i \circ \boldsymbol{f}) \\ &= \boldsymbol{f}_*[\boldsymbol{X}_1, \boldsymbol{X}_2]u^i, \end{aligned}$$

thereby establishing the claim. □

Our next goal is to derive an alternative characterization of the bracket, one that involves the flow of the first vector field. For this, we need the following:

Lemma 2.9.2. *Let I be an interval containing 0. Then any function $f : I \to \mathbb{R}$ may be written as $f(t) = f(0) + tg(t)$, where $g(0) = f'(0)$.*

Proof. Fix any $t \in I$, and define $\varphi : [0,1] \to \mathbb{R}$ by $\varphi(s) = f(st)$. Then

$$f(t) - f(0) = \varphi(1) - \varphi(0) = \int_0^1 \varphi'(s)\,\mathrm{d}s = t\int_0^1 f'(st)\,\mathrm{d}s = tg(t),$$

where $g(t) = \int_0^1 f'(st)\,\mathrm{d}s$. □

Theorem 2.9.3. *Given vector fields $\boldsymbol{X}, \boldsymbol{Y}$ on an open set $U \subset \mathbb{R}^n$, and $\boldsymbol{p} \in U$,*

$$[\boldsymbol{X}, \boldsymbol{Y}](\boldsymbol{p}) = \lim_{t\to 0} \frac{\boldsymbol{\Phi}_{-t*}\boldsymbol{Y}(\boldsymbol{\Phi}_t(\boldsymbol{p})) - \boldsymbol{Y}(\boldsymbol{p})}{t},$$

where $\boldsymbol{\Phi}$ denotes the flow of $\boldsymbol{X}$, and $\boldsymbol{\Phi}_t(\boldsymbol{q}) := \boldsymbol{\Phi}(t, \boldsymbol{q})$, $\boldsymbol{q} \in U$.

Proof. The plan is to show that the two vectors in the above identity have the same components. The i-th component of the one on the left is

$$\left([\boldsymbol{X}, \boldsymbol{Y}]u^i\right)(\boldsymbol{p}) = \left(\boldsymbol{X}(\boldsymbol{p})\boldsymbol{Y} - \boldsymbol{Y}(\boldsymbol{p})\boldsymbol{X}\right)(u^i).$$

As usual, denote by $\tilde{\boldsymbol{Y}} : U \to \mathbb{R}^n$ the map $\mathcal{I}^{-1} \circ \boldsymbol{Y}$ representing the vector field $\boldsymbol{Y}$. The i-th component Z^i of the right side prior to taking limits is

$$\frac{1}{t}\left[u^i \circ D\boldsymbol{\Phi}_{-t}(\boldsymbol{\Phi}_t(\boldsymbol{p}))(\tilde{\boldsymbol{Y}}(\boldsymbol{\Phi}_t(\boldsymbol{p}))) - u^i \circ \tilde{\boldsymbol{Y}}(\boldsymbol{p})\right]$$
$$= \frac{1}{t}\left[D(u^i \circ \boldsymbol{\Phi}_{-t})(\boldsymbol{\Phi}_t(\boldsymbol{p}))(\tilde{\boldsymbol{Y}}(\boldsymbol{\Phi}_t(\boldsymbol{p}))) - u^i \circ \tilde{\boldsymbol{Y}}(\boldsymbol{p})\right]$$

by the chain rule and the fact that $Du^i = u^i$, u^i being linear. Now, if $h(t, \boldsymbol{p}) = (u^i \circ \boldsymbol{\Phi})(t, \boldsymbol{p})$, then by the previous lemma, $h(t, \boldsymbol{p}) = u^i(\boldsymbol{p}) + tg(t, \boldsymbol{p})$, where $g(0, \boldsymbol{p}) = D_1 h(0, \boldsymbol{p}) = D_1(u^i \circ \boldsymbol{\Phi})(0, \boldsymbol{p})$. By Theorem 2.8.1 (2), the latter equals $u^i(\tilde{\boldsymbol{X}}(\boldsymbol{p}))$, which may be written as $\boldsymbol{X}(\boldsymbol{p})(u^i)$. Setting $g_t(\boldsymbol{q}) = g(t, \boldsymbol{q})$, we conclude that

$$u^i \circ \boldsymbol{\Phi}_{-t} = u^i - tg_{-t}, \qquad g_0 = \boldsymbol{X}u^i.$$

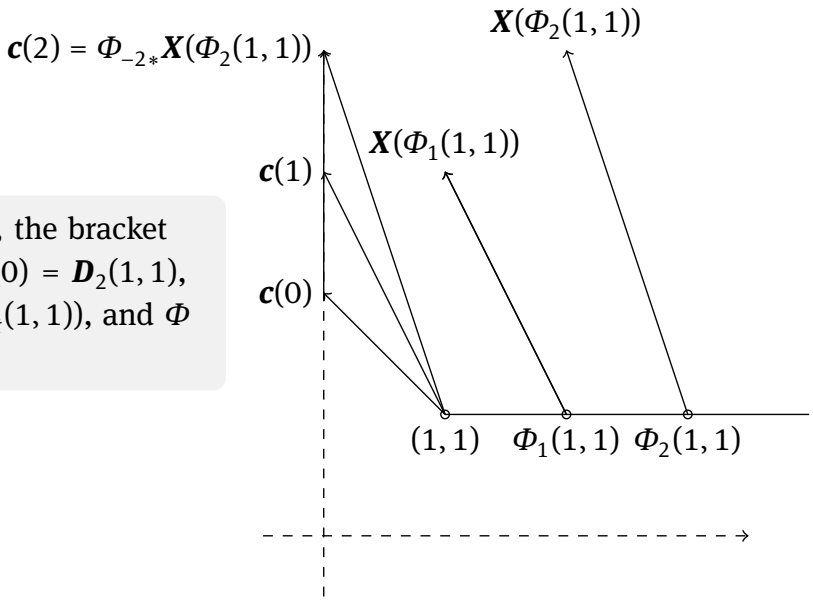
Thus,

$$Z^i = \frac{1}{t}\left[Du^i(\boldsymbol{\Phi}_t(\boldsymbol{p}))(\tilde{\boldsymbol{Y}}(\boldsymbol{\Phi}_t(\boldsymbol{p}))) - u^i(\tilde{\boldsymbol{Y}}(\boldsymbol{p}))\right] - Dg_{-t}(\boldsymbol{\Phi}_t(\boldsymbol{p}))(\tilde{\boldsymbol{Y}}(\boldsymbol{\Phi}_t(\boldsymbol{p})))$$
$$= \frac{1}{t}\left[(u^i \circ \tilde{\boldsymbol{Y}})(\boldsymbol{\Phi}_t(\boldsymbol{p})) - u^i(\tilde{\boldsymbol{Y}}(\boldsymbol{p}))\right] - Dg_{-t}(\boldsymbol{\Phi}_t(\boldsymbol{p}))(\tilde{\boldsymbol{Y}}(\boldsymbol{\Phi}_t(\boldsymbol{p}))).$$

Taking limits as $t \to 0$ and recalling that $t \mapsto \boldsymbol{\Phi}_t(\boldsymbol{p})$ is the integral curve of $\boldsymbol{X}$ passing through $\boldsymbol{p}$ at time 0, we deduce that the i-th component of the right side is

$$\boldsymbol{X}(\boldsymbol{p})(u^i \circ \tilde{\boldsymbol{Y}}) - Dg_0(\boldsymbol{p})(\tilde{\boldsymbol{Y}}(\boldsymbol{p})) = \boldsymbol{X}(\boldsymbol{p})(\boldsymbol{Y}u^i) - D(\boldsymbol{X}u^i)(\boldsymbol{p})\tilde{\boldsymbol{Y}}(\boldsymbol{p})$$
$$= \boldsymbol{X}(\boldsymbol{p})(\boldsymbol{Y}u^i) - \boldsymbol{Y}(\boldsymbol{p})(\boldsymbol{X}u^i),$$

as claimed. □



For $X = u^1 D_2 - u^2 D_1$, the bracket $[D_1, X](1,1)$ equals $c'(0) = D_2(1,1)$, where $c(t) = \Phi_{-t*} X(\Phi_t(1,1))$, and Φ is the flow of D_1

Our last goal in this section is to use the above theorem and prove that the bracket of two vector fields is identically zero if and only if their flows commute. We begin with the following:

Lemma 2.9.3. *Let U be an open set in $\mathbb{R}^n$, $f : U \to f(U) \subset \mathbb{R}^n$ a diffeomorphism, and X a vector field on U with flow Φ. Then X is f-related to itself if and only if $f \circ \Phi_t = \Phi_t \circ f$. Here, as always, $\Phi_t = \Phi(t, \cdot)$.*

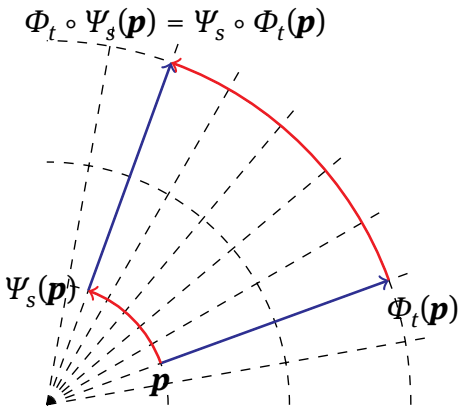
Proof. Fix any $q \in f(U)$, and consider the curve γ, where

$$\gamma(t) = (f \circ \Phi_t \circ f^{-1})(q).$$

The curve $t \mapsto c(t) := (\Phi_t \circ f^{-1})(q) = (f^{-1} \circ \gamma)(t)$ is by definition an integral curve of X, so that

$$\dot{\gamma} = f_*(\dot{c}) = f_* \circ X \circ c = (f_* \circ X \circ f^{-1}) \circ \gamma.$$

Thus, the vector field $f_* \circ X \circ f^{-1}$ has flow Ψ, where $\Psi_t = f \circ \Phi_t \circ f^{-1}$. The claim follows, since X is f-related to itself if and only if $f_* \circ X \circ f^{-1} = X$; i.e., if and only if $\Phi_t = f \circ \Phi_t \circ f^{-1}$. □



The flows Φ of $X = u^1 D_1 + u^2 D_2$ (in blue) and Ψ of $Y = u^1 D_2 - u^2 D_1$ (in red) commute

Theorem 2.9.4. *Let $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ denote the flows of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Then $[\boldsymbol{X}, \boldsymbol{Y}] \equiv \boldsymbol{0}$ if and only if these flows commute; i.e., $\boldsymbol{\Phi}_t \circ \boldsymbol{\Psi}_s = \boldsymbol{\Psi}_s \circ \boldsymbol{\Phi}_t$ for all s, t.*

Proof. If $\boldsymbol{\Phi}_t \circ \boldsymbol{\Psi}_s = \boldsymbol{\Psi}_s \circ \boldsymbol{\Phi}_t$, then $\boldsymbol{Y}$ is $\boldsymbol{\Phi}_t$-related to itself for every t by Lemma 2.9.3; i.e., $\boldsymbol{\Phi}_{t*}\boldsymbol{Y} = \boldsymbol{Y} \circ \boldsymbol{\Phi}_t$, and by Theorem 2.9.3,

$$[\boldsymbol{X}, \boldsymbol{Y}](\boldsymbol{p}) = \lim_{t\to 0} \frac{\boldsymbol{\Phi}_{-t*}\boldsymbol{Y}(\boldsymbol{\Phi}_t(\boldsymbol{p})) - \boldsymbol{Y}(\boldsymbol{p})}{t} = \lim_{t\to 0} \frac{\boldsymbol{Y}(\boldsymbol{p}) - \boldsymbol{Y}(\boldsymbol{p})}{t} = \boldsymbol{0}$$

for every $\boldsymbol{p}$. Conversely, suppose the vector fields have vanishing bracket, and for any fixed $\boldsymbol{p}$, consider the curve $\boldsymbol{c}$ in the tangent space $\mathbb{R}^n_{\boldsymbol{p}}$ of $\mathbb{R}^n$ at $\boldsymbol{p}$ given by $\boldsymbol{c}(t) = \boldsymbol{\Phi}_{-t*}\boldsymbol{Y}(\boldsymbol{\Phi}_t(\boldsymbol{p}))$. We will be done (again by the lemma) once we show that $\boldsymbol{c}(t) = \boldsymbol{c}(0) = \boldsymbol{Y}(\boldsymbol{p})$ for all t; i.e., once we establish that $\boldsymbol{c}' \equiv \boldsymbol{0}$. Vanishing of the Lie bracket implies, by Theorem 2.9.3, that $\boldsymbol{c}'(0) = \boldsymbol{0}$. But if $\boldsymbol{q} = \boldsymbol{\Phi}_t(\boldsymbol{p})$ for any given fixed t, then using Exercise 2.28, we obtain

$$\begin{aligned}
\boldsymbol{c}'(t) &= \lim_{h\to 0} \frac{\boldsymbol{c}(t+h) - \boldsymbol{c}(t)}{h} \\
&= \lim_{h\to 0} \frac{1}{h} \left[\boldsymbol{\Phi}_{-(t+h)*} \circ \boldsymbol{Y} \circ \boldsymbol{\Phi}_{t+h}(\boldsymbol{p}) - \boldsymbol{\Phi}_{-t*} \circ \boldsymbol{Y} \circ \boldsymbol{\Phi}_t(\boldsymbol{p})\right] \\
&= \lim_{h\to 0} \frac{1}{h} \boldsymbol{\Phi}_{-t*} \left[(\boldsymbol{\Phi}_{-h*} \circ \boldsymbol{Y} \circ \boldsymbol{\Phi}_h)(\boldsymbol{\Phi}_t(\boldsymbol{p})) - \boldsymbol{Y}(\boldsymbol{\Phi}_t(\boldsymbol{p}))\right] \\
&= \boldsymbol{\Phi}_{-t*} \lim_{h\to 0} \frac{1}{h} \left[(\boldsymbol{\Phi}_{-h*} \circ \boldsymbol{Y} \circ \boldsymbol{\Phi}_h)(\boldsymbol{q}) - \boldsymbol{Y}(\boldsymbol{q})\right] = \boldsymbol{\Phi}_{-t*}\boldsymbol{c}'(0) \\
&= \boldsymbol{0}.
\end{aligned}$$

□

More generally, given $\boldsymbol{p} \in \mathbb{R}^n$, define a curve $\boldsymbol{c}_{\boldsymbol{p}}$ on an interval around 0 by

$$\boldsymbol{c}_{\boldsymbol{p}}(t) = (\boldsymbol{\Psi}_{-t} \circ \boldsymbol{\Phi}_{-t} \circ \boldsymbol{\Psi}_t \circ \boldsymbol{\Phi}_t)(\boldsymbol{p}).$$

If the bracket of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is identically zero, then $\boldsymbol{c}_{\boldsymbol{p}}(t) = \boldsymbol{p}$ for all t. When the bracket is nonzero, the curve $\boldsymbol{c}_{\boldsymbol{p}}$ is no longer constant. In the exercises, however, the reader is asked to show that $\boldsymbol{c}'_{\boldsymbol{p}}(0) = \boldsymbol{0}$. It can be shown that the acceleration of $\boldsymbol{c}_{\boldsymbol{p}}$ at 0 measures the bracket; specifically, for any smooth function $f : U \to \mathbb{R}$ defined on a neighborhood U of $\boldsymbol{p}$,

$$(f \circ \boldsymbol{c}_{\boldsymbol{p}})''(0) = 2[\boldsymbol{X}, \boldsymbol{Y}](\boldsymbol{p})(f).$$

2.10 Partitions of unity

In this section, we introduce a concept that plays a key role both in the theory of integration and that of manifolds.

Let $A \subset \mathbb{R}^n$. Recall that a collection $\mathcal{C}$ of (not necessarily open) subsets of $\mathbb{R}^n$ is said to be a *cover* of A if A is contained in the union of all the sets in $\mathcal{C}$. $\mathcal{C}$ is said to be *locally finite* if every $\boldsymbol{a} \in A$ admits a neighborhood that intersects only finitely many sets in $\mathcal{C}$.

Definition 2.10.1. A *partition of unity* for $A \subset \mathbb{R}^n$ is a collection Φ of differentiable functions defined on some open set containing A such that
(1) $0 \le \varphi \le 1$ for all $\varphi \in \Phi$;
(2) The collection $\{\operatorname{supp} \varphi \mid \varphi \in \Phi\}$ is a locally finite cover of A;
(3) $\sum_{\varphi \in \Phi} \varphi(\boldsymbol{a}) = 1$ for all $\boldsymbol{a} \in A$.

Even though Φ may be an infinite collection, the sum in (3) makes sense because by (2), only finitely many φ are nonzero on some neighborhood of any given point. In fact, only finitely many φ are nonzero on any given compact set $C \subset A$: by (2), any $\boldsymbol{p} \in C$ has a neighborhood on which only finitely many φ are not identically zero. Since C is compact, it can be covered by finitely many such sets.

Theorem 2.10.1. *Any $A \subset \mathbb{R}^n$ admits a partition of unity Φ. Furthermore, if $\{U_\alpha\}_{\alpha \in J}$ is an open cover of A, then Φ may be chosen to be* subordinate *to the cover; i.e., for any $\varphi \in \Phi$, there exists $\alpha \in J$ such that* $\operatorname{supp} \varphi \subset U_\alpha$.

Proof. Both statements will be considered at the same time; in other words, we assume some open cover$\{U_\alpha\}_{\alpha \in J}$ of A is given (if none is given, we take it to be $\{\mathbb{R}^n\}$). For each $\boldsymbol{a} \in A$, choose some U_α that contains it, and some bounded open neighborhood $V_{\boldsymbol{a}}$ of $\boldsymbol{a}$ whose closure is contained in U_α. The collection $\{V_{\boldsymbol{a}} \mid \boldsymbol{a} \in A\}$ is an open cover of A, which, by Theorem 1.7.5, contains a countable subcover $\{V_1, V_2, \dots\}$. By construction, for each $i \in \mathbb{N}$ there exists some $\alpha_i \in J$ such that $\overline{V}_i \subset U_{\alpha_i}$. We may now appeal to Theorem 2.2.5 to assert the existence of smooth functions $\tilde{\varphi}_i : \mathbb{R}^n \to \mathbb{R}$, $i \in \mathbb{N}$, satisfying

$$0 \le \tilde{\varphi}_i \le 1, \qquad \tilde{\varphi}_{i|\overline{V}_i} \equiv 1, \qquad \operatorname{supp} \tilde{\varphi}_i \subset U_{\alpha_i}.$$

Set $\varphi_1 = \tilde{\varphi}_1$, and

$$\varphi_i = (1 - \tilde{\varphi}_1) \cdots (1 - \tilde{\varphi}_{i-1}) \tilde{\varphi}_i, \qquad i > 1.$$

Notice that $\operatorname{supp} \varphi_i \subset \operatorname{supp} \tilde{\varphi}_i \subset U_{\alpha_i}$, so that the collection of supports is subordinate to the original cover. Furthermore, it is locally finite: given any $\boldsymbol{a} \in A$, choose some V_i that contains it. Then $\tilde{\varphi}_i$ equals 1 on V_i, and by definition, φ_k vanishes on V_i for all $k > i$. Finally, an easy induction argument shows that

$$\sum_{i=1}^{k} \varphi_i = 1 - (1 - \tilde{\varphi}_1) \cdots (1 - \tilde{\varphi}_k), \qquad k \in \mathbb{N},$$

so that on V_k (where $1 - \tilde{\varphi}_k$ is identically zero),

$$\sum_{i=1}^{\infty} \varphi_{i|V_k} = \sum_{i=1}^{k} \varphi_{i|V_k} = 1 - 0 = 1.$$

Since the V_k cover A, $\sum_i \varphi_i$ is identically 1 on A. □

Remarks 2.10.1. (i) The proof of the Theorem shows that when the cover $\{U_1, \dots, U_k\}$ of A is finite, the partition of unity Φ may be assumed to consist of the same

number of elements; i.e. $\Phi = \{\varphi_1, \ldots, \varphi_k\}$, with supp $\varphi_i \subset U_i$. It also shows that in general, the partition of unity is countable.

(ii) We are now in a position to prove the claim made is Section 2.2, that given any open set $U \subset \mathbb{R}^n$, any $\boldsymbol{p} \in U$, and any (not necessarily compact) neighborhood V of $\boldsymbol{p}$ whose closure is contained in U, there exists a smooth function $\varphi : \mathbb{R}^n \to \mathbb{R}$, $0 \leq \varphi \leq 1$, that equals 1 on the closure on V and has its support inside U: indeed, since $\{U, \mathbb{R}^n \setminus \overline{V}\}$ is an open cover of $\mathbb{R}^n$, there exists a partition of unity $\{\psi_1, \psi_2\}$ subordinate to this cover, with supp $\psi_1 \subset U$ and supp $\psi_2 \subset \mathbb{R}^n \setminus \overline{V}$. Setting $\varphi = \psi_1$ yields the claim: its support lies in U, and since ψ_2 vanishes on $\overline{V}$, φ must equal 1 there.

2.11 Exercises

2.1. Find the derivative of each of the following maps:

(1) $f(x, y, z) = (x^y)^z$;

(2) $g(x, y, z) = x^{(y^z)}$;

(3) $\boldsymbol{h}(x, y, z) = (x \cos(ye^z), \sin \sqrt{x^2 + z^2 + 1})$.

(4) $\boldsymbol{k}(x, y, z, w) = ((x - y)e^{(z+w^2)}, \cos(\sin(\log y^4 + z^2 + 5)), 1/(x^2 + y^2 + z^2 + w^2 + 1))$.

2.2. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x, y) = x^2 y + \cos e^{x+y}$, and $\boldsymbol{c} : \mathbb{R} \to \mathbb{R}^2$ by $\boldsymbol{c}(t) = (t - 1, t^2 + 1)$.

(a) Use the chain rule to determine $(f \circ \boldsymbol{c})'(t)$ and $D(\boldsymbol{c} \circ f)(x, y)$;

(b) Check your answers in (a) by computing $f \circ \boldsymbol{c}$, $\boldsymbol{c} \circ f$ and evaluating their derivatives directly.

2.3. Suppose $g : \mathbb{R}^3 \to \mathbb{R}$ has Jacobian matrix $\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$ at $(1, 1, 0)$. Find $Df(1, 1, 1)$, if $f(x, y, z) = g(x^2, yz, x - y)$.

2.4. Suppose $\boldsymbol{c} : I \to \mathbb{R}^n$ is a curve such that the position vector $\boldsymbol{c}(t)$ is always orthogonal to the velocity vector $\boldsymbol{c}'(t)$. Show that the image of $\boldsymbol{c}$ lies in some sphere centered at the origin.

2.5. Let $f : \mathbb{R}^2 \to \mathbb{R}$, and define $g : \mathbb{R}^2 \to \mathbb{R}$ by $g(r, \theta) = f(r \cos \theta, r \sin \theta)$. Express the partial derivatives of g in terms of those of f.

Many Calculus texts write these partial derivatives in polar coordinates as

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x} \cos \theta + \frac{\partial f}{\partial y} \sin \theta,$$

and another expression for the derivative of f with respect to θ. Explain why, even though the notation is suggestive, the formula is incorrect or at least ambiguous.

2.6. Answer the same question as in the previous problem, but for a function $f : \mathbb{R}^3 \to \mathbb{R}$ of 3 variables, and for spherical coordinates.

2.7. If E is a vector space, and $\boldsymbol{f} : E \to \mathbb{R}^m$ is a map, we say $\boldsymbol{f}$ *is differentiable* if $\boldsymbol{f} \circ L : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable for some isomorphism $L : \mathbb{R}^n \to E$.
(a) Prove that this definition makes sense; i.e. it is independent of the choice of L.
(b) Explain why one recovers the usual concept of differentiability when $E = \mathbb{R}^n$.
(c) Show that any linear transformation $T : E \to \mathbb{R}^m$ is differentiable.

2.8. Let E be an inner product space. Prove that the function $f : E \to \mathbb{R}$ given by $f(\boldsymbol{u}) = |\boldsymbol{u}|$ is differentiable at any point other than the origin, and is not differentiable at $\mathbf{0}$ (see Exercise 2.7). Find $Df(\boldsymbol{u})$.

2.9. Let E be an inner product space, $f : E \to \mathbb{R}$ a function satisfying $|f(\boldsymbol{u})| \leq |\boldsymbol{u}|^\alpha$ for all $\boldsymbol{u} \in E$ and some $\alpha > 1$. Show that f is differentiable at $\mathbf{0}$ and find $Df(\mathbf{0})$. (Notice that the conclusion is false if $\alpha = 1$ by Exercise 2.8.)

2.10. Determine whether $f : \mathbb{R}^n \to \mathbb{R}$, where

$$f(\boldsymbol{a}) = \begin{cases} \frac{\sin |\boldsymbol{a}|^3}{|\boldsymbol{a}|} & \text{if } \boldsymbol{a} \neq \mathbf{0} \\ 0 & \text{if } \boldsymbol{a} = \mathbf{0}, \end{cases}$$

is differentiable at the origin, and if yes, find $Df(\mathbf{0})$.

2.11. Suppose $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ satisfies $\boldsymbol{f}(t\boldsymbol{a}) = t\boldsymbol{f}(a)$ for all $\boldsymbol{a} \in \mathbb{R}^n$ and $t \in \mathbb{R}$. Prove that if $\boldsymbol{f}$ is differentiable at the origin, then $\boldsymbol{f}$ is linear. *Hint:* It is enough to show that $\boldsymbol{f} = D\boldsymbol{f}(\mathbf{0})$.

2.12. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *homogeneous of degree* k if $f(t\boldsymbol{a}) = t^k f(\boldsymbol{a})$ for all $\boldsymbol{a} \in \mathbb{R}^n$, $t \in \mathbb{R}$, and some $k \in \mathbb{N}$. Show that if f is homogeneous of degree k and differentiable, then

$$f = \frac{1}{k} \sum_{i=1}^{n} u^i D_i f.$$

2.13. Show, by means of an example, that in an arbitrary metric space X, a contraction $X \to X$ does not necessarily have a fixed point (in this case, X cannot of course be complete).

2.14. Let U denote a convex open set in $\mathbb{R}^n$, $\boldsymbol{f} : U \to \mathbb{R}^m$ a differentiable map, and $\boldsymbol{a}, \boldsymbol{b} \in U$. Prove the *mean value theorem*: For any $\boldsymbol{u} \in \mathbb{R}^m$, there exists some $\boldsymbol{c}$ on the line segment joining $\boldsymbol{a}$ and $\boldsymbol{b}$ such that

$$\langle \boldsymbol{u}, \boldsymbol{f}(\boldsymbol{b}) - \boldsymbol{f}(\boldsymbol{a}) \rangle = \langle \boldsymbol{u}, D\boldsymbol{f}(\boldsymbol{c})(\boldsymbol{b} - \boldsymbol{a}) \rangle.$$

Hint: Set $\boldsymbol{v} = \boldsymbol{b} - \boldsymbol{a}$, and apply the ordinary mean value theorem to g on $[0, 1]$, where $g(t) := \langle \boldsymbol{u}, \boldsymbol{f}(\boldsymbol{a} + t\boldsymbol{v}) \rangle$.

2.15. (a) Show that the mean value theorem from Exercise 2.14 does indeed generalize the ordinary mean value theorem.

(b) Suppose U is a convex open set in $\mathbb{R}^n$, $\boldsymbol{a}, \boldsymbol{b} \in U$, $f : U \to \mathbb{R}$ a differentiable function. Prove that there exists some $\boldsymbol{c}$ on the line segment joining $\boldsymbol{a}$ and $\boldsymbol{b}$ such that

$$f(\boldsymbol{b}) - f(\boldsymbol{a}) = Df(\boldsymbol{c})(\boldsymbol{b} - \boldsymbol{a}) = \langle \nabla f(\boldsymbol{c}), \boldsymbol{b} - \boldsymbol{a} \rangle.$$

2.16. Prove that a linear isometry of an inner product space has determinant ± 1.

2.17. Show that if $T : V \to W$ is a linear isometry and L is a self-adjoint (respectively skew-adjoint) operator on W, then $T^{-1} \circ L \circ T$ is a self-adjoint (resp. skew-adjoint) operator on V.

2.18. The *orthogonal group* $O(n) \subset M_{n,n}$ is the collection of all $n \times n$ orthogonal matrices. Suppose $\boldsymbol{c} : I \to O(n)$ is a smooth curve with $\boldsymbol{c}(0) = I_n$. Prove that $\boldsymbol{c}'(0)$ is skew-adjoint: $\boldsymbol{c}'(0) + \boldsymbol{c}'^T(0) = \mathbf{0}$.

2.19. (a) Find the derivative of the determinant $\det : M_{n,n} \cong \mathbb{R}^{n^2} \to \mathbb{R}$.

(b) The *special linear group* is the subset $Sl(n) \subset M_{n,n}$ consisting of all matrices with determinant 1. Suppose $\boldsymbol{c} : I \to Sl(n)$ is a smooth curve with $\boldsymbol{c}(0) = I_n$. Prove that $\boldsymbol{c}'(0)$ has trace equal to zero.

2.20. Prove that there exists a differentiable function f defined on some neighborhood of $(1, 0) \in \mathbb{R}^2$ satisfying $x \log f(x, y) + y f(x, y) = 0$.

2.21. Find and classify all critical points of f, if $f(x, y) = e^{x^2 - y^2 + 1}$.

2.22. Find and classify all critical points of f, where f is given by $f(x, y) = (x^2 - y^2)e^{-(x^2+y^2)^2}$.

2.23. Show that the function f given by

$$f(x, y) = \begin{cases} \frac{\sin(xy) - xy}{x^2 + y^2} & \text{if } (x, y) \neq \mathbf{0}, \\ 0 & \text{if } (x, y) = \mathbf{0} \end{cases}$$

is continuous everywhere. *Hint:* Look at the Taylor polynomial of the function $(x, y) \mapsto \sin(xy)$ at $\mathbf{0}$.

2.24. Recall from Chapter 1 that a *polynomial of degree k* on $\mathbb{R}^n$ is a function f of the form

$$f = a_0 + \sum_{j=1}^{k} \sum_{1 \leq i_1, \ldots, i_j \leq n} a_{i_1 \ldots i_j} u^{i_1} \cdots u^{i_j}, \qquad a_0, a_{i_1 \ldots i_j} \in \mathbb{R}.$$

Prove that the Taylor polynomial of degree k at any point of such a function f equals f.

2.25. Write $2x^2 - y^2 + 3xy - 1$ as a polynomial in $(x - 1)$ and $(y + 1)$.

2.26. Let b denote a scalar product on $\mathbb{R}^n$, L the associated self-adjoint operator, so that $b(\boldsymbol{u}, \boldsymbol{v}) = \langle L\boldsymbol{u}, \boldsymbol{v} \rangle$. The *quadratic form associated to b* is the function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(\boldsymbol{a}) = b(\boldsymbol{a}, \boldsymbol{a})$.

(a) Show that the Jacobian matrix $[Df(\boldsymbol{a})]$ of f at $\boldsymbol{a}$ equals $2(L\boldsymbol{a})^T$.
(b) Prove that the Hessian of f is $2L$.
(c) Part (a) implies that the set of critical points of f is the kernel of L, and in particular the origin is always a critical point. Suppose the kernel is trivial, so that the origin is the only critical point. *Without using Theorem 2.7.3*, show that the origin is an absolute minimum if all eigenvalues of L are positive, an absolute maximum if they are all negative, and a saddle point in all other cases.

2.27. A map $\boldsymbol{f} : \mathbb{R}^n \supset U \to \mathbb{R}^{n+k}$ that has rank n at every point in U is called an *immersion* of U into $\mathbb{R}^{n+k}$. An injective immersion is called an *imbedding* if $\boldsymbol{f}^{-1} : \boldsymbol{f}(U) \to U$ is continuous.
(a) Show that an immersion $\boldsymbol{f} : U \to \mathbb{R}^{n+k}$ is always locally an imbedding; i.e., any $\boldsymbol{a} \in U$ admits a neighborhood V such that the restriction of $\boldsymbol{f}$ to V is an imbedding.
(b) Let $\boldsymbol{c} : (0, 2\pi) \to \mathbb{R}^2$ be given by $\boldsymbol{c}(t) = (\sin t, \sin 2t)$. The image of $\boldsymbol{c}$ is called a *lemniscate*. Show that $\boldsymbol{c}$ is an injective immersion but not an imbedding.

2.28. If $\boldsymbol{X}$ is a vector field with flow $\boldsymbol{\Phi}$, define a map $\boldsymbol{\Phi}_t$ by $\boldsymbol{\Phi}_t(\boldsymbol{p}) = \boldsymbol{\Phi}(t, \boldsymbol{p})$, for all t and $\boldsymbol{p}$ for which the formula makes sense. Prove that $\boldsymbol{\Phi}_t \circ \boldsymbol{\Phi}_s = \boldsymbol{\Phi}_s \circ \boldsymbol{\Phi}_t = \boldsymbol{\Phi}_{t+s}$.

2.29. Let U be open in $\mathbb{R}^n$, $\boldsymbol{p} \in U$, $\boldsymbol{u} = (\boldsymbol{p}, \boldsymbol{v}) \in \mathbb{R}^n_{\boldsymbol{p}}$.
(a) If $\boldsymbol{X}$ is a smooth vector field on U represented by the map $\boldsymbol{g}$, show that $D_{\boldsymbol{u}}\boldsymbol{X} = (\boldsymbol{p}, D\boldsymbol{g}(\boldsymbol{p})\boldsymbol{v})$.
(b) If $f : U \to \mathbb{R}$ is smooth, show that the derivative of f in direction $\boldsymbol{u}$ (as defined in Definition 2.8.7) equals $Df(\boldsymbol{p})\boldsymbol{v}$.

2.30. Determine the velocity and acceleration vector fields of the curve $\boldsymbol{c} : I \to \mathbb{R}^3$, $\boldsymbol{c}(t) = (a\cos t, a\sin t, bt)$, $a, b > 0$, which parametrizes a helix.

2.31. Given a function $f : \mathbb{R}^n \to \mathbb{R}$, extend the Hessian of f on $\mathbb{R}^n$ to any tangent space by identifying both spaces in the usual way; i.e., for $\boldsymbol{a} \in \mathbb{R}^n$, define $\tilde{H}_f : \mathbb{R}^n_{\boldsymbol{a}} \to \mathbb{R}^n_{\boldsymbol{a}}$ by $\tilde{H}_f \mathcal{I}_{\boldsymbol{a}} \boldsymbol{u} = H_f \boldsymbol{u}$, $\boldsymbol{u} \in \mathbb{R}^n$, and extend the Hessian form similarly. For the sake of simplicity, we denote $\tilde{H}_f$ by H_f again.
(a) Show that $H_f \boldsymbol{u} = D_{\boldsymbol{u}} \nabla f$ for $\boldsymbol{u} \in T\mathbb{R}^n$.
(b) Given $\boldsymbol{a}, \boldsymbol{u} \in \mathbb{R}^n$, prove that $h_f(\mathcal{I}_{\boldsymbol{a}}\boldsymbol{u}, \mathcal{I}_{\boldsymbol{a}}\boldsymbol{u}) = (f \circ \boldsymbol{c})''(0)$, where $\boldsymbol{c}(t) = \boldsymbol{a} + t\boldsymbol{u}$.

2.32. Suppose $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ is a smooth map such that $|D\boldsymbol{f}(\boldsymbol{p})\boldsymbol{u}| = |\boldsymbol{u}|$ for all $\boldsymbol{p}, \boldsymbol{u} \in \mathbb{R}^n$. Prove that $\boldsymbol{f}$ is a diffeomorphism which preserves distances; i.e., $|\boldsymbol{f}(\boldsymbol{p}) - \boldsymbol{f}(\boldsymbol{q})| = |\boldsymbol{p} - \boldsymbol{q}|$ for any $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^n$. *Hint:* $\boldsymbol{f}$ preserves the length of curves, so it cannot increase distances. The same is true for any (*a priori* only local) inverse, so it cannot decrease them either.

2.33. Determine the flow of the position vector field $\boldsymbol{P} = \sum_i u^i D_i$. Is $\boldsymbol{P}$ a complete vector field?

2.34. This problem investigates integral curves of vector fields on $\mathbb{R}^n$ of the form

$$\boldsymbol{X} = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij} u^j \right) \boldsymbol{D}_i, \qquad a_{ij} = a_{ji} \in \mathbb{R},$$

without using Example 2.8.1.

(a) Prove that $\boldsymbol{X}(\boldsymbol{u}) = (\boldsymbol{u}, A\boldsymbol{u})$, where $A = (a_{ij})$ is a symmetric matrix, and that $\boldsymbol{c} : I \to \mathbb{R}^n$ is an integral curve of $\boldsymbol{X}$ if and only if $\boldsymbol{c}' = A\boldsymbol{c}$.

(b) By the spectral theorem, there exists an orthogonal matrix P and a diagonal matrix

$$D = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

such that $A = PDP^T$. Show that the integral curve $\boldsymbol{c}$ of $\boldsymbol{X}$ with $\boldsymbol{c}(0) = \boldsymbol{a} \in \mathbb{R}^n$ is given by

$$\boldsymbol{c}(t) = P \cdot \begin{bmatrix} e^{\lambda_1 t} & 0 & \dots & 0 \\ 0 & e^{\lambda_2 t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & e^{\lambda_n t} \end{bmatrix} \cdot P^T \boldsymbol{a}.$$

In particular, $\boldsymbol{X}$ is complete.

2.35. Use the results of the previous exercise to find the integral curve $\boldsymbol{c}$ of the vector field $\boldsymbol{X}$, where $\boldsymbol{X} = (u^2 - u^3)\boldsymbol{D}_1 + (u^1 - u^3)\boldsymbol{D}_2 + (u^1 - u^2)\boldsymbol{D}_3$, if $\boldsymbol{c}(0) = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}^T$.

2.36. Show that $\mathbb{R}^3$ is a Lie algebra with the cross product.

2.37. Let $\boldsymbol{X}$ be a parallel vector field on $\mathbb{R}^n$.

(a) Prove that there exists some $\boldsymbol{u} \in \mathbb{R}^n$ such that $\boldsymbol{X}(\boldsymbol{p}) = \mathcal{I}_{\boldsymbol{p}} \boldsymbol{u}$ for all $\boldsymbol{p} \in \mathbb{R}^n$.

(b) Show that a vector field $\boldsymbol{Y}$ satisfies $[\boldsymbol{X}, \boldsymbol{Y}] \equiv \boldsymbol{0}$ if and only if $\boldsymbol{X}(\boldsymbol{a} + t\boldsymbol{u}) = (\mathcal{I}_{\boldsymbol{a}+t\boldsymbol{u}} \circ \mathcal{I}_{\boldsymbol{a}}^{-1})\boldsymbol{X}(\boldsymbol{a})$ for all $\boldsymbol{a} \in \mathbb{R}^n$, $t \in \mathbb{R}$. Interpret this geometrically.

2.38. Determine the flow $\boldsymbol{\Phi}$ of the vector field $\boldsymbol{Y}$, where

$$\boldsymbol{Y} = -u^3 \boldsymbol{D}_1 - u^4 \boldsymbol{D}_2 + u^1 \boldsymbol{D}_3 + u^2 \boldsymbol{D}_4,$$

see Example 2.8.1. Interpret $\boldsymbol{\Phi}_t : \mathbb{R}^4 \to \mathbb{R}^4$ geometrically.

2.39. Consider the vector fields $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ on $\mathbb{R}^4$ given by

$$\begin{aligned} \boldsymbol{X} &= -u^2 \boldsymbol{D}_1 + u^1 \boldsymbol{D}_2 + u^4 \boldsymbol{D}_3 - u^3 \boldsymbol{D}_4, \\ \boldsymbol{Y} &= -u^3 \boldsymbol{D}_1 - u^4 \boldsymbol{D}_2 + u^1 \boldsymbol{D}_3 + u^2 \boldsymbol{D}_4, \\ \boldsymbol{Z} &= -u^4 \boldsymbol{D}_1 + u^3 \boldsymbol{D}_2 - u^2 \boldsymbol{D}_3 + u^1 \boldsymbol{D}_4. \end{aligned}$$

Prove that the set $\{a\boldsymbol{X} + b\boldsymbol{Y} + c\boldsymbol{Z} \mid a, b, c \in \mathbb{R}\}$ is a 3-dimensional Lie algebra.

2.40. Let $\boldsymbol{X}$, $\boldsymbol{Y}$ denote vector fields on $\mathbb{R}^n$ with flows $\boldsymbol{\Phi}_t$ and $\boldsymbol{\Psi}_t$ respectively. Given $\boldsymbol{p} \in \mathbb{R}^n$, define a curve $\boldsymbol{c}$ on a neighborhood of $\mathbf{0}$ by

$$\boldsymbol{c}(t) = (\boldsymbol{\Psi}_{-t} \circ \boldsymbol{\Phi}_{-t} \circ \boldsymbol{\Psi}_t \circ \boldsymbol{\Phi}_t)(\boldsymbol{p}).$$

We have seen that if $[\boldsymbol{X}, \boldsymbol{Y}] \equiv \mathbf{0}$, then $\boldsymbol{c}(t) = \boldsymbol{p}$ for all t. The object of this exercise is to show that even when the bracket of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is not zero (so that $\boldsymbol{c}$ is not a constant curve), its derivative $\boldsymbol{c}'(0)$ nevertheless vanishes. To see this, define maps $\boldsymbol{F}_i : U \to \mathbb{R}^n$ on a neighborhood $U \subset \mathbb{R}^2$ of the origin by

$$\begin{aligned}
\boldsymbol{F}_1(t,s) &= (\boldsymbol{\Psi}_t \circ \boldsymbol{\Phi}_s)(\boldsymbol{p}),\\
\boldsymbol{F}_2(t,s) &= (\boldsymbol{\Phi}_{-t} \circ \boldsymbol{\Psi}_s \circ \boldsymbol{\Phi}_s)(\boldsymbol{p}),\\
\boldsymbol{F}_3(t,s) &= (\boldsymbol{\Psi}_{-t} \circ \boldsymbol{\Phi}_{-s} \circ \boldsymbol{\Psi}_s \circ \boldsymbol{\Phi}_s)(\boldsymbol{p}).
\end{aligned}$$

(a) Prove that

$$\boldsymbol{F}_{1*}\boldsymbol{D}_1 = \boldsymbol{Y} \circ \boldsymbol{F}_1, \quad \boldsymbol{F}_{2*}\boldsymbol{D}_1 = -\boldsymbol{X} \circ \boldsymbol{F}_2, \quad \boldsymbol{F}_{3*}\boldsymbol{D}_1 = -\boldsymbol{Y} \circ \boldsymbol{F}_3,$$

and that $\boldsymbol{F}_{1*}\boldsymbol{D}_2(0,s) = (\boldsymbol{X} \circ \boldsymbol{F}_1)(0,s)$.

(b) Notice that $\boldsymbol{c}(t) = \boldsymbol{F}_3(t,t)$, $\boldsymbol{F}_2(0,t) = \boldsymbol{F}_1(t,t)$, and $\boldsymbol{F}_3(0,t) = \boldsymbol{F}_2(t,t)$. Use the chain rule to show that $\boldsymbol{c}'(0) = \mathbf{0}$.

2.41. Two Lie algebras $(V_1, [,]_1)$ and $(V_2, [,]_2)$ are said to be isomorphic if there exists a vector space isomorphism $L : V_1 \to V_2$ that preserves Lie brackets; i.e.,

$$L[\boldsymbol{u}, \boldsymbol{v}]_1 = [L\boldsymbol{u}, L\boldsymbol{v}]_2, \qquad \boldsymbol{u}, \boldsymbol{v} \in V_1.$$

In this case, L is called a Lie algebra isomorphism. Show that the Lie algebras from Exercises 2.36 and 2.39 are isomorphic. Give an example of a 3-dimensional Lie algebra that is not isomorphic to these. Thus, unlike vector spaces, Lie algebras of the same dimension need not be isomorphic.

2.42. (a) Show that if L is a self-adjoint operator, then so is e^L.

(b) Show that if L is a skew-adjoint operator, then e^L is orthogonal.

2.43. This exercise generalizes the discussion from Example 2.8.1 regarding the system of ODEs $\boldsymbol{c}' = A\boldsymbol{c}$ to the case when A is no longer constant.

(a) Prove that $\exp : M_{n,n} \cong \mathbb{R}^{n^2} \to M_{n,n}$ is differentiable with derivative $D\exp(\mathbf{0})A = \exp(\mathbf{0})A = A$ at the origin.

(b) Show that if $AB = BA$, then $D\exp(B)A = \exp(B)A = A\exp(B)$.

(c) If $\boldsymbol{A} : \mathbb{R} \to M_{n,n}$ is a smooth curve, prove that $\boldsymbol{c} : \mathbb{R} \to M_{n,n}$, where $\boldsymbol{c}(t) = \exp(\boldsymbol{A}(t))$, has derivative $\boldsymbol{c}'(t) = \boldsymbol{A}'(t)\boldsymbol{c}(t)$, provided $\boldsymbol{A}(t)\boldsymbol{A}'(t) = \boldsymbol{A}'(t)\boldsymbol{A}(t)$ for all t.

(d) With $\boldsymbol{A}$ as in (c), show that the system of linear ODEs $\boldsymbol{c}'(t) = \boldsymbol{A}(t)\boldsymbol{c}(t)$ has as solution

$$\boldsymbol{c}(t) = \exp\left(\int_0^t \boldsymbol{A}\right)\boldsymbol{c}(0),$$

provided $\boldsymbol{A}(s)\boldsymbol{A}(t) = \boldsymbol{A}(t)\boldsymbol{A}(s)$ for all s, t. Here, the (i,j)-th entry of $\int_0^t \boldsymbol{A}$ is

$$u^{ij} \circ \int_0^t \boldsymbol{A} = \int_0^t u^{ij} \circ \boldsymbol{A}.$$

2.44. Use the results from the previous exercise to solve the system

$$x'(t) = -x(t) - ty(t)$$
$$y'(t) = tx(t) - y(t)$$

2.45. Suppose $\boldsymbol{\Phi} : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^n$ is differentiable. For each $t \in \mathbb{R}$, define $\boldsymbol{\Phi}_t : \mathbb{R}^n \to \mathbb{R}^n$ by $\boldsymbol{\Phi}_t(\boldsymbol{a}) = \boldsymbol{\Phi}(t, \boldsymbol{a})$, $\boldsymbol{a} \in \mathbb{R}^n$. $\{\boldsymbol{\Phi}_t\}_{t\in\mathbb{R}}$ is called a *one-parameter group of diffeomorphisms of* $\mathbb{R}^n$ if
(1) $\boldsymbol{\Phi}_0 = 1_{\mathbb{R}^n}$, and
(2) $\boldsymbol{\Phi}_{s+t} = \boldsymbol{\Phi}_s \circ \boldsymbol{\Phi}_t$, $s, t \in \mathbb{R}$.
Prove that one-parameter groups are in 1-1 correspondence with complete vector fields on $\mathbb{R}^n$.

2.46. Let C be a closed set in $\mathbb{R}^n$, and $f : C \to \mathbb{R}$ a smooth function. Show that if U is any open set containing C, then there exists a smooth function $g : \mathbb{R}^n \to \mathbb{R}$, such that g agrees with f on C, and the support of g lies in U. Show this is no longer necessarily true if C is not assumed to be closed. In particular, any smooth function that is defined on a closed subset of Euclidean space is extendable to a smooth function on all of Euclidean space.

2.47. Let $\boldsymbol{X}$ be a smooth vector field on a closed set $C \subset \mathbb{R}^n$ (as for functions, this means that there exists a smooth vector field $\tilde{\boldsymbol{X}}$ on some open set U containing C such that the restriction of $\tilde{\boldsymbol{X}}$ to C equals $\boldsymbol{X}$). Show that $\boldsymbol{X}$ is extendable to a smooth vector field on $\mathbb{R}^n$; i.e., there exists a smooth vector field on $\mathbb{R}^n$ whose restriction to C equals $\boldsymbol{X}$.

3 Manifolds

Differential geometry arguably began with the study of surfaces in three-dimensional Euclidean space. In this chapter, we give a precise meaning to the word "surface", and generalize this concept to higher dimensions.

3.1 Submanifolds of Euclidean space

Definition 3.1.1. A subset $M \subset \mathbb{R}^{n+k}$ is said to be an *n-dimensional submanifold* of $\mathbb{R}^{n+k}$ if each $\boldsymbol{p} \in M$ admits an open neighborhood U in $\mathbb{R}^{n+k}$ and there exists a one-to-one differentiable map $\boldsymbol{h} : \mathbb{R}^n \supset V \to \mathbb{R}^{n+k}$ defined on some open set V, such that
(1) $\boldsymbol{h}$ has maximal rank (= n) everywhere;
(2) $\boldsymbol{h}(V) = U \cap M$, and
(3) $\boldsymbol{h}^{-1} : U \cap M \to V$ is continuous.

Thus, loosely speaking, an n-dimensional submanifold of $\mathbb{R}^{n+k}$ is a subset M of $\mathbb{R}^{n+k}$ such that each point of M has an open neighborhood (in M) that "looks like" an open set in $\mathbb{R}^n$. In the terminology of Exercise 2.27, M is a submanifold if for any $\boldsymbol{p} \in M$ there exists an imbedding of an open set in $\mathbb{R}^n$ into $\mathbb{R}^{n+k}$ whose image is an open neighborhood of $\boldsymbol{p}$ in M. Condition (3) in the above definition is equivalent to requiring that open sets in V get mapped by $\boldsymbol{h}$ to open sets of M. A subset M for which the first two conditions hold but the third one does not is called an *immersed submanifold*.

The pair $(V, \boldsymbol{h})$ in the definition is called a *local parametrization* of M around $\boldsymbol{p}$. The inverse of a parametrization, or more precisely, the pair $(\boldsymbol{h}(V), \boldsymbol{h}^{-1})$ is called a *chart* of M around $\boldsymbol{p}$. It is common practice to denote charts by $(U, \boldsymbol{x})$, $(V, \boldsymbol{y})$, and so on. A collection of charts whose domains form an open cover of M is called an *atlas* of M. For the sake of brevity, and when there is no possible confusion about what the "ambient space" $\mathbb{R}^{n+k}$ is, we sometimes say that M is an n-dimensional manifold, or just an n-manifold. The dimension is often implicitly specified by adding a superscript, as in "M^n".

Notice also that any submanifold of $\mathbb{R}^{n+k}$ is a submanifold of $\mathbb{R}^{n+k+l}$ for $l \in \mathbb{N}$ if one composes every parametrization with the map $\mathbb{R}^{n+k}$ to $\mathbb{R}^{n+k+l}$ which sends $\boldsymbol{p}$ to $(\boldsymbol{p}, \boldsymbol{0}) \in \mathbb{R}^{n+k} \times \mathbb{R}^l$. It is customary to choose the ambient space with the smallest dimension.

A 2-dimensional submanifold of $\mathbb{R}^3$ is often called a *surface*. More generally, a *hypersurface* is an $(n-1)$-dimensional submanifold of $\mathbb{R}^n$ for some $n \in \mathbb{N}$.

Any open set U in $\mathbb{R}^n$ is a trivial example of a submanifold of $\mathbb{R}^n$, with the atlas $\{(U, \imath)\}$, where $\imath : U \to \mathbb{R}^n$ is the inclusion map. More interesting manifolds are described below:

Examples 3.1.1. (i) Let $f : V \to \mathbb{R}$ be a differentiable function on an open set V in $\mathbb{R}^2$. Then the graph

$$M = \{(\boldsymbol{a}, f(\boldsymbol{a})) \mid \boldsymbol{a} \in V\}$$

of f is a 2-dimensional submanifold of $\mathbb{R}^3$: The map $\boldsymbol{h} : V \to \mathbb{R}^3$, where $\boldsymbol{h}(\boldsymbol{a}) = (\boldsymbol{a}, f(\boldsymbol{a}))$, is certainly one-to-one with maximal rank everywhere, and the set U in the definition may be taken to be $V \times \mathbb{R}$. In the same way, if V_0 is open in V, then $\boldsymbol{h}(V_0) = M \cap (V_0 \times \mathbb{R})$ is open in M, so that $\boldsymbol{h}^{-1}$ is continuous. More generally, the graph of a function defined on an open set in $\mathbb{R}^n$ is an n-dimensional submanifold of $\mathbb{R}^{n+1}$.

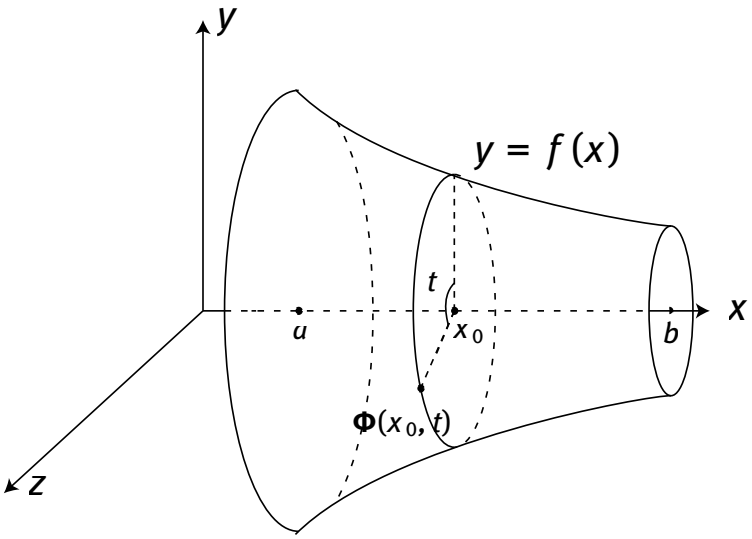


Fig. 3.1: Parametrizing a surface of revolution

(ii) If $f : (a, b) \to \mathbb{R}$ is a positive function, the *surface of revolution* obtained by revolving the graph of f around the x-axis is the 2-dimensional submanifold of $\mathbb{R}^3$ described by two parametrizations $\boldsymbol{h} : (a, b)\times(0, 2\pi) \to \mathbb{R}^3$, and $\boldsymbol{k} : (a, b)\times(\varepsilon, 2\pi + \varepsilon) \to \mathbb{R}^3$, $0 < \varepsilon < 2\pi$, both given by the same formula

$$\boldsymbol{h}(x, t) = \boldsymbol{k}(x, t) = (x, f(x)\cos t, f(x)\sin t).$$

The first parametrization covers the whole surface except for the original graph, which is why a second one is needed. The Jacobian matrix of either equals

$$\begin{bmatrix} 1 & 0 \\ f'(x)\cos t & -f(x)\sin t \\ f'(x)\sin t & f(x)\cos t \end{bmatrix},$$

which has rank 2, unless f is zero somewhere, but that was ruled out at the beginning. This establishes the first condition; the other two are straightforward, and left to the reader.

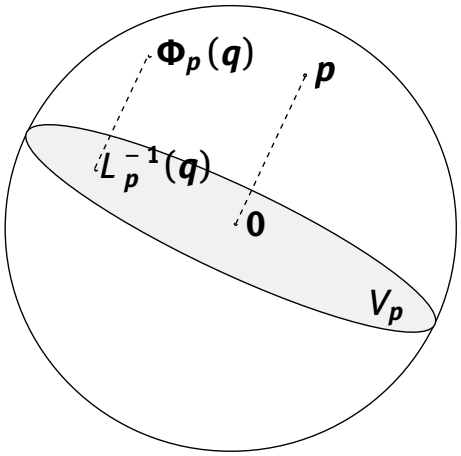


Fig. 3.2: A local parametrization of $S^n(r)$ around $\boldsymbol{p}$

(iii) The sphere $S^n(r) = \{\boldsymbol{p} \in \mathbb{R}^{n+1} \mid |\boldsymbol{p}| = r\}$ of radius $r > 0$ centered at the origin is an n-dimensional submanifold of $\mathbb{R}^{n+1}$. A local parametrization $(V, \boldsymbol{h})$ around the north pole $r\boldsymbol{e}_{n+1} = (0, \dots, 0, r)$ is given by

$$\boldsymbol{h}(\boldsymbol{a}) = (\boldsymbol{a}, \sqrt{1 - |\boldsymbol{a}|^2}), \qquad \boldsymbol{a} \in V := B_r(\boldsymbol{0}) \subset \mathbb{R}^n.$$

The image of $\boldsymbol{h}$ is the open northern hemisphere. This parametrization can be modified to yield one around any $\boldsymbol{p} \in S^n(r)$: Let $L_{\boldsymbol{p}} : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ be any linear isometry that maps $\boldsymbol{p}$ to $r\boldsymbol{e}_{n+1}$ (for example, extend $\boldsymbol{p}/r$ to an orthonormal basis $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n, \boldsymbol{p}/r$ of $\mathbb{R}^{n+1}$, define $L_{\boldsymbol{p}}(\boldsymbol{u}_i) = \boldsymbol{e}_i$ for $i \le n$, $L_{\boldsymbol{p}}(\boldsymbol{p}/r) = \boldsymbol{e}_{n+1}$, and extend linearly; i.e., $L_{\boldsymbol{p}}$ maps the vector $\sum_{i=1}^n a_i \boldsymbol{u}_i + a_{n+1}(\boldsymbol{p}/r)$ to $(a_1, \dots, a_{n+1})$). If $V_{\boldsymbol{p}}$ denotes the set of all $\boldsymbol{q} \in \boldsymbol{p}^\perp$ with norm less than r, then $L_{\boldsymbol{p}}$ maps $V_{\boldsymbol{p}}$ isometrically onto the domain V of the parametrization $(V, \boldsymbol{h})$ above. The pair $(V, \boldsymbol{h}_{\boldsymbol{p}})$ is now a parametrization around $\boldsymbol{p}$, if $\boldsymbol{h}_{\boldsymbol{p}} = L_{\boldsymbol{p}}^{-1} \circ \boldsymbol{h}$.

(iv) Although the above parametrizations are easy to visualize, they are not the most efficient in the sense that a fairly large number of them is needed to cover the whole sphere. It is possible to give an atlas with only two charts: *stereographic projection* $\boldsymbol{x} : S^n(r) \setminus \{r\boldsymbol{e}_{n+1}\} \to \mathbb{R}^n$ from the north pole $r\boldsymbol{e}_{n+1} = (\boldsymbol{0}, r)$ is the map that

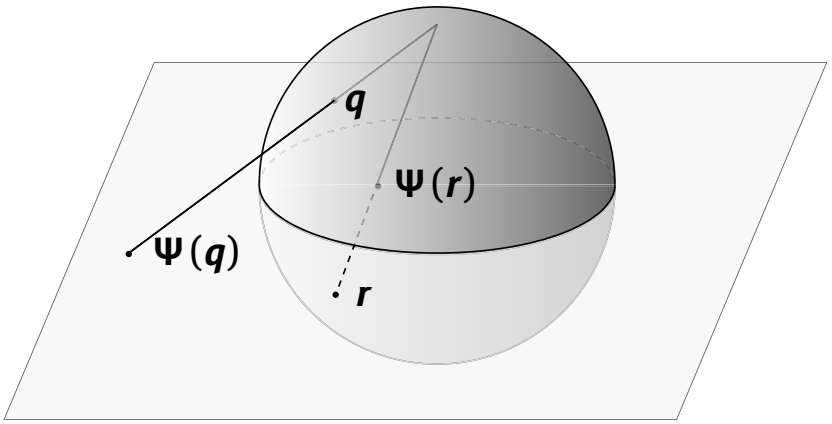


Fig. 3.3: Stereographic projection from the north pole

assigns to each $\boldsymbol{q}$ of $S^n(r)$ distinct from the north pole the unique point where the line from the north pole to $\boldsymbol{q}$ intersects the hyperplane $\mathbb{R}^n \times 0 = \{(\boldsymbol{p}, 0) \in \mathbb{R}^{n+1} \mid \boldsymbol{p} \in \mathbb{R}^n\}$.

To find the point of intersection, notice that the line through the north pole and any other $\boldsymbol{q} \in S^n(r)$ can be parametrized by the curve

$$t \mapsto (\mathbf{0}, r) + t(\boldsymbol{q} - (\mathbf{0}, r)) = t\boldsymbol{q} + (\mathbf{0}, (1-t)r),$$

with $\mathbf{0}$ denoting the origin in $\mathbb{R}^n$. Its image intersects the equatorial hyperplane when $t = r/(r - u^{n+1}(\boldsymbol{q}))$, so that

$$\boldsymbol{x}(\boldsymbol{q}) = \frac{r}{r - u^{n+1}(\boldsymbol{q})}(u^1(\boldsymbol{q}), \dots, u^n(\boldsymbol{q})), \qquad \boldsymbol{q} \in S^n(r) \setminus \{r\boldsymbol{e}_{n+1}\}.$$

The inverse $\boldsymbol{h} : \mathbb{R}^n \to S^n(r)$ of $\boldsymbol{x}$ is the parametrization given by

$$\boldsymbol{h}(\boldsymbol{p}) = \frac{1}{|\boldsymbol{p}|^2 + r^2}(2r^2\boldsymbol{p}, (|\boldsymbol{p}|^2 - r^2)r), \qquad \boldsymbol{p} \in \mathbb{R}^n.$$

Combining stereographic projection from the north pole with the one from the south pole now yields an atlas of the sphere. The atlas is minimal in terms of the amount of charts that are needed. A space admitting an atlas with only one chart – as is the case of a sphere with one point deleted – is considered trivial. This perspective will change once curvature is introduced.

(v) The curve $\boldsymbol{c} : (0, 2\pi) \to \mathbb{R}^2$, with $\boldsymbol{c}(t) = (\sin t, \sin(2t))$, parametrizes a lemniscate. The image is an immersed submanifold M of $\mathbb{R}^2$: the origin in the plane is the point $\boldsymbol{c}(\pi)$, and the set $\boldsymbol{c}(3\pi/4, 5\pi/4)$ is not an open neighborhood of the origin in M, since any such neighborhood must contain points $\boldsymbol{c}(t)$ for values of t arbitrarily close to 0 and 2π.

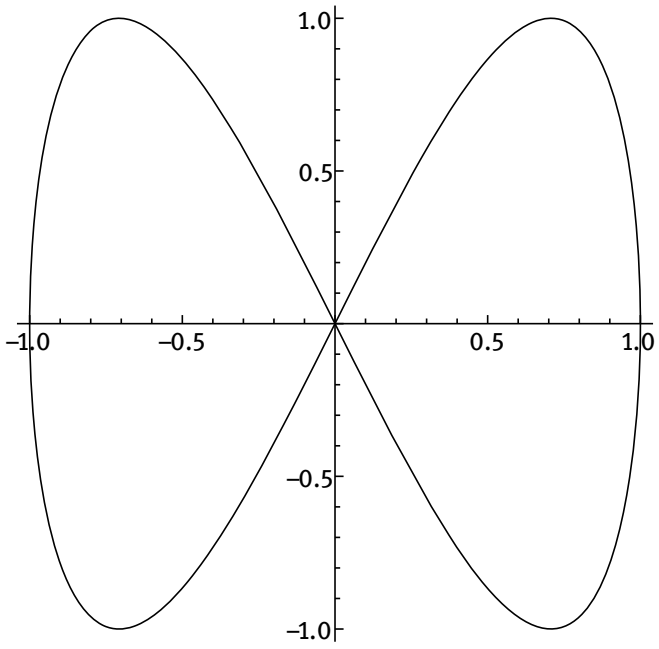


Fig. 3.4: The lemniscate in (v)

(vi) If M is an n-manifold, then so is any open subset U of M: in fact, any parametrization $(V, \boldsymbol{h})$ of M around some $\boldsymbol{p} \in U$, when restricted to $V \cap \boldsymbol{h}^{-1}(U)$, is a parametrization of U around $\boldsymbol{p}$.

(vii) If $M^n \subset \mathbb{R}^k$ and $N^m \subset \mathbb{R}^l$ are manifolds, then the Cartesian product

$$M \times N = \{(\boldsymbol{p}, \boldsymbol{q}) \in \mathbb{R}^k \times \mathbb{R}^l \mid \boldsymbol{p} \in M,\ \boldsymbol{q} \in N\}$$

is a (n+m)-dimensional submanifold of $\mathbb{R}^{k+l}$: any parametrizations $\boldsymbol{h} : \mathbb{R}^n \supset U \to \mathbb{R}^k$ of M and $\boldsymbol{k} : \mathbb{R}^m \supset V \to \mathbb{R}^l$ of N generate a parametrization $\boldsymbol{h} \times \boldsymbol{k}$ of $M \times N$, where $\boldsymbol{h} \times \boldsymbol{k}(\boldsymbol{a}, \boldsymbol{b}) = (\boldsymbol{h}(\boldsymbol{a}), \boldsymbol{k}(\boldsymbol{b}))$.

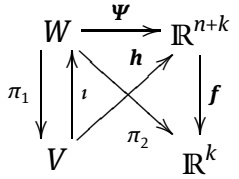
There is a particularly useful and simple way of constructing manifolds: $\boldsymbol{q} \in \mathbb{R}^k$ is said to be a *regular value* of a map $\boldsymbol{f} : \mathbb{R}^n \supset U \to \mathbb{R}^k$ if the derivative $\boldsymbol{f}_*$ has maximal rank at every $\boldsymbol{p} \in \boldsymbol{f}^{-1}(\boldsymbol{q})$.

Theorem 3.1.1. *Let U be open in $\mathbb{R}^{n+k}$, and $\boldsymbol{f} : U \to \mathbb{R}^k$. If $\boldsymbol{a} \in \mathbb{R}^k$ is a regular value of $\boldsymbol{f}$ and $M = \boldsymbol{f}^{-1}(\boldsymbol{a})$ is nonempty, then M is an n-dimensional submanifold of $\mathbb{R}^{n+k}$.*

Proof. We shall construct a parametrization $\boldsymbol{h}$ around any $\boldsymbol{p} \in M$. It may be assumed first of all that $\boldsymbol{a} = \boldsymbol{0}$, for if the result holds when $\boldsymbol{a} = \boldsymbol{0}$, we may, in the general case, apply it to $\boldsymbol{g}$, where $\boldsymbol{g}(\boldsymbol{p}) = \boldsymbol{f}(\boldsymbol{p}) - \boldsymbol{a}$. Similarly, $\boldsymbol{p}$ may be assumed to be $\boldsymbol{0}$.
By the implicit function theorem, there exists a neighborhood W of $\boldsymbol{0}$ in $\mathbb{R}^{n+k}$ and a diffeomorphism $\boldsymbol{\Psi} : W \to \boldsymbol{\Psi}(W) \subset \mathbb{R}^{n+k}$ such that $\boldsymbol{f} \circ \boldsymbol{\Psi}$ equals the restriction to W of the projection $\pi_2 : \mathbb{R}^{n+k} = \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^k$ onto the second factor. Let π_1 denote the projection onto the first factor $\mathbb{R}^n$, $V = \pi_1(W)$, $\imath : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^k$ the inclusion $\imath(\boldsymbol{u}) = (\boldsymbol{u}, \boldsymbol{0})$, and define $\boldsymbol{h} : V \to \mathbb{R}^{n+k}$ to be $\boldsymbol{\Psi} \circ \imath$. $\boldsymbol{h}$ is one-to-one of maximal rank, being the composition of two maps that enjoy those properties. Furthermore,

$$\boldsymbol{f} \circ \boldsymbol{h} = \boldsymbol{f} \circ \boldsymbol{\Psi} \circ \imath = \pi_2 \circ \imath = \boldsymbol{0},$$

so that $\boldsymbol{h}(V) \subset M$.



We claim that $\boldsymbol{h}(V) = M \cap \boldsymbol{\Psi}(W)$; more generally, if $V_0 \subset V$ is open, we claim that

$$\boldsymbol{h}(V_0) = M \cap (\boldsymbol{\Psi}(\pi_{1|W}^{-1}(V_0))).$$

(If V is a subset of the domain U of a map f, the notation $f_{|U}$ refers to the *restriction* of f to U; i.e., $f_{|U} := f \circ \jmath$, where $\jmath : U \hookrightarrow V$ is the inclusion map.) Notice that since $\boldsymbol{\Psi}(\pi_{1|W}^{-1}(V_0))$ is open in $\mathbb{R}^{n+k}$, the claim, once established, will imply that $\boldsymbol{h}$ maps open sets to open sets in M, thereby concluding the proof.

One inclusion is already clear:

$$\boldsymbol{h}(V_0) = \boldsymbol{\Psi} \circ \imath(V_0) \subset \boldsymbol{\Psi}(\pi_{1|W}^{-1}(V_0)),$$

and, of course, it is also contained in M, since $\boldsymbol{h}(V)$ is. Conversely, consider any $\boldsymbol{q} \in M \cap (\boldsymbol{\Psi}(\pi_{1|W}^{-1}(V_0)))$. Then $\boldsymbol{q} = \boldsymbol{\Psi}(\boldsymbol{r})$ for some unique $\boldsymbol{r} \in \pi_{1|W}^{-1}(V_0)$, and $\pi_2(\boldsymbol{r}) = (\boldsymbol{f} \circ \boldsymbol{\Psi})(\boldsymbol{r}) = \boldsymbol{f}(\boldsymbol{q}) = \boldsymbol{0}$. Thus,

$$\boldsymbol{r} = (\tilde{\boldsymbol{r}}, \boldsymbol{0}) = \imath(\tilde{\boldsymbol{r}})$$

for a unique $\tilde{\boldsymbol{r}} = \pi_1(\boldsymbol{r}) \in V_0$, and $\boldsymbol{q} = \boldsymbol{\Psi}(\boldsymbol{r}) = \boldsymbol{h}(\tilde{\boldsymbol{r}}) \in \boldsymbol{h}(V_0)$, as claimed. □

Example 3.1.2. Given $r > 0$, the function $f : \mathbb{R}^{n+1} \to \mathbb{R}$, with $f(\boldsymbol{a}) = |\boldsymbol{a}|^2 - r^2$, has 0 as regular value, since $[Df(\boldsymbol{p})] = 2\boldsymbol{p} \neq \boldsymbol{0}$ if $\boldsymbol{p} \in f^{-1}(0) = S^n(r)$. This yields a shorter proof of the fact that the sphere is a submanifold of Euclidean space.

Definition 3.1.2. Let M be an n-dimensional submanifold of Euclidean space $\mathbb{R}^{n+k}$. The *tangent space $M_{\boldsymbol{p}}$ of M at $\boldsymbol{p} \in M$* is the collection of velocity vectors $\dot{\boldsymbol{c}}(0)$ of all curves $\boldsymbol{c} : I \to M$ defined on some open interval I containing 0 such that $\boldsymbol{p} = \boldsymbol{c}(0)$.

Thus, $M_{\boldsymbol{p}}$ is a subset of $\mathbb{R}^{n+k}_{\boldsymbol{p}}$. What is not immediately clear is that it is actually a subspace:

Theorem 3.1.2. *If M is an n-dimensional submanifold of $\mathbb{R}^{n+k}$, then $M_{\boldsymbol{p}}$ is an n-dimensional subspace of $\mathbb{R}^{n+k}_{\boldsymbol{p}}$ for every $\boldsymbol{p} \in M$.*

Proof. Consider a parametrization $(V, \boldsymbol{h})$ around $\boldsymbol{p}$, $V \subset \mathbb{R}^n$. By composing $\boldsymbol{h}$ with a translation if necessary, it may be assumed that $\boldsymbol{0} \in V$, and that $\boldsymbol{h}(\boldsymbol{0}) = \boldsymbol{p}$. If $\boldsymbol{c} : I \to M$ is a curve passing through $\boldsymbol{p}$ at 0, then, after restricting the domain of $\boldsymbol{c}$ if necessary, $\boldsymbol{h}^{-1} \circ \boldsymbol{c}$ is a curve in V, and

$$\dot{\boldsymbol{c}}(0) = \boldsymbol{c}_{*0}D(0) = (\boldsymbol{h} \circ \boldsymbol{h}^{-1} \circ \boldsymbol{c})_{*0}D(0) = \boldsymbol{h}_{*\boldsymbol{0}}(\boldsymbol{h}^{-1} \circ \boldsymbol{c})_{*0}D(0) = \boldsymbol{h}_{*\boldsymbol{0}}\boldsymbol{v},$$

where $\boldsymbol{v} = (\boldsymbol{h}^{-1} \circ \boldsymbol{c})_{*0}D(0) \in \mathbb{R}^n_{\boldsymbol{0}}$ is the velocity vector of the curve $\boldsymbol{h}^{-1} \circ \boldsymbol{c}$ in V. Thus, $M_{\boldsymbol{p}} \subset \boldsymbol{h}_{*\boldsymbol{0}}\mathbb{R}^n_{\boldsymbol{0}}$. Conversely, if $\boldsymbol{c}$ is a curve in V with $\boldsymbol{c}(0) = \boldsymbol{0}$, then $\boldsymbol{h} \circ \boldsymbol{c}$ is a curve in M passing through $\boldsymbol{p}$ at 0, so that $\boldsymbol{h}_*(\dot{\boldsymbol{c}}(0)) \in M_{\boldsymbol{p}}$. We have therefore shown that

$$M_{\boldsymbol{p}} = \boldsymbol{h}_*\mathbb{R}^n_{\boldsymbol{0}}. \tag{3.1.1}$$

Since $\boldsymbol{h}$ has rank n, this concludes the proof. □

If instead of parametrizations, the manifold is given by $\boldsymbol{f}^{-1}(\boldsymbol{0})$, where $\boldsymbol{0}$ is a regular value of $\boldsymbol{f}$, the following description of tangent spaces holds:

Proposition 3.1.1. *If $M = \boldsymbol{f}^{-1}(\boldsymbol{a})$, where $\boldsymbol{a}$ is a regular value of the map $\boldsymbol{f} : \mathbb{R}^{n+k} \supset U \to \mathbb{R}^k$, then*

$$M_{\boldsymbol{p}} = \ker \boldsymbol{f}_{*\boldsymbol{p}}$$

for any $\boldsymbol{p} \in M$.

Proof. Since both sides of the above identity are vector spaces of the same dimension n, it suffices to show that one is contained in the other. So consider the velocity vector $\boldsymbol{u} = \dot{\boldsymbol{c}}(0) \in M_{\boldsymbol{p}}$ of a curve $\boldsymbol{c}$ in M at time 0. Since $\boldsymbol{c}(I) \subset M = \boldsymbol{f}^{-1}(\boldsymbol{a})$, $\boldsymbol{f} \circ \boldsymbol{c} \equiv \boldsymbol{a}$. Thus,

$$\boldsymbol{f}_{*\boldsymbol{p}}\boldsymbol{u} = (\boldsymbol{f} \circ \boldsymbol{c})_{*0}D(0) = \boldsymbol{0},$$

and $\boldsymbol{u} \in \ker \boldsymbol{f}_{*\boldsymbol{p}}$. This shows that $M_{\boldsymbol{p}} \subset \ker \boldsymbol{f}_{*\boldsymbol{p}}$, as claimed. □

Examples 3.1.3. (i) If M is the graph of a function $f : \mathbb{R}^n \supset U \to \mathbb{R}$, then M admits the parametrization $\boldsymbol{h} : U \to \mathbb{R}^{n+1}$, $\boldsymbol{h}(\boldsymbol{a}) = (\boldsymbol{a}, f(\boldsymbol{a}))$. By (3.1.1), the tangent space of M at $\boldsymbol{h}(\boldsymbol{a})$ is the image of $\boldsymbol{h}_{*\boldsymbol{a}}$. Now,

$$[D\boldsymbol{h}(\boldsymbol{a})] = \begin{bmatrix} 1_{\mathbb{R}^n} \\ [Df(\boldsymbol{a})] \end{bmatrix},$$

so that $D\boldsymbol{h}(\boldsymbol{a})\boldsymbol{e}_i = \boldsymbol{e}_i + D_i f(\boldsymbol{a})\boldsymbol{e}_{n+1}$. It follows that

$$M_{\boldsymbol{h}(\boldsymbol{a})} = \operatorname{span}\{\boldsymbol{D}_i(\boldsymbol{h}(\boldsymbol{a})) + (D_i f)(\boldsymbol{a})\boldsymbol{D}_{n+1}(\boldsymbol{h}(\boldsymbol{a})) \mid i = 1, \ldots, n\}.$$

Another way of describing the tangent space at $\boldsymbol{h}(a)$ is to notice that $M = g^{-1}(0)$, where $g : \mathbb{R}^{n+1} \supset U \times \mathbb{R} \to \mathbb{R}$ is given by $g(a_1, \ldots, a_{n+1}) = a_{n+1} - f(a_1, \ldots, a_n)$. Since

$$Dg(a_1, \ldots, a_{n+1}) = \begin{bmatrix} -D_1 f(a_1, \ldots, a_n) & \ldots & -D_n f(a_1, \ldots, a_n) & 1 \end{bmatrix},$$

0 is a regular value of g and Proposition 3.1.1 says that the tangent space is the orthogonal complement of the gradient of g: for $\boldsymbol{a}, \boldsymbol{u} \in \mathbb{R}^{n+1}$,

$$\langle \nabla g(\boldsymbol{a}), \mathcal{I}_{\boldsymbol{a}}\boldsymbol{u} \rangle = \langle Dg(\boldsymbol{a})^T, \boldsymbol{u} \rangle = Dg(\boldsymbol{a})\boldsymbol{u},$$

so that $\mathcal{I}_{\boldsymbol{a}}\boldsymbol{u}$ is orthogonal to $\nabla g(\boldsymbol{a})$ if and only if $\boldsymbol{u}$ belongs to the kernel of $Dg(\boldsymbol{a})$, or equivalently $\mathcal{I}_{\boldsymbol{a}}\boldsymbol{u}$ belongs to the kernel of $g_{*\boldsymbol{a}}$. The reader should check that the two descriptions of the tangent space agree.

(ii) If M is the sphere of radius r centered at the origin in $\mathbb{R}^{n+1}$, then $M = f^{-1}(0)$, where $f(\boldsymbol{a}) = |\boldsymbol{a}|^2 - r^2$. It was computed earlier that $[Df(\boldsymbol{a})] = 2\boldsymbol{a}^T$. This means that

$$\ker Df(\boldsymbol{a}) = \{\boldsymbol{u} \in \mathbb{R}^{n+1} \mid \boldsymbol{a}^T\boldsymbol{u} = \langle \boldsymbol{a}, \boldsymbol{u} \rangle = 0\} = \boldsymbol{a}^{\perp},$$

and therefore

$$S^n(r)_{\boldsymbol{a}} = \mathcal{I}_{\boldsymbol{a}}\boldsymbol{a}^{\perp}.$$

(iii) Suppose $\boldsymbol{h} : \mathbb{R}^n \supset U \to M$ and $\boldsymbol{k} : \mathbb{R}^m \supset V \to N$ are local parametrizations of M and N respectively. The canonical isomorphism

$$\begin{aligned} \mathbb{R}^n_{\boldsymbol{p}} \times \mathbb{R}^m_{\boldsymbol{q}} &\xrightarrow{\cong} (\mathbb{R}^n \times \mathbb{R}^m)_{(\boldsymbol{p},\boldsymbol{q})} \\ ((\boldsymbol{p}, \boldsymbol{u}), (\boldsymbol{q}, \boldsymbol{v})) &\longmapsto ((\boldsymbol{p}, \boldsymbol{q}), (\boldsymbol{u}, \boldsymbol{v})) \end{aligned}$$

induces a canonical isomorphism $M_{\boldsymbol{p}} \times N_{\boldsymbol{q}} \cong (M \times N)_{(\boldsymbol{p},\boldsymbol{q})}$ by means of the commutative diagram

$$\begin{array}{ccc} \mathbb{R}^n_{\boldsymbol{p}} \times \mathbb{R}^m_{\boldsymbol{q}} & \xrightarrow{\boldsymbol{h}_{*\boldsymbol{p}} \times \boldsymbol{k}_{*\boldsymbol{q}}} & M_{\boldsymbol{p}} \times N_{\boldsymbol{q}} \\ \cong\downarrow & & \downarrow\cong \\ (\mathbb{R}^n \times \mathbb{R}^m)_{(\boldsymbol{p},\boldsymbol{q})} & \xrightarrow[(\boldsymbol{h},\boldsymbol{k})_{*(\boldsymbol{p},\boldsymbol{q})}]{} & (M \times N)_{(\boldsymbol{p},\boldsymbol{q})} \end{array}$$

3.2 Differentiable maps on manifolds

Definition 3.2.1. Let $M^n, N^l \subset \mathbb{R}^k$ be manifolds, $\imath : N \hookrightarrow \mathbb{R}^k$ the inclusion map. A map $\boldsymbol{f} : N \to M$ is said to be *differentiable* or *smooth* if for every $\boldsymbol{p} \in N$ and any parametrization $(V, \boldsymbol{h})$ of N around $\boldsymbol{p}$, the map

$$\imath \circ \boldsymbol{f} \circ \boldsymbol{h} : V \to \mathbb{R}^k$$

is differentiable.

Examples and Remarks 3.2.1. (i) The identity map $1_M : M \to M$ is differentiable, since this is equivalent to requiring that every parametrization be differentiable as a map into the ambient Euclidean space.

(ii) In the above definition, it suffices to check smoothness of $\imath \circ \boldsymbol{f} \circ \boldsymbol{h}$ for *some* parametrization $(V, \boldsymbol{h})$ around every $\boldsymbol{p}$: If $\tilde{\boldsymbol{h}}$ is another parametrization around $\boldsymbol{p}$, then locally $\imath \circ \boldsymbol{f} \circ \tilde{\boldsymbol{h}} = (\imath \circ \boldsymbol{f} \circ \boldsymbol{h}) \circ (\boldsymbol{h}^{-1} \circ \tilde{\boldsymbol{h}})$ which is a composition of smooth maps by the following:

Theorem 3.2.1. *Let $(V_i, \boldsymbol{h}_i)$ be two parametrizations around $\boldsymbol{p} \in M^n$. Then*

$$\boldsymbol{h}_2^{-1} \circ \boldsymbol{h}_1 : \mathbb{R}^n \supset V_1 \cap \boldsymbol{h}_1^{-1}(\boldsymbol{h}_2(V_2)) \to \boldsymbol{h}_2^{-1}(\boldsymbol{h}_1(V_1) \cap \boldsymbol{h}_2(V_2)) \subset \mathbb{R}^n$$

is a diffeomorphism.

Proof. The map is bijective, continuous, and has continuous inverse. Once we show it is differentiable, it will follow by symmetry that the inverse is also differentiable, since the inverse is obtained by interchanging the indices. It may be assumed, after composing with (necessarily differentiable) translations if need be, that $\boldsymbol{p} = \boldsymbol{0} = \boldsymbol{h}_i(\boldsymbol{0})$. The implicit function theorem guarantees the existence of diffeomorphisms $\boldsymbol{F}_i$ in a neighborhood of $\boldsymbol{0}$ such that $\boldsymbol{F}_i \circ \boldsymbol{h}_i = \imath$, with $\imath : \mathbb{R}^n \to \mathbb{R}^{n+k}$ mapping $\boldsymbol{a}$ to $(\boldsymbol{a}, \boldsymbol{0})$. Furthermore, if π is the projection $\mathbb{R}^{n+k} = \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$ onto the first factor, then

$$(\pi \circ \boldsymbol{F}_i) \circ \boldsymbol{h}_i = \pi \circ \imath \tag{3.2.1}$$

equals the identity 1_{V_i} on V_i, so that $\pi \circ \boldsymbol{F}_i$ is a left inverse for $\boldsymbol{h}_i$. Thus,

$$\boldsymbol{h}_2^{-1} \circ \boldsymbol{h}_1 = \pi \circ \boldsymbol{F}_2 \circ \boldsymbol{h}_1, \tag{3.2.2}$$

which is a composition of differentiable maps. □

(iii) The above theorem implies that if $(V, \boldsymbol{h})$ is a local parametrization of M, then $\boldsymbol{h}^{-1}$ is differentiable as a map from $\boldsymbol{h}(V)$ (which is a manifold, since it is open in M). In other words, using the terminology of the previous section, charts are differentiable maps.

If, however, we view a local parametrization of M as a map into Euclidean space, then it admits no differentiable inverse (since the inverse is not defined on an open set of the ambient Euclidean space). Nevertheless, the inverse may be extended to a differentiable map on an open set:

Proposition 3.2.1. *If $(V, \boldsymbol{h})$ is a local parametrization of $M^n \subset \mathbb{R}^{n+k}$, then there exists an open set $U \subset \mathbb{R}^{n+k}$ and a differentiable map $\boldsymbol{G} : U \to \mathbb{R}^n$ such that $\boldsymbol{G} \circ \boldsymbol{h} = 1_V$.*

Proof. As usual, we may assume that V contains $\mathbf{0} \in \mathbb{R}^n$ and that $\boldsymbol{h}(\mathbf{0}) = \mathbf{0} \in \mathbb{R}^{n+k}$. If $\boldsymbol{F}$ is the diffeomorphism (guaranteed by the implicit function theorem) that satisfies $\boldsymbol{F} \circ \boldsymbol{h} = \imath$, set $\boldsymbol{G} = \pi \circ \boldsymbol{F}$ (with $\imath$ and π denoting the same maps used in the proof of the above theorem). The statement then follows as in (3.2.1). □

One useful property of manifolds is the existence of "bump functions" in a neighborhood of any point:

Proposition 3.2.2. *If $(U, \boldsymbol{x})$ is a chart of M^n around $\boldsymbol{p}$, then there exists a differentiable function $\psi : M \to \mathbb{R}$ such that*
(1) $0 \leq \psi \leq 1$;
(2) $\psi \equiv 1$ *on some neighborhood of* $\boldsymbol{p}$, *and*
(3) *the support of ψ is contained in U.*

Proof. By Theorem 2.2.5, there exists a smooth $\varphi : \mathbb{R}^n \to \mathbb{R}$ with values in $[0, 1]$, support in $\boldsymbol{x}(U)$, that equals 1 in some neighborhood of $\boldsymbol{x}(\boldsymbol{p})$. ψ may now be taken to equal $\varphi \circ \boldsymbol{x}$ on U and zero outside U. This function is smooth because it is so in a neighborhood of any point: if the point is in U, then the neighborhood may be taken to be U; otherwise the point lies in the complement of the support of $\varphi \circ \boldsymbol{x}$, and this is an open set on which the function vanishes. □

Notice that (3.2.2) also yields a formula for the derivative of $\boldsymbol{h}_2^{-1} \circ \boldsymbol{h}_1$: even though the inverse of $\boldsymbol{h}_2$ is not differentiable, a formula similar to the chain rule holds:

$$D(\boldsymbol{h}_2^{-1} \circ \boldsymbol{h}_1) = D(\pi \circ \boldsymbol{F}_2) \circ D\boldsymbol{h}_1 = \pi \circ D\boldsymbol{F}_2 \circ D\boldsymbol{h}_1.$$

The above considerations are in a sense a special case of the following theorem, which roughly speaking, asserts that an n-dimensional manifold in $\mathbb{R}^{n+k}$ looks locally, up to a diffeomorphism of the ambient space, like $\mathbb{R}^n \times \{\mathbf{0}\} \subset \mathbb{R}^n \times \mathbb{R}^k$.

Theorem 3.2.2. *M is an n-dimensional submanifold of $\mathbb{R}^{n+k}$ if and only if for every $\boldsymbol{p} \in M$, there exists an open neighborhood U of $\boldsymbol{p}$ in R^{n+k} and a diffeomorphism $\boldsymbol{F} : U \to \boldsymbol{F}(U)$ such that*

$$\boldsymbol{F}(U \cap M) = \boldsymbol{F}(U) \cap (\mathbb{R}^n \times \{\boldsymbol{0}\}).$$

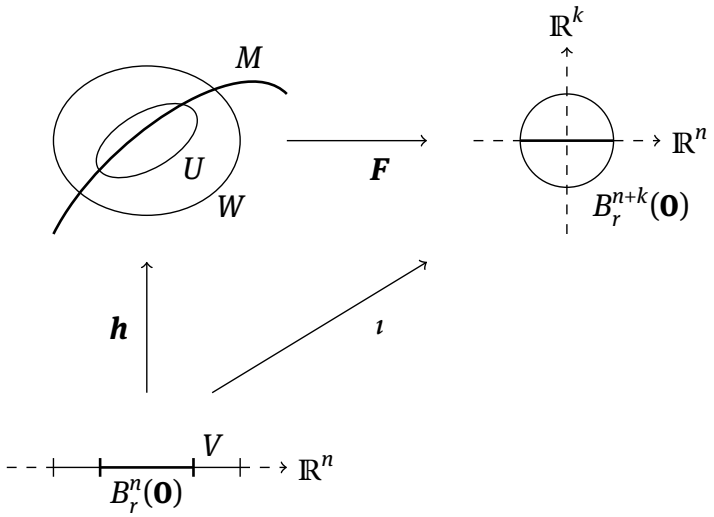
Proof. Suppose M satisfies the above condition. We proceed to construct a parametrization $(V, \boldsymbol{h})$ around any $\boldsymbol{p} \in M$. Let $\pi : \mathbb{R}^{n+k} = \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$ denote the projection, $\imath : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^k$ the map sending $\boldsymbol{a}$ to $(\boldsymbol{a}, \boldsymbol{0})$. If $\boldsymbol{F}$ is a diffeomorphism as in the statement, define

$$V = \pi(\boldsymbol{F}(U) \cap (\mathbb{R}^n \times \{\boldsymbol{0}\})), \qquad \boldsymbol{h} = \boldsymbol{F}^{-1} \circ \imath.$$

$\boldsymbol{h}$ is one-to-one with maximal rank everywhere because $\boldsymbol{F}$ and $\imath$ have those properties. Its inverse is continuous, for if W is an open subset of V, then $\boldsymbol{h}(W) = M \cap \boldsymbol{F}^{-1}(W \times \mathbb{R}^k)$ is open in M. Furthermore,

$$\begin{aligned}\boldsymbol{h}(V) &= \boldsymbol{F}^{-1}(\imath(V)) = \boldsymbol{F}^{-1}((\imath \circ \pi)(\boldsymbol{F}(U) \cap \mathbb{R}^n \times \{\boldsymbol{0}\})) \\ &= \boldsymbol{F}^{-1}(\boldsymbol{F}(U) \cap \mathbb{R}^n \times \{\boldsymbol{0}\}) = \boldsymbol{F}^{-1}(\boldsymbol{F}(U \cap M)) \\ &= U \cap M,\end{aligned}$$

and M is therefore an n-manifold.
Conversely, suppose M is a manifold, $\boldsymbol{p} \in M$, $(V, \boldsymbol{h})$ a parametrization around $\boldsymbol{p}$, so that $\boldsymbol{h}(V)$ equals the intersection of M with some open neighborhood W of $\boldsymbol{h}(\boldsymbol{p})$ in $\mathbb{R}^{n+k}$. As usual, we may assume that $\boldsymbol{p} = \boldsymbol{0} \in \mathbb{R}^n$, $\boldsymbol{h}(\boldsymbol{p}) = \boldsymbol{0} \in \mathbb{R}^{n+k}$. By the implicit function theorem, there exists a diffeomorphism $\boldsymbol{F}$ of a neighborhood of $\boldsymbol{0}$ in $\mathbb{R}^{n+k}$ such that $\boldsymbol{F} \circ \boldsymbol{h} = \imath$.



For $r > 0$, $B_r^n(\boldsymbol{0})$ and $B_r^{n+k}(\boldsymbol{0})$ will denote the open balls of radius r around $\boldsymbol{0}$ in $\mathbb{R}^n$ and $\mathbb{R}^{n+k}$ respectively. Choose r small enough that $B_r^n(\boldsymbol{0}) \subset V$, that $B_r^{n+k}(\boldsymbol{0})$ is contained in the image of $\boldsymbol{F}$, and that $U := \boldsymbol{F}^{-1}(B_r^{n+k}(\boldsymbol{0}))$ is contained in W. Then $B_r^n(\boldsymbol{0}) = \boldsymbol{h}^{-1}(U)$, so that $\boldsymbol{h}(B_r^n(\boldsymbol{0})) = U \cap M$. We will denote the restriction $\boldsymbol{F}_{|U}$ of $\boldsymbol{F}$ to U by $\boldsymbol{F}$ for simplicity. Then $\boldsymbol{F} : U \to \boldsymbol{F}(U)$ is by definition a diffeomorphism, and

$$\begin{aligned}\boldsymbol{F}(U \cap M) &= \boldsymbol{F}(\boldsymbol{h}(B_r^n(\boldsymbol{0}))) = \imath(B_r^n(\boldsymbol{0})) = B_r^{n+k}(\boldsymbol{0}) \cap (\mathbb{R}^n \times \{\boldsymbol{0}\}) \\ &= \boldsymbol{F}(U) \cap (\mathbb{R}^n \times \{\boldsymbol{0}\}).\end{aligned}$$

□

Recall that Theorem 3.1.1 provides a convenient way of constructing manifolds: if $\boldsymbol{f}: U \subset \mathbb{R}^{n+k} \to \mathbb{R}^k$ has $\mathbf{0} \in \mathbb{R}^k$ as a regular value, then the pre-image of $\mathbf{0}$, if nonempty, is an n-manifold. Theorem 3.2.2 shows that every manifold is at least *locally*, if not globally, obtainable in this way:

Corollary 3.2.1. *$M \subset \mathbb{R}^{n+k}$ is an n-dimensional manifold if and only if for every $\boldsymbol{p} \in M$, there exists a neighborhood U of $\boldsymbol{p}$ in $\mathbb{R}^{n+k}$, and a map $\boldsymbol{f}: U \to \mathbb{R}^k$ having $\boldsymbol{0}$ as a regular value, such that*

$$U \cap M = \boldsymbol{f}^{-1}(\boldsymbol{0}).$$

Proof. Suppose M is a manifold. Given $\boldsymbol{p} \in M$, there exists, by Theorem 3.2.2, a neighborhood U of $\boldsymbol{p}$ in the ambient Euclidean space, and a diffeomorphism $\boldsymbol{F}: U \to \boldsymbol{F}(U)$ such that $U \cap M = U \cap \boldsymbol{F}^{-1}(\mathbb{R}^n \times \{0\})$. It follows that if $\pi_2: \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^k$ is projection, then $\boldsymbol{f} := \pi_2 \circ \boldsymbol{F}$ has $\mathbf{0}$ as regular value (being a composition of maps of maximal rank), and $U \cap M = \boldsymbol{f}^{-1}(\mathbf{0})$. The converse is left as an exercise. □

A smooth map $\boldsymbol{f}: M \to N$ that has a differentiable inverse is called a *diffeomorphism*.

Definition 3.2.2. Let $M^n, N^l \subset \mathbb{R}^k$ be manifolds, $\boldsymbol{f}: M \to N$ a smooth map. The *derivative* of $\boldsymbol{f}$ at $\boldsymbol{p} \in M$ is the linear transformation $\boldsymbol{f}_{*\boldsymbol{p}}: M_{\boldsymbol{p}} \to N_{\boldsymbol{f}(\boldsymbol{p})}$ defined as follows: if $\boldsymbol{u} = \dot{\boldsymbol{c}}(0) \in M_{\boldsymbol{p}}$ for some curve $\boldsymbol{c}$ in M, then $\boldsymbol{f}_{*\boldsymbol{p}}\boldsymbol{u} = \dot{\boldsymbol{\gamma}}(0)$, where $\boldsymbol{\gamma} = \boldsymbol{f} \circ \boldsymbol{c}$.

The same argument used when defining the tangent space of M shows that this does not depend on the particular choice of curve. It is furthermore straightforward to check that if $\boldsymbol{g}: N \to P$ is also differentiable, then

$$(\boldsymbol{g} \circ \boldsymbol{f})_{*\boldsymbol{p}} = \boldsymbol{g}_{*\boldsymbol{f}(\boldsymbol{p})} \circ \boldsymbol{f}_{*\boldsymbol{p}}.$$

In particular, if $\boldsymbol{h}$ and $\boldsymbol{k}$ are parametrizations of M and N mapping $\mathbf{0}$ to $\boldsymbol{p}$ and $\boldsymbol{f}(\boldsymbol{p})$ respectively, then the following diagram commutes:

$$\begin{array}{ccc} M_{\boldsymbol{p}} & \xrightarrow{\boldsymbol{f}_{*\boldsymbol{p}}} & N_{\boldsymbol{f}(\boldsymbol{p})} \\ {\scriptstyle \boldsymbol{h}_{*\mathbf{0}}}\uparrow & & \uparrow{\scriptstyle \boldsymbol{k}_{*\mathbf{0}}} \\ \mathbb{R}^n_{\mathbf{0}} & \xrightarrow[(\boldsymbol{k}^{-1}\circ \boldsymbol{f}\circ \boldsymbol{h})_{*\mathbf{0}}]{} & \mathbb{R}^l_{\mathbf{0}} \end{array}$$

Notice that using the alternative notation $\{\mathbf{0}\} \times \mathbb{R}^k$ for the tangent space of $\mathbb{R}^k$ at $\mathbf{0}$,

$$\boldsymbol{f}_{*\boldsymbol{p}}(\boldsymbol{p}, D\boldsymbol{h}(\mathbf{0})\boldsymbol{u}) = (\boldsymbol{f}(\boldsymbol{p}), D(\boldsymbol{f} \circ \boldsymbol{h})(\mathbf{0})\boldsymbol{u}), \qquad \boldsymbol{u} \in \mathbb{R}^n. \tag{3.2.3}$$

In the special case of a function $f: M \to \mathbb{R}$, there is a further concept which plays an important role:

Definition 3.2.3. The *differential* of $f: M \to \mathbb{R}$ at $\boldsymbol{p} \in M$ is the linear map $\mathrm{d}f(\boldsymbol{p}): M_{\boldsymbol{p}} \to \mathbb{R}$ defined by $\mathrm{d}f(\boldsymbol{p})\boldsymbol{u} = (f \circ \boldsymbol{c})'(0)$, where $\boldsymbol{c}$ is any curve in M with $\dot{\boldsymbol{c}}(0) = \boldsymbol{u}$.

Thus $\mathrm{d}f(\boldsymbol{p})$ is an element of the dual space of $M_{\boldsymbol{p}}$. When M is Euclidean space,

$$\mathrm{d}f(\boldsymbol{p})(\boldsymbol{u}) = \boldsymbol{u}(f) = \langle \nabla f(\boldsymbol{p}), \boldsymbol{u} \rangle,$$

where $\boldsymbol{u}(f)$ is the derivative of f with respect to $\boldsymbol{u}$ as introduced in Definition 2.8.7. In particular,

$$\mathrm{d}f(\boldsymbol{p}) = \sum_i D_i f(\boldsymbol{p})\, \mathrm{d}u^i(\boldsymbol{p}),$$

which in classical notation reads

$$\mathrm{d}f = \sum_i \frac{\partial f}{\partial x^i}\, \mathrm{d}x^i,$$

for $n > 1$, and

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

when $n = 1$. More generally, in an arbitrary manifold,

$$\mathrm{d}f(\boldsymbol{p})\boldsymbol{u} = \mathcal{I}^{-1}_{f(\boldsymbol{p})}(f_{*\boldsymbol{p}}\boldsymbol{u}),$$

which incidentally also shows that the differential of f is independent of the curve chosen in Definition 3.2.3.

It is useful to keep the notation for derivative of a function f in direction $\boldsymbol{u}$ in the context of manifolds. Summarizing, we have:

$$\mathrm{d}f(\boldsymbol{p})\boldsymbol{u} = \boldsymbol{u}(f), \qquad f : M \to \mathbb{R}, \quad \boldsymbol{p} \in M, \quad \boldsymbol{u} \in M_{\boldsymbol{p}}. \tag{3.2.4}$$

We end this section with an application of Corollary 3.2.1 called the *method of Lagrange multipliers*. Consider the following rather trivial problem: find the highest and lowest points on the sphere $S^2 \subset \mathbb{R}^3$ of radius 1 centered at the origin. These are of course the north and south poles respectively. Now, the height function, u^3, has gradient $\boldsymbol{D}_3$, and the poles are precisely those points where this gradient is orthogonal to the tangent plane. Alternatively, $S^2 = f^{-1}(0)$, where $f = (u^1)^2 + (u^2)^2 + (u^3)^2 - 1$, and since ∇f spans the orthogonal complement of the tangent space of S^2 at every point, ∇u^3 is a multiple of ∇f at those points where f has a maximum or minimum: indeed if $\boldsymbol{c}$ is any curve in S^2 that passes through, say, the highest point $\boldsymbol{p}$ at time 0, then

$$\langle \nabla(u^3)(\boldsymbol{p}), \dot{\boldsymbol{c}}(0) \rangle = (u^3 \circ \boldsymbol{c})'(0) = 0;$$

this says that the gradient of u^3 is orthogonal to the tangent space at $\boldsymbol{p}$. More generally, we have the following:

Proposition 3.2.3. *Let $\boldsymbol{f} : \mathbb{R}^{n+k} \supset U \to \mathbb{R}^k$ be a map that has $\boldsymbol{0}$ as a regular value, so that $M = \boldsymbol{f}^{-1}(\boldsymbol{0})$ is an n-dimensional submanifold of $\mathbb{R}^{n+k}$. If a function $g : U \to \mathbb{R}$, when restricted to M, has a maximum or minimum at $\boldsymbol{p} \in M$, then there exist $\lambda_i \in \mathbb{R}$, $i = 1, \ldots, k$, such that*

$$\nabla g(\boldsymbol{p}) = \sum_{i=1}^{k} \lambda_i \nabla f^i(\boldsymbol{p}),$$

with $f^i := u^i \circ \boldsymbol{f}$.

The numbers λ_i are called *Lagrange multipliers*. When $k = 1$, then $\boldsymbol{f} = f$ is real-valued, and the identity becomes $\nabla g = \lambda \nabla f$ as in the example above.

Proof. Let $\boldsymbol{u} \in M_{\boldsymbol{p}}$, $\boldsymbol{c} : I \to M$ a curve with $\dot{\boldsymbol{c}}(0) = \boldsymbol{u}$. Then $g \circ \boldsymbol{c}$ has a maximum or minimum at 0, and

$$0 = (g \circ \boldsymbol{c})'(0) = \langle \nabla g(\boldsymbol{p}), \boldsymbol{u} \rangle.$$

Since $\boldsymbol{u}$ is arbitrary, $\nabla g(\boldsymbol{p})$ is orthogonal to $M_{\boldsymbol{p}}$. Now, $\nabla f^i(\boldsymbol{p})$ is also orthogonal to $M_{\boldsymbol{p}}$ for each i (recall that $f^i \circ \boldsymbol{c} \equiv 0$) and they are linearly independent because $\boldsymbol{f}$ has maximal rank on M. Thus, $\nabla f^i(\boldsymbol{p})$ span the orthogonal complement of $M_{\boldsymbol{p}}$, and the statement follows. □

Example 3.2.1. We will use Lagrange multipliers to prove that the geometric mean of n positive numbers is no larger than its arithmetic mean; i.e.,

$$\sqrt[n]{x_1 \cdots x_n} \le \frac{x_1 + \cdots + x_n}{n}, \qquad x_1, \ldots, x_n > 0.$$

Consider the first quadrant $U = \{\boldsymbol{p} \mid u^i(\boldsymbol{p}) > 0, \quad i = 1, \ldots, n\}$ in $\mathbb{R}^n$, and for any fixed $c > 0$, the function $f = \sum_i u^i - c : \mathbb{R}^n \to \mathbb{R}$. Let $g = (u^1 \cdots u^n)^{1/n} : U \to \mathbb{R}$. Notice that g has no minimum on U, since $g > 0$ on U, and $g(\boldsymbol{p}) \to 0$ if $u^i(\boldsymbol{p}) \to 0$. On the other hand, g has a maximum on $U \cap f^{-1}(0)$: it certainly has one on $\bar{U} \cap f^{-1}(0)$ which is a closed and bounded set, hence compact. Since g is zero on the boundary of U, this maximum lies in the interior of U. Thus, if the equation $\nabla g = \lambda \nabla f$ is satisfied at only one point in U, that point must be the maximum. We compute

$$\nabla g = \frac{1}{n}(u^1 \cdots u^n)^{\frac{1}{n}-1} \sum_{i=1}^{n} (u^1 \cdots \hat{u^i} \cdots u^n) \boldsymbol{D}_i,$$

where the accent "ˆ" indicates the corresponding term is deleted. Similarly, $\nabla f = \sum_i \boldsymbol{D}_i$, so $\nabla g = \lambda \nabla f$ iff

$$\frac{1}{n}(u^1 \cdots u^n)^{\frac{1}{n}-1} u^1 \cdots \hat{u^i} \cdots u^n = \lambda, \qquad i = 1, \ldots, n.$$

This can only hold if $u^i = u^j$ for all i and j. In this case, their common value is c/n, and the corresponding value of g is $c/n = (u^1 + \cdots + u^n)/n$. Since it occurs at a single point, it must be the maximum, and the inequality is proved.

3.3 Vector fields on manifolds

Definition 3.3.1. A *vector field* on a manifold M^n is a map $\boldsymbol{X}$ which assigns to each $\boldsymbol{p} \in M$ an element $\boldsymbol{X}(\boldsymbol{p}) \in M_{\boldsymbol{p}}$ of the tangent space of M at $\boldsymbol{p}$, and the map is differentiable in the following sense: for any $\boldsymbol{p} \in M$ and any parametrization $(V, \boldsymbol{h})$ around $\boldsymbol{p}$, there exists a differentiable (in the usual sense) vector field $\tilde{\boldsymbol{X}}$ on V such that

$$\boldsymbol{X} \circ \boldsymbol{h} = \boldsymbol{h}_* \circ \tilde{\boldsymbol{X}}; \tag{3.3.1}$$

i.e., $\boldsymbol{X}(\boldsymbol{h}(\boldsymbol{q})) = \boldsymbol{h}_{*\boldsymbol{q}}\tilde{\boldsymbol{X}}(\boldsymbol{q})$ for every $\boldsymbol{q} \in V$. Just as in the case of Euclidean space, vector fields satisfying (3.3.1) for some map $\boldsymbol{h}$ are said to be $\boldsymbol{h}$-*related*. If, as is the case for a parametrization, $\boldsymbol{h}$ is invertible, we also write $\boldsymbol{X} = \boldsymbol{h}_* \circ \tilde{\boldsymbol{X}} \circ \boldsymbol{h}^{-1}$.

Examples and Remarks 3.3.1. (i) (3.3.1) can be rephrased as saying that $\boldsymbol{X}$ is differentiable if for any parametrization $(V, \boldsymbol{h})$, the map $(\boldsymbol{h})_*^{-1} \circ \boldsymbol{X} \circ \boldsymbol{h}$ is a differentiable vector field on $V \subset \mathbb{R}^n$.

(ii) In Definition 3.3.1, it is enough to require that (3.3.1) hold for *some* parametrization $(V, \boldsymbol{h})$ around every $\boldsymbol{p} \in M$. For if $(U, \boldsymbol{k})$ is another parametrization around $\boldsymbol{p}$, let $\bar{\boldsymbol{X}} = \boldsymbol{k}_*^{-1} \circ \boldsymbol{X} \circ \boldsymbol{k}$. Then $\boldsymbol{X} \circ \boldsymbol{k} = \boldsymbol{k}_* \circ \bar{\boldsymbol{X}}$, and $\bar{\boldsymbol{X}}$ is differentiable, since locally

$$\bar{\boldsymbol{X}} = (\boldsymbol{k}^{-1} \circ \boldsymbol{h})_* \circ \tilde{\boldsymbol{X}} \circ (\boldsymbol{h}^{-1} \circ \boldsymbol{k})$$

is a composition of differentiable maps.

(iii) The *coordinate vector fields* associated to a chart $(U, \boldsymbol{x})$ of M^n are the vector fields

$$\frac{\partial}{\partial x^i} := \boldsymbol{x}_*^{-1} \circ \boldsymbol{D}_i \circ \boldsymbol{x}, \qquad i = 1, \ldots, n \tag{3.3.2}$$

on U. By (i) and (ii) they are differentiable, being $\boldsymbol{h}$-related to $\boldsymbol{D}_i$, for the parametrization $(\boldsymbol{x}(U), \boldsymbol{h} = \boldsymbol{x}^{-1})$.

(iv) As an illustration of (iii), recall spherical coordinates from Examples 2.5.1, which assign to each $(x, y, z) \in \mathbb{R}^3$ with $x \neq 0$ the point (ρ, θ, φ), with

$$\begin{aligned}
\rho(x, y, z) &= (x^2 + y^2 + z^2)^{1/2}, \\
\theta(x, y, z) &= \arctan \frac{y}{x} + c, \\
\varphi(x, y, z) &= \arccos \frac{z}{(x^2 + y^2 + z^2)^{1/2}},
\end{aligned}$$

where c depends on the quadrant where $(x, y) \in \mathbb{R}^2$ lies, cf. Section 4.6.1. (As observed earlier, they are actually defined on $\mathbb{R}^3 \setminus \{\boldsymbol{0}\}$; we exclude the plane $x = 0$ only to have a unified formula). The map $\boldsymbol{G}$ with $G^1 = \rho$, $G^2 = \theta$, and $G^3 = \varphi$ is a local diffeomorphism of $\mathbb{R}^3$ with inverse

$$\boldsymbol{G}^{-1}(a_1, a_2, a_3) = (a_1 \cos a_2 \sin a_3, a_1 \sin a_2 \sin a_3, a_1 \cos a_3).$$

Since for any $a > 0$, the sphere $S^2(a)$ of radius a centered at the origin equals $(G^1)^{-1}(a)$, the proof of Corollary 3.2.1 shows that the map $\boldsymbol{x} := (\theta, \varphi)$ is a chart on any such sphere. In order to describe the coordinate vector fields $\partial/\partial\varphi$, $\partial/\partial\theta$, notice that the inverse of the chart is the parametrization $\boldsymbol{h}$, where

$$\boldsymbol{h} = (a \cos u^1 \sin u^2, a \sin u^1 \sin u^2, a \cos u^2);$$

i.e., $\boldsymbol{h}(s, t) = \boldsymbol{G}^{-1}(a, s, t)$. Thus, by (3.3.2),

$$\begin{aligned}
\frac{\partial}{\partial\theta} \circ \boldsymbol{h} = \boldsymbol{h}_* \boldsymbol{D}_1 &= -a \sin u^1 \sin u^2 (\boldsymbol{D}_1 \circ \boldsymbol{h}) + a \cos u^1 \sin u_2 (\boldsymbol{D}_2 \circ \boldsymbol{h}) \\
&= -h^2 (\boldsymbol{D}_1 \circ \boldsymbol{h}) + h^1 (\boldsymbol{D}_2 \circ \boldsymbol{h}).
\end{aligned}$$

Since $h^i \circ \boldsymbol{h}^{-1} = u^i \circ \boldsymbol{h} \circ \boldsymbol{h}^{-1} = u^i$, composing the above identity with $\boldsymbol{h}^{-1}$ yields

$$\frac{\partial}{\partial\theta} = -u^2 \boldsymbol{D}_1 + u^1 \boldsymbol{D}_2. \tag{3.3.3}$$

The computations for $\partial/\partial\varphi$ are similar, only slightly more involved:

$$\begin{aligned}\frac{\partial}{\partial\varphi} \circ \boldsymbol{h} = \boldsymbol{h}_* \boldsymbol{D}_2 = a\cos u^1 \cos u^2 (\boldsymbol{D}_1 \circ \boldsymbol{h}) + a \sin u^1 \cos u^2 (\boldsymbol{D}_2 \circ \boldsymbol{h}) \\ - a \sin u^2 (\boldsymbol{D}_3 \circ \boldsymbol{h}).\end{aligned} \tag{3.3.4}$$

Since u^2 takes values in $(0, \pi)$, its sine is nonnegative, so that

$$\sin u^2 = |\sin u^2| = ((\cos u^1 \sin u^2)^2 + (\sin u^1 \sin u^2)^2)^{1/2},$$

and the first function on the right side of (3.3.4) may be written

$$\begin{aligned}a\cos u^1 \cos u^2 &= \frac{(a\cos u^1 \sin u^2)(a \cos u^2)}{((a\cos u^1 \sin u^2)^2 + (a \sin u^1 \sin u^2)^2)^{\frac{1}{2}}} \\ &= \frac{h^1 h^3}{((h^1)^2 + (h^2)^2)^{\frac{1}{2}}}.\end{aligned}$$

Similarly,

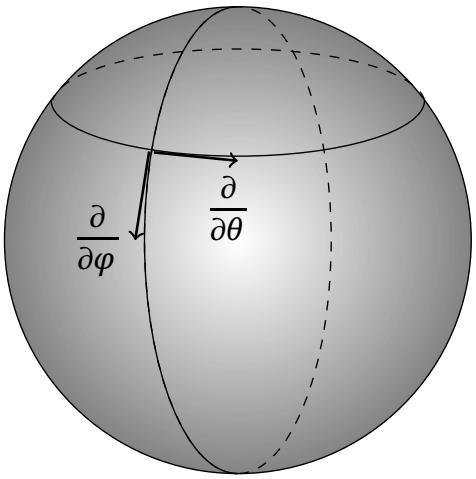
$$a \sin u^1 \cos u^2 = \frac{h^2 h^3}{((h^1)^2 + (h^2)^2)^{\frac{1}{2}}}, \quad -a \sin u^2 = -((h^1)^2 + (h^2)^2)^{\frac{1}{2}},$$

and therefore

$$\begin{aligned}\frac{\partial}{\partial\varphi} \circ \boldsymbol{h} = \frac{1}{((h^1)^2 + (h^2)^2)^{\frac{1}{2}}} (h^1 h^3 (\boldsymbol{D}_1 \circ \boldsymbol{h}) + h^2 h^3 (\boldsymbol{D}_2 \circ \boldsymbol{h}) \\ - ((h^1)^2 + (h^2)^2)(\boldsymbol{D}_3 \circ \boldsymbol{h})),\end{aligned}$$

or equivalently,

$$\frac{\partial}{\partial\varphi} = \frac{1}{((u^1)^2 + (u^2)^2)^{\frac{1}{2}}} (u^1 u^3 \boldsymbol{D}_1 + u^2 u^3 \boldsymbol{D}_2 - ((u^1)^2 + (u^2)^2)\boldsymbol{D}_3). \tag{3.3.5}$$



Coordinate vector fields on a sphere

Notice that $\partial/\partial\theta$ and $\partial/\partial\varphi$ are mutually orthogonal everywhere, and are orthonormal along the equator.

(v) A vector field $\boldsymbol{X}$ on M and a function $f : M \to \mathbb{R}$ can be combined to yield a new vector field $f\boldsymbol{X}$, by setting $(f\boldsymbol{X})(\boldsymbol{p}) := f(\boldsymbol{p})\boldsymbol{X}(\boldsymbol{p})$, $\boldsymbol{p} \in M$. They can also be combined to form a new function $\boldsymbol{X}f$, where $(\boldsymbol{X}f)(\boldsymbol{p}) := \boldsymbol{X}(\boldsymbol{p})(f)$.

The following proposition generalizes the fact that coordinate vector fields are differentiable.

Proposition 3.3.1. *A map $\boldsymbol{p} \mapsto \boldsymbol{X}(\boldsymbol{p}) \in M_{\boldsymbol{p}}$ is a vector field (in the sense that $\boldsymbol{X}$ is differentiable) if and only if for any $\boldsymbol{p} \in M$ there exists a chart $(U, \boldsymbol{x})$ of M around $\boldsymbol{p}$ such that*

$$\boldsymbol{X}_{|U} = \sum_i f^i \frac{\partial}{\partial x^i}$$

for some differentiable functions f^i on U.

Proof. The restriction to U of the map $\boldsymbol{X}$ is smooth iff in terms of the above chart, the map $\boldsymbol{x}_* \circ \boldsymbol{X} \circ \boldsymbol{x}^{-1}$ is differentiable on $V = \boldsymbol{x}(U)$; i.e., iff it can be written as $\sum_i f^i \boldsymbol{D}_i$ for smooth functions f^i on V. This is equivalent to requiring that

$$\boldsymbol{X}_{|U} \circ \boldsymbol{x}^{-1} = \boldsymbol{x}_*^{-1}\left(\sum_i f^i D_i\right) = \sum_i (f^i \circ \boldsymbol{x}^{-1})\boldsymbol{x}_*^{-1}\boldsymbol{D}_i;$$

i.e.,

$$\boldsymbol{X}_{|U} = \sum_i f^i \frac{\partial}{\partial x^i},$$

by (3.3.2). □

There is another, sometimes more useful way of determining differentiability:

Theorem 3.3.1. *Let M denote an n-dimensional submanifold of $\mathbb{R}^{n+k}$. A map $\boldsymbol{X}$ that assigns to each $\boldsymbol{p} \in M$ a vector $\boldsymbol{X}(\boldsymbol{p}) \in M_{\boldsymbol{p}}$ is a vector field (once again, in the sense that the map is differentiable) if and only if for any $\boldsymbol{p} \in M$ there exists a vector field $\boldsymbol{Y}$ on an open neighborhood U of $\boldsymbol{p}$ in $\mathbb{R}^{n+k}$ such that the restrictions of both vector fields agree on the induced neighborhood of $\boldsymbol{p}$ in M: i.e., $\boldsymbol{Y}_{|M\cap U} = \boldsymbol{X}_{|M\cap U}$.*

Proof. Suppose $\boldsymbol{X}$ is a vector field on M, and consider a parametrization $(V, \boldsymbol{h})$ around some $\boldsymbol{p} \in M$. As usual, we may assume to simplify matters that $\boldsymbol{p} = \boldsymbol{0} = \boldsymbol{h}(\boldsymbol{0})$. Thus, the map $\tilde{\boldsymbol{X}} = (\boldsymbol{h}^{-1})_* \circ \boldsymbol{X} \circ \boldsymbol{h}$ is differentiable on V. Write $\tilde{\boldsymbol{X}} = \sum_i f^i \boldsymbol{D}_i$, where f^i are differentiable functions on V, and extend $\tilde{\boldsymbol{X}}$ to a vector field $\tilde{\boldsymbol{Y}}$ on $V \times \mathbb{R}^k$ by setting

$$\tilde{\boldsymbol{Y}} = \sum_i (f^i \circ \pi)\boldsymbol{D}_i,$$

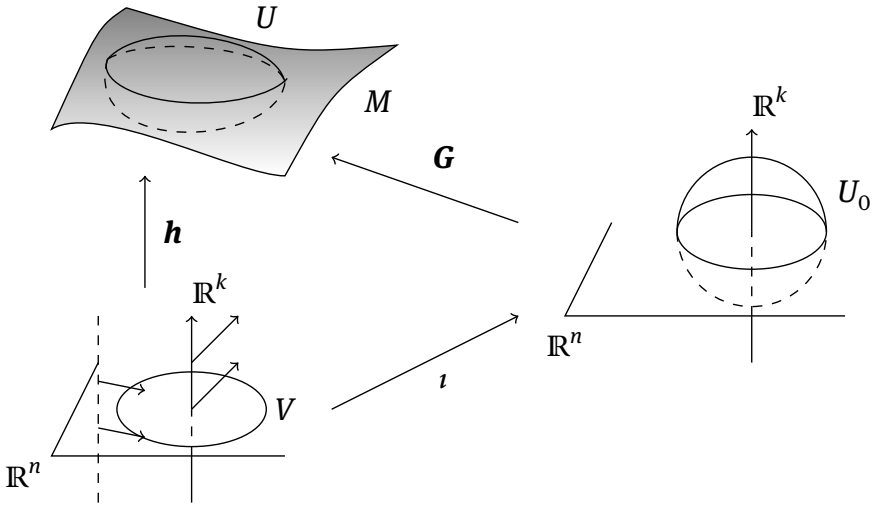
with $\pi : V \times \mathbb{R}^k \to V$ denoting projection. Observe that if $\imath : V \to V \times \mathbb{R}^k$ maps $\boldsymbol{a} \in V$ to $(\boldsymbol{a}, \boldsymbol{0})$, then

$$\tilde{\boldsymbol{Y}} \circ \imath = \imath_* \circ \tilde{\boldsymbol{X}}_{|V}.$$

By Theorem 3.2.2, there exists a diffeomorphism $\boldsymbol{G}$ on a neighborhood U_0 of $\mathbf{0}$ in $\mathbb{R}^{n+k}$ such that

$$\begin{aligned} \boldsymbol{h} &= \boldsymbol{G} \circ \imath, \\ U_0 \cap (\mathbb{R}^n \times \{\mathbf{0}\}) &= \imath(V), \text{ and} \\ \boldsymbol{G}^{-1}(U \cap M) &= \boldsymbol{G}^{-1}(U) \cap (\mathbb{R}^n \times \{\mathbf{0}\}), \end{aligned}$$

with U denoting $\boldsymbol{G}(U_0)$.



Define a vector field $\boldsymbol{Y}$ on U by setting $\boldsymbol{Y} = \boldsymbol{G}_* \circ \tilde{\boldsymbol{Y}} \circ \boldsymbol{G}^{-1}$. We claim $\boldsymbol{Y}$ satisfies the conclusion of the theorem. Indeed,

$$\boldsymbol{Y} \circ \boldsymbol{h} = \boldsymbol{G}_* \circ \tilde{\boldsymbol{Y}} \circ \boldsymbol{G}^{-1} \circ \boldsymbol{h} = \boldsymbol{G}_* \circ \tilde{\boldsymbol{Y}} \circ \imath = \boldsymbol{G}_* \circ \imath_* \circ \tilde{\boldsymbol{X}} = \boldsymbol{h}_* \circ \boldsymbol{X} = \boldsymbol{X} \circ \boldsymbol{h}.$$

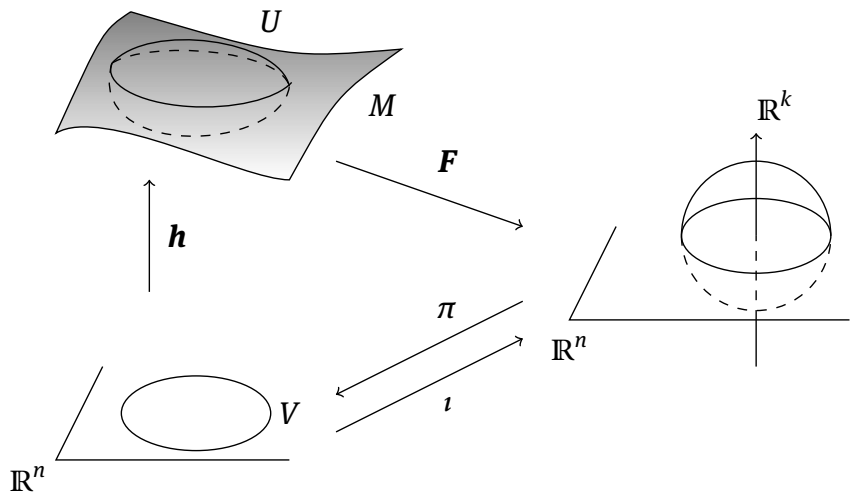
Thus, the restrictions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ to $\boldsymbol{h}(V)$ agree. Since $\boldsymbol{h}(V) = (\boldsymbol{G} \circ \imath)(V) = U \cap M$, the claim follows.

For the converse, let $\boldsymbol{p} \in M$, which as usual may be assumed to be $\mathbf{0}$. By assumption, there exists an open neighborhood U of $\boldsymbol{p}$ in $\mathbb{R}^{n+k}$, and a vector field $\boldsymbol{Y}$ on U such that the restrictions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ to $M \cap U$ agree. It must be shown that there exists a parametrization $(V, \boldsymbol{h})$ of M with $\boldsymbol{p} \in \boldsymbol{h}(V)$ such that $\boldsymbol{h}_*^{-1} \circ \boldsymbol{X} \circ \boldsymbol{h}$ is differentiable.

Now, by Theorem 3.2.2, after restricting U if necessary, there exists a diffeomorphism $\boldsymbol{F} : U \to \boldsymbol{F}(U)$ with the property that

$$\boldsymbol{F}(U \cap M) = \boldsymbol{F}(U) \cap (\mathbb{R}^n \times \{\mathbf{0}\}).$$

Furthermore, if $\pi : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$ is projection, if $\imath : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^k$ maps $\boldsymbol{a}$ to $(\boldsymbol{a}, \mathbf{0})$ and $V = (\pi \circ \boldsymbol{F})(U \cap M)$, then $(V, \boldsymbol{h})$, where $\boldsymbol{h} = \boldsymbol{F}^{-1} \circ \imath$, is a parametrization of M around $\boldsymbol{p}$.



Define $\bar{\boldsymbol{X}} = \boldsymbol{F}_* \circ \boldsymbol{Y} \circ \boldsymbol{F}^{-1}$. $\bar{\boldsymbol{X}}$ is differentiable and its restriction to $\mathbb{R}^n \times \{\boldsymbol{0}\}$ maps by construction each $\boldsymbol{p}$ to a vector tangent to $\mathbb{R}^n$. Thus, there exists a vector field $\tilde{\boldsymbol{X}}$ on V such that $\boldsymbol{X} \circ \imath = \imath_* \circ \tilde{\boldsymbol{X}}$: indeed, if $\bar{\boldsymbol{X}} = \sum_i g^i \boldsymbol{D}_i$, then each g_i is smooth, and therefore so is the map $\tilde{\boldsymbol{X}} = \sum (g^i \circ \imath) \boldsymbol{D}_i$. This in turn shows that $\boldsymbol{X}$ is differentiable, since

$$\boldsymbol{h}_* \circ \tilde{\boldsymbol{X}} = (\boldsymbol{F}^{-1} \circ \imath)_* \circ \tilde{\boldsymbol{X}} = \boldsymbol{F}_*^{-1} \circ \bar{\boldsymbol{X}} \circ \imath = \boldsymbol{Y} \circ \boldsymbol{F}^{-1} \circ \imath = \boldsymbol{Y} \circ \boldsymbol{h} = \boldsymbol{X} \circ \boldsymbol{h}. \qquad \square$$

Either the definition of differentiable vector field or the characterization given in Theorem 3.3.1 can now be used to extend the concept of flow to vector fields on manifolds. As in Euclidean space, an *integral curve* of a vector field $\boldsymbol{X}$ on a manifold M is a curve $\boldsymbol{c} : I \to M$ such that $\dot{\boldsymbol{c}} = \boldsymbol{X} \circ \boldsymbol{c}$. Theorem 3.3.1 immediately implies the existence and uniqueness of integral curves, and thus of local flows. If one prefers to use the definition, it suffices to observe that locally, a vector field on a manifold is $\boldsymbol{h}$-related to one in Euclidean space for some parametrization $\boldsymbol{h}$, and appeal to the following:

Proposition 3.3.2. *Let $\boldsymbol{X}, \tilde{\boldsymbol{X}}$ be $\boldsymbol{h}$-related vector fields, $\boldsymbol{h}_* \circ \tilde{\boldsymbol{X}} = \boldsymbol{X} \circ \boldsymbol{h}$.*
(1) *If $\boldsymbol{c}$ is an integral curve of $\tilde{\boldsymbol{X}}$, then $\boldsymbol{h} \circ \boldsymbol{c}$ is an integral curve of $\boldsymbol{X}$.*
(2) *If $\tilde{\boldsymbol{\Psi}}_t, \boldsymbol{\Psi}_s$ are flows of $\tilde{\boldsymbol{X}}, \boldsymbol{X}$ respectively, then $\boldsymbol{h} \circ \tilde{\boldsymbol{\Psi}}_t = \boldsymbol{\Psi}_t \circ \boldsymbol{h}$.*

Proof. If $\boldsymbol{c}$ is an integral curve of $\tilde{\boldsymbol{X}}$, then

$$\dot{\boldsymbol{h} \circ \boldsymbol{c}} = \boldsymbol{h}_* \dot{\boldsymbol{c}} = \boldsymbol{h}_* \circ \tilde{\boldsymbol{X}} \circ \boldsymbol{c} = (\boldsymbol{X} \circ \boldsymbol{h}) \circ \boldsymbol{c} = \boldsymbol{X} \circ (\boldsymbol{h} \circ \boldsymbol{c}),$$

which is the first claim. The second claim is then clear: for any $\boldsymbol{p} \in M$, the curve $t \mapsto \boldsymbol{h} \circ \tilde{\boldsymbol{\Psi}}_t(\boldsymbol{p})$ is by (1) an integral curve of $\boldsymbol{X}$ passing through $\boldsymbol{h}(\boldsymbol{p})$ at $t = 0$, and so is $t \mapsto \boldsymbol{\Psi}_t \circ \boldsymbol{h}(\boldsymbol{p})$. They are therefore one and the same. $\square$

The work needed to group local flows into a single global one was already done in the proof of Theorem 2.8.3. Even though it was stated in the setting of Euclidean space, the argument makes no use of that setting, and works equally well in the more general one of manifolds. We restate it for future reference. For $\boldsymbol{p} \in M$, let $I_{\boldsymbol{p}}$ denote the largest open interval on which the integral curve $\boldsymbol{\Psi}_{\boldsymbol{p}}$ passing through $\boldsymbol{p}$ at time 0 is defined.

Theorem 3.3.2. *If $\boldsymbol{X}$ is a vector field on a manifold M, there exists a unique open set $W \subset \mathbb{R} \times M$ and a unique differentiable map $\boldsymbol{\Psi} : W \to M$ such that*
(1) $I_{\boldsymbol{p}} \times \{\boldsymbol{p}\} = W \cap (\mathbb{R} \times \{\boldsymbol{p}\})$ *for all* $\boldsymbol{p} \in M$, *and*
(2) $\boldsymbol{\Psi}(t, \boldsymbol{p}) = \boldsymbol{\Psi}_{\boldsymbol{p}}(t)$ *for* $(t, \boldsymbol{p}) \in W$.

The map $\boldsymbol{\Psi}$ in the above theorem is called *the flow* of the vector field $\boldsymbol{X}$, as opposed to *a flow*, or *a local flow*, which refers to a restriction of $\boldsymbol{\Psi}$ to a subset of its domain. As in Euclidean space, $\boldsymbol{X}$ is said to be *complete* if its flow has domain $\mathbb{R}\times M$, or equivalently, if its integral curves are defined on all of $\mathbb{R}$. It turns out that any vector field on a compact manifold is complete. In order to show this, we will need the following extension theorem:

Theorem 3.3.3. *Let $\boldsymbol{c} : [a, b) \to M$ be an integral curve of a vector field $\boldsymbol{X}$ on M. If there exists a sequence $\{t_k\}$ in $[a, b)$ that converges to b for which $\{\boldsymbol{c}(t_k)\}$ converges to some $\boldsymbol{p} \in M$, then $\boldsymbol{c}$ may be extended to a continuous curve on $[a, b]$; i.e., the curve $\bar{\boldsymbol{c}} : [a, b] \to M$, with $\bar{\boldsymbol{c}}(t) = \boldsymbol{c}(t)$ if $t < b$ and $\bar{\boldsymbol{c}}(b) = \boldsymbol{p}$, is continuous. Furthermore, if $\boldsymbol{\gamma} : I \to M$ is the maximal integral curve of $\boldsymbol{X}$ with $\boldsymbol{\gamma}(b) = \boldsymbol{p}$, then I contains $[a, b]$ and the restriction of $\boldsymbol{\gamma}$ to $[a, b]$ equals $\bar{\boldsymbol{c}}$.*

Proof. By using a chart around $\boldsymbol{p}$ as in Theorem 3.3.1, we may translate the situation to Euclidean space. So assume $M = \mathbb{R}^n$ and $\boldsymbol{p} = \boldsymbol{0}$. Suppose the curve $\bar{\boldsymbol{c}}$ from the statement is not continous. Then there exists a sequence $r_k \to b$ such that $\{\boldsymbol{c}(r_k)\}$ does not converge to $\boldsymbol{0}$; i.e., there exists $\delta > 0$ such that for any $N \in \mathbb{N}$, $|\boldsymbol{c}(r_k)| > \delta$ for some $k > N$. Let $\varepsilon \in (0, \delta)$. By passing to subsequences if necessary, we may assume $|\boldsymbol{c}(t_k)| < \varepsilon$ and $\boldsymbol{c}(r_k) > \delta$ for all k. The contradiction arises, roughly speaking, from the fact that $|t_k - r_k| \to 0$ but $|\boldsymbol{c}(t_k) - \boldsymbol{c}(r_k)|$ does not, so that the speed $|\dot{\boldsymbol{c}}|$ of $\boldsymbol{c}$ and hence the norm of $\boldsymbol{X}$ is unbounded on a compact neighborhood of the origin. More precisely, we construct subsequences as follows: set $k_1 = 1$, and choose some $\tilde{k}_1 \in \mathbb{N}$ such that $r_{\tilde{k}_1} > t_{k_1}$ (such a $\tilde{k}_1$ exists since otherwise $\{r_k\}$ would not converge to b). For the same reason, there exists some $k_2 > k_1$ with $t_{k_2} > r_{\tilde{k}_1}$. Continue in this fashion to obtain subsequences $\{t_{k_i}\}$ and $\{r_{\tilde{k}_i}\}$ that converge to b and satisfy $t_{k_i} < r_{\tilde{k}_i} < t_{k_{i+1}}$ for all i. Now, $|\boldsymbol{c}(t_{k_i})| < \varepsilon < \delta$, but $|\boldsymbol{c}(r_{\tilde{k}_i})| > \delta$, so that by continuity $|\boldsymbol{c}(t)|$ must equal δ for some $t \in (t_{k_i}, r_{\tilde{k}_i})$. If s_i denotes the supremum of all such t, then $|\boldsymbol{c}(s_i)| = \delta$ by continuity again.
The length of the restriction of $\boldsymbol{c}$ to $[t_{k_i}, s_i]$ satisfies

$$\int_{t_{k_i}}^{s_i} |\dot{\boldsymbol{c}}| \geq |\boldsymbol{c}(s_i) - \boldsymbol{c}(t_{k_i})| \geq |\boldsymbol{c}(s_i)| - |\boldsymbol{c}(t_{k_i})| \geq \delta - \varepsilon$$

by the triangle inequality. Thus, for any natural number m,

$$\sum_{i=1}^{m} \int_{t_{k_i}}^{s_i} |\dot{\boldsymbol{c}}| \geq m(\delta - \varepsilon). \tag{3.3.6}$$

Since $\boldsymbol{c}[t_{k_i}, s_i]$ is contained in the closure $\overline{B_\delta(\mathbf{0})}$ of the ball of radius δ about the origin, this means that the portion of $\boldsymbol{c}$ that lies inside it has unbounded length. But this is impossible, because the norm of $\boldsymbol{X}$ is bounded on the compact $\overline{B_\delta(\mathbf{0})}$: specifically, if $|\boldsymbol{X}| \leq M$ on $\overline{B_\delta(\mathbf{0})}$, then $|\dot{\boldsymbol{c}}(t)| = |\boldsymbol{X}(\boldsymbol{c}(t))| \leq M$ for $t \in [t_{k_i}, s_i]$, and

$$\sum_{i=1}^{m} \int_{t_{k_i}}^{s_i} |\dot{\boldsymbol{c}}| \leq M \sum_i (s_i - t_{k_i}) \leq M(b-a),$$

which contradicts (3.3.6). This shows that $\boldsymbol{c}$ is extendable to a continuous curve $\bar{\boldsymbol{c}} : [a,b] \to M$ by setting $\bar{\boldsymbol{c}}(b) = \mathbf{0}$. Furthermore,

$$\lim_{t \to b^-} \boldsymbol{X}(\bar{\boldsymbol{c}}(t)) = \lim_{t \to b^-} \dot{\bar{\boldsymbol{c}}}(t) = \boldsymbol{X}(\mathbf{0}) \tag{3.3.7}$$

by continuity of $\boldsymbol{X}$ and $\bar{\boldsymbol{c}}$.
Consider the curve $\boldsymbol{\gamma}_1 : I \to M$ that equals $\bar{\boldsymbol{c}}$ on $[a,b]$ and $\boldsymbol{\gamma}$ on $I \setminus [a,b]$. This curve is continuous and differentiable (with $\dot{\boldsymbol{\gamma}}_1 = \boldsymbol{X} \circ \boldsymbol{\gamma}_1$) everywhere except perhaps at b. To show smoothness at b (and thereby conclude the proof of the theorem), let $\{t_k\}$ be any sequence in $[a,b)$ that converges to b. By the mean value theorem, there exists $s_k \in (t_k, b)$ such that

$$\frac{(u^i \circ \boldsymbol{\gamma}_1)(b) - (u^i \circ \boldsymbol{\gamma}_1)(t_k)}{b - t_k} = (u^i \circ \boldsymbol{\gamma}_1)'(s_k).$$

But $\dot{\boldsymbol{\gamma}}_1(s_k) = \dot{\bar{\boldsymbol{c}}}(s_k) \to \boldsymbol{X}(\mathbf{0})$ by (3.3.7). Similarly, if $b < s_k < t_k$ and $t_k \to b$, then $\dot{\boldsymbol{\gamma}}_1(s_k) = \dot{\boldsymbol{\gamma}}(s_k) \to \boldsymbol{X}(\mathbf{0})$. This shows that $(u^i \circ \boldsymbol{\gamma}_1)'(b)$ exists. The same argument then applies to derivatives of higher order to show that $\boldsymbol{\gamma}_1$ is smooth at b. □

Corollary 3.3.1. *Any vector field on a compact manifold is complete; i.e., its integral curves are defined on all of* $\mathbb{R}$.

Proof. Let $\boldsymbol{X}$ denote a vector field on a compact manifold M, and $\boldsymbol{c}$ the maximal integral curve of $\boldsymbol{X}$ with $\boldsymbol{c}(0)$ equal to some $\boldsymbol{p} \in M$. It is enough to show that $[0,\infty)$ is contained in the domain of $\boldsymbol{c}$; the same argument applied to the integral curve $t \mapsto \boldsymbol{c}(-t)$ of $-\boldsymbol{X}$ then shows that $(-\infty, 0]$ is also contained in the domain. With this in mind, let I denote the set of all $t \in [0,\infty)$ such that $\boldsymbol{c}$ is defined on $[0,t)$. Then I is by definition a nonempty open interval in $[0,\infty)$. Theorem 3.3.3 implies I is also closed. By Proposition 1.7.1, $I = [0,\infty)$. □

The Lie bracket of vector fields on a manifold M is defined in exactly the same way it was done for Euclidean space, and all relevant theorems proved in Section 2.9 carry over to the manifold setting. In fact, the Lie bracket of $\boldsymbol{X}$ and $\boldsymbol{Y}$ at a point $\boldsymbol{p} \in M$ depends only on the values of these fields in a neighborhood of $\boldsymbol{p}$, and if we take this neighborhood to be the domain of a chart $(U, \boldsymbol{x})$, then the vector fields $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ on $\boldsymbol{x}(U)$, where $\tilde{\boldsymbol{X}} = \boldsymbol{x}_* \circ \boldsymbol{X}_{|U} \circ \boldsymbol{x}^{-1}$ (and similarly for $\tilde{\boldsymbol{Y}}$) are $\boldsymbol{h}$-related to $\boldsymbol{X}$ and $\boldsymbol{Y}$, if $\boldsymbol{h} = \boldsymbol{x}^{-1}$.

By Theorem 2.9.2, the restriction of the bracket to U is given by

$$[\boldsymbol{X}, \boldsymbol{Y}]_{|U} = \boldsymbol{h}_* \circ [\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}] \circ \boldsymbol{x}.$$

For example, the Lie bracket of coordinate vector fields is always zero, since these fields are related to the standard coordinate vector fields $\boldsymbol{D}_i$ in Euclidean space.

3.4 Lie groups

The material in this section will not be used in the sequel, and the reader may skip it without loss of continuity. It has been included because it describes a large class of examples of manifolds, which, in addition, possess a rich additional structure. It also illustrates many of the techniques we developed for vector fields.

In modern algebra, a *group* is a pair $(G, \cdot)$, where G is a nonempty set, and $\cdot : G \times G \to G$ is a map, called the group product, satisfying the following properties:

(1) $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ for all $a, b, c \in G$;
(2) There exists an element $e \in G$ such that $a \cdot e = e \cdot a = a$ for all $a \in G$;
(3) For any $a \in G$ there exists an element $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e$.

e is called the identity element, a^{-1} the inverse of a. We will often write ab instead of $a \cdot b$. Examples of groups are plentiful: the set of nonzero reals with the usual multiplication is one. Any vector space is a group with vector addition. So is the set $GL(n)$ of all $n \times n$ invertible matrices with the usual matrix multiplication. In the latter case, the identity element is the identity matrix I_n.

Definition 3.4.1. Let G be a group that admits in addition a manifold structure. If the group product $G \times G \to G$ and the inverse map $G \to G$ (which sends a to a^{-1}) are differentiable, then G is called a *Lie group*.

The three examples given earlier are Lie groups: vector addition in a vector space E is differentiable as a bilinear map on $E \times E$, and in the other cases, the group product is differentiable because it is the restriction of a bilinear map to an open subset of a vector space. The inverse maps are also easily seen to be differentiable; in the case of $GL(n)$ for example, this follows from the alternative formula for the inverse of a matrix given in the proof of Theorem 1.3.6. Another important example is the following subgroup of $GL(n)$ (a subgroup of a group G is a subset that is a group in its own right with the restriction of the group product in G):

Proposition 3.4.1. *The orthogonal group* $O(n) = \{A \in M_n \mid AA^T = I_n\}$ *is a Lie group of dimension* $n(n-1)/2$.

Proof. Denote by S_n the subspace of M_n that consists of all symmetric matrices. S_n has as basis the set $\{A_{ij} \mid 1 \le i \le j \le n\}$, where A_{ij} has entries 1 in the (i, j) and (j, i) slots, and zero elsewhere. Its dimension equals the number of elements in the set of all pairs (i, j) where $1 \le i \le j \le n$. This is just $1 + 2 + \cdots + n = n(n+1)/2$.

Next, consider the map $\boldsymbol{f} : GL(n) \to S_n$ given by $\boldsymbol{f}(A) = AA^T$. We claim it has maximal rank at any $A \in \boldsymbol{f}^{-1}(I_n)$, so that by Theorem 3.1.1, $O(n) = \boldsymbol{f}^{-1}(I_n)$ is a manifold of dimension $n^2 - n(n+1)/2 = n(n-1)/2$. Now, $\boldsymbol{f}$ is the restriction to $GL(n)$ of the composition $m \circ \imath$, where $\imath : M_n \to M_n \times M_n$ is given by $\imath(A) = (A, A^T)$, and m denotes the multiplication map $m : M_n \times M_n \to M_n$, $m(A, B) = AB$. But $\imath$ is linear, and m is bilinear, so the results from Chapter 2 imply that for $A \in GL(n)$, $M \in M_n$,

$$\begin{aligned} D\boldsymbol{f}(A)M &= Dm(A, A^T) \circ D\imath(A)M = Dm(A, A^T) \circ \imath(M) \\ &= Dm(A, A^T)(M, M^T) = m(A, M^T) + m(M, A^T) \\ &= AM^T + MA^T. \end{aligned}$$

This means that for $A \in O(n)$, $D\boldsymbol{f}(A)$ is onto S_n: indeed, given a symmetric matrix S, if $M = (1/2)SA$, then

$$D\boldsymbol{f}(A)M = \frac{1}{2}A(SA)^T + \frac{1}{2}SAA^T = \frac{1}{2}(S^T + S) = S.$$

Thus, $O(n)$ is a manifold. It is a Lie group because the product and inverse map are the composition of the corresponding smooth operations on $GL(n)$ with the differentiable inclusion map of $O(n)$ into the open subset $GL(n)$ of M_n. □

Given an element a in a Lie group G, define *left translation* by a to be the map

$$\begin{aligned} L_a : G &\to G, \\ b &\mapsto ab. \end{aligned}$$

It is differentiable, being the composition of the group product with the map $\imath_a : G \to G \times G$, $\imath_a(b) = (a, b)$, and is in fact a diffeomorphism of G with inverse $L_{a^{-1}}$. Notice also that $L_a \circ L_b = L_{ab}$ for $a, b \in G$.

Definition 3.4.2. A vector field $\boldsymbol{X}$ on a Lie group G is said to be *left-invariant* if it is L_g-related to itself for any $g \in G$; i.e.,

$$L_{g*}\boldsymbol{X} = \boldsymbol{X} \circ L_g, \qquad g \in G.$$

Recall that the collection of all vector fields on G is a (infinite-dimensional) Lie algebra with the Lie bracket of vector fields, cf. Proposition 2.9.2. It follows from the definition that the bracket of two left-invariant vector fields is again left-invariant, so that the collection $\mathfrak{g}$ of all such vector fields is also a Lie algebra. This one, though, is finite-dimensional. In fact, we have the following:

Theorem 3.4.1. *The Lie algebra $\mathfrak{g}$ of a Lie group G is a vector space naturally isomorphic to G_e.*

Proof. Define $\boldsymbol{h} : \mathfrak{g} \to G_e$ by $\boldsymbol{h}(\boldsymbol{X}) = \boldsymbol{X}(e)$ for $\boldsymbol{X} \in \mathfrak{g}$. This map is clearly linear. If $\boldsymbol{h}(\boldsymbol{X}) = \boldsymbol{X}(e) = \boldsymbol{0}$, then for any $g \in G$, $\boldsymbol{X}(g) = L_{g*}\boldsymbol{X}(e) = \boldsymbol{0}$, so $\boldsymbol{h}$ has trivial kernel. To

see that it is onto, consider an arbitrary $\boldsymbol{u} \in G_e$. Define a vector field $\boldsymbol{X}$ on G by setting $\boldsymbol{X}(g) = L_{g*}\boldsymbol{u}$. $\boldsymbol{X}$ is left-invariant, since for any $a, b \in G$,

$$L_{a*}\boldsymbol{X}(b) = L_{a*}L_{b*}\boldsymbol{v} = L_{(ab)*}\boldsymbol{u} = \boldsymbol{X}(ab) = \boldsymbol{X} \circ L_a(b).$$

Furthermore, $\boldsymbol{hX} = \boldsymbol{u}$ by construction. This completes the argument. □

We next investigate integral curves of left-invariant vector fields. Given groups G, H, a *group homomorphism* from G to H is a map $\boldsymbol{f} : G \to H$ which preserves group products; i.e., $\boldsymbol{f}(ab) = \boldsymbol{f}(a)\boldsymbol{f}(b)$ for all $a, b \in G$. Any homomorphism maps the identity element $e \in G$ to the identity $e \in H$, because $\boldsymbol{f}(e) = \boldsymbol{f}(e \cdot e) = \boldsymbol{f}(e) \cdot \boldsymbol{f}(e)$; multiplying both sides by $\boldsymbol{f}(e)^{-1}$ then yields $\boldsymbol{f}(e) = e$. This in turn implies that $\boldsymbol{f}(a^{-1}) = \boldsymbol{f}(a)^{-1}$ for any $a \in G$.

In the following theorem, $\mathbb{R}$ is the Lie group with the usual addition.

Theorem 3.4.2. *Any smooth Lie group homomorphism $\boldsymbol{c} : \mathbb{R} \to G$ is the integral curve passing through e at 0 of the left-invariant vector field $\boldsymbol{X}$ with $\boldsymbol{X}(e) = \dot{\boldsymbol{c}}(0)$. The integral curve of $\boldsymbol{X}$ that passes through $g \in G$ at 0 is $L_g \circ \boldsymbol{c}$.*

Conversely, if $\boldsymbol{c}$ is an integral curve of $\boldsymbol{X} \in \mathfrak{g}$ with $\boldsymbol{c}(0) = e$, then $\boldsymbol{c} : \mathbb{R} \to G$ is a Lie group homomorphism. In particular, left-invariant vector fields are complete.

Proof. Let $\boldsymbol{c} : \mathbb{R} \to G$ be a homomorphism, and fix any $t_0 \in \mathbb{R}$. The curve $\boldsymbol{\gamma} : \mathbb{R} \to G$ given by $\boldsymbol{\gamma}(t) = \boldsymbol{c}(t + t_0)$ satisfies $\boldsymbol{\gamma}(t) = \boldsymbol{c}(t_0)\boldsymbol{c}(t) = L_{\boldsymbol{c}(t_0)}\boldsymbol{c}(t)$. Thus if $\boldsymbol{X} \in \mathfrak{g}$ is the left-invariant vector field that equals $\dot{\boldsymbol{c}}(0)$ at e, then

$$\dot{\boldsymbol{c}}(t_0) = \dot{\boldsymbol{\gamma}}(0) = L_{\boldsymbol{c}(t_0)*}\dot{\boldsymbol{c}}(0) = L_{\boldsymbol{c}(t_0)*}\boldsymbol{X}(e) = \boldsymbol{X}(\boldsymbol{c}(t_0)).$$

This shows that $\boldsymbol{c}$ is indeed the integral curve of $\boldsymbol{X}$ that passes through e at 0. Furthermore, if $g \in G$ and $\boldsymbol{c}_1 = L_g \circ \boldsymbol{c}$, then $\boldsymbol{c}_1(0) = g$ and

$$\dot{\boldsymbol{c}}_1(t) = L_{g*}\dot{\boldsymbol{c}}(t) = L_{g*}\boldsymbol{X}(\boldsymbol{c}(t)) = \boldsymbol{X}(L_g \circ \boldsymbol{c}(t)) = \boldsymbol{X}(\boldsymbol{c}_1(t)).$$

Conversely, let $\boldsymbol{c} : (a, b) \to G$ be the maximal integral curve of some $\boldsymbol{X} \in \mathfrak{g}$ passing through e at 0. Given $t, \tilde{t} \in \mathbb{R}$ such that $t, \tilde{t}, t + \tilde{t} \in (a, b)$, we must have $\boldsymbol{c}(t + \tilde{t}) = \boldsymbol{c}(t)\boldsymbol{c}(\tilde{t})$ because the curves $t \mapsto \boldsymbol{c}(t + \tilde{t})$ and $t \mapsto \boldsymbol{c}(t)\boldsymbol{c}(\tilde{t})$ are both integral curves of $\boldsymbol{X}$ and they agree at 0. It only remains to show that $(a, b) = \mathbb{R}$. But if, say, $b < \infty$, choose $\tilde{b} \in (0, b)$; the curve $\tilde{\boldsymbol{c}} : (a + \tilde{b}, b + \tilde{b}) \to G$ defined by $\tilde{\boldsymbol{c}}(t) = \boldsymbol{c}(\tilde{b})\boldsymbol{c}(t - \tilde{b})$ is an integral curve of $\boldsymbol{X}$ that coincides with $\boldsymbol{c}$ at $\tilde{b}$, and must therefore coincide with $\boldsymbol{c}$ everywhere. Thus, the domain of $\boldsymbol{c}$ may be extended beyond b, contradicting our assumption. □

Example 3.4.1. Let us illustrate these various concepts for the orthogonal group $G = O(n)$ introduced earlier. For the sake of brevity, the identity matrix I_n will be denoted by e. We begin with the Lie algebra of G, which may be identified with G_e. By Proposition 3.1.1, $\imath_* G_e = \ker \boldsymbol{f}_{*e}$, where $\boldsymbol{f} : GL(n) \to M_n$ is given by $\boldsymbol{f}(M) = MM^T$, and $\imath : G \hookrightarrow GL(n)$ denotes inclusion. We already computed that $D\boldsymbol{f}(A)M = AM^T + MA^T$, so that

$$G_e = \ker \boldsymbol{f}_{*e} = \{\mathcal{I}_e M \mid M + M^T = 0\}$$

is isomorphic to the space of skew-symmetric matrices. The left-invariant field $\boldsymbol{X}$ with $\boldsymbol{X}(e) = \mathcal{I}_e M$ is given by

$$\boldsymbol{X}(A) = L_{A*e}\boldsymbol{X}(e) = L_{A_*e}\mathcal{I}_e M = \mathcal{I}_A DL_A(e)M = \mathcal{I}_A(AM), \qquad A \in G$$

since $L_A : M_n \to M_n$ is linear, so that $DL_A(e) = L_A$. The integral curve $\boldsymbol{c}$ of $\boldsymbol{X}$ that passes through e at time 0 has as expression $\boldsymbol{c}(t) = \exp(tM)$ by the theorem, since $\boldsymbol{c} : \mathbb{R} \to G$ is a homomorphism with $\dot{\boldsymbol{c}}(0) = \mathcal{I}_e M$. Notice that $\boldsymbol{c}$ does indeed have its image in G, because

$$\boldsymbol{c}(t)\boldsymbol{c}(t)^T = \exp(tM)\exp(tM^T) = \exp(t(M + M^T)) = \exp(\mathbf{0}) = e$$

by Exercise 1.50. It follows that the integral curve of $\boldsymbol{X}$ that passes through A at time zero is $t \mapsto A\exp(tM)$.

Next, we investigate the Lie bracket $[\boldsymbol{X}, \boldsymbol{Y}]$ for $\boldsymbol{X}, \boldsymbol{Y} \in \mathfrak{g}$ with $\boldsymbol{X}(e) = \mathcal{I}_e M$ and $\boldsymbol{Y}(e) = \mathcal{I}_e N$. Since $[\boldsymbol{X}, \boldsymbol{Y}] \in \mathfrak{g}$,

$$[\boldsymbol{X}, \boldsymbol{Y}](A) = \mathcal{I}_A(A\mathcal{I}_e^{-1}[\boldsymbol{X}, \boldsymbol{Y}](e)), \qquad A \in G,$$

so that we need only identify $[\boldsymbol{X}, \boldsymbol{Y}](e) = D_{\boldsymbol{X}(e)}\boldsymbol{Y} - D_{\boldsymbol{Y}(e)}\boldsymbol{X}$. Now, the curve $\boldsymbol{c}$, where $\boldsymbol{c}(t) = e + tM$, has velocity vector $\mathcal{I}_e M$ at 0, so that $(D_{\boldsymbol{X}}\boldsymbol{Y})(e) = (\boldsymbol{Y}\dot{\circ}\boldsymbol{c})(0)$. Furthermore, $(\boldsymbol{Y} \circ \boldsymbol{c})(t) = \mathcal{I}_{\boldsymbol{c}(t)}(\boldsymbol{c}(t)N) = \mathcal{I}_{\boldsymbol{c}(t)}((e + tM)N)$, which implies $(D_{\boldsymbol{X}}\boldsymbol{Y})(e) = \mathcal{I}_e(MN)$. The Lie bracket is therefore given by

$$[\boldsymbol{X}, \boldsymbol{Y}](e) = \mathcal{I}_e(MN - NM), \qquad \boldsymbol{X}(e) = \mathcal{I}_e M, \quad \boldsymbol{Y}(e) = \mathcal{I}_e N.$$

In the computation of the Lie bracket, nowhere did we use the fact that $\boldsymbol{X}$ and $\boldsymbol{Y}$ belong to the Lie algebra of $O(n)$. The same argument shows that the above formula actually holds for the Lie algebra of $GL(n)$ whose Lie algebra is isomorphic to M_n. Notice also that $GL(1)$ is just $\mathbb{R} \setminus \{0\}$ with ordinary multiplication.

Remark 3.4.1. The exponential map on M_n is extendable to the Lie algebra $\mathfrak{g}$ of any Lie group G: given $\boldsymbol{X} \in \mathfrak{g}$, define $\exp(\boldsymbol{X}) = \boldsymbol{c}_{\boldsymbol{X}}(1)$, where $\boldsymbol{c}_{\boldsymbol{X}}$ is the integral curve of $\boldsymbol{X}$ passing through the identity at time 0. It follows that this integral curve is given by $t \mapsto \exp(t\boldsymbol{X})$ for all t: to see this, let $t_0 \in \mathbb{R}$, and consider the curve $\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}(t) = \boldsymbol{c}_{\boldsymbol{X}}(t_0 t)$. Then

$$\dot{\boldsymbol{\gamma}}(t) = t_0\dot{\boldsymbol{c}}_{\boldsymbol{X}}(t_0 t) = (t_0\boldsymbol{X} \circ \boldsymbol{c}_{\boldsymbol{X}})(t_0 t) = (t_0\boldsymbol{X} \circ \boldsymbol{\gamma})(t),$$

so that $\boldsymbol{\gamma}$ is the integral curve of $t_0\boldsymbol{X}$ through e at time 0. By uniqueness, $\boldsymbol{\gamma} = \boldsymbol{c}_{t_0\boldsymbol{X}}$. Thus, $\boldsymbol{c}_{\boldsymbol{X}}(t_0) = \boldsymbol{\gamma}(1) = \exp(t_0\boldsymbol{X})$. Since t_0 was arbitrary, the claim follows.

$\exp : \mathfrak{g} \to G$ shares many of the properties of the matrix exponential map. For example, $\exp(-\boldsymbol{X}) = (\exp \boldsymbol{X})^{-1}$: the term on the left equals, by the above, $\boldsymbol{c}_{\boldsymbol{X}}(-1)$, whereas the one on the right is $\boldsymbol{c}_{\boldsymbol{X}}(1)^{-1}$. But $\boldsymbol{c}_{\boldsymbol{X}}$ is a homomorphism, so $e = \boldsymbol{c}_{\boldsymbol{X}}(0) = \boldsymbol{c}_{\boldsymbol{X}}(1 - 1) = \boldsymbol{c}_{\boldsymbol{X}}(1)\,\boldsymbol{c}_{\boldsymbol{X}}(-1)$, which means that $\boldsymbol{c}_{\boldsymbol{X}}(-1)$ is the inverse of $\boldsymbol{c}_{\boldsymbol{X}}(1)$.

More generally, recall that for $A, B \in M_n$, if $AB = BA$, then $\exp(A + B) = \exp A \exp B$. By the above example, identifying M_n with the Lie algebra of $GL(n)$, this says that for $\boldsymbol{X}, \boldsymbol{Y} \in \mathfrak{g}$, $\exp(\boldsymbol{X} + \boldsymbol{Y}) = \exp(\boldsymbol{X})\exp(\boldsymbol{Y})$ whenever $[\boldsymbol{X}, \boldsymbol{Y}] = \mathbf{0}$. This is also true in arbitrary Lie algebras. A proof is outlined in Exercise 3.22.

3.5 The tangent bundle

The main concepts – derivatives and vector fields – introduced in the last three sections involved maps taking values in tangent spaces. These tangent spaces can be grouped together to form a new manifold, and the corresponding maps become differentiable in the process. If M^n is a submanifold of $\mathbb{R}^{n+k}$, define the *tangent bundle* TM of M to be

$$TM = \bigcup_{\boldsymbol{p}\in M} M_{\boldsymbol{p}}.$$

Each tangent space $M_{\boldsymbol{p}}$ is a subset of $\mathbb{R}^{n+k}_{\boldsymbol{p}} = \{\boldsymbol{p}\} \times \mathbb{R}^{n+k} \subset \mathbb{R}^{n+k} \times R^{n+k}$, so that the tangent bundle of M is contained in Euclidean space of dimension $2(n+k)$.

Theorem 3.5.1. *The tangent bundle TM of M^n is a $2n$-dimensional manifold.*

Proof. Any parametrization $(U, \boldsymbol{h})$ of M induces one of TM: Define

$$\begin{aligned} \boldsymbol{F} : U \times \mathbb{R}^n &\to TM, \\ (\boldsymbol{q}, \boldsymbol{u}) &\mapsto (\boldsymbol{h}(\boldsymbol{q}), D\boldsymbol{h}(\boldsymbol{q})\boldsymbol{u}). \end{aligned}$$

The first $n+k$ component functions of $\boldsymbol{F}$ are those of $\boldsymbol{h}$, and hence are differentiable. So are the last $n+k$, since

$$u^{n+k+j} \circ \boldsymbol{F}(\boldsymbol{q}, \boldsymbol{u}) = \sum_{i=1}^{n} D_i h^j(\boldsymbol{q}) u_i.$$

Thus, $\boldsymbol{F}$ is differentiable, and has maximal rank because $\boldsymbol{h}$ has. It remains to show that $\boldsymbol{F}^{-1}$ is continuous. So consider a convergent sequence

$$(\boldsymbol{h}(\boldsymbol{p}_n), D\boldsymbol{h}(\boldsymbol{p}_n)\boldsymbol{u}_n) \to (\boldsymbol{h}(\boldsymbol{p}), D\boldsymbol{h}(\boldsymbol{p})\boldsymbol{u})$$

in the image of $\boldsymbol{F}$. It must be shown that $\boldsymbol{p}_n \to \boldsymbol{p}$ and $\boldsymbol{u}_n \to \boldsymbol{u}$. The first sequence converges by continuity of $\boldsymbol{h}^{-1}$. Since $\boldsymbol{h}$ is continuously differentiable, this, in turn, implies that

$$D\boldsymbol{h}(\boldsymbol{p}_n) \to D\boldsymbol{h}(\boldsymbol{p}), \tag{3.5.1}$$

in the sense that $|D\boldsymbol{h}(\boldsymbol{p}_n) - D\boldsymbol{h}(\boldsymbol{p})| \to 0$ with the norm from Definition 1.4.1. By Proposition 3.2.1, there exists an open set $V \subset \mathbb{R}^{2(n+k)}$ and a smooth map $\boldsymbol{G} : V \to \mathbb{R}^{n+k}$ such that $\boldsymbol{G} \circ \boldsymbol{h} = 1_U$. Applying $D\boldsymbol{G}(\boldsymbol{p})$ to (3.5.1), we then obtain

$$D\boldsymbol{G}(\boldsymbol{p})D\boldsymbol{h}(\boldsymbol{p}_n) \to 1_{\mathbb{R}^{n+k}}. \tag{3.5.2}$$

Now,

$$|\boldsymbol{u}_n - \boldsymbol{u}| \le |\boldsymbol{u}_n - D\boldsymbol{G}(\boldsymbol{p})D\boldsymbol{h}(\boldsymbol{p}_n)\boldsymbol{u}_n| + |D\boldsymbol{G}(\boldsymbol{p})D\boldsymbol{h}(\boldsymbol{p}_n)\boldsymbol{u}_n - \boldsymbol{u}|.$$

The first term on the right is no larger than $|1_{\mathbb{R}^{n+k}} - D\boldsymbol{G}(\boldsymbol{p})D\boldsymbol{h}(\boldsymbol{p}_n)||\boldsymbol{u}_n|$, which goes to zero by (3.5.2) (we are implicitly using here the fact that $\{\boldsymbol{u}_n\}$ is bounded, which follows from convergence of $\{D\boldsymbol{h}(\boldsymbol{p}_n)\boldsymbol{u}_n\}$ and $\{D\boldsymbol{h}(\boldsymbol{p}_n)\}$). The second term can be written

$$|D\boldsymbol{G}(\boldsymbol{p})D\boldsymbol{h}(\boldsymbol{p}_n)\boldsymbol{u}_n - D\boldsymbol{G}(\boldsymbol{p})D\boldsymbol{h}(\boldsymbol{p})\boldsymbol{u}| \le |D\boldsymbol{G}(\boldsymbol{p})||D\boldsymbol{h}(\boldsymbol{p}_n)\boldsymbol{u}_n - D\boldsymbol{h}(\boldsymbol{p})\boldsymbol{u}|,$$

which also goes to zero since $D\boldsymbol{h}(\boldsymbol{p}_n)\boldsymbol{u}_n \to D\boldsymbol{h}(\boldsymbol{p})\boldsymbol{u}$ by hypothesis. This shows that $\boldsymbol{F}^{-1}$ is continuous. □

One could also have argued Theorem 3.5.1 in terms of charts instead of parametrizations. The reader is invited to check that a chart $(U, \boldsymbol{x})$ of M induces a chart $(\pi^{-1}(U), \bar{\boldsymbol{x}})$ of TM where

$$\bar{\boldsymbol{x}} = (\boldsymbol{x}, \mathrm{d}x^1, \ldots, \mathrm{d}x^n) \circ \pi. \tag{3.5.3}$$

Given a smooth map $\boldsymbol{f} : M \to N$ between manifolds M and N, we define a map $\boldsymbol{f}_* : TM \to TN$ between their tangent bundles by setting

$$\boldsymbol{f}_*\boldsymbol{u} = \boldsymbol{f}_{*\boldsymbol{p}}\boldsymbol{u}, \quad \text{if } \boldsymbol{u} \in M_{\boldsymbol{p}}.$$

Corollary 3.5.1. *If $\boldsymbol{f} : M \to N$ is differentiable, then so is $\boldsymbol{f}_* : TM \to TN$.*

Proof. It must be shown that if $(U \times \mathbb{R}^n, \boldsymbol{F})$ is a parametrization of TM as in Theorem 3.5.1, then $\boldsymbol{f}_* \circ \boldsymbol{F}$ is differentiable in the usual sense. But

$$(\boldsymbol{f}_* \circ \boldsymbol{F})(\boldsymbol{p}, \boldsymbol{u}) = \boldsymbol{f}_*(\boldsymbol{h}(\boldsymbol{p}), D\boldsymbol{h}(\boldsymbol{p})\boldsymbol{u}) = ((\boldsymbol{f} \circ \boldsymbol{h})(\boldsymbol{p}), D(\boldsymbol{f} \circ \boldsymbol{h})(\boldsymbol{p})\boldsymbol{u})$$

by (3.2.3), which establishes the claim. □

The *bundle projection* $\pi_M : TM \to M$ is the map that sends a vector to its base point; i.e., any $\boldsymbol{v} \in TM$ belongs to some $M_{\boldsymbol{p}}$ for a unique $\boldsymbol{p} \in M$, and $\pi_M(\boldsymbol{v})$ is defined to be this $\boldsymbol{p}$. It is an easy exercise to show that π_M is differentiable. When $M = \mathbb{R}^n$,

$$T\mathbb{R}^n = \bigcup_{\boldsymbol{p}\in\mathbb{R}^n} \mathbb{R}^n_{\boldsymbol{p}} = \bigcup_{\boldsymbol{p}\in\mathbb{R}^n} \{\boldsymbol{p}\} \times \mathbb{R}^n = \mathbb{R}^n \times \mathbb{R}^n,$$

so that the bundle projection is the projection $\pi_1 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ onto the first factor. If π_2 is the projection onto the second factor, then $\boldsymbol{u} = \mathcal{I}_{\boldsymbol{p}}\boldsymbol{v} \in T\mathbb{R}^n$ if and only if $\pi_1(\boldsymbol{u}) = \boldsymbol{p}$ and $\pi_2(\boldsymbol{u}) = \boldsymbol{v}$. In general, notice that for $\boldsymbol{f} : N \to M$, the diagram

$$\begin{array}{ccc} TN & \xrightarrow{\boldsymbol{f}_*} & TM \\ {\scriptstyle \pi_N}\downarrow & & \downarrow{\scriptstyle \pi_M} \\ N & \xrightarrow[\boldsymbol{f}]{} & M \end{array}$$

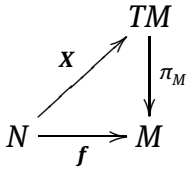
commutes.

Corollary 3.5.2. *A vector field $\boldsymbol{X}$ on M is differentiable as a map $\boldsymbol{X} : M \to TM$.*

Proof. $\boldsymbol{X}$ is differentiable if $\boldsymbol{X} \circ \boldsymbol{h}$ is smooth for any local parametrization $(U, \boldsymbol{h})$ of M. But this is immediate from (3.3.1). □

Definition 3.5.1. Let N, M denote manifolds, $\boldsymbol{f} : N \to M \subset \mathbb{R}^{n+k}$ a map. A *vector field along $\boldsymbol{f}$* is a differentiable map $\boldsymbol{X} : N \to TM$ that assigns to each $\boldsymbol{p} \in N$ a vector

$\boldsymbol{X}(\boldsymbol{p}) \in M_{\boldsymbol{f}(\boldsymbol{p})}$; i.e., $\boldsymbol{X}$ is a differentiable map for which the diagram



commutes.

The above definition clearly generalizes that of vector fields along maps between Euclidean spaces given in Chapter 2.

Remark 3.5.1. Any vector field $\boldsymbol{X}$ on N induces a vector field along $\boldsymbol{f} : N \to M$, namely $\boldsymbol{p} \mapsto \boldsymbol{f}_{*\boldsymbol{p}}\boldsymbol{X}(\boldsymbol{p})$. Similarly, any vector field $\boldsymbol{Y}$ on M induces one along $\boldsymbol{f}$, namely $\boldsymbol{Y} \circ \boldsymbol{f}$. In general, given $\boldsymbol{X}$ as above, there need not exist any vector field $\boldsymbol{Y}$ on M such that $\boldsymbol{f}_*\boldsymbol{X} = \boldsymbol{Y} \circ \boldsymbol{f}$ (i.e., such that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $\boldsymbol{f}$-related). If $\boldsymbol{f}_{*\boldsymbol{p}}$ is one-to-one at $\boldsymbol{p} \in N$ however, then $\boldsymbol{f}$ is one-to-one on a neighborhood of $\boldsymbol{p}$ by the inverse function theorem; thus, there exists a neighborhood U of $\boldsymbol{p}$ and a vector field $\tilde{\boldsymbol{X}}$ on $\boldsymbol{f}(U)$ that is $\boldsymbol{f}$-related to $\boldsymbol{X}$ (namely, $\tilde{\boldsymbol{X}} = \boldsymbol{f}_*\boldsymbol{X} \circ (\boldsymbol{f}_{|U})^{-1}$). Smoothness of $\boldsymbol{X}$ in the definition above is then equivalent to smoothness of $\tilde{\boldsymbol{X}}$.

3.6 Covariant differentiation

In Definition 2.8.7, we introduced the covariant derivative $D_{\boldsymbol{u}}\boldsymbol{X}$ of a vector field $\boldsymbol{X}$ in Euclidean space $\mathbb{R}^n$ with respect to a vector $\boldsymbol{u} \in \mathbb{R}^n_{\boldsymbol{p}}$. We wish to extend this concept to a vector field $\boldsymbol{X}$ on a manifold $M \subset \mathbb{R}^n$. Merely adopting the same formula does not work, because although $D_{\boldsymbol{u}}\boldsymbol{X} \in \mathbb{R}^n_{\boldsymbol{p}}$, it need not belong to the tangent space of M at $\boldsymbol{p}$. The best we can do is to project this vector back onto $M_{\boldsymbol{p}}$. This can be done using the Riemannian metric on $\mathbb{R}^n$: there is a decomposition

$$\mathbb{R}^n_{\boldsymbol{p}} = M_{\boldsymbol{p}} \oplus M^{\perp}_{\boldsymbol{p}}$$

of $\mathbb{R}^n_{\boldsymbol{p}}$ as a direct sum of the tangent space of M at that point with its orthogonal complement, and any $\boldsymbol{u} \in \mathbb{R}^n_{\boldsymbol{p}}$ has a corresponding unique decomposition

$$\boldsymbol{u} = \boldsymbol{u}^{\top} + \boldsymbol{u}^{\perp} \in M_{\boldsymbol{p}} \oplus M^{\perp}_{\boldsymbol{p}},$$

thus inducing maps $\top : \mathbb{R}^n_{\boldsymbol{p}} \to M_{\boldsymbol{p}}$, $\perp : \mathbb{R}^n_{\boldsymbol{p}} \to M^{\perp}_{\boldsymbol{p}}$, that project $\boldsymbol{u}$ onto $\boldsymbol{u}^{\top}$ and $\boldsymbol{u}^{\perp}$ respectively. $M^{\perp}_{\boldsymbol{p}}$ is called the *normal space* of M at $\boldsymbol{p}$.

Definition 3.6.1. If M is a submanifold of $\mathbb{R}^n$, $\boldsymbol{X}$ a vector field on some open set $U \subset M$, $\boldsymbol{p} \in U$, and $\boldsymbol{u} \in M_{\boldsymbol{p}}$, the *covariant derivative* of $\boldsymbol{X}$ with respect to $\boldsymbol{u}$ is

$$\nabla_{\boldsymbol{u}}\boldsymbol{X} := (D_{\boldsymbol{u}}\boldsymbol{X})^{\top}. \tag{3.6.1}$$

More generally, as in the Euclidean setting, the same formula defines the covariant derivative of a vector field $\boldsymbol{X}$ along a map $\boldsymbol{f} : N \to M$ with respect to some $\boldsymbol{u} \in N_{\boldsymbol{p}}$. Such an $\boldsymbol{X}$ will sometimes be called a *vector field in M along $\boldsymbol{f}$* to emphasize the fact that $\boldsymbol{X}$ takes values in the tangent bundle of M rather than that of the ambient space. In particular, if $\boldsymbol{X}$ is a vector field in M along a curve $\boldsymbol{c} : I \to M$, its covariant derivative at t is

$$(\nabla_{\boldsymbol{D}}\boldsymbol{X})(t) := \nabla_{\boldsymbol{D}(t)}\boldsymbol{X} = (\boldsymbol{X}'(t))^{\top}. \tag{3.6.2}$$

Likewise, given vector fields $\boldsymbol{X}, \boldsymbol{Y}$ on M, one obtains a new vector field $\nabla_{\boldsymbol{X}}\boldsymbol{Y}$ by defining $(\nabla_{\boldsymbol{X}}\boldsymbol{Y})(\boldsymbol{p}) = \nabla_{\boldsymbol{X}(\boldsymbol{p})}\boldsymbol{Y}$. It must, of course, be checked that $\nabla_{\boldsymbol{X}}\boldsymbol{Y}$ is differentiable. For this, we will use the following:

Theorem 3.6.1. *If M^n is a submanifold of $\mathbb{R}^{n+k}$, then for any $\boldsymbol{p} \in M$, there exists an open neighborhood U of $\boldsymbol{p}$ in $\mathbb{R}^{n+k}$, and (differentiable) vector fields $\boldsymbol{N}_1, \dots, \boldsymbol{N}_k$ on U such that $\boldsymbol{N}_1(\boldsymbol{q}), \dots, \boldsymbol{N}_k(\boldsymbol{q})$ form an orthonormal basis of the normal space of M at each $\boldsymbol{q} \in U \cap M$.*

Proof. By Corollary 3.2.1, there exists a neighborhood U of $\boldsymbol{p}$ such that $U \cap M = \boldsymbol{f}^{-1}(\boldsymbol{0})$ for some map $\boldsymbol{f} : U \to \mathbb{R}^k$ that has $\boldsymbol{0}$ as regular value. Since the tangent space of M at any point of U equals the kernel of $\boldsymbol{f}_*$ at that point, and since $\boldsymbol{u} \in \ker \boldsymbol{f}_{*\boldsymbol{q}}$ if and only if $0 = D_{\boldsymbol{u}}f^i = \langle \nabla f^i(\boldsymbol{q}), \boldsymbol{u}\rangle$ for $i = 1, \dots, k$ (see Examples 2.8.2 (iv)), it follows that the vector fields ∇f^i on U are linearly independent vector fields that span the normal space of M at every point of $U \cap M$. Apply the Gram-Schmidt orthogonalization process (Theorem 1.4.2), observing that if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are differentiable, then so is $\operatorname{proj}_{\boldsymbol{X}} \boldsymbol{Y}$, to obtain differentiable fields $\boldsymbol{N}_1, \dots, \boldsymbol{N}_k$ on U that form an orthonormal basis of the normal space of M at each $\boldsymbol{q} \in U \cap M$. □

To see that $\nabla_{\boldsymbol{X}}\boldsymbol{Y}$ is differentiable if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are, recall that the latter may be extended to an open neighborhood U of any $\boldsymbol{p}$ (Theorem 3.3.1). This neighborhood may be assumed to be the one in Theorem 3.6.1. Then $D_{\boldsymbol{X}}\boldsymbol{Y}$ is differentiable on U, and so is

$$D_{\boldsymbol{X}}\boldsymbol{Y} - \sum_{i=1}^{k} \langle D_{\boldsymbol{X}}\boldsymbol{Y}, \boldsymbol{N}_i\rangle \boldsymbol{N}_i.$$

But by definition this vector field equals $\nabla_{\boldsymbol{X}}\boldsymbol{Y}$ at any point of $U \cap M$. This shows that $\nabla_{\boldsymbol{X}}\boldsymbol{Y}$ is smooth.

The exact same argument shows that if $\boldsymbol{X}$ is a vector field on N and $\boldsymbol{Y}$ is a vector field along $\boldsymbol{f} : N \to M$, then the vector field $\nabla_{\boldsymbol{X}}\boldsymbol{Y}$ along $\boldsymbol{f}$ is smooth.

The properties of covariant derivatives in Euclidean space that were established in Chapter 2 also hold on manifolds:

Theorem 3.6.2. *Let $\boldsymbol{X}, \boldsymbol{Y}$ denote vector fields on an open subset U of a manifold M, $f : U \to \mathbb{R}$, $a \in \mathbb{R}$. Given $\boldsymbol{p} \in U$, $\boldsymbol{u} \in M_{\boldsymbol{p}}$,*

(1) $\nabla_{\boldsymbol{u}}(a\boldsymbol{X} + \boldsymbol{Y}) = a\nabla_{\boldsymbol{u}}\boldsymbol{X} + \nabla_{\boldsymbol{u}}\boldsymbol{Y}$;

(2) $\nabla_{a\boldsymbol{u}+\boldsymbol{v}}\boldsymbol{X} = a\nabla_{\boldsymbol{u}}\boldsymbol{X} + \nabla_{\boldsymbol{v}}\boldsymbol{X}$;

(3) $\nabla_{\boldsymbol{u}} f\boldsymbol{X} = \boldsymbol{u}(f)\boldsymbol{X}(\boldsymbol{p}) + f(\boldsymbol{p})\nabla_{\boldsymbol{u}}\boldsymbol{X}$;

(4) $\nabla_{f\boldsymbol{X}}\boldsymbol{Y} = f\nabla_{\boldsymbol{X}}\boldsymbol{Y}$;
(5) $\nabla_{\boldsymbol{X}}\boldsymbol{Y} - \nabla_{\boldsymbol{Y}}\boldsymbol{X} = [\boldsymbol{X}, \boldsymbol{Y}]$;
(6) *Given* $\boldsymbol{g} : N \to M$, $\boldsymbol{v} \in TN$, $\nabla_{\boldsymbol{v}}(\boldsymbol{X} \circ \boldsymbol{g}) = \nabla_{\boldsymbol{g}_*\boldsymbol{v}}\boldsymbol{X}$;
(7) *Let* $\boldsymbol{X}$, $\boldsymbol{Y}$ *denote vector fields in* M *along a map* $\boldsymbol{f} : N \to M$, $\boldsymbol{p} \in N$, $\boldsymbol{u} \in N_{\boldsymbol{p}}$. *Then*

$$\boldsymbol{u}\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \langle \nabla_{\boldsymbol{u}}\boldsymbol{X}, \boldsymbol{Y}(\boldsymbol{p})\rangle + \langle \boldsymbol{X}(\boldsymbol{p}), \nabla_{\boldsymbol{u}}\boldsymbol{Y}\rangle.$$

Proof. These properties follow from the corresponding ones in Euclidean space that were established in Theorem 2.8.4, together with linearity of the projection $\top$ for the first four. In the same way, (6) is a consequence of (2.8.4). To prove the identity (5), consider any $\boldsymbol{p} \in M$, and extend the vector fields $\boldsymbol{X}$ and $\boldsymbol{Y}$ to a neighborhood of $\boldsymbol{p}$ in the ambient Euclidean space. Then $[\boldsymbol{X}, \boldsymbol{Y}](\boldsymbol{p}) = D_{\boldsymbol{X}(\boldsymbol{p})}\boldsymbol{Y} - D_{\boldsymbol{Y}(\boldsymbol{p})}\boldsymbol{X}$. But the left side of this identity is tangent to M. Thus,

$$[\boldsymbol{X}, \boldsymbol{Y}](\boldsymbol{p}) = (D_{\boldsymbol{X}(\boldsymbol{p})}\boldsymbol{Y} - D_{\boldsymbol{Y}(\boldsymbol{p})}\boldsymbol{X})^{\top} = \nabla_{\boldsymbol{X}(\boldsymbol{p})}\boldsymbol{Y} - \nabla_{\boldsymbol{Y}(\boldsymbol{p})}\boldsymbol{X}.$$

For the last one, notice that since $\boldsymbol{Y}(\boldsymbol{p})$ is tangent to M,

$$\begin{aligned}\langle D_{\boldsymbol{u}}\boldsymbol{X}, \boldsymbol{Y}(\boldsymbol{p})\rangle &= \langle (D_{\boldsymbol{u}}\boldsymbol{X})^{\top} + (D_{\boldsymbol{u}}\boldsymbol{X})^{\perp}, \boldsymbol{Y}(\boldsymbol{p})\rangle = \langle (D_{\boldsymbol{u}}\boldsymbol{X})^{\top}, \boldsymbol{Y}(\boldsymbol{p})\rangle \\ &= \langle \nabla_{\boldsymbol{u}}\boldsymbol{Y}(\boldsymbol{p}), \boldsymbol{Y}(\boldsymbol{p})\rangle,\end{aligned}$$

with a similar identity holding for the other term. The claim then follows from Theorem 2.8.4. □

Definition 3.6.2. A vector field $\boldsymbol{X}$ in M along a curve $\boldsymbol{c} : I \to M$ is said to be *parallel* if

$$\nabla_{\boldsymbol{D}}\boldsymbol{X} = \boldsymbol{0}.$$

Notice that if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are parallel along $\boldsymbol{c}$, then $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle$ is a constant function by part (7) of Theorem 3.6.2. In particular, parallel vector fields have constant norm.

Theorem 3.6.3. *Let* $\boldsymbol{c} : [0, a] \to M$ *be a curve in a manifold* $M^n \subset \mathbb{R}^{n+k}$. *Then*
(1) *For any* $\boldsymbol{u} \in M_{\boldsymbol{c}(0)}$, *there exists one and only one parallel field* $\boldsymbol{X}$ *along* $\boldsymbol{c}$ *such that* $\boldsymbol{X}(0) = \boldsymbol{u}$.
(2) *The (well-defined by (1)) map* $M_{\boldsymbol{c}(0)} \to M_{\boldsymbol{c}(a)}$ *which sends* $\boldsymbol{u} \in M_{\boldsymbol{c}(0)}$ *to* $\boldsymbol{X}(a)$, *where* $\boldsymbol{X}$ *is the parallel field along* $\boldsymbol{c}$ *with* $\boldsymbol{X}(0) = \boldsymbol{u}$ *is a linear isometry, called* parallel translation along $\boldsymbol{c}$.

Proof. The image of $\boldsymbol{c}$ can be broken up into finitely many pieces, each of which lie in an open set admitting an orthonormal basis of vector fields normal to M. Since it suffices to prove the theorem for each such piece, we may assume the whole image lies in such a set, so that there exist vector fields $\boldsymbol{N}_1, \ldots, \boldsymbol{N}_k$ such that $\{\boldsymbol{N}_i \circ \boldsymbol{c}(t) \mid 1 \le i \le k\}$ is an orthonormal basis of $M^{\perp}_{\boldsymbol{c}(t)}$ for each $t \in [0, a]$. A vector field $\boldsymbol{X}$ along $\boldsymbol{c}$ is parallel

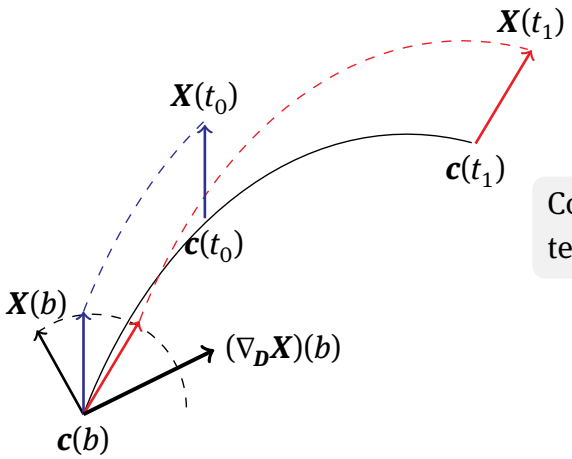
iff

$$\begin{aligned}\mathbf{0} = \nabla_{\boldsymbol{D}}\boldsymbol{X} &= \boldsymbol{X}' - \sum_j \langle \boldsymbol{X}', \boldsymbol{N}_j \circ \boldsymbol{c}\rangle \boldsymbol{N}_j \circ \boldsymbol{c} \\ &= \boldsymbol{X}' - \sum_j \big(\langle \boldsymbol{X}, \boldsymbol{N}_j \circ \boldsymbol{c}\rangle' - \langle \boldsymbol{X}, (\boldsymbol{N}_j \circ \boldsymbol{c})'\rangle\big) \boldsymbol{N}_j \circ \boldsymbol{c} \\ &= \boldsymbol{X}' + \sum_j \langle \boldsymbol{X}, (\boldsymbol{N}_j \circ \boldsymbol{c})'\rangle \boldsymbol{N}_j \circ \boldsymbol{c}.\end{aligned}$$

Writing $\boldsymbol{X} = \sum X^i \boldsymbol{D}_i \circ \boldsymbol{c}$ in components, with $X^i = \langle \boldsymbol{X}, \boldsymbol{D}_i \circ \boldsymbol{c}\rangle$, and doing the same for $\boldsymbol{N}_j$, the above equation becomes a system of ordinary differential equations

$$X^{i\prime} + \sum_{j,l} X^l (N_j^l \circ \boldsymbol{c})(N_j^i \circ \boldsymbol{c}) = 0, \quad i = 1, \dots, n+k.$$

This system is linear, in the sense that a sum of two solutions as well as a scalar multiple of a solution are again solutions. The theory of linear differential equations then guarantees the existence of a unique collection X^i of solutions defined on $[0, a]$ with initial conditions $X^i(0) = u^i$, where $u^i = \langle \boldsymbol{u}, \boldsymbol{D}_i \circ \boldsymbol{c}(0)\rangle$. The second part of Theorem 3.6.3 is an immediate consequence of the remark preceding the statement of the theorem. □



Covariant derivative in terms of parallel translation

Conversely, covariant derivatives can be expressed in terms of parallel translation:

Theorem 3.6.4. *Let $\boldsymbol{c} : [0, a] \to M$ be a curve in a manifold $M^n \subset \mathbb{R}^{n+k}$, and $\boldsymbol{X}$ a vector field in M along $\boldsymbol{c}$. For $b \in (0, a)$, let $\boldsymbol{X}^b$ denote the parallel vector field along $\boldsymbol{c}$ that equals $\boldsymbol{X}(b)$ at b. Then*

$$(\nabla_{\boldsymbol{D}}\boldsymbol{X})(b) = \lim_{t \to b} \frac{\boldsymbol{X}^t(b) - \boldsymbol{X}(b)}{t - b}.$$

Proof. By Theorem 3.6.3, there exist parallel vector fields $\boldsymbol{Y}_i$ in M along $\boldsymbol{c}$ such that $\boldsymbol{Y}_1(t), \dots, \boldsymbol{Y}_n(t)$ form a basis of $M_{\boldsymbol{c}(t)}$ for each $t \in [0, a]$. Thus, $\boldsymbol{X} = \sum_i f_i \boldsymbol{Y}_i$ for smooth functions $f_i : [0, a] \to \mathbb{R}$, and the left side of the above identity can be written as

$$(\nabla_{\boldsymbol{D}}\boldsymbol{X})(b) = \sum_i f_i'(b) \boldsymbol{Y}_i(b).$$

By definition of $\boldsymbol{X}^t$, $\boldsymbol{X}^t(b) = \sum_i f_i(t)\boldsymbol{Y}_i(b)$, so that the right side reads

$$\lim_{t\to b} \frac{\boldsymbol{X}^t(b) - \boldsymbol{X}(b)}{t-b} = \sum_i \lim_{t\to b} \frac{f_i(t) - f_i(b)}{t-b} \boldsymbol{Y}_i(b).$$

Comparing both expressions now implies the claim. □

The above theorem roughly says that in order to compute the covariant derivative of $\boldsymbol{X}$ at b, one parallel translates for each t close to b the vector $\boldsymbol{X}(t)$ to the tangent space of M at $\boldsymbol{c}(b)$. This yields a curve in the vector space $M_{\boldsymbol{c}(b)}$, and its derivative at b is the desired vector.

Example 3.6.1. Let us explore parallel translation along a parametrized circle of latitude in the sphere $M = S^2(1)$. The circle at height $a \in (-1, 1)$ may be parametrized by $\boldsymbol{c} : \mathbb{R} \to M$, where

$$\boldsymbol{c}(t) = (b\cos t, b\sin t, a), \qquad b := \sqrt{1-a^2}.$$

The vector fields $\boldsymbol{X}$, $\boldsymbol{Y}$ along $\boldsymbol{c}$, with

$$\begin{aligned}\boldsymbol{X} &= \frac{1}{|\dot{\boldsymbol{c}}|}\dot{\boldsymbol{c}} = -(\sin)\boldsymbol{D}_1 \circ \boldsymbol{c} + (\cos)\boldsymbol{D}_2 \circ \boldsymbol{c},\\ \boldsymbol{Y} &= -a(\cos)\boldsymbol{D}_1 \circ \boldsymbol{c} - a(\sin)\boldsymbol{D}_2 \circ \boldsymbol{c} + b\boldsymbol{D}_3 \circ \boldsymbol{c},\end{aligned}$$

form an orthonormal basis of the tangent space of the sphere at $\boldsymbol{c}(t)$, which, together with the unit normal vector field

$$\boldsymbol{N} = b(\cos)\boldsymbol{D}_1 \circ \boldsymbol{c} + b(\sin)\boldsymbol{D}_2 \circ \boldsymbol{c} + a\boldsymbol{D}_3 \circ \boldsymbol{c}$$

yield an orthonormal basis of the tangent space of $\mathbb{R}^3$ at $\boldsymbol{c}(t)$. Since $\boldsymbol{X}' = -(\cos)\boldsymbol{D}_1 \circ \boldsymbol{c} - (\sin)\boldsymbol{D}_2 \circ \boldsymbol{c}$, and $\langle \boldsymbol{X}', \boldsymbol{N}\rangle = -b$, we deduce

$$\begin{aligned}\nabla_{\boldsymbol{D}}\boldsymbol{X} &= \boldsymbol{X}' - \langle \boldsymbol{X}', \boldsymbol{N}\rangle \boldsymbol{N}\\ &= (-\cos + b^2\cos)\boldsymbol{D}_1 \circ \boldsymbol{c} + (-\sin + b^2\sin)\boldsymbol{D}_2 \circ \boldsymbol{c} + ab\boldsymbol{D}_3 \circ \boldsymbol{c}\\ &= a\boldsymbol{Y},\end{aligned}$$

where the last equality uses the fact that $b^2 - 1 = -a^2$. Similarly,

$$\nabla_{\boldsymbol{D}}\boldsymbol{Y} = -a\boldsymbol{X}.$$

Thus, the vector fields $\boldsymbol{X}$ and $\boldsymbol{Y}$ are parallel only when the circle in question is a great circle (namely, the equator, corresponding to $a = 0$). In this case, parallel translation along $\boldsymbol{c}$ from $\boldsymbol{c}(0)$ to $\boldsymbol{c}(2\pi) = \boldsymbol{c}(0)$ is the identity. In general, any vector field $\boldsymbol{Z}$ in M along $\boldsymbol{c}$ may be expressed as

$$\boldsymbol{Z} = f\boldsymbol{X} + g\boldsymbol{Y}, \qquad f := \langle \boldsymbol{Z}, \boldsymbol{X}\rangle, \quad g := \langle \boldsymbol{Z}, \boldsymbol{Y}\rangle,$$

so that $\boldsymbol{Z}$ is parallel when

$$\mathbf{0} = \nabla_{\boldsymbol{D}}(f\boldsymbol{X} + g\boldsymbol{Y}) = f'\boldsymbol{X} + af\boldsymbol{Y} + g'\boldsymbol{Y} - ag\boldsymbol{X}.$$

This yields the system of linear differential equations

$$f' - ag = 0, \qquad g' + af = 0,$$

which may be written as

$$\begin{bmatrix} f' \\ g' \end{bmatrix} = A \begin{bmatrix} f \\ g \end{bmatrix}, \qquad A = \begin{bmatrix} 0 & a \\ -a & 0 \end{bmatrix}.$$

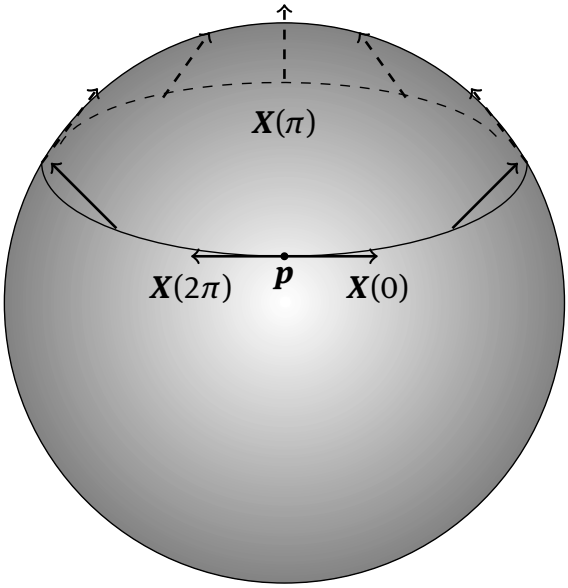
By Example 2.8.1, it admits as general solution

$$\begin{bmatrix} f(t) \\ g(t) \end{bmatrix} = e^{tA} \begin{bmatrix} f(0) \\ g(0) \end{bmatrix} = \begin{bmatrix} \cos(at) & \sin(at) \\ -\sin(at) & \cos(at) \end{bmatrix} \begin{bmatrix} f(0) \\ g(0) \end{bmatrix}.$$

In other words, the parallel field $\boldsymbol{Z}$ along $\boldsymbol{c}$ with "initial condition" $\boldsymbol{Z}(0) = c_1\boldsymbol{X}(0) + c_2\boldsymbol{Y}(0)$ is given by

$$\boldsymbol{Z}(t) = (c_1\cos(at) + c_2\sin(at))\boldsymbol{X}(t) + (c_2\cos(at) - c_1\sin(at))\boldsymbol{Y}(t).$$

Thus, parallel translation along $\boldsymbol{c}$ from $\boldsymbol{c}(0)$ to $\boldsymbol{c}(2\pi) = \boldsymbol{c}(0)$ consists of rotation by angle $2\pi a$. In particular, and as remarked earlier, parallel translation along the equator is the identity. We will see that curvature, a concept to be soon introduced, is responsible for parallel translation along closed curves differing from the identity.



Parallel translation along the circle of latitude through $\boldsymbol{p}$ results in a rotation about the origin by angle π in $S^2_{\boldsymbol{p}}$ when $\boldsymbol{p}$ has height 1/2.

3.7 Geodesics

A fundamental concept in differential geometry is that of curves of shortest length between two points in a manifold. Such curves (or rather certain parametrizations of such curves) are called *geodesics*.

Definition 3.7.1. A curve $\boldsymbol{c} : I \to M$ is said to be a *geodesic* if its acceleration is everywhere orthogonal to M; i.e., if

$$\nabla_{\boldsymbol{D}}\dot{\boldsymbol{c}} = (\dot{\boldsymbol{c}}')^{\top} = \boldsymbol{0}.$$

Thus, from the surface's point of view, geodesics are those curves that have no acceleration. In particular, they have constant speed $|\dot{\boldsymbol{c}}|$:

$$|\dot{\boldsymbol{c}}|^{2\prime} = \langle \dot{\boldsymbol{c}}, \dot{\boldsymbol{c}} \rangle' = 2\langle \nabla_{\boldsymbol{D}}\dot{\boldsymbol{c}}, \dot{\boldsymbol{c}} \rangle = 0.$$

Geodesics are also invariant under linear reparametrizations: if $\boldsymbol{c}$ is a geodesic and $t \mapsto \varphi(t) := at + b$, $a, b \in \mathbb{R}$, then so is $\boldsymbol{c} \circ \varphi$: in fact,

$$\boldsymbol{c}\dot{\circ}\varphi = (\boldsymbol{c} \circ \varphi)_* \boldsymbol{D} = \boldsymbol{c}_* \circ \varphi_* \boldsymbol{D} = \varphi' \boldsymbol{c}_* (\boldsymbol{D} \circ \varphi) = a\dot{\boldsymbol{c}} \circ \varphi,$$

so that by Theorem 3.6.2,

$$\nabla_{\boldsymbol{D}}(\boldsymbol{c}\dot{\circ}\varphi) = a\nabla_{\boldsymbol{D}}(\dot{\boldsymbol{c}} \circ \varphi) = a\nabla_{\varphi_* \boldsymbol{D}}\dot{\boldsymbol{c}} = a^2(\nabla_{\boldsymbol{D}}\dot{\boldsymbol{c}}) \circ \varphi \tag{3.7.1}$$

also vanishes. In particular, any nonconstant geodesic admits a reparametrization by arc length which is again a geodesic. The latter is called a *normal geodesic*. A similar definition can be made for Euclidean space itself: a curve $\boldsymbol{c}$ in $\mathbb{R}^n$ is said to be a geodesic if $\dot{\boldsymbol{c}}' \equiv 0$; equivalently, $\boldsymbol{c}$ has vanishing acceleration. It is easy to see that $\boldsymbol{c}$ is a geodesic in $\mathbb{R}^n$ if and only if $\boldsymbol{c}(t) = \boldsymbol{p} + t\boldsymbol{u}$ for some $\boldsymbol{p}, \boldsymbol{u} \in \mathbb{R}^n$.

Theorem 3.7.1. *For any $\boldsymbol{p} \in M$ and any $\boldsymbol{v} \in M_{\boldsymbol{p}}$, there exists a unique geodesic $\boldsymbol{c}_{\boldsymbol{v}}$ such that $\boldsymbol{c}_{\boldsymbol{v}}(0) = \boldsymbol{p}$ and $\dot{\boldsymbol{c}}_{\boldsymbol{v}}(0) = \boldsymbol{v}$.*

Proof. Any $\boldsymbol{p} \in M$ has a neighborhood U on which there exists an orthonormal basis $\boldsymbol{N}_i$ of vector fields normal to M. The geodesic equation in U then reads

$$\dot{\boldsymbol{c}}' - \sum_i \langle \boldsymbol{N}_i \circ \boldsymbol{c}, \dot{\boldsymbol{c}}' \rangle \boldsymbol{N}_i \circ \boldsymbol{c} = 0.$$

Since

$$\langle \boldsymbol{N}_i \circ \boldsymbol{c}, \dot{\boldsymbol{c}}' \rangle = \langle \boldsymbol{N}_i \circ \boldsymbol{c}, \dot{\boldsymbol{c}} \rangle' - \langle (\boldsymbol{N}_i \circ \boldsymbol{c})', \dot{\boldsymbol{c}} \rangle = -\langle (\boldsymbol{N}_i \circ \boldsymbol{c})', \dot{\boldsymbol{c}} \rangle,$$

this equation can be rewritten

$$\dot{\boldsymbol{c}}' + \sum_i \langle (\boldsymbol{N}_i \circ \boldsymbol{c})', \dot{\boldsymbol{c}} \rangle \boldsymbol{N}_i \circ \boldsymbol{c} = 0. \tag{3.7.2}$$

If $M^n \subset \mathbb{R}^{n+l}$, let $c^k = u^k \circ \boldsymbol{c}$ as usual, and $N_i^k = \langle \boldsymbol{N}_i, \boldsymbol{D}_k \rangle$. (3.7.2) is then equivalent to the second order linear system

$$c^{k\prime\prime} + \sum_{i=1}^{l} \sum_{j=1}^{n+l} (N_i^j \circ \boldsymbol{c})'(N_i^k \circ \boldsymbol{c}) c^{j\prime} = 0, \qquad k = 1, \ldots, n + l. \tag{3.7.3}$$

Standard results from the theory of ordinary differential equations now guarantee the existence of an interval I containing 0, and solutions $c^k : I \to \mathbb{R}$ of (3.7.3) satisfying arbitrary initial conditions $c^k(0) = u^k(\boldsymbol{p})$ and $c^{k\prime}(0) = \boldsymbol{v}(u^k)$. □

It should be noted that the system of equations (3.7.3) is not linear, so that the solutions need not be defined on the whole real line.

Example 3.7.1 (Geodesics on the sphere). Let $M = S^n(r) \subset \mathbb{R}^{n+1}$. Example 3.6.1 implies that for $n = 2$ and $r = 1$ the equatorial curve $t \mapsto (\cos t, \sin t, 0)$ is a geodesic. More generally, consider any $\boldsymbol{p} \in M$, and recall that the tangent space of M at $\boldsymbol{p}$ is just $\mathcal{I}_{\boldsymbol{p}}\boldsymbol{p}^{\perp}$. Let $\boldsymbol{u} \in \mathbb{R}^{n+1}$ be a unit vector orthogonal to $\boldsymbol{p}$, and define $\boldsymbol{c}_{\boldsymbol{u}} : \mathbb{R} \to M$ by

$$\boldsymbol{c}_{\boldsymbol{u}}(t) = (\cos \frac{t}{r})\boldsymbol{p} + (\sin \frac{t}{r})r\boldsymbol{u}.$$

We claim that $\boldsymbol{c}_{\boldsymbol{u}}$ is the geodesic in M with $\boldsymbol{c}_{\boldsymbol{u}}(0) = \boldsymbol{p}$, $\dot{\boldsymbol{c}}_{\boldsymbol{u}}(0) = \mathcal{I}_{\boldsymbol{p}}\boldsymbol{u}$. Indeed, $\boldsymbol{c}_{\boldsymbol{u}}'' = -(1/r^2)\boldsymbol{c}_{\boldsymbol{u}}$, so that $\dot{\boldsymbol{c}}_{\boldsymbol{u}}' = \mathcal{I}_{\boldsymbol{c}_{\boldsymbol{u}}}\boldsymbol{c}_{\boldsymbol{u}}''$ is orthogonal to M. This means that $\boldsymbol{c}_{\boldsymbol{u}}$ is a geodesic in the sphere, and the initial conditions are easily verified.

Implicit in the proof of Theorem 3.7.1 is the fact that a geodesic $\boldsymbol{c}$ "depends smoothly" on initial conditions $\boldsymbol{c}(0)$ and $\dot{\boldsymbol{c}}(0)$. This is also implied by the following fact, which we next establish: the velocity fields of geodesics are integral curves of a certain vector field on TM. To see this, recall that if $M^n \subset \mathbb{R}^{n+l}$, then $TM \subset M \times \mathbb{R}^{n+l} \subset \mathbb{R}^{2(n+l)}$ is a $2n$-dimensional manifold. If $\pi : TM \subset M \times \mathbb{R}^{n+l} \to M$ and $\pi_2 : TM \to \mathbb{R}^{n+l}$ denote the projections onto the two factors, then any point $\boldsymbol{u} \in TM$ can be written

$$\boldsymbol{u} = (\pi(\boldsymbol{u}), \pi_2(\boldsymbol{u})) = \mathcal{I}_{\pi(\boldsymbol{u})}\pi_2(\boldsymbol{u}). \tag{3.7.4}$$

Given a curve $\boldsymbol{c} : I \to M$, its velocity field $\dot{\boldsymbol{c}} : I \to TM$ is a curve in TM with

$$\dot{\boldsymbol{c}}(t) = (\boldsymbol{c}(t), \boldsymbol{c}'(t)) = \mathcal{I}_{\boldsymbol{c}(t)}\boldsymbol{c}'(t) \in M_{\boldsymbol{c}(t)}.$$

Repeating the above process yields a curve $\ddot{\boldsymbol{c}} : I \to T(TM)$ given by

$$\ddot{\boldsymbol{c}}(t) = (\dot{\boldsymbol{c}}(t), \boldsymbol{c}'(t), \boldsymbol{c}''(t)) = \mathcal{I}_{\dot{\boldsymbol{c}}(t)}(\boldsymbol{c}'(t), \boldsymbol{c}''(t)) \in (TM)_{\dot{\boldsymbol{c}}(t)}.$$

As usual, we denote by $\boldsymbol{c}_{\boldsymbol{v}}$ the geodesic with initial tangent vector $\boldsymbol{v}$. Define a map $S : TM \to T(TM)$ by $S(\boldsymbol{v}) = \ddot{\boldsymbol{c}}_{\boldsymbol{v}}(0)$.

Theorem 3.7.2. (1) *The map S is differentiable; i.e., S is a vector field on TM, called the* geodesic spray *of M.*

(2) *$\boldsymbol{c}$ is a geodesic in M if and only if $\dot{\boldsymbol{c}}$ is an integral curve of S.*

Proof. We first show that S is differentiable at any $\boldsymbol{p} \in M$. Denote by U a neighborhood of $\boldsymbol{p}$ on which there exists an orthonormal basis $\boldsymbol{N}_1, \ldots, \boldsymbol{N}_l$ of unit normal fields to M. We claim that the restriction of S to TU is given by

$$\boldsymbol{v} \mapsto \mathcal{I}_{\boldsymbol{v}}\left(\mathcal{I}_{\pi(\boldsymbol{v})}^{-1}\boldsymbol{v}, -\sum_i \langle \nabla_{\boldsymbol{v}}\boldsymbol{N}_i, \boldsymbol{v}\rangle \mathcal{I}_{\pi(\boldsymbol{v})}^{-1}\boldsymbol{N}_i \circ \pi(\boldsymbol{v})\right), \tag{3.7.5}$$

so that S is smooth at $\boldsymbol{p}$. To see this, denote by $\tilde{S}$ the local vector field determined by (3.7.5). If $\boldsymbol{\gamma}$ is an integral curve of $\tilde{S}$, and $\boldsymbol{c} = \pi \circ \boldsymbol{\gamma}$, then

$$\dot{\boldsymbol{\gamma}} = \tilde{S} \circ \boldsymbol{\gamma} = \mathcal{I}_{\boldsymbol{\gamma}}\left(\mathcal{I}_{\boldsymbol{c}}^{-1}\boldsymbol{\gamma}, -\sum_i \langle \nabla_{\boldsymbol{\gamma}}\boldsymbol{N}_i, \boldsymbol{\gamma}\rangle \mathcal{I}_{\boldsymbol{c}}^{-1}\boldsymbol{N}_i \circ \boldsymbol{c}\right). \tag{3.7.6}$$

On the other hand, $\boldsymbol{\gamma} = (\boldsymbol{c}, \pi_2 \circ \boldsymbol{\gamma})$, so that

$$\dot{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}, \boldsymbol{c}', (\pi_2 \circ \boldsymbol{\gamma})') = \mathcal{I}_{\boldsymbol{\gamma}}(\boldsymbol{c}', (\pi_2 \circ \boldsymbol{\gamma})'). \tag{3.7.7}$$

Comparing with the first part of (3.7.6), we deduce that

$$\boldsymbol{c}' = \mathcal{I}_{\boldsymbol{c}}^{-1}\boldsymbol{\gamma}, \text{ or equivalently, } \boldsymbol{\gamma} = \mathcal{I}_{\boldsymbol{c}}\boldsymbol{c}' = \dot{\boldsymbol{c}}. \tag{3.7.8}$$

The latter just says that $\boldsymbol{\gamma}$ is the velocity field of the curve $\boldsymbol{c}$ in M. Comparing the second part of the expressions for $\dot{\boldsymbol{\gamma}}$ in (3.7.6) and (3.7.7) yields

$$(\pi_2 \circ \boldsymbol{\gamma})' = -\sum_i \langle \nabla_{\boldsymbol{\gamma}} \boldsymbol{N}_i, \boldsymbol{\gamma} \rangle \mathcal{I}_{\boldsymbol{c}}^{-1} \boldsymbol{N}_i \circ \boldsymbol{c}.$$

Now, by (3.7.4) and (3.7.8),

$$(\pi_2 \circ \boldsymbol{\gamma})' = (\mathcal{I}_{\boldsymbol{c}}^{-1} \circ \boldsymbol{\gamma})' = (\mathcal{I}_{\boldsymbol{c}}^{-1} \circ \dot{\boldsymbol{c}})' = \boldsymbol{c}'',$$

so that

$$\dot{\boldsymbol{c}}' = \mathcal{I}_{\boldsymbol{c}}\boldsymbol{c}'' = -\sum_i \langle \nabla_{\boldsymbol{\gamma}} \boldsymbol{N}_i, \boldsymbol{\gamma} \rangle \boldsymbol{N}_i \circ \boldsymbol{c} = -\sum_i \langle \nabla_{\dot{\boldsymbol{c}}} \boldsymbol{N}_i, \dot{\boldsymbol{c}} \rangle \boldsymbol{N}_i \circ \boldsymbol{c}.$$

This is just the geodesic equation (3.7.2) for $\boldsymbol{c}$.
Summarizing, integral curves of $\tilde{S}$ are velocity fields of geodesics of M. Thus, for any $\boldsymbol{v} \in TU$, $\tilde{S}(\boldsymbol{v}) = \tilde{S}(\dot{\boldsymbol{c}}_{\boldsymbol{v}}(0)) = \ddot{\boldsymbol{c}}_{\boldsymbol{v}}(0)$, and $\tilde{S}$ is the restriction of S to TU, as claimed. This shows that S is differentiable. The second statement of the theorem was established in the course of proving the first one. □

The concept of geodesic spray provides with yet a third way of describing geodesics, one which is both useful and important: we shall construct a differentiable map exp from an open subset of TM to M that for each $\boldsymbol{v} \in TM$ maps the ray $t \mapsto t\boldsymbol{v}$ in the tangent space $M_{\pi(\boldsymbol{v})}$ to the geodesic $\boldsymbol{c}_{\boldsymbol{v}}$ in M. For this, recall that by Theorem 3.3.2, there exists an open subset W of $\mathbb{R}\times TM$ containing $\{0\}\times TM$, and a differentiable map $\boldsymbol{\Psi} : W \to TM$ (namely, the flow of the geodesic spray vector field) such that for any $\boldsymbol{u} \in TM$, if $\boldsymbol{\Psi}_{\boldsymbol{u}}$ denotes the curve $t \mapsto \boldsymbol{\Psi}(t, \boldsymbol{u})$, then the geodesic $\boldsymbol{c}_{\boldsymbol{u}}$ in direction $\boldsymbol{u}$ equals $\pi \circ \boldsymbol{\Psi}_{\boldsymbol{u}}$. Let

$$\widetilde{TM} = \{\boldsymbol{u} \in TM \mid (1, \boldsymbol{u}) \in W\}.$$

Notice that $\widetilde{TM}$ is open in TM (because W is), and contains a neighborhood of the *zero section* of M (the zero section is defined as the image of the zero vector field $\boldsymbol{Z}$, $\boldsymbol{Z}(\boldsymbol{p}) = \boldsymbol{0} \in M_{\boldsymbol{p}}$ for $\boldsymbol{p} \in M$).

Definition 3.7.2. The *exponential map* of M is the map $\exp : \widetilde{TM} \to M$ given by

$$\exp(\boldsymbol{u}) = (\pi \circ \boldsymbol{\Psi})(1, \boldsymbol{u}) = \boldsymbol{c}_{\boldsymbol{u}}(1).$$

For $\boldsymbol{p} \in M$, let $\tilde{M}_{\boldsymbol{p}} = \widetilde{TM} \cap M_{\boldsymbol{p}}$, and denote by $\exp_{\boldsymbol{p}}$ the restriction of exp to $\tilde{M}_{\boldsymbol{p}}$. Recall also the isomorphism $\mathcal{I}_{\boldsymbol{u}}$ between Euclidean space and its tangent space at any point

$\boldsymbol{u}$. It can be written as $\mathcal{I}_{\boldsymbol{u}}(\boldsymbol{v}) = \dot{\boldsymbol{\gamma}}_{\boldsymbol{v}}(0)$, where $\boldsymbol{\gamma}_{\boldsymbol{v}}(t) = \boldsymbol{u} + t\boldsymbol{v}$. For a manifold M, given $\boldsymbol{p} \in M$, the tangent space $M_{\boldsymbol{p}}$ is both a vector space and a manifold, so $M_{\boldsymbol{p}}$ also has a tangent space at any $\boldsymbol{u} \in M_{\boldsymbol{p}}$. In the same vein, we introduce a canonical isomorphism $\mathcal{I}_{\boldsymbol{u}}$ between $M_{\boldsymbol{p}}$ and its tangent space at $\boldsymbol{u}$ by the formula above. The second statement in the following theorem then says that up to this isomorphism, the derivative $\exp_{\boldsymbol{p}*\boldsymbol{0}}$ of $\exp_{\boldsymbol{p}}$ at the origin $\boldsymbol{Z}(\boldsymbol{p}) = \boldsymbol{0}$ in $M_{\boldsymbol{p}}$ is just the identity.

Theorem 3.7.3. (1) $\boldsymbol{c}_{\boldsymbol{v}}(t) = \exp(t\boldsymbol{v})$ *for any* $\boldsymbol{v} \in TM$;
(2) exp *is differentiable, and for* $\boldsymbol{p} \in M$, $\boldsymbol{v} \in M_{\boldsymbol{p}}$, $\exp_{\boldsymbol{p}*}\, \mathcal{I}_{\boldsymbol{Z}(\boldsymbol{p})}\boldsymbol{v} = \boldsymbol{v}$.

Proof. For the first statement, notice that if t_0 belongs to the domain of $\boldsymbol{c}_{\boldsymbol{v}}$, then the curve $t \mapsto \boldsymbol{\gamma}(t) := \boldsymbol{c}_{\boldsymbol{v}}(t_0 t)$ is a geodesic (being a linear reparametrization of $\boldsymbol{c}_{\boldsymbol{v}}$) with initial velocity $\dot{\boldsymbol{\gamma}}(0) = t_0 \dot{\boldsymbol{c}}_{\boldsymbol{v}}(0) = t_0 \boldsymbol{v}$, and by uniqueness, $\boldsymbol{\gamma} = \boldsymbol{c}_{t_0 \boldsymbol{v}}$. Thus,

$$\boldsymbol{\Psi}_{t_0 \boldsymbol{v}}(1) = \boldsymbol{\Psi}_{\boldsymbol{v}}(t_0), \qquad (t_0, \boldsymbol{v}) \in W,$$

so that

$$\exp(t_0 \boldsymbol{v}) = \pi \circ \boldsymbol{\Psi}_{t_0 \boldsymbol{v}}(1) = \pi \circ \boldsymbol{\Psi}_{\boldsymbol{v}}(t_0) = \boldsymbol{c}_{\boldsymbol{v}}(t_0)$$

for any t_0 in the domain of $\boldsymbol{c}_{\boldsymbol{v}}$, thereby establishing (1). For (2), let

$$\begin{aligned} \iota_1 : \widetilde{TM} &\hookrightarrow W, \\ \boldsymbol{u} &\mapsto (1, \boldsymbol{u}). \end{aligned}$$

Then exp, being a composition $\pi \circ \boldsymbol{\Psi} \circ \iota_1$ of differentiable maps, is also smooth. Finally, if $\boldsymbol{\gamma}_{\boldsymbol{v}}$ denotes the ray $t \mapsto t\boldsymbol{v}$ in the tangent space $M_{\pi(\boldsymbol{v})}$, and $\boldsymbol{0}$ the origin there, then $\dot{\boldsymbol{\gamma}}_{\boldsymbol{v}}(0) = \mathcal{I}_{\boldsymbol{0}}\boldsymbol{v}$, so that

$$\boldsymbol{v} = \dot{\boldsymbol{c}}_{\boldsymbol{v}}(0) = \exp_{\boldsymbol{p}*}(\dot{\boldsymbol{\gamma}}_{\boldsymbol{v}}(0)) = \exp_{\boldsymbol{p}*}\, \mathcal{I}_{\boldsymbol{0}}\boldsymbol{v}. \qquad \square$$

In particular, for every $\boldsymbol{p} \in M$, $\exp_{\boldsymbol{p}}$ has maximal rank at the origin in $M_{\boldsymbol{p}}$. The inverse function theorem then implies:

Corollary 3.7.1. *For any* $\boldsymbol{p} \in M$, *there exists a neighborhood* U *of the origin in* $M_{\boldsymbol{p}}$ *such that* $\exp_{\boldsymbol{p}}$ *maps* U *diffeomorphically onto a neighborhood of* $\boldsymbol{p}$ *in* M.

Remark 3.7.1. There is a stronger version of Corollary 3.7.1 that will be used later on. Consider the map $(\pi, \exp) : \widetilde{TM} \to M \times M$, $(\pi, \exp)(\boldsymbol{v}) = (\pi(\boldsymbol{v}), \exp(\boldsymbol{v}))$, where π is the bundle projection. We have already seen that any local chart $(U, \boldsymbol{x})$ of M around a point $\boldsymbol{p} \in M$ induces a chart $(\pi^{-1}(U), \bar{\boldsymbol{x}})$ of TM around $\pi^{-1}(\boldsymbol{p})$, where

$$\bar{\boldsymbol{x}} = (\boldsymbol{x}, \mathrm{d}x^1, \ldots, \mathrm{d}x^n) \circ \pi,$$

cf. (3.5.3). It also induces one of $M \times M$ around $(\boldsymbol{p}, \boldsymbol{p})$, namely $\boldsymbol{x} \times \boldsymbol{x}$. Now, if $\boldsymbol{v} = (\boldsymbol{p}, \boldsymbol{u}) \in TM \subset M \times \mathbb{R}^{n+k}$, then $\pi(\boldsymbol{p}, \boldsymbol{u}) = \boldsymbol{p}$; i.e., π is projection onto the first factor. This means that the matrix of the derivative of $(\pi, \exp)$ in the corresponding basis of coordinate

vector fields at any point of the zero section has the same top n rows as the $2n \times 2n$ identity matrix I_{2n}. Furthermore, we claim that

$$\frac{\partial}{\partial \bar{\boldsymbol{x}}^{n+j}}(\mathbf{0}_{\boldsymbol{p}}) = (\imath_* \circ \mathcal{I}_{\mathbf{0}_{\boldsymbol{p}}}) \frac{\partial}{\partial x^j}(\boldsymbol{p}),$$

where $\imath : M_{\boldsymbol{p}} \hookrightarrow TM$ is inclusion and $\mathbf{0}_{\boldsymbol{p}}$ denotes the zero vector in $M_{\boldsymbol{p}}$. If we accept this claim for the moment, then Theorem 3.7.3 (2) implies that

$$\exp_* \frac{\partial}{\partial \bar{\boldsymbol{x}}^{n+j}}(\mathbf{0}_{\boldsymbol{p}}) = (\exp_* \circ \imath_* \circ \mathcal{I}_{\mathbf{0}_{\boldsymbol{p}}}) \frac{\partial}{\partial x^j}(\boldsymbol{p}) = \exp_{\boldsymbol{p}_*} \mathcal{I}_{\mathbf{0}_{\boldsymbol{p}}} \frac{\partial}{\partial x^j}(\boldsymbol{p}) = \frac{\partial}{\partial x^j}(\boldsymbol{p}),$$

so that the matrix of $(\pi, \exp)_*$ has the form

$$\begin{bmatrix} I_n & 0 \\ * & I_n \end{bmatrix}.$$

Here all four entries are $n \times n$ matrices, with I_n the identity and 0 the zero matrix.

To verify the claim, consider the ray $\boldsymbol{\gamma}_j : \mathbb{R} \to M_{\boldsymbol{p}}$ with $\boldsymbol{\gamma}_j(t) = t\, \partial/\partial x^j(\boldsymbol{p})$. Then

$$(\imath \circ \boldsymbol{\gamma}_j)_* \boldsymbol{D}(0) = \sum_{i=1}^{2n} D(0)(\bar{x}^i \circ \imath \circ \boldsymbol{\gamma}_j) \frac{\partial}{\partial \bar{\boldsymbol{x}}^i}(\mathbf{0}_{\boldsymbol{p}}).$$

For $i \leq n$, $\bar{x}^i \circ \imath \circ \boldsymbol{\gamma}_j$ is the constant map $x^i(\boldsymbol{p})$, whereas

$$(\bar{x}^{n+i} \circ \imath \circ \boldsymbol{\gamma}_j)(t) = \mathrm{d}x^i\left(t \frac{\partial}{\partial x^j}(\boldsymbol{p})\right) = t$$

when $i = j$ and 0 otherwise. Thus, $(\imath \circ \boldsymbol{\gamma}_j)_* \boldsymbol{D}(0) = (\partial/\partial \bar{x}^{n+j})(\mathbf{0}_{\boldsymbol{p}})$. The claim now follows, since by definition, $(\imath \circ \boldsymbol{\gamma}_j)_* \boldsymbol{D}(0) = (\imath_* \circ \mathcal{I}_{\mathbf{0}_{\boldsymbol{p}}}) \partial/\partial x^j(\boldsymbol{p})$.

In conclusion, the derivative has maximal rank at each point of the zero section $\boldsymbol{Z}$, so that $(\pi, \exp)$ is a diffeomorphism in a neighborhood of each $\boldsymbol{Z}(\boldsymbol{p})$, $\boldsymbol{p} \in M$. Furthermore, the map itself is one-to-one when restricted to the zero section, since $(\pi, \exp)\boldsymbol{Z}(\boldsymbol{p}) = (\boldsymbol{p}, \boldsymbol{p})$. It can be shown that under these conditions, $(\pi, \exp)$ is actually one-to-one on a neighborhood of the zero section. The proof of this last fact involves technical details outside the scope of this text. We nevertheless summarize this for future use: $(\pi, \exp) : \widetilde{TM} \to M \times M$ *maps a neighborhood of the zero section in TM diffeomorphically onto a neighborhood of the diagonal* $\Delta = \{(\boldsymbol{p}, \boldsymbol{p}) \mid \boldsymbol{p} \in M\}$ in $M \times M$.

3.8 The second fundamental tensor

Recall that given vector spaces V and W, a map $M : V^k \to W$ is said to be multilinear if it is linear in each component; i.e. if for any $(k-1)$-tuple $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}) \in V^{k-1}$ and any $i \in \{1, \ldots, k\}$, the map $T \circ \imath_{i, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}} : V \to W$ is linear, where

$$\begin{aligned} \imath_{i, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}} &: V \to V^k, \\ &\boldsymbol{u} \mapsto (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}, \boldsymbol{u}, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_{k-1}). \end{aligned}$$

We adopt the convention that $V^0 = \mathbb{R}$. For $l = 0$ or 1, a multilinear map $T : V^k \to V^l$ is called a *tensor of type* (k, l) on V.

Definition 3.8.1. Let $l = 0$ or 1. A *tensor field of type* (k, l) on a manifold M is a map T such that for each $\boldsymbol{p} \in M$, $T(\boldsymbol{p}) : (M_{\boldsymbol{p}})^k \to (M_{\boldsymbol{p}})^l$ is a tensor of type (k, l) on the tangent space of M at $\boldsymbol{p}$. T is furthermore assumed to be smooth in the sense that for vector fields $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ on M the function (if $l = 0$) or the vector field (if $l = 1$) $T(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)$, given by

$$T(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)(\boldsymbol{p}) = T(\boldsymbol{p})(\boldsymbol{X}_1(\boldsymbol{p}), \ldots, \boldsymbol{X}_k(\boldsymbol{p})), \tag{3.8.1}$$

is smooth in the usual sense.

For example, on any manifold M, the map $\boldsymbol{p} \mapsto g(\boldsymbol{p})$, where $g(\boldsymbol{p})$ denotes the inner product on the tangent space of M at $\boldsymbol{p}$, is a tensor field of type $(2, 0)$, called the *first fundamental tensor field of* M.

We will denote by $\mathfrak{X}M$ the vector space of vector fields on M, and by $(\mathfrak{X}M)^0$ the vector space of all functions $f : M \to \mathbb{R}$. Thus, any tensor field T of type (k, l) defines a multilinear map $T : (\mathfrak{X}M)^k \to (\mathfrak{X}M)^l$ via (3.8.1). This map is actually linear over functions: given $f : M \to \mathbb{R}$,

$$T(\boldsymbol{X}_1, \ldots, f\boldsymbol{X}_i, \ldots, \boldsymbol{X}_k) = fT(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i, \ldots, \boldsymbol{X}_k). \tag{3.8.2}$$

It turns out that this property characterizes tensor fields:

Theorem 3.8.1. *A multilinear map $T : (\mathfrak{X}M)^k \to (\mathfrak{X}M)^l$ is a tensor field if and only if it is linear over functions; i.e., if and only if it satisfies* (3.8.2).

Proof. The condition is necessary by the above remark. Conversely, suppose T is multilinear and linear over functions. We claim that if $\boldsymbol{X}_i(\boldsymbol{p}) = \boldsymbol{Y}_i(\boldsymbol{p})$ for all i, then $T(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)(\boldsymbol{p}) = T(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k)(\boldsymbol{p})$. To see this, assume for simplicity that $k = 1$ and $l = 0$, the general case being analogous. It suffices to establish that if $\boldsymbol{X}(\boldsymbol{p}) = \boldsymbol{0}$, then $T(\boldsymbol{X})(\boldsymbol{p}) = 0$: indeed, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are as above, then $\boldsymbol{X} - \boldsymbol{Y}$ is a vector field that vanishes at $\boldsymbol{p}$, so that $T(\boldsymbol{X})(\boldsymbol{p}) - T(\boldsymbol{Y})(\boldsymbol{p}) = T(\boldsymbol{X} - \boldsymbol{Y})(\boldsymbol{p}) = 0$.
So consider a chart $(U, \boldsymbol{x})$ around $\boldsymbol{p}$, and write $\boldsymbol{X}_{|U} = \sum f^i \, \partial/\partial x^i$ with $f^i(\boldsymbol{p}) = 0$. By Proposition 3.2.2, there exists a function φ on M with support in U, that equals one on a neighborhood of $\boldsymbol{p}$ and takes values in $[0, 1]$. Define vector fields $\boldsymbol{X}_i$ on M by setting them equal to $\varphi \, \partial/\partial x^i$ on U and to 0 outside U. Each $\boldsymbol{X}_i$ is differentiable on M: its restriction to U is given by a differentiable vector field, and if $\boldsymbol{q}$ lies outside U, there is an open neighborhood of $\boldsymbol{q}$ on which $\boldsymbol{X}_i$ is identically zero, implying that $\boldsymbol{X}_i$ is smooth at $\boldsymbol{q}$. Similarly, let g^i be the functions that equal φf^i on U and 0 outside U. Notice that $\sum g^i \boldsymbol{X}_i = \varphi^2 \boldsymbol{X}$, because $\sum_i g^i \boldsymbol{X}_i = \sum_i g^i \varphi \, \partial/\partial x^i = \varphi^2 \sum_i f^i \, \partial/\partial x^i = \varphi^2 \boldsymbol{X}$ on U, and both fields are zero outside U. Thus, $\boldsymbol{X} = \varphi^2 \boldsymbol{X} + (1 - \varphi^2)\boldsymbol{X} = \sum_i g^i \boldsymbol{X}_i + (1 - \varphi^2)\boldsymbol{X}$. Recalling that $g^i(\boldsymbol{p}) = 0$ and $\varphi(\boldsymbol{p}) = 1$, we conclude that

$$(T\boldsymbol{X})(\boldsymbol{p}) = \left(\sum_i g^i(\boldsymbol{p})(T\boldsymbol{X}_i)(\boldsymbol{p}) \right) + (1 - \varphi^2(\boldsymbol{p}))(T\boldsymbol{X})(\boldsymbol{p}) = 0,$$

establishing the claim. We may therefore define for each $\boldsymbol{p} \in M$ a multilinear map $T(\boldsymbol{p})$ by

$$T(\boldsymbol{p})(\boldsymbol{u}_1, \dots, \boldsymbol{u}_k) := T(\boldsymbol{X}_1, \dots, \boldsymbol{X}_k)(\boldsymbol{p})$$

for any vector fields $\boldsymbol{X}_i$ with $\boldsymbol{X}_i(\boldsymbol{p}) = \boldsymbol{u}_i$. The map $\boldsymbol{p} \mapsto T(\boldsymbol{p})$ is clearly smooth. □

Let $M^n \subset \mathbb{R}^{n+k}$ be a submanifold of Euclidean space, and consider the map $S : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(\mathbb{R}^{n+k})_{|M}$ from the Cartesian product of the space of vector fields on M with itself to the space of vector fields on $\mathbb{R}^{n+k}$ restricted to M given by

$$S(\boldsymbol{X}, \boldsymbol{Y}) = (D_{\boldsymbol{X}}\boldsymbol{Y})^{\perp} = D_{\boldsymbol{X}}\boldsymbol{Y} - \nabla_{\boldsymbol{X}}\boldsymbol{Y}. \tag{3.8.3}$$

S is bilinear, and linear over functions in the first component by properties of the covariant derivative. Furthermore, S is symmetric, because

$$S(\boldsymbol{X}, \boldsymbol{Y}) - S(\boldsymbol{Y}, \boldsymbol{X}) = (D_{\boldsymbol{X}}\boldsymbol{Y} - D_{\boldsymbol{Y}}\boldsymbol{X})^{\perp} = [\boldsymbol{X}, \boldsymbol{Y}]^{\perp} = 0,$$

since the bracket of fields tangent to M is a field tangent to M by Theorem 2.9.3. Thus, it is also linear over functions in the second component, and in light of Theorem 3.8.1, the formula

$$S(\boldsymbol{x}, \boldsymbol{y}) = (D_{\boldsymbol{X}}\boldsymbol{Y})^{\perp}(\boldsymbol{p}), \qquad \boldsymbol{x}, \boldsymbol{y} \in M_{\boldsymbol{p}}, \quad \boldsymbol{p} \in M, \tag{3.8.4}$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are any vector fields on M that equal $\boldsymbol{x}$ and $\boldsymbol{y}$ at $\boldsymbol{p}$, defines a tensor field on M, called the *second fundamental tensor of* M. It is worth emphasizing that the fields $\boldsymbol{X}$ and $\boldsymbol{Y}$ only need to be defined in a neighborhood of the point being considered: just as in the proof of Theorem 3.8.1, if U is the domain of a chart around $\boldsymbol{p}$, $\boldsymbol{X}$ is a vector field on U, and φ a nonnegative function with support in U that equals 1 on a neighborhood of $\boldsymbol{p}$, then the field $\tilde{\boldsymbol{X}}$ that equals $\varphi\boldsymbol{X}$ on U and zero outside U is a vector field on M that equals $\boldsymbol{X}$ on a neighborhood of $\boldsymbol{p}$.

Definition 3.8.2. Let $\boldsymbol{u}$ denote a vector orthogonal to $M_{\boldsymbol{p}}$. The *second fundamental tensor with respect to* $\boldsymbol{u}$ is the self-adjoint linear transformation $S_{\boldsymbol{u}} : M_{\boldsymbol{p}} \to M_{\boldsymbol{p}}$ given by

$$S_{\boldsymbol{u}}\boldsymbol{x} = -(D_{\boldsymbol{x}}\boldsymbol{U})^{\top}, \qquad \boldsymbol{x} \in M_{\boldsymbol{p}},$$

where $\boldsymbol{U}$ is any locally defined normal vector field to M with $\boldsymbol{U}(\boldsymbol{p}) = \boldsymbol{u}$ (here, normal to M means that $\boldsymbol{U}(\boldsymbol{q}) \perp M_{\boldsymbol{q}}$ for all $\boldsymbol{q}$ in the domain of $\boldsymbol{U}$).

The above does not depend on the particular extension $\boldsymbol{U}$ because it is tensorial in $\boldsymbol{U}$: If $f : M \to \mathbb{R}$ is a function, then

$$(D_{\boldsymbol{v}}(f\boldsymbol{U}))^{\top} = (\boldsymbol{v}f)\boldsymbol{U}^{\top} + f(\boldsymbol{p})(D_{\boldsymbol{v}}\boldsymbol{U})^{\top} = f(\boldsymbol{p})(D_{\boldsymbol{v}}\boldsymbol{U})^{\top},$$

since $\boldsymbol{U}$ is normal to M, so that $\boldsymbol{U}^{\top}$ is zero. Another way of seeing this is to consider the associated symmetric (as we will see in a moment) bilinear form on $M_{\boldsymbol{p}}$ given by

$$(\boldsymbol{x}, \boldsymbol{y}) \mapsto \langle S_{\boldsymbol{u}}\boldsymbol{x}, \boldsymbol{y}\rangle, \qquad \boldsymbol{x}, \boldsymbol{y} \in M_{\boldsymbol{p}}, \tag{3.8.5}$$

which is called the *second fundamental form with respect to* $\boldsymbol{u}$. If $\boldsymbol{Y}$ is a local extension of $\boldsymbol{y}$, then

$$\langle S_{\boldsymbol{u}}\boldsymbol{x},\boldsymbol{y}\rangle = -\langle D_{\boldsymbol{x}}\boldsymbol{U},\boldsymbol{y}\rangle = -D_{\boldsymbol{x}}\langle \boldsymbol{U},\boldsymbol{Y}\rangle + \langle \boldsymbol{U}(\boldsymbol{p}), D_{\boldsymbol{x}}\boldsymbol{Y}\rangle = \langle \boldsymbol{u}, S(\boldsymbol{x},\boldsymbol{y})\rangle,$$

which, once again, only depends on the value of $\boldsymbol{U}$ at the point $\boldsymbol{p}$. This also shows that the second fundamental form is symmetric as claimed.

Example 3.8.1. Consider the sphere $M = S^n(r) \subset \mathbb{R}^{n+1}$ of radius r. One of the two unit normal fields is $\boldsymbol{N} = (1/r)\boldsymbol{P}$, where $\boldsymbol{P}$ is the position vector field. The second fundamental tensor at $\boldsymbol{p}$ is given by

$$\begin{aligned} S(\boldsymbol{x},\boldsymbol{y}) &= (D_{\boldsymbol{x}}\boldsymbol{Y})^{\perp} = \langle D_{\boldsymbol{x}}\boldsymbol{Y}, \frac{1}{r}\boldsymbol{P}(\boldsymbol{p})\rangle \frac{1}{r}\boldsymbol{P}(\boldsymbol{r}) \\ &= \left(\boldsymbol{x}\langle \boldsymbol{Y},\boldsymbol{P}\rangle - \langle \boldsymbol{Y}(\boldsymbol{p}), D_{\boldsymbol{x}}\boldsymbol{P}\rangle\right)\frac{1}{r^2}\boldsymbol{P}(\boldsymbol{p}) \\ &= -\langle \boldsymbol{y}, D_{\boldsymbol{x}}\boldsymbol{P}\rangle \frac{1}{r^2}\boldsymbol{P}(\boldsymbol{p}), \end{aligned}$$

so that by Examples 2.8.2 (ii),

$$S(\boldsymbol{x},\boldsymbol{y}) = -\frac{1}{r}\langle \boldsymbol{x},\boldsymbol{y}\rangle \boldsymbol{n}, \tag{3.8.6}$$

where $\boldsymbol{n}$ is the outward-pointing unit normal $\boldsymbol{N}(\boldsymbol{p})$. Thus, for any $\boldsymbol{y} \in M_{\boldsymbol{p}}$,

$$\langle S_{\boldsymbol{n}}\boldsymbol{x},\boldsymbol{y}\rangle = \langle \boldsymbol{n}, S(\boldsymbol{x},\boldsymbol{y})\rangle = -\frac{1}{r}\langle \boldsymbol{x},\boldsymbol{y}\rangle,$$

which implies

$$S_{\boldsymbol{n}}\boldsymbol{x} = -\frac{1}{r}\boldsymbol{x}, \qquad \boldsymbol{x} \in M_{\boldsymbol{p}}. \tag{3.8.7}$$

3.9 Curvature

Definition 3.9.1. The *curvature tensor* of M is the tensor field R of type (3,1) given by

$$R(\boldsymbol{v},\boldsymbol{w})\boldsymbol{z} = S_{S(\boldsymbol{w},\boldsymbol{z})}\boldsymbol{v} - S_{S(\boldsymbol{v},\boldsymbol{z})}\boldsymbol{w}, \qquad \boldsymbol{v},\boldsymbol{w},\boldsymbol{z} \in M_{\boldsymbol{p}}, \quad \boldsymbol{p} \in M,$$

where S denotes the second fundamental tensor of M.

Since S is a tensor field, it is clear that R is one also.

Example 3.9.1 (Curvature tensor of a sphere). Let $M = S^n(r)$ denote the sphere of radius r centered at the origin. The second fundamental tensor S of M is given by (3.8.6), and for the outer unit normal vector $\boldsymbol{n}$, $S_{\boldsymbol{n}}\boldsymbol{v} = -\frac{1}{r}\boldsymbol{v}$. Thus,

$$\begin{aligned} R(\boldsymbol{v},\boldsymbol{w})\boldsymbol{z} &= S_{S(\boldsymbol{w},\boldsymbol{z})}\boldsymbol{v} - S_{S(\boldsymbol{v},\boldsymbol{z})}\boldsymbol{w} = S_{-\frac{1}{r}\langle \boldsymbol{w},\boldsymbol{z}\rangle \boldsymbol{n}}\boldsymbol{v} - S_{-\frac{1}{r}\langle \boldsymbol{v},\boldsymbol{z}\rangle \boldsymbol{n}}\boldsymbol{w} \\ &= \frac{1}{r^2}\left(\langle \boldsymbol{w},\boldsymbol{z}\rangle \boldsymbol{v} - \langle \boldsymbol{v},\boldsymbol{z}\rangle \boldsymbol{w}\right). \end{aligned}$$

There is another characterization of the curvature tensor that is often useful. First, though, observe that if $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ are vector fields on an open set $U \subset \mathbb{R}^n$, then

$$(D_{\boldsymbol{X}}D_{\boldsymbol{Y}} - D_{\boldsymbol{Y}}D_{\boldsymbol{X}} - D_{[\boldsymbol{X},\boldsymbol{Y}]})\boldsymbol{Z} = 0. \tag{3.9.1}$$

To see this, write $Z = \sum_i f^i \boldsymbol{D}_i$, and recall that the coordinate vector fields $\boldsymbol{D}_i$ are parallel. Thus,

$$\begin{aligned}(D_{\boldsymbol{X}}D_{\boldsymbol{Y}} - D_{\boldsymbol{Y}}D_{\boldsymbol{X}} - D_{[\boldsymbol{X},\boldsymbol{Y}]})Z &= \sum_i ((\boldsymbol{X}\boldsymbol{Y} - \boldsymbol{Y}\boldsymbol{X} - [\boldsymbol{X},\boldsymbol{Y}])f^i)\boldsymbol{D}_i \\ &= \boldsymbol{0}\end{aligned}$$

by definition of the Lie bracket. In light of the following theorem, (3.9.1) just says that the curvature tensor of Euclidean space is identically zero; i.e., Euclidean space is *flat*.

Theorem 3.9.1. *For vector fields $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ on a manifold M,*

$$R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{Z} = \nabla_{\boldsymbol{X}}\nabla_{\boldsymbol{Y}}\boldsymbol{Z} - \nabla_{\boldsymbol{Y}}\nabla_{\boldsymbol{X}}\boldsymbol{Z} - \nabla_{[\boldsymbol{X},\boldsymbol{Y}]}\boldsymbol{Z}.$$

Proof. For the sake of clarity, we will sometimes write $D^{\perp}_{\boldsymbol{X}}\boldsymbol{Y}$ instead of $(D_{\boldsymbol{X}}\boldsymbol{Y})^{\perp}$, and similarly for $^{\top}$. With this in mind, notice that

$$\nabla_{\boldsymbol{X}}\nabla_{\boldsymbol{Y}}\boldsymbol{Z} = \nabla_{\boldsymbol{X}}(D^{\top}_{\boldsymbol{Y}}\boldsymbol{Z}) = D^{\top}_{\boldsymbol{X}}(D_{\boldsymbol{Y}}\boldsymbol{Z} - S(\boldsymbol{Y},\boldsymbol{Z})) = D^{\top}_{\boldsymbol{X}}D_{\boldsymbol{Y}}\boldsymbol{Z} + S_{S(\boldsymbol{Y},\boldsymbol{Z})}\boldsymbol{X}.$$

Thus, the right side of the identity in the statement equals

$$\begin{aligned}&D^{\top}_{\boldsymbol{X}}D_{\boldsymbol{Y}}\boldsymbol{Z} + S_{S(\boldsymbol{Y},\boldsymbol{Z})}\boldsymbol{X} - D^{\top}_{\boldsymbol{Y}}D_{\boldsymbol{X}}\boldsymbol{Z} - S_{S(\boldsymbol{X},\boldsymbol{Z})}\boldsymbol{Y} - D^{\top}_{[\boldsymbol{X},\boldsymbol{Y}]}\boldsymbol{Z} \\ &= (D_{\boldsymbol{X}}D_{\boldsymbol{Y}} - D_{\boldsymbol{Y}}D_{\boldsymbol{X}} - D_{[\boldsymbol{X},\boldsymbol{Y}]})^{\top}\boldsymbol{Z} + S_{S(\boldsymbol{Y},\boldsymbol{Z})}\boldsymbol{X} - S_{S(\boldsymbol{X},\boldsymbol{Z})}\boldsymbol{Y} \\ &= S_{S(\boldsymbol{Y},\boldsymbol{Z})}\boldsymbol{X} - S_{S(\boldsymbol{X},\boldsymbol{Z})}\boldsymbol{Y} \\ &= R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{Z},\end{aligned}$$

as claimed. □

More generally, we have the following:

Theorem 3.9.2. *Let $\boldsymbol{f} : N \to M$ be differentiable, $\boldsymbol{U}$, $\boldsymbol{V}$ vector fields on N, and $\boldsymbol{X}$ a vector field along $\boldsymbol{f}$. Then*

$$R(\boldsymbol{f}_*\boldsymbol{U},\boldsymbol{f}_*\boldsymbol{V})\boldsymbol{X} = \nabla_{\boldsymbol{U}}\nabla_{\boldsymbol{V}}\boldsymbol{X} - \nabla_{\boldsymbol{V}}\nabla_{\boldsymbol{U}}\boldsymbol{X} - \nabla_{[\boldsymbol{U},\boldsymbol{V}]}\boldsymbol{X}.$$

Proof. The statement being a local one, we may work in the domain U of a chart on M with coordinate vector fields $\boldsymbol{X}_i$. The restriction of any vector field along $\boldsymbol{f}$ to $\boldsymbol{f}^{-1}(U)$ may be written as $\sum_i h^i \boldsymbol{X}_i \circ \boldsymbol{f}$ for some functions $h^i : \boldsymbol{f}^{-1}(U) \to \mathbb{R}$. Since both sides of the identity we are proving are tensorial, we need only consider vector fields of the form $\boldsymbol{X}_i \circ \boldsymbol{f}$; i.e., we may assume that there exist vector fields $\tilde{\boldsymbol{U}}$, $\tilde{\boldsymbol{V}}$, and $\tilde{\boldsymbol{X}}$ on U such that $\boldsymbol{f}_*\boldsymbol{U} = \tilde{\boldsymbol{U}} \circ \boldsymbol{f}$, $\boldsymbol{f}_*\boldsymbol{V} = \tilde{\boldsymbol{V}} \circ \boldsymbol{f}$, and $\boldsymbol{X} = \tilde{\boldsymbol{X}} \circ \boldsymbol{f}$ on $\boldsymbol{f}^{-1}(U)$. Thus, on $\boldsymbol{f}^{-1}(U)$,

$$\begin{aligned}R(\boldsymbol{f}_*\boldsymbol{U},\boldsymbol{f}_*\boldsymbol{V})\boldsymbol{X} &= R(\tilde{\boldsymbol{U}} \circ \boldsymbol{f}, \tilde{\boldsymbol{V}} \circ \boldsymbol{f})\tilde{\boldsymbol{X}} \circ \boldsymbol{f} = [R(\tilde{\boldsymbol{U}},\tilde{\boldsymbol{V}})\tilde{\boldsymbol{X}}] \circ \boldsymbol{f} \\ &= [\nabla_{\tilde{\boldsymbol{U}}}\nabla_{\tilde{\boldsymbol{V}}}\tilde{\boldsymbol{X}} - \nabla_{\tilde{\boldsymbol{V}}}\nabla_{\tilde{\boldsymbol{U}}}\tilde{\boldsymbol{X}} - \nabla_{[\tilde{\boldsymbol{U}},\tilde{\boldsymbol{V}}]}\tilde{\boldsymbol{X}}] \circ \boldsymbol{f}.\end{aligned}$$

By repeated use of Theorem 3.6.2 (6), the first term on the above line may be rewritten as:

$$\begin{aligned}[\nabla_{\tilde{U}}\nabla_{\tilde{V}}\tilde{\boldsymbol{X}}]\circ\boldsymbol{f} &= \nabla_{\tilde{U}\circ\boldsymbol{f}}\nabla_{\tilde{V}}\tilde{\boldsymbol{X}} = \nabla_{\boldsymbol{f}_*\boldsymbol{U}}\nabla_{\tilde{V}}\tilde{\boldsymbol{X}} = \nabla_{\boldsymbol{U}}(\nabla_{\tilde{V}}\tilde{\boldsymbol{X}}\circ\boldsymbol{f})\\ &= \nabla_{\boldsymbol{U}}\nabla_{\boldsymbol{f}_*\boldsymbol{V}}\tilde{\boldsymbol{X}} = \nabla_{\boldsymbol{U}}\nabla_{\boldsymbol{V}}(\tilde{\boldsymbol{X}}\circ\boldsymbol{f})\\ &= \nabla_{\boldsymbol{U}}\nabla_{\boldsymbol{V}}\boldsymbol{X},\end{aligned}$$

and a similar expression holds for the second term. Finally, the last term may, in view of Theorem 2.9.2, be rewritten as follows:

$$\left(\nabla_{[\tilde{U},\tilde{V}]}\tilde{\boldsymbol{X}}\right)\circ\boldsymbol{f} = \nabla_{[\tilde{U},\tilde{V}]\circ\boldsymbol{f}}\tilde{\boldsymbol{X}} = \nabla_{\boldsymbol{f}_*[\boldsymbol{U},\boldsymbol{V}]}\tilde{\boldsymbol{X}} = \nabla_{[\boldsymbol{U},\boldsymbol{V}]}(\tilde{\boldsymbol{X}}\circ\boldsymbol{f}) = \nabla_{[\boldsymbol{U},\boldsymbol{V}]}\boldsymbol{X}.$$

This establishes the result. □

Proposition 3.9.1. *The following identities hold for vector fields $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}$ on M:*
(1) $R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{Z} = -R(\boldsymbol{Y},\boldsymbol{X})\boldsymbol{Z}$;
(2) $R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{Z} + R(\boldsymbol{Y},\boldsymbol{Z})\boldsymbol{X} + R(\boldsymbol{Z},\boldsymbol{X})\boldsymbol{Y} = \boldsymbol{0}$;
(3) $\langle R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{Z},\boldsymbol{U}\rangle = -\langle R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{U},\boldsymbol{Z}\rangle$, *and*
(4) $\langle R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{Z},\boldsymbol{U}\rangle = \langle R(\boldsymbol{Z},\boldsymbol{U})\boldsymbol{X},\boldsymbol{Y}\rangle$.

Proof. (1) is an immediate consequence of the definition. Since R is a tensor field, and is therefore linear over functions, it is enough to prove the identities for coordinate vector fields, or more generally, for vector fields with vanishing Lie brackets. Thus,

$$\nabla_{\boldsymbol{X}}\boldsymbol{Y} = \nabla_{\boldsymbol{Y}}\boldsymbol{X}, \qquad R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{Z} = \nabla_{\boldsymbol{X}}\nabla_{\boldsymbol{Y}}\boldsymbol{Z} - \nabla_{\boldsymbol{Y}}\nabla_{\boldsymbol{X}}\boldsymbol{Z},$$

with similar identities holding for the other terms. (2) now easily follows. For (3), it is enough to show that $\langle R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{V},\boldsymbol{V}\rangle = 0$ for any vector field $\boldsymbol{V}$, since this will imply that

$$\begin{aligned}0 &= \langle R(\boldsymbol{X},\boldsymbol{Y})(\boldsymbol{Z}+\boldsymbol{U}),\boldsymbol{Z}+\boldsymbol{U}\rangle\\ &= \langle R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{Z},\boldsymbol{Z}\rangle + \langle R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{U},\boldsymbol{U}\rangle + \langle R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{Z},\boldsymbol{U}\rangle + \langle R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{U},\boldsymbol{Z}\rangle\\ &= \langle R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{Z},\boldsymbol{U}\rangle + \langle R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{U},\boldsymbol{Z}\rangle.\end{aligned}$$

Now,

$$\begin{aligned}\langle\nabla_{\boldsymbol{X}}\nabla_{\boldsymbol{Y}}\boldsymbol{Z},\boldsymbol{Z}\rangle &= \boldsymbol{X}\langle\nabla_{\boldsymbol{Y}}\boldsymbol{Z},\boldsymbol{Z}\rangle - \langle\nabla_{\boldsymbol{Y}}\boldsymbol{Z},\nabla_{\boldsymbol{X}}\boldsymbol{Z}\rangle\\ &= \frac{1}{2}\boldsymbol{X}\boldsymbol{Y}\langle\boldsymbol{Z},\boldsymbol{Z}\rangle - \langle\nabla_{\boldsymbol{Y}}\boldsymbol{Z},\nabla_{\boldsymbol{X}}\boldsymbol{Z}\rangle,\end{aligned}$$

so that

$$\langle R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{Z},\boldsymbol{Z}\rangle = \frac{1}{2}(\boldsymbol{X}\boldsymbol{Y}-\boldsymbol{Y}\boldsymbol{X})\langle\boldsymbol{Z},\boldsymbol{Z}\rangle = \frac{1}{2}[\boldsymbol{X},\boldsymbol{Y}]\langle\boldsymbol{Z},\boldsymbol{Z}\rangle = \boldsymbol{0}$$

because $[\boldsymbol{X},\boldsymbol{Y}] = \boldsymbol{0}$.
The last identity can be derived from the other three:

$$\begin{aligned}\langle R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{Z},\boldsymbol{U}\rangle &= -\langle R(\boldsymbol{Y},\boldsymbol{X})\boldsymbol{Z},\boldsymbol{U}\rangle \text{ by (1)}\\ &= \langle R(\boldsymbol{X},\boldsymbol{Z})\boldsymbol{Y},\boldsymbol{U}\rangle + \langle R(\boldsymbol{Z},\boldsymbol{Y})\boldsymbol{X},\boldsymbol{U}\rangle \text{ by (2).}\end{aligned} \tag{3.9.2}$$

Similarly,

$$\begin{aligned}\langle R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{Z},\boldsymbol{U}\rangle &= -\langle R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{U},\boldsymbol{Z}\rangle \text{ by (3)}\\ &= \langle R(\boldsymbol{Y},\boldsymbol{U})\boldsymbol{X},\boldsymbol{Z}\rangle + \langle R(\boldsymbol{U},\boldsymbol{X})\boldsymbol{Y},\boldsymbol{Z}\rangle \text{ by (2).}\end{aligned} \tag{3.9.3}$$

Adding (3.9.2) and (3.9.3) implies

$$\begin{aligned}\langle 2R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{Z},\boldsymbol{U}\rangle = \langle R(\boldsymbol{X},\boldsymbol{Z})\boldsymbol{Y},\boldsymbol{U}\rangle + \langle R(\boldsymbol{Z},\boldsymbol{Y})\boldsymbol{X},\boldsymbol{U}\rangle + \langle R(\boldsymbol{Y},\boldsymbol{U})\boldsymbol{X},\boldsymbol{Z}\rangle\\ + \langle R(\boldsymbol{U},\boldsymbol{X})\boldsymbol{Y},\boldsymbol{Z}\rangle.\end{aligned} \tag{3.9.4}$$

Swapping $\boldsymbol{X}$ with $\boldsymbol{Z}$ and $\boldsymbol{Y}$ with $\boldsymbol{U}$ in (3.9.4) yields

$$\begin{aligned}\langle 2R(\boldsymbol{Z},\boldsymbol{U})\boldsymbol{X},\boldsymbol{Y}\rangle = \langle R(\boldsymbol{Z},\boldsymbol{X})\boldsymbol{U},\boldsymbol{Y}\rangle + \langle R(\boldsymbol{X},\boldsymbol{U})\boldsymbol{Z},\boldsymbol{Y}\rangle + \langle R(\boldsymbol{U},\boldsymbol{Y})\boldsymbol{Z},\boldsymbol{X}\rangle\\ + \langle R(\boldsymbol{Y},\boldsymbol{Z})\boldsymbol{U},\boldsymbol{X}\rangle.\end{aligned} \tag{3.9.5}$$

Finally, applying (1) and (3) to each term on the right side of (3.9.5) yields the right side of (3.9.4). This proves (4). □

Let $\boldsymbol{p} \in M$, E a 2-dimensional subspace of $M_{\boldsymbol{p}}$. If $\boldsymbol{v}$ and $\boldsymbol{w}$ form a basis of E, then $|\boldsymbol{v}|^2|\boldsymbol{w}|^2 - \langle \boldsymbol{v},\boldsymbol{w}\rangle^2 > 0$ by the Cauchy-Schwartz inequality and Exercise 1.20. The *sectional curvature* of E is the number

$$K_E = \frac{\langle R(\boldsymbol{v},\boldsymbol{w})\boldsymbol{w},\boldsymbol{v}\rangle}{|\boldsymbol{v}|^2|\boldsymbol{w}|^2 - \langle \boldsymbol{v},\boldsymbol{w}\rangle^2} \tag{3.9.6}$$

To see that this definition does not depend on the particular basis, denote by k the *curvature form* of R; i.e., the map $k : \mathfrak{X}M \times \mathfrak{X}M \to \mathbb{R}$, where $k(\boldsymbol{X},\boldsymbol{Y}) = \langle R(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{Y},\boldsymbol{X}\rangle$. Notice that if R_S denotes the curvature tensor of the unit sphere, then the corresponding curvature form k_S is precisely the denominator of (3.9.6) by Example 3.9.1, and

$$K_E = \frac{k(\boldsymbol{v},\boldsymbol{w})}{k_S(\boldsymbol{v},\boldsymbol{w})}.$$

Suppose $\boldsymbol{u} = a_{11}\boldsymbol{v} + a_{12}\boldsymbol{w}$ and $\boldsymbol{v} = a_{21}\boldsymbol{v} + a_{22}\boldsymbol{w}$ form another basis of E. A straightforward computation shows that

$$k(\boldsymbol{u},\boldsymbol{v}) = (\det A)^2 k(\boldsymbol{v},\boldsymbol{w}),$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

is the transition matrix between the two bases. Since this identity does not depend on the particular curvature tensor, it also holds for k_S, and we conclude that the sectional curvature is indeed well-defined. Furthermore, if $\boldsymbol{v}$ and $\boldsymbol{w}$ are orthonormal, then

$$K_E = \langle R(\boldsymbol{v},\boldsymbol{w})\boldsymbol{w},\boldsymbol{v}\rangle.$$

Thus, for example, the sectional curvature of any plane tangent to a sphere of radius r is $1/r^2$. We therefore say that the sphere of radius r has *constant curvature* $1/r^2$. More generally, a *space of constant curvature* $\kappa \in \mathbb{R}$ is a space where every plane has sectional curvature κ. As remarked earlier, Euclidean space has constant zero curvature.

3.10 Sectional curvature and the length of small circles

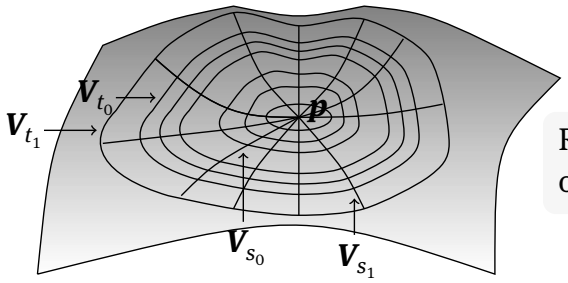
There is also a more geometric interpretation of the sectional curvature K_P of a plane $P \subset M_{\boldsymbol{p}}$. Suppose $\{\boldsymbol{v}, \boldsymbol{w}\}$ is an orthonormal basis of $M_{\boldsymbol{p}}$, and consider the variation $\boldsymbol{V} : [0, a] \times [0, 2\pi] \to M$ by geodesics, where

$$\boldsymbol{V}(t, s) = \exp_p[t(\cos s\boldsymbol{v} + \sin s\boldsymbol{w})].$$

When M is 2-dimensional, then for small t, the image of $\boldsymbol{V}_t$, where $\boldsymbol{V}_t(s) = \boldsymbol{V}(t, s)$, is the set of points at distance t from $\boldsymbol{p}$. By abuse of notation, $\boldsymbol{V}_t$ will be called the *circle of radius* t, and $\boldsymbol{V}_s$, where $\boldsymbol{V}_s(t) = \boldsymbol{V}(t, s)$, a "radial" geodesic. The length of the circle of radius t is

$$L(t) = \int_0^{2\pi} |\dot{\boldsymbol{V}}_t(s)|\, ds.$$

$\dot{\boldsymbol{V}}_t(s)$ is also $\boldsymbol{Y}_s(t)$, where $\boldsymbol{Y}_s$ is the vector field $\boldsymbol{Y}_s(t) = \boldsymbol{V}_* \boldsymbol{D}_2(t, s)$ along the geodesic $\boldsymbol{V}_s$.



Radial geodesics $\boldsymbol{V}_s$ and "circles" $\boldsymbol{V}_t$ of points at constant distance from $\boldsymbol{p}$

In order to estimate $|\boldsymbol{Y}_s|$, we will need the following

Lemma 3.10.1. *Let $\boldsymbol{Y}$ be a vector field along a curve $\boldsymbol{c}$. If $\boldsymbol{E}_i$ is the parallel vector field along $\boldsymbol{c}$ that equals $\boldsymbol{Y}^{(i)}(0)$ when $t = 0$, then*

$$\boldsymbol{Y}(t) = \sum_{i=0}^{k} \frac{t^i}{i!} \boldsymbol{E}_i(t) + \boldsymbol{Z}(t),$$

where $\boldsymbol{Z}$ is a vector field along $\boldsymbol{c}$ satisfying $|\boldsymbol{Z}(t)|/t^k \to 0$ as $t \to 0$.

Proof. Choose a basis $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_n$ of orthonormal parallel fields along $\boldsymbol{c}$, so that $\boldsymbol{Y} = \sum f_l \boldsymbol{F}_l$, where $f_l = \langle \boldsymbol{Y}, \boldsymbol{F}_l \rangle$. By Taylor's theorem,

$$f_l(t) = \sum_{i=1}^{k} \frac{t^i}{i!} f_l^{(i)}(0) + o(t^k),$$

with $o(t^k)$ designating some expression that, when divided by t^k, goes to zero as $t \to 0$. Thus,

$$\boldsymbol{Y}(t) = \sum_{i,l} \left(\frac{t^i}{i!} f_l^{(i)}(0) + o(t^k) \right) \boldsymbol{F}_l(t).$$

The claim now follows, since $\boldsymbol{E}_i = \sum_l f_l^{(i)}(0) \boldsymbol{F}_l$. □

We will use the Lemma with $k = 3$: by definition, $\boldsymbol{Y}_s(0) = \boldsymbol{0}$, and

$$\boldsymbol{Y}_s'(0) = \nabla_{\boldsymbol{D}_1} \boldsymbol{V}_* \boldsymbol{D}_2(0,s) = \nabla_{\boldsymbol{D}_2} \boldsymbol{V}_* \boldsymbol{D}_1(0,s) = -\sin s\, \boldsymbol{v} + \cos s\, \boldsymbol{w},$$

since $\boldsymbol{V}_* \boldsymbol{D}_1(0,s) = \cos s\, \boldsymbol{v} + \sin s\, \boldsymbol{w}$. In order to compute $\boldsymbol{Y}_s'(0)$, notice that by Theorem 3.9.2,

$$\begin{aligned} R(\boldsymbol{V}_* \boldsymbol{D}_2, \boldsymbol{V}_* \boldsymbol{D}_1) \boldsymbol{V}_* \boldsymbol{D}_1 &= \nabla_{\boldsymbol{D}_2} \nabla_{\boldsymbol{D}_1} \boldsymbol{V}_* \boldsymbol{D}_1 - \nabla_{\boldsymbol{D}_1} \nabla_{\boldsymbol{D}_2} \boldsymbol{V}_* \boldsymbol{D}_1 \\ &\quad - \nabla_{[\boldsymbol{D}_2, \boldsymbol{D}_1]} \boldsymbol{V}_* \boldsymbol{D}_1 \\ &= -\nabla_{\boldsymbol{D}_1} \nabla_{\boldsymbol{D}_2} \boldsymbol{V}_* \boldsymbol{D}_1 = -\nabla_{\boldsymbol{D}_1} \nabla_{\boldsymbol{D}_1} \boldsymbol{V}_* \boldsymbol{D}_2. \end{aligned}$$

Thus, $\boldsymbol{Y}_s'' = -R(\boldsymbol{Y}_s, \dot{\boldsymbol{V}}_s)\dot{\boldsymbol{V}}_s$, and since $\boldsymbol{Y}_s(0) = \boldsymbol{0}$, $\boldsymbol{Y}_s''(0) = \boldsymbol{0}$ also. It remains to evaluate $\boldsymbol{Y}^{(3)} = -\big(R(\boldsymbol{Y}_s, \dot{\boldsymbol{V}}_s)\dot{\boldsymbol{V}}_s\big)'$ at 0. To do so, consider a basis $\boldsymbol{E}_i$ of parallel fields along $\boldsymbol{V}_s$ and write $\boldsymbol{Y}_s = \sum f_i \boldsymbol{E}_i$ in terms of this basis. Then

$$\begin{aligned} \boldsymbol{Y}_s^{(3)}(0) &= -\sum \big(f_i R(\boldsymbol{E}_i, \dot{\boldsymbol{V}}_s)\dot{\boldsymbol{V}}_s\big)'(0) \\ &= -\sum_i \Big(f_i(0)\big(R(\boldsymbol{E}_i, \dot{\boldsymbol{V}}_s)\dot{\boldsymbol{V}}_s\big)'(0) + f_i'(0)\big(R(\boldsymbol{E}_i, \dot{\boldsymbol{V}}_s)\dot{\boldsymbol{V}}_s\big)(0)\Big) \\ &= -\sum f_i'(0) R(\boldsymbol{E}_i, \dot{\boldsymbol{V}}_s)\dot{\boldsymbol{V}}_s(0) = -R(\boldsymbol{Y}_s'(0), \dot{\boldsymbol{V}}_s(0))\dot{\boldsymbol{V}}_s(0) \end{aligned}$$

since $f_i(0) = 0$. Lemma 3.10.1 now implies that $\boldsymbol{Y}_s$ may be written as

$$\boldsymbol{Y}_s(t) = t\boldsymbol{E}_1(t) + \frac{t^3}{3!}\boldsymbol{E}_3(t) + \boldsymbol{Z}(t),$$

where $\boldsymbol{E}_1(0) = \boldsymbol{w}_s := -\sin s\, \boldsymbol{v} + \cos s\, \boldsymbol{w}$, $\boldsymbol{E}_3(0) = -R(\boldsymbol{w}_s, \boldsymbol{v}_s)\boldsymbol{v}_s$, and $\boldsymbol{v}_s := \cos s\, \boldsymbol{v} + \sin s\, \boldsymbol{w}$. Furthermore, $|\boldsymbol{Z}(t)|/t^3 \to 0$ as $t \to 0$. It follows that

$$\begin{aligned} |\boldsymbol{Y}_s|^2(t) &= t^2 - \frac{2t^4}{3!}K_P + \left(\frac{t^3}{3!}\right)^2 |R(\boldsymbol{w}_s, \boldsymbol{v}_s)\boldsymbol{v}_s|^2 + |\boldsymbol{Z}|^2(t) + 2t\langle \boldsymbol{E}_1, \boldsymbol{Z}\rangle(t) \\ &\quad + \frac{2t^3}{3!}\langle \boldsymbol{E}_3, \boldsymbol{Z}\rangle(t) \\ &= t^2\left(1 - \frac{2t^2}{3!}K_P - g(t)\right), \end{aligned}$$

where g is a function such that $g(t)/t^2 \to 0$ as $t \to 0$. Since $(1-x)^{1/2} = 1 - (x/2) + o(x)$, we obtain for $t > 0$

$$\begin{aligned} |\boldsymbol{Y}_s(t)| &= t\left(1 - \frac{2t^2}{3!}K_P - g(t)\right)^{1/2} = t\left(1 - \frac{t^2}{3!}K_P + o(t^2)\right) \\ &= t - \frac{t^3}{3!}K_P + o(t^3). \end{aligned} \tag{3.10.1}$$

We are now able to state our main result, which roughly says that M has positive (resp. negative) sectional curvature at $\boldsymbol{p} \in M$ if and only if small circles centered at $\boldsymbol{p}$ have smaller (resp. larger) circumference than circles of the same radius in Euclidean space:

Theorem 3.10.1. *Let M be a manifold, $\boldsymbol{p}$ a point in M, and P a plane in $M_{\boldsymbol{p}}$. For $t > 0$, denote by C_t the "circle" in M obtained by exponentiating the circle of radius t centered at the origin in P. Then the length of C_t is*

$$L(C_t) = 2\pi t\left(1 - \frac{K_P}{3!}t^2 + o(t^2)\right).$$

In particular, the sectional curvature of P is

$$K_P = \lim_{t\to 0}\frac{6}{t^2}\left(1 - \frac{L(C_t)}{2\pi t}\right).$$

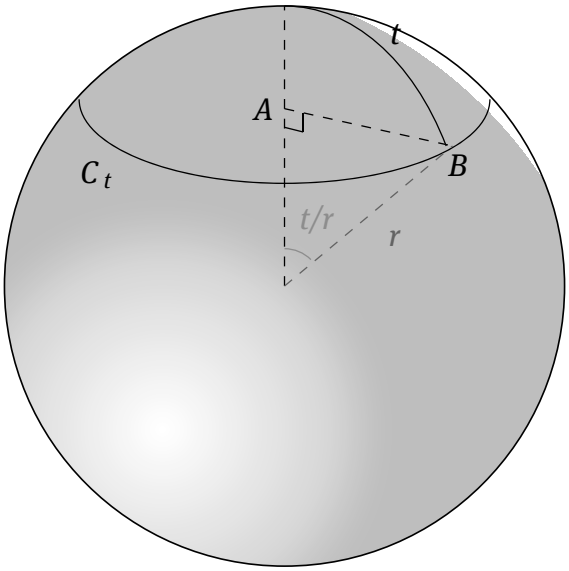
Proof. This is immediate from (3.10.1), since $L(C_t) = \int_0^{2\pi} |\boldsymbol{Y}_s(t)|\, ds$ as remarked earlier. □

Even though the above formula for the sectional curvature is more qualitative than anything, there are some special cases where it can be used to actually evaluate the curvature:

Example 3.10.1. Consider a sphere of radius $r > 0$. A geodesic of length $t < \pi r$ is an arc of a circle with radius r and central angle t/r. Elementary trigonometry implies that C_t has length $2\pi r \sin(t/r)$. Thus, any plane has curvature

$$K = \lim_{t\to 0} 6\frac{t - \sin(t/r)}{t^3} = \frac{1}{r^2}$$

after applying l'Hospital's rule three times.



the circle C_t has radius $AB = r \sin\frac{t}{r}$

3.11 Isometries

Maps that preserve the inner product play a fundamental role in differential geometry:

Definition 3.11.1. A map $\boldsymbol{f} : M \to N$ between manifolds is said to be *isometric* if for all $\boldsymbol{p} \in M$ and $\boldsymbol{v}, \boldsymbol{w} \in M_{\boldsymbol{p}}$, $\langle \boldsymbol{f}_* \boldsymbol{v}, \boldsymbol{f}_* \boldsymbol{w} \rangle = \langle \boldsymbol{v}, \boldsymbol{w} \rangle$. An isometric diffeomorphism is called an *isometry*.

Equivalently, $\boldsymbol{f} : M \to N$ is isometric if $|\boldsymbol{f}_* \boldsymbol{v}| = |\boldsymbol{v}|$ for all $\boldsymbol{v} \in TM$: indeed, if $\boldsymbol{f}$ preserves norms, then it must preserve inner products, for

$$\langle \boldsymbol{f}_* \boldsymbol{u}, \boldsymbol{f}_* \boldsymbol{v} \rangle = \frac{1}{2} \left(|\boldsymbol{f}_* (\boldsymbol{u} + \boldsymbol{v})|^2 - |\boldsymbol{f}_* \boldsymbol{u}|^2 - |\boldsymbol{f}_* \boldsymbol{v}|^2 \right)$$

by linearity of $\boldsymbol{f}_*$. The converse is also clear. In particular, $\boldsymbol{f}_{*\boldsymbol{p}} : M_{\boldsymbol{p}} \to N_{\boldsymbol{f}(\boldsymbol{p})}$ is injective in this case, and $\dim M \leq \dim N$.

If $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ is an isometry, and M is a submanifold of $\mathbb{R}^n$, then the image $N := \boldsymbol{f}(M)$ is also a manifold, and the restriction $\boldsymbol{f} : M \to N$ is an isometry: in fact, any local parametrization $\boldsymbol{h}$ of M induces a local parametrization $\boldsymbol{f} \circ \boldsymbol{h}$ of N, so that N is a submanifold. Since the inner product on the tangent bundle of M is the restriction of that on $\mathbb{R}^n$, the restriction of $\boldsymbol{f}$ to M is clearly an isometry. The isometries of Euclidean space are called *Euclidean motions*, and are easy to describe:

Proposition 3.11.1. *$\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ is a Euclidean motion if and only if it is the composition of an orthogonal transformation A with a translation $\boldsymbol{v} \mapsto \boldsymbol{v} + \boldsymbol{a}$ by some vector $\boldsymbol{a}$; i.e., $\boldsymbol{f}(\boldsymbol{v}) = A\boldsymbol{v} + \boldsymbol{a}$.*

Proof. Clearly, compositions of orthogonal transformations with translations are isometries: If $\boldsymbol{f}(\boldsymbol{v}) = A\boldsymbol{v} + \boldsymbol{a}$, then for $\boldsymbol{u} = \mathcal{I}_{\boldsymbol{p}} \boldsymbol{v} \in \mathbb{R}^n_{\boldsymbol{p}}$,

$$|\boldsymbol{f}_* \boldsymbol{u}| = |D\boldsymbol{f}(\boldsymbol{p})\boldsymbol{v}| = |A\boldsymbol{v}| = |\boldsymbol{v}| = |\boldsymbol{u}|.$$

Conversely, suppose $\boldsymbol{f}$ is a Euclidean motion. Notice first of all that $\boldsymbol{f}$ preserves distances between points: by hypothesis, $|\dot{\boldsymbol{f} \circ \boldsymbol{c}}| = |\boldsymbol{f}_* \circ \dot{\boldsymbol{c}}| = |\dot{\boldsymbol{c}}|$, so that $\ell(\boldsymbol{f} \circ \boldsymbol{c}) = \ell(\boldsymbol{c})$, and $\boldsymbol{f}$ preserves the length of curves. Thus, if $\boldsymbol{c}$ is the line segment from $\boldsymbol{p}$ to $\boldsymbol{q}$, then $\boldsymbol{f} \circ \boldsymbol{c}$ is a curve from $\boldsymbol{f}(\boldsymbol{p})$ to $\boldsymbol{f}(\boldsymbol{q})$ of length $|\boldsymbol{p} - \boldsymbol{q}|$. This implies that $|\boldsymbol{f}(\boldsymbol{p}) - \boldsymbol{f}(\boldsymbol{q})| \leq |\boldsymbol{p} - \boldsymbol{q}|$ for any $\boldsymbol{p}, \boldsymbol{q}$. But $\boldsymbol{f}^{-1}$ is also an isometry, so by the same reasoning,

$$|\boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{p})) - \boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{q}))| \leq |\boldsymbol{f}(\boldsymbol{p}) - \boldsymbol{f}(\boldsymbol{q})| \leq |\boldsymbol{p} - \boldsymbol{q}|,$$

and all terms in the above inequality are the same; i.e., $\boldsymbol{f}$ preserves distances as claimed.
If $\boldsymbol{a} = \boldsymbol{f}(\boldsymbol{0})$, and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^n$ denotes the map given by $\boldsymbol{g}(\boldsymbol{p}) = \boldsymbol{f}(\boldsymbol{p}) - \boldsymbol{a}$, then $D\boldsymbol{g} = D\boldsymbol{f}$, so that $\boldsymbol{g}$ is also an isometry. In particular,

$$|\boldsymbol{g}(\boldsymbol{v})| = |\boldsymbol{g}(\boldsymbol{v}) - \boldsymbol{g}(\boldsymbol{0})| = |\boldsymbol{v} - \boldsymbol{0}| = |\boldsymbol{v}|,$$

and it only remains to show that $\boldsymbol{g}$ is linear. But if $\boldsymbol{g}$ preserves norms, it also preserves inner products, because

$$\begin{aligned}\langle \boldsymbol{g}(\boldsymbol{v}), \boldsymbol{g}(\boldsymbol{w})\rangle &= \frac{1}{2}\left(|\boldsymbol{g}(\boldsymbol{v})|^2 + |\boldsymbol{g}(\boldsymbol{w})|^2 - |\boldsymbol{g}(\boldsymbol{v}) - \boldsymbol{g}(\boldsymbol{w})|^2\right)\\ &= \frac{1}{2}\left(|\boldsymbol{v}|^2 + |\boldsymbol{w}|^2 - |\boldsymbol{v} - \boldsymbol{w}|^2\right)\\ &= \langle \boldsymbol{v}, \boldsymbol{w}\rangle.\end{aligned}$$

It follows that for any $a \in \mathbb{R}$ and $\boldsymbol{z} \in \mathbb{R}^n$,

$$\begin{aligned}\langle \boldsymbol{g}(a\boldsymbol{v} + \boldsymbol{w}) - a\boldsymbol{g}(\boldsymbol{v}) - \boldsymbol{g}(\boldsymbol{w}), \boldsymbol{z}\rangle &= \langle \boldsymbol{g}(a\boldsymbol{v} + \boldsymbol{w}) - a\boldsymbol{g}(\boldsymbol{v}) - \boldsymbol{g}(\boldsymbol{w}), \boldsymbol{g}(\boldsymbol{g}^{-1}(\boldsymbol{z}))\rangle\\ &= \langle \boldsymbol{g}(a\boldsymbol{v} + \boldsymbol{w}), \boldsymbol{g}(\boldsymbol{g}^{-1}(\boldsymbol{z}))\rangle\\ &\quad - a\langle \boldsymbol{g}(\boldsymbol{v}), \boldsymbol{g}(\boldsymbol{g}^{-1}(\boldsymbol{z}))\rangle - \langle \boldsymbol{g}(\boldsymbol{w}), \boldsymbol{g}(\boldsymbol{g}^{-1}(\boldsymbol{z}))\rangle\\ &= \langle a\boldsymbol{v} + \boldsymbol{w}, \boldsymbol{g}^{-1}(\boldsymbol{z})\rangle - a\langle \boldsymbol{v}, \boldsymbol{g}^{-1}(\boldsymbol{z})\rangle\\ &\quad - \langle \boldsymbol{w}, \boldsymbol{g}^{-1}(\boldsymbol{z})\rangle\\ &= 0.\end{aligned}$$

Now take $\boldsymbol{z} = \boldsymbol{g}(a\boldsymbol{v} + \boldsymbol{w}) - a\boldsymbol{g}(\boldsymbol{v}) - \boldsymbol{g}(\boldsymbol{w})$ in the above equation to deduce that $\boldsymbol{g}(a\boldsymbol{v} + \boldsymbol{w}) - a\boldsymbol{g}(\boldsymbol{v}) - \boldsymbol{g}(\boldsymbol{w}) = \boldsymbol{0}$. Thus, $\boldsymbol{g}$ is linear, as claimed. □

We've already remarked that if M^n is a submanifold of $\mathbb{R}^{n+k}$ and $\boldsymbol{f}$ is a Euclidean motion, then the restriction $\boldsymbol{f} : M \to \boldsymbol{f}(M)$ of $\boldsymbol{f}$ to M is an isometry. It is by no means true that every isometry of M with another manifold is the restriction of a Euclidean motion: for example, let $M = (0, 2\pi) \times \{0\} \subset \mathbb{R}^2$. Then M is isometric to a circle of unit radius with one point removed via $\boldsymbol{f}$, where $\boldsymbol{f}(t, 0) = (\cos t, \sin t)$, but $\boldsymbol{f}$ is not a Euclidean motion. We shall, however, see in a later chapter that if $k = 1$ and $\boldsymbol{f}$ also preserves the second fundamental form, then it is the restriction of a Euclidean motion (this is also true when $k > 1$ under additional assumptions). In order to do so, we must first investigate how covariant derivatives behave under isometric maps. This will in fact lead us to an alternative expression for the covariant derivative, one which is of interest in its own right.

Let $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ denote vector fields on M. Since $\boldsymbol{Z}$ is tangent to M, $\langle \nabla_{\boldsymbol{X}}\boldsymbol{Y}, \boldsymbol{Z}\rangle = \langle D_{\boldsymbol{X}}\boldsymbol{Y}, \boldsymbol{Z}\rangle$. Thus, by Theorem 3.6.2,

$$\boldsymbol{X}\langle \boldsymbol{Y}, \boldsymbol{Z}\rangle = \langle D_{\boldsymbol{X}}\boldsymbol{Y}, \boldsymbol{Z}\rangle + \langle \boldsymbol{Y}, D_{\boldsymbol{X}}\boldsymbol{Z}\rangle. \tag{3.11.1}$$

In the same way, but also using Proposition 2.9.1,

$$\begin{aligned}\boldsymbol{Y}\langle \boldsymbol{Z}, \boldsymbol{X}\rangle &= \langle D_{\boldsymbol{Y}}\boldsymbol{Z}, \boldsymbol{X}\rangle + \langle \boldsymbol{Z}, D_{\boldsymbol{Y}}\boldsymbol{X}\rangle\\ &= \langle D_{\boldsymbol{Y}}\boldsymbol{Z}, \boldsymbol{X}\rangle + \langle \boldsymbol{Z}, D_{\boldsymbol{X}}\boldsymbol{Y}\rangle - \langle \boldsymbol{Z}, [\boldsymbol{X}, \boldsymbol{Y}]\rangle,\end{aligned} \tag{3.11.2}$$

and

$$\boldsymbol{Z}\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \langle D_{\boldsymbol{X}}\boldsymbol{Z}, \boldsymbol{Y}\rangle + \langle [\boldsymbol{Z}, \boldsymbol{X}], \boldsymbol{Y}\rangle + \langle \boldsymbol{X}, D_{\boldsymbol{Y}}\boldsymbol{Z}\rangle - \langle \boldsymbol{X}, [\boldsymbol{Y}, \boldsymbol{Z}]\rangle. \tag{3.11.3}$$

Adding the first two equations (3.11.1), (3.11.2), and subtracting the third (3.11.3) yields

$$\begin{aligned}\boldsymbol{X}\langle \boldsymbol{Y},\boldsymbol{Z}\rangle + \boldsymbol{Y}\langle \boldsymbol{Z},\boldsymbol{X}\rangle - \boldsymbol{Z}\langle \boldsymbol{X},\boldsymbol{Y}\rangle = 2\langle D_{\boldsymbol{X}}\boldsymbol{Y},\boldsymbol{Z}\rangle + \langle \boldsymbol{X},[\boldsymbol{Y},\boldsymbol{Z}]\rangle - \langle \boldsymbol{Y},[\boldsymbol{Z},\boldsymbol{X}]\rangle \\ - \langle \boldsymbol{Z},[\boldsymbol{X},\boldsymbol{Y}]\rangle,\end{aligned}$$

so that

$$\begin{aligned}\langle D_{\boldsymbol{X}}\boldsymbol{Y},\boldsymbol{Z}\rangle = \langle \nabla_{\boldsymbol{X}}\boldsymbol{Y},\boldsymbol{Z}\rangle = \frac{1}{2}\Big\{\boldsymbol{X}\langle \boldsymbol{Y},\boldsymbol{Z}\rangle + \boldsymbol{Y}\langle \boldsymbol{Z},\boldsymbol{X}\rangle - \boldsymbol{Z}\langle \boldsymbol{X},\boldsymbol{Y}\rangle \\ + \langle \boldsymbol{Z},[\boldsymbol{X},\boldsymbol{Y}]\rangle + \langle \boldsymbol{Y},[\boldsymbol{Z},\boldsymbol{X}]\rangle - \langle \boldsymbol{X},[\boldsymbol{Y},\boldsymbol{Z}]\rangle\Big\}.\end{aligned} \tag{3.11.4}$$

Notice that (3.11.4) actually characterizes the covariant derivative $\nabla_{\boldsymbol{X}}\boldsymbol{Y}$, since for any $\boldsymbol{p} \in M$ one can find a local orthonormal basis $\boldsymbol{Z}_i$ of vector fields in a neighborhood of $\boldsymbol{p}$ in M, and

$$\nabla_{\boldsymbol{X}(\boldsymbol{p})}\boldsymbol{Y} = \sum_i \langle \nabla_{\boldsymbol{X}(\boldsymbol{p})}\boldsymbol{Y}, \boldsymbol{Z}_i(\boldsymbol{p})\rangle \boldsymbol{Z}_i(\boldsymbol{p}).$$

Now, (3.11.1) actually holds for vector fields along maps. In particular, given a map $\boldsymbol{f} : N \to M$ and vector fields $\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{Z}}$ on N,

$$\tilde{\boldsymbol{X}}\langle \boldsymbol{f}_*\tilde{\boldsymbol{Y}}, \boldsymbol{f}_*\tilde{\boldsymbol{Z}}\rangle = \langle \nabla_{\tilde{\boldsymbol{X}}}\boldsymbol{f}_*\tilde{\boldsymbol{Y}}, \boldsymbol{f}_*\tilde{\boldsymbol{Z}}\rangle + \langle \boldsymbol{f}_*\tilde{\boldsymbol{Y}}, \nabla_{\tilde{\boldsymbol{X}}}\boldsymbol{f}_*\tilde{\boldsymbol{Z}}\rangle.$$

Furthermore, by Theorem 2.9.1,

$$\boldsymbol{f}_*[\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}] = \nabla_{\tilde{\boldsymbol{X}}}\boldsymbol{f}_*\tilde{\boldsymbol{Y}} - \nabla_{\tilde{\boldsymbol{Y}}}\boldsymbol{f}_*\tilde{\boldsymbol{X}}.$$

Thus, the same argument that lead to (3.11.4) yields the following generalization of it:

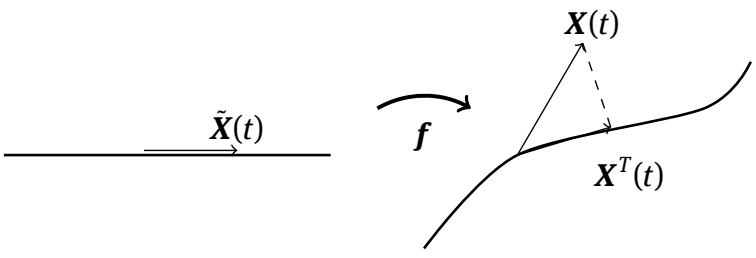
$$\begin{aligned}\langle \nabla_{\tilde{\boldsymbol{X}}}\boldsymbol{f}_*\tilde{\boldsymbol{Y}}, \boldsymbol{f}_*\tilde{\boldsymbol{Z}}\rangle = \frac{1}{2}\Big\{\tilde{\boldsymbol{X}}\langle \boldsymbol{f}_*\tilde{\boldsymbol{Y}}, \boldsymbol{f}_*\tilde{\boldsymbol{Z}}\rangle + \tilde{\boldsymbol{Y}}\langle \boldsymbol{f}_*\tilde{\boldsymbol{Z}}, \boldsymbol{f}_*\tilde{\boldsymbol{X}}\rangle - \tilde{\boldsymbol{Z}}\langle \boldsymbol{f}_*\tilde{\boldsymbol{X}}, \boldsymbol{f}_*\tilde{\boldsymbol{Y}}\rangle \\ + \langle \boldsymbol{f}_*\tilde{\boldsymbol{Z}}, \boldsymbol{f}_*[\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}]\rangle + \langle \boldsymbol{f}_*\tilde{\boldsymbol{Y}}, \boldsymbol{f}_*[\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{X}}]\rangle - \langle \boldsymbol{f}_*\tilde{\boldsymbol{X}}, \boldsymbol{f}_*[\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{Z}}]\rangle\Big\}.\end{aligned} \tag{3.11.5}$$

The next lemma essentially identifies the tangential component of a vector field along an isometric map:

Lemma 3.11.1. *Given an isometric map $\boldsymbol{f} : N \to M$, there exists, for any vector field $\boldsymbol{X}$ along $\boldsymbol{f}$, a unique vector field $\boldsymbol{X}^T$ along $\boldsymbol{f}$ such that $\boldsymbol{X}^T = \boldsymbol{f}_*\tilde{\boldsymbol{X}}$ for some vector field $\tilde{\boldsymbol{X}}$ on N, and*

$$\langle \boldsymbol{X}, \boldsymbol{f}_*\tilde{\boldsymbol{Y}}\rangle = \langle \boldsymbol{X}^T, \boldsymbol{f}_*\tilde{\boldsymbol{Y}}\rangle$$

for any vector field $\tilde{\boldsymbol{Y}}$ on N.



Proof. We shall construct $\boldsymbol{X}^T$ locally. Given $\boldsymbol{p} \in N$, consider an orthonormal basis $\tilde{\boldsymbol{X}}_i$ of vector fields in a neighborhood U of $\boldsymbol{p}$ in N, and set

$$\varphi_i = \langle \boldsymbol{X}, \boldsymbol{f}_* \tilde{\boldsymbol{X}}_i \rangle : U \to \mathbb{R}.$$

The restriction $\tilde{\boldsymbol{Y}}_{|U}$ to U of a vector field $\tilde{\boldsymbol{Y}}$ on N then equals $\sum_i \langle \tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}_i \rangle \tilde{\boldsymbol{X}}_i$, and

$$\boldsymbol{f}_* \tilde{\boldsymbol{Y}}_{|U} = \sum_i \langle \tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}_i \rangle \boldsymbol{f}_* \tilde{\boldsymbol{X}}_i.$$

Thus,

$$\langle \boldsymbol{X}, \boldsymbol{f}_* \tilde{\boldsymbol{Y}} \rangle = \sum_i \langle \tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}_i \rangle \langle \boldsymbol{X}, \boldsymbol{f}_* \tilde{\boldsymbol{X}}_i \rangle = \langle \sum_i \varphi_i \tilde{\boldsymbol{X}}_i, \tilde{\boldsymbol{Y}} \rangle,$$

and we may take $\boldsymbol{X}^T_{|U} = \boldsymbol{f}_* \tilde{\boldsymbol{X}}$, where $\tilde{\boldsymbol{X}} = \sum_i \varphi_i \tilde{\boldsymbol{X}}_i$.
For uniqueness, if $\boldsymbol{Z} = \boldsymbol{f}_* \tilde{\boldsymbol{Z}}$ satisfies $\langle \boldsymbol{Z}, \boldsymbol{f}_* \tilde{\boldsymbol{Y}} \rangle = \langle \boldsymbol{X}^T, \boldsymbol{f}_* \tilde{\boldsymbol{Y}} \rangle$ for every vector field $\tilde{\boldsymbol{Y}}$ on U, then

$$\langle \tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{Y}} \rangle = \langle \boldsymbol{Z}, \boldsymbol{f}_* \tilde{\boldsymbol{Y}} \rangle = \langle \boldsymbol{X}^T, \boldsymbol{f}_* \tilde{\boldsymbol{Y}} \rangle = \langle \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}} \rangle,$$

so that $\langle \tilde{\boldsymbol{Z}} - \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}} \rangle = 0$ for any such $\tilde{\boldsymbol{Y}}$. Now take $\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{Z}} - \tilde{\boldsymbol{X}}$ to conclude that $\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{X}}$, and therefore also $\boldsymbol{Z} = \boldsymbol{X}^T$. □

The next theorem roughly says that covariant derivatives are preserved under isometric maps:

Theorem 3.11.1. *Suppose $\boldsymbol{f} : N \to M$ is isometric, and $\boldsymbol{X}$, $\boldsymbol{Y}$ are vector fields on N. Then*

$$(\nabla_{\boldsymbol{X}} \boldsymbol{f}_* \boldsymbol{Y})^T = \boldsymbol{f}_* \nabla_{\boldsymbol{X}} \boldsymbol{Y}.$$

Proof. Notice that although the same notation is used in the equation above, the covariant derivative on the left is the one on M, and the one on the right is that on N. It must be shown that $\langle \nabla_{\boldsymbol{X}} \boldsymbol{f}_* \boldsymbol{Y}, \boldsymbol{f}_* \boldsymbol{Z} \rangle = \langle \nabla_{\boldsymbol{X}} \boldsymbol{Y}, \boldsymbol{Z} \rangle$ for any vector field $\boldsymbol{Z}$ on N. The first inner product is given by (3.11.5), if one deletes all tildes in that equation. Since $\boldsymbol{f}$ is isometric, all $\boldsymbol{f}_*$ on the right side of that identity may be dropped. What remains is the right side of (3.11.4), thereby establishing the claim. □

Suppose furthermore that N and M have the same dimension. The above theorem then says that

$$\nabla_{\boldsymbol{X}} \boldsymbol{f}_* \boldsymbol{Y} = \boldsymbol{f}_* \nabla_{\boldsymbol{X}} \boldsymbol{Y} \tag{3.11.6}$$

for any vector fields $\boldsymbol{X}$ and $\boldsymbol{Y}$ on N. This implies that if $\boldsymbol{X}$ is a parallel vector field along a curve $\boldsymbol{c}$ in N, then $\boldsymbol{f}_* \boldsymbol{X}$ is a parallel vector field along $\boldsymbol{f} \circ \boldsymbol{c}$. In particular, $\boldsymbol{f}$ maps a geodesic $\boldsymbol{c}$ in N to a geodesic $\boldsymbol{f} \circ \boldsymbol{c}$ in M; equivalently, $\exp_M \circ \boldsymbol{f}_* = \boldsymbol{f} \circ \exp_N$. Moreover, by Theorem 3.9.2, $\boldsymbol{f}$ preserves the curvature tensors in the sense that $\boldsymbol{f}_* R_N(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{Z} = R_M(\boldsymbol{f}_* \boldsymbol{X}, \boldsymbol{f}_* \boldsymbol{Y})\boldsymbol{f}_* \boldsymbol{Z}$ for any vector fields $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ on N. This, of course, also implies that the sectional curvature of a plane $P \subset TN$ equals that of $\boldsymbol{f}_*(P) \subset TM$. Summarizing, we have proved:

Theorem 3.11.2. *Suppose $\boldsymbol{f} : N \to M$ is an isometric map between manifolds of the same dimension. Then*

(1) $\exp_M \circ \boldsymbol{f}_* = \boldsymbol{f} \circ \exp_N$.

(2) $\boldsymbol{f}_* R_N(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{Z} = R_M(\boldsymbol{f}_*\boldsymbol{X}, \boldsymbol{f}_*\boldsymbol{Y})\boldsymbol{f}_*\boldsymbol{Z}$ *for any vector fields* $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ *on* N.

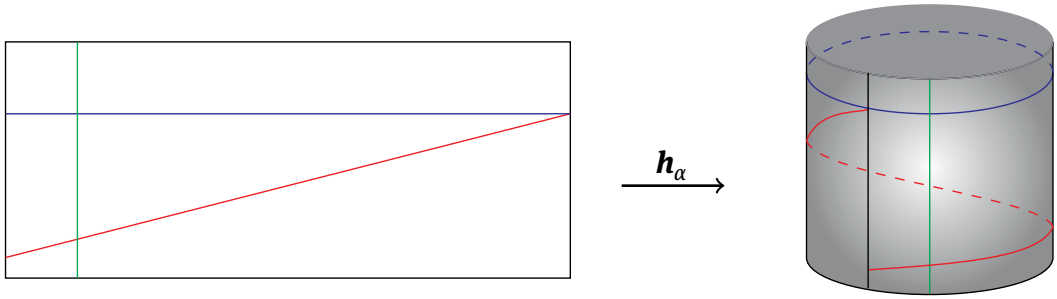
(3) *The sectional curvature of a plane $P \subset TN$ equals that of $\boldsymbol{f}_*(P) \subset TM$.*

Example 3.11.1. Consider the cylinder $M = \{(x, y, z) \mid x^2 + y^2 = r^2\}$ of radius $r > 0$ in $\mathbb{R}^3$. Given any $\alpha \in \mathbb{R}$, the map $\boldsymbol{h}_\alpha : (r\alpha, r(\alpha + 2\pi)) \times \mathbb{R} \to M$, where $\boldsymbol{h}_\alpha(u, v) = (r\cos(u/r), r\sin(u/r), v)$ is a local parametrization of M. Its image is all of M except for a line parallel to the z-axis. M can therefore be covered by two parametrizations corresponding to judiciously chosen values of α. The domain of $\boldsymbol{h}_\alpha$ is an open subset of $\mathbb{R}^2$ and has constant sectional curvature zero. Furthermore, the two vector fields

$$\boldsymbol{h}_\alpha * \boldsymbol{D}_1(u, v) = -\sin\frac{u}{r}\boldsymbol{D}_1 \circ \boldsymbol{h}_\alpha(u, v) + \cos\frac{u}{r}\boldsymbol{D}_2 \circ \boldsymbol{h}_\alpha(u, v),$$

$$\boldsymbol{h}_\alpha * \boldsymbol{D}_2(u, v) = \boldsymbol{D}_3 \circ \boldsymbol{h}_\alpha(u, v),$$

are everywhere orthonormal, so that $\boldsymbol{h}_\alpha$ is isometric. It follows that M is flat. The geodesics of M are the images of the geodesics in the plane and are therefore helices in M.



Isometries can be useful in identifying geodesics of M: Suppose N is a manifold contained in M. N is said to be *totally geodesic in M* if every geodesic of N is also a geodesic of M. For example, $S^n \times \{0\}$ is totally geodesic in $S^{n+1} \subset \mathbb{R}^{n+2}$ (but not in $\mathbb{R}^{n+2}$).

Proposition 3.11.2. *Let M be a manifold and $\boldsymbol{f} : M \to M$ an isometry. If the fixed point set $N = \{\boldsymbol{p} \in M \mid \boldsymbol{f}(\boldsymbol{p}) = \boldsymbol{p}\}$ of $\boldsymbol{f}$ is a manifold, then it is totally geodesic in M.*

Proof. Consider any $\boldsymbol{p} \in N$ and $\boldsymbol{u} \in N_{\boldsymbol{p}}$. Denote by $\boldsymbol{c}_{\boldsymbol{u}}$ the geodesic of M with $\dot{\boldsymbol{c}}_{\boldsymbol{u}}(0) = \boldsymbol{u}$. Since $\boldsymbol{f}$ is an isometry, $\boldsymbol{f} \circ \boldsymbol{c}_{\boldsymbol{u}}$ is also a geodesic, and $\boldsymbol{f} \dot{\circ} \boldsymbol{c}_{\boldsymbol{u}}(0) = \boldsymbol{f}_*\boldsymbol{u} = \boldsymbol{u}$ because $\boldsymbol{u} \in N_{\boldsymbol{p}}$ and $\boldsymbol{f}_*$ is the identity on $N_{\boldsymbol{p}}$. By uniqueness of geodesics, $\boldsymbol{f} \circ \boldsymbol{c}_{\boldsymbol{u}} = \boldsymbol{c}_{\boldsymbol{u}}$, and $\boldsymbol{c}_{\boldsymbol{u}}$ has its image contained in N. By Theorem 6.1.1, it is also a geodesic in N. Since geodesics are uniquely determined by their tangent vector at the origin and $\boldsymbol{u}$ was arbitrary, the result follows. □

The above proposition provides for example another way of describing geodesics on spheres: We may first of all assume the sphere is centered at the origin, since translations are isometries, and thus map geodesics to geodesics. So let $\boldsymbol{p} \in M = S^n(r)$,

$\boldsymbol{u} \in \mathbb{R}^{n+1}$ a unit vector orthogonal to $\boldsymbol{p}$, so that $\mathcal{I}_{\boldsymbol{p}}\boldsymbol{u} \in M_{\boldsymbol{p}}$. We wish to describe the geodesic starting at $\boldsymbol{p}$ with initial tangent vector $\mathcal{I}_{\boldsymbol{p}}\boldsymbol{u}$.

If E is a subspace of $\mathbb{R}^{n+1}$, recall that *reflection* in E is the linear transformation of Euclidean space that maps $\boldsymbol{v} + \boldsymbol{w} \in E \oplus E^{\perp} = \mathbb{R}^{n+1}$ to $\boldsymbol{v} - \boldsymbol{w}$. It is an isometry because $\boldsymbol{v} \perp \boldsymbol{w}$, so that $|\boldsymbol{v} + \boldsymbol{w}| = |\boldsymbol{v} - \boldsymbol{w}|$. Reflection in the plane P spanned by $\boldsymbol{p}$ and $\boldsymbol{u}$ is then an isometry of Euclidean space which maps M to itself, so that its restriction to M is an isometry of M. Furthermore, the fixed point set of this restriction is the great circle $P \cap M$ on the sphere, and this circle is therefore totally geodesic. Since this is true for any point $\boldsymbol{p} \in M$ and unit vector in the tangent space at $\boldsymbol{p}$, all great circles are the images of geodesics. As seen earlier, the geodesic above may be parametrized by $t \mapsto (\cos(t/r))\boldsymbol{p} + (\sin(t/r))(r\boldsymbol{u})$.

3.12 Exercises

3.1. Let $R, h > 0$. An *open cone* with base R and height h is the subset of $\mathbb{R}^3$ obtained by rotating the half-open line segment

$$\{(t, 0, (-h/R)t + h) \mid 0 \leq t < R\}$$

joining $(0, 0, h)$ and $(R, 0, 0)$ about the z-axis. Notice that the line segment contains the first point but not the second.

(a) Show that even though a cone is commonly called a surface, it is not a submanifold of $\mathbb{R}^3$.

(b) If the tip $(0, 0, h)$ of the cone is removed, is the resulting set a manifold?

3.2. Recall that a *hypersurface* is a manifold with dimension one less than that of the ambient Euclidean space.

(a) Let $\boldsymbol{f} : \mathbb{R}^{n+k} \supset U \to \mathbb{R}^k$. Show that if $\boldsymbol{0}$ is a regular value of $\boldsymbol{f}$, then 0 is a regular value of each $u^i \circ \boldsymbol{f}$, $i = 1, \ldots, k$. Conclude that $\boldsymbol{f}^{-1}(\boldsymbol{0})$ is the intersection of k hypersurfaces.

(b) Suppose $f_1, \ldots, f_k : U \to \mathbb{R}$ all have 0 as regular value, and set $M_i = f_i^{-1}(0)$. Show that $M = \cap_{i=1}^{k} M_i$ is an n-dimensional manifold of $\mathbb{R}^{n+k}$ if and only if the vectors $\nabla f_1(\boldsymbol{p}), \ldots, \nabla f_k(\boldsymbol{p})$ are linearly independent for every $\boldsymbol{p} \in M$.

3.3. Prove the "if" part of Corollary 3.2.1: $M \subset \mathbb{R}^{n+k}$ is an n-dimensional manifold if for every $\boldsymbol{p} \in M$, there exists a neighborhood U of $\boldsymbol{p}$ in $\mathbb{R}^{n+k}$, and a map $\boldsymbol{f} : U \to \mathbb{R}^k$ having $\boldsymbol{0}$ as a regular value, such that $U \cap M = \boldsymbol{f}^{-1}(\boldsymbol{0})$.

3.4. Let M^n be a manifold, $(U, \boldsymbol{x})$ a chart of M around some $\boldsymbol{p} \in M$. Show that $\{\mathrm{d}x^i(\boldsymbol{p}) \mid i = 1, \ldots, n\}$ is the basis of $M_{\boldsymbol{p}}^*$ dual to $\{\partial/\partial x^i(\boldsymbol{p})\}$.

3.5. Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(\boldsymbol{a}) = |\boldsymbol{a}|^2$.

(a) Find $\mathrm{d}f$.

(b) Find $\mathrm{d}g$, where $g := f_{|S^{n-1}}$.

3.6. Let $f, g : M \to \mathbb{R}$, $\varphi : \mathbb{R} \to \mathbb{R}$. Prove that
(a) $\mathrm{d}(f + g) = \mathrm{d}f + \mathrm{d}g$.
(b) $\mathrm{d}(fg) = f\,\mathrm{d}g + g\,\mathrm{d}f$.
(c) $\mathrm{d}(\varphi \circ f) = (\varphi' \circ f)\,\mathrm{d}f$.

3.7. If r, θ denote polar coordinates on the plane, express $\mathrm{d}r$ and $\mathrm{d}\theta$ in terms of $\mathrm{d}u^1$ and $\mathrm{d}u^2$.

3.8. Determine $\mathrm{d}f$, if $f : \mathbb{R}^3 \to \mathbb{R}$ is given by:
(a) $f(x, y, z) = \cos e^{x^2 y + z}$.
(b) $f(\boldsymbol{a}) = \langle \boldsymbol{p}, \boldsymbol{a} \rangle$ for some $\boldsymbol{p} \in \mathbb{R}^3$.

3.9. A *differential 1-form* on a manifold M is just a tensor field of type (1,0).
(a) Let α be a differential 1-form on M, and (U, x) a chart of M. Prove that the restriction of α to U can be written

$$\alpha_{|U} = \sum_i \alpha_i \,\mathrm{d}x^i$$

for some functions $\alpha_i : U \to \mathbb{R}$. How is α_i defined?
(b) Use the first fundamental tensor field of M to show that there is a one-to-one correspondence between differential 1-forms and vector fields on M, see also Corollary 1.4.2.

3.10. Let M be a manifold, $f : M \to \mathbb{R}$, and $(U, \boldsymbol{x})$ a chart of M, so that the restriction of $\mathrm{d}f$ to U can be written

$$\mathrm{d}f_{|U} = \sum_i \mathrm{d}f\left(\frac{\partial}{\partial x^i}\right)\frac{\partial}{\partial x^i} = \sum_i \frac{\partial}{\partial x^i}(f)\frac{\partial}{\partial x^i}.$$

Show that $\partial/\partial x^i(f) = D_i(f \circ \boldsymbol{x}^{-1}) \circ \boldsymbol{x}$.

3.11. (a) Use the method of Lagrange multipliers to determine the maximum of $g = \sum_{i=1}^n u^i u^{n+i} : \mathbb{R}^{2n} \to \mathbb{R}$ on the manifold $\boldsymbol{f}^{-1}(\boldsymbol{0})$, where

$$\boldsymbol{f} = \left(\sum_{i=1}^n (u^i)^2 - 1, \sum_{i=1}^n (u^{n+i})^2 - 1\right) : \mathbb{R}^{2n} \to \mathbb{R}^2.$$

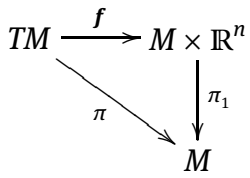
(b) Given any $(x_1, \ldots, x_n, y_1, \ldots, y_n) \in \mathbb{R}^n \times \mathbb{R}^n$ with $\sum x_i^2, \sum y_i^2 \neq 0$, the point

$$\boldsymbol{p} = \left(\frac{x_1}{(\sum x_i^2)^{1/2}}, \ldots, \frac{x_n}{(\sum x_i^2)^{1/2}}, y_1/(\sum y_i^2)^{1/2}, \ldots, y_n/(\sum y_i^2)^{1/2}\right)$$

belongs to $\boldsymbol{f}^{-1}(\boldsymbol{0})$. Use this to give another proof of the Cauchy-Schwarz inequality from Theorem 1.4.1.

3.12. Given a manifold M, the *bundle projection* is the map $\boldsymbol{\pi} : TM \to M$ that maps $\boldsymbol{u} \in TM$ to $\boldsymbol{p}$, if $\boldsymbol{u} \in M_{\boldsymbol{p}}$. Show that the bundle projection is differentiable.

3.13. An n-dimensional manifold M is said to be *parallelizable* if its tangent bundle is diffeomorphic to $M \times \mathbb{R}^n$ by means of a diffeomorphism $\boldsymbol{f}$ which makes the diagram



commute. Here, π is the bundle projection, and π_1 is projection onto the first factor. Show that M is parallelizable if and only if it admits n vector fields which are linearly independent at every point of M. Conclude that every Lie group is parallelizable. In contrast, it can be shown that a 2-dimensional sphere is not parallelizable; in fact, every vector field on it must vanish somewhere.

3.14. Let M, N be manifolds. Given $\boldsymbol{p} \in M$, define $\jmath_{\boldsymbol{p}} : N \to M \times N$ by $\jmath_{\boldsymbol{p}}(\boldsymbol{q}) = (\boldsymbol{p}, \boldsymbol{q})$. Similarly, for $\boldsymbol{q} \in N$, define $\imath_{\boldsymbol{q}} : M \to M \times N$ by $\imath_{\boldsymbol{q}}(\boldsymbol{p}) = (\boldsymbol{p}, \boldsymbol{q})$.

(a) Prove that the map $\imath_{\boldsymbol{q}*} + \jmath_{\boldsymbol{p}*} : M_{\boldsymbol{p}} \times N_{\boldsymbol{q}} \xrightarrow{\cong} (M \times N)_{(\boldsymbol{p},\boldsymbol{q})}$ given by $(\imath_{\boldsymbol{q}*} + \jmath_{\boldsymbol{p}*})(\boldsymbol{u}, \boldsymbol{v}) = \imath_{\boldsymbol{q}*}\boldsymbol{u} + \jmath_{\boldsymbol{p}*}\boldsymbol{v}$, is an isomorphism.

(b) Show that the isomorphism from (a) has as inverse $(\pi_{M*(\boldsymbol{p},\boldsymbol{q})}, \pi_{N*(\boldsymbol{p},\boldsymbol{q})})$, where $\pi_M : M \times N \to M$ and $\pi_N : M \times N \to N$ denote the projections.

3.15. Let $(U, \boldsymbol{x})$ denote a chart on M. Show that if $\boldsymbol{X}$ is a vector field on U, then

$$\boldsymbol{X} = \sum_i \mathrm{d}x^i(\boldsymbol{X}) \frac{\partial}{\partial x^i}.$$

Conclude that $\mathrm{d}x^i$ is a tensor field on U (i.e., that it is differentiable).

3.16. Show that if $(U, \boldsymbol{x})$ and $(V, \boldsymbol{y})$ are two charts of M, then on $U \cap V$,

$$\frac{\partial}{\partial x^i} = \sum_j (D_i(u^j \circ \boldsymbol{y} \circ \boldsymbol{x}^{-1}) \circ \boldsymbol{x}) \frac{\partial}{\partial y^j}.$$

Notice that according to Exercise 3.10, this is equivalent to

$$\frac{\partial}{\partial x^i} = \sum_j \frac{\partial y^j}{\partial x^i} \frac{\partial}{\partial y^j}.$$

3.17. Let $\boldsymbol{X}$ be a vector field on $\mathbb{R}^n$ with flow $\boldsymbol{\Phi}_t$. Define a map $\boldsymbol{F} : U \to \mathbb{R}^n$ on a neighborhood U of the origin by

$$\boldsymbol{F}(a_1, \ldots, a_n) = \boldsymbol{\Phi}_{a_1}(0, a_2, \ldots, a_n).$$

(a) Show that $\boldsymbol{F}_* \boldsymbol{D}_1 = \boldsymbol{X} \circ \boldsymbol{F}$.t'

(b) Show that $\boldsymbol{F}_* \boldsymbol{D}_i(\boldsymbol{0}) = \boldsymbol{D}_i(\boldsymbol{0})$ for $i > 1$.

(c) Let M be a manifold, $\boldsymbol{p} \in M$, and $\boldsymbol{X}$ a vector field on M with $\boldsymbol{X}(\boldsymbol{p}) \neq \boldsymbol{0}$. Prove that there exists a chart $(U, \boldsymbol{x})$ around $\boldsymbol{p}$ such that

$$\boldsymbol{X}_{|U} = \frac{\partial}{\partial x^1}.$$

3.18. Let $\boldsymbol{X}$ be a vector field on a manifold M that vanishes outside a compact set. Show that $\boldsymbol{X}$ is complete.

3.19. Let $a < b$ denote two regular values of $f : \mathbb{R}^n \to \mathbb{R}$. Suppose that $f^{-1}(a)$ and $f^{-1}(b)$ are nonempty, so that they are $(n-1)$-dimensional manifolds. Prove that if $f^{-1}[a,b]$ is compact and contains no critical points of f, then $f^{-1}(a)$ is diffeomorphic to $f^{-1}(b)$. *Hint:* Consider a vector field $\boldsymbol{X}$ that equals $\nabla f/|\nabla f|$ on $f^{-1}[a,b]$ and vanishes outside a compact set. $\boldsymbol{X}$ is then complete by Exercise 3.18, so that $\boldsymbol{\Phi}_{b-a}$ is defined, where $\boldsymbol{\Phi}_t$ is the flow of $\boldsymbol{X}$.

3.20. Recall that $M_{m,n} \cong \mathbb{R}^{mn}$ denotes the space of all $m \times n$ matrices. The goal of this exercise is to show that the subset $M_{m,n}(k)$ consisting of all matrices of rank k is a submanifold of $\mathbb{R}^{mn}$ with dimension $k(m+n-k)$.

(a) Show that if $M \in M_{m,n}$ has rank at least k, then there exist elementary matrices $P \in M_{m,m}$ and $Q \in M_{n,n}$ such that

$$PMQ = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where A is an invertible $k \times k$ matrix, $B \in M_{k,n-k}$, $C \in M_{m-k,k}$, and $D \in M_{m-k,n-k}$.

(b) Let A be a nonsingular $k \times k$ matrix. Prove that

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \in M_{m,n}$$

has rank k if and only if $D = CA^{-1}B$. *Hint:* the matrix

$$\begin{bmatrix} A & B \\ 0 & -CA^{-1}B + D \end{bmatrix} = \begin{bmatrix} I_k & 0 \\ -CA^{-1} & I_{m-k} \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

has the same rank as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}.$$

(c) Given $M_0 \in M_{m,n}(k)$, choose elementary matrices P and Q such that

$$PM_0Q = \begin{bmatrix} A_0 & B_0 \\ C_0 & D_0 \end{bmatrix}$$

with A_0 invertible. There exists an open neighborhood V of A_0 in $M_{k,k}$ such that every A in V is invertible. Denote by U the open neighborhood of M_0 in $M_{m,n}(k)$ consisting of all matrices of the form

$$M = P^{-1} \begin{bmatrix} A & B \\ C & D \end{bmatrix} Q^{-1},$$

with $A \in V$, $B \in M_{k,n-k}$, $C \in M_{m-k,k}$, and $D \in M_{m-k,n-k}$.

Identify $R^{k(m+n-k)} = R^{k^2+k(n-k)+k(m-k)}$ with the subspace of $M_{m,n}$ that consists of all matrices of the form

$$\begin{bmatrix} A & B \\ C & 0 \end{bmatrix}, \qquad A \in M_{k,k}, \quad B \in M_{k,n-k}, \quad C \in M_{m-k,k}.$$

If

$$\begin{aligned} \boldsymbol{\pi} : R^{mn} &\longrightarrow \mathbb{R}^{k(m+n-k)} \\ \begin{bmatrix} A & B \\ C & D \end{bmatrix} &\longmapsto \begin{bmatrix} A & B \\ C & 0 \end{bmatrix} \end{aligned}$$

denotes projection, define

$$\begin{aligned} \boldsymbol{x} : U &\longrightarrow \mathbb{R}^{k(m+n-k)} \\ M &\longmapsto \pi(PMQ). \end{aligned} \tag{3.12.1}$$

Show that the collection of all $(U, \boldsymbol{x})$ in (3.12.1) defines an atlas on $M_{m,n}(k)$.

3.21. Let $m : G \times G \to G$ denote the multiplication map on a Lie group G.

(a) Show that for $(\boldsymbol{v}, \mathbf{0}) \in G_e \times G_e$, $m_*(\boldsymbol{v}, \mathbf{0}) = \boldsymbol{v}$.

(b) Suppose $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are two curves in G that pass through e at time zero. Show that if $\boldsymbol{c}(t) = \boldsymbol{c}_1(t) \cdot \boldsymbol{c}_2(t)$, then

$$\dot{\boldsymbol{c}}(0) = \dot{\boldsymbol{c}}_1(0) + \dot{\boldsymbol{c}}_2(0).$$

3.22. Suppose G is a Lie group, $\boldsymbol{X}, \boldsymbol{Y} \in \mathfrak{g}$ with flows $\boldsymbol{\Phi}_t$ and $\boldsymbol{\Psi}_t$ respectively.

(a) Prove that $(\boldsymbol{\Phi}_t \circ \boldsymbol{\Psi}_t)(e) = \boldsymbol{\Psi}_t(e) \cdot \boldsymbol{\Phi}_t(e)$.

(b) Let $\boldsymbol{c}_{\boldsymbol{X}}$ denote the integral curve of $\boldsymbol{X} \in \mathfrak{g}$ that passes through e at time 0. Show that if $[\boldsymbol{X}, \boldsymbol{Y}] = \mathbf{0}$ for $\boldsymbol{X}, \boldsymbol{Y} \in \mathfrak{g}$, then $\boldsymbol{c}_{\boldsymbol{X}}(t) \cdot \boldsymbol{c}_{\boldsymbol{Y}}(t) = \boldsymbol{c}_{\boldsymbol{Y}}(t) \cdot \boldsymbol{c}_{\boldsymbol{X}}(t)$ for all t.

(c) Suppose $\boldsymbol{X}, \boldsymbol{Y} \in \mathfrak{g}$ have vanishing bracket. Use Exercise 3.21 to prove that the curve $\boldsymbol{c}_{\boldsymbol{X}} \cdot \boldsymbol{c}_{\boldsymbol{Y}} : \mathbb{R} \to G$ is a homomorphism with velocity vector $\boldsymbol{X}(e) + \boldsymbol{Y}(e)$ at zero. Conclude that $\exp(\boldsymbol{X} + \boldsymbol{Y}) = \exp(\boldsymbol{X}) \exp(\boldsymbol{Y})$.

3.23. If g is an element of a Lie group G, *conjugation by* g is the diffeomorphism $\tau_g : G \to G$ given by $\tau_g a = gag^{-1}$, $a \in G$. Its derivative at the identity is therefore an isomorphism $\tau_{g*e} : G_e \to G_e$. In analogy with $GL(n)$, the space of all isomorphisms of $G_e \cong \mathfrak{g}$ with itself is denoted $GL(\mathfrak{g})$, and is a Lie group under composition. The map

$$\begin{aligned} \mathrm{Ad} : G &\longrightarrow GL(\mathfrak{g}), \\ g &\longmapsto \mathrm{Ad}_g = \tau_{g*e} \end{aligned}$$

is called the *adjoint representation* of G.

(a) Prove that Ad is a Lie group homomorphism.

(b) Suppose $G = GL(n)$, and identify $\mathfrak{g}$ with M_n. Show that for $A \in G$ and $M \in \mathfrak{g}$, $\mathrm{Ad}_A M = AMA^{-1}$. In particular, Ad is differentiable. It can be shown that the adjoint representation in any Lie group is differentiable.

(c) The derivative of Ad at the identity is denoted $\mathrm{ad} : \mathfrak{g} \to M_n(\mathfrak{g})$. Prove that for $\boldsymbol{X}, \boldsymbol{Y} \in \mathfrak{g}$, $\mathrm{ad}_{\boldsymbol{X}}(\boldsymbol{Y}) = [\boldsymbol{X}, \boldsymbol{Y}]$ when $G = GL(n)$. It can be shown that this identity is valid in any Lie group.

3.24. Prove that on a compact manifold, the exponential map at any point is defined on the whole tangent space; equivalently, any geodesic has all of $\mathbb{R}$ as its domain.

3.25. A *geodesic vector field* on a manifold M is a vector field $\boldsymbol{X}$ that satisfies $\nabla_{\boldsymbol{X}}\boldsymbol{X} \equiv \boldsymbol{0}$.

(a) Prove that $\boldsymbol{X}$ is geodesic if and only if its integral curves are geodesics in M.

(b) Let $\boldsymbol{X}$ be a vector field on an open set $U \subset \mathbb{R}^n$ represented by $\boldsymbol{f} : U \to \mathbb{R}^n$, so that $\boldsymbol{X}(\boldsymbol{p}) = \mathcal{I}_{\boldsymbol{p}}\boldsymbol{f}(\boldsymbol{p})$. Show that $\boldsymbol{X}$ is geodesic if and only if

$$\langle \nabla f^i, \boldsymbol{f} \rangle \equiv \boldsymbol{0}, \qquad 1 \le i \le n.$$

In particular, any parallel vector field on U is geodesic. Are these the only ones? *Hint:* Consider $\boldsymbol{P}/|\boldsymbol{P}|$ on $\mathbb{R}^n \setminus \{\boldsymbol{0}\}$, where $\boldsymbol{P}$ is the position vector field.

3.26. A vector field $\boldsymbol{X}$ on a manifold M is called a *Killing field* if its flow $\{\boldsymbol{\Phi}_t\}$ consists of local isometries of M.

(a) Show that $\boldsymbol{X}$ is Killing if and only if for any vector field $\boldsymbol{Y}$ that is $\boldsymbol{\Phi}_t$-related to itself for all t, $\langle \boldsymbol{Y}, \boldsymbol{Y} \rangle$ is constant.

(b) Show that in (a) the condition of being $\boldsymbol{\Phi}_t$-related to itself is equivalent to $[\boldsymbol{X}, \boldsymbol{Y}] \equiv \boldsymbol{0}$.

(c) Show that $\boldsymbol{X}$ is Killing if and only $\langle D_{\boldsymbol{u}}\boldsymbol{X}, \boldsymbol{u} \rangle = 0$ for all $\boldsymbol{u} \in TM$; equivalently, the operator $\boldsymbol{u} \mapsto D_{\boldsymbol{u}}\boldsymbol{X}$ on $M_{\boldsymbol{p}}$ is skew-adjoint for all $\boldsymbol{p} \in M$.

3.27. Let M^2 be a 2-dimensional submanifold of $\mathbb{R}^3$, $\boldsymbol{n}$ a unit normal vector field to M in a neighborhood of $\boldsymbol{p} \in M$, and S the second fundamental tensor field of M with respect to $\boldsymbol{n}$.

(a) Show that $R(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{z} = \langle S\boldsymbol{y}, \boldsymbol{z} \rangle S\boldsymbol{x} - \langle S\boldsymbol{x}, \boldsymbol{z} \rangle S\boldsymbol{y}$ for $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in M_{\boldsymbol{p}}$.

(b) Prove that if $\boldsymbol{x}$ and $\boldsymbol{y}$ form an orthonormal basis of $M_{\boldsymbol{p}}$, then the sectional curvature $K(\boldsymbol{p})$ of M at $\boldsymbol{p}$ is given by

$$K(\boldsymbol{p}) = \langle S\boldsymbol{x}, \boldsymbol{x} \rangle \langle S\boldsymbol{y}, \boldsymbol{y} \rangle - \langle S\boldsymbol{x}, \boldsymbol{y} \rangle^2.$$

(c) Let $(U, \boldsymbol{x})$ be a chart of M around $\boldsymbol{p}$, and consider the six functions on U given by

$$E = \left\langle \frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^1} \right\rangle \qquad F = \left\langle \frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^2} \right\rangle \qquad G = \left\langle \frac{\partial}{\partial x^2}, \frac{\partial}{\partial x^2} \right\rangle$$
$$l = \left\langle S\frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^1} \right\rangle \qquad n = \left\langle S\frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^2} \right\rangle \qquad m = \left\langle S\frac{\partial}{\partial x^2}, \frac{\partial}{\partial x^2} \right\rangle$$

Prove that the sectional curvature on U is given by the function $K : U \to \mathbb{R}$, where

$$K = \frac{lm - n^2}{EG - F^2}.$$

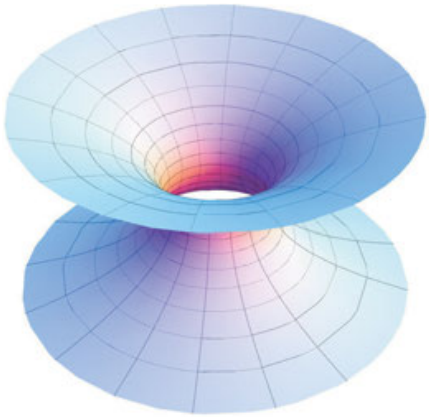


Fig. 3.5: A catenoid

3.28. Let M^2 be a 2-dimensional submanifold of $\mathbb{R}^3$, $\boldsymbol{n}$ and S as in Exercise 3.27. The *mean curvature* of M at a point $\boldsymbol{p} \in M$ is the trace of $S(\boldsymbol{p})$. A surface is said to be *minimal* if its mean curvature is zero everywhere.
Let $a > 0$. The image of the map $\boldsymbol{h} : [0, 2\pi] \times \mathbb{R} \to \mathbb{R}^3$, where

$$\boldsymbol{h}(u, v) = \left(a \cosh \frac{v}{a} \cos u, a \cosh \frac{v}{a} \sin u, v\right),$$

is called a *catenoid*. A catenoid can be simulated by dipping two circles in a soapy solution and pulling them slowly apart. Compute the curvature of a catenoid and show that it is a minimal surface.

3.29. Since S^1 is a submanifold of $\mathbb{R}^2$, $S^1 \times S^1$ is in a natural way a submanifold of $\mathbb{R}^4 = \mathbb{R}^2 \times \mathbb{R}^2$, called a *flat torus*, and denoted T^2.
(a) Show that T^2 is indeed flat; i.e., its curvature is identically zero.
(b) Let $0 < r < R$. Prove that the map $\boldsymbol{h} : [0, 2\pi] \times [0, 2\pi] \to \mathbb{R}^3$, where

$$\boldsymbol{h}(u, v) = \left((R + r \cos u) \cos v, (R + r) \cos u) \sin v, r \sin u\right),$$

is an immersion, and that its image M is an (imbedded) 2-dimensional submanifold of $\mathbb{R}^3$. M is also called a torus.
(c) Show that M and T^2 are diffeomorphic but not isometric.

3.30. Given $a, b, c > 0$, consider the ellipsoid

$$M^2 = \left\{(x, y, z) \in \mathbb{R}^3 \mid \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1\right\}$$

in $\mathbb{R}^3$.
(a) Show that

$$\boldsymbol{N} = \frac{u^1}{a^2}\boldsymbol{D}_1 + \frac{u^2}{b^2}\boldsymbol{D}_2 + \frac{u^3}{c^2}\boldsymbol{D}_2$$

is a vector field normal to M.

(b) Let $\boldsymbol{p} = (x, y, z) \in M$, and denote by s the second fundamental form of M at $\boldsymbol{p}$ with respect to $\boldsymbol{N}(\boldsymbol{p})/|\boldsymbol{N}(\boldsymbol{p})|$. Prove that

$$s(\boldsymbol{x}, \boldsymbol{y}) = \left(\frac{x^2}{a^4} + \frac{y^2}{b^4} + \frac{z^2}{c^4} \right)^{-\frac{1}{2}} \left(\frac{x_1 y_1}{a^2} + \frac{x_2 y_2}{b^2} + \frac{x_3 y_3}{c^2} \right),$$

for $\boldsymbol{x} = \sum_{i=1}^{3} x_i \boldsymbol{D}_i(\boldsymbol{p})$, $\boldsymbol{y} = \sum_{i=1}^{3} y_i \boldsymbol{D}_i(\boldsymbol{p})$.

(c) Show that the sectional curvature of M at $\boldsymbol{p} = (x, y, z)$ equals

$$K = \left[abc \left(\frac{x^2}{a^4} + \frac{y^2}{b^4} + \frac{z^2}{c^4} \right) \right]^{-2}.$$

3.31. Given $a, b, c > 0$, consider the hyperboloid

$$M^2 = \left\{ (x, y, z) \in \mathbb{R}^3 \mid \frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1 \right\}$$

in $\mathbb{R}^3$.

(a) Show that

$$\boldsymbol{N} = \frac{u^1}{a^2} \boldsymbol{D}_1 + \frac{u^2}{b^2} \boldsymbol{D}_2 - \frac{u^3}{c^2} \boldsymbol{D}_2$$

is a vector field normal to M.

(b) Let $\boldsymbol{p} = (x, y, z) \in M$, and denote by s the second fundamental form of M at $\boldsymbol{p}$ with respect to $\boldsymbol{N}(\boldsymbol{p})/|\boldsymbol{N}(\boldsymbol{p})|$. Prove that

$$s(\boldsymbol{x}, \boldsymbol{y}) = \left(\frac{x^2}{a^4} + \frac{y^2}{b^4} + \frac{z^2}{c^4} \right)^{-\frac{1}{2}} \left(\frac{x_1 y_1}{a^2} + \frac{x_2 y_2}{b^2} - \frac{x_3 y_3}{c^2} \right),$$

for $\boldsymbol{x} = \sum_{i=1}^{3} x_i \boldsymbol{D}_i(\boldsymbol{p})$, $\boldsymbol{y} = \sum_{i=1}^{3} y_i \boldsymbol{D}_i(\boldsymbol{p})$.

(c) Show that the sectional curvature of M at $\boldsymbol{p} = (x, y, z)$ equals

$$K = -\left[abc \left(\frac{x^2}{a^4} + \frac{y^2}{b^4} + \frac{z^2}{c^4} \right) \right]^{-2}.$$

3.32. Describe three mutually orthogonal geodesics (i.e., their tangent vectors are orthogonal at the points of intersection) on the ellipsoid from exercise 3.30.

3.33. Let $0 \leq a < b$, and $f : (a, b) \to \mathbb{R}$ be a smooth function. If $a = 0$, suppose that f is extendable to a with $f^{(n)}(a) = 0$ for all n.

(a) Show that $M = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_n = f((x_1^2 + \cdots + x_n^2)^{1/2})\}$ is an $(n-1)$-dimensional submanifold of $\mathbb{R}^n$.

(b) Let A denote the annulus $\{\boldsymbol{u} \in \mathbb{R}^{n-1} \mid a < |\boldsymbol{u}| < b\}$. Show that for $\boldsymbol{u} \in A$, the curve $t \mapsto (t\boldsymbol{u}, f(|t\boldsymbol{u}|))$ is, after reparametrization, a geodesic of M.

3.34. A more general surface of revolution than the one introduced in Section 3.1 is that generated by a curve rather than by the graph of a function: let $\boldsymbol{c} = (c^1, 0, c^2) : I \to$

$\mathbb{R}^3$ denote a curve in the x-z plane. Its image, when rotated about the z-axis, generates a surface M which can be parametrized by $\boldsymbol{h} : I \times [0, 2\pi] \to \mathbb{R}^3$, with

$$\boldsymbol{h}(u, v) = (c^1(u) \cos v, c^1(u) \sin v, c^2(u)).$$

(a) Show that any meridian $t \mapsto \boldsymbol{h}(t, v_0)$, $v_0 \in [0, 2\pi]$, is a geodesic.

(b) Let $\boldsymbol{\gamma} : [0, 2\pi] \to M$ denote a parallel; i.e, $\boldsymbol{\gamma}(t) = \boldsymbol{h}(u_0, t)$ for some $u_0 \in I$. Prove that $\boldsymbol{\gamma}$ is a geodesic if and only if u_0 is a critical point of c^1. Geometrically, this means that the tangent line to $\boldsymbol{c}$ is vertical at the point $\boldsymbol{c}(u_0)$ where $\boldsymbol{c}$ intersects the parallel circle.

(c) If $\boldsymbol{c}$ is given by

$$\boldsymbol{c}(t) = (t, 0, -\frac{h}{R}t + h), \qquad t \in (0, R), \quad h > 0,$$

so that the image of $\boldsymbol{c}$ is a line segment joining $(0, 0, h)$ to $(R, 0, 0)$, then the resulting surface is a cone, and by (b), parallels are not geodesics. Determine explicitly the geodesic of M that points in the parallel direction (i.e., the geodesic whose initial tangent vector equals that of some parallel $\boldsymbol{\gamma}$ in (b)). *Hint:* Show first that the cone is flat.

4 Integration on Euclidean space

We have postponed the subject of integration until now because our aim is to not only integrate real-valued functions on Euclidean space, but also differential forms on manifolds. The latter are particularly suited to formulate Stokes' theorem. We begin with the former, which closely parallels integration of single-variable functions.

4.1 The integral of a function over a box

Recall that a partition P of an interval $[a,b] \subset \mathbb{R}$ is just a finite subset of $[a,b]$ that contains the endpoints of the interval. This subset is then ordered: $a = t_0 < t_1 < \cdots < t_k = b$, and each $[t_{i-1}, t_i]$, $i = 1,\ldots,k$, is referred to as a *subinterval* of the partition. Given partitions P_i of $[a_i, b_i]$, $i = 1,\ldots,n$, the product $P = P_1 \times \cdots \times P_n$ is called a *partition* of the box $A = [a_1,b_1] \times [a_2,b_2] \times \cdots \times [a_n,b_n] \subset \mathbb{R}^n$. If P_i partitions $[a_i,b_i]$ into k_i subintervals, then P partitions A into $k_1 k_2 \cdots k_n$ boxes, which will be called the *subboxes* of P. Each subbox is then a Cartesian product $J_1 \times \cdots \times J_n$, where J_i is a subinterval of the partition P_i. By abuse of notation, we write $B \in P$ if B is a subbox of P. Definition 1.6.1 implies that the volume of A equals $\Pi_{i=1}^n (b_i - a_i)$.

Just as the integral $\int_a^b f$ of a nonnegative function f is meant to represent the area under the graph of f between a and b, the integral $\int_A f$ of a nonnegative function f of n variables will be interpreted as the volume of the $(n+1)$-dimensional solid that lies under the graph of f and above A. In the same vein, we begin by approximating this volume by sums of volumes of boxes.

Let A be a box as above, P a partition of A, and f a bounded real-valued function whose domain contains A. For each $B \in P$, set

$$m_B(f) = \inf\{f(\boldsymbol{b}) \mid \boldsymbol{b} \in B\}, \qquad M_B(f) = \sup\{f(\boldsymbol{b}) \mid \boldsymbol{b} \in B\}.$$

The *lower* and *upper sums* of f for P are respectively given by

$$L(f,P) = \sum_{B \in P} m_B(f)\,\mathrm{vol}(B), \qquad U(f,P) = \sum_{B \in P} M_B(f)\,\mathrm{vol}(B).$$

Of course $L(f,P) \le U(f,P)$, but a stronger property holds. A partition $\tilde{P}$ is said to be a *refinement* of P if $P \subset \tilde{P}$. Two easy but important observations are in order: firstly, if $\tilde{P}$ refines P, then $L(f,P) \le L(f,\tilde{P})$ and $U(f,\tilde{P}) \le U(f,P)$. To establish the first inequality, notice that any $B \in P$ equal a union of subboxes $\tilde{B}_1,\ldots,\tilde{B}_k$ of $\tilde{P}$, so that

$$m_B(f)\,\mathrm{vol}(B) = m_B(f) \sum_{i=1}^k \mathrm{vol}(\tilde{B}_i) \le \sum_{i=1}^k m_{\tilde{B}_i}(f)\,\mathrm{vol}(\tilde{B}_i).$$

(The last inequality follows from the fact $m_B(f) \le m_{\tilde{B}_i}(f)$ since $\tilde{B}_i \subset B$). The sum over all B in P of the left side is $L(f,P)$, and the corresponding sum of the right side equals $L(f,\tilde{P})$. The argument for upper sums is similar.

The second observation is that given any two partitions of A, there exists one that refines both: indeed, if $P = P_1 \times \cdots \times P_n$ and $\tilde{P} = \tilde{P}_1 \times \cdots \times \tilde{P}_n$ are the two partitions, then $(P_1 \cup \tilde{P}_1) \times \cdots \times (P_n \cup \tilde{P}_n)$ is one such refinement.

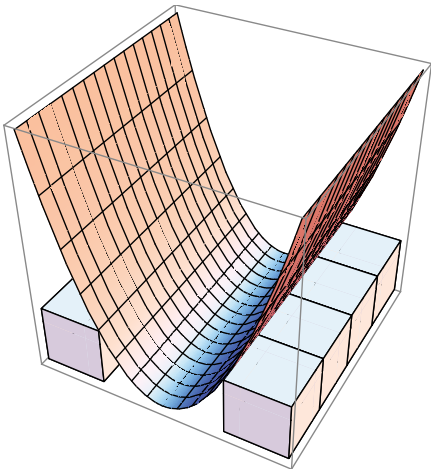


Fig. 4.1: A lower sum for $(x, y) \mapsto x^2$ on $[-2, 2] \times [-2, 2]$, using a partition with squares of side length 1.

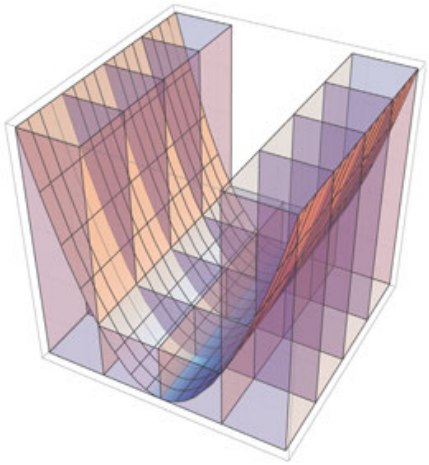


Fig. 4.2: The corresponding upper sum.

Proposition 4.1.1. *For any two partitions P and $\tilde{P}$ of a box A, $L(f, P) \leq U(f, \tilde{P})$.*

Proof. If $\bar{P}$ is a refinement of both P and $\tilde{P}$, then

$$L(f, P) \leq L(f, \bar{P}) \leq U(f, \bar{P}) \leq U(f, \tilde{P}).$$ □

By Proposition 4.1.1, the supremum $L(f, A)$ of the collection of lower sums of f for all possible partitions of A exists, and is called the *lower integral of f over A*. Similarly, the infimum $U(f, A)$ of all upper sums exists, and is called the *upper integral of f over A*. Again by the proposition, we always have that $L(f, A) \leq U(f, A)$. Furthermore, by a fundamental property of suprema – see Appendix A, given any $\varepsilon > 0$, there exists a partition P for which $L(f, P) > L(f, A) - \varepsilon$. A similar property holds for infima.

Definition 4.1.1. Let f be a bounded function on a box A. f is said to be *integrable over* A if the lower and upper integrals of f over A are equal. In this case, their common value is defined to be the *integral of* f *over* A, denoted $\int_A f$.

The following is often a useful tool in deciding whether or not a function is integrable:

Theorem 4.1.1. *Let f denote a bounded function over a box A. f is integrable over A if and only if for any $\varepsilon > 0$, there exists a partition P of A such that $U(f,P) - L(f,P) < \varepsilon$.*

Proof. Suppose f is integrable over A, and $\varepsilon > 0$ is given. There exists a partition P of A such that $U(f,P) < U(f,A) + \varepsilon/2$. Similarly, there exist a partition P' of A with $L(f,P') > L(f,A) - \varepsilon/2$. Since $U(f,A) = L(f,A)$, $U(f,P) - L(f,P') < \varepsilon$. Thus, if $\tilde{P}$ refines both P and P', then

$$U(f,\tilde{P}) - L(f,\tilde{P}) \leq U(f,P) - L(f,P') < \varepsilon.$$

Conversely, suppose that for any $\varepsilon > 0$, there exists a partition P of A such that $U(f,P) - L(f,P) < \varepsilon$. Then the nonnegative number $U(f,A) - L(f,A)$ is less than ε for any $\varepsilon > 0$, and must therefore equal 0; i.e, f is integrable. □

Examples and Remarks 4.1.1. (i) Let $c \in \mathbb{R}$, and $f : \mathbb{R}^n \to \mathbb{R}$ denote the constant function $f(\boldsymbol{a}) = c$, $\boldsymbol{a} \in \mathbb{R}^n$. Then for *any* box B, $m_B(f) = M_B(f) = c$, so that for any partition P of A, $L(f,P) = U(f,P) = c \cdot \mathrm{vol}(A)$. f is therefore integrable, and $\int_A f = c \cdot \mathrm{vol}(A)$.

(ii) Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x,y) = xy$, and the square $A = [0,1]\times[0,1]$. Denote by P_n the partition of A into n^2 squares of equal sides; i.e., $P_n = \{(i/n, j/n) \mid 0 \leq i,j \leq n\}$. If B_{ij} denotes the subsquare $[(i-1)/n, i/n] \times [(j-1)/n, j/n]$, then the minimum of f on B_{ij} occurs at the lower left corner $((i-1)/n, (j-1)/n)$, and the maximum at the upper right corner $(i/n, j/n)$, so that $m_{B_{ij}}(f) = (i-1)(j-1)/n^2$ and $M_{B_{ij}(f)} = ij/n^2$. Thus,

$$L(f,P_n) = \sum_{i,j=1}^{n} \frac{(i-1)(j-1)}{n^4} = \frac{1}{n^4}\sum_{j=1}^{n}(j-1)\frac{n(n-1)}{2} = \frac{n^2(n-1)^2}{4n^4}$$
$$= \frac{1}{4}\left(1 - \frac{1}{n}\right)^2,$$

and similarly,

$$U(f,P_n) = \sum_{i,j=1}^{n} \frac{ij}{n^4} = \frac{1}{4}\left(1 + \frac{1}{n}\right)^2.$$

It follows that $U(f,P_n) - L(f,P_n) = 1/n$, and f is integrable by Theorem 4.1.1. Furthermore, $L(f,P_n) \leq 1/4 \leq U(f,P_n)$ for all n, so that $\int_A f = 1/4$.

(iii) Let $A = [0,1] \times [0,1]$ as in (ii), and $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x,y) = 1$ if x and y are both rational, and 0 otherwise. Given any partition P of A, an arbitrary box B of P will contain points whose coordinates are rational, and points with at least one

irrational coordinate. Thus, $m_B(f) = 0$ and $M_B(f) = 1$. This implies that $L(f, A) = L(f, P) = 0$ and $U(f, A) = U(f, P) = 1$, so that f is not integrable.

(iv) A word about notation: It is common practice, when given a specific formula for a function, to include that formula in the integral. For example, the integral of the function f in (ii) is also denoted $\int_A xy\,dx\,dy$, or $\int_A u^1u^2\,du^1\,du^2$. This is rather unfortunate since du^1 is really the differential of the function u^1 and has a completely different meaning here. Nevertheless, this practice is so widespread that we will conform to it.

Theorem 4.1.2. *Suppose f and g are integrable over A.*

(1) *If $f \le g$, then $\int_A f \le \int_A g$;*
(2) *If m and M are lower and upper bounds respectively of f on A, then $m\,\mathrm{vol}(A) \le \int_B f \le M\,\mathrm{vol}(A)$;*
(3) *$|f|$ is integrable, and $|\int_A f| \le \int_A |f|$;*
(4) *$f + g$ is integrable, and $\int_A (f + g) = \int_A f + \int_A g$;*
(5) *Given $c \in \mathbb{R}$, cf is integrable, and $\int_A cf = c\int_A f$.*

Proof. If $f \le g$, then $L(f, P) \le L(g, P)$ for any partition P of A. The first statement follows by taking suprema over all partitions of A. The second one follows from the first, since f is squeezed in between the constant functions m and M. For the third one, observe first that for any box B,

$$|f|(\boldsymbol{a}) - |f|(\boldsymbol{b}) \le M_B(f) - m_B(f), \qquad \boldsymbol{a}, \boldsymbol{b} \in B. \tag{4.1.1}$$

This inequality is (tediously) verified on a case by case basis: for example, suppose $f(\boldsymbol{a}) \le 0 \le f(\boldsymbol{b})$. Then $M_B(f) \ge 0$, and

$$|f(\boldsymbol{a})| - |f(\boldsymbol{b})| = -f(\boldsymbol{a}) - f(\boldsymbol{b}) \le -f(\boldsymbol{a}) \le -m_B(f) \le M_B(f) - m_B(f).$$

Fixing an arbitrary $\boldsymbol{b}$ in (4.1.1) and taking the supremum over all $\boldsymbol{a} \in B$ implies that $M_B(|f|) - |f|(\boldsymbol{b}) \le M_B(f) - m_B(f)$, or equivalently, that

$$|f|(\boldsymbol{b}) \ge M_B(|f|) - (M_B(f) - m_B(f)), \qquad \boldsymbol{b} \in B.$$

Finally, taking the infimum over all $\boldsymbol{b} \in B$ and rearranging terms yields

$$M_B(|f|) - m_B(|f|) \le M_B(f) - m_B(f).$$

Thus, given any partition P of A, we have

$$U(|f|, P) - L(|f|, P) \le U(f, P) - L(f, P).$$

It now follows from Theorem 4.1.1 that $|f|$ is integrable whenever f is. The inequality $|\int_A f| \le \int_A |f|$ is obtained by applying the first assertion of the theorem to $-|f| \le f \le |f|$. The last two statements – like the others – have proofs that are similar to those for functions of one variable, and are left as an exercise. □

In the context of Theorem 4.1.2, it is also true that a product of integrable functions is integrable, as well as a quotient (provided the bottom function is nonzero). This turns out to be an immediate consequence of results from the next section.

4.2 Integrability and discontinuities

Our next goal is to characterize those functions that are integrable, and define integration over regions that are more complicated than boxes. This, in turn, will lead us to seek a more general concept of integral.

Definition 4.2.1. A set $A \subset \mathbb{R}^n$ is said to have *measure zero* if for any $\varepsilon > 0$, A can be covered by a countable collection $B_1, B_2, \dots$ of open boxes with $\sum_{i=1}^{\infty} \operatorname{vol}(B_i) < \varepsilon$.

Since a closed box has the same volume as its interior, a set of measure zero may also be covered by a countable collection of closed boxes satisfying the above volume condition. The converse is easily verified, so that one may replace open by closed in the definition. Given a collection $\{B_i\}$ of boxes, we will refer to the number $\sum_i \operatorname{vol}(B_i)$ (if it exists) as the *total volume* of the collection. This is just for convenience's sake, since already for a finite collection, the total volume will in general be larger than the volume of the union if the boxes intersect.

Examples 4.2.1. (i) Any countable set of points has measure zero: if $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots$ is a sequence and ε is a positive number, let B_i be a box of volume smaller than $\varepsilon/2^i$ containing $\boldsymbol{a}_i$. Then this collection has total volume less than ε. For example, the set $\mathbb{Q}$ of rationals has measure zero.

(ii) A countable union of sets A_i of measure zero has measure zero: given $\varepsilon > 0$, cover each A_i by boxes $B_{i1}, B_{i2}, \dots$ such that $\sum_j \operatorname{vol}(B_{ij}) < \varepsilon/2^i$. The collection $\{B_{ij} \mid i, j = 1, 2, \dots\}$ is then a countable cover of $\cup A_i$ (see Appendix A), and may therefore be arranged into a sequence $C_1, C_2, \dots$. There are, of course, many ways to do this, but if $\sum \operatorname{vol}(C_i)$ converges, then any rearrangement also converges to the same limit, since the convergence is absolute. To see that it does converge, let $k \in \mathbb{N}$. Then there exists an integer l such that all the *terms* $C_1, \dots, C_k$ appear in the *sequences* B_{1i} through B_{li}, $i = 1, \dots, \infty$. Thus,

$$\begin{aligned} \operatorname{vol}(C_1) + \cdots + \operatorname{vol}(C_k) &\le \sum_{i=1}^{\infty} \operatorname{vol}(B_{1i}) + \cdots + \sum_{i=1}^{\infty} \operatorname{vol}(B_{li}) \\ &< \frac{\varepsilon}{2} + \cdots + \frac{\varepsilon}{2^l} \\ &< \varepsilon. \end{aligned}$$

This means that the sequence of partial sums is bounded above by ε, and therefore $\sum_{i=1}^{\infty} \operatorname{vol}(C_i) \le \varepsilon$.

(iii) A (non-degenerate) box does not have measure zero. This may seem an obvious statement, but with only the definition to work with, it does require an argument. To see this, it suffices to consider a closed box A. Suppose $B_1, B_2, \dots$ is a collection of boxes covering A. We claim that $\sum \operatorname{vol}(B_i) \ge \operatorname{vol}(A)$. Indeed, A is compact, so a finite subcollection, which we may rename as $B_1, \dots, B_k$, covers A. This subcollection defines a partition P of A; namely, P consists of all points whose i-th

coordinate is either the beginning or the end point of the projection of some $B_j \cap A$ onto the i-th coordinate axis. Specifically, if $B_j \cap A = \Pi_{i=1}^n [a_i^j, b_i^j]$, then

$$P = \{(x_1, \ldots, x_n) \mid \text{for each } i = 1, \ldots, n,\ x_i = a_i^j \text{ or } b_i^j \text{ for some } j\}.$$

By construction, any box in the partition P is contained in each of the B_j it intersects, and the union of these boxes equals A, so that

$$\sum_i \text{vol}(B_i) \geq \sum_{i=1}^k \text{vol}(B_i) \geq \sum_{B \in P} \text{vol}(B) = \text{vol}(A).$$

This establishes the assertion, for if A had measure zero, then by the above inequality, A would have volume less than any positive number ε, and would then be a degenerate box.

(iv) The collection of all irrational numbers between 0 and 1 does not have measure zero: if it did, then by (i) and (ii), the interval $[0, 1]$ would have measure zero, contradicting (iii).

(v) We will later see that a ball of radius r in $\mathbb{R}^n$ has volume $k_n r^n$, where k_n is a constant depending on n, see also Exercise 4.25. Assuming this, we could use open or closed balls instead of boxes in the definition of measure zero, since

$$[-r, r]^n \subset B_{r\sqrt{n}}(\mathbf{0}) \subset [-r\sqrt{n}, r\sqrt{n}]^n.$$

Remark 4.2.1. The type of construction referred to in part (iii) of the above examples will be used many times. One variant of it is worth emphasizing: If $B_1, \ldots, B_k$ are boxes contained in a larger box B, then there exists a partition P of B such that each B_i is a union of subboxes of P. The proof is the same as the one used in (iii).

Now that the preliminaries have been dealt with, we are in a position to characterize those functions that are integrable. In light of our work with upper and lower sums, it shouldn't come as a surprise to learn that they are the ones with not too many jumps. Specifically:

Theorem 4.2.1. *Let $f : A \to \mathbb{R}$ denote a function that is bounded on the closed box A. Then f is integrable over A if and only if the set of discontinuities of f has measure zero.*

Proof. Denote by D the set of points where f is discontinuous, and for $\delta > 0$, by D_δ the set of all $\boldsymbol{a} \in A$ such that $M_U(f) - m_U(f) \geq \delta$ for any neighborhood U of $\boldsymbol{a}$. Then $D_\delta \subset D$ for any δ, and $D = \cup_{i=1}^\infty D_{1/i}$ by Exercise 1.53.
Suppose first that f is integrable. To show that D has measure zero, it suffices to show that each $D_{1/k}$ has measure zero. Now, given any partition P of A, $D_{1/k}$ may be written as a union

$$D_{1/k} = \left(D_{1/k} \cap \bigcup_{B \in P} B^0 \right) \bigcup \left(D_{1/k} \cap \bigcup_{B \in P} \partial B \right)$$

of two sets, the second of which has measure zero. Thus, given $\varepsilon > 0$, it suffices to produce a partition P of A for which those boxes that intersect $D_{1/k}$ in their interior have

total volume less than ε. We claim that any partition P of A for which $U(f,P) - L(f,P) < \varepsilon/k$ will do. Indeed, denote by $\mathcal{C}$ the collection of boxes in P that intersect $D_{1/k}$ in their interior. Then $M_B(f) - m_B(f) \geq 1/k$ for each $B \in \mathcal{C}$, so that

$$\begin{aligned}\sum_{B\in\mathcal{C}} \operatorname{vol}(B) &\leq k \cdot \sum_{B\in\mathcal{C}} (M_B(f) - m_B(f)) \operatorname{vol}(B) \\ &\leq k \cdot \sum_{B\in P} (M_B(f) - m_B(f)) \operatorname{vol}(B) = k(U(f,P) - L(f,P)) \\ &< \varepsilon,\end{aligned}$$

and $D_{1/k}$ has measure zero, as claimed.

Conversely, suppose D has measure zero. Given $\varepsilon > 0$, we must exhibit a partition P of A satisfying $U(f,P) - L(f,P) < \varepsilon$. Let M be an upper bound of $|f|$ on A, and set $\delta := \varepsilon/(2M + \operatorname{vol}(A))$. Cover D by open boxes B_i with $\sum_i \operatorname{vol}(B_i) < \delta$. Similarly, for each $\boldsymbol{a} \in A \setminus D$, choose some box $C_{\boldsymbol{a}}$ that contains $\boldsymbol{a}$ in its interior and satisfies $M_{C_{\boldsymbol{a}}}(f) - m_{C_{\boldsymbol{a}}}(f) < \delta$. The two collections yield a cover of A. Choose a finite subcover, and a partition P of A such that each subbox in P is contained in one of the boxes from the subcover, see Remark 4.2.1.

If B is a subbox contained in some B_i, then $(M_B(f) - m_B(f)) \operatorname{vol}(B) < 2M \operatorname{vol}(B)$, and if B is contained in some $C_{\boldsymbol{a}}$, then $(M_B(f) - m_B(f)) \operatorname{vol}(B)$ is less than $\delta \operatorname{vol}(B)$. Thus,

$$U(f,P) - L(f,P) < 2M\delta + \delta \operatorname{vol}(A) = \varepsilon. \qquad \square$$

We are now in a position to integrate over regions that are more general than boxes:

Definition 4.2.2. The *characteristic function* χ_A of a bounded set $A \subset \mathbb{R}^n$ is defined by $\chi_A(\boldsymbol{a}) = 1$ if $\boldsymbol{a} \in A$ and $\chi_A(\boldsymbol{a}) = 0$ otherwise.

A is said to be *Jordan-measurable* if $\int_B \chi_A$ exists, with B denoting some box that contains A. In this case, the value of this integral is called the n-dimensional *volume* of A.

Remark 4.2.2. If A is a bounded Jordan-measurable set of measure zero, then it has zero volume: consider any partition P of some box B containing A. The infimum of χ_A on any subbox of P must be zero, since the only other possibility is 1. The latter is ruled out, for it would imply that the subbox is contained in A, contradicting the fact that A has measure zero. Thus $L(\chi_A, P) = 0$ for any P, and the claim follows.

Notice, however, that there exist bounded sets of measure zero which are not Jordan-measurable. One such is $A = \mathbb{Q} \cap [0,1] \subset \mathbb{R}$.

Definition 4.2.3. A bounded $f : A \to \mathbb{R}$ is said to be *integrable over* A if $\int_B f \cdot \chi_A$ exists, with B denoting any box containing A. The value of this integral is denoted $\int_A f$.

Clearly, the above definition does not depend on the particular box B, and f is integrable if $\int_B f \cdot \chi_A$ exists for *some* box B containing A.

Theorem 4.2.2. *A bounded set A is Jordan-measurable if and only if the boundary of A has measure zero.*

Proof. If B is a closed box whose interior contains A, then the restriction of χ_A to B has exactly the boundary of A as its set of discontinuities, since any neighborhood of a point on the boundary contains points of A (where χ_A equals 1) and points outside A (where it equals zero). The claim now follow from Theorem 4.2.1. □

Examples 4.2.2. (i) Any region in $\mathbb{R}^2$ whose boundary is (the image of) a bounded curve is Jordan-measurable: let $\boldsymbol{c} : [0,b] \to \mathbb{R}^2$ be a piecewise-smooth curve parametrizing the boundary, and let $L = \int_0^b |\boldsymbol{c}'|$ denote its length. Given $\varepsilon > 0$, we must find a cover of the boundary by rectangles with total volume (which we call area in dimension 2) less than ε. So let n be an integer large enough so that $(n+1)/n^2 < \varepsilon/9L^2$. Since the function $f : [0,b] \to \mathbb{R}$ given by $f(t) = \int_0^t |\boldsymbol{c}'|$ is continuous, there exists a partition $t_0 = 0 < t_1 < \cdots < t_n = b$ of $[0,b]$ such that $f(t_i) = iL/n$, $0 \le i \le n$; i.e., the portion of the curve joining the points $\boldsymbol{p}_{i-1} = \boldsymbol{c}(t_{i-1})$ and $\boldsymbol{p}_i = \boldsymbol{c}(t_i)$ has length L/n. In particular, the distance between these two is no larger than L/n, so that *any* point on the boundary is at distance less than or equal to L/n from some $\boldsymbol{p}_i$. It follows that the collection $\mathcal{C}$ of squares centered at $\boldsymbol{p}_i$ with sides of length $3L/n$ covers $\mathcal{C}$, and

$$\sum_{S\in\mathcal{C}} \mathrm{vol}(S) = (n+1)\frac{9L^2}{n^2} < \varepsilon,$$

which establishes the claim.

(ii) One would hope that any bounded open set in $\mathbb{R}^n$ is Jordan-measurable. This is not the case, however, even in dimension 1, a fact that will lead us to generalize the concept of integral in order to avoid these exceptions. One such is obtained by arranging all rational numbers in $(0,1)$ in a sequence $x_1, x_2, \ldots$, and setting

$$U_i = \left(x_i - \frac{1}{2^{i+2}}, x_i + \frac{1}{2^{i+2}}\right) \cap (0,1), \qquad U = \cup_{i=1}^{\infty} U_i.$$

Notice that U is open, so that its boundary ∂U is disjoint from U. Furthermore, $\partial U \subset [0,1]$, since any $x \notin [0,1]$ has a neighborhood (namely $(-\infty, 0)$ or $(1,\infty)$) that does not intersect $[0,1]$, and therefore does not intersect U either. Thus, $\partial U \subset [0,1] \setminus U$. Conversely, if $x \in [0,1] \setminus U$, then any neighborhood of x intersects both U (since it must contain some rational) and the complement of U (since it contains x). Thus, $\partial U = [0,1] \setminus U$, and $[0,1]$ equals the disjoint union of U and ∂U. This means that ∂U does not have measure zero: indeed,

$$\mathrm{vol}(U) \le \sum_i \mathrm{vol}(U_i) \le \sum_{i=1}^{\infty} \frac{1}{2^{i+1}} = \frac{1}{2},$$

so that if ∂U could be covered by intervals of total length less than $1/2$, then $[0,1]$ could be covered by intervals of total length smaller than 1, contradicting Examples 4.2.1 (iii).

(iii) Let U, V denote open sets in $\mathbb{R}^n$. A map $h : U \to V$ is said to be a *homeomorphism* if it is bijective, continuous, and has continuous inverse. If h is a homeomorphism

and $A \subset U$ is a compact Jordan-measurable set, then so is $h(A)$. The compactness property was proved in Chapter 1. The Jordan-measurability can be seen as follows: the fact that h is a homeomorphism easily implies that $h(\partial A) = \partial h(A)$. The restriction of h to A is uniformly continuous. Thus, for any $\varepsilon > 0$, there exists some $\delta > 0$ such that any ball of radius δ centered at a point of ∂A is mapped by h into a ball of radius ε centered at a point of $\partial h(A)$. The claim now follows from Examples 4.2.1 (v); alternatively, one could avoid any mention of volume of balls by rephrasing continuity in terms of boxes.

In order to avoid the awkward situation from part (ii) in the above example, we will use partitions of unity when the region of integration is an open set. Let f be a function with open domain U that is bounded in some neighborhood of any point in U, and suppose the set of discontinuities of f has measure zero. Let $\{U_i\}$ denote an *admissible* open cover of U (meaning that each U_i is contained in U), and $\boldsymbol{\Phi} = \{\varphi_j\}$ a countable partition of unity subordinate to the cover. An arbitrary $\varphi \in \boldsymbol{\Phi}$ is zero outside some compact set in U, so that $\varphi|f|$ is integrable on any closed box that contains U. Suppose that the series $\sum_i \int_U \varphi|f|$ converges. Since $|\int_U \varphi_i f| \leq \int_U \varphi_i|f|$, this in turn implies absolute convergence of $\sum_i \int_U \varphi_i f$. Now, $\sum \varphi_i \equiv 1$, so it is tempting to define $\int_U f$ as the sum of this series. It must first be checked, though, that a different partition of unity $\boldsymbol{\Psi} = \{\psi_j\}$ produces the same sum. This is not difficult to show, for if $\varphi \in \Phi$, then $\int_U \varphi f = \int_U \sum_j \psi_j \varphi f$. Furthermore, each point in the support of φ has a neighborhood on which only finitely many ψ_j are nonzero. Since the support is compact, only finitely many ψ_j are nonzero on it. Thus,

$$\int_U \varphi f = \sum_{j=1}^{\infty} \int_U \psi_j \varphi f,$$

and consequently,

$$\sum_{i=1}^{\infty} \int_U \varphi_i f = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \int_U \psi_j \varphi_i f.$$

Applying this identity to $|f|$ shows that the convergence is absolute. Similarly,

$$\sum_{j=1}^{\infty} \int_U \psi_j f = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \int_U \psi_j \varphi_i f.$$

Absolute convergence implies that the order of summation may be interchanged (see Examples 4.2.1 (ii)), so that

$$\sum_{\varphi \in \boldsymbol{\Phi}} \int_U \varphi f = \sum_{\psi \in \boldsymbol{\Psi}} \int_U \psi f.$$

Notice that the sum is not only independent of the partition of unity, it is also independent of the open cover to which it is subordinate, since this cover was not used in the argument.

Lemma 4.2.1. *Let $U \subset \mathbb{R}^n$ be an open bounded set, $f : U \to \mathbb{R}$ a function that is bounded and has a set of discontinuities of measure zero. Given any admissible open cover of U and partition of unity $\{\varphi_i\}_{i\in\mathbb{N}}$ subordinate to this cover, the series*

$$\sum_{i=1}^{\infty} \int_U \varphi_i |f|$$

converges, and the sum is independent of the particular cover and partition of unity chosen.

Proof. Let B denote a closed box that contains U. Each $\varphi_i|f|$ extends to a function (denoted in the same way) on B which is smooth outside U by setting it equal to zero there. If M denotes an upper bound of $|f|$ on U, the partial sums of the series are bounded above, because

$$\sum_{i=1}^{n} \int_U \varphi_i |f| = \int_U \sum_{i=1}^{n} \varphi_i |f| = \int_B \sum_{i=1}^{n} \varphi_i |f| \leq \int_B |f| \leq M \operatorname{vol}(B).$$

Thus, the series converges. Independence of the cover and partition of unity was established earlier. □

Definition 4.2.4. Let $f : U \to \mathbb{R}$, where U and f satisfy the hypotheses of Lemma 4.2.1. The *generalized integral of f over U* is defined to be the number

$$\int_U f := \sum_{i=1}^{\infty} \int_U \varphi_i f,$$

where $\{\varphi_i\}$ is any partition of unity subordinate to some admissible open cover of U.

Of course, if this definition is to be of any value, it should coincide with the old one when the region of integration is Jordan-measurable.

Theorem 4.2.3. *If f is a bounded function which is integrable over some bounded Jordan-measurable set A, then the original $\int_A f$ coincides with the generalized one from Definition 4.2.4.*

Proof. Let $\boldsymbol{\Phi} = \{\varphi_1, \varphi_2, \dots\}$ be a partition of unity for A, $\varepsilon > 0$, and $\int_A f$ denote the original integral. The claim follows once we show that there exists a natural number N such that

$$\left| \int_A f - \sum_{j=1}^{k} \int_A \varphi_j f \right| < \varepsilon \text{ for all } k \geq N.$$

So consider a box B that contains A, and let M denote an upper bound of $|f|$ on B. Since ∂A is compact and has measure zero, there exist boxes $B_1, \dots, B_k \subset B$ covering ∂A with total volume less than ε/M. Take a partition P of B where each B_i is a union of subboxes of P (see Remark 4.2.1), and denote by C the union of the subboxes of P that

lie inside A. Then

$$\operatorname{vol}(A \setminus C) \leq \sum_{i=1}^{k} \operatorname{vol}(B_i) < \frac{\varepsilon}{M}.$$

Furthermore, by compactness of C, the collection of all $\varphi \in \boldsymbol{\Phi}$ that are not identically zero on C is finite. This means that there exists an integer N such that for $k \geq N$, the restriction of φ_k to C is zero, and therefore, $\int_A \varphi_k f = \int_{A\setminus C} \varphi_k f$. But then for any $k \geq N$,

$$\begin{aligned}\left|\int_A f - \sum_{j=1}^{k} \int_A \varphi_j f\right| &= \left|\int_A \left(f - \sum_{j=1}^{k} \varphi_j f\right)\right| \leq \int_A \left|f - \sum_{j=1}^{k} \varphi_j f\right| \\ &= \int_A \left(1 - \sum_{j=1}^{k} \varphi_j\right) |f| \leq M \int_A \left(1 - \sum_{j=1}^{k} \varphi_j\right) \\ &= M \int_A \sum_{j>k} \varphi_j \leq M \int_{A\setminus C} 1 = M \operatorname{vol}(A \setminus C) \\ &< \varepsilon. \end{aligned}$$

□

As before, we define the *volume* $\operatorname{vol}(A)$ of $A \subset \mathbb{R}^n$ to be $\int_A 1$. Thus, the volume of a bounded Jordan-measurable set coincides with the original notion of volume, but Lemma 4.2.1 now guarantees that *any bounded open set*, even a non Jordan-measurable one (see Examples 4.2.2), has a well-defined volume.

Remarks 4.2.3. (i) The proof of Theorem 4.2.3 shows that if A is a Jordan-measurable bounded set, then for any $\varepsilon > 0$ there exists a compact Jordan-measurable subset C of A such that $\operatorname{vol}(A \setminus C) < \varepsilon$.

(ii) If U is open in $\mathbb{R}^n$, and $f, g : U \to \mathbb{R}$ are equal everywhere except on a set of measure zero, then f is integrable on U if and only if g is, and in that case, both integrals are equal. This is because $h = f - g$ vanishes except on a set of measure zero and is therefore integrable. Thus, if f is integrable, then so is $g = f - h$, and vice-versa.

4.3 Fubini's theorem

Any box B in $\mathbb{R}^{m+n}$ decomposes as a product $B_1 \times B_2 = \pi_1(B) \times \pi_2(B) \subset \mathbb{R}^m \times \mathbb{R}^n$ of boxes, with π_i denoting the orthogonal projection of $\mathbb{R}^m \times \mathbb{R}^n$ onto each factor, $i = 1, 2$. Given $f : B \to \mathbb{R}$, each $\boldsymbol{x} \in B_1$ induces a function $f_{\boldsymbol{x}} : B_2 \to \mathbb{R}$ by setting $f_{\boldsymbol{x}}(\boldsymbol{y}) = f(\boldsymbol{x}, \boldsymbol{y})$. As an elementary example, let $B = [0, 2] \times [0, 2]$, and $f(x, y) = (x + 1)y^2$. Then $f_0(x) = y^2$ and $f_1(x) = 2y^2$.

Suppose next that for each $\boldsymbol{x} \in B_1$, $f_{\boldsymbol{x}}$ is integrable on B_2. This yields a new function g on B_1 by setting $g(\boldsymbol{x}) = \int_{B_2} f_{\boldsymbol{x}}$. This expression is commonly denoted $\int_{B_2} f(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y}$. In the above example,

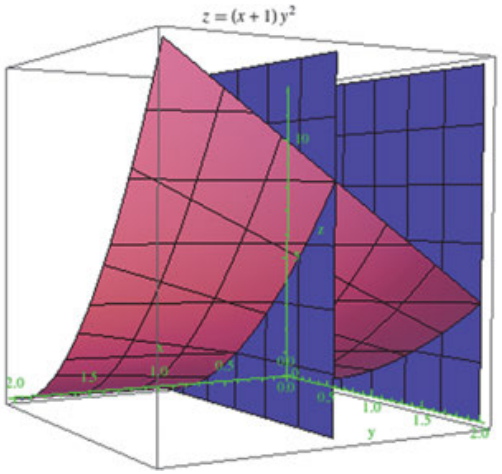


Fig. 4.3: Cross-sectional areas $\int_{B_2} f(x,y)\mathrm{y}$ for $x = 1$ and $x = 0$.

$$\int_{B_2} f(x,y)\,\mathrm{d}y = \int_0^2 (x+1)y^2\,\mathrm{d}y = (x+1)\frac{y^3}{3}\Big|_{y=0}^{y=2} = \frac{8(x+1)}{3}.$$

If this function g is integrable over B_1, its integral is called an *iterated integral*, and is denoted $\int_{B_1}\int_{B_2} f(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x}$. In our example, the reader can verify that this integral equals 32/3, and that reversing the order of integration (that is, evaluating $\int_{B_2}\int_{B_1} f(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}$ instead) yields the same result. This is no coincidence. Fubini's theorem implies that for continuous f, both iterated integrals are equal, and furthermore equal the integral $\int_B f$ of the original function over B, thereby substantially simplifying the evaluation of these integrals. The theorem actually holds for arbitrary integrable functions, although the statement becomes more complicated because iterated integrals do not always exist.

Recall that a bounded function $f : B \to \mathbb{R}$ always admits lower and upper integrals $L(f,B)$, $U(f,B)$, regardless of whether it is integrable. Also recall that a partition P of $B_1 \times B_2$ induces partitions $P_i = \pi_i(P)$ of B_i, $i = 1,2$, and $P = P_1 \times P_2$.

Lemma 4.3.1. *Let B_1 and B_2 denote rectangles in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively, and $f : B_1 \times B_2 \to \mathbb{R}$ be a bounded function. For each $\boldsymbol{x} \in B_1$, define $f_{\boldsymbol{x}} : B_2 \to \mathbb{R}$ by $f_{\boldsymbol{x}}(\boldsymbol{y}) = f(\boldsymbol{x},\boldsymbol{y})$, and denote by $g_1 : B_1 \to \mathbb{R}$ the function given by $g_1(\boldsymbol{x}) = L(f_{\boldsymbol{x}}, B_2)$. Then for any partition P of B,*

$$L(f,P) \le L(g_1,P_1) \le U(g_1,P_1) \le U(f,P),$$

where $P_1 = \pi_1(P)$ is the induced partition of B_1.

Proof. Observe that if S_i are subboxes of P_i, $i = 1,2$, and $\boldsymbol{x} \in S_1$, then

$$m_{S_1\times S_2}(f) \le m_{\{\boldsymbol{x}\}\times S_2}(f) = m_{S_2}(f_{\boldsymbol{x}}).$$

Consequently,

$$\sum_{S_2\in P_2} m_{S_1\times S_2}(f)\,\mathrm{vol}(S_2) \le \sum_{S_2\in P_2} m_{S_2}(f_{\boldsymbol{x}})\,\mathrm{vol}(S_2) = L(f_{\boldsymbol{x}},P_2) \le g_1(\boldsymbol{x}).$$

Since this holds for every $\boldsymbol{x} \in S_1$,

$$\sum_{S_2 \in P_2} m_{S_1 \times S_2}(f) \operatorname{vol}(S_2) \le m_{S_1}(g_1),$$

so that

$$\begin{aligned} L(f,P) &= \sum_{S_1 \times S_2} m_{S_1 \times S_2}(f) \operatorname{vol}(S_1 \times S_2) \\ &= \sum_{S_1 \in P_1} \Big(\sum_{S_2 \in P_2} m_{S_1 \times S_2}(f) \operatorname{vol}(S_2) \Big) \operatorname{vol}(S_1) \le \sum_{S_1 \in P_1} m_{S_1}(g_1) \operatorname{vol}(S_1) \\ &= L(g_1, P_1). \end{aligned} \tag{4.3.1}$$

Next, if $h_1 : B_1 \to \mathbb{R}$ is given by $h_1(\boldsymbol{x}) = U(f_{\boldsymbol{x}}, P_2)$, then a similar argument, using h_1 instead of g_1, shows that $U(h_1, P_1) \le U(f, P)$. But $g_1 \le h_1$, so that

$$U(g_1, P_1) \le U(f, P).$$

This, together with (4.3.1), yields the claim. □

Remarks 4.3.1. (i) With minor modifications, the proof of the lemma shows that

$$L(f,P) \le L(h_1, P_1) \le U(h_1, P_1) \le U(f,P),$$

where h_1 is the function that was used in the proof, $h_1(\boldsymbol{x}) = U(f_{\boldsymbol{x}}, P_2)$.

(ii) Define, for each $\boldsymbol{y} \in B_2$, a function $f_{\boldsymbol{y}} : B_1 \to \mathbb{R}$ by $f_{\boldsymbol{y}}(\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{y})$, and denote by $g_2, h_2 : B_2 \to \mathbb{R}$ the functions given by

$$g_2(\boldsymbol{y}) = L(f_{\boldsymbol{y}}, B_1), \qquad h_2(\boldsymbol{y}) = U(f_{\boldsymbol{y}}, B_1).$$

Arguments similar to those used in the proof of the lemma imply that

$$L(f,P) \le L(g_2, P_2) \le U(g_2, P_2) \le U(f,P),$$

and

$$L(f,P) \le L(h_2, P_2) \le U(h_2, P_2) \le U(f,P).$$

Theorem 4.3.1 (Fubini's Theorem). *Suppose $f : B_1 \times B_2 \subset \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is integrable. Then, with notation as in Lemma 4.3.1 and the remark following it, the functions g_i, h_i are integrable on B_i, and*

$$\int_{B_1 \times B_2} f = \int_{B_i} g_i = \int_{B_i} h_i, \qquad i = 1, 2.$$

In particular if f is continuous, then

$$\int_{B_1 \times B_2} f = \int_{B_1} \int_{B_2} f(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}\boldsymbol{x} = \int_{B_2} \int_{B_1} f(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y}.$$

Proof. We limit ourselves to establishing $\int_{A\times B} f = \int_{B_1} g_1$ in the first identity, since they are all proved in essentially the same way. By Lemma 4.3.1,

$$L(f,B) \le L(g_1,B_1) \le U(g_1,B_1) \le U(f,B).$$

But f is integrable, so $L(f,B) = U(f,B)$, and the claim follows.
For the second identity, observe that if f is continuous, then so is $f_{\boldsymbol{x}}$ for any $\boldsymbol{x} \in B_1$, and

$$g_1(\boldsymbol{x}) = L(f_{\boldsymbol{x}}, B_2) = \int_{B_2} f_{\boldsymbol{x}} = \int_{B_2} f(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}.$$

Therefore,

$$\int_{B_1\times B_2} f = \int_{B_1} g_1 = \int_{B_1}\int_{B_2} f(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x}.$$

Using g_2 instead of g_1 shows that the order of integration may be reversed. □

Examples 4.3.1. (i) The function f, where $f(x,y) = 2x^3e^{x^2y}$, is integrable over $R = [0,1]\times[0,1]$ since it is continuous. Trying to evaluate the iterated integral $\int_0^1\int_0^1 2x^3e^{x^2y}\,\mathrm{d}x\,\mathrm{d}y$ doesn't look auspicious. If we reverse the order of integration, however, then

$$\int_0^1 2x^3e^{x^2y}\,\mathrm{d}y = 2xe^{x^2y}\Big|_{y=0}^{y=1} = 2x(e^{x^2}-1),$$

so that $\int_R f = \int_0^1 2x(e^{x^2}-1)\,\mathrm{d}x = e^{x^2}-x^2\big|_0^1 = e-2$.

(ii) It may well happen that a function is integrable, but one of the iterated integrals does not exist: consider

$$f : [0,1]\times[0,1] \to \mathbb{R},$$

$$(x,y) \mapsto \begin{cases} 1 & \text{if } x = y = 0,\\ 0 & \text{if } x \text{ or } y \text{ is irrational},\\ \frac{1}{n} & \text{if } y \text{ is rational and } x = \frac{m}{n} \neq 0, \end{cases}$$

where m, n are integers with no common factor and $n > 0$. We claim that the set of discontinuities of f is $(\mathbb{Q}\cap[0,1])\times[0,1]$, so that f is integrable. To see this, consider a point (x_0,y_0) in this set, with $x_0 = p/q$. If y_0 is irrational, then $f(x_0,y_0) = 0$, but $f(x_0,y_k) = 1/q$ for any sequence $\{y_k\}$ of rationals that converges to y_0. This shows that f is discontinuous at that point. If y_0 is rational, then $f(x_0,y_0) \neq 0$ but any neighborhood of (x_0,y_0) contains points where f vanishes, so again f is discontinuous there. Next, we check that f is continuous at any other point; i.e., at any (x_0,y_0) with x_0 irrational. So consider a sequence $(x_k,y_k) \to (x_0,y_0)$. If x_k is rational for only finitely many k, then for large enough k, $f(x_k,y_k) = 0 = f(x_0,y_0)$. Otherwise, there exists a subsequence $x_{k_n} = p_n/q_n$ of $\{x_k\}$ consisting of rationals. It is not difficult to prove that in this case, $q_n \to \infty$ (see Examples and Remarks (ii)

in Appendix A), so that $f(x_{k_n}, y_{k_n})$ is either zero or $1/q_n$, and in any case converges to zero. Since those terms that are not in the subsequence are mapped to zero, f is indeed continuous at (x_0, y_0).

Thus, f is integrable, and its integral is zero, since any lower sum is. However, if $x = m/n$ is a nonzero rational, then $f_x(y) = 1/n$ when y is rational and 0 otherwise, so that f_x is not integrable and $\int_0^1 f(x,y)\,\mathrm{d}y$ does not exist. The other integral $\int_0^1 f(x,y)\,\mathrm{d}x$ does, however, exist, and equals zero: this is clear if y is irrational, since f_y is then identically zero, and follows from the example in Appendix A mentioned above when y is rational.

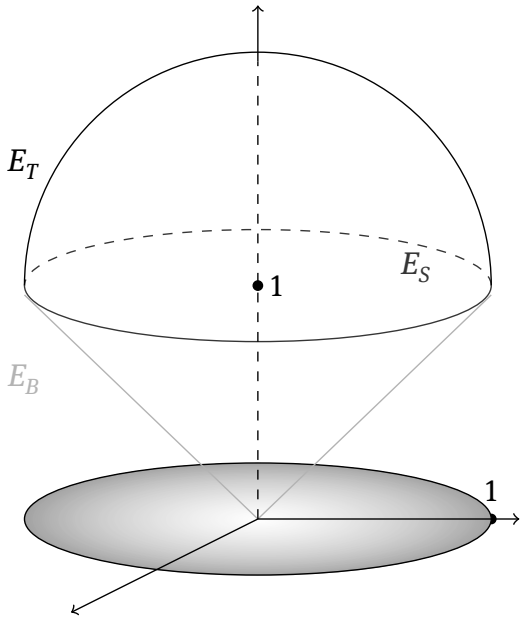
Fubini's theorem is applicable to regions that are more general than boxes: The *j-th coordinate plane* in $\mathbb{R}^n$, $j = 1, \dots, n$, is the set Π_j of all points $\boldsymbol{a} \in \mathbb{R}^n$ with $u^j(\boldsymbol{a}) = 0$. It is canonically isomorphic with $\mathbb{R}^{n-1}$ via

$$\begin{aligned} \imath_j : \mathbb{R}^{n-1} &\to \Pi_j, \\ (a_1, \dots, a_{n-1}) &\mapsto (a_1, \dots, a_{j-1}, 0, a_j, \dots, a_{n-1}), \end{aligned}$$

and we routinely identify the two. Define the projection π_j of $\mathbb{R}^n$ onto the j-th coordinate plane by $\pi_j(\boldsymbol{a}) = (u^1(\boldsymbol{a}), \dots, u^{j-1}(\boldsymbol{a}), u^{j+1}(\boldsymbol{a}), \dots, u^n(\boldsymbol{a}))$. A nonempty set $E \subset \mathbb{R}^n$ is said to be *projectable* if there exists a compact Jordan-measurable set $\tilde{E} \subset \mathbb{R}^{n-1}$, some $j \in \{1, \dots, n\}$ and continuous functions $g_i : \tilde{E} \to \mathbb{R}$, $i = 1, 2$, such that

$$E = \{\boldsymbol{a} \in \mathbb{R}^n \mid \pi_j(\boldsymbol{a}) \in \tilde{E} \text{ and } (g_1 \circ \pi_j)(\boldsymbol{a}) \le u^j(\boldsymbol{a}) \le (g_2 \circ \pi_j)(\boldsymbol{a})\}.$$

Geometrically speaking, a projectable set E consists of the region lying between the graphs of two functions defined on the projection of E onto some coordinate plane. Notice that such a set is entirely determined (and we say it is generated) by j, $\tilde{E}$, g_1, and g_2.



Projectable set with
$g_1(x,y) = \sqrt{x^2 + y^2}$,
$g_2(x,y) = 1 + \sqrt{1 - x^2 - y^2}$,
$\tilde{E} = \bar{B}_1(\boldsymbol{0}) \subset \mathbb{R}^2$, $j = 3$.

Lemma 4.3.2. *A projectable set is Jordan-measurable.*

Proof. The claim follows from Theorem 4.2.2 once we establish that the boundary of E has measure zero. So suppose E is generated by j, $\tilde{E}$, g_1, and g_2. We may assume without loss of generality that $j = n$. Now, the boundary of E decomposes as a union of three compact overlapping sets, a "top" $E_T = \{(\boldsymbol{x}, g_2(\boldsymbol{x})) \mid \boldsymbol{x} \in \tilde{E}\}$, a "bottom" $E_B = \{(\boldsymbol{x}, g_1(\boldsymbol{x})) \mid \boldsymbol{x} \in \tilde{E}\}$, and a "side" $E_S = \{(\boldsymbol{x}, x_n) \mid \boldsymbol{x} \in \partial E, g_1(\boldsymbol{x}) \le x_n \le g_2(\boldsymbol{x})\}$.
Consider first the top. Let $\varepsilon > 0$, and choose some box $B \subset \mathbb{R}^{n-1}$ that contains $\tilde{E}$. Since g_2 is uniformly continuous on the compact set $\tilde{E}$, there exists some $\delta > 0$ such that $|g_2(\boldsymbol{a}) - g_2(\boldsymbol{b})| < \varepsilon/(4\,\mathrm{vol}(B))$ whenever $\boldsymbol{a}, \boldsymbol{b} \in \tilde{E}$ are at a distance less than δ from each other. Next, divide B into subboxes $B_1, \ldots, B_k$ of diameter less than δ, and choose some $\boldsymbol{a}_i \in B_i$, $i = 1, \ldots, k$. Then the sets

$$B_i \times [g_2(\boldsymbol{a}_i) - \frac{\varepsilon}{3\,\mathrm{vol}(B)}, g_2(\boldsymbol{a}_i) + \frac{\varepsilon}{3\,\mathrm{vol}(B)}], \qquad i = 1, \ldots, k$$

form a collection of n-dimensional boxes covering E_T with total volume $\sum_i 2\varepsilon\,\mathrm{vol}(B_i)/(3\,\mathrm{vol}(B)) = 2\varepsilon/3 < \varepsilon$. A similar argument shows that the bottom E_B can also be covered by boxes with arbitrarily small total volume.
It remains to consider the side E_S. Set

$$m = \min\{g_1(\boldsymbol{x}) \mid \boldsymbol{x} \in \partial\tilde{E}\}, \qquad M = \max\{g_2(\boldsymbol{x}) \mid \boldsymbol{x} \in \partial\tilde{E}\}.$$

Let $\varepsilon > 0$. Since $\tilde{E}$ is Jordan-measurable, there exist $(n-1)$-dimensional boxes $\tilde{B}_1, \ldots, \tilde{B}_l$ covering the boundary of $\tilde{E}$ with total volume less than $\varepsilon/(M - m)$ (at least if $m < M$. If $m = M$, then $g_1 \equiv g_2$ equal a constant α, and $E_S = \partial\tilde{E} \times \{\alpha\}$ certainly has measure zero). Then

$$E_S \subset \bigcup_{i=1}^{l} \tilde{B}_i \times [m, M],$$

and $\sum_i \mathrm{vol}(\tilde{B}_i \times [m, M]) < \varepsilon$. □

Proposition 4.3.1. *Suppose $E \subset \mathbb{R}^n$ is projectable, generated by $j, \tilde{E}, g_1$, and g_2. For each $\boldsymbol{x} = (x_1, \ldots, x_{n-1}) \in \tilde{E}$, define $f_{\boldsymbol{x}} : [g_1(\boldsymbol{x}), g_2(\boldsymbol{x})] \to \mathbb{R}$ by*

$$f_{\boldsymbol{x}}(t) = f(x_1, \ldots, x_{j-1}, t, x_j, \ldots, x_{n-1}).$$

If $f : E \to \mathbb{R}$ is integrable, then

$$\int_E f = \int_{\tilde{E}} \tilde{f}, \text{ where } \tilde{f}(\boldsymbol{x}) = \int_{g_1(\boldsymbol{x})}^{g_2(\boldsymbol{x})} f_{\boldsymbol{x}}, \quad \boldsymbol{x} \in \tilde{E}.$$

Proof. For simplicity of notation, assume that $j = n$. Consider any box $B = [a_1, b_1] \times \cdots \times [a_n, b_n]$ that contains E, and define $g : B \to \mathbb{R}$ by setting it equal to f inside E, and

zero otherwise. If $\tilde{B} = \pi_n(B) \subset \mathbb{R}^{n-1}$, then by Fubini's theorem,

$$\begin{aligned}\int_E f &= \int_B g = \int_{\tilde{B}} \left(\int_{a_n}^{b_n} g(\boldsymbol{x}, t)\, \mathrm{d}t \right) \mathrm{d}\boldsymbol{x} = \int_{\tilde{E}} \left(\int_{a_n}^{b_n} g(\boldsymbol{x}, t)\, \mathrm{d}t \right) \mathrm{d}\boldsymbol{x} \\ &= \int_{\tilde{E}} \left(\int_{g_1(\boldsymbol{x})}^{g_2(\boldsymbol{x})} f(\boldsymbol{x}, t)\, \mathrm{d}t \right) \mathrm{d}\boldsymbol{x} \\ &= \int_{\tilde{E}} \tilde{f}. \end{aligned}$$

□

Examples 4.3.2. (i) Suppose we wish to find the volume of the 3-dimensional region R that lies inside the cylinder $x^2 + y^2 = 9$ and between the planes $z = 1$ and $y + z = 5$. Let D denote the disk $\{(x, y) \mid x^2 + y^2 \leq 9\}$ in $\mathbb{R}^2$. Then

$$R = \{(x, y, z) \mid (x, y) \in D \text{ and } 1 \leq z \leq 5 - y\},$$

so that

$$\mathrm{vol}(R) = \int_D \left(\int_1^{5-y} \mathrm{d}z \right) \mathrm{d}y\, \mathrm{d}x = \int_D (4 - y)\, \mathrm{d}y\, \mathrm{d}x.$$

Now, D itself is projectable; in fact,

$$D = \{(x, y) \mid -3 \leq x \leq 3, -\sqrt{9 - x^2} \leq y \leq \sqrt{9 - x^2}\},$$

and

$$\begin{aligned}\mathrm{vol}(R) &= \int_{-3}^{3} \int_{-\sqrt{9-x^2}}^{\sqrt{9-x^2}} (4 - y)\, \mathrm{d}y\, \mathrm{d}x = \int_{-3}^{3} 8\sqrt{9 - x^2}\, \mathrm{d}x \\ &= 8\left(\frac{x}{2}\sqrt{9 - x^2} + \frac{9}{2} \arcsin \frac{x}{3} \right)\Bigg|_{-3}^{3} = 36\pi. \end{aligned}$$

(ii) We have so far dealt with integration over bounded regions. More generally, if $A \subset \mathbb{R}^n$ and $f : A \to \mathbb{R}$, we define

$$\int_A f = \lim_{k \to \infty} \int_{A \cap [-k,k]^n} f,$$

provided the sequence on the right converges.

As a simple application that will be used in the proof of Sard's theorem, suppose $A \subset \mathbb{R}^n = \mathbb{R} \times \mathbb{R}^{n-1}$ is a set such that its intersection $A_t = A \cap (\{t\} \times \mathbb{R}^{n-1})$ with each hyperplane $u^1 = t$, $t \in \mathbb{R}$, has ($(n-1)$-dimensional) volume zero. Then A has volume zero. To see this, notice first of all that by the above definition, A may be assumed to be bounded: indeed, if the claim holds for bounded sets, and A is

an unbounded set satisfying the above hypothesis, then $\mathrm{vol}(A \cap [-k,k]^n) = 0$ for every k, and so is the limit as $k \to \infty$. So consider a box $B = [a_1, b_1] \times B_1$ in $\mathbb{R}^n$ that contains A. Then

$$\mathrm{vol}(A) = \int_A 1 = \int_B \chi_A = \int_{a_1}^{b_1} \left(\int_{B_1} \chi_{A_t} \right) \mathrm{d}t$$

is zero since by assumption each $\int_{B_1} \chi_{A_t} = 0$.

(ii) In Section 1.6, we defined the volume of the parallelepiped P spanned by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^n$ to be the absolute value of the determinant of the matrix that has $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ as columns. This may be rephrased as follows: notice that $P = L(B)$, where $B = [0,1]^n$, and $L : \mathbb{R}^n \to \mathbb{R}^n$ is the linear transformation determined by $L\boldsymbol{e}_i = \boldsymbol{x}_i$, $i = 1, \ldots, n$. The above formula for the volume of P may therefore be written as

$$\mathrm{vol}\, L(B) = |\det L| \cdot \mathrm{vol}(B). \tag{4.3.2}$$

For the sake of consistency, it must be checked that this definition coincides with the definition of volume as an integral. We will establish, somewhat more generally, that (4.3.2) holds for any box $B = [a_1, b_1] \times \cdots \times [a_n, b_n]$ in $\mathbb{R}^n$. First of all, notice that L may be assumed to be an isomorphism, since otherwise both sides of the above identity vanish. Next, since any isomorphism is a composition of elementary transformations (see Appendix B) and the determinant of a composition is the product of the determinants, it is enough to prove (4.3.2) for elementary transformations. Recall that these transformations come in three flavors:

(1) $L\boldsymbol{e}_i = \boldsymbol{e}_j$, $L\boldsymbol{e}_j = \boldsymbol{e}_i$ for some $1 \le i < j \le n$, and $L\boldsymbol{e}_k = e_k$ if $k \neq i, j$;
(2) There exists $1 \le i \le n$ and $\alpha \neq 0$ such that $L\boldsymbol{e}_i = \alpha \boldsymbol{e}_i$, and $L\boldsymbol{e}_j = \boldsymbol{e}_j$ for $j \neq i$;
(3) There exist distinct $i, j \in \{1, \ldots, n\}$, $\alpha \in \mathbb{R}$, such that $L\boldsymbol{e}_i = \boldsymbol{e}_i + \alpha \boldsymbol{e}_j$, and $L\boldsymbol{e}_k = \boldsymbol{e}_k$ when $k \neq i$.

If L is of type 1, then $L(B)$ is obtained by interchanging edges i and j in B, and both have the same volume because the latter equals the product of the lengths of its edges. Since $|\det L| = 1$, (4.3.2) holds in this case. If L is of type 2, then $L(B)$ is a box that has the same edges as B except for the i-th one, which has length $|\alpha|$ times that of B; but $\alpha = \det L$, so again (4.3.2) is true in this case. Finally, if L is of type 3, then by what was established for type 1, we may assume that $L\boldsymbol{e}_i = \boldsymbol{e}_i$ if $i \neq n-1$ and $L\boldsymbol{e}_{n-1} = \boldsymbol{e}_{n-1} + \alpha \boldsymbol{e}_n$. Set

$$B_1 = [a_1, b_1] \times \ldots [a_{n-2}, b_{n-2}], \qquad B_2 = [a_{n-1}, b_{n-1}] \times [a_n, b_n],$$

so that $B = B_1 \times B_2$. By Fubini's theorem,

$$\mathrm{vol}\, L(B) = \int_{L(B_1 \times B_2)} 1 = \int_{B_1 \times L(B_2)} 1 = \mathrm{vol}\, B_1 \cdot \mathrm{vol}\, L(B_2),$$

and since $\det L = 1$, it remains to show that B_2 and $L(B_2)$ have the same volume; i.e., that (4.3.2) holds for a type 3 elementary transformation $L : \mathbb{R}^2 \to \mathbb{R}^2$. But if $B = [a, b] \times [c, d]$, then $L(B)$ is the projectable region

$$L(B) = \{(x, y) \mid a \le x \le b,\ \alpha x + c \le y \le \alpha x + d\},$$

and

$$\operatorname{vol} L(B) = \int_a^b \left(\int_{\alpha x + c}^{\alpha x + d} 1 \right) \mathrm{d}x = \int_a^b (d - c) = \operatorname{vol} B.$$

4.4 Sard's theorem

Recall from the section on Taylor polynomials that $\boldsymbol{a} \in U \subset \mathbb{R}^n$ is said to be a *critical point* of a map $\boldsymbol{f} : U \to \mathbb{R}^m$ if $D\boldsymbol{f}(\boldsymbol{a}) : \mathbb{R}^n \to \mathbb{R}^m$ is either not onto or does not exist, and that in this case, $\boldsymbol{f}(\boldsymbol{a})$ is called a *critical value* of $\boldsymbol{f}$. As an application of Fubini's theorem, we discuss a remarkable result of Sard, which asserts that the set of critical values of a C^k map $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ has measure zero, if k is large enough (depending on n and m). This is of course not surprising (and not difficult to prove) when $n < m$: after all, even in the worst case scenario when $\boldsymbol{f}$ has maximal rank everywhere, the set of critical values is all of $\boldsymbol{f}(U)$, which is locally an immersed n-dimensional submanifold of $\mathbb{R}^m$. Our approach of the argument follows that of [11]. It has the advantage of being relatively short, but comes at additional cost, namely $\boldsymbol{f}$ will be assumed to have continuous partial derivatives of any order.

Theorem 4.4.1 (Sard). *Let $U \subset \mathbb{R}^n$, and $\boldsymbol{f} : U \to \mathbb{R}^m$ a C^∞ map; i.e., the component functions $u^i \circ \boldsymbol{f}$ have continuous partial derivatives of any order. Then the set of critical values of $\boldsymbol{f}$ has measure zero.*

Proof. The argument will be by induction on the dimension n of the domain, beginning with $n = 0$. Recall that the zero-dimensional space $\mathbb{R}^0 = \{0\}$, so the statement is true in this case. Assume then that the statement holds in dimensions less than n.
Denote by C the set of critical points of $\boldsymbol{f}$, and by C_k the subset consisting of those points where all partial derivatives of (the components of) $\boldsymbol{f}$ of order $\le k$ vanish, $k \ge 1$. Thus, $C \supset C_k \supset C_{k+1}$, $k \in \mathbb{N}$. We will repeatedly make use of the following observation: in order to show that a given set $\boldsymbol{f}(A)$ has measure zero, it is enough to show that every point in A admits a neighborhood V such that $\boldsymbol{f}(V \cap A)$ has measure zero. This is because the resulting cover of A contains a countable subcover by Theorem 1.7.5, and a countable union of measure zero sets has measure zero.

Claim 1: $\boldsymbol{f}(C \setminus C_1)$ has measure zero. To see this, consider $\boldsymbol{p} \in C \setminus C_1$. By assumption, $D_i f^j(\boldsymbol{p}) \neq 0$ for some $1 \le i, j \le n$. It may be assumed that $i = j = 1$, since interchanging two coordinates in the domain or in the range of $\boldsymbol{f}$ amounts to composing $\boldsymbol{f}$ on the right

or on the left with a diffeomorphism $\boldsymbol{F}$; in the former case, $\boldsymbol{f} \circ \boldsymbol{F}$ has the same critical values as $\boldsymbol{f}$, and in the latter case, $(\boldsymbol{F} \circ \boldsymbol{f})(C)$ has measure zero if and only if $\boldsymbol{f}(C)$ has measure zero. Now, the map $\boldsymbol{h} : U \to \mathbb{R}^n$ given by

$$\boldsymbol{h}(\boldsymbol{a}) = (f^1(\boldsymbol{a}), a_2, \ldots, a_n), \qquad a_i = u^i(\boldsymbol{a}),$$

has rank n at $\boldsymbol{p}$: indeed, its Jacobian at that point is a matrix whose entries below the diagonal all vanish, so that its determinant equals the product of the diagonal elements, namely $D_1 f^1(\boldsymbol{p}) \neq 0$. By the inverse function theorem, there exists a neighborhood V of $\boldsymbol{p}$ in U such that the restriction $\boldsymbol{h} : V \to \boldsymbol{h}(V)$ is a diffeomorphism. Set $\boldsymbol{g} := \boldsymbol{f} \circ \boldsymbol{h}^{-1} : \boldsymbol{h}(V) \to \mathbb{R}^m$. Since $\boldsymbol{h}$ is a diffeomorphism, the set A of critical values of $\boldsymbol{g}$ coincides with the set of critical values of the restriction of $\boldsymbol{f}$ to V; i.e., $A = \boldsymbol{f}(V \cap C)$, and it remains to show that A has measure zero. Notice that $u^1 \circ \boldsymbol{h} = f^1$, so that

$$u^1 \circ \boldsymbol{g} = u^1 \circ \boldsymbol{f} \circ \boldsymbol{h}^{-1} = f^1 \circ \boldsymbol{h}^{-1} = u^1 \circ \boldsymbol{h} \circ \boldsymbol{h}^{-1} = u^1.$$

In other words, $\boldsymbol{g}$ maps each hyperplane $u^1 =$ constant into itself. Denote by $\boldsymbol{g}_\alpha$ the restriction of $\boldsymbol{g}$ to the (portion of the) hyperplane

$$H_\alpha = \{\boldsymbol{p} \in U \mid u^1(\boldsymbol{p}) = \alpha\}, \qquad \alpha \in \mathbb{R}.$$

Observe that for each $\boldsymbol{p} \in H_\alpha$,
(1) the Jacobian matrix of $\boldsymbol{g}$ at $\boldsymbol{p}$ has $\boldsymbol{e}_1^T$ as its first row, and
(2) $D_i g^j(\boldsymbol{p}) = D_i g_\alpha^j(\boldsymbol{p})$ for all $i, j \neq 1$.

The latter statement says that if $\boldsymbol{g}_\alpha$ is viewed as a map from $\mathbb{R}^{n-1} \cong H_\alpha$ to itself, then the Jacobian of $\boldsymbol{g}_\alpha$ at $\boldsymbol{p}$ is obtained by deleting the first row and column from that of $\boldsymbol{g}$ at $\boldsymbol{p}$:

$$D\boldsymbol{g}(\boldsymbol{p}) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ * & & & \\ \vdots & & D\boldsymbol{g}_\alpha(\boldsymbol{p}) & \\ * & & & \end{bmatrix}$$

Together with the former, this means that the columns of $[D\boldsymbol{g}(\boldsymbol{p})]$ are linearly dependent if and only if those of $[D\boldsymbol{g}_\alpha(\boldsymbol{p})]$ are: for if M^i denotes the i-th column of the first matrix, and M_α^i that of the second, then the first entry of the column vector $\sum a_i M^i$ is a_1. Thus, if $\sum a_i M^i = \boldsymbol{0}$, then $a_1 = 0$ and $\sum a_i M^i = \sum_{i>1} a_i M^i \in \mathbb{R}^m$. This vector has zero as its first entry, and deleting this entry yields $\sum_i a_i M_\alpha^{i-1} = \boldsymbol{0} \in \mathbb{R}^{m-1}$. Summarizing, $\boldsymbol{p} \in H_\alpha$ is a critical point of $\boldsymbol{g}_\alpha$ iff it is a critical point of $\boldsymbol{g}$. According to our induction hypothesis, the set of critical values of $\boldsymbol{g}_\alpha$ has measure zero, so that the intersection of the set of critical values of $\boldsymbol{g}$ with each hyperplane $u^j = \alpha$ has measure zero. By Examples 4.3.2 (ii), A itself then has measure zero.

Claim 2: $\boldsymbol{f}(C_k \setminus C_{k+1})$ has measure zero for $k \geq 1$. Given any $\boldsymbol{p} \in C_k \setminus C_{k+1}$, we shall once again exhibit an open neighborhood V of $\boldsymbol{p}$ such that $\boldsymbol{f}(V \cap C_k)$ has measure zero. By

assumption, there exists a partial derivative of (a component function of) $\boldsymbol{f}$ of order k, which we denote by φ, such that $\varphi(\boldsymbol{p}) = 0$ but $D_i\varphi(\boldsymbol{p}) \neq 0$ for some i between 1 and n. As before, the map $\boldsymbol{h} : U \to \mathbb{R}^n$, given by

$$\boldsymbol{h}(\boldsymbol{a}) = (a_1, \ldots, a_{i-1}, \varphi(\boldsymbol{a}), a_{i+1}, \ldots, a_n),$$

has rank n at $\boldsymbol{p}$ and there exists a neighborhood V of $\boldsymbol{p}$ on which the restriction $\boldsymbol{h} : V \to \boldsymbol{h}(V)$ is a diffeomorphism. Set once again $\boldsymbol{g} = \boldsymbol{f} \circ \boldsymbol{h}^{-1} : \boldsymbol{h}(V) \to \mathbb{R}^m$. Since φ vanishes when restricted to $V \cap C_k$, $\boldsymbol{h}(V \cap C_k)$ lies in the coordinate hyperplane $u^i = 0$. If $\boldsymbol{g}_0$ denotes the restriction of $\boldsymbol{g}$ to this hyperplane, then every point of $\boldsymbol{h}(V \cap C_k)$ is a critical point of $\boldsymbol{g}_0$ because $V \cap C_k$ is contained in the set of critical points of $\boldsymbol{f}$. By the induction hypothesis, the critical values of $\boldsymbol{g}_0$ form a set of measure zero. Thus, $\boldsymbol{f}(V \cap C_k) = \boldsymbol{g}_0(\boldsymbol{h}(V \cap C_k))$ has measure zero, as claimed.

Claim 3: $\boldsymbol{f}(C_k)$ has measure zero if k is large enough. To see this, consider a box $B = \Pi_{i=1}^n [a_i, b_i]$ with edges of common length $R = b_i - a_i$ small enough that B is contained in U. It suffices to show that if $k > (n/m) - 1$, then $\boldsymbol{f}(C_k \cap B)$ has measure zero. Applying Remark 2.7.1 to each component function of $\boldsymbol{f}$, we see that there exists $\beta > 0$ such that

$$|R(\boldsymbol{x}, \boldsymbol{h})| := |\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x})| \leq \beta |\boldsymbol{h}|^{k+1}, \qquad \boldsymbol{x} \in C_k \cap B, \quad \boldsymbol{x} + \boldsymbol{h} \in B. \tag{4.4.1}$$

Given any natural number $l > 1$, partition each $[a_i, b_i]$ into l intervals of equal length R/l. The corresponding partition of B consists of l^n subboxes with diameter $\sqrt{n}R/l$. Let $\boldsymbol{x} \in C_k \cap B$, and $\tilde{B}$ be a subbox containing $\boldsymbol{x}$. Then any other point in $\tilde{B} \cap C_k$ is of the form $\boldsymbol{x} + \boldsymbol{h}$, where $|\boldsymbol{h}| \leq \sqrt{n}R/l$. It follows from (4.4.1) that $\boldsymbol{f}(\tilde{B} \cap C_k)$ is contained in a box centered at $\boldsymbol{f}(\boldsymbol{x})$ with all sides of common length $2\beta(\sqrt{n}R/l)^{k+1}$. Since this box lives in $\mathbb{R}^m$, it has volume $\alpha/l^{m(k+1)}$, where α is a constant that does not depend on l. Thus, $\boldsymbol{f}(B \cap C_k)$ is contained in l^n boxes with total volume

$$V \leq \alpha l^{n-m(k+1)}.$$

The exponent of l is a negative integer, so that V can be made arbitrarily small by choosing l large enough. As observed earlier, this implies that $\boldsymbol{f}(C_k)$ has measure zero. To conclude, set $C_0 := C$. If $k > (n/m) - 1$, then

$$\boldsymbol{f}(C) = \left(\bigcup_{i=1}^{k} \boldsymbol{f}(C_{i-1} \setminus C_i) \right) \cup \boldsymbol{f}(C_k)$$

is a finite union of sets of measure zero. This completes the proof. □

Remark 4.4.1. Sard's theorem holds more generally for maps between manifolds: A subset A of a manifold M^k is said to have measure zero if A can be written as a countable union $\cup A_n$ where each A_n lies in the domain of some chart $(U_n, \boldsymbol{x}_n)$, and $\boldsymbol{x}_n(A_n)$ has measure zero in $\mathbb{R}^k$. Now, if $\boldsymbol{f} : M \to N$ is a $\mathcal{C}^\infty$ map between manifolds M and N, then for any charts $(U, \boldsymbol{x})$ of M and $(V, \boldsymbol{y})$ of N, the set of critical values of $\boldsymbol{y} \circ \boldsymbol{f} \circ \boldsymbol{x}^{-1}$ has measure zero. It follows that the set of critical values of $\boldsymbol{f}$ has measure zero.

4.5 The change of variables theorem

For functions of a single variable, the chain rule implies the well-known change of variables theorem:

If $g : [a,b] \to \mathbb{R}$ is continuously differentiable, and f is a continuous function whose domain contains $g([a,b])$, then

$$\int_{g(a)}^{g(b)} f = \int_a^b (f \circ g)g'. \tag{4.5.1}$$

The proof is easy: if F is an antiderivative of f, then $F \circ g$ is an antiderivative of $(f \circ g)g'$, so that both sides equal $F(g(b)) - F(g(a))$. Before stating the generalization of this theorem to higher dimensions – let alone proving it, which turns out to be much more involved – we observe that it may be reformulated as follows:

$$\int_{g(a,b)} f = \int_{(a,b)} (f \circ g)|g'|, \tag{4.5.2}$$

at least if g is one-to-one. Indeed, g is then either increasing, in which case $g' \geq 0$ and both identities are the same, or decreasing, so that $g(a,b) = (g(b), g(a))$, $g' \leq 0$, and both sides of (4.5.2) are the negative of those in (4.5.1).

The higher-dimensional analogue of (4.5.2) is given by the following:

Theorem 4.5.1. *Let $U \subset \mathbb{R}^n$ be bounded, and suppose $\boldsymbol{g} : U \to \mathbb{R}^n$ is a continuously differentiable map, which is injective on an open subset whose complement in U has measure zero. If $f : \boldsymbol{g}(U) \to \mathbb{R}$ is integrable, then*

$$\int_{\boldsymbol{g}(U)} f = \int_U (f \circ \boldsymbol{g})|\det D\boldsymbol{g}|.$$

The proof will be handled in a series of steps. First of all, U may be assumed to be open and $\boldsymbol{g}$ injective on U by Remarks 4.2.3 (ii) together with the fact that the image $\boldsymbol{g}(A)$ of a set A of measure zero has measure zero if $\boldsymbol{g}$ is continuous and A lies in some compact set, see Exercise 4.3. Similarly, by Sard's theorem, we may suppose that the Jacobian determinant of $\boldsymbol{g}$ is nowhere zero. Next, we point out two observations that will be used repeatedly in the proof:

Observation 4.5.1. The class of maps $\boldsymbol{g}$ for which the theorem holds is closed under composition.

To see this, let $\boldsymbol{g}_1 : U \to \mathbb{R}^n$, $\boldsymbol{g}_2 : V \to \mathbb{R}^n$, where $\boldsymbol{g}_1(U) \subset V$. If the claim is true for each of these maps, then

$$\begin{aligned}\int\limits_{(\boldsymbol{g}_2\circ\boldsymbol{g}_1)(U)} f &= \int\limits_{\boldsymbol{g}_2(\boldsymbol{g}_1(U))} f = \int\limits_{\boldsymbol{g}_1(U)} (f\circ\boldsymbol{g}_2)|\det D\boldsymbol{g}_2| \\ &= \int\limits_U (f\circ\boldsymbol{g}_2\circ\boldsymbol{g}_1)|((\det D\boldsymbol{g}_2)\circ\boldsymbol{g}_1)|\det D\boldsymbol{g}_1| \\ &= \int\limits_U (f\circ\boldsymbol{g}_2\circ\boldsymbol{g}_1)|\det D(\boldsymbol{g}_2\circ\boldsymbol{g}_1)|.\end{aligned}$$

Observation 4.5.2. It suffices to prove that for any $\boldsymbol{a} \in U$, there exists an open neighborhood $W \subset U$ of $\boldsymbol{a}$ for which the theorem holds.

Indeed, if the statement is true in this case, then there exists an admissible open cover $\{W_k \mid k \in A \subset \mathbb{N}\}$ of U such that the theorem holds on each W_k. Then $\{\boldsymbol{g}(W_k) \mid k \in A\}$ is an admissible open cover of $\boldsymbol{g}(U)$. If $\{\varphi_k \mid k \in A\}$ is a partition of unity subordinate to $\{\boldsymbol{g}(W_k) \mid k \in A\}$, then again by hypothesis,

$$\int\limits_{\boldsymbol{g}(W_k)} \varphi_k f = \int\limits_{W_k} (\varphi_k f)\circ\boldsymbol{g}|\det D\boldsymbol{g}|. \tag{4.5.3}$$

Now, φ_k vanishes outside $\boldsymbol{g}(W_k)$, so that, since $\boldsymbol{g}$ is injective, $\varphi_k\circ\boldsymbol{g}$ is zero outside W_k (and in particular, $\{\varphi_k\circ\boldsymbol{g} \mid k \in A\}$ is a partition of unity subordinate to U). Thus, (4.5.3) becomes

$$\int\limits_{\boldsymbol{g}(U)} \varphi_k f = \int\limits_U (\varphi_k f)\circ\boldsymbol{g}|\det D\boldsymbol{g}|. \tag{4.5.4}$$

As noted above, the collection $\{\varphi_k\circ\boldsymbol{g} \mid k \in A\}$ is a partition of unity subordinate to U. The definition of the generalized integral then implies that

$$\int\limits_{\boldsymbol{g}(U)} f = \sum_{k\in A}\int\limits_{\boldsymbol{g}(U)} \varphi_k f = \sum_{k\in A}\int\limits_U (\varphi_k\circ\boldsymbol{g})(f\circ\boldsymbol{g})|\det D\boldsymbol{g}| = \int\limits_U (f\circ\boldsymbol{g})|\det D\boldsymbol{g}|.$$

Now that these observations are out of the way, we proceed with the proof of the change of variables formula by first considering constant functions. Since the integral of a constant function c over a set is c times the volume of the set, it suffices to consider the constant function 1:

Lemma 4.5.1. *Let U be open in $\mathbb{R}^n$, $\boldsymbol{g} : U \to \mathbb{R}^n$ a one-to-one continuously differentiable map with nowhere zero Jacobian determinant. Then*

$$\operatorname{vol}\boldsymbol{g}(U) = \int\limits_{\boldsymbol{g}(U)} 1 = \int\limits_U |\det D\boldsymbol{g}|.$$

Proof. We proceed by induction on n. The case $n = 1$ follows from (4.5.2), so assume the claim holds in dimension $n - 1$. By Observation 4.5.2, it suffices to establish the

result for some neighborhood of an arbitrary point $\boldsymbol{a} \in U$. Furthermore, we may suppose that $[D\boldsymbol{g}(\boldsymbol{a})] = I_n$: if the lemma is valid for maps with Jacobian matrix equal to the identity at $\boldsymbol{a}$, then it must hold for the map $D\boldsymbol{g}(\boldsymbol{a})^{-1} \circ \boldsymbol{g}$. On the other hand, it is also true for the linear transformation $D\boldsymbol{g}(\boldsymbol{a})$ by Examples and Remarks 4.3.2 (iv). By Observation 4.5.1, it then holds for $\boldsymbol{g} = D\boldsymbol{g}(\boldsymbol{a}) \circ (D\boldsymbol{g}(\boldsymbol{a})^{-1} \circ \boldsymbol{g})$.
In order to use the induction hypothesis, we rewrite $\boldsymbol{g}$ as a composition $\boldsymbol{f} \circ \boldsymbol{h}$ as follows: let $\pi_1 : \mathbb{R}^n = \mathbb{R}^{n-1} \times \mathbb{R} \to \mathbb{R}^{n-1}$ denote projection, and define $\boldsymbol{h} : U \to \mathbb{R}^n$ by $\boldsymbol{h} = (\pi_1 \circ \boldsymbol{g}, u^n)$; i.e., for $\boldsymbol{x} = (x_1, \dots, x_n)$, $\boldsymbol{h}(\boldsymbol{x}) = (g^1(\boldsymbol{x}), \dots, g^{n-1}(\boldsymbol{x}), x_n)$. Since $D\boldsymbol{g}(\boldsymbol{a}) = 1_{\mathbb{R}^n}$, $D\boldsymbol{h}(\boldsymbol{a}) = 1_{\mathbb{R}^n}$ and $\boldsymbol{h}$ is invertible in some neighborhood W of $\boldsymbol{a}$. Set $\tilde{W} = \boldsymbol{h}(W)$, and define $\boldsymbol{f} : \tilde{W} \to \mathbb{R}^n$ by $\boldsymbol{f} = (\pi_1, g^n \circ (\boldsymbol{h}_{|W})^{-1})$. Then on W

$$\boldsymbol{f} \circ \boldsymbol{h} = (\pi_1 \circ \boldsymbol{h}, g^n) = (\pi_1 \circ \boldsymbol{g}, g^n) = \boldsymbol{g}.$$

Furthermore,

$$D\boldsymbol{f}(\boldsymbol{h}(\boldsymbol{a})) = D\boldsymbol{f}(\boldsymbol{h}(\boldsymbol{a})) \circ D\boldsymbol{h}(\boldsymbol{a}) = D\boldsymbol{g}(\boldsymbol{a}) = 1_{\mathbb{R}^n},$$

so that $\boldsymbol{f}$ is injective with Jacobian matrix of rank n on some neighborhood V of $\boldsymbol{h}(\boldsymbol{a})$. Set $U = \boldsymbol{h}^{-1}(V)$, restrict $\boldsymbol{h}$ to U and $\boldsymbol{f}$ to V. Then $\boldsymbol{h} : U \to V$, $\boldsymbol{f} : V \to \boldsymbol{f}(V)$ are diffeomorphisms, $\boldsymbol{g}_{|U} = \boldsymbol{f} \circ \boldsymbol{h}$, and it suffices, by the two observations above, to establish the claim for $\boldsymbol{f}$ and $\boldsymbol{h}$ on open boxes containing $\boldsymbol{h}(\boldsymbol{a})$ and $\boldsymbol{a}$ respectively. We begin with $\boldsymbol{h}$: consider an open box $B = B_1 \times (a_n, b_n) \subset U$ containing $\boldsymbol{a}$, where $B_1 = \pi_1(B)$. By Examples 4.2.2 (iii) and Exercise 1.29, $\boldsymbol{h}(B \times [a_n, b_n])$ and $\boldsymbol{h}(B \times \{t\})$ are Jordan-measurable, and Fubini's theorem implies

$$\int\limits_{\boldsymbol{h}(B)} 1 = \int\limits_{a_n}^{b_n} \left(\int\limits_{\boldsymbol{h}(B_1 \times \{t\})} 1 \, \mathrm{d}\boldsymbol{x} \right) \mathrm{d}t. \tag{4.5.5}$$

Now, for each $t \in (a_n, b_n)$, the map $\boldsymbol{h}_t : B_1 \to \mathbb{R}^{n-1}$, where $\boldsymbol{h}_t(\boldsymbol{x}) = (\pi_1 \circ \boldsymbol{g})(\boldsymbol{x}, t)$, is injective, and

$$[D\boldsymbol{h}(\boldsymbol{x}, t)] = \begin{bmatrix} & & & * \\ & D\boldsymbol{h}_t(\boldsymbol{x}) & & \vdots \\ & & & * \\ 0 & \dots & 0 & 1 \end{bmatrix},$$

so that $\det D\boldsymbol{h}_t(\boldsymbol{x}) \neq 0$. By the induction hypothesis,

$$\int\limits_{\boldsymbol{h}_t(B_1)} 1 \, \mathrm{d}\boldsymbol{x} = \int\limits_{B_1} |D\boldsymbol{h}_t| \, \mathrm{d}\boldsymbol{x} = \int\limits_{B_1} |D\boldsymbol{h}(\boldsymbol{x}, t)| \, \mathrm{d}\boldsymbol{x}.$$

But $\boldsymbol{h}_t(B_1) = \boldsymbol{h}(B_1 \times \{t\})$, so that (4.5.5) now yields

$$\int\limits_{\boldsymbol{h}(B)} 1 = \int\limits_{a_n}^{b_n} \left(\int\limits_{B_1} |D\boldsymbol{h}(\boldsymbol{x}, t)| \, \mathrm{d}\boldsymbol{x} \right) \mathrm{d}t = \int\limits_{B} |D\boldsymbol{h}|.$$

This proves the claim for $\boldsymbol{h}$, and we now turn our attention to $\boldsymbol{f}$. Let $B = B_1 \times (a_n, b_n)$ be a box in $\mathbb{R}^{n-1} \times \mathbb{R}$, and for each $\boldsymbol{x} \in B$, define $f_{\boldsymbol{x}} : (a_n, b_n) \to \mathbb{R}$ by $f_{\boldsymbol{x}}(t) = f^n(\boldsymbol{x}, t) = (g^n \circ \boldsymbol{h}^{-1})(\boldsymbol{x}, t)$. Since $\boldsymbol{f}(\boldsymbol{x}, t) = (\boldsymbol{x}, f_{\boldsymbol{x}}(t))$,

$$[D\boldsymbol{f}(\boldsymbol{x}, t)] = \begin{bmatrix} & & & 0 \\ & I_{n-1} & & \vdots \\ & & & 0 \\ * & \dots & * & f'_{\boldsymbol{x}}(t) \end{bmatrix},$$

and in particular, $|\det D\boldsymbol{f}(\boldsymbol{x}, t)| = |f'_{\boldsymbol{x}}(t)|$. Now,

$$\boldsymbol{f}(B) = \bigcup_{\boldsymbol{x} \in B_1} \{\boldsymbol{x}\} \times f_{\boldsymbol{x}}(a_n, b_n),$$

so that

$$\begin{aligned} \int_{\boldsymbol{f}(B)} 1 &= \int_{B_1} \left(\int_{f_{\boldsymbol{x}}(a_n, b_n)} 1 \, \mathrm{d}t \right) \mathrm{d}\boldsymbol{x} = \int_{B_1} \left(\int_{(a_n, b_n)} |f'_{\boldsymbol{x}}(t)| \, \mathrm{d}t \right) \\ &= \int_{B_1} \left(\int_{a_n}^{b_n} |\det D\boldsymbol{f}(\boldsymbol{x}, t)| \, \mathrm{d}t \right) \mathrm{d}\boldsymbol{x} \\ &= \int_B |\det D\boldsymbol{f}|. \end{aligned}$$

This completes the proof of the lemma. □

Proof of Theorem 4.5.1. By Observation 4.5.2, it suffices to show that for any $\boldsymbol{a} \in U$, there exists a neighborhood $W \subset U$ of $\boldsymbol{a}$ such that the theorem holds on W. So consider an open box B that contains $\boldsymbol{g}(\boldsymbol{a})$, and let $W = \boldsymbol{g}^{-1}(B)$. If P is a partition of B, then

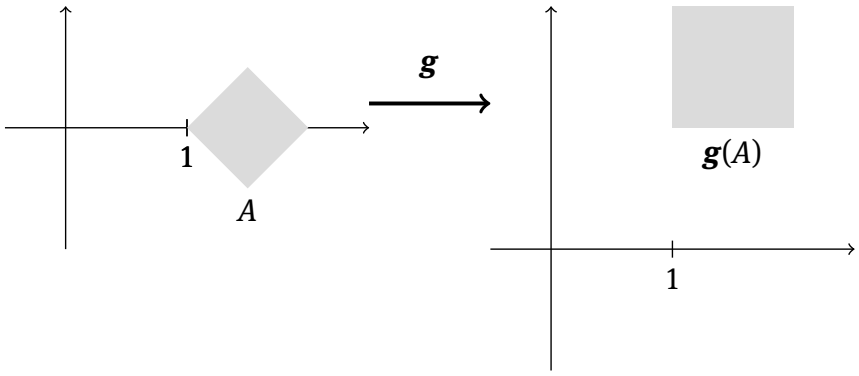
$$\begin{aligned} L(f, P) &= \sum_{\tilde{B} \in P} m_{\tilde{B}}(f) \operatorname{vol}(\tilde{B}) = \sum_{\tilde{B} \in P} m_{\tilde{B}}(f) \int_{\tilde{B}^0} 1 \\ &= \sum_{\tilde{B} \in P} m_{\tilde{B}}(f) \int_{\boldsymbol{g}^{-1}(\tilde{B}^0)} (1 \circ \boldsymbol{g}) |\det D\boldsymbol{g}| \\ &= \sum_{\tilde{B} \in P} \int_{\boldsymbol{g}^{-1}(\tilde{B}^0)} (m_{\tilde{B}}(f)) \circ \boldsymbol{g} |\det D\boldsymbol{g}| \\ &\leq \int_W (f \circ \boldsymbol{g}) |\det D\boldsymbol{g}|, \end{aligned}$$

and therefore $\int_{\boldsymbol{g}(W)} f \leq \int_W (f \circ \boldsymbol{g}) |\det D\boldsymbol{g}|$. A similar argument using $U(f, P)$ instead of $L(f, P)$ and $M_{\tilde{B}}(f)$ instead of $m_{\tilde{B}}(f)$ implies that the inequality holds in the other direction. Thus,

$$\int_{\boldsymbol{g}(W)} f = \int_W (f \circ \boldsymbol{g}) |\det D\boldsymbol{g}|.$$

This shows that the theorem holds for W and establishes the result. □

Example 4.5.1. Consider the integral $\int_A (x - y)/(x + y)\,\mathrm{d}x\,\mathrm{d}y$, where A denotes the region bounded by the square with vertices $(1,0)$, $(3/2,1/2)$, $(2,0)$ and $(3/2,-1/2)$. Although it can be evaluated with Fubini's theorem (notice that the region is projectable), the computation is lengthy.



Alternatively, observe that $\boldsymbol{g} : \mathbb{R}^2 \to \mathbb{R}^2$, where $\boldsymbol{g}(x,y) = (x - y, x + y)$, is an isomorphism, so we may apply the change of variables theorem to the function $f(u,v) = u/v$ to obtain

$$\int_A \frac{x-y}{x+y}\,\mathrm{d}x\,\mathrm{d}y = \int_A (f \circ \boldsymbol{g}) = \frac{1}{|\det D\boldsymbol{g}|} \int_A (f \circ \boldsymbol{g})|\det D\boldsymbol{g}|$$
$$= \frac{1}{|\det D\boldsymbol{g}|} \int_{\boldsymbol{g}(A)} f.$$

Now, the Jacobian determinant of $\boldsymbol{g}$ equals 2, and A is bounded by the lines $x - y = 1$, $x - y = 2$, $x + y = 1$, and $x + y = 2$. Thus, $\boldsymbol{g}(A) = [1,2] \times [1,2]$, and

$$\int_A \frac{x-y}{x+y}\,\mathrm{d}x\,\mathrm{d}y = \frac{1}{2}\int_1^2 \left(\int_1^2 \frac{u}{v}\,\mathrm{d}u \right) \mathrm{d}v = \frac{1}{2}\int_1^2 u \ln 2\,\mathrm{d}u = \frac{3}{4}\ln 2.$$

4.6 Cylindrical and spherical coordinates

As an application of the results from the previous section, we discuss some changes of variables that are useful when working with regions that exhibit radial or spherical symmetry.

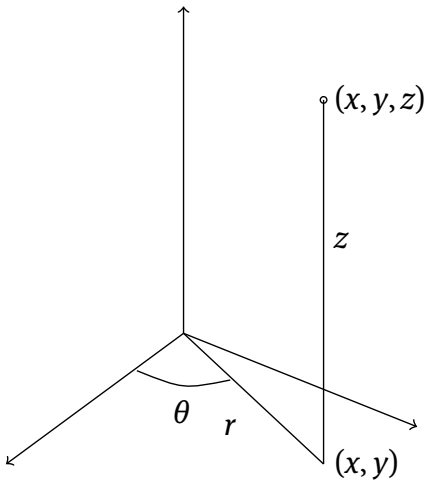
4.6.1 Cylindrical coordinates

Polar coordinates in $\mathbb{R}^2$ are given by the map $(r,\theta) : \mathbb{R}^2 \setminus \{\mathbf{0}\} \to \mathbb{R}^2$, where r assigns to a point $\boldsymbol{p}$ its distance to the origin $\mathbf{0}$, and θ is the angle in $[0, 2\pi)$ between the positive x-axis and the line segment $\mathbf{0}\boldsymbol{p}$. The polar angle θ requires some care in writing down,

because tan is only invertible when restricted to an interval of length π. Formally, $r = \sqrt{(u^1)^2 + (u^2)^2}$, whereas

$$\theta = \begin{cases} \arctan(u^2/u^1) & \text{if } u^1 > 0, u^2 \geq 0, \\ \pi/2 & \text{if } u^1 = 0 \text{ and } u^2 > 0, \\ \pi + \arctan(u^2/u^1) & \text{if } u^1 < 0, \\ 3\pi/2 & \text{if } u^1 = 0 \text{ and } u^2 < 0, \\ 2\pi + \arctan(u^2/u^1) & \text{if } u^1 > 0 \text{ and } u^2 < 0. \end{cases}$$

Polar coordinates extend to *cylindrical coordinates* in $\mathbb{R}^3$ via $(r, \theta, u^3) : \{(x, y, z) \in \mathbb{R}^3 \mid (x, y) \neq (0, 0)\} \to \mathbb{R}^3$.



Cylindrical coordinates

They are particularly useful when working with surfaces that are invariant under rotation about the z-axis. If $V \subset \mathbb{R}^3$ is a region that can be expressed as U in cylindrical coordinates, then $V = \boldsymbol{g}(U)$, where $\boldsymbol{g}$ is the inverse of cylindrical coordinates,

$$\begin{aligned} \boldsymbol{g} : [0, \infty) \times [0, 2\pi) \times \mathbb{R} &\to \mathbb{R}^3, \\ (r, \theta, z) &\mapsto (r\cos\theta, r\sin\theta, z). \end{aligned}$$

$\boldsymbol{g}$ is injective on $(0, \infty) \times [0, 2\pi) \times \mathbb{R}$ and differentiable on $(0, \infty) \times (0, 2\pi) \times \mathbb{R}$ with Jacobian

$$[D\boldsymbol{g}(r, \theta, z)] = \begin{bmatrix} \cos\theta & -r\sin\theta & 0 \\ \sin\theta & r\cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus, if $V \subset \mathbb{R}^3$ is expressed as U in cylindrical coordinates, i.e., if $V = \boldsymbol{g}(U)$, and if f is integrable on V, then

$$\int_V f = \int_U (f \circ \boldsymbol{g})(r, \theta, z) r \, dr \, d\theta \, dz. \tag{4.6.1}$$

Integration in polar coordinates is a special case of the above: suppose f is a function of 2 variables defined on a region V in the plane. Since the inverse $\boldsymbol{h}$ of polar coordinates (r,θ) equals $\pi_1 \circ \boldsymbol{g}_{|\mathbb{R}^2\times\{0\}}$, where $\pi_1 = (u^1,u^2) : \mathbb{R}^3 \to \mathbb{R}^2$ is projection, we have

$$[D\boldsymbol{h}(r,\theta)] = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}.$$

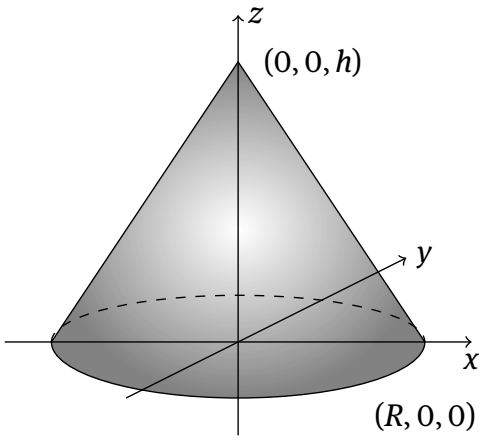
Thus, if $V \subset \mathbb{R}^2$ is expressed as U in polar coordinates, i.e., if $V = \boldsymbol{g}(U)$, and if f is integrable on V, then

$$\int_V f = \int_U (f \circ \boldsymbol{g})(r,\theta) r \, \mathrm{d}r \, \mathrm{d}\theta. \tag{4.6.2}$$

Examples 4.6.1. (i) A *right circular cone* of height h and radius R is the surface (the term is used loosely here) in $\mathbb{R}^3$ obtained by rotating the line segment

$$\{(t,0,(-h/R)t + h) \mid 0 \le t \le R\}$$

joining $(R,0,0)$ and $(0,0,h)$ about the z-axis.



A cone with base radius R and height h

The volume of this cone (or rather the volume of the 3-dimensional region bounded by the cone and the plane $z = 0$) is therefore equal to $\mathrm{vol}\, A$, where

$$A = \left\{(x,y,z) \mid 0 \le \sqrt{x^2 + y^2} \le R, \quad 0 \le z \le -\frac{h}{R}\sqrt{x^2 + y^2} + h\right\}.$$

Thus, the interior A^0 of A can be expressed as $\boldsymbol{g}(U)$, where U is the projectable region

$$U = \left\{(r,\theta,z) \mid 0 \le r < R, \quad 0 \le \theta < 2\pi, \quad 0 < z < h\left(1 - \frac{r}{R}\right)\right\}.$$

By Fubini's and the change of variables theorems,

$$\begin{aligned}\operatorname{vol}(A) = \operatorname{vol}(A^0) = \operatorname{vol}(\boldsymbol{g}(U)) &= \int_{\boldsymbol{g}(U)} 1 = \int_U r\,\mathrm{d}z\,\mathrm{d}r\,\mathrm{d}\theta \\ &= \int_0^{2\pi}\int_0^R \int_0^{h(1-\frac{r}{R})} r\,\mathrm{d}z\,\mathrm{d}r\,\mathrm{d}\theta = 2\pi \int_0^R \left(rh - r^2\frac{h}{R}\right)\mathrm{d}r \\ &= \frac{\pi R^2 h}{3}.\end{aligned}$$

(ii) Suppose we are asked to determine the area of the planar region R bounded by the curve $(x^2 + y^2)^3 = (x^2 - y^2)^2$. This region is better visualized in polar coordinates, where the equation of the curve becomes

$$r^6 = (r^2\cos^2\theta - r^2\sin^2\theta)^2 = r^4\cos^2(2\theta), \quad \text{or } r = |\cos 2\theta|$$

since $r \geq 0$.

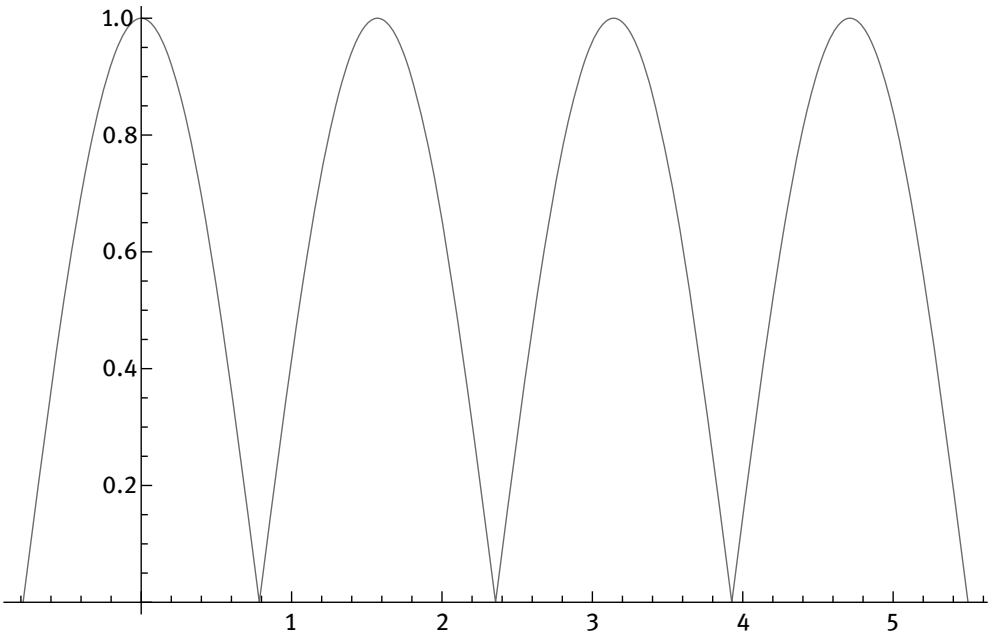


Fig. 4.4: $r = |\cos 2\theta|$

The curve is invariant under reflection in both coordinate axes (that is, the equation is unchanged when θ is replaced by $-\theta$ or by $\pi - \theta$), and represents the boundary of a 4-leaved rose. For example, the leaf

$$A = \{(r,\theta) \mid -\frac{\pi}{4} \leq \theta \leq \frac{\pi}{4}, \quad 0 \leq r \leq \cos 2\theta\}$$

in the region $-\pi/4 \leq \theta \leq \pi/4$ starts out at the origin when $\theta = -\pi/4$, and r increases until it reaches a maximum of 1 at $\theta = 0$, which corresponds to the point (1,0) in Cartesian coordinates. This is the bottom half of the leaf, and the top

is obtained by reflecting in the x-axis. Reflection in the y-axis yields a second leaf. Finally, notice that the curve is also invariant under reflection in the line $y = x$ because the equation is unchanged when θ is replaced by $\pi/2 - \theta$. Said reflection then reveals the two remaining leaves.

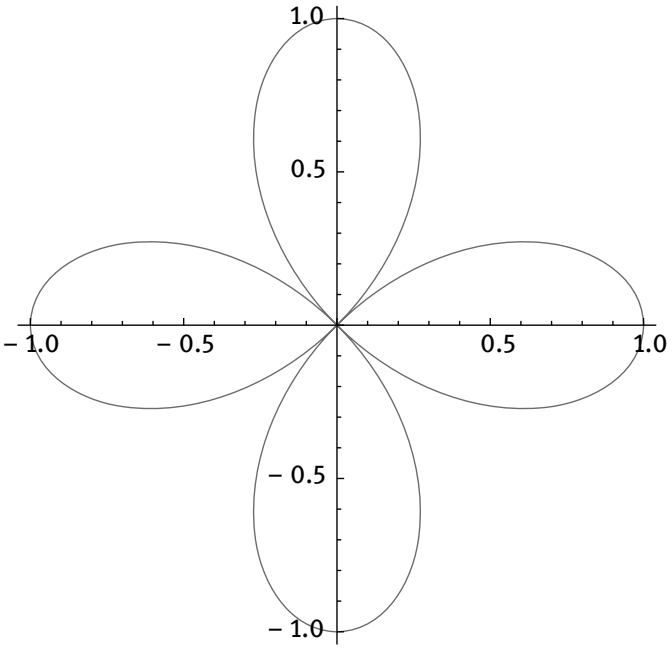


Fig. 4.5: $r = |\cos 2\theta|$ in polar coordinates

Thus,

$$\begin{aligned}\operatorname{area}(R) &= \int_R 1 \cdot r \, \mathrm{d}r \, \mathrm{d}\theta = 4 \int_A r \, \mathrm{d}r \, \mathrm{d}\theta = 4 \int_{-\pi/4}^{\pi/4} \int_0^{\cos 2\theta} r \, \mathrm{d}r \, \mathrm{d}\theta \\ &= 4 \int_{-\pi/4}^{\pi/4} \frac{1}{2} \cos^2(2\theta) \, \mathrm{d}\theta = \int_{-\pi/4}^{\pi/4} (1 + \cos 4\theta) \, \mathrm{d}\theta = [\theta + \frac{1}{4} \sin 4\theta]_{-\pi/4}^{\pi/4} \\ &= \frac{\pi}{2}.\end{aligned}$$

4.6.2 Spherical coordinates

Like cylindrical coordinates, spherical coordinates generalize polar ones, but in a different direction. Denote by $\arccos : [-1, 1] \to [0, \pi]$ the inverse of the restriction of cos to the interval $[0, \pi]$. Let $\rho : \mathbb{R}^3 \to \mathbb{R}$ denote the distance function to the origin, $\rho(\boldsymbol{x}) = |\boldsymbol{x}|$, and $\varphi : \mathbb{R}^3 \setminus \{\mathbf{0}\} \to [0, \pi]$ the angle with the positive z-axis,

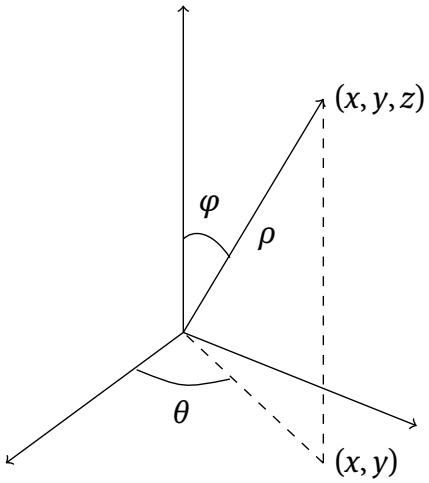
$$\varphi(\boldsymbol{x}) = \arccos(\langle \boldsymbol{x}, \boldsymbol{e}_3 \rangle / |\boldsymbol{x}|), \qquad \boldsymbol{x} \in \mathbb{R}^3 \setminus \{\mathbf{0}\}.$$

Thus,

$$\boldsymbol{x} = (\pi(\boldsymbol{x}), \rho(\boldsymbol{x}) \cos \varphi(\boldsymbol{x})),$$

where $\pi : \mathbb{R}^3 = \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^2$ is projection onto the x-y plane. Since $\pi(\boldsymbol{x}) \in \mathbb{R}^2$, it may be expressed in polar coordinates ; in fact, $r(\pi(\boldsymbol{x})) = \rho(\boldsymbol{x}) \sin \varphi(\boldsymbol{x})$, so that

$$\boldsymbol{x} = (\rho(\boldsymbol{x}) \cos \theta(\pi(\boldsymbol{x})) \sin \varphi(\boldsymbol{x}), \rho(\boldsymbol{x}) \sin \theta(\pi(\boldsymbol{x})) \sin \varphi(\boldsymbol{x}), \rho(\boldsymbol{x}) \cos \varphi(\boldsymbol{x})). \tag{4.6.3}$$



Spherical coordinates

Spherical coordinates are given by the map $(\rho, \theta, \varphi) : \mathbb{R}^3 \setminus \{\mathbf{0}\} \to \mathbb{R}^3$, with ρ, φ as above, and θ the same function from cylindrical coordinates. If

$$\begin{aligned} \boldsymbol{g} : [0, \infty) \times [0, 2\pi) \times [0, \pi] &\to \mathbb{R}^3, \\ (a_1, a_2, a_3) &\mapsto a_1(\sin a_3 \cos a_2, \sin a_3 \sin a_2, \cos a_3), \end{aligned}$$

then (4.6.3) says that (ρ, θ, φ) is invertible with inverse $\boldsymbol{g}$. $\boldsymbol{g}$ has Jacobian matrix

$$[D\boldsymbol{g}(\rho, \theta, \varphi)] = \begin{bmatrix} \sin\varphi \cos\theta & -\rho \sin\varphi \sin\theta & \rho \cos\varphi \cos\theta \\ \sin\varphi \sin\theta & \rho \sin\varphi \cos\theta & \rho \cos\varphi \sin\theta \\ \cos\varphi & 0 & -\rho \sin\varphi \end{bmatrix},$$

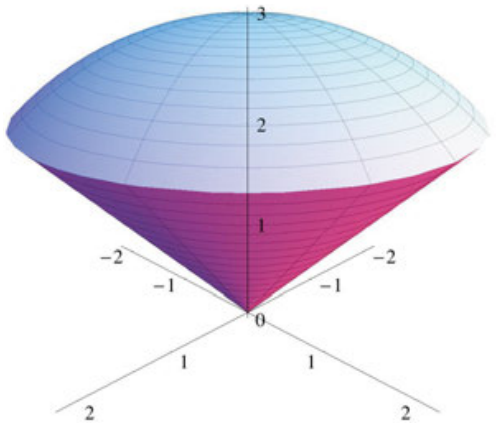
which has determinant $-\rho^2 \sin\varphi$. Thus, if $V = \boldsymbol{g}(U)$ and f is integrable over V, then

$$\int_V f = \int_U (f \circ \boldsymbol{g})(\rho, \theta, \varphi) \rho^2 \sin\varphi \, \mathrm{d}\rho \, \mathrm{d}\theta \, \mathrm{d}\varphi. \tag{4.6.4}$$

Examples and Remarks 4.6.1. (i) Suppose we are asked to determine the volume of the 'ice cream cone' consisting of the region V bounded by the sphere $x^2 + y^2 + z^2 = R^2$ and the cone $z = \sqrt{x^2 + y^2}$. First of all, notice that in spherical coordinates, the sphere and cone have equations $\rho = R$ and $\varphi = \pi/4$ respectively (more generally the equation $\varphi = \alpha$, $\alpha \in (0, \pi)$ describes a cone with tip at the origin and axis the z-axis). This means that $V^0 = \boldsymbol{g}(U)$, where $U = (0, R) \times [0, 2\pi) \times [0, \pi/4)$,

so that

$$\begin{aligned}\operatorname{vol} V = \operatorname{vol} V^0 &= \int_{\boldsymbol{g}(U)} 1 = \int_U \rho^2 \sin\varphi \, \mathrm{d}\rho \, \mathrm{d}\varphi \, \mathrm{d}\theta \\ &= \int_0^{2\pi}\int_0^{\pi/4}\int_0^R \rho^2 \sin\varphi \, \mathrm{d}\rho \, \mathrm{d}\varphi \, \mathrm{d}\theta = 2\pi\frac{R^3}{3}\int_0^{\pi/4} \sin\varphi \, \mathrm{d}\varphi \\ &= 2\pi\frac{R^3}{3}\left(1 - \frac{\sqrt{2}}{2}\right).\end{aligned}$$



Notice that the volume of the whole ball is obtained by replacing the upper bound of $\pi/4$ in the last integral with π, which results in $4\pi R^3/3$.

(ii) More generally, one can compute the volume of the ball $B^n(R)$ of radius $R > 0$ around $\mathbf{0}$ in $\mathbb{R}^n$ as follows: first of all, notice that the diffeomorphism $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^n$, $\boldsymbol{g}(\boldsymbol{x}) = R\boldsymbol{x}$, maps $B^n(1)$ onto $B^n(R)$. Thus, by the change of variables theorem,

$$\operatorname{vol}(B^n(R)) = \int_{B^n(1)} |\det D\boldsymbol{g}| = \int_{B^n(1)} R^n = R^n \operatorname{vol}(B^n(1)). \tag{4.6.5}$$

Next, write $B^n(1) = \{(\boldsymbol{x}, \boldsymbol{a}) \mid \boldsymbol{x} \in B^2(1), \boldsymbol{a} \in \mathbb{R}^{n-2}, |\boldsymbol{a}|^2 + |\boldsymbol{x}|^2 \leq 1\}$, and express $B^2(1)$ in polar coordinates to obtain with (4.6.5)

$$\begin{aligned}\operatorname{vol}(B^n(1)) &= \int_0^{2\pi}\int_0^1 \operatorname{vol}(B^{n-2}(\sqrt{1-r^2}))r \, \mathrm{d}r \, \mathrm{d}\theta \\ &= 2\pi\int_0^1 (1-r^2)^{\frac{n-2}{2}} \operatorname{vol}(B^{n-2}(1))r \, \mathrm{d}r \\ &= 2\pi \operatorname{vol}(B^{n-2}(1))\int_0^1 (1-r^2)^{\frac{n-2}{2}} r \, \mathrm{d}r \\ &= \frac{2\pi}{n}\operatorname{vol}(B^{n-2}(1)).\end{aligned}$$

An easy induction argument together with (4.6.5) then yields

$$\operatorname{vol} B^{2n+1}(R) = \frac{\pi^n 2^{n+1} R^{2n+1}}{(2n+1)(2n-1)\cdots 5\cdot 3}, \qquad \operatorname{vol} B^{2n}(R) = \frac{\pi^n R^{2n}}{n!}.$$

(iii) Spherical coordinates in $\mathbb{R}^3$ can be extended in exactly the same way to $\mathbb{R}^4$: change the notation for φ to φ_1, and define $\varphi_2 : \mathbb{R}^4 \setminus \{\mathbf{0}\} \to [0,\pi]$ by

$$\varphi_2(\boldsymbol{x}) = \arccos(\langle \boldsymbol{x}, \boldsymbol{e}_4\rangle/|\boldsymbol{x}|), \qquad \boldsymbol{x} \in \mathbb{R}^4 \setminus \{\mathbf{0}\}.$$

Then $\boldsymbol{x} = (\pi(\boldsymbol{x}), \rho(\boldsymbol{x})\cos\varphi_2(\boldsymbol{x}))$, where ρ denotes the distance function to the origin in $\mathbb{R}^4$ and $\pi : \mathbb{R}^3 \times \mathbb{R} \to \mathbb{R}^3$ the projection. Using 3-dimensional spherical coordinates for $\pi(\boldsymbol{x})$, we obtain

$$\boldsymbol{x} = (\rho\cos\theta\sin\varphi_1\sin\varphi_2, \rho\sin\theta\sin\varphi_1\sin\varphi_2, \rho\cos\varphi_1\sin\varphi_2, \rho\cos\varphi_2)(\boldsymbol{x}).$$

A straightforward if tedious induction argument now yields spherical coordinates $(\rho, \theta, \varphi_1, \dots, \varphi_{n-2}) : \mathbb{R}^n \setminus \{\mathbf{0}\} \to \mathbb{R}^n$ on $\mathbb{R}^n$ with inverse $\boldsymbol{g}$, where

$$\begin{aligned}\boldsymbol{g}(\rho, \theta, \varphi_1, \dots, \varphi_{n-2}) = \rho(&\cos\theta\sin\varphi_1\cdots\sin\varphi_{n-2}, \sin\theta\sin\varphi_1\cdots\sin\varphi_{n-2},\\ &\cos\varphi_1\sin\varphi_2\cdots\sin\varphi_{n-2}, \dots, \cos\varphi_{n-2}),\end{aligned} \tag{4.6.6}$$

and

$$|\det D\boldsymbol{g}|(\rho, \theta, \varphi_1, \dots, \varphi_{n-2}) = \rho^{n-1}\sin^{n-2}\varphi_{n-2}\sin^{n-3}\varphi_{n-3}\cdots\sin\varphi_1. \tag{4.6.7}$$

This provides an alternative method for deriving the volume of a ball in $\mathbb{R}^n$, see Exercise 4.25.

(iv) Recall that for a continuous function f of one variable, the *improper integral* $\int_a^\infty f$ is defined to be $\lim_{x\to\infty}\int_a^x f$, provided the limit exists, in which case the improper integral is said to *converge*. Similarly, $\int_{-\infty}^a f = \lim_{x\to-\infty}\int_x^a f$, and if both exist, we define $\int_{-\infty}^\infty f = \int_{-\infty}^a f + \int_a^\infty f$ (notice that the particular value of a is irrelevant). Consider the function f given by $f(x) = e^{-x^2}$. Even though f has an antiderivative F, it is well known that there is no formula for F other than $F(x) = \int_a^x e^{-t^2}$ for some $a \in \mathbb{R}$. Nevertheless, we will compute explicitly $\int_{-\infty}^\infty f$. First of all, observe that $\int_0^\infty f$ exists, because $e^{-x^2} \le 1/x^2$ for large x, and $\int_a^\infty 1/x^2\,\mathrm{d}x$ converges. Furthermore, $\int_{-\infty}^\infty f = \lim_{R\to\infty}\int_{-R}^{\mathbb{R}} f$ since f is even. If $S_R = [-R,R]\times[-R,R]$, then

$$\left(\int_{-R}^R f\right)^2 = \left(\int_{-R}^R e^{-x^2}\,\mathrm{d}x\right)\left(\int_{-R}^R e^{-y^2}\,\mathrm{d}y\right) = \int_{S_R} e^{-(x^2+y^2)}\,\mathrm{d}x\,\mathrm{d}y \tag{4.6.8}$$

by Fubini's theorem. Let B_R denote the disk of radius R centered at the origin. Observe that if $g(x,y) = e^{-(x^2+y^2)}$, then $\lim_{R\to\infty}\int_{B_R} g = \lim_{R\to\infty}\int_{S_R} g$ in the sense

that if one exists, then so does the other and the two are equal: indeed, $S_{R/\sqrt{2}} \subset B_R \subset S_{2R}$, and $|g| \le e^{-R^2}$ on $S_{2R} \setminus B_R$. Similarly, $|g| \le e^{-R^2/2}$ on $B_R \setminus S_{R/\sqrt{2}}$, so that

$$\int_{B_R} g - \int_{S_{R/\sqrt{2}}} g \le e^{-R^2/2} R^2 (\pi - 2) \overset{R\to\infty}{\longrightarrow} 0,$$

and

$$\int_{S_{2R}} g - \int_{B_R} g \le e^{-R^2} R^2 (4 - \pi) \overset{R\to\infty}{\longrightarrow} 0,$$

thereby proving the claim. Using polar coordinates,

$$\int_{B_R} g = \int_0^{2\pi} \int_0^R e^{-r^2} r \, \mathrm{d}r \, \mathrm{d}\theta = \pi(1 - e^{-R^2}),$$

so that by (4.6.8),

$$\int_{-\infty}^{\infty} e^{-x^2} \, \mathrm{d}x = \left(\lim_{R\to\infty} \int_{B_R} g\right)^{\frac{1}{2}} = \sqrt{\pi}.$$

4.7 Some applications

Before discussing applications of integration to concepts from physics, we observe that if $f : [a, b] \to \mathbb{R}$ is integrable, then its integral may be evaluated by considering only partitions P_n of $[a, b]$ into n subintervals

$$I_j^n = [a + (j - 1)\frac{b - a}{n}, a + j\frac{b - a}{n}]$$

of equal length $l(I_j^n) = (b - a)/n$. If x_j^n is any point in I_j^n, the expression

$$\sum_{j=1}^{n} f(x_j^n) \, l(I_j^n)$$

is called a *Riemann sum* of f for P_n. Since this sum is sandwiched between $L(f, P_n)$ and $U(f, P_n)$, and the latter two sequences converge to the integral of f,

$$\int_a^b f = \lim_{n\to\infty} \sum_{j=1}^{n} f(x_j^n) \, l(I_j^n).$$

The same is of course true for integrals of functions of more than one variable, if one considers partitions by boxes of the same size. This turns out to be useful in defining physical concepts as limits of “approximations”.

4.7.1 Mass

Consider a body occupying a region E in 3-space, made of a not necessarily homogenous material. Its *density* $\rho(\boldsymbol{r})$ at an interior point $\boldsymbol{r} \in E$ is

$$\rho(\boldsymbol{r}) = \lim_{\varepsilon \to 0} \frac{\text{mass}(B_\varepsilon(\boldsymbol{r}))}{\text{vol}(B_\varepsilon(\boldsymbol{r}))},$$

provided the limit exists. We assume that the density is defined everywhere and is, in fact, continuous. In order to evaluate the mass of the whole body, we begin by approximating it with Riemann sums. Suppose first that E is a box, and consider a partition P_n of E by n^3 boxes of equal size $B_1, \ldots, B_{n^3}$, obtained by partitioning each side of the box into n subintervals. If n is large enough that each B_i is very small, the continuity of ρ ensures that for any $\boldsymbol{r}_i \in B_i$, $\rho(\boldsymbol{r}_i)\,\text{vol}(B_i)$ is a fair approximation of the mass of B_i. Thus, $\sum_{i=1}^{n^3} \rho(\boldsymbol{r}_i)\,\text{vol}(B_i)$ is an approximation to the mass m of the body, and it is expected that the approximation gets better as n is larger. Since this is a Riemann sum of the continuous function ρ for P_n, its limit as $n \to \infty$ equals the integral of ρ over E, and we define the mass m of the body to equal

$$m = \int_E \rho.$$

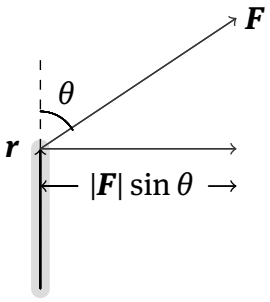
This is easily extended to a more general body, by enclosing E inside some box and integrating $\rho\chi_E$ over the box. The above formula is therefore still valid.

4.7.2 Center of mass

When trying to place a two-dimensional rigid object such as a tray on the tip of a thin vertical rod, a little experimentation shows that there is one and only one point on the object which, when in contact with the rod, leaves the object balanced. This point is called the center of mass of the object, and can be defined more generally for a three-dimensional body. In order to determine this point, the concept of *torque* is useful.

In physics, the torque $\boldsymbol{\tau}$ of a force $\boldsymbol{F}$ is a vector that measures the tendency of that force to rotate an object about an axis or a point. If the point where the torque is measured is located at the origin, and the object is at the end of a lever arm at position $\boldsymbol{r}$, then the magnitude of the torque is experimentally determined to be proportional to both the length of the lever arm and to the magnitude of the component of the force orthogonal to the arm. In other words, if θ is the angle between $\boldsymbol{F}$ and $\boldsymbol{r}$, the magnitude of the torque is proportional to $|\boldsymbol{r}||\boldsymbol{F}| \sin\theta$. This is also the magnitude of $\boldsymbol{r} \times \boldsymbol{F}$, and the torque is therefore defined to be

$$\boldsymbol{\tau} = \boldsymbol{r} \times \boldsymbol{F}.$$



If we now have a discrete system of point masses m_i at positions $\boldsymbol{r}_i$, $i = 1, \ldots, n$, then the center of mass of the system is that point $\boldsymbol{R}$ with respect to which the total torque of gravity on the point masses is zero, so that rotationally speaking, the system behaves as though all mass were concentrated at the center of mass. Since the torque due to the i-th object is $(\boldsymbol{r}_i - \boldsymbol{R}) \times -m_i g \boldsymbol{k}$, where g is the gravitational constant, the total torque is $\sum_i (\boldsymbol{r}_i - \boldsymbol{R}) \times -m_i g \boldsymbol{k}$. Setting this equal to zero implies that the vector $\sum_i m_i (\boldsymbol{r}_i - \boldsymbol{R})$ is parallel to $\boldsymbol{k}$. If this is to hold for any rotation of the system about its center of mass, then the vector itself must be zero, since this amounts to replacing $\boldsymbol{k}$ by an arbitrary vector. Thus, the center of mass is located at

$$\boldsymbol{R} = \frac{1}{M} \sum_i m_i \boldsymbol{r}_i,$$

where $M = \sum_i m_i$ is the total mass.

The above discussion for discrete mass distributions generalizes to continuous mass distributions: consider a solid occupying a Jordan-measurable region E in space, with possibly variable density ρ. Assume first that E is a box, and partition it into subboxes B_i of equal size, $i = 1, \ldots, n^3$. If $\boldsymbol{r}_i$ denotes the center of gravity of B_i and B_i is small enough, we approximate the mass of B_i by $\rho(\boldsymbol{r}_i) \Delta V$, where ΔV is the common volume of the subboxes. Denoting by $\boldsymbol{R}$ the position vector of the center of mass, the torque due to B_i experienced at the center of mass is approximately $(\boldsymbol{r}_i - \boldsymbol{R}) \times \rho(\boldsymbol{r}_i) \Delta V g \boldsymbol{k}$. Eliminating $g\boldsymbol{k}$ as above, we see that the vector $\sum_i (\boldsymbol{r}_i - \boldsymbol{R}) \rho(\boldsymbol{r}_i) \Delta V$ must go to zero as $n \to \infty$. Its three components are limits of Riemann sums, and if $\boldsymbol{R} = \begin{bmatrix} \bar{x} & \bar{y} & \bar{z} \end{bmatrix}^T$, then the first component equals $\int_E (x - \bar{x}) \rho(x, y, z) \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z$. Setting this equal to zero and doing the same with the other components, we obtain for the coordinates of the center of mass

$$\bar{x} = \frac{1}{M} \int_E x \rho(x, y, z) \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z,$$
$$\bar{y} = \frac{1}{M} \int_E y \rho(x, y, z) \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z,$$
$$\bar{z} = \frac{1}{M} \int_E z \rho(x, y, z) \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z.$$

The same argument used in discussing the mass of an object shows that the above identities hold for any Jordan-measurable body E, not just box-like ones.

When the object is homogeneous, that is, when the density is constant, the center of mass is called the *centroid*.

Example 4.7.1. Let us find the centroid of a solid hemisphere E of radius a. Since the density is constant, the mass M equals density times the volume of E, and the first coordinate of the centroid is

$$\bar{x} = \frac{1}{\operatorname{vol}(E)} \int_E x \, dx \, dy \, dz = \frac{3}{2\pi a^3} \int_E x \, dx \, dy \, dz,$$

with similar formulae for the other coordinates. Assuming the solid is the Northern hemisphere of a sphere centered at the origin, we would expect, by symmetry, that $\bar{x} = \bar{y} = 0$. This is indeed the case: using spherical coordinates,

$$\begin{aligned}
\int_E x \, dx \, dy \, dz &= \int_0^{2\pi} \int_0^{\pi/2} \int_0^a (\rho \sin\varphi \cos\theta) \cdot \rho^2 \sin\varphi \, d\theta \, d\varphi \, d\rho \\
&= \int_0^{2\pi} \cos\theta \, d\theta \int_0^{\pi/2} \sin^2\varphi \, d\varphi \int_0^a \rho^3 \, d\rho \\
&= 0
\end{aligned}$$

because the first integral on the second to last line vanishes. A similar calculation yields $\bar{y} = 0$. Finally,

$$\begin{aligned}
\int_E z \, dx \, dy \, dz &= \int_0^{2\pi} \int_0^{\pi/2} \int_0^a (\rho \cos\varphi) \cdot \rho^2 \sin\varphi \, d\theta \, d\varphi \, d\rho \\
&= 2\pi \int_0^{\pi/2} \sin\varphi \cos\varphi \, d\varphi \int_0^a \rho^3 \, d\rho = 2\pi \frac{\sin^2\varphi}{2} \Big|_0^{\pi/2} \frac{\rho^4}{4} \Big|_0^a \\
&= \frac{\pi a^4}{4},
\end{aligned}$$

so that $\bar{z} = (3a)/8$.

4.7.3 Moment of inertia

According to Newton's second law, a force must be applied to an object moving along a straight line in order to change its velocity. The magnitude F of the force is proportional to the object's acceleration a, and the constant of proportionality is the mass m of the object: $F = ma$.

A similar law holds for objects that are rotating about an axis. In this case, torque has to be applied in order to change the object's angular velocity: recall that if the

object is at distance r from the axis, and F is the (norm of the) component of the force tangent to the object's circular path, then the amount of torque measured at the axis is $\tau = rF$. Now, if m denotes the object's mass, then its acceleration is $a = F/m$, and its angular acceleration is $\alpha = a/r = F/(mr)$. Therefore, the applied torque

$$\tau = (mr^2)\alpha$$

is proportional to the angular acceleration. The constant of proportionality, $I = mr^2$, is called the *moment of inertia* of the object about the axis.

The discussion generalizes to discrete mass distributions and continuous ones in exactly the same way as was done for the center of mass. Thus, the moment of inertia of a solid body occupying a region E with density ρ about an axis is given by

$$I = \int_E \rho r^2, \tag{4.7.1}$$

where $r : E \to [0, \infty)$ is the distance function to the axis.

Example 4.7.2. Let us find the moment of inertia of a solid ball with constant density and radius a about any axis through its center. By symmetry, the ball may be assumed to be centered at the origin and the axis is the z-axis. Using spherical coordinates, and denoting by $\tilde{\rho}$ the density of the ball to avoid confusion with the spherical coordinate ρ,

$$\begin{aligned} I &= \int_{\{(x,y,z)|x^2+y^2+z^2\le a^2\}} \tilde{\rho}\,(x^2 + y^2)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \\ &= \tilde{\rho} \int_0^{2\pi}\int_0^a\int_0^\pi (\rho^2 \sin^2\varphi)\cdot\rho^2 \sin\varphi\,\mathrm{d}\varphi\,\mathrm{d}\rho\,\mathrm{d}\theta = 2\pi\tilde{\rho}\int_0^a \rho^4\,\mathrm{d}\rho \int_0^\pi \sin^3\varphi\,\mathrm{d}\varphi \\ &= 2\pi\tilde{\rho}\cdot\frac{a^5}{5}\cdot\frac{4}{3}. \end{aligned}$$

Since $\tilde{\rho}$ is constant, the mass m of the ball equals $(4/3)\pi a^3\tilde{\rho}$, so the moment of inertia may also be written as $I = (2/5)ma^2$.

4.8 Exercises

4.1. Prove the last two assertions of Theorem 4.1.2.

4.2. Prove or disprove:

(a) If f is a bounded function on $A \subset \mathbb{R}^n$ and $|f|$ is integrable on A, then so is f.

(b) If U is open in $\mathbb{R}^n$ and $f : U \to \mathbb{R}$ is continuous at some $\boldsymbol{p} \in U$, then f is integrable on $B_r(\boldsymbol{p})$ for small enough $r > 0$.

(c) If $A \subset \mathbb{R}^n$ has measure zero, then A is bounded.

(d) If $A \subset \mathbb{R}^n$ has measure zero, then the boundary of A has measure zero.

4.3. Let A be a compact set of measure zero in $\mathbb{R}^n$. Show that if $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^k$ is continuous, then $\boldsymbol{g}(A)$ has measure zero in $\mathbb{R}^k$.

4.4. It was shown in Remark 4.2.2 that if $A \subset \mathbb{R}^n$ is Jordan-measurable and has measure zero, then it has volume zero. Show that the converse is also true, so that for Jordan-measurable sets, the two concepts are equivalent.

4.5. Show that any compact submanifold of $\mathbb{R}^n$ has measure zero in $\mathbb{R}^n$.

4.6. $A \subset \mathbb{R}^n$ is said to have *content zero* if for any $\varepsilon > 0$, there exists a finite cover $\{B_1, \ldots, B_k\}$ of A by boxes with $\sum_i \operatorname{vol}(B_i) < \varepsilon$.
(a) Show that if A is compact and has measure zero, then it has content zero.
(b) Give examples of sets of measure zero that do not have content zero.

4.7. Let $f : U \to \mathbb{R}^n$ be a function which is integrable over some box $A = \Pi_{i=1}^n [a_i, b_i]$ in U. For each positive integer k, consider the partition P_k of A obtained by partitioning each $[a_i, b_i]$ into k intervals of equal length, and choose some point $\boldsymbol{a}_B$ in each subbox $B \in P_k$. Prove that

$$\lim_{k\to\infty} \sum_{B\in P_k} f(\boldsymbol{a}_B) \operatorname{vol}(B) = \int_A f.$$

4.8. Prove that if U is a bounded open set in $\mathbb{R}^n$, then there exists a sequence of smooth functions $f_k : \mathbb{R}^n \to \mathbb{R}$ such that

$$\lim_{k\to\infty} \int_{R^n} f_k = \operatorname{vol}(U).$$

4.9. Given $f : [a, b] \to \mathbb{R}$, define $g : [a, b] \times [c, d] \to \mathbb{R}$ by $g(x, y) = f(x)$. Using only the definition of integral, show that g is integrable if and only f is, and if it is, then $\int_{[a,b]\times[c,d]} g = (d - c) \int_{[a,b]} f$.

4.10. Let U be an open set in $\mathbb{R}^n$, and suppose $f : U \to \mathbb{R}$ is integrable on U and continuous at some $\boldsymbol{p} \in U$.
(a) If $C_r(\boldsymbol{p})$ denotes the cube

$$[u^1(\boldsymbol{p}) - r/2, u^1(\boldsymbol{p}) + r/2] \times \cdots \times [u^n(\boldsymbol{p}) - r/2, u^n(\boldsymbol{p}) + r/2]$$

with sides of length r centered at $\boldsymbol{p}$, show that

$$\lim_{r\to 0+} \frac{1}{\operatorname{vol}(C_r(\boldsymbol{p}))} \int_{C_r(\boldsymbol{p})} f = f(\boldsymbol{p}).$$

(b) Prove that

$$\lim_{r\to 0+} \frac{1}{\operatorname{vol}(B_r(\boldsymbol{p}))} \int_{B_r(\boldsymbol{p})} f = f(\boldsymbol{p}).$$

4.11. Let $A \subset \mathbb{R}^n$ be compact, connected, and Jordan-measurable. If $f : A \to \mathbb{R}$ is continuous, show that there exists some $\boldsymbol{a} \in A$ such that

$$\int_A f = f(\boldsymbol{a}) \cdot \mathrm{vol}(A).$$

4.12. Let $f : [a, b] \times [c, d] \to \mathbb{R}$ be continuous.
(a) If $g(x, t) = \int_c^t f(x, y)\,\mathrm{d}y$, find $D_1 G$ and $D_2 G$.
(b) If $h(s, t) = \int_a^s g(x, t)\,\mathrm{d}x = \int_a^s (\int_c^t f(x, y)\,\mathrm{d}y)\,\mathrm{d}x$, find $D_{12} h$.

4.13. Let $f : [0, 1] \times [0, 1] \to \mathbb{R}$ be continuous. Prove that

$$\int_0^1 (\int_0^x f(x, y)\,\mathrm{d}y)\,\mathrm{d}x = \int_0^1 (\int_y^1 f(x, y)\,\mathrm{d}x)\,\mathrm{d}y.$$

4.14. Let A denote a Jordan-measurable set in $\mathbb{R}^n$, and $B = A \times [0, 1] \subset \mathbb{R}^{n+1}$. Given $\boldsymbol{u} \in \mathbb{R}^n$,define

$$B_{\boldsymbol{u}} = \{\boldsymbol{a} + t(\boldsymbol{u} + \boldsymbol{e}_{n+1}) \mid \boldsymbol{a} \in A, t \in [0, 1]\} \subset \mathbb{R}^{n+1}.$$

Thus, $B = B_{\boldsymbol{0}}$ is a right solid cylinder over A of height 1, and $B_{\boldsymbol{u}}$ is a slanted cylinder over A of the same height. Prove that $\mathrm{vol}\, B_{\boldsymbol{u}} = \mathrm{vol}\, B$ for any $\boldsymbol{u} \in \mathbb{R}^n$.

4.15. (a) Define $f : (0, 1) \to \mathbb{R}$ by

$$f(x) = -n, \qquad x \in (1 - \frac{1}{n}, 1 - \frac{1}{n+1}).$$

Show that $\lim_{x \to 1} \int_{(0,x)} f$ exists, but $\int_{(0,1)} f$ does not.
(b) Prove that if $f : (0, 1) \to \mathbb{R}$ is continuous, then $\lim_{x \to 1} \int_{(0,x)} f$ exists if and only if $\int_{(0,1)} f$ does, and if they do exist, then they coincide. Give an example of a continuous $f : (0, 1) \to \mathbb{R}$ that is not integrable over $(0, 1)$.

4.16. Define $f : [0, 1] \times [0, 1] \to \mathbb{R}$ by

$$f(x, y) = \begin{cases} 1 & \text{if } x \text{ is rational,} \\ y & \text{otherwise.} \end{cases}$$

(a) Show that $\int_0^1 \int_0^1 f(x, y)\,\mathrm{d}y\,\mathrm{d}x$ exists, and find it.
(b) Prove that $\int_0^1 \int_0^1 f(x, y)\,\mathrm{d}x\,\mathrm{d}y$ does not exist.
(c) Show that f is not integrable over its domain.

4.17. Explain why one may assume that the Jacobian determinant of $\boldsymbol{g}$ is nowhere zero in the proof of the change of variables theorem.

4.18. Prove that Euclidean motions preserve volume; i.e., if $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ is a Euclidean motion, and $A \subset \mathbb{R}^n$ is Jordan-measurable, then $\mathrm{vol}(\boldsymbol{f}(A)) = \mathrm{vol}(A)$.

4.19. Let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable.

(a) Suppose that $|\det D\boldsymbol{f}(\boldsymbol{p})| > 1$ for some $\boldsymbol{p}$. Show that $\mathrm{vol}(\boldsymbol{f}(B_r(\boldsymbol{p}))) > \mathrm{vol}(B_r(\boldsymbol{p}))$ for sufficiently small $r > 0$.

(b) More generally, prove that if $|\det D\boldsymbol{f}(\boldsymbol{p})| \neq 0$, then

$$|\det D\boldsymbol{f}(\boldsymbol{p})| = \lim_{r\to 0+} \frac{\mathrm{vol}(\boldsymbol{f}(B_r(\boldsymbol{p})))}{\mathrm{vol}(B_r(\boldsymbol{p}))}.$$

4.20. Evaluate $\int_A f$, if $f(x,y) = \cos\frac{x}{x+y}$ and $A \subset \mathbb{R}^2$ is the region inside the triangle with vertices $(0,0)$, $(0,1)$, and $(1,0)$. *Hint:* Use the change of variables $u = x$, $v = x + y$.

4.21. The set of points (x,y) in the plane satisfying $(x^2 + y^2 - y)^2 - x^2 - y^2 = 0$ is called a *cardioid*. Sketch this curve and determine the area of the region it encloses.

4.22. Use an appropriate change of variables to evaluate $\int_A f$, if $f(x,y) = \sin(2x^2 + y^2)$, and $A = \{(x,y) \in \mathbb{R}^2 \mid y \geq 0, 2x^2 + y^2 \leq 2\}$.

4.23. Evaluate once again the volume of the cone from Examples 4.6.1 (i), but using spherical coordinates rather than cylindrical ones.

4.24. Determine the volume of the region in $\mathbb{R}^3$ that is bounded by the paraboloid $z = x^2 + y^2$ and the sphere $x^2 + y^2 + z^2 = 2$.

4.25. (a) Use induction and integration by parts to show that

$$\int_0^\pi \sin^{2n+1} x\,\mathrm{d}x = \frac{2^{n+1}n!}{(2n+1)(2n-1)\cdots 5\cdot 3},$$

$$\int_0^\pi \sin^{2n} x\,\mathrm{d}x = \pi\frac{(2n-1)(2n-3)\cdots 5\cdot 3}{2^n n!}.$$

(b) Use (4.6.7) to prove that the volume of a ball $B^n(R)$ of radius R in $\mathbb{R}^n$ is given by

$$\mathrm{vol}\, B^{2n+1}(R) = \frac{\pi^n 2^{n+1} R^{2n+1}}{(2n+1)(2n-1)\cdots 5\cdot 3}, \qquad \mathrm{vol}\, B^{2n}(R) = \frac{\pi^n R^{2n}}{n!}.$$

4.26. Let $a_i > 0$, $1 \leq i \leq n$. Use an appropriate change of variables to evaluate the volume of the region in $\mathbb{R}^n$ bounded by the ellipsoid

$$\{(x_1,\ldots,x_n) \in \mathbb{R}^n \mid \sum_{i=1}^n \frac{x_i^2}{a_i^2} = 1\}.$$

4.27. The *gamma function* is defined by

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,\mathrm{d}t, \qquad 0 < x < \infty.$$

(a) Show that this improper integral converges for all $x > 0$. (Notice that if $x < 1$, the integral is also improper at 0; i.e., it must be shown that

$$\int_0^a t^{x-1}e^{-t}\,\mathrm{d}t := \lim_{\varepsilon\to 0+}\int_\varepsilon^a t^{x-1}e^{-t}\,\mathrm{d}t$$

converges for some and hence all $a > 0$).

(b) Show that $\Gamma(1) = 1$.

(c) Use the change of variables $t = u^2$ and Examples and Remarks 4.6.1 (iv) to prove that $\Gamma(1/2) = \sqrt{\pi}$.

(d) Use integration by parts to show that $\Gamma(x + 1) = x\Gamma(x)$, and deduce that $\Gamma(n) = (n - 1)!$ for all $n \in \mathbb{N}$.

(e) In Examples and Remarks 4.6.1 (ii), a formula was derived for the volume $\mathrm{vol}(B^n(R))$ of a ball of radius R in $\mathbb{R}^n$. There were actually two formulas, one for even n and the other for odd n. Prove that they can be unified into one by means of the gamma function:

$$\mathrm{vol}(B^n(R)) = \frac{2R^n\pi^{n/2}}{n\Gamma(n/2)}.$$

4.28. (a) Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} 1 & \text{if } |x| \le 1, \\ \frac{1}{x^2} & \text{if } |x| \ge 1. \end{cases}$$

Show that $\lim_{r\to\infty}\int_{-r}^{r} f$ exists.

(b) Let $g : \mathbb{R}^2 \to \mathbb{R}$ be given by

$$g(\boldsymbol{x}) = \begin{cases} 1 & \text{if } |\boldsymbol{x}| \le 1, \\ \frac{1}{|\boldsymbol{x}|^2} & \text{if } |\boldsymbol{x}| \ge 1. \end{cases}$$

Show that $\lim_{r\to\infty}\int_{[-r,r]\times[-r,r]} g$ does not exist.

4.29. Reprove Theorem 2.3.1 by using the methods from this chapter; i.e., show that if $f, D_2 f : [a, b] \times [c, d] \to \mathbb{R}$ are continuous, then the function

$$\varphi : [c, d] \longrightarrow \mathbb{R},$$
$$y \longmapsto \int_a^b f(x, y)\,\mathrm{d}x$$

is differentiable on (c, d) and

$$\varphi'(y_0) = \int_a^b D_2 f(x, y_0)\,\mathrm{d}x, \qquad y_0 \in (c, d).$$

Hint: $\varphi(y) = \int_a^b \left(\int_c^y D_2 f(x, t)\,\mathrm{d}t + f(x, c)\right)\mathrm{d}x$. Notice also that the hypotheses may be somewhat weakened.

4.30. This exercise uses Sard's theorem to show that any smooth map $\boldsymbol{f} : \mathbb{R}^n \supset U \to \mathbb{R}^m$ can be "approximated" by an immersion if $m \geq 2n$; more specifically, given any $\varepsilon > 0$, there exists an $m \times n$ matrix $A = (a_{ij})$ with $|a_{ij}| < \varepsilon$ for all i and j, such that the map

$$\begin{aligned} U &\longrightarrow \mathbb{R}^m \\ \boldsymbol{x} &\longmapsto \boldsymbol{f}(\boldsymbol{x}) + A\boldsymbol{x} \end{aligned}$$

is an immersion. Recall from Exercise 3.20 that the collection $M_{m,n}(k)$ of all $m \times n$ matrices of rank k is a manifold of dimension $k(m + n - k)$. Define

$$\begin{aligned} \boldsymbol{g}_k : U \times M_{m,n}(k) &\longrightarrow M_{m,n} \\ (\boldsymbol{u}, B) &\longmapsto B - [D\boldsymbol{f}(\boldsymbol{u})]. \end{aligned}$$

(a) Show that if $m \geq 2n$ and $k < n$, then the dimension of the domain of $\boldsymbol{g}_k$ is less that that of its image. *Hint:* the function $t \mapsto t(m + n - t)$ is increasing if $t < n < m$.

(b) Show that the image of $\boldsymbol{g}_k$ has measure zero under the assumptions from part (a), and prove the claim made at the beginning of the exercise.

(c) Show that if $\boldsymbol{f} : \mathbb{R}^n \supset U \to \mathbb{R}^m$ is smooth, with U bounded and $m \geq 2n$, then for any $\varepsilon > 0$ there exists an immersion $\boldsymbol{g} : U \to \mathbb{R}^m$ such that $|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})| < \varepsilon$ for all $\boldsymbol{x} \in U$.

This result can be used in proving the so-called "Whitney imbedding theorem" which states that any n-dimensional manifold M can be imbedded in $\mathbb{R}^{2n+1}$ (see for example [14]); i.e., there exists an injective map $\boldsymbol{f} : M \to \mathbb{R}^{2n+1}$ of maximal rank everywhere with continuous inverse. In particular $\boldsymbol{f}(M)$ is a submanifold of $\mathbb{R}^{2n+1}$ (any parametrization $\boldsymbol{h}$ of M generates a parametrization of $\boldsymbol{f} \circ \boldsymbol{h}$ of $\boldsymbol{f}(M)$) which is diffeomorphic to M (via $\boldsymbol{f}$), so that M itself may be considered to be sitting inside $\mathbb{R}^{2n+1}$ regardless of the dimension of the original Euclidean space it is a submanifold of.

It should be noted that manifolds can be defined as abstract sets that are not contained in Euclidean space. Whitney's theorem asserts that our definition is equivalent to that one.

5 Differential Forms

Now that we are familiar with integration on Euclidean space, we would like to translate this to manifolds. Functions will be replaced by more exotic objects called differential forms. Their main advantage lies in that they have the change of variables formula built-in, once the concept of orientation is introduced. We begin by recalling the more general concept of tensor field.

5.1 Tensors and tensor fields

Definition 5.1.1. Let V be a vector space. Given $k \in \mathbb{N}$, a *k-tensor* on V is a multilinear map $T : V^k \to \mathbb{R}$, where V^k denotes the k-fold Cartesian product of V with itself.

The collection $\mathcal{T}_k(V)$ of all k-tensors on V is clearly a vector space under the usual addition of maps and scalar multiplication. For example, $\mathcal{T}_1(V)$ is just the dual space V^*. By convention, $\mathcal{T}_0(V)$ is defined to be $\mathbb{R}$. An inner product on V is a 2-tensor (with some additional properties).

Remark 5.1.1. k-tensors are actually a special case of what we defined to be tensors in Chapter 3; i.e., what we call a k-tensor here is just a tensor of order $(0,k)$ in the previous terminology. We will only deal with this particular subset in this chapter.

Definition 5.1.2. The *tensor product* of $T \in \mathcal{T}_k(V)$ with $\tilde{T} \in \mathcal{T}_l(V)$ is the $(k+l)$-tensor $T \otimes \tilde{T}$ given by

$$(T \otimes \tilde{T})(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k+l}) = T(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) \cdot \tilde{T}(\boldsymbol{v}_{k+1}, \ldots, \boldsymbol{v}_{k+l}), \qquad \boldsymbol{v}_i \in V.$$

The following properties are easy consequences of the definition, and their verification is left to the reader: for tensors T_i, $a \in \mathbb{R}$,

$$\begin{aligned}
(T_1 \otimes T_2) \otimes T_3 &= T_1 \otimes (T_2 \otimes T_3);\\
T_1 \otimes (T_2 + T_3) &= T_1 \otimes T_2 + T_1 \otimes T_3;\\
(T_1 + T_2) \otimes T_3 &= T_1 \otimes T_3 + T_2 \otimes T_3;\\
a(T_1 \otimes T_2) &= (aT_1) \otimes T_2 = T_1 \otimes (aT_2),
\end{aligned}$$

where of course the tensors being added in the second and third identities are assumed to have the same order. In view of the first property, either side will be denoted $T_1 \otimes T_2 \otimes T_3$. Observe, though, that the tensor product is, in general, not commutative; i.e., $T_1 \otimes T_2$ need not equal $T_2 \otimes T_1$.

Tensor products yield explicit bases for the spaces $\mathcal{T}_k(V)$:

Theorem 5.1.1. *Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ denote a basis of V. If $\alpha^1, \ldots, \alpha^n \in V^* = \mathcal{T}_1(V)$ is the dual basis (i.e., $\alpha^i(\boldsymbol{v}_j) = \delta_{ij}$), then*

$$\{\alpha^{i_1} \otimes \cdots \otimes \alpha^{i_k} \mid 1 \le i_1, \ldots, i_k \le n\}$$

is a basis of $\mathcal{T}_k(V)$, which therefore has dimension n^k.

Proof. Set $T_{i_1\cdots i_k} = T(\boldsymbol{v}_{i_1}, \ldots, \boldsymbol{v}_{i_k}) \in \mathbb{R}$. We claim that

$$T = \sum_{i_1,\ldots,i_k=1}^{n} T_{i_1\cdots i_k} \alpha^{i_1} \otimes \cdots \otimes \alpha^{i_k},$$

which will show that the set in the statement indeed spans $\mathcal{T}_k(V)$. In order to establish the claim, it suffices to check that both sides agree when evaluated on basis elements, since both are multilinear: for if $\boldsymbol{w}_i = \sum_j a_{ij}\boldsymbol{v}_j$ and M is any k-tensor, then

$$\begin{aligned} M(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k) &= M(\sum_{j_1=1}^{n} a_{1j_1}\boldsymbol{v}_{j_1}, \ldots, \sum_{j_k=1}^{n} a_{kj_k}\boldsymbol{v}_{j_k}) \\ &= \sum_{j_1,\ldots,j_k=1}^{n} a_{1j_1}\cdots a_{kj_k} M(\boldsymbol{v}_{j_1}, \ldots, \boldsymbol{v}_{j_k}), \end{aligned}$$

so that M is entirely determined by what it does to basis vectors. Now, $\alpha^i(\boldsymbol{v}_j) = \delta_{ij}$, so that

$$\begin{aligned} &\sum_{i_1,\ldots,i_k=1}^{n} T_{i_1\cdots i_k} \alpha^{i_1} \otimes \cdots \otimes \alpha^{i_k}(\boldsymbol{v}_{j_1}, \ldots, \boldsymbol{v}_{j_k}) \\ &\quad = \sum_{i_1,\ldots,i_k=1}^{n} T_{i_1\cdots i_k} \alpha^{i_1}(\boldsymbol{v}_{j_1}) \cdots \alpha^{i_k}(\boldsymbol{v}_{j_k}) = T_{j_1\cdots j_k} \\ &\quad = T(\boldsymbol{v}_{j_1}, \ldots, \boldsymbol{v}_{j_k}), \end{aligned}$$

thereby establishing the claim. Linear independence is similar: suppose that $\sum a_{i_1\cdots i_k} \alpha^{i_1} \otimes \cdots \otimes \alpha^{i_k} = 0$. As in the last calculation, given any $j_1, \ldots, j_k \in \{1, \ldots, n\}$, applying both sides to $\boldsymbol{v}_{j_1}, \ldots, \boldsymbol{v}_{j_k}$ yields $a_{j_1\cdots j_k} = 0$. This concludes the proof of the theorem. □

As in Chapter 3, we define a *k-tensor field* on a manifold M^n to be a map T that assigns to each $\boldsymbol{p} \in M$ a k-tensor $T(\boldsymbol{p}) \in \mathcal{T}_k(M_{\boldsymbol{p}})$ which is smooth in the sense that for any vector fields $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ on M, the function $T(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)$ which assigns to $\boldsymbol{p} \in M$ the number $T(\boldsymbol{p})(\boldsymbol{X}_1(\boldsymbol{p}), \ldots, \boldsymbol{X}_k(\boldsymbol{p}))$ is smooth in the usual sense. If $(U, \boldsymbol{x})$ is a local chart of M, then the restriction of T to U is smooth if and only if the functions

$$T_{i_1\cdots i_k} := T(\frac{\partial}{\partial x^{i_1}}, \ldots, \frac{\partial}{\partial x^{i_k}}) : U \to \mathbb{R}, \qquad 1 \le i_1, \ldots, i_k \le n$$

are differentiable . In fact, they must by definition be differentiable if T is smooth. Conversely, if these functions are differentiable, and x^i denotes as usual $u^i \circ \boldsymbol{x}$, then

$$T_{|U} = \sum_{i_1,\ldots,i_k=1}^{n} T_{i_1\cdots i_k}\, \mathrm{d}x^{i_1} \otimes \cdots \otimes \mathrm{d}x^{i_k}$$

is smooth because each $\mathrm{d}x^i$ is: recall that, given any vector field $\boldsymbol{X}$ on U, $\boldsymbol{X} = \sum_i \mathrm{d}x^i(\boldsymbol{X})\partial/\partial x^i$ (see Exercise 3.15), so that $\mathrm{d}x^i(\boldsymbol{X})$ is indeed a differentiable function on U.

The collection of all k-tensor fields on M is clearly a vector space, denoted $\mathcal{T}_k(M)$.

Definition 5.1.3. Let $\boldsymbol{f} : M \to N$ be a differentiable map, T a k-tensor field on N, $k > 0$. The *pullback* of T by $\boldsymbol{f}$ is the k-tensor field $\boldsymbol{f}^* T$ on M given by

$$(\boldsymbol{f}^* T)(\boldsymbol{p})(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) = T(\boldsymbol{f}(\boldsymbol{p}))(\boldsymbol{f}_* \boldsymbol{v}_1, \ldots, \boldsymbol{f}_* \boldsymbol{v}_k), \qquad \boldsymbol{p} \in M, \quad \boldsymbol{v}_i \in M_p.$$

A zero-tensor field on N is just a function $h : N \to \mathbb{R}$. In this case, we define $\boldsymbol{f}^* h$ to equal $h \circ \boldsymbol{f}$.

The following properties of the pullback are easily verified and left as an exercise:

Proposition 5.1.1. *Let $\boldsymbol{f} : M \to N$ be a differentiable map between manifolds M and N, S, T k-tensor fields on N, R an l-tensor field on N, and $h : N \to \mathbb{R}$ a function. Then*
(1) $\boldsymbol{f}^*(aS + bT) = a\boldsymbol{f}^* S + b\boldsymbol{f}^* T, \quad a, b \in \mathbb{R}$,
(2) $\boldsymbol{f}^*(R \otimes S) = \boldsymbol{f}^* R \otimes \boldsymbol{f}^* S$,
(3) $\boldsymbol{f}^*(hT) = (h \circ \boldsymbol{f})\boldsymbol{f}^* T$.

Example 5.1.1. A *Riemannian metric* on M is a 2-tensor field $\boldsymbol{g}$ on M such that $\boldsymbol{g}(\boldsymbol{p})$ is an inner product on $M_{\boldsymbol{p}}$ for all $\boldsymbol{p} \in M$. The *standard Riemannian metric* on $\mathbb{R}^n$ is $\boldsymbol{g} = \sum_i \mathrm{d}u^i \otimes \mathrm{d}u^i$. This is the inner product we've been using all along on each tangent space, since by definition $\boldsymbol{g}(\boldsymbol{D}_i, \boldsymbol{D}_j) = \delta_{ij}$, so that $\{\boldsymbol{D}_i \mid i = 1, \ldots, n\}$ is an orthonormal basis when evaluated at any point. If $M \subset \mathbb{R}^n$, the *standard Riemannian metric* on M is $\imath^* \boldsymbol{g}$, where $\imath : M \hookrightarrow \mathbb{R}^n$ is the inclusion map and $\boldsymbol{g}$ is the standard Riemannian metric on Euclidean space. Again, it is by definition the restriction to each $M_{\boldsymbol{p}}$ of the inner product on $\mathbb{R}^n_{\boldsymbol{p}}$, and coincides with the first fundamental tensor field on M defined in Chapter 3.

Remark 5.1.2. In physics, tensors are usually defined in a more convoluted way. For (relative) simplicity, we only discuss Cartesian tensors; i.e., tensors defined on Euclidean space $\mathbb{R}^n$. Consider two ordered bases $\boldsymbol{v}_i$ and $\boldsymbol{w}_j$ of $\mathbb{R}^n$, and let T be a tensor of order k on $\mathbb{R}^n$ in our sense of the word. Denote by $T_{i_1 \ldots i_k} = T(\boldsymbol{v}_{i_1}, \ldots, \boldsymbol{v}_{i_k})$ the components of T in the first basis, and by $\tilde{T}_{j_1 \ldots j_k} = T(\boldsymbol{w}_{j_1}, \ldots, \boldsymbol{w}_{j_k})$ the components of the same tensor in the second one. If $L = [L_{ij}]$ is the change of basis matrix, then

$$\begin{aligned} \tilde{T}_{j_1 \ldots j_k} &= T(\boldsymbol{w}_{j_1}, \ldots, \boldsymbol{w}_{j_k}) = T\Big(\sum_{i_1} L_{j_1 i_1} \boldsymbol{v}_{i_1}, \ldots, \sum_{i_k} L_{j_k i_k} \boldsymbol{v}_{i_k}\Big) \\ &= \sum_{i_1, i_2, \ldots, i_k} L_{j_1 i_1} L_{j_2 i_2} \cdots L_{j_k i_k} T(\boldsymbol{v}_{i_1}, \ldots, \boldsymbol{v}_{i_k}), \end{aligned}$$

so that

$$\tilde{T}_{j_1 \ldots j_k} = \sum_{i_1, i_2, \ldots, i_k} L_{j_1 i_1} L_{j_2 i_2} \cdots L_{j_k i_k} T_{i_1 \ldots i_k}. \tag{5.1.1}$$

A Cartesian tensor of order k is then traditionally defined as an "object" determined by n^k components in any given "coordinate system" (meaning basis), such that components in different coordinate systems are related by (5.1.1). Needless to say, such a condition can be quite difficult to check in a given instance.

5.2 Alternating tensors and forms

Recall from Chapter 1 that a k-tensor T on a vector space V is said to be alternating or *skew-symmetric* if

$$T(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_j, \ldots, \boldsymbol{v}_k) = -T(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_j, \ldots, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_k),$$

for all $1 \leq i \neq j \leq k$, $\boldsymbol{v}_l \in V$, $l = 1, \ldots, k$, and *symmetric* if the above equation holds with the minus sign removed. It is easily seen that any 2-tensor T can be written as a sum $T_s + T_a$ of a symmetric tensor T_s and an alternating one T_a: set

$$T_s(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}((T(\boldsymbol{x}, \boldsymbol{y}) + T(\boldsymbol{y}, \boldsymbol{x})), \qquad T_a(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}(T(\boldsymbol{x}, \boldsymbol{y}) - T(\boldsymbol{y}, \boldsymbol{x})),$$

for $\boldsymbol{x}, \boldsymbol{y} \in V$. Both symmetric and alternating parts can be generalized to k-tensors for arbitrary k. We only outline the latter, since we are particularly interested in alternating tensors for now: given a k-tensor T, define a new tensor T_a by

$$T_a(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) = \frac{1}{k!} \sum_{\sigma \in S_k} \varepsilon(\sigma) T(\boldsymbol{v}_{\sigma(1)}, \ldots, \boldsymbol{v}_{\sigma(k)}), \qquad \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in V,$$

with the terminology from Chapter 1: S_k denotes the set of permutations on k elements, and $\varepsilon(\sigma)$ is the sign of the permutation σ. Notice that for $k = 2$, the formula coincides with the 2-tensor T_a defined earlier. We claim that T_a is alternating. To see this, let τ denote the transposition (i, j). Since $S_k = \{\sigma \circ \tau \mid \sigma \in S_k\}$,

$$\begin{aligned}
T_a(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_j, \ldots, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_k) &= T_a(\boldsymbol{v}_{\tau(1)}, \ldots, \boldsymbol{v}_{\tau(i)}, \ldots, \boldsymbol{v}_{\tau(j)}, \ldots, \boldsymbol{v}_{\tau(k)}) \\
&= \frac{1}{k!} \sum_{\sigma \in S_k} \varepsilon(\sigma) T(\boldsymbol{v}_{(\sigma\circ\tau)(1)}, \ldots, \boldsymbol{v}_{(\sigma\circ\tau)(k)}) \\
&= -\frac{1}{k!} \sum_{\sigma \in S_k} \varepsilon(\sigma)\varepsilon(\tau) T(\boldsymbol{v}_{(\sigma\circ\tau)(1)}, \ldots, \boldsymbol{v}_{(\sigma\circ\tau)(k)}) \\
&= -\frac{1}{k!} \sum_{\sigma \in S_k} \varepsilon(\sigma \circ \tau) T(\boldsymbol{v}_{(\sigma\circ\tau)(1)}, \ldots, \boldsymbol{v}_{(\sigma\circ\tau)(k)}) \\
&= -\frac{1}{k!} \sum_{\sigma \in S_k} \varepsilon(\sigma) T(\boldsymbol{v}_{\sigma(1)}, \ldots, \boldsymbol{v}_{\sigma(k)}) \\
&= -T_a(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_j, \ldots, \boldsymbol{v}_k).
\end{aligned}$$

The collection of all alternating k-tensors on V is a vector space denoted $\Lambda_k(V)$. Its elements are called *k-forms*. The following properties are easily proven, and their verification is left to the reader:

Proposition 5.2.1. (1) *If $\alpha \in \Lambda_k(V)$, then $\alpha_a = \alpha$. In particular, for any tensor T on V, $(T_a)_a = T_a$;*

(2) *The map $\mathcal{T}_k(V) \to \Lambda_k(V)$ which sends T to T_a is linear; i.e., $(S + T)_a = S_a + T_a$, $(cT)_a = cT_a$ for all $S, T \in \mathcal{T}_k(V)$, $c \in \mathbb{R}$;*

(3)

$$\begin{aligned}(S \otimes (T_1 + T_2))_a &= (S \otimes T_1)_a + (S \otimes T_2)_a,\\ ((T_1 + T_2) \otimes S)_a &= (T_1 \otimes S)_a + (T_2 \otimes S)_a,\\ c(S \otimes T)_a &= (cS \otimes T)_a = (S \otimes cT)_a\end{aligned}$$

for all $T, T_1, T_2 \in \mathcal{T}_k(V)$, $S \in \mathcal{T}_l(V)$, $c \in \mathbb{R}$;

Since the tensor product of two forms is, in general, no longer alternating, we modify it as follows:

Definition 5.2.1. Given $\alpha \in \Lambda_k(V)$, $\beta \in \Lambda_l(V)$, their *wedge product* or *exterior product* is the $(k + l)$-form $\alpha \wedge \beta \in \Lambda_{k+l}(V)$ given by

$$\alpha \wedge \beta = \frac{(k + l)!}{k!\, l!}(\alpha \otimes \beta)_a.$$

The reason for including the factorial term in the definition of the wedge product will be revealed shortly. Notice that if $\alpha, \alpha^1, \alpha^2 \in \Lambda_k(V)$, $\beta \in \Lambda_l(V)$, and $c \in \mathbb{R}$, then

$$\begin{aligned}(\alpha^1 + \alpha^2) \wedge \beta &= \alpha^1 \wedge \beta + \alpha^2 \wedge \beta,\\ \beta \wedge (\alpha^1 + \alpha^2) &= \beta \wedge \alpha^1 + \beta \wedge \alpha^2,\\ c(\alpha \wedge \beta) &= (c\alpha) \wedge \beta = \alpha \wedge (c\beta).\end{aligned}$$

by the third statement in Proposition 5.2.1. It is also true that $(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma)$, but a few more properties are needed in order to prove this:

Lemma 5.2.1. *For* $\alpha^1, \ldots, \alpha^k \in \Lambda_1(V) = \mathcal{T}_1(V)$, $\sigma \in S_k$,

$$(\alpha^1 \otimes \cdots \otimes \alpha^k)_a = \frac{1}{k!} \sum_{\sigma \in S_k} \varepsilon(\sigma) \alpha^{\sigma(1)} \otimes \cdots \otimes \alpha^{\sigma(k)},$$

$$\begin{aligned}(\alpha^{\sigma(1)} \otimes \cdots \otimes \alpha^{\sigma(k)})_a &= \varepsilon(\sigma)(\alpha^1 \otimes \cdots \otimes \alpha^k)_a,\\ (\alpha^1 \otimes \cdots \otimes \alpha^k)_a &= ((\alpha^1 \otimes \cdots \alpha^{k-l})_a \otimes \alpha^{k-l+1} \otimes \cdots \otimes \alpha^k)_a\\ &= (\alpha^1 \otimes \cdots \alpha^{k-l} \otimes (\alpha^{k-l+1} \otimes \cdots \otimes \alpha^k)_a)_a.\end{aligned}$$

In particular, $(S_a \otimes T)_a = (S \otimes T_a)_a = (S \otimes T)_a$ *for all* $S \in \mathcal{T}_k(V)$, $T \in \mathcal{T}_l(V)$.

Proof. For the first identity, observe that

$$\begin{aligned}(\alpha^1 \otimes \cdots \otimes \alpha^k)_a(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) &= \frac{1}{k!} \sum_{\sigma \in S_k} \varepsilon(\sigma) \alpha^1(\boldsymbol{v}_{\sigma(1)}) \cdots \alpha^k(\boldsymbol{v}_{\sigma(k)})\\ &= \frac{1}{k!} \sum_{\sigma \in S_k} \varepsilon(\sigma)(\alpha^{\sigma^{-1}(1)} \otimes \cdots \otimes \alpha^{\sigma^{-1}(k)})(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k).\end{aligned}$$

Furthermore, a permutation has the same sign as its inverse, so that

$$\begin{aligned}(\alpha^1 \otimes \cdots \otimes \alpha^k)_a &= \frac{1}{k!} \sum_{\sigma \in S_k} \varepsilon(\sigma) \alpha^{\sigma^{-1}(1)} \otimes \cdots \otimes \alpha^{\sigma^{-1}(k)} \\ &= \frac{1}{k!} \sum_{\sigma^{-1} \in S_k} \varepsilon(\sigma^{-1}) \alpha^{\sigma^{-1}(1)} \otimes \cdots \otimes \alpha^{\sigma^{-1}(k)} \\ &= \frac{1}{k!} \sum_{\sigma \in S_k} \varepsilon(\sigma) \alpha^{\sigma(1)} \otimes \cdots \otimes \alpha^{\sigma(k)}.\end{aligned}$$

For the second identity, we have by the one just proved,

$$\begin{aligned}(\alpha^{\sigma(1)} \otimes \cdots \otimes \alpha^{\sigma(k)})_a &= \frac{1}{k!} \sum_{\tau \in S_k} \varepsilon(\tau) \alpha^{(\tau \circ \sigma)(1)} \otimes \cdots \otimes \alpha^{(\tau \circ \sigma)(k)} \\ &= \varepsilon(\sigma) \frac{1}{k!} \sum_{\tau \in S_k} \varepsilon(\tau \circ \sigma) \alpha^{(\tau \circ \sigma)(1)} \otimes \cdots \otimes \alpha^{(\tau \circ \sigma)(k)} \\ &= \varepsilon(\sigma) \frac{1}{k!} \sum_{\tau \circ \sigma \in S_k} \varepsilon(\tau \circ \sigma) \alpha^{(\tau \circ \sigma)(1)} \otimes \cdots \otimes \alpha^{(\tau \circ \sigma)(k)} \\ &= \varepsilon(\sigma)(\alpha^1 \otimes \cdots \otimes \alpha^k)_a.\end{aligned}$$

The two equalities in the third identity are proved in the same way, so we only argue the first one:

$$\begin{aligned}&[(\alpha^1 \otimes \cdots \otimes \alpha^{k-l})_a \otimes \alpha^{k-l+1} \otimes \cdots \otimes \alpha^l]_a \\ &\quad = \frac{1}{(k-l)!} \sum_{\sigma \in S_{k-l}} \varepsilon(\sigma)(\alpha^{\sigma(1)} \otimes \cdots \otimes \alpha^{\sigma(k-l)} \otimes \alpha^{k-l+1} \otimes \cdots \otimes \alpha^l)_a \\ &\quad = \frac{1}{(k-l)!} \sum_{\sigma \in S_{k-l}} \varepsilon^2(\sigma)(\alpha^1 \otimes \cdots \otimes \alpha^{k+l})_a \\ &\quad = (\alpha^1 \otimes \cdots \otimes \alpha^{k+l})_a,\end{aligned}$$

using the second identity on the third line. Finally, the last statement follows from the third identity, since the map $T \mapsto T_a$ is linear and each k-tensor T is a linear combination of basis elements of the form $\alpha^{i_1} \otimes \cdots \otimes \alpha^{i_k}$ by Theorem 5.1.1. □

The final statement in the above Lemma also implies that the wedge product is associative, meaning:

Theorem 5.2.1. *If $\alpha \in \Lambda_k(V)$, $\beta \in \Lambda_l(V)$, and $\gamma \in \Lambda_m(V)$, then*

$$(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma) = \frac{(k+l+m)!}{k!\, l!\, m!} (\alpha \otimes \beta \otimes \gamma)_a.$$

Proof.

$$\begin{aligned}(\alpha \wedge \beta) \wedge \gamma &= \frac{(k+l+m)!}{(k+l)!\, m!}((\alpha \wedge \beta) \otimes \gamma)_a \\ &= \frac{(k+l+m)!}{(k+l)!\, m!}\frac{(k+l)!}{k!\, l!}((\alpha \otimes \beta)_a \otimes \gamma)_a \\ &= \frac{(k+l+m)!}{k!\, l!\, m!}(\alpha \otimes \beta \otimes \gamma)_a.\end{aligned}$$

A similar argument shows that $\alpha \wedge (\beta \wedge \gamma)$ also equals the last term in the above identity. □

In view of the above theorem, we write $\alpha \wedge \beta \wedge \gamma$ for $(\alpha \wedge \beta) \wedge \gamma$ or $\alpha \wedge (\beta \wedge \gamma)$, and similarly for products of higher order. Notice that the first two identities from Lemma 5.2.1, together with associativity of the wedge product, immediately imply

$$\begin{aligned}\alpha^1 \wedge \cdots \wedge \alpha^k &= \sum_{\sigma \in S_k} \varepsilon(\sigma)\, \alpha^{\sigma(1)} \otimes \cdots \otimes \alpha^{\sigma(k)} \\ \alpha^{\sigma(1)} \wedge \cdots \wedge \alpha^{\sigma(k)} &= \varepsilon(\sigma)\, \alpha^1 \wedge \cdots \wedge \alpha^k\end{aligned} \tag{5.2.1}$$

for $\alpha^1, \ldots, \alpha^k \in \Lambda_1(V)$, $\sigma \in S_k$. The second identity, in particular, lets us identify a basis of $\Lambda_k(V)$:

Theorem 5.2.2. *Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ denote a basis of V, and $\alpha^1, \ldots, \alpha^n \in \Lambda_1(V)$ the dual basis. Then the set*

$$\{\alpha^{i_1} \wedge \cdots \wedge \alpha^{i_k} \mid 1 \le i_1 < i_2 < \cdots < i_k \le n\}$$

is a basis of $\Lambda_k(V)$, which therefore has dimension $\binom{n}{k}$, where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}.$$

Proof. If $\alpha \in \Lambda_k(V)$, then α is a k-tensor, and by Theorem 5.1.1,

$$\alpha = \sum_{i_1, \ldots, i_k = 1}^{n} \alpha_{i_1 \ldots i_k} \alpha^{i_1} \otimes \cdots \otimes \alpha^{i_k}.$$

Thus,

$$\begin{aligned}\alpha = \alpha_a &= \sum_{j_1, \ldots, j_k = 1}^{n} \alpha_{j_1 \ldots j_k} (\alpha^{j_1} \otimes \cdots \otimes \alpha^{j_k})_a \\ &= \frac{1}{k!} \sum_{j_1, \ldots, j_k = 1}^{n} \alpha_{j_1 \ldots j_k} \alpha^{j_1} \wedge \cdots \wedge \alpha^{j_k}.\end{aligned}$$

Let $i_1 < i_2 < \cdots < i_k$ denote $j_1, \ldots, j_k$ written in increasing order. Then $\alpha^{j_1} \wedge \cdots \wedge \alpha^{j_k} = \pm \alpha^{i_1} \wedge \cdots \wedge \alpha^{i_k}$ by (5.2.1), and the collection in the statement of the theorem spans $\Lambda_k(V)$. To see that it is linearly independent, suppose that

$$\sum_{1 \le j_1 < \cdots < j_k \le n} \alpha_{j_1 \ldots j_k} \alpha^{j_1} \wedge \cdots \wedge \alpha^{j_k} = 0. \tag{5.2.2}$$

Observe that if $i_1 < \cdots < i_k$, then

$$\alpha^{j_1} \wedge \cdots \wedge \alpha^{j_k}(\boldsymbol{v}_{i_1}, \ldots, \boldsymbol{v}_{i_k}) = \sum_{\sigma \in S_k} \varepsilon(\sigma) \alpha^{j_1}(\boldsymbol{v}_{\sigma(i_1)}) \cdots \alpha^{j_k}(\boldsymbol{v}_{\sigma(i_k)}), \tag{5.2.3}$$

where S_k denotes all the permutations of $\{i_1, \ldots, i_k\}$. Since $\alpha^{j_l}(\boldsymbol{v}_{i_l}) = \delta_{i_l j_l}$, the only nonzero terms in this sum are those for which $\sigma(i_l) = j_l$, $l = 1, \ldots, k$, in which case they equal $\varepsilon(\sigma)$. In particular, as sets, $\{i_1, \ldots, i_k\} = \{\sigma(i_1), \ldots, \sigma(i_k)\} = \{j_1, \ldots, j_k\}$. But $i_1 < \cdots < i_k$ and $j_1 < \cdots < j_k$, so all terms vanish except when $i_1 = j_1, \ldots, i_k = j_k$, and the sum then equals 1. Thus, for any $i_1 < \cdots < i_k$, applying both sides of (5.2.2) to $\boldsymbol{v}_{i_1}, \ldots, \boldsymbol{v}_{i_k}$ yields

$$0 = \sum_{1 \le j_1 < \cdots < j_k \le n} \alpha_{j_1 \ldots j_k} \alpha^{j_1} \wedge \cdots \wedge \alpha^{j_k}(\boldsymbol{v}_{i_1}, \ldots, \boldsymbol{v}_{i_k}) = \alpha_{i_1 \ldots i_k},$$

which establishes linear independence.

To conclude the proof, it remains to show that the dimension is $\binom{n}{k}$ as claimed. The size of the basis is the number of distinct k elements chosen from a set of n elements written in increasing order. Now the number of distinct ordered sets of k elements is $n(n-1)\cdots(n-k+1)$, since there are n choices for the first one, $n-1$ for the second, and so on. Since there are $k!$ ways of ordering a given set of k elements, the claim follows. □

The theorem says in particular that $\Lambda_n(V)$, where n is the dimension of V, is one-dimensional. This we already knew of course from Theorem 1.3.1: on $\mathbb{R}^n$, any n-form is a multiple of the determinant, so $\Lambda_n(\mathbb{R}^n)$ is one-dimensional. A choice of basis for V induces an isomorphism of V with $\mathbb{R}^n$, which in turn induces one of $\Lambda_n(V)$ with $\Lambda_n(\mathbb{R}^n)$.

Observe also that if $\{\varepsilon^i\}$ is the basis of $\mathbb{R}^{n*}$ dual to $\{\boldsymbol{e}_i\}$, then

$$\varepsilon^1 \wedge \cdots \wedge \varepsilon^n = \det. \tag{5.2.4}$$

Indeed, the form on the left equals c times det for some $c \in \mathbb{R}$. Thus,

$$\varepsilon^1 \wedge \cdots \wedge \varepsilon^n(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n) = c \cdot \det(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n) = c.$$

But by (5.2.3), the term on the left equals 1. The factorials in the definition of the wedge product were chosen in order for (5.2.4) to hold.

Definition 5.2.2. An *orientation* on an n-dimensional vector space V is a choice of a nonzero $\omega \in \Lambda_n(V)$. Given any two nonzero $\omega_1, \omega_2 \in \Lambda_n(V)$, $\omega_1 = c \cdot \omega_2$ for some $c \neq 0$. If $c > 0$, we say ω_1 and ω_2 determine the *same orientation*. Otherwise, they determine opposite orientations.

Clearly, there are only two possible orientations on a given vector space. The *standard orientation* on $\mathbb{R}^n$ is the one induced by the determinant. In an oriented vector space V, if $\omega \in \Lambda_n(V)$ induces the given orientation, an ordered basis $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ of V (i.e., an n-tuple consisting of basis elements) is said to be *positively oriented* if $\omega(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) > 0$.

The following result deals with *decomposable* elements of $\Lambda_k(V)$; i.e., elements of the form $\alpha^1 \wedge \cdots \wedge \alpha^k$, where $\alpha^i \in V^*$ (as opposed to a sum of such terms):

Corollary 5.2.1. *Let V be an n-dimensional vector space, $\boldsymbol{v}_i \in V$, $\alpha^j \in V^*$, $1 \leq i, j \leq k \leq n$. Then*

(1) $(\alpha^1 \wedge \cdots \wedge \alpha^k)(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) = \det(\alpha_i(\boldsymbol{v}_j))$;

(2) $\alpha^1 \wedge \cdots \wedge \alpha^k \neq 0$ *if and only if* $\alpha^1, \ldots, \alpha^k$ *are linearly independent;*

(3) *There is a bijective correspondence between one-dimensional decomposable subspaces of $\Lambda_k(V)$ and k-dimensional subspaces of V.*

Proof. For (1), (5.2.1) implies that

$$(\alpha^1 \wedge \cdots \wedge \alpha^k)(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) = \sum_{\sigma \in S_k} \varepsilon(\sigma) \alpha^{\sigma(1)}(\boldsymbol{v}_1) \cdots \alpha^{\sigma(k)}(\boldsymbol{v}_k) = \det(\alpha^i(\boldsymbol{v}_j)$$

by definition of the determinant.

For (2), if $\alpha^1, \ldots, \alpha^k$ are linearly dependent, then one of them, say α^1, is a linear combination $\sum c_j \alpha^j$ of the others. But then

$$\alpha^1 \wedge \cdots \wedge \alpha^k = \sum_{j=2}^{k} c_j \alpha^j \wedge \alpha_2 \wedge \cdots \wedge \alpha^k = 0$$

since α^j appears twice in each term inside the summation. On the other hand, if they are linearly independent, then they can be extended to a basis of V^*, and by Theorem 5.2.2, $\alpha^1 \wedge \cdots \wedge \alpha^k$ is one of the corresponding basis elements of $\Lambda_k(V)$. Being a basis element, it cannot be zero.

For the third statement, the correspondence

$$\operatorname{span}\{\alpha^1 \wedge \cdots \wedge \alpha^k\} \longleftrightarrow \operatorname{span}\{\alpha^1, \ldots, \alpha^k\}$$

where $\alpha^1, \ldots, \alpha^k$ are linearly independent in V^* is a bijective one between one-dimensional decomposable subspaces of $\Lambda_k(V)$ and k-dimensional subspaces of V^* by (2). The claim follows in view of the isomorphism between a space and its dual. □

Let us return for a moment to the topic of orientation. Suppose that in addition to an orientation, V is endowed with an inner product. If $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ is a positively oriented orthonormal basis of V, then there exists a unique n-form α on V such that $\alpha(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) = 1$; in fact, $\alpha = \alpha^1 \wedge \cdots \wedge \alpha^n$, where $\alpha^1, \ldots, \alpha^n$ is the basis of V^* dual to $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$. α is called the *volume form* of V. The reason we used a definite rather than indefinite article for volume form is because $\alpha(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n) = 1$ for *any* positively oriented orthonormal basis $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$ of V, which means that such an α is unique:

Theorem 5.2.3. *Let V be an n-dimensional vector space.*

(1) *If $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ is a basis of V, and $\alpha \in \Lambda_n(V)$, then*

$$\alpha(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n) = \det(a_{ij}) \alpha(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n),$$

for any $\boldsymbol{w}_i = \sum_j a_{ij} \boldsymbol{v}_j \in V$.

(2) *Suppose that in addition, V is oriented and endowed with an inner product. If ω is the corresponding volume form, then for any positively oriented orthonormal basis $\boldsymbol{w}_1, \dots, \boldsymbol{w}_n$ of V, $\omega(\boldsymbol{w}_1, \dots, \boldsymbol{w}_n) = 1$.*

Proof. Define an n-form β on $\mathbb{R}^n$ by

$$\beta\left(\begin{bmatrix} b_{11} \\ \vdots \\ b_{n1} \end{bmatrix}, \dots, \begin{bmatrix} b_{1n} \\ \vdots \\ b_{nn} \end{bmatrix}\right) = \alpha\left(\sum_i b_{i1}\boldsymbol{v}_i, \dots, \sum_i b_{in}\boldsymbol{v}_i\right),$$

for $b_{ij} \in \mathbb{R}$, $1 \le i, j \le n$. By Theorem 1.3.1, $\beta = c \cdot \det$ for some $c \in \mathbb{R}$, and

$$c = c \cdot \det(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) = \beta(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) = \alpha(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n).$$

Thus,

$$\alpha(\boldsymbol{w}_1, \dots, \boldsymbol{w}_n) = \beta\left(\begin{bmatrix} a_{11} \\ \vdots \\ a_{n1} \end{bmatrix}, \dots, \begin{bmatrix} a_{1n} \\ \vdots \\ a_{nn} \end{bmatrix}\right) = \det(a_{ij})\alpha(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n),$$

which establishes the first claim. For the second one, let $\omega = \omega^1 \wedge \cdots \wedge \omega^n$, and $\mathcal{B} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ the positively oriented orthonormal basis of V dual to $(\omega^1, \dots, \omega^n)$. If $\mathcal{C}$ is the basis $(\boldsymbol{w}_1, \dots, \boldsymbol{w}_n)$, then by what was just proved,

$$\omega(\boldsymbol{w}_1, \dots, \boldsymbol{w}_n) = \det[1_V]_{\mathcal{C},\mathcal{B}} \cdot \omega(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n) = \det[1_V]_{\mathcal{C},\mathcal{B}},$$

and it remains to show this determinant equals 1. Now, $[1_V]_{\mathcal{C},\mathcal{B}}$ is also the matrix with respect to $\mathcal{B}$ of the operator L which maps $\boldsymbol{v}_i$ to $\boldsymbol{w}_i$, $i = 1, \dots, n$, because

$$[\boldsymbol{w}_i]_{\mathcal{B}} = [1_V]_{\mathcal{C},\mathcal{B}}[\boldsymbol{w}_i]_{\mathcal{C}} = [1_V]_{\mathcal{C},\mathcal{B}}\,\boldsymbol{e}_i, \qquad i = 1, \dots, n,$$

on the one hand, and $[\boldsymbol{w}_i]_{\mathcal{B}} = [L]_{\mathcal{B}}\boldsymbol{e}_i$ by definition of L on the other. But L maps an orthonormal basis to another one, so it is a linear isometry. In particular, $[L]_{\mathcal{B}}[L]_{\mathcal{B}}^T = I_n$, which implies that $(\det L)^2 = 1$, so that $\det L = \pm 1$. Summarizing, $\omega(\boldsymbol{w}_1, \dots, \boldsymbol{w}_n) = \pm 1$. Finally, $\omega(\boldsymbol{w}_1, \dots, \boldsymbol{w}_n) > 0$ since $\mathcal{C}$ is positively oriented, so it must equal one. This completes the proof. □

2-forms on V possess an additional property, which in turn has several consequences explored in the exercises:

Proposition 5.2.2. *Given any nonzero form $\alpha \in \Lambda_2(V)$, there exists a basis $\alpha^1, \dots, \alpha^n$ of $\Lambda_1(V)$ such that*

$$\alpha = \alpha^1 \wedge \alpha^2 + \alpha^3 \wedge \alpha^4 + \cdots + \alpha^{2k-1} \wedge \alpha^{2k}.$$

Proof. It suffices to construct a basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ of V such that

$$\alpha(\boldsymbol{v}_{2i-1}, \boldsymbol{v}_{2i}) = 1 \text{ if } i \le k, \qquad \alpha(\boldsymbol{v}_i, \boldsymbol{v}_j) = 0 \text{ if } i \text{ or } j > k, \text{ or if } |j - i| > 1,$$

since the dual basis will then satisfy the claim. We argue by induction on the dimension of V. If the dimension is 1, there is nothing to prove, so assume the claim is true in all dimensions less than n. Since $\alpha \neq 0$, there exist $\boldsymbol{v}_1, \boldsymbol{w}_2 \in V$ such that $\alpha(\boldsymbol{v}_1, \boldsymbol{w}_2) \neq 0$. If $\boldsymbol{v}_2 = \boldsymbol{w}_2/\alpha(\boldsymbol{v}_1, \boldsymbol{w}_2)$, then $\alpha(\boldsymbol{v}_1, \boldsymbol{v}_2) = 1$. Let W be the subspace of V spanned by $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, and define

$$Z = \{\boldsymbol{v} \in V \mid \alpha(\boldsymbol{v}, \boldsymbol{v}_1) = \alpha(\boldsymbol{v}, \boldsymbol{v}_2) = 0\}.$$

By construction, Z is a subspace of V that intersects W in the zero vector only. Furthermore, Z has dimension at least $n-2$: indeed, $Z = \ker \beta^1 \cap \ker \beta^2$, where $\beta^i \in V^*$ is given by $\beta^i(\boldsymbol{v}) = \alpha(\boldsymbol{v}, \boldsymbol{v}_i)$, $i = 1, 2$, $\boldsymbol{v} \in V$, and each kernel has dimension $n-1$, see also Exercise 1.21. By that same exercise,

$$\begin{aligned} \dim Z &= \dim \ker \beta^1 + \dim \ker \beta^2 - \dim(\ker \beta^1 + \ker \beta^2) \\ &\geq \dim \ker \beta^1 + \dim \ker \beta^2 - \dim V \\ &= n - 2. \end{aligned}$$

Since W is two-dimensional and $Z \cap W = \{\boldsymbol{0}\}$, $V = W \oplus Z$. The result now follows by the induction hypothesis applied to the restriction of α to Z. □

Even though the tensor product is not commutative, i.e., $S \otimes T$ need not equal $T \otimes S$, the wedge product is "almost" commutative:

$$\alpha \wedge \beta = (-1)^{kl} \beta \wedge \alpha, \qquad \alpha \in \Lambda_k(V), \quad \beta \in \Lambda_l(V). \tag{5.2.5}$$

This is easily seen by writing both forms in terms of a basis and using (5.2.1), see Exercise 5.5.

Examples 5.2.1. (i) Let $\alpha \in \Lambda_1(V)$, $\beta \in \Lambda_2(V)$. Then $\alpha \wedge \beta = \beta \wedge \alpha$, and for any $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V$,

$$\begin{aligned} (\alpha \wedge \beta)(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) &= \frac{1}{2}[\alpha(\boldsymbol{x})\beta(\boldsymbol{y}, \boldsymbol{z}) - \alpha(\boldsymbol{x})\beta(\boldsymbol{z}, \boldsymbol{y}) - \alpha(\boldsymbol{y})\beta(\boldsymbol{x}, \boldsymbol{z}) + \alpha(\boldsymbol{y})\beta(\boldsymbol{z}, \boldsymbol{x}) \\ &\qquad + \alpha(\boldsymbol{z})\beta(\boldsymbol{x}, \boldsymbol{y}) - \alpha(\boldsymbol{z})\beta(\boldsymbol{y}, \boldsymbol{x})] \\ &= \alpha(\boldsymbol{x})\beta(\boldsymbol{y}, \boldsymbol{z}) + \alpha(\boldsymbol{y})\beta(\boldsymbol{z}, \boldsymbol{x}) + \alpha(\boldsymbol{z})\beta(\boldsymbol{x}, \boldsymbol{y}) \\ &= \circlearrowright\, (\alpha \otimes \beta)(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}), \end{aligned}$$

where $\circlearrowright$ denotes cyclic summation: given a 3-tensor T,

$$\circlearrowright\, T(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = T(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) + T(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}) + T(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{y}).$$

(ii) Let $\alpha, \beta \in \Lambda^1(V) = V^*$. Recall from Corollary 5.2.1(2) that if $\alpha \neq 0$ and $\alpha \wedge \beta = 0$, then $\beta \in \operatorname{span}\{\alpha\}$. The following generalization is known as Cartan's lemma: Suppose $\alpha^1, \ldots, \alpha^k \in V^*$ are linearly independent. If $\beta^1, \ldots, \beta^k \in V^*$ satisfy

$$\sum_{i=1}^{k} \alpha^i \wedge \beta^i = 0,$$

then each $\beta^i \in \operatorname{span}\{\alpha^1, \ldots, \alpha^k\}$. Furthermore, if $\beta^i = \sum_j a_{ij}\alpha^j$, then $a_{ij} = a_{ji}$.

To see this, extend the α^i's to a basis $\alpha^1, \dots, \alpha^n$ of V^*, and write

$$\beta^i = \sum_{j=1}^{k} a_{ij}\alpha^j + \sum_{j=k+1}^{n} b_{ij}\alpha^j.$$

Then

$$0 = \sum_{i=1}^{k} \alpha^i \wedge \beta^i = \sum_{1\le i<j\le k} (a_{ij} - a_{ji})\alpha^i \wedge \alpha^j + \sum_{i\le k<j} b_{ij}\alpha^i \wedge \alpha^j,$$

and the claim follows from linear independence of the $\alpha^i \wedge \alpha^j$, $i < j$.

5.3 Differential forms

The same procedure used in going from tensors to tensor fields (or from vectors to vector fields) can be applied to forms. The result is not, however, traditionally called a "form field":

Definition 5.3.1. A *differential k-form*, or simply a differential form, on a manifold M is a map ω that assigns to each $\boldsymbol{p} \in M$ an element $\omega(\boldsymbol{p})$ of $\Lambda_k(M_{\boldsymbol{p}})$. ω is assumed to be smooth in the sense that for any vector fields $\boldsymbol{X}_1, \dots, \boldsymbol{X}_k$ on M, the function

$$\begin{aligned} \omega(\boldsymbol{X}_1, \dots, \boldsymbol{X}_k) : M &\longrightarrow \mathbb{R}, \\ \boldsymbol{p} &\longmapsto \omega(\boldsymbol{p})(\boldsymbol{X}_1(\boldsymbol{p}), \dots, \boldsymbol{X}_k(\boldsymbol{p})) \end{aligned}$$

is differentiable.

Since a differential form is a tensor field, the results established for tensor fields hold in this context as well. Thus, for example, smoothness may be rephrased by requiring that for any $\boldsymbol{p} \in M$, there be a chart $(U, \boldsymbol{x})$ around $\boldsymbol{p}$ such that the functions $\omega(\partial/\partial x^{i_1}, \dots, \partial/\partial x^{i_k})$ are differentiable on U. Furthermore, if we denote these functions by $\omega_{i_1 \dots i_k}$, then the restriction of ω to U is given by

$$\omega_{|U} = \sum_{1\le i_1<\cdots<i_k\le n} \omega_{i_1 \dots i_k}\, \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}.$$

Similarly, the collection of all differential k-forms on M is a vector space, denoted $\Lambda_k(M)$. Finally, Proposition 5.1.1 implies that given $\boldsymbol{f} : M \to N$, α, β k-forms on N, γ an l-form on N, $h : \mathbb{N} \to \mathbb{R}$, and real numbers a, b,

$$\begin{aligned} \boldsymbol{f}^*(a\alpha + b\beta) &= a\boldsymbol{f}^*\alpha + b\boldsymbol{f}^*\beta, \\ \boldsymbol{f}^*(\alpha \wedge \gamma) &= \boldsymbol{f}^*\alpha \wedge \boldsymbol{f}^*\gamma, \\ \boldsymbol{f}^*(h\alpha) &= (h \circ \boldsymbol{f})\boldsymbol{f}^*\alpha. \end{aligned} \tag{5.3.1}$$

There is one construction that is specific to forms as opposed to general tensor fields: let us agree to call a function on M a differential 0-form. Taking differentials of func-

tions is then a linear map $\mathrm{d} : \Lambda_0(M) \to \Lambda_1(M)$. We now extend it to an operator $\mathrm{d} : \Lambda_k(M) \to \Lambda_{k+1}(M)$ for any nonnegative integer k as follows: let α be a differential k-form, $\boldsymbol{p} \in M$. In order to define $\mathrm{d}\alpha(\boldsymbol{p})$, consider a chart $(U, \boldsymbol{x})$ of M around $\boldsymbol{p}$. Then as noted earlier, the restriction of α to U may be written as $\alpha_{|U} = \sum_{1 \le i_1 < \cdots < i_k \le n} \alpha_{i_1 \ldots i_k}\, \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}$. Define

$$(\mathrm{d}\alpha)_{|U} := \sum_{1 \le i_1 < \cdots < i_k \le n} \mathrm{d}\alpha_{i_1 \ldots i_k} \wedge \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}. \tag{5.3.2}$$

It must be checked that this formula is independent of the chosen chart. To see this, we first claim that d satisfies the following properties at $\boldsymbol{p}$:

(1) $\mathrm{d}\alpha(\boldsymbol{p}) \in \Lambda_{k+1}(M_{\boldsymbol{p}})$;
(2) if $\alpha = \beta$ on some neighborhood of $\boldsymbol{p}$, then $\mathrm{d}\alpha(\boldsymbol{p}) = \mathrm{d}\beta(\boldsymbol{p})$;
(3) $\mathrm{d}(a\alpha + b\beta)(\boldsymbol{p}) = a\,\mathrm{d}\alpha(\boldsymbol{p}) + b\,\mathrm{d}\beta(\boldsymbol{p}), \qquad a, b \in \mathbb{R}, \quad \beta \in \Lambda_k(M)$;
(4) $\mathrm{d}(\alpha \wedge \beta)(\boldsymbol{p}) = \mathrm{d}\alpha(\boldsymbol{p}) \wedge \beta(\boldsymbol{p}) + (-1)^k \alpha(\boldsymbol{p}) \wedge \mathrm{d}\beta(\boldsymbol{p}), \qquad \beta \in \Lambda_l(M)$;
(5) $\mathrm{d}(\mathrm{d}f)(\boldsymbol{p}) = 0$ for $f \in \Lambda_0(M)$.

The first three items follow directly from (5.3.2). To prove the fourth one, it suffices to check it in the case that $\alpha = f\,\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}$ and $\beta = g\,\mathrm{d}x^{j_1} \wedge \cdots\, \mathrm{d}x^{j_l}$ by the linearity of d established in (3). Furthermore, the identity certainly holds if α and/or β are functions (i.e., zero forms). Define $\gamma = \mathrm{d}x^{j_1} \wedge \cdots\, \mathrm{d}x^{j_l}$, so that $\beta = g\gamma$. Then

$$\begin{aligned}
\mathrm{d}(\alpha \wedge \beta)(\boldsymbol{p}) &= \mathrm{d}(fg\,\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k} \wedge \gamma)(\boldsymbol{p}) \\
&= (\mathrm{d}f(\boldsymbol{p})g(\boldsymbol{p}) + f(\boldsymbol{p})\,\mathrm{d}g(\boldsymbol{p})) \wedge (\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k} \wedge \gamma)(\boldsymbol{p}) \\
&= (\mathrm{d}f \wedge \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k})(\boldsymbol{p}) \wedge (g\gamma)(\boldsymbol{p}) \\
&\quad + (-1)^k (f\,\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k})(\boldsymbol{p}) \wedge (\mathrm{d}g \wedge \gamma)(\boldsymbol{p}) \\
&= \mathrm{d}\alpha(\boldsymbol{p}) \wedge \beta(\boldsymbol{p}) + (-1)^k \alpha(\boldsymbol{p}) \wedge \mathrm{d}\beta(\boldsymbol{p}).
\end{aligned}$$

The last identity can be seen by writing $(\mathrm{d}f)_{|U} = \sum_i \partial/\partial x^i(f)\,\mathrm{d}x^i$, so that

$$\begin{aligned}
\mathrm{d}(\mathrm{d}f)(\boldsymbol{p}) &= \left(\sum_i \mathrm{d}\left(\frac{\partial}{\partial x^i} f\right) \wedge \mathrm{d}x^i \right)(\boldsymbol{p}) = \sum_{i,j} \left(\left(\frac{\partial}{\partial x^j} \frac{\partial}{\partial x^i} f\right) \mathrm{d}x^j \wedge \mathrm{d}x^i \right)(\boldsymbol{p}) \\
&= \sum_{j<i} \left(\left(\frac{\partial}{\partial x^i} \frac{\partial}{\partial x^j} - \frac{\partial}{\partial x^j} \frac{\partial}{\partial x^i} f\right) \mathrm{d}x^j \wedge \mathrm{d}x^i \right)(\boldsymbol{p}) \\
&= 0.
\end{aligned}$$

We can now justify the claim made earlier that the definition of d does not depend on the chosen chart: indeed, suppose $\tilde{\mathrm{d}}$ is the operator obtained by using a different chart. It suffices to show that

$$\tilde{\mathrm{d}}(f\,\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}) = \mathrm{d}f \wedge \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}$$

for any function f by linearity of $\tilde{\mathrm{d}}$. But $\tilde{\mathrm{d}}f = \mathrm{d}f$ and $\tilde{\mathrm{d}}$ satisfies properties (1) through (5), so that

$$\tilde{\mathrm{d}}(f\,\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}) = \mathrm{d}f \wedge \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k} + f\,\tilde{\mathrm{d}}(\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k})$$

by (4). It remains to show that the last term in the above identity is zero. But again by (4), this term is a sum over j of expressions of the form

$$\pm\, \mathrm{d}x^{i_1} \wedge \cdots \wedge \tilde{\mathrm{d}}(\mathrm{d}x^{i_j}) \wedge \cdots \wedge \mathrm{d}x^{i_k},$$

and $\tilde{\mathrm{d}}(\mathrm{d}x^{i_j}) = 0$ by (5). This shows that d is well defined.

Theorem 5.3.1. *There exists a unique linear map* $\mathrm{d} : \Lambda_k(M) \to \Lambda_{k+1}(M)$, $k = 0, 1, 2, \ldots$, *called the* exterior derivative operator, *that satisfies*

(1) $\mathrm{d}(\alpha \wedge \beta) = (\mathrm{d}\alpha) \wedge \beta + (-1)^k \alpha \wedge (\mathrm{d}\beta), \qquad \alpha \in \Lambda_k(M)$,

(2) $\mathrm{d} \circ \mathrm{d} = 0$, *and*

(3) *for* $f \in \Lambda_0(M)$, $\mathrm{d}f$ *is the differential of* f.

Proof. Existence has already been established. For uniqueness, it suffices to show that a linear map as in the statement necessarily satisfies the five properties listed earlier in the definition of d. All but the second property are clear. So let $\tilde{\mathrm{d}}$ be a map satisfying the conditions of the statement, and suppose that $\alpha = \beta$ on a neighborhood U of $\boldsymbol{p}$. Set $\gamma = \alpha - \beta$. Then $\gamma_{|U} \equiv 0$, and we must show that $\tilde{\mathrm{d}}\gamma(\boldsymbol{p}) = 0$. To see this is so, consider a neighborhood V of $\boldsymbol{p}$ whose closure lies in U, and a function φ with values between zero and one, which equals 1 on V and has support in U. Then $\gamma = (1 - \varphi)\gamma$, so that

$$\tilde{\mathrm{d}}\gamma(\boldsymbol{p}) = \mathrm{d}(1 - \varphi)(\boldsymbol{p}) \wedge \gamma(\boldsymbol{p}) + (1 - \varphi)(\boldsymbol{p})\, \tilde{\mathrm{d}}\gamma(\boldsymbol{p}) = 0. \qquad \square$$

In the case of $\mathbb{R}^n$, the definition we gave of exterior derivative is intrinsic, but on a generic manifold it is somewhat unsatisfactory to rely on a definition in terms of charts. The following property, which could have been used as a definition, seeks to remedy this:

Theorem 5.3.2. *For* $\alpha \in \Lambda_k(M)$, $\boldsymbol{X}_i \in \mathfrak{X}M$, $i = 0, \ldots, k$,

$$\begin{aligned}\mathrm{d}\alpha(\boldsymbol{X}_0, \ldots, \boldsymbol{X}_k) = &\sum_{i=0}^{k} (-1)^i \boldsymbol{X}_i(\alpha(\boldsymbol{X}_0, \ldots, \hat{\boldsymbol{X}}_i, \ldots, \boldsymbol{X}_k)) \\ &\sum_{i<j} (-1)^{i+j} \alpha([\boldsymbol{X}_i, \boldsymbol{X}_j], \boldsymbol{X}_0, \ldots, \hat{\boldsymbol{X}}_i, \ldots, \hat{\boldsymbol{X}}_j, \ldots, \boldsymbol{X}_k).\end{aligned}$$

(We use $\hat{\boldsymbol{X}}_i$ *to indicate that* $\boldsymbol{X}_i$ *does not appear in the list).*

Proof. We prove the identity for $k = 1$, the general case being similar; i.e., we will show that

$$\mathrm{d}\alpha(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{X}(\alpha(\boldsymbol{Y})) - \boldsymbol{Y}(\alpha(\boldsymbol{X})) - \alpha([\boldsymbol{X}, \boldsymbol{Y}]).$$

We first establish that the right side of the above equation is tensorial in $\boldsymbol{X}$ and $\boldsymbol{Y}$. It is certainly linear over vector fields, so that by Theorem 3.8.1 we only need to check linearity over functions. Now, for $f : M \to \mathbb{R}$, replacing $\boldsymbol{X}$ by $f\boldsymbol{X}$ in the last two terms

(the first one is trivially linear) on the right yields

$$\begin{aligned} -\boldsymbol{Y}(\alpha(f\boldsymbol{X})) - \alpha[f\boldsymbol{X}, \boldsymbol{Y}] &= -\boldsymbol{Y}(f(\alpha\boldsymbol{X})) - \alpha[f\boldsymbol{X}, \boldsymbol{Y}] \\ &= -(\boldsymbol{Y}f)\alpha(\boldsymbol{X}) - f\boldsymbol{Y}(\alpha\boldsymbol{X}) - \alpha(f[\boldsymbol{X}, \boldsymbol{Y}] - (\boldsymbol{Y}f)\boldsymbol{X}) \\ &= -f(\boldsymbol{Y}(\alpha(\boldsymbol{X})) - \alpha[\boldsymbol{X}, \boldsymbol{Y}]), \end{aligned}$$

which shows that it is indeed tensorial in the first argument. Since α is skew-symmetric, it is also tensorial in the second argument. In light of this, the identity need only be established for coordinate vector fields. So let $(U, \boldsymbol{x})$ be a chart. Then the restriction of α to U equals $\sum_l \alpha(\partial/\partial x^l)\,\mathrm{d}x^l$, so that

$$\begin{aligned} \mathrm{d}\alpha_{|U} &= \sum_l \mathrm{d}\left(\alpha\frac{\partial}{\partial x^l}\right) \wedge \mathrm{d}x^l = \sum_{k,l} \frac{\partial}{\partial x^k}\left(\alpha\frac{\partial}{\partial x^l}\right)\mathrm{d}x^k \wedge \mathrm{d}x^l \\ &= \sum_{k<l}\left(\frac{\partial}{\partial x^k}(\alpha\frac{\partial}{\partial x^l}) - \frac{\partial}{\partial x^l}(\alpha\frac{\partial}{\partial x^k})\right)\mathrm{d}x^k \wedge \mathrm{d}x^l. \end{aligned}$$

This means that

$$\mathrm{d}\alpha\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) = \frac{\partial}{\partial x^i}\left(\alpha\frac{\partial}{\partial x^j}\right) - \frac{\partial}{\partial x^j}\left(\alpha\frac{\partial}{\partial x^i}\right),$$

which is the desired identity for coordinate vector fields. □

Another important property of the exterior derivative is that it commutes with pullbacks:

Theorem 5.3.3. *Let M, N be manifolds with exterior derivative operators d^M and d^N respectively. Given a map $\boldsymbol{f} : M \to N$, $\boldsymbol{f}^* \circ \mathrm{d}^N = \mathrm{d}^M \circ \boldsymbol{f}^*$.*

Proof. Having emphasized the fact that the two exterior derivative operators live in different spaces, we omit the superscript for brevity. Let us first consider the case of a zero-form $h : N \to \mathbb{R}$; it must be established that $\boldsymbol{f}^*\,\mathrm{d}h = \mathrm{d}(h \circ \boldsymbol{f})$. This is nothing more than the chain rule together with an exercise in notation, once we recall from Chapter 3 that by definition, given $\boldsymbol{q} \in N$, $\boldsymbol{u} \in N_{\boldsymbol{q}}$,

$$\mathrm{d}h(\boldsymbol{q})\boldsymbol{u} = \mathcal{I}^{-1}_{h(\boldsymbol{q})}(h_{*\boldsymbol{q}}\boldsymbol{u}).$$

Indeed, for $\boldsymbol{p} \in M$ and $\boldsymbol{v} \in M_{\boldsymbol{p}}$,

$$\begin{aligned} \boldsymbol{f}^*\,\mathrm{d}h(\boldsymbol{p})(\boldsymbol{v}) &= \mathrm{d}h(\boldsymbol{f}(p))(\boldsymbol{f}_{*\boldsymbol{p}}\boldsymbol{v}) = \mathcal{I}^{-1}_{(h\circ\boldsymbol{f})(\boldsymbol{p})}h_*(\boldsymbol{f}_{*\boldsymbol{p}}\boldsymbol{v}) = \mathcal{I}^{-1}_{(h\circ\boldsymbol{f})(\boldsymbol{p})}((h\circ\boldsymbol{f})_{*\boldsymbol{p}}\boldsymbol{v}) \\ &= d(h\circ\boldsymbol{f})(\boldsymbol{p})(\boldsymbol{v}), \end{aligned}$$

as claimed. In the general case, linearity of both sides in the identity allows us to only consider a k-form on N of type $h\,\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}$. Then

$$\begin{aligned}
\mathrm{d}(\boldsymbol{f}^* h\,\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}) &= \mathrm{d}\left((h \circ \boldsymbol{f})\boldsymbol{f}^*\,\mathrm{d}x^{i_1} \wedge \cdots \wedge \boldsymbol{f}^*\,\mathrm{d}x^{i_k}\right) \\
&= \mathrm{d}\left((h \circ \boldsymbol{f})\,\mathrm{d}\boldsymbol{f}^* x^{i_1} \wedge \cdots \wedge \mathrm{d}\boldsymbol{f}^* x^{i_k}\right) \\
&= \mathrm{d}(h \circ \boldsymbol{f}) \wedge \mathrm{d}\boldsymbol{f}^* x^{i_1} \wedge \cdots \wedge \mathrm{d}\boldsymbol{f}^* x^{i_k} \text{ since } \mathrm{d} \circ \mathrm{d} = 0 \\
&= \boldsymbol{f}^*\,\mathrm{d}h \wedge \boldsymbol{f}^*\,\mathrm{d}x^{i_1} \wedge \cdots \wedge \boldsymbol{f}^*\,\mathrm{d}x^{i_k} \\
&= \boldsymbol{f}^*(\,\mathrm{d}h \wedge \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}) \\
&= \boldsymbol{f}^*\,\mathrm{d}(h\,\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k}).
\end{aligned}$$

This completes the proof of the Theorem. □

5.4 Integration on manifolds

Our goal in this section is to define the integral of a form of degree k on a differentiable manifold of dimension k. The form will always be assumed to have compact support, and the manifold will need to be endowed with an orientation, in the following sense:

Definition 5.4.1. An n-dimensional manifold M is said to be *orientable* if it admits a nowhere-zero n-form.

Given any two nowhere-zero n-forms ω_1 and ω_2 on M^n, there exists a function $f : M \to \mathbb{R}$, which is either positive everywhere or negative everywhere, such that $\omega_1 = f\omega_2$. The two forms are said to be equivalent if $f > 0$. Thus, the collection of all non-zero n-forms splits into two disjoint subsets, called equivalence classes. An *orientation* of M is a choice of one of these two classes. Clearly, an orientable manifold has exactly two orientations, and the property of being orientable may be thought of as being able to orient all tangent spaces in a continuous manner. The *standard orientation* of $\mathbb{R}^n$ is the equivalence class containing $\mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^n$. This is often described as the orientation *induced by* $\mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^n$.

Proposition 5.4.1. *M is orientable if and only if it admits an atlas consisting of charts whose transition functions have positive Jacobian determinant; i.e, for $(U, \boldsymbol{x})$ and $(V, \boldsymbol{y})$ in the atlas,* $\det D(\boldsymbol{y} \circ \boldsymbol{x}^{-1}) > 0$ *if* $U \cap V \neq \emptyset$.

Proof. Suppose M is orientable, and consider an n-form ω on M which vanishes nowhere. Choose any atlas of M. By reordering the components of the coordinate maps, if necessary, it may be assumed that for any chart $(U, \boldsymbol{x})$, $f_{\boldsymbol{x}} := \omega(\partial/\partial x^1, \ldots, \partial/\partial x^n) > 0$. Since ω, when restricted to U, equals $f_{\boldsymbol{x}}\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n$, we have for overlapping charts $(U, \boldsymbol{x})$ and $(V, \boldsymbol{y})$,

$$\frac{1}{f_{\boldsymbol{y}}}\omega = \mathrm{d}y^1 \wedge \cdots \wedge \mathrm{d}y^n = \frac{f_{\boldsymbol{x}}}{f_{\boldsymbol{y}}}\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n$$

on $U \cap V$. By Theorem 5.2.3, $f_{\boldsymbol{x}}/f_{\boldsymbol{y}} = \det D(\boldsymbol{y} \circ \boldsymbol{x}^{-1})$, and the latter is therefore positive. Conversely, suppose there exists an atlas $\{(U_\alpha, \boldsymbol{x}_\alpha)\}$ whose transition functions have positive Jacobian. Consider a partition of unity $\{\varphi_i\}$ subordinate to the atlas, and for each i choose a chart $(U_i, \boldsymbol{x}_i)$ in the atlas that contains the support of φ_i. Define forms ω_i on M by

$$\omega_i(\boldsymbol{p}) = \begin{cases} \varphi_i(\boldsymbol{p}) \left(\mathrm{d}x_i^1 \wedge \cdots \wedge \mathrm{d}x_i^n \right)(\boldsymbol{p}), & \text{if } \boldsymbol{p} \in U_i \\ 0 & \text{otherwise.} \end{cases}$$

Then the form $\omega = \sum_i \omega_i$ cannot vanish anywhere, because if $\boldsymbol{q} \in U_j$, then $\omega(\boldsymbol{q})$ applied to the n-tuple $(\partial/\partial x_j^1, \ldots, \partial/\partial x_j^n)$ of vector fields at $\boldsymbol{q}$ is a sum of nonnegative terms, at least one of which is positive (namely $\omega_i(\boldsymbol{q})$ for any i such that $\varphi_i(\boldsymbol{q}) > 0$). □

One example of nonorientable manifold is the Möbius strip described in Chapter 7.

Given manifolds M_i with orientations respectively induced by n-forms ω_i, $i = 1, 2$, a diffeomorphism $\boldsymbol{f} : M_1 \to M_2$ is said to be *orientation-preserving* if $\boldsymbol{f}^* \omega_2$ induces the same orientation of M_1 as ω_1. An orientation of M also induces an orientation of any open subset U of M, namely the one defined by $\imath^* \omega$, where ω represents the orientation of M and $\imath : U \hookrightarrow M$ denotes the inclusion map. A chart $(U, \boldsymbol{x})$ of an oriented manifold M is said to be *positive* if $\boldsymbol{x} : U \to \boldsymbol{x}(U)$ is orientation-preserving for the standard orientation of $\mathbb{R}^n$.

In order to define integration of an n-form on a manifold M^n, we begin with the special case $M = \mathbb{R}^n$:

Definition 5.4.2. Let ω denote an n-form with compact support on $\mathbb{R}^n$, $U \subset \mathbb{R}^n$. The *integral of ω over U* is defined to be the ordinary integral over U of the function $\omega(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n)$.

In other words, if we write $\omega = f \, \mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^n$, then $\int_U \omega = \int_U f$.

Next, suppose M^n is an oriented manifold, $(U, \boldsymbol{x})$ a positive chart of M, and ω an n-form on M with support in U. Define

$$\int_M \omega = \int_U \omega := \int_{\boldsymbol{x}(U)} (\boldsymbol{x}^{-1})^* \omega, \tag{5.4.1}$$

with the last integral as in Definition 5.4.2. It must be checked that this definition makes sense; i.e., that it does not depend on the particular chart. This is essentially due to the change of variables theorem:

Theorem 5.4.1. *If $(U, \boldsymbol{x})$ and $(V, \boldsymbol{y})$ are positive charts on M^n, and ω is an n-form with compact support in $U \cap V$, then*

$$\int_{\boldsymbol{x}(U \cap V)} (\boldsymbol{x}^{-1})^* \omega = \int_{\boldsymbol{y}(U \cap V)} (\boldsymbol{y}^{-1})^* \omega.$$

Proof. First of all, observe that

$$\begin{aligned}\int_{\boldsymbol{x}(U\cap V)} (\boldsymbol{x}^{-1})^*\omega &= \int_{\boldsymbol{x}(U\cap V)} \omega\circ\boldsymbol{x}^{-1}\left(\boldsymbol{x}_*^{-1}\boldsymbol{D}_1,\dots,\boldsymbol{x}_*^{-1}\boldsymbol{D}_n\right)\\ &= \int_{\boldsymbol{x}(U\cap V)} \omega\circ\boldsymbol{x}^{-1}\left(\frac{\partial}{\partial x^1}\circ\boldsymbol{x}^{-1},\dots,\frac{\partial}{\partial x^n}\circ\boldsymbol{x}^{-1}\right)\\ &= \int_{\boldsymbol{x}(U\cap V)} \left[\omega\left(\frac{\partial}{\partial x^1},\dots,\frac{\partial}{\partial x^n}\right)\right]\circ\boldsymbol{x}^{-1},\end{aligned}$$

and similarly for the term involving $\boldsymbol{y}$, so that the identity to be established becomes

$$\int_{\boldsymbol{x}(U\cap V)} \omega\left(\frac{\partial}{\partial x^1},\dots,\frac{\partial}{\partial x^n}\right)\circ\boldsymbol{x}^{-1} = \int_{\boldsymbol{y}(U\cap V)} \omega\left(\frac{\partial}{\partial y^1},\dots,\frac{\partial}{\partial y^n}\right)\circ\boldsymbol{y}^{-1}.$$

Now, by Exercise 3.16,

$$\frac{\partial}{\partial y^i} = \sum_j \left(D_i(u^j\circ\boldsymbol{x}\circ\boldsymbol{y}^{-1})\circ\boldsymbol{y}\right)\frac{\partial}{\partial x^j}.$$

Together with Theorem 5.2.3, this yields

$$\omega\left(\frac{\partial}{\partial y^1},\dots,\frac{\partial}{\partial y^n}\right) = (\det D(\boldsymbol{x}\circ\boldsymbol{y}^{-1})\circ\boldsymbol{y})\,\omega\left(\frac{\partial}{\partial x^1},\dots,\frac{\partial}{\partial x^n}\right).$$

But then by the change of variables theorem,

$$\begin{aligned}&\int_{\boldsymbol{y}(U\cap V)} \omega\left(\frac{\partial}{\partial y^1},\dots,\frac{\partial}{\partial y^n}\right)\circ\boldsymbol{y}^{-1}\\ &\quad= \int_{\boldsymbol{y}(U\cap V)} \det D(\boldsymbol{x}\circ\boldsymbol{y}^{-1})\,\omega\left(\frac{\partial}{\partial x^1},\dots,\frac{\partial}{\partial x^n}\right)\circ\boldsymbol{y}^{-1}\\ &\quad= \int_{\boldsymbol{y}(U\cap V)} \left|\det D(\boldsymbol{x}\circ\boldsymbol{y}^{-1})\right|\left[\omega\left(\frac{\partial}{\partial x^1},\dots,\frac{\partial}{\partial x^n}\right)\circ\boldsymbol{x}^{-1}\right]\circ\boldsymbol{x}\circ\boldsymbol{y}^{-1}\\ &\quad= \int_{\boldsymbol{x}(U\cap V)} \omega\left(\frac{\partial}{\partial x^1},\dots,\frac{\partial}{\partial x^n}\right)\circ\boldsymbol{x}^{-1}.\end{aligned}$$

This establishes the claim. □

Notice how the above proof used the fact that both charts were positive.

The general case (when the form does not have support inside the domain of a single chart) uses a partition of unity:

Definition 5.4.3. Let ω denote an n-form with compact support on M^n, and $\{(U_i,\boldsymbol{x}_i)\}$ a countable atlas with subordinate partition of unity $\{\varphi_i\}$. For each i, define an n-form ω_i by $\omega_i = \varphi_i\omega$ on U_i and $\omega_i = 0$ outside U_i. The *integral of ω over M* is

$$\int_M \omega = \sum_i \int_M \omega_i.$$

The proof that this definition does not depend on the chosen partition of unity is similar to that in Chapter 4, but more straightforward since the sum is a finite one.

Examples and Remarks 5.4.1. (i) Let M^2 be a surface – i.e., a two-dimensional manifold – in $\mathbb{R}^3$. Any chart $(U, \boldsymbol{x})$ of M induces a unit vector field $\boldsymbol{N}$ on U orthogonal to M, namely

$$\boldsymbol{N} = \frac{1}{|\frac{\partial}{\partial x^1} \times \frac{\partial}{\partial x^2}|} \frac{\partial}{\partial x^1} \times \frac{\partial}{\partial x^2}.$$

If $(V, \boldsymbol{y})$ is another chart, then by Exercise 3.16,

$$\frac{\partial}{\partial x^i} = \sum_j (D_i(u^j \circ \boldsymbol{y} \circ \boldsymbol{x}^{-1}) \circ \boldsymbol{x}) \frac{\partial}{\partial x^j},$$

so that

$$\frac{\partial}{\partial x^1} \times \frac{\partial}{\partial x^2} = \det(D(\boldsymbol{y} \circ \boldsymbol{x}^{-1}) \circ \boldsymbol{x}) \frac{\partial}{\partial y^1} \times \frac{\partial}{\partial y^2}.$$

It follows that positive charts determine the same unit normal field on the intersection of their domains. Thus, an orientable surface admits a global unit normal field. Conversely, if M admits a unit normal field $\boldsymbol{N}$, then it is orientable: the 2-form ω on M, given by $\omega(\boldsymbol{X}, \boldsymbol{Y}) = \det(\boldsymbol{N}, \boldsymbol{X}, \boldsymbol{Y})$ is nowhere zero. We will later generalize this to hypersurfaces.

(ii) The *volume form* of an oriented n-manifold M is the n-form η on M such that $\eta(\boldsymbol{p})(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) = 1$ for any positively oriented orthonormal basis $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ of $M_{\boldsymbol{p}}$, $\boldsymbol{p} \in M$. It is well-defined by Theorem 5.2.3. In fact, if α is any n-form inducing the orientation, then $f\alpha$, where $f(\boldsymbol{p}) = 1/\alpha(\boldsymbol{p})(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ for any positive orthonormal basis $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ of M_p, is the volume form of M. When M is compact, its *volume* is $\int_M \eta$.

(iii) By Theorem 1.4.4, there is a bijection $\flat : \mathfrak{X}M \to \Lambda_1(M)$ given by $\boldsymbol{X}^\flat(\boldsymbol{Y}) = \langle \boldsymbol{X}, \boldsymbol{Y} \rangle$ for vector fields $\boldsymbol{X}$, $\boldsymbol{Y}$ on M, see also Exercise 3.9. It is customary in physics to consider so-called 'line integrals of vector fields in $\mathbb{R}^n$ along oriented curves'. This can be interpreted in the present context as follows: let M be a one-dimensional oriented manifold in $\mathbb{R}^n$, $\boldsymbol{X}$ a vector field on some open set U containing M, so that $\boldsymbol{X}^\flat$ is a one-form on U. If $\imath : M \hookrightarrow \mathbb{R}^n$ denotes inclusion, then $\imath^* \boldsymbol{X}^\flat$ is a one-form on M, and the line integral of $\boldsymbol{X}$ along the curve is defined to be $\int_M \imath^* \boldsymbol{X}^\flat$. This, in turn, is easily extendible to the case when the image of the curve $\boldsymbol{c} : I \to \mathbb{R}^n$ is not a manifold: define the integral of $\boldsymbol{X}$ along $\boldsymbol{c}$ to be $\int_I \boldsymbol{c}^* \boldsymbol{X}^\flat$.
In physics, a common kind of vector field is a force field $\boldsymbol{F}$, which, when acting on a body of mass m, produces an acceleration $\dot{\boldsymbol{c}}'$ satisfying $\boldsymbol{F}(\boldsymbol{c}(t)) = m\, \dot{\boldsymbol{c}}'(t)$ according to Newton's second law of motion. The *work* W done by $\boldsymbol{F}$ in moving the object from $\boldsymbol{c}(a)$ to $\boldsymbol{c}(b)$ is defined to be $\int_{[a,b]} \boldsymbol{c}^* \boldsymbol{F}^\flat$. The *kinetic energy* of the object at time t is $K(\boldsymbol{c}(t)) = (1/2)m|\dot{\boldsymbol{c}}|^2$. It follows that the work done by the force equals the

change in kinetic energy:

$$W = \int_{[a,b]} \boldsymbol{c}^* \boldsymbol{F}^\flat = \int_a^b \boldsymbol{F}^\flat(\dot{\boldsymbol{c}}) = \int_a^b \langle m\dot{\boldsymbol{c}}', \dot{\boldsymbol{c}} \rangle = \frac{m}{2} \int_a^b |\dot{\boldsymbol{c}}|^{2\prime}$$
$$= K(\boldsymbol{c}(b)) - K(\boldsymbol{c}(a)).$$

(iv) Let ω denote the one-form on $\mathbb{R}^2 \setminus \{0\}$ given by

$$\omega = \frac{-u^2}{((u^1)^2 + (u^2)^2)^{1/2}} \, \mathrm{d}u^1 + \frac{u^1}{((u^1)^2 + (u^2)^2)^{1/2}} \, \mathrm{d}u^2.$$

If $\imath_r : S^1(r) \hookrightarrow \mathbb{R}^2$ is the inclusion map from the circle of radius r and center the origin into the plane, then $\imath_r^* \omega$ is the volume form of $S^1(r)$ for the induced (counterclockwise orientation): indeed the metric dual

$$\omega^\sharp = \frac{-u^2}{((u^1)^2 + (u^2)^2)^{1/2}} \boldsymbol{D}_1 + \frac{u^1}{((u^1)^2 + (u^2)^2)^{1/2}} \boldsymbol{D}_2$$

of ω is tangent to all circles (being orthogonal to the position vector field), has length 1, and $\omega(\omega^\sharp)$ is easily computed to equal one.

The curve $\boldsymbol{c} : (0, 2\pi) \to \mathbb{R}^2$, $\boldsymbol{c}(t) = (r \cos t, r \sin t)$, is a parametrization of $S^1(r) \setminus \{r\boldsymbol{e}_1\}$, and

$$\dot{\boldsymbol{c}} = -r \sin \boldsymbol{D}_1 \circ \boldsymbol{c} + r \cos \boldsymbol{D}_2 \circ \boldsymbol{c} = r\omega^\sharp \circ \boldsymbol{c}.$$

Thus, the volume of $S^1(r)$ equals

$$\int_{S^1(r)} \imath_r^* \omega = \int_0^{2\pi} \omega(\dot{\boldsymbol{c}}) = \int_0^{2\pi} r\omega(\omega^\sharp) = \int_0^{2\pi} r = 2\pi r,$$

as expected.

5.5 Manifolds with boundary

The *upper half-space* H^n is the set of all points $\boldsymbol{p} \in \mathbb{R}^n$ such that $u^n(\boldsymbol{p}) \geq 0$. Even though it is closed in $\mathbb{R}^n$, we extended the notion of differentiability to maps defined on sets such as these; namely, a map is differentiable on H^n if it is extendable to a map that is smooth on some open set containing H^n.

In order to discuss Stokes' theorem, we need the following concept, which generalizes that of manifold:

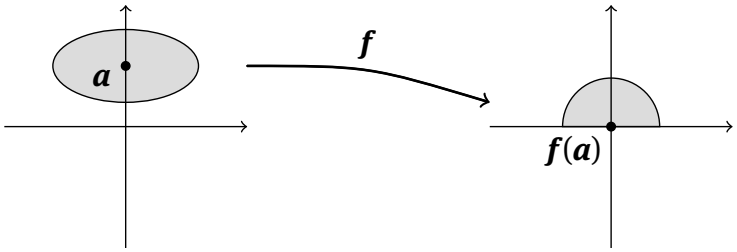
Definition 5.5.1. A subset $M \subset \mathbb{R}^{n+k}$ of Euclidean space is said to be an *n-dimensional manifold with boundary* if every point $\boldsymbol{p} \in M$ admits a neighborhood U in $\mathbb{R}^{n+p}$ with either one of the following properties:

(1) There exists an open set V in $\mathbb{R}^n$, a one-to-one differentiable map $\boldsymbol{h} : V \to \mathbb{R}^{n+k}$ of maximal rank everywhere such that
   - $\boldsymbol{h}(V) = U \cap M$, and
   - $\boldsymbol{h}^{-1}$ is continuous;

(2) There exists an open set V in H^n, a one-to-one differentiable map $\boldsymbol{h} : V \to \mathbb{R}^{n+k}$ of maximal rank everywhere such that
   - $\boldsymbol{h}(V) = U \cap M$,
   - $\boldsymbol{h}^{-1}$ is continuous, and
   - $(u^n \circ \boldsymbol{h}^{-1})(\boldsymbol{p}) = 0$.

$\boldsymbol{h}$ is called a *local parametrization* of M, and its inverse a *chart*.

When all points of M satisfy the first condition, one recovers the usual definition of a manifold. It is important to realize that if a point $\boldsymbol{p}$ satisfies the second condition for some parametrization $\boldsymbol{h}$, then it satisfies it for every parametrization: for if there were some other parametrization $\tilde{\boldsymbol{h}}$ of the first type, with, say, $\tilde{\boldsymbol{h}}(\boldsymbol{a}) = \boldsymbol{p}$ and $u^n(\boldsymbol{a}) \neq 0$, then $\boldsymbol{f} := \boldsymbol{h}^{-1} \circ \tilde{\boldsymbol{h}}$ would be a map of maximal rank on a neighborhood of $\boldsymbol{a}$ in $\mathbb{R}^n$. The image of arbitrarily small open neighborhoods of $\boldsymbol{a}$ would not be open in $\mathbb{R}^n$, contradicting the inverse function theorem.



In view of this, we may define the *boundary* ∂M of M to be the set of all points satisfying the second condition. A word of caution is in order here: even though the notation is the same, the boundary of a manifold M need not coincide with the topological boundary of the set M as defined in Chapter 1, see Exercise 5.24.

One of the simplest examples of manifold with boundary is a closed metric ball in $\mathbb{R}^n$. The boundary sphere is then also a manifold, with dimension lower by 1. This is always the case:

Proposition 5.5.1. *Let M be an n-dimensional manifold with boundary. If ∂M is non-empty, then it is an $(n-1)$-dimensional manifold.*

Proof. Consider $\boldsymbol{p} \in \partial M$. Any parametrization $\boldsymbol{h} : U \to \boldsymbol{h}(U)$ of a neighborhood $\boldsymbol{h}(U)$ of $\boldsymbol{p}$ in M induces a parametrization $\tilde{\boldsymbol{h}}$ of a neighborhood of $\boldsymbol{p}$ in ∂M: Let $\tilde{U} = \{\boldsymbol{u} \in \mathbb{R}^n \mid u^n(\boldsymbol{u}) = 0\}$. $\tilde{U}$ is naturally identified with a subset of $\mathbb{R}^{n-1} = \mathbb{R}^{n-1} \times \{0\} \subset \mathbb{R}^n$, and the restriction $\tilde{\boldsymbol{h}}$ of $\boldsymbol{h}$ to $\tilde{U}$ is differentiable, has maximal rank, and has continuous inverse because $\boldsymbol{h}$ enjoys these properties. □

In light of the above proposition, if $\boldsymbol{p} \in \partial M$, then the tangent space of ∂M at $\boldsymbol{p}$ is well defined. It is also convenient to have a notion of tangent space of M itself at $\boldsymbol{p}$. In order to do so, consider a parametrization $\boldsymbol{h}$ with $\boldsymbol{h}(\boldsymbol{a}) = \boldsymbol{p}$. Let $\tilde{\boldsymbol{h}}$ be an extension of $\boldsymbol{h}$ to an open set in $\mathbb{R}^n$, and define $M_{\boldsymbol{p}} = \tilde{\boldsymbol{h}}_*(\mathbb{R}^n_{\boldsymbol{a}})$. It must of course be checked that this is independent of the extension. If $\bar{\boldsymbol{h}}$ is another extension of $\boldsymbol{h}$, choose a neighborhood U of $\boldsymbol{a}$ in $\mathbb{R}^n$ small enough so that both extensions are defined on U. Then $\bar{\boldsymbol{h}}^{-1} \circ \tilde{\boldsymbol{h}}$ is the identity on $U \cap H^n$, and therefore the same is true for the derivative $D(\bar{\boldsymbol{h}}^{-1} \circ \tilde{\boldsymbol{h}})$ on the intersection of U with the interior of H^n. Choosing any sequence $\boldsymbol{a}_k \in U \cap (H^n)^0$ converging to $\boldsymbol{a}$, we have that $D(\bar{\boldsymbol{h}}^{-1} \circ \tilde{\boldsymbol{h}})(\boldsymbol{a})$ is the identity by continuity, and thus, $D(\bar{\boldsymbol{h}})(\boldsymbol{a}) = D(\tilde{\boldsymbol{h}})(\boldsymbol{a})$ as claimed. We will for convenience's sake denote either by $D\boldsymbol{h}(\boldsymbol{a})$.

Consider any $\boldsymbol{p} \in \partial M$ and parametrization $\boldsymbol{h}$ of a neighborhood of $\boldsymbol{p}$. Since $M_{\boldsymbol{p}}$ has dimension n, and $(\partial M)_{\boldsymbol{p}}$ is a subspace of $M_{\boldsymbol{p}}$ of dimension $n-1$, there are exactly two unit vectors in $M_{\boldsymbol{p}}$ orthogonal to $(\partial M)_{\boldsymbol{p}}$. If $\boldsymbol{x} = \boldsymbol{h}^{-1}$, then they cannot lie in the kernel of $\mathrm{d}x^n(\boldsymbol{p})$ because this kernel is precisely $(\partial M)_{\boldsymbol{p}}$. Thus, there is one, and only one unit vector $\boldsymbol{n} \in M_{\boldsymbol{p}}$ orthogonal to $(\partial M)_{\boldsymbol{p}}$ with $\boldsymbol{n}(x^n) < 0$. $\boldsymbol{n}$ is called the *outward unit normal at* $\boldsymbol{p}$. This procedure can be done at any point of ∂M, and the resulting map $\boldsymbol{N} : \partial M \to TM$ which assigns to each element in the boundary the outward unit normal vector is called the *outward unit normal field*. This field can be used to induce an orientation of the boundary of M. In order to do so, we introduce the following terminology:

Definition 5.5.2. Let $\boldsymbol{X}$ be a vector field on a manifold M. *Interior multiplication by* $\boldsymbol{X}$ is the map $i(\boldsymbol{X}) : \Lambda_k(M) \to \Lambda_{k-1}(M)$ given by

$$(i(\boldsymbol{X})\alpha)(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}) = \alpha(\boldsymbol{X}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}), \qquad \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1} \in \mathfrak{X}M.$$

Suppose now that M is an oriented manifold with boundary, with orientation induced by some nowhere-zero n-form ω. If $\boldsymbol{N}$ is the outward unit normal field, then $i(\boldsymbol{N})\omega$ is a nowhere-zero form on ∂M, and thus induces an orientation on ∂M. This orientation is said to be the one *induced by* M. Notice that if η is the volume form of M, then $i(\boldsymbol{N})\eta$ is the volume form of ∂M.

As a simple example, consider the closed disk D of radius 1 about the origin in $\mathbb{R}^2$. It is a 2-dimensional manifold with boundary. The open ball has a natural orientation as an open set in $\mathbb{R}^2$, namely that induced by the determinant. The induced orientation on the boundary S^1 is the one induced by the (restriction of the) one-form ω from Examples and Remarks 5.4.1 (iii), and is usually referred to as the counterclockwise orientation.

Another important special case is that of a one-dimensional manifold with boundary. For simplicity, let us consider the case when M^1 is parametrized by a curve $\boldsymbol{c} : [0, a] \to M$ with $|\dot{\boldsymbol{c}}| \equiv 1$. ∂M is a zero-dimensional manifold consisting of $\boldsymbol{c}(0)$ and $\boldsymbol{c}(1)$. Since a zero-form is a function, two nowhere zero forms induce the same orientation on ∂M if they both have the same sign when evaluated at each point. Thus, an orientation is an assignment of a sign to each point in the boundary. If $\dot{\boldsymbol{c}}$ represents

the orientation of M, then the induced orientation of ∂M is $(\boldsymbol{c}(1), +)$, $(\boldsymbol{c}(0), -)$, because the outward unit normal is $\dot{\boldsymbol{c}}(1)$ at $\boldsymbol{c}(1)$ and $-\dot{\boldsymbol{c}}(0)$ at $\boldsymbol{c}(0)$.

If α is zero-form on an oriented compact zero-dimensional manifold $M = \{\boldsymbol{a}_1, \dots, \boldsymbol{a}_k\}$, define

$$\int_M \alpha := \sum_{i=1}^{k} \operatorname{sign}(\boldsymbol{a}_i)\alpha(\boldsymbol{a}_i),$$

where sign denotes the sign induced by the orientation. The reader is invited to verify that with this notation, if $M = [a, b] \subset \mathbb{R}$ with the standard orientation, and $f : M \to \mathbb{R}$ is a zero-form, then the Fundamental Theorem of Calculus reads

$$\int_M \mathrm{d}f = \int_{\partial M} \imath^* f,$$

where $\imath : \partial M \hookrightarrow M$ denotes inclusion (recall that for a zero-form f, $\imath^* f = f \circ \imath$ is the restriction of f to ∂M).

5.6 Stokes' theorem

In the previous section, we rewrote the Fundamental Theorem of Calculus as

$$\int_M \mathrm{d}f = \int_{\partial M} \imath^* f.$$

The generalization to higher dimensions is known as Stokes' Theorem. Perhaps unsurprisingly, the key ingredients in its proof consist of the above theorem and Fubini's.

Theorem 5.6.1 (Stokes' Theorem). *Let M be an oriented n-dimensional manifold with boundary, ω an $(n-1)$-form with compact support in M. Then*

$$\int_M \mathrm{d}\omega = \int_{\partial M} \imath^* \omega,$$

with $\imath : \partial M \hookrightarrow M$ denoting inclusion.

Proof. By definition of the integral and the fact that ω can be written as a finite sum of forms each of which has support inside some chart, it may be assumed that there is a local positive parametrization $\boldsymbol{h} : H^n \supset U \to M$ whose image contains the support of ω.

By Theorem 5.3.3, $\int_M \mathrm{d}\omega = \int_U \boldsymbol{h}^* \mathrm{d}\omega = \int_U \mathrm{d}\boldsymbol{h}^* \omega$. Now,

$$\boldsymbol{h}^* \omega = \sum_{i=1}^{n} f^i \, \mathrm{d}u^1 \wedge \dots \mathrm{d}u^{n-1} \wedge \mathrm{d}u^{n+1} \wedge \cdots \wedge \mathrm{d}u^n,$$

where $f^i = \boldsymbol{h}^* \omega(\boldsymbol{D}_1, \dots, \boldsymbol{D}_{i-1}, \boldsymbol{D}_{i+1}, \dots, \boldsymbol{D}_n)$. By Theorem 5.3.3,

$$\int_M \mathrm{d}\omega = \int_U \mathrm{d}\boldsymbol{h}^* \omega = \int_U \sum_i (-1)^i D_i f^i,$$

since $\mathrm{d}\boldsymbol{h}^*\omega = (\sum_i (-1)^i D_i f^i)\,\mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^n$. Extend f^i to smooth functions on H^n by setting them equal to zero outside U, and denote them both by the same symbol. We consider two possibilities:

Suppose first that $\boldsymbol{h}(U)$ does not intersect the boundary of M. Then $\imath^*\omega$, and its integral over the boundary of M vanish. Consider a box $B = [a_1, b_1] \times \cdots \times [a_n, b_n]$ that contains U in its interior. Since the support of ω is compact, it may be assumed that $a_n > 0$; i.e., the box does not intersect the boundary of H^n. If B_i denotes the lower-dimensional box $[a_1, b_1] \times \cdots \times [a_{i-1}, b_{i-1}] \times [a_{i+1}, b_{i+1}] \times \ldots [a_n, b_n]$, then Fubini's theorem and the fundamental theorem of Calculus imply that $\int_B D_i f^i = \int_{B_i} g^i$, where

$$\begin{aligned} g^i(u^1, \ldots, u^{n-1}) = f^i(u^1, \ldots, u^{i-1}, b_i, u^i, \ldots, u^{n-1}) \\ - f^i(u^1, \ldots, u^{i-1}, a_i, u^i, \ldots, u^{n-1}). \end{aligned}$$

But then g^i is identically zero, since f^i vanishes on the boundary of the box. This proves the theorem in this case.

Suppose next that $\boldsymbol{h}(U)$ does intersect the boundary of M. As before, extend $\boldsymbol{h}^*\omega$ to all of H^n by setting it equal to zero outside U, and enclose it in a box $B = [a_1, b_1] \times \cdots \times [0, b_n]$ large enough so that U is contained in $(a_1, b_1) \times \cdots \times [0, b_n)$. We will, as above, denote by B_i the lower-dimensional box obtained by deleting the i-th edge of B. If $\bar{\boldsymbol{h}}$ denotes the restriction of $\boldsymbol{h}$ to the boundary ∂H^n of H^n, then

$$\bar{\boldsymbol{h}}^*\imath^*\omega = f_0^n\,\mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^{n-1},$$

where

$$f_0^n(u^1, \ldots, u^{n-1}) = f^n(u^1, \ldots, u^{n-1}, 0),$$

see also Examples and Remarks 4.1.1 (iv) regarding notation. Now, the volume form on $\partial H^n = \mathbb{R}^{n-1} \times \{0\}$ induced by H^n is given by

$$\begin{aligned} \mu &= i(-\boldsymbol{D}_n)\,\mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^n \\ &= \sum_{j=1}^{n} \mu(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_{j-1}, \boldsymbol{D}_{j+1}, \ldots, \boldsymbol{D}_n)\,\mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^{j-1} \wedge \mathrm{d}u^{j+1} \wedge \cdots \wedge \mathrm{d}u^n. \end{aligned}$$

The only non-vanishing term in this sum occurs when $j = n$, so that

$$\begin{aligned} \mu &= \mu(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_{n-1})\,\mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^{n-1} \\ &= \mathrm{d}u^1 \wedge \ldots\, \mathrm{d}u^n(-\boldsymbol{D}_n, \boldsymbol{D}_1, \ldots, \boldsymbol{D}_{n-1})\,\mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^{n-1} \\ &= (-1)^n\,\mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^{n-1}. \end{aligned}$$

Thus,

$$\int_{\partial M} \imath^*\omega = \int_{B_n} (-1)^n f^n(u^1, \ldots, u^{n-1}, 0)\,\mathrm{d}u^1 \ldots \mathrm{d}u^{n-1}. \tag{5.6.1}$$

On the other hand, we have, as before,

$$\int_M \mathrm{d}\omega = \sum_{i=1}^{n} (-1)^{i+1} \int_B D_i f^i = \sum_{i=1}^{n} (-1)^{i+1} \int_{B_i} g^i.$$

Since $g^i = 0$ for $i < n$,

$$\begin{aligned}\int_M \mathrm{d}\omega &= (-1)^{n+1} \int_{B_n} (f^n(u^1, \dots, u^{n-1}, b_n) \\ &\quad - f^n(u^1, \dots, u^{n-1}, 0))\, \mathrm{d}u^1 \dots \mathrm{d}u^{n-1} \\ &= \int_{B_n} (-1)^n f^n(u^1, \dots, u^{n-1}, 0)\, \mathrm{d}u^1 \dots \mathrm{d}u^{n-1}.\end{aligned}$$

Comparing with (5.6.1) now completes the proof. □

Examples and Remarks 5.6.1. (i) If M^n is a manifold without boundary, then $\int_M \mathrm{d}\omega = 0$ for any n-form with compact support in M.

(ii) The compact support hypothesis in Stokes' theorem cannot be removed: consider for example the interior of a unit disk $M^2 = \{(a, b) \in \mathbb{R}^2 \mid a^2 + b^2 < 1\}$, and the 2-form $\alpha = \mathrm{d}u^1 \wedge \mathrm{d}u^2$ on M. We have $\alpha = \mathrm{d}(u^1\, \mathrm{d}u^2)$, so if Stokes' theorem were to hold, then by (i), the integral of α over M would vanish. By the definition of integral, however, $\int_M \alpha$ equals the area π of M^2.

(iii) Consider the one-form

$$\omega = \frac{-u^2}{(u^1)^2 + (u^2)^2}\, \mathrm{d}u^1 + \frac{u^1}{(u^1)^2 + (u^2)^2}\, \mathrm{d}u^2$$

on $\mathbb{R}^2 \setminus \{(0, 0)\}$. It is related to the polar angle function θ which is defined on $\mathbb{R}^2 \setminus \{(a, 0) \mid a \geq 0\}$. In fact, in each of the four open quadrants in the plane, θ can be written as $\arctan(u^2/u^1) + c$, where c is some constant depending on the quadrant, cf. Section 4.6.1. A direct computation shows that $\mathrm{d}(\arctan(u^2/u^1))) = \omega$. There is, however, no zero-form f defined on *all* of $\mathbb{R}^2$ such that $\mathrm{d}f = \omega$: otherwise, with $\imath : S^1 \hookrightarrow \mathbb{R}^2$ denoting the inclusion of the circle S^1 with radius 1 around the origin, we would have that

$$\int_{S^1} \imath^* \omega = \int_{S^1} \imath^* (\mathrm{d}f) = \int_{S^1} \mathrm{d}(\imath^* f) = 0$$

by (i). On the unit circle, however, this one-form coincides with that from Examples 5.4.1 (iii), so that $\int_{S^1} \imath^* \omega = 2\pi$. A differential form ω is said to be *closed* if $\mathrm{d}\omega = 0$, and *exact* if $\omega = \mathrm{d}\alpha$ for some form α of degree one less. Since $\mathrm{d} \circ \mathrm{d} = 0$, any exact form is closed, but this example shows the converse is false in general. It also illustrates yet again how Stokes' theorem can fail without the compactness assumption: if M denotes the manifold with boundary that consists of the closed unit disk about the origin with the origin deleted, then $\int_{\partial M} \imath^* \omega = 2\pi$, but $\int_M \mathrm{d}\omega = 0$, since ω is closed.

(iv) The form $\imath^* \omega$ on ∂M is by definition just the restriction of ω to the boundary. For this reason, Stokes' theorem is often stated simply as $\int_M \mathrm{d}\omega = \int_{\partial M} \omega$.

5.7 Classical versions of Stokes' theorem

We present some classical theorems found in multivariable calculus texts that are all special cases of Stokes' theorem. Since most books do not use differential forms, each statement needs some degree of reinterpretation.

Theorem 5.7.1 (Green's Theorem). *Suppose $M \subset \mathbb{R}^2$ is a compact 2-dimensional manifold with boundary, with the usual orientation. If P and Q are differentiable, then*

$$\int_{\partial M} P\,\mathrm{d}x + Q\,\mathrm{d}y = \iint_M (D_1 Q - D_2 P)\,\mathrm{d}A.$$

Proof. As observed in Examples and Remarks 5.4.1 (ii), the left side is interpreted in physics as the line integral along ∂M of the vector field $\boldsymbol{F} = P\boldsymbol{D}_1 + Q\boldsymbol{D}_2$. In the present context, the identity means

$$\int_{\partial M} \imath^*(P\,\mathrm{d}u^1 + Q\,\mathrm{d}u^2) = \int_M (D_1 Q - D_2 P)\,\mathrm{d}u^1 \wedge \mathrm{d}u^2.$$

This is precisely Stokes' theorem for the 1-form $\omega = \boldsymbol{F}^\flat = P\,\mathrm{d}u^1 + Q\,\mathrm{d}u^2$. □

Before stating the second theorem, two more concepts must be defined: first of all, if M is an oriented 2-dimensional manifold in $\mathbb{R}^3$, the *positive unit normal field* $\boldsymbol{N}$ of M is the unit normal field (out of two) such that the restriction of $i(\boldsymbol{N})\,\mathrm{d}u^1 \wedge \mathrm{d}u^2 \wedge \mathrm{d}u^3$ to M equals the volume form η of M. Extend the determinant and cross product to each tangent space via the canonical isomorphism $\mathcal{I}_{\boldsymbol{u}} : \mathbb{R}^3 \to \mathbb{R}^3_{\boldsymbol{u}}$. If $\boldsymbol{N} = \sum N_i \boldsymbol{D}_i$, then for a positive local orthonormal basis $\boldsymbol{X}_1, \boldsymbol{X}_2$ of vector fields on M,

$$\begin{aligned} 1 = \eta(\boldsymbol{X}_1, \boldsymbol{X}_2) &= \mathrm{d}u^1 \wedge \mathrm{d}u^2 \wedge \mathrm{d}u^3(\boldsymbol{N}, \boldsymbol{X}_1, \boldsymbol{X}_2) = \det\begin{bmatrix} \boldsymbol{N} & \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix} \\ &= \langle \boldsymbol{N}, \boldsymbol{X}_1 \times \boldsymbol{X}_2 \rangle \\ &= (N^1\,\mathrm{d}u^2 \wedge \mathrm{d}u^3 + N^2\,\mathrm{d}u^3 \wedge \mathrm{d}u^1 + N^3\,\mathrm{d}u^1 \wedge \mathrm{d}u^2)(\boldsymbol{X}_1, \boldsymbol{X}_2), \end{aligned}$$

where the last equality was obtained by expanding the determinant along the first row. Thus, the volume form of an oriented 2-dimensional submanifold of $\mathbb{R}^3$ with positive unit normal $\boldsymbol{N} = \sum_i N^i \boldsymbol{D}_i$ is

$$\eta = N^1\,\mathrm{d}u^2 \wedge \mathrm{d}u^3 + N^2\,\mathrm{d}u^3 \wedge \mathrm{d}u^1 + N^3\,\mathrm{d}u^1 \wedge \mathrm{d}u^2.$$

The second concept is that of the *curl* of a vector field. The curl of $\boldsymbol{F} = \sum_i F^i \boldsymbol{D}_i$ is the vector field

$$\operatorname{curl} \boldsymbol{F} = (D_2 F^3 - D_3 F^2)\boldsymbol{D}_1 + (D_3 F^1 - D_1 F^3)\boldsymbol{D}_2 + (D_1 F^2 - D_2 F^1)\boldsymbol{D}_3,$$

cf. also the exercises for a more motivating definition.

With this in mind, the classical Stokes' theorem may now be stated as follows:

Theorem 5.7.2 (Stokes' theorem). *Let $M^2 \subset \mathbb{R}^3$ denote a compact oriented 2-dimensional submanifold of $\mathbb{R}^3$ with boundary and positive unit normal field $\boldsymbol{N}$. If $\boldsymbol{F}$ is a differentiable vector field on an open set in $\mathbb{R}^3$ containing M, then*

$$\int_M \langle \operatorname{curl} \boldsymbol{F}, \boldsymbol{N} \rangle \, \mathrm{d}S = \int_{\partial M} \langle \boldsymbol{F}, \, \mathrm{d}\boldsymbol{c} \rangle.$$

Proof. The right side is standard notation for the line integral of $\boldsymbol{F}$ along the boundary. Thus, if $\boldsymbol{F} = \sum_i F^i \boldsymbol{D}_i$, and $\omega = \boldsymbol{F}^\flat = \sum_i F^i \, \mathrm{d}u^i$ is the dual 1-form, then the identity that needs to be established is

$$\int_M \langle \operatorname{curl} \boldsymbol{F}, \boldsymbol{N} \rangle \, \eta = \int_{\partial M} \imath^* \omega,$$

where η is the volume form of M. In view of our version of Stokes', it must then be shown that $\langle \operatorname{curl} \boldsymbol{F}, \boldsymbol{N} \rangle \eta = \mathrm{d}\omega$. Notice that for orthonormal vector fields $\boldsymbol{X}_i$ on M,

$$\det \begin{bmatrix} \langle \boldsymbol{F}, \boldsymbol{N} \rangle \boldsymbol{N} & \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix} = \det \begin{bmatrix} \boldsymbol{F} & \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix} \tag{5.7.1}$$

since $\langle \boldsymbol{F}, \boldsymbol{N} \rangle \boldsymbol{N}$ is the component of $\boldsymbol{F}$ orthogonal to $TM = \operatorname{span}\{\boldsymbol{X}_1, \boldsymbol{X}_2\}$. Thus,

$$\begin{aligned} \langle \operatorname{curl} \boldsymbol{F}, \boldsymbol{N} \rangle \, \eta(\boldsymbol{X}_1, \boldsymbol{X}_2) &= \det \begin{bmatrix} \langle \operatorname{curl} \boldsymbol{F}, \boldsymbol{N} \rangle \boldsymbol{N} & \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix} \\ &= \det \begin{bmatrix} \operatorname{curl} \boldsymbol{F} & \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix}. \end{aligned}$$

Writing out this last expression yields

$$\begin{aligned} ((D_2 F^3 - D_3 F^2) \, \mathrm{d}u^2 \wedge \mathrm{d}u^3 + (D_3 F^1 - D_1 F^3) \, \mathrm{d}u^3 \wedge \mathrm{d}u^1 \\ + (D_1 F^2 - D_2 F^1) \, \mathrm{d}u^1 \wedge \mathrm{d}u^2)(\boldsymbol{X}_1, \boldsymbol{X}_2), \end{aligned}$$

which is precisely $\mathrm{d}\omega(\boldsymbol{X}_1, \boldsymbol{X}_2)$. □

The next result involves yet another concept commonly used in physics: the *divergence* of a vector field $\boldsymbol{X} = \sum_i X^i \boldsymbol{D}_i$ in $\mathbb{R}^3$ is the function $\operatorname{div} \boldsymbol{X} = D_1 X^1 + D_2 X^2 + D_3 X^3$. It can actually be defined for vector fields on a manifold of arbitrary dimension, and is explored further in the exercises.

Theorem 5.7.3 (Divergence theorem). *Let M be a compact 3-dimensional manifold with boundary in $\mathbb{R}^3$, with positive unit normal field $\boldsymbol{N}$. If $\boldsymbol{F}$ is a differentiable vector field on M, then*

$$\int_M \operatorname{div} \boldsymbol{F} \, \mathrm{d}V = \int_{\partial M} \langle \boldsymbol{F}, \boldsymbol{N} \rangle \, \mathrm{d}S.$$

Proof. The identity to be established is

$$\int_M \operatorname{div} \boldsymbol{F} \, \mathrm{d}u^1 \wedge \mathrm{d}u^2 \wedge \mathrm{d}u^3 = \int_{\partial M} \langle \boldsymbol{F}, \boldsymbol{N} \rangle \eta,$$

where η denotes the volume form of the boundary. If $\boldsymbol{F} = \sum_i F^i \boldsymbol{D}_i$, let

$$\omega = F^1\, \mathrm{d}u^2 \wedge \mathrm{d}u^3 + F^2\, \mathrm{d}u^3 \wedge \mathrm{d}u^1 + F^3\, \mathrm{d}u^1 \wedge \mathrm{d}u^2,$$

so that $\mathrm{d}\omega = \operatorname{div} \boldsymbol{F}\, \mathrm{d}u^1 \wedge \mathrm{d}u^2 \wedge \mathrm{d}u^3$. Now, by (5.7.1), given vector fields $\boldsymbol{X}_i$ on ∂M,

$$\begin{aligned}\langle \boldsymbol{F}, \boldsymbol{N}\rangle\, \eta(\boldsymbol{X}_1, \boldsymbol{X}_2) &= \det\begin{bmatrix}\langle \boldsymbol{F}, \boldsymbol{N}\rangle \boldsymbol{N} & \boldsymbol{X}_1 & \boldsymbol{X}_2\end{bmatrix} = \det\begin{bmatrix}\boldsymbol{F} & \boldsymbol{X}_1 & \boldsymbol{X}_2\end{bmatrix} \\ &= (F^1\, \mathrm{d}u^1 \wedge \mathrm{d}u^2 + F^2\, \mathrm{d}u^3 \wedge \mathrm{d}u^1 \\ &\quad + F^3\, \mathrm{d}u^1 \wedge \mathrm{d}u^2)(\boldsymbol{X}_1, \boldsymbol{X}_2) \\ &= \imath^* \omega(\boldsymbol{X}_1, \boldsymbol{X}_2),\end{aligned}$$

so that $\langle \boldsymbol{F}, \boldsymbol{N}\rangle \eta = \imath^* \omega$, and the identity is once again that from Stokes' theorem. □

The above theorems provide physical interpretations for the notions of curl and divergence. Consider for example a fluid with constant density ρ in a region of 3-space. If $\boldsymbol{v}$ is the velocity field of the fluid, $\boldsymbol{F} = \rho \boldsymbol{v}$, and $\boldsymbol{N}$ a unit field, then $\langle \boldsymbol{F}, \boldsymbol{N}\rangle$ represents the rate of flow (in the physical sense, not to be confused with the flow of a vector field) per unit area in direction $\boldsymbol{N}$. If $\boldsymbol{a}$ is a point in the fluid, $B_r(\boldsymbol{a})$ the set of points at distance less than r from $\boldsymbol{a}$, let

$$m_r = \inf\{\operatorname{div} \boldsymbol{F}(\boldsymbol{b}) \mid \boldsymbol{b} \in \overline{B_r}(\boldsymbol{a})\}, \qquad M_r = \sup\{\operatorname{div} \boldsymbol{F}(\boldsymbol{b}) \mid \boldsymbol{b} \in \overline{B_r}(\boldsymbol{a})\}.$$

Denote by $S_r(\boldsymbol{a})$ the sphere of radius r centered at $\boldsymbol{a}$. By the divergence theorem,

$$m_r \leq \frac{1}{\operatorname{vol}(B_r(\boldsymbol{a}))} \int_{B_r(\boldsymbol{a})} \operatorname{div} \boldsymbol{F} = \frac{1}{\operatorname{vol}(B_r(\boldsymbol{a}))} \int_{S_r(\boldsymbol{a})} \langle \boldsymbol{F}, \boldsymbol{N}\rangle \eta \leq M_r,$$

so that

$$\operatorname{div} \boldsymbol{F}(\boldsymbol{a}) = \lim_{r \to 0} \frac{1}{\operatorname{vol}(B_r(\boldsymbol{a}))} \int_{S_r(\boldsymbol{a})} \langle \boldsymbol{F}, \boldsymbol{N}\rangle \eta.$$

In this case, the divergence may be interpreted as the net rate of outward flow per unit volume at $\boldsymbol{a}$. $\boldsymbol{a}$ is called a *source* if $\operatorname{div} \boldsymbol{F}(\boldsymbol{a}) > 0$, and a *sink* if it is negative. A fluid is said to be *incompressible* if it has everywhere vanishing divergence. More generally, the reader is asked to show in the exercises that a vector field on $\mathbb{R}^n$ has everywhere vanishing divergence if and only if the (mathematical) flow $\boldsymbol{\Phi}_t$ of the vector field preserves volumes; i.e., $\boldsymbol{\Phi}_t^* \omega = \omega$ for all t, where ω is the volume form of $\mathbb{R}^n$.

The curl has a similar interpretation, but in terms of rotation of a fluid. This is also discussed in the exercises.

Examples 5.7.1. (i) One of Maxwell's equations in electromagnetism is

$$\operatorname{curl} \boldsymbol{H} = \boldsymbol{J},$$

where $\boldsymbol{H}$ denotes the magnetizing field and $\boldsymbol{J}$ the current density. It can be used to prove Ampère's law, which states that if C is a closed curve, then the current I

crossing any surface M bounded by C is

$$I = \int_C \langle \boldsymbol{H},\, \mathrm{d}\boldsymbol{c}\rangle.$$

To see this, let M be a 2-dimensional manifold with boundary $\partial M = C$. As in the case of a fluid, $\langle \boldsymbol{J}, \boldsymbol{N}\rangle$ represents the current per unit area that crosses a surface orthogonal to $\boldsymbol{N}$. Then, by Stokes' theorem, the current crossing M is

$$I = \int_M \langle \boldsymbol{J}, \boldsymbol{N}\rangle \,\mathrm{d}S = \int_M \langle \operatorname{curl} \boldsymbol{H}, \boldsymbol{N}\rangle \,\mathrm{d}S = \int_C \langle \boldsymbol{H},\, \mathrm{d}\boldsymbol{c}\rangle.$$

(ii) Yet another of Maxwell's equations is that if $\boldsymbol{E}$ is the electric field created by a charge distribution with charge density ρ, then $\operatorname{div} \boldsymbol{E} = \rho/\varepsilon_0$, where ε_0 is a universal constant. One can use it to prove Gauss's law, which says that if M is a compact 3-dimensional manifold with boundary ∂M, then the total electric charge inside ∂M (i.e., in M) is

$$Q = \varepsilon_0 \int_{\partial M} \langle \boldsymbol{E}, \boldsymbol{N}\rangle \,\mathrm{d}S.$$

Indeed, by the divergence theorem and the definition of Q,

$$Q = \int_M \rho \,\mathrm{d}V = \varepsilon_0 \int_M \operatorname{div} \boldsymbol{E} = \varepsilon_0 \int_{\partial M} \langle \boldsymbol{E}, \boldsymbol{N}\rangle \,\mathrm{d}S.$$

5.7.1 An application: the polar planimeter

The fundamental theorem of Calculus and its generalization, Stokes' theorem, illustrate how the behavior of a function on an open set is controlled by its behavior on the boundary of that set. This has some important practical applications: One such is an instrument, called a *planimeter*, that calculates the area of the region bounded by a simple closed curve in the plane. It does so by merely tracing the boundary curve. There are several types of planimeters. The first one was invented in 1814 in Bavaria, and the one we discuss here, the polar planimeter, was created by Jakob Amsler in 1856. Although the underlying principle relies on Green's theorem, it seems that this was only established almost a century later, cf. [10, 3].

Our planimeter consists of a pole that is fixed at the origin in the plane. A pole arm which can rotate about the pole is attached to it, and at the end of the pole arm is a pivot connected to the tracer arm. The tracer arm terminates in a wheel that is perpendicular to the arm, so that the wheel can only roll about the pivot. It can, however, also slide if moved parallel to the arm. The wheel then traces out the boundary curve by alternatively rolling and sliding, but only the rolling part is recorded.

In order to simplify the computations, our planimeter has both arms of equal length R. Thus, the curve that is traced has to lie within a disk of radius $2R$ about the pole. We will see that if l is the total distance the wheel has rolled after tracing the

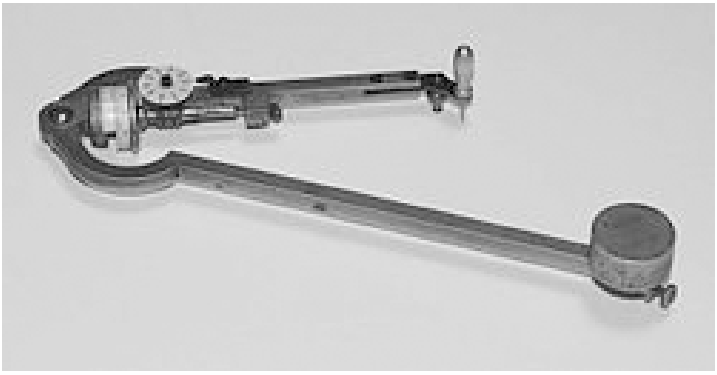
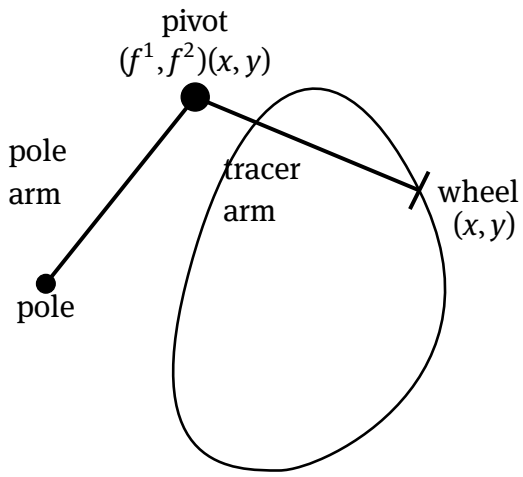


Fig. 5.1: A polar planimeter [1]

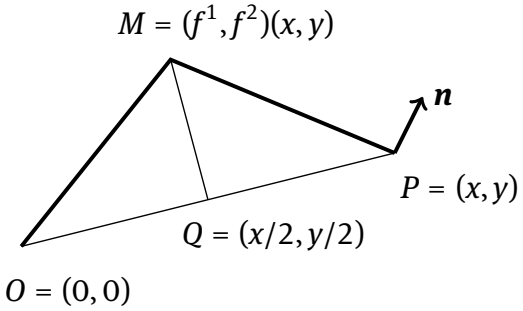
boundary curve, then the area A of the region enclosed by the curve is

$$A = lR.$$



Simple geometry dictates that if the wheel is at a given point (x, y), there are exactly two possible locations for the pivot; these are mutual reflections in the line connecting the pole to the wheel. One way to see this is that the pivot must lie at one of the two intersections of circles with radius R, the first one centered at the origin and the other at (x, y).

In order to find the coordinates (f^1, f^2) of the pivot M, we use the figure below, and denote by $\overrightarrow{QM}$ the vector version of the point $M - Q$. Thus, $|\overrightarrow{QM}|^2 = R^2 - (x^2 + y^2)/4$, and $\overrightarrow{QM}$ is orthogonal to $\overrightarrow{OP} = \begin{bmatrix} x & y \end{bmatrix}$.



There are two possible directions for this orthogonal vector, and we choose a counterclockwise rotation by $\pi/2$. This transforms $\begin{bmatrix} x & y \end{bmatrix}$ into $\begin{bmatrix} -y & x \end{bmatrix}$, so that

$$\overrightarrow{QM} = \left(\frac{R^2 - (x^2 + y^2)/4}{(x^2 + y^2)} \right)^{1/2} \begin{bmatrix} -y & x \end{bmatrix},$$

and since $\overrightarrow{OM} = \overrightarrow{OQ} + \overrightarrow{QM}$, the coordinates of the pivot M are

$$f^1(x,y) = \frac{1}{2}\left[x - y \left(\frac{4R^2 - x^2 - y^2}{x^2 + y^2} \right)^{1/2} \right],$$
$$f^2(x,y) = \frac{1}{2}\left[y + x \left(\frac{4R^2 - x^2 - y^2}{x^2 + y^2} \right)^{1/2} \right].$$

It immediately follows that

$$\overrightarrow{MP} = \frac{1}{2}\left[x + y \left(\tfrac{4R^2 - x^2 - y^2}{x^2 + y^2} \right)^{1/2} \quad y - x \left(\tfrac{4R^2 - x^2 - y^2}{x^2 + y^2} \right)^{1/2} \right],$$

and performing once again a counterclockwise rotation by $\pi/2$, we see that the unit normal vector field orthogonal to the tracer arm $\overrightarrow{MP}$ is $\boldsymbol{n} = P\boldsymbol{D}_1 + \boldsymbol{D}_2$, where

$$P(x,y) = \frac{1}{2R}\left[-y + x \left(\frac{4R^2 - x^2 - y^2}{x^2 + y^2} \right)^{1/2} \right],$$
$$Q(x,y) = \frac{1}{2R}\left[x + y \left(\frac{4R^2 - x^2 - y^2}{x^2 + y^2} \right)^{1/2} \right].$$

Now, the total distance that the wheel has rolled is the line integral of $\boldsymbol{n}$ along the boundary curve of the region M, and according to Green's theorem, this integral equals

$$\int_M (D_1 Q - D_2 P)\, \mathrm{d}u^1 \wedge \mathrm{d}u^2.$$

But

$$D_1 Q(x,y) = \frac{1}{2R} + y D_1 g(x,y), \qquad D_2 P(x,y) = -\frac{1}{2R} + x D_2 g(x,y), \tag{5.7.2}$$

where

$$g(x,y) = \left(\frac{4R^2 - x^2 - y^2}{x^2 + y^2} \right)^{1/2}.$$

A straightforward calculation yields that

$$D_1 g(x,y) = -\frac{4R^2 x}{(x^2 + y^2)^{3/2}(4R^2 - x^2 - y^2)^{1/2}}.$$

Since the formula for g is symmetric in x and y, $D_2 g$ is obtained by interchanging x and y in the formula for $D_1 g$. Thus $yD_1 g(x,y) = xD_2 g(x,y)$, and (5.7.2) implies that the integrand in Green's theorem is $D_1 Q - D_2 P = 1/R$. This means that the distance the wheel has rolled equals the area of M divided by R, as claimed.

5.8 Closed forms and exact forms

Recall that a differential form α on M is said to be closed if $\mathrm{d}\alpha = 0$, and exact if there is some form β such that $\mathrm{d}\beta = \alpha$. Every exact form is closed, but not every closed form is exact, cf. Examples and Remarks 5.6.1 (iii). We will show, however, that a closed form is always locally exact.

Remark 5.8.1. A one-form α on a Riemannian manifold is exact if and only if its dual vector field is the gradient of some function: the equation $\alpha = \mathrm{d}f$ is equivalent to $\alpha^\sharp = \nabla f$. Alternatively, for a vector field $\boldsymbol{X}$, $\boldsymbol{X} = \nabla f$ if and only if $\boldsymbol{X}^\flat = \mathrm{d}f$.

This concept is important enough in physics to warrant its own terminology: a force (i.e., a vector field) $\boldsymbol{F}$ is said to be *conservative* if there is some function f such that $\boldsymbol{F} = \nabla f$. In this case, the function $P = -f$ is called a *potential energy* function for f. Such a function is only defined up to an additive constant (if P is a potential energy, then so is $P + c$ for any constant c), but this is usually irrelevant, since one is mostly interested in the change in potential energy. In fact, the reason such forces are called conservative is that they satisfy the *conservation of energy law*: the sum of potential and kinetic energy is constant.

To see why, recall from Examples and Remarks 5.4.1 (ii) that if an object of mass m moves along a curve $\boldsymbol{c} : [a, b] \to M$ under the action of a (not necessarily conservative) force $\boldsymbol{F}$, then its kinetic energy at $\boldsymbol{c}(t)$ is $K(\boldsymbol{c}(t)) = (m|\dot{\boldsymbol{c}}(t)|^2)/2$ and the work done by $\boldsymbol{F}$ in moving the object from $\boldsymbol{c}(a)$ to $\boldsymbol{c}(b)$ is

$$W = \int_{\boldsymbol{c}} \boldsymbol{F}^b = K(\boldsymbol{c}(b)) - K(\boldsymbol{c}(a)).$$

If in addition $\boldsymbol{F}$ is conservative, then the work also equals

$$W = \int_{\boldsymbol{c}} \boldsymbol{F}^b = \int_{\boldsymbol{c}} -\mathrm{d}P = P(\boldsymbol{c}(a)) - P(\boldsymbol{c}(b)),$$

so that $(P + K)(\boldsymbol{c}(a)) = (P + K)(\boldsymbol{c}(b))$, as claimed.

A typical example of a conservative force is the gravitational force: an object of mass M that is placed at the origin in 3-space exerts a gravitational force on nearby objects. The force acting on an object of mass m located at $\boldsymbol{a} \in \mathbb{R}^3 \setminus \{\boldsymbol{0}\}$ is

$$\boldsymbol{F}(\boldsymbol{a}) = -\frac{GMm}{|\boldsymbol{a}|^3}\boldsymbol{P}(\boldsymbol{a}),$$

where G is a universal constant and $\boldsymbol{P}$ denotes the position vector field $\boldsymbol{P} = \sum_i u^i \boldsymbol{D}_i$. Thus,

$$\boldsymbol{F} = \nabla\left(\frac{GMm}{|\boldsymbol{P}|}\right),$$

as a straightforward calculation shows.

Denote by $Z^k(M)$ the space of all closed k-forms on M, and by $B^k(M)$ that of exact k-forms. $B^k(M)$ is a subspace of $Z^k(M)$. In general, if W is a subspace of a vector space V, the *quotient space* V/W is the collection of all elements of the form $\boldsymbol{v} + W$, $\boldsymbol{v} \in V$. Two elements $\boldsymbol{v}_1 + W$ and $\boldsymbol{v}_2 + W$ are said to be equal if $\boldsymbol{v}_1 - \boldsymbol{v}_2 \in W$. V/W inherits a vector space structure from the operations in V, if we define

$$(\boldsymbol{v}_1 + W) + (\boldsymbol{v}_2 + W) = (\boldsymbol{v}_1 + \boldsymbol{v}_2) + W, \qquad c(\boldsymbol{v} + W) = (c\boldsymbol{v}) + W,$$

for $\boldsymbol{v}_i \in V$, $c \in \mathbb{R}$. It must be checked that these operations are well defined: for example, if $\boldsymbol{v}_1 + W = \boldsymbol{v}_2 + W$, then for any $\boldsymbol{v}_3$, the identity

$$(\boldsymbol{v}_1 + W) + (\boldsymbol{v}_3 + W) = (\boldsymbol{v}_2 + W) + (\boldsymbol{v}_3 + W)$$

must hold. The left side is $\boldsymbol{v}_1 + \boldsymbol{v}_3 + W$, whereas the right side equals $\boldsymbol{v}_2 + \boldsymbol{v}_3 + W$. Since $\boldsymbol{v}_1 + \boldsymbol{v}_3 - (\boldsymbol{v}_2 + \boldsymbol{v}_3) = \boldsymbol{v}_1 - \boldsymbol{v}_2$ lies in W, both sides indeed agree. A similar argument works for scalar multiplication. In fact, with these operations, the projection $\boldsymbol{V \to V/W}$ which maps $\boldsymbol{v}$ to $\boldsymbol{v} + W$ is a linear transformation that is by construction surjective and has W as kernel, so that if V is finite-dimensional, then V/W has dimension $\dim V - \dim W$.

Definition 5.8.1. Given a positive integer k, the *k-th de Rham cohomology space* of M is $H^k(M) = Z^k(M)/B^k(M)$. $H^0(M)$ is defined as $Z^0(M)$.

The reader is invited to verify that if M is connected, then $H^0(M) \cong \mathbb{R}$, see Exercise 5.22. Even though determining cohomology spaces can, in general, be challenging, there is one case that can already be settled:

Proposition 5.8.1. *If M is a compact, oriented n-dimensional manifold without boundary, $n > 0$, then $H^n(M) \neq 0$.*

Proof. Let ω be the volume form of M. ω is closed because there are no $(n+1)$-forms. By definition, if $(U, \boldsymbol{x})$ is any positive chart of M, then $\int_{\boldsymbol{x}(U)} (\boldsymbol{x}^{-1})^* \omega > 0$. Thus, $\int_M \omega > 0$, so that ω cannot be exact (if $\omega = \mathrm{d}\alpha$, then $\int_M \omega = \int_{\partial M} \alpha = 0$ because the boundary is empty). □

Our next goal is to show that the pullback of $\boldsymbol{f} : M \to N$ induces a linear transformation $\boldsymbol{f}^* : H^k(N) \to H^k(M)$. In general, if W_i is a subspace of V_i, $i = 1, 2$, then any linear transformation $L : V_1 \to V_2$ that maps W_1 to W_2 induces a linear map $L : V_1/W_1 \to V_2/W_2$: define $L(\boldsymbol{v} + W_1) = (L\boldsymbol{v}) + W_2$. One only needs to check that this definition makes sense, because linearity follows from linearity of the original map. But if $\boldsymbol{u} + W_1 = \boldsymbol{v} + W_1$, then $\boldsymbol{u} - \boldsymbol{v} \in W_1$, so that $L\boldsymbol{u} - L\boldsymbol{v} = L(\boldsymbol{u} - \boldsymbol{v}) \in W_2$; i.e., $L\boldsymbol{u} + W_2 = L\boldsymbol{v} + W_2$, and the map is well defined.

Proposition 5.8.2. *A differentiable map $\boldsymbol{f} : M \to N$ induces linear transformations $\boldsymbol{f}^* : H^k(N) \to H^k(M)$ for all k.*

Proof. By Theorem 5.3.3, the pullback $\boldsymbol{f}^* : \Lambda_k(N) \to \Lambda_k(M)$ maps closed forms to closed forms, and exact forms to exact forms. The first property means that $\boldsymbol{f}^*$ restricts to a map $\boldsymbol{f}^* : Z^k(N) \to Z^k(M)$, and the second one implies the claim by the remark before the Proposition. □

By definition of the pullback, if $\boldsymbol{f} : M_1 \to M_2$ and $\boldsymbol{g} : M_2 \to M_3$, then $\boldsymbol{g} \circ \boldsymbol{f} : M_1 \to M_3$ has as pullback

$$(\boldsymbol{g} \circ \boldsymbol{f})^* = \boldsymbol{f}^* \circ \boldsymbol{g}^* : H^k(M_3) \to H^k(M_1).$$

One consequence is that diffeomorphic manifolds have isomorphic cohomology spaces: the pullback 1_M^* of the identity map $1_M : M \to M$ is the identity on the cohomology of M, so that if $\boldsymbol{f} : M \to N$ is a diffeomorphism, then $\boldsymbol{f}^* \circ \boldsymbol{f}^{-1*} = 1_{H^k(N)}$ and $\boldsymbol{f}^{-1*} \circ \boldsymbol{f}^* = 1_{H^k(M)}$. Thus, $\boldsymbol{f}^*$ is an isomorphism with inverse $\boldsymbol{f}^{-1*}$.

An equivalent formulation is that if two manifolds have different cohomology, then they are not diffeomorphic. It turns out that a much stronger property holds. In order to state it, we will need the following:

Definition 5.8.2. Two maps $\boldsymbol{f}, \boldsymbol{g} : M \to N$ are said to be *homotopic* if there exists a map $\boldsymbol{H} : M \times [0,1] \to N$ such that $\boldsymbol{H} \circ \imath_0 = \boldsymbol{f}$ and $\boldsymbol{H} \circ \imath_1 = \boldsymbol{g}$, where $\imath_t : M \to M \times [0,1]$ maps $\boldsymbol{p} \in M$ to $(\boldsymbol{p}, t)$, $0 \le t \le 1$. In this case, $\boldsymbol{H}$ is called a *homotopy* between $\boldsymbol{f}$ and $\boldsymbol{g}$.

Roughly speaking, a homotopy smoothly deforms $\boldsymbol{f}$ into $\boldsymbol{g}$ by a "curve of maps" $t \mapsto \boldsymbol{H} \circ \imath_t$ joining $\boldsymbol{f}$ to $\boldsymbol{g}$. Two spaces M and N are said to be *homotopy equivalent* if there exist maps $\boldsymbol{f} : M \to N$ and $\boldsymbol{g} : N \to M$ such that $\boldsymbol{g} \circ \boldsymbol{f}$ and $\boldsymbol{f} \circ \boldsymbol{g}$ are homotopic to the identity maps on the two spaces. Spaces that are diffeomorphic are of course homotopy equivalent, but the latter concept is much weaker: for example, $\mathbb{R}^n$ is homotopy equivalent to the one point subset consisting of the origin. In fact, if $\boldsymbol{f} : \mathbb{R}^n \to \{\boldsymbol{0}\}$ maps everything to the origin, and $\boldsymbol{g} : \{\boldsymbol{0}\} \to \mathbb{R}^n$ is inclusion, then $\boldsymbol{f} \circ \boldsymbol{g}$ is the identity on $\{\boldsymbol{0}\}$, whereas $\boldsymbol{g} \circ \boldsymbol{f}$ is homotopic to the identity on $\mathbb{R}^n$ via $\boldsymbol{H} : \mathbb{R}^n \times [0,1] \to \mathbb{R}^n$, where $\boldsymbol{H}(\boldsymbol{p}, t) = (1-t)\boldsymbol{p}$. In general, a manifold that is homotopy equivalent to a one point subset is said to be *contractible*; alternatively, M is contractible if the identity map on M is homotopic to a constant map (i.e., to one that maps all of M to a single point).

Let π_M and t denote the projections of the product $M \times [0,1]$ onto the factors, so that $(\pi_{M*}, t_*) : (M \times [0,1])_{(\boldsymbol{p},t_0)} \to M_{\boldsymbol{p}} \times [0,1]_{t_0}$ is an isomorphism, cf. Exercise 3.14. Denote by $\tilde{\boldsymbol{D}}$ the vector field on $M \times [0,1]$ given by $\tilde{\boldsymbol{D}}(\boldsymbol{p}, t_0) = (\pi_{M*}, t_*)^{-1}_{(\boldsymbol{p},t_0)}(\boldsymbol{0}, \boldsymbol{D}(t_0))$. Alternatively, $\tilde{\boldsymbol{D}}(\boldsymbol{p}, t_0) = \jmath_{\boldsymbol{p}*}\boldsymbol{D}(t_0)$, where $\jmath_{\boldsymbol{p}} : [0,1] \to M \times [0,1]$ maps t_0 to $(\boldsymbol{p}, t_0)$.

Lemma 5.8.1. *Any $\omega \in \Lambda_k(M \times [0,1])$ can be uniquely written as $\omega = \omega_1 + \mathrm{d}t \wedge \chi$, where ω_1 and χ are k and $(k-1)$-forms on $M \times [0,1]$ respectively such that $i(\tilde{\boldsymbol{D}})\omega_1 = i(\tilde{\boldsymbol{D}})\chi = 0$.*

Proof. Because of the isomorphism (π_{M*}, t_*), it suffices to check that if a vector space V decomposes as $W \times \mathbb{R}$, then any $\omega \in \Lambda_k(V)$ can be uniquely written as $\omega_1 + \alpha \wedge \eta$, where $i(t)\omega_1, i(t)\eta = 0$ for $t \in \mathbb{R}$ and $\alpha \in \mathbb{R}^* \setminus \{0\}$. But this is clear, since if $\alpha_1, \dots, \alpha_n$ is a basis of W^*, then $\alpha_1, \dots, \alpha_n, \alpha$ is a basis of V^*. □

Writing $\omega = \omega_1 + \mathrm{d}t \wedge \chi \in \Lambda_k(M \times [0,1])$ as in Lemma 5.8.1, define a linear operator $I : A_k(M \times [0,1]) \to A_{k-1}(M)$ by

$$I\omega(\boldsymbol{p})(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}) = \int_0^1 \chi(\boldsymbol{p}, t)(\imath_{t*}\boldsymbol{v}_1, \ldots, \imath_{t*}\boldsymbol{v}_{k-1})\,\mathrm{d}t. \tag{5.8.1}$$

Proposition 5.8.3. *If ω is a k-form on $M \times [0,1]$, then $\imath_1^*\omega - \imath_0^*\omega = \mathrm{d}(I\omega) + I(\mathrm{d}\omega)$.*

Proof. Let $(U, \boldsymbol{x})$ be a chart on M, so that $(U \times [0,1])(\bar{\boldsymbol{x}}, t))$ is a chart on $M \times [0,1]$, where $\bar{\boldsymbol{x}} = \boldsymbol{x} \circ \pi_M$. By Lemma 5.8.1, the restriction of ω can be written as a sum of terms of two types:
(1) $f\,\mathrm{d}\bar{x}^{i_1} \wedge \cdots \wedge \mathrm{d}\bar{x}^{i_k}$ for some function f on U, and
(2) $f\,\mathrm{d}t \wedge \mathrm{d}\bar{x}^{i_1} \wedge \cdots \wedge \mathrm{d}\bar{x}^{i_{k-1}}$.

Since I is linear, it suffices to consider the case when ω is one of the above types.

If ω is of type 1, then

$$\mathrm{d}\omega = (\text{terms not involving } \mathrm{d}t) + (\partial f/\partial t)\,\mathrm{d}t \wedge \mathrm{d}\bar{x}^{i_1} \wedge \cdots \wedge \mathrm{d}\bar{x}^{i_k}.$$

Now, for any $t \in [0,1]$, $\pi_M \circ \imath_t = 1_M$, so that $\imath_t^*\,\mathrm{d}\bar{x}^i = \mathrm{d}(\bar{x}^i \circ \imath_t) = \mathrm{d}x^i$, and

$$\begin{aligned}
&I(\mathrm{d}\omega)(\boldsymbol{p})\left(\frac{\partial}{\partial x^{j_1}}, \ldots, \frac{\partial}{\partial x^{j_k}}\right) \\
&\quad = \int_0^1 \frac{\partial f}{\partial t}(\boldsymbol{p}, t)(\mathrm{d}\bar{x}^{i_1} \wedge \cdots \wedge \mathrm{d}\bar{x}^{i_k})(\boldsymbol{p}, t)\left(\imath_{t*}\frac{\partial}{\partial x^{j_1}}, \ldots, \imath_{t*}\frac{\partial}{\partial x^{j_k}}\right)\mathrm{d}t \\
&\quad = \int_0^1 \frac{\partial f}{\partial t}(\boldsymbol{p}, t)(\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k})(\boldsymbol{p})\left(\frac{\partial}{\partial x^{j_1}}, \ldots, \frac{\partial}{\partial x^{j_k}}\right)\mathrm{d}t \\
&\quad = \left(\int_0^1 \frac{\partial f}{\partial t}(\boldsymbol{p}, t)\,\mathrm{d}t\right)(\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k})(\boldsymbol{p})\left(\frac{\partial}{\partial x^{j_1}}, \ldots, \frac{\partial}{\partial x^{j_k}}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
I(\mathrm{d}\omega)(\boldsymbol{p}) &= (f(\boldsymbol{p}, 1) - f(\boldsymbol{p}, 0))(\mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_k})(\boldsymbol{p}) \\
&= \imath_1^*\omega(\boldsymbol{p}) - \imath_0^*\omega(\boldsymbol{p}),
\end{aligned}$$

which proves the result in this case, since $I\omega = 0$.
Suppose now that ω is of type 2; then $\imath_1^*\omega = \imath_0^*\omega = 0$ because $\imath_{t_0}^*\,\mathrm{d}t = 0$ for any t_0. On

the other hand,

$$\begin{aligned} I(\,\mathrm{d}\omega)(\boldsymbol{p}) &= I\Big(\sum_{l=1}^{n}\frac{\partial f}{\partial \bar{x}^l}\,\mathrm{d}\bar{x}^l\wedge\,\mathrm{d}t\wedge\,\mathrm{d}\bar{x}^{i_1}\wedge\cdots\wedge\,\mathrm{d}\bar{x}^{i_k}\Big)(\boldsymbol{p})\\ &= I\Big(-\sum_{l=1}^{n}\frac{\partial f}{\partial \bar{x}^l}\,\mathrm{d}t\wedge\,\mathrm{d}\bar{x}^l\wedge\,\mathrm{d}\bar{x}^{i_1}\wedge\cdots\wedge\,\mathrm{d}\bar{x}^{i_k}\Big)(\boldsymbol{p})\\ &= -\sum_l\Big(\int_0^1\frac{\partial f}{\partial \bar{x}^l}(\boldsymbol{p},t)\;\mathrm{d}t\Big)(\,\mathrm{d}x^l\wedge\,\mathrm{d}x^{i_1}\wedge\cdots\wedge\,\mathrm{d}x^{i_k})(\boldsymbol{p}), \end{aligned}$$

while

$$\begin{aligned} d(I\omega)(\boldsymbol{p}) &= \mathrm{d}\Big(\Big[\int_0^1 f(\boldsymbol{p},t)\,\mathrm{d}t\Big]\,\mathrm{d}x^{i_1}\wedge\cdots\wedge\,\mathrm{d}x^{i_k}\Big)(\boldsymbol{p})\\ &= \sum_l\frac{\partial}{\partial x^l}\Big(\int_0^1 f(\boldsymbol{p},t)\;\mathrm{d}t\Big)(\,\mathrm{d}x^l\wedge\,\mathrm{d}x^{i_1}\wedge\cdots\wedge\,\mathrm{d}x^{i_k})(\boldsymbol{p})\\ &= \sum_l\Big(\int_0^1\frac{\partial f}{\partial \bar{x}^l}(\boldsymbol{p},t)\;\mathrm{d}t\Big)(\,\mathrm{d}x^l\wedge\,\mathrm{d}x^{i_1}\wedge\cdots\wedge\,\mathrm{d}x^{i_k})(\boldsymbol{p}), \end{aligned}$$

so that $I(\,\mathrm{d}\omega)+\,\mathrm{d}(I\omega)=0$. □

Theorem 5.8.1. *If $\boldsymbol{f}_0,\boldsymbol{f}_1 : M\to N$ are homotopic, then the induced linear maps $\boldsymbol{f}_0^*,\boldsymbol{f}_1^* : H^k(N)\to H^k(M)$ are equal for every k.*

Proof. Let $\boldsymbol{H} : M\times[0,1]\to N$ be a homotopy, with $\boldsymbol{H}\circ\iota_j=\boldsymbol{f}_j$, $j=0,1$. Then for any closed k-form ω on N,

$$\boldsymbol{f}_1^*\omega-\boldsymbol{f}_0^*\omega=(\iota_1^*-\iota_0^*)\boldsymbol{H}^*\omega=(\,\mathrm{d}I+I\,\mathrm{d})\boldsymbol{H}^*\omega=\,\mathrm{d}I\boldsymbol{H}^*\omega+I\boldsymbol{H}^*\,\mathrm{d}\omega=\,\mathrm{d}I\boldsymbol{H}^*\omega$$

is exact, so that $\boldsymbol{f}_1^*\omega+B^k(M)=\boldsymbol{f}_0^*\omega+B^k(M)$. □

Corollary 5.8.1. *If M is contractible, then $H^k(M)=\{\boldsymbol{0}\}$ for all $k\geq 1$.*

Proof. If M is contractible, then there is a point $\boldsymbol{p}_0\in M$ such that the constant map $\boldsymbol{f} : M\to M$, $\boldsymbol{f}(\boldsymbol{p})=\boldsymbol{p}_0$, is homotopic to the identity map 1_M of M. By Theorem 5.8.1, $1_M^*,\boldsymbol{f}^* : H^k(M)\to H^k(M)$ are equal. But 1_M^* is the identity map on $H^k(M)$, whereas $\boldsymbol{f}^*\omega=0$ for any $\omega\in\Lambda_k(M)$ if $k\geq 1$ because $\boldsymbol{f}_*=\boldsymbol{0}$. □

In particular, a compact oriented manifold of positive dimension n cannot be contractible, since its n-th cohomology is nontrivial.

Corollary 5.8.2 (Poincaré Lemma). *Let α be a closed k-form on a manifold M, $k\geq 1$. Then any $\boldsymbol{p}\in M$ has a neighborhood on which the restriction of α is exact.*

Proof. Let $(V,\boldsymbol{h})$ be a local parametrization of M with $\boldsymbol{h}(\boldsymbol{0})=\boldsymbol{p}$, and $\varepsilon>0$ small enough that $B_\varepsilon(\boldsymbol{0})\subset V$. Then $U=\boldsymbol{h}(B_\varepsilon(\boldsymbol{0}))$ is a contractible neighborhood of $\boldsymbol{p}$. If $\iota : U\hookrightarrow M$

denotes inclusion, then the restriction $\imath^*\alpha$ of α to U is a closed form on a contractible manifold, and is therefore exact by the previous corollary. □

We emphasize again that the conclusion of the Poincaré Lemma is local: the 2-form ω from Examples and Remarks 5.6.1 (iii) is the differential of the polar angle function θ in a neighborhood of any point, even though it is not exact.

5.9 Exercises

5.1. Prove that a k-tensor ($k > 1$) T on V is alternating if and only if

$$T(\dots, \boldsymbol{v}, \dots, \boldsymbol{v}, \dots) = 0, \qquad \boldsymbol{v} \in V.$$

5.2. Given a k-tensor T on V, define T_s by

$$T_s(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k) = \frac{1}{k!} \sum_{\sigma \in S_k} T(\boldsymbol{v}_{\sigma(1)}, \dots, \boldsymbol{v}_{\sigma(k)}),$$

for $\boldsymbol{v}_i \in V$. Show that T_s is a symmetric k-tensor. Is it true that $T = T_s + T_a$?

5.3. Prove Proposition 5.1.1.

5.4. Prove that if α is alternating, then $\alpha_a = \alpha$, and conclude that for any k-tensor T, $(T_a)_a = T_a$.

5.5. Show that $\alpha \wedge \beta = (-1)^{kl}\beta \wedge \alpha$ for $\alpha \in \Lambda_k(V)$ and $\beta \in \Lambda_l(V)$. *Hint:* It is enough to show this for $\alpha = \alpha^1 \wedge \cdots \wedge \alpha^k$ and $\beta = \beta^1 \wedge \cdots \wedge \beta^l$, where $\alpha^i, \beta^j \in V^*$.

5.6. Prove that any form on V is decomposable if V has dimension ≤ 3. Show that this is false if $\dim V \geq 4$. *Hint:* Use Proposition 5.2.2.

5.7. Prove that $\alpha \in \Lambda_2(V)$ is decomposable if and only if $\alpha \wedge \alpha = 0$. *Hint:* Use Proposition 5.2.2.

5.8. Let V be an n-dimensional vector space. Show that any n-form on V is the volume form of some inner product and some orientation on V. Are the inner product and orientation unique?

5.9. Show that if W is a subspace of V, then the operation

$$c(\boldsymbol{v} + W) = (c\boldsymbol{v}) + W, \qquad c \in \mathbb{R}, \quad \boldsymbol{v} \in V,$$

on the set $V/W = \{\boldsymbol{v} + W \mid \boldsymbol{v} \in V\}$ is well defined; i.e., if $\boldsymbol{u}_i + W = \boldsymbol{v}_i + W$, then $(c\boldsymbol{u}_i) + W = (c\boldsymbol{v}_i) + W$. What is the zero vector in the space V/W?

5.10. Let V be an n-dimensional oriented inner product space. This exercise constructs an isomorphism $\star : \Lambda_k(V) \to \Lambda_{n-k}(V)$, $k = 1, \dots, n-1$, called the *Hodge star*

operator, that preserves decomposable elements. In particular, it shows that every element in $\Lambda_{n-1}(V) \cong \Lambda_1(V) = V^*$ is decomposable.

Recall that a linear transformation is entirely determined by its values on a basis. So let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ be a positively oriented orthonormal basis, $\alpha^1, \dots, \alpha^n$ its dual basis, and $\omega = \alpha^1 \wedge \cdots \wedge \alpha^n \in \Lambda_n(V)$ the volume form of V.

(a) Show that for any basis element

$$\alpha \in \mathcal{B} = \{\alpha^{i_1} \wedge \cdots \wedge \alpha^{i_k} \mid 1 \leq i_1 < \cdots < i_k \leq n\}$$

of the induced basis $\mathcal{B}$ of $\Lambda_k(V)$, there exists a unique $\beta \in \Lambda_{n-k}(V)$ such that $\alpha \wedge \beta = \omega$. Define $\star\alpha = \beta$ and extend linearly to all $\Lambda_k(V)$. *Hint:* β or $-\beta$ belongs to the induced basis of $\Lambda_{n-k}(V)$.

(b) Prove that $\star \circ \star = (-1)^{k(n-k)} 1_{\Lambda_k(V)}$. In particular, $\star$ is an isomorphism.

(c) Prove that if α is decomposable, then so is $\star\alpha$. *Hint:* a decomposable α corresponds to a k-dimensional subspace W of V. Consider its orthogonal complement.

(d) Show that $\star$ does not depend on the particular choice of basis.

5.11. Let V be a vector space with basis $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3\}$. According to Exercise 5.10, every element of $\Lambda_2(V)$ is decomposable. If α_i is the basis dual to $\boldsymbol{v}_i$, write $\alpha_1 \wedge \alpha_2 + \alpha_2 \wedge \alpha_3 + \alpha_1 \wedge \alpha_3$ as a decomposable element.

5.12. Let V be an oriented n-dimensional inner product space with volume form ω and musical isomorphisms $\sharp, \flat$. Given $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{n-1} \in V$, define their *cross product* to be the vector

$$\boldsymbol{v}_1 \times \cdots \times \boldsymbol{v}_{n-1} := \left(\star(\boldsymbol{v}_1^\flat \wedge \cdots \wedge \boldsymbol{v}_{n-1}^\flat)\right)^\sharp \in V.$$

Here, $\star$ is the Hodge star operator from Exercise 5.10.

(a) Show that $\langle \boldsymbol{v}_1 \times \cdots \times \boldsymbol{v}_{n-1}, \boldsymbol{u} \rangle = \omega(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{n-1}, \boldsymbol{u}), \boldsymbol{u} \in V$.

(b) Prove that when $V = \mathbb{R}^3$ with the standard inner product and orientation, this definition coincides with the usual cross product.

5.13. Recall that the musical isomorphisms in $\mathbb{R}^n$ extend naturally to each tangent space and its dual, and thus establish a bijective correspondence between vector fields and 1-forms. The same is true for vector fields and one-forms on a manifold. Similarly, if M^n is an oriented manifold, then the Hodge star operator from Exercise 5.10 on each tangent space induces a bijective correspondence $\star : \Lambda_k(M) \to \Lambda_{n-k}(M)$.

(a) The *gradient* of $f : M \to \mathbb{R}$ is the vector field ∇f on M given by $(df)^\sharp$. Show that if $M = \mathbb{R}^n$, this coincides with the usual definition of gradient.

(b) Given a vector field $\boldsymbol{X}$ on an oriented manifold M, the *divergence* of $\boldsymbol{X}$ is the function $\operatorname{div} \boldsymbol{X} = \star\, \mathrm{d} \star \boldsymbol{X}^\flat$. Prove that when $M = \mathbb{R}^n$ and $\boldsymbol{X} = \sum_i X^i \boldsymbol{D}_i$, then

$$\operatorname{div} \boldsymbol{X} = \sum_j D_j(X^j).$$

(c) Suppose $\boldsymbol{X}$ is a vector field on $\mathbb{R}^3$. The *curl* of $\boldsymbol{X}$ is the vector field $\operatorname{curl}\boldsymbol{X} = (\star\, \mathrm{d}\boldsymbol{X}^\flat)^\sharp$. Show that if $X^i = \langle \boldsymbol{X}, \boldsymbol{D}_i\rangle$, then

$$\operatorname{curl}\boldsymbol{X} = (D_2X^3 - D_3X^2)\boldsymbol{D}_1 + (D_3X^1 - D_1X^3)\boldsymbol{D}_2 + (D_1X^2 - D_2X^1)\boldsymbol{D}_3.$$

5.14. Let $\boldsymbol{X}$, $\boldsymbol{Y}$ denote vector fields on $\mathbb{R}^3$, $f : \mathbb{R}^3 \to \mathbb{R}$. Prove the following identities:
(a) $\operatorname{curl}(\boldsymbol{X} + \boldsymbol{Y}) = \operatorname{curl}(\boldsymbol{X}) + \operatorname{curl}(\boldsymbol{Y})$ (see previous exercise for the definition of curl).
(b) $\operatorname{curl}(f\boldsymbol{X}) = f \operatorname{curl}\boldsymbol{X} + \nabla f \times \boldsymbol{X}$.
(c) $\operatorname{div}(f\boldsymbol{X}) = f \operatorname{div}\boldsymbol{X} + \langle \nabla f, \boldsymbol{X}\rangle$.
(d) $\operatorname{div}(\boldsymbol{X} \times \boldsymbol{Y}) = \langle \operatorname{curl}\boldsymbol{X}, \boldsymbol{Y}\rangle - \langle \operatorname{curl}\boldsymbol{Y}, \boldsymbol{X}\rangle$.
(e) $\operatorname{div}\operatorname{curl}\boldsymbol{X} = 0$.
(f) $\operatorname{curl}\nabla f = \boldsymbol{0}$.

5.15. The next two problems examine in more detail the concept of divergence; they show, in particular, that a vector field $\boldsymbol{X}$ has vanishing divergence if and only if its flow $\boldsymbol{\Phi}_t$ preserves the volume form ω; i.e., $\boldsymbol{\Phi}_t^*\omega = \omega$ for all t. For simplicity, we limit ourselves to the case $M = \mathbb{R}^n$.
The *Lie derivative of a k-form α with respect to a vector field $\boldsymbol{X}$* is the k-form $L_{\boldsymbol{X}}\alpha$ defined by

$$L_{\boldsymbol{X}}\alpha(\boldsymbol{p}) = \lim_{t\to 0} \frac{1}{t}\left(\boldsymbol{\Phi}_t^*\omega(\boldsymbol{p}) - \omega(\boldsymbol{p})\right),$$

where $\boldsymbol{\Phi}_t$ denotes the flow of $\boldsymbol{X}$.
This problem shows that the divergence of $\boldsymbol{X}$ satisfies $L_{\boldsymbol{X}}\omega = (\operatorname{div}\boldsymbol{X})\omega$, where ω is the volume form of $\mathbb{R}^n$. The next problem deals with the statement about vanishing divergence.
(a) Show that $\boldsymbol{\Phi}_t^*\omega = (\det D\boldsymbol{\Phi}_t)\omega$. Thus, $L_{\boldsymbol{X}}\omega(\boldsymbol{p}) = f'(0)\omega(\boldsymbol{p})$, where

$$f(t) = \det M(t), \qquad M(t) = D\boldsymbol{\Phi}_t(\boldsymbol{p}).$$

(b) Use the chain rule to prove that $f'(t) = \sum_i \langle M_i'(t), \boldsymbol{e}_i\rangle$, where M_i denotes the i-th column of M.
(c) Let $\boldsymbol{\Phi} : \mathbb{R}^{n+1} \supset U \to \mathbb{R}^n$ be given by $\boldsymbol{\Phi}(t, \boldsymbol{q}) = \boldsymbol{\Phi}_t(\boldsymbol{q})$. Show that

$$M_i'(t) = D(D\boldsymbol{\Phi}(t,\boldsymbol{p})\boldsymbol{e}_i)\boldsymbol{e}_1 = D(D\boldsymbol{\Phi}(t,\boldsymbol{p})\boldsymbol{e}_1)\boldsymbol{e}_i,$$

so that

$$M_i'(t) = D(\pi_2(\boldsymbol{X}\circ\boldsymbol{\Phi}(t,\boldsymbol{p})))\boldsymbol{e}_i,$$

where $\pi_2 : T\mathbb{R}^n = \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ denotes projection. Conclude that $L_{\boldsymbol{X}}\omega = (\operatorname{div}\boldsymbol{X})\omega$.

5.16. This problem uses notation and results from the previous one.
(a) Show that if the volume form ω of $\mathbb{R}^n$ is invariant under the flow Φ_t of a vector field $\boldsymbol{X}$ (in the sense that $\Phi_t^*\omega = \omega$), then the divergence of $\boldsymbol{X}$ is identically zero.
(b) Conversely, suppose that a vector field with flow Φ_t has zero divergence. Given $\boldsymbol{p} \in \mathbb{R}^n$, consider as before $f(t) = \det(D\Phi_t)(\boldsymbol{p})$, so that $f'(0) = 0$. Prove that $f'(t) = 0$

for all t, and conclude that $\Phi_t^* \omega = \omega$. *Hint:* Use an argument similar to that in the last part of the proof of Theorem 2.9.4.

5.17. The spectral theorem describes a canonical form for self-adjoint operators on an inner product space V. This problem describes one for skew-adjoint operators; i.e., operators L satisfying $\langle L\boldsymbol{u}, \boldsymbol{v}\rangle = -\langle L\boldsymbol{v}, \boldsymbol{u}\rangle$ for any $\boldsymbol{u}, \boldsymbol{v} \in V$.

Let V be an inner product space, and $\mathfrak{o}(V)$ the space of skew-adjoint operators on V.

(a) Prove that the map $\mathfrak{o}(V) \to \Lambda_2(V)$ that assigns to $L \in \mathfrak{o}(V)$ the element $\alpha \in \Lambda_2(V)$ given by $\alpha(\boldsymbol{v}, \boldsymbol{w}) = \langle L\boldsymbol{v}, \boldsymbol{w}\rangle$, is an isomorphism.

(b) Use Proposition 5.2.2 to show that for any $L \in \mathfrak{o}(V)$, there exists an orthonormal basis in which the matrix of L has the form

$$\begin{bmatrix} 0 & -\lambda_1 & 0 & 0 & \cdots \\ \lambda_1 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & -\lambda_2 & \cdots \\ 0 & 0 & \lambda_2 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

with the last row and column consisting of zeros if the dimension of V is odd.

5.18. An (orthonormal) *moving frame* on $\mathbb{R}^n$ is an orthonormal basis of vector fields $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ on $\mathbb{R}^n$. The term "basis" is of course an abuse of notation; what is really meant is that when evaluated at any point, these vector fields form a basis of the tangent space at that point. Moving frames were introduced by E. Cartan to study the geometry of manifolds.

Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ denote a moving frame on $\mathbb{R}^n$. Given $\boldsymbol{p} \in \mathbb{R}^n$ and $1 \le i, j \le n$, the map

$$\begin{aligned} \mathbb{R}^n_{\boldsymbol{p}} &\longrightarrow \mathbb{R} \\ \boldsymbol{u} &\longmapsto \langle \nabla_{\boldsymbol{u}} \boldsymbol{X}_i, \boldsymbol{X}_j(\boldsymbol{p})\rangle \end{aligned}$$

is linear, and thus defines a one-form $\omega_i^j(\boldsymbol{p})$ on $\mathbb{R}^n_{\boldsymbol{p}}$. The n^2 forms ω_i^j are called the *connection forms* of $\mathbb{R}^n$ in the moving frame.

(a) Show that $\omega_i^j = -\omega_j^i$.

(b) Let $X_i^j = \langle \boldsymbol{X}_i, \boldsymbol{D}_j\rangle$. By definition,

$$\nabla_{\boldsymbol{X}} \boldsymbol{X}_i = \sum_k \omega_i^k(\boldsymbol{X}) \boldsymbol{X}_k = \sum_{k,j} \omega_i^k(\boldsymbol{X}) X_k^j \boldsymbol{D}_j.$$

Use this to prove that $\mathrm{d}X_i^j = \sum_k \omega_i^k X_k^j$.

(c) Let $\omega_1, \dots, \omega_n$ denote the one-forms that are dual to the moving frame: $\omega_i = \boldsymbol{X}_i^\flat$, so that $\omega_i = \sum_j X_i^j \, \mathrm{d}u^j$. Prove the *structural equations of* $\mathbb{R}^n$:

$$\mathrm{d}\omega_i = -\sum_k \omega_i^k \wedge \omega_k \tag{5.9.1}$$

$$\mathrm{d}\omega_i^j = \sum_k \omega_i^k \wedge \omega_k^j. \tag{5.9.2}$$

5.19. A *Riemannian manifold* is a pair $(M, \boldsymbol{g})$, where M is a manifold and $\boldsymbol{g}$ a Riemannian metric on M, cf. Example 5.1.1. Although we have mostly been using the metric induced from the ambient Euclidean space, there are many other interesting examples: given $\kappa \in \mathbb{R}$, denote by U_κ the open set

$$U_\kappa = \begin{cases} \mathbb{R}^n, & \text{if } \kappa \geq 0; \\ \{\boldsymbol{a} \in \mathbb{R}^n \mid |\boldsymbol{a}|^2 < \frac{1}{|\kappa|}, & \text{if } \kappa < 0\}. \end{cases}$$

Show that the tensor field $\boldsymbol{g}_\kappa$ on U_κ, where

$$\boldsymbol{g}_\kappa(\boldsymbol{a}) = \frac{4}{(1 + \kappa|\boldsymbol{a}|^2)^2} \sum_{i=1}^{n} \mathrm{d}u^i \otimes \mathrm{d}u^i, \qquad \boldsymbol{a} \in U_\kappa,$$

is a Riemannian metric.
Notice that when $\kappa = 0$, we recover standard Euclidean space, which is flat. In general, one can define the curvature tensor of a Riemannian metric in the same way we did for the metric induced from Euclidean space. The same is true for isometries, and as in Euclidean space, the curvature tensor is preserved under isometries. The next exercise shows that when $\kappa > 0$, $(U_\kappa, \boldsymbol{g}_\kappa)$ has constant curvature κ. This turns out to be true also when $\kappa < 0$.

5.20. If $(M_1, \boldsymbol{g}_1)$ and $(M_2, \boldsymbol{g}_2)$ are Riemannian manifolds (see Exercise 5.19), an *isometry* between them is a diffeomorphism $\boldsymbol{f} : M_1 \to M_2$ such that $\boldsymbol{f}^* \boldsymbol{g}_2 = \boldsymbol{g}_1$; i.e.,

$$\boldsymbol{g}_2(\boldsymbol{f}_{*\boldsymbol{p}}\boldsymbol{u}, \boldsymbol{f}_{*\boldsymbol{p}}\boldsymbol{v}) = \boldsymbol{g}_1(\boldsymbol{u}, \boldsymbol{v}), \qquad \boldsymbol{p} \in M_1, \quad \boldsymbol{u}, \boldsymbol{v} \in M_{1\boldsymbol{p}}.$$

When this is the case, the two Riemannian manifolds are said to be isometric. Notice that when the metrics are those induced by the ambient Euclidean space, this definition coincides with the usual one.
Prove that when $\kappa > 0$, the Riemannian manifold $(U_\kappa, \boldsymbol{g}_\kappa)$ from Exercise 5.19 is isometric to a sphere of radius $1/\sqrt{\kappa}$ with a point removed (and therefore with constant curvature κ) in $\mathbb{R}^{n+1}$ with the usual metric. *Hint:* Consider stereographic projection.

5.21. In Examples and Remarks 5.4.1 (ii), we defined the integral of a 1-form $\omega \in \Lambda_k(\mathbb{R}^n)$ along a curve $\boldsymbol{c} : I \to \mathbb{R}^n$ to be

$$\int_{\boldsymbol{c}} \omega = \int_I \boldsymbol{c}^* \omega.$$

Let ω be a 1-form on a connected open set U in $\mathbb{R}^n$, and consider the following three statements:
(1) ω is exact in U.
(2) If $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are curves in U with the same beginning and endpoints, then $\int_{\boldsymbol{c}_1} \omega = \int_{\boldsymbol{c}_2} \omega$.
(3) If $\boldsymbol{c}$ is a closed curve (i.e., beginning and endpoints coincide) in U, then $\int_{\boldsymbol{c}} \omega = 0$.

(a) Prove that 1 implies 2.
(b) Conclude that all three statements are equivalent. *Hint:* to show that 2 implies 1, fix any $\boldsymbol{p} \in U$, and define a function f on U by $f(\boldsymbol{a}) = \int_{\boldsymbol{c}} \omega$, where $\boldsymbol{c}$ is any curve in U beginning at $\boldsymbol{p}$ and ending at $\boldsymbol{a}$. Show that $\omega = \mathrm{d}f$.

5.22. Recall that the 0-th cohomology space of M is

$$H^0(M) = Z^0(M) = \{f : M \to \mathbb{R} \mid \mathrm{d}f = 0\}.$$

Prove that if M is connected, then $H^0(M) \cong \mathbb{R}$.

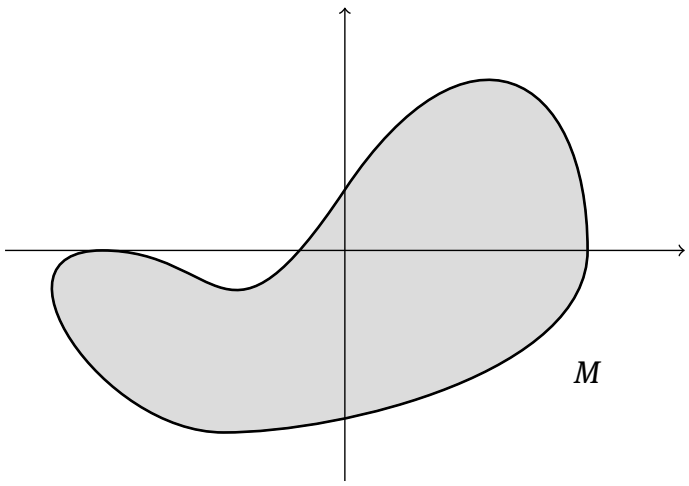
5.23. Let $\boldsymbol{a} \in \mathbb{R}^n$, $r > 0$, and M be the n-dimensional manifold with boundary $\overline{B_r}(\boldsymbol{a})$. Show that every closed k-form on M is exact.

5.24. Give examples of manifolds with boundary where the manifold boundary does not coincide with the topological boundary of the set M from Chapter 1. Is one of these boundaries always contained in the other?

5.25. Let M, N be open sets in $\mathbb{R}^n$ whose topological boundaries are smooth $(n-1)$-dimensional manifolds, so that $\bar{M}$ and $\bar{N}$ are manifolds with boundaries ∂M and ∂N respectively.
(a) Show that if $\bar{N} \subset M$, then $\int_{\partial N} \omega = \int_{\partial M} \omega$ for any closed $(n-1)$-form ω. (It should really be the pullback of ω by the respective inclusion maps, of course). *Hint:* apply Stokes' theorem to $\bar{M} \setminus N$.
(b) Show that if

$$\omega = \frac{-u^2}{(u^1)^2 + (u^2)^2}\,\mathrm{d}u^1 + \frac{u^1}{(u^1)^2 + (u^2)^2}\,\mathrm{d}u^2,$$

then $\int_{\partial M} \omega = 2\pi$, with M the region pictured below. *Hint:* see Examples and Remarks 5.6.1 (iii).



(c) Prove that if M is the region above but translated so that it no longer contains the origin, then $\int_{\partial M} \omega = 0$.

5.26. The higher-dimensional version of the 1-form ω on $\mathbb{R}^2 \setminus \{\mathbf{0}\}$ from part (b) in the previous problem is the 2-form

$$\Omega = \frac{1}{((u^1)^2 + (u^2)^2 + (u^3)^2)}(u^1\,\mathrm{d}u^2 \wedge \mathrm{d}u^3 + u^2\,\mathrm{d}u^3 \wedge \mathrm{d}u^1 + u^3\,\mathrm{d}u^1 \wedge \mathrm{d}u^2)$$

on $\mathbb{R}^3 \setminus \{\mathbf{0}\}$.

(a) Show that Ω is closed.

(b) Extend the concepts of determinant and cross product to the tangent space of $\mathbb{R}^3$ at any point $\boldsymbol{p}$ by means of the canonical isomorphism $\mathcal{I}_{\boldsymbol{p}} : \mathbb{R}^3 \to \mathbb{R}^3_{\boldsymbol{p}}$. Show that for $\alpha = u^1\,\mathrm{d}u^2 \wedge \mathrm{d}u^3 + u^2\,\mathrm{d}u^3 \wedge \mathrm{d}u^1 + u^3\,\mathrm{d}u^1 \wedge \mathrm{d}u^2$,

$$\alpha(\boldsymbol{p})(\boldsymbol{u},\boldsymbol{v}) = \det\begin{bmatrix}\boldsymbol{P}(\boldsymbol{p}) & \boldsymbol{u} & \boldsymbol{v}\end{bmatrix} = \langle \boldsymbol{P}(\boldsymbol{p}), \boldsymbol{u}\times\boldsymbol{v}\rangle, \quad \boldsymbol{p}\in\mathbb{R}^3\setminus\{\mathbf{0}\}, \quad \boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^3_{\boldsymbol{p}},$$

where $\boldsymbol{P} = \sum_i u^i \boldsymbol{D}_i$ is the position vector field.

(c) Conclude that if μ is the volume form of $S^2(r)$ and $\imath : S^2(r) \hookrightarrow \mathbb{R}^3 \setminus \{\mathbf{0}\}$ denotes the inclusion map, then

$$\mu = r^2\imath^*\Omega, \text{ so that } \int_{S^2(r)} \imath^*\Omega = 4\pi.$$

Why does this imply that Ω is not exact?

(d) Let M be a bounded, open set in $\mathbb{R}^3$ whose topological boundary ∂M is a smooth manifold. Show that the integral of Ω over ∂M equals zero if $\bar{M}$ does not contain the origin, and 4π otherwise.

5.27. Let α be the 1-form on $\mathbb{R}^3$ given by

$$\alpha = \ln((u^2)^2 + 1)\,\mathrm{d}u^1 + u^1u^2u^3\,\mathrm{d}u^2 - (u^2)^2 e^{u^1u^2}\,\mathrm{d}u^3,$$

and M the manifold with boundary consisting of the upper hemisphere of the unit sphere centered at the origin, oriented so that the position vector field is the positive unit normal field. If $\imath : M \hookrightarrow \mathbb{R}^3$ denotes inclusion, determine $\int_M \imath^*\,\mathrm{d}\alpha$. *Hint:* Direct computation is difficult, and applying Stokes' theorem isn't easy either. The integral of α along the boundary ∂M may, however, be computed by applying Stokes' theorem to *any* manifold that has ∂M as boundary.

5.28. Suppose that a is a regular value of $f : \mathbb{R}^3 \to \mathbb{R}$, and that $M = f^{-1}(a) \neq \emptyset$, so that M is a 2-dimensional manifold. Let α be the 2-form on $\mathbb{R}^3$ given by

$$\alpha = D_1 f\,\mathrm{d}u^2 \wedge \mathrm{d}u^3 + D_2 f\,\mathrm{d}u^3 \wedge \mathrm{d}u^1 + D_3 f\,\mathrm{d}u^1 \wedge \mathrm{d}u^2.$$

Show that if $\imath : M \hookrightarrow \mathbb{R}^3$ is inclusion, then $\imath^*\alpha$ is a nowhere-zero 2-form on M, so that the latter is orientable. In fact, prove that $(1/|\nabla f|)\imath^*\alpha$ is the volume form of this orientation. *Hint:* the 2-form α from Exercise 5.26 (b) is the special case when $f(\boldsymbol{x}) = (1/2)|\boldsymbol{x}|^2$.

5.29. The divergence of a vector field on $\mathbb{R}^n$ was defined in Exercise 5.13 for arbitrary n. Formulate and prove a divergence theorem in $\mathbb{R}^n$.

5.30. The *Laplacian* of $f : \mathbb{R}^n \to \mathbb{R}$ is the function $\Delta f = \operatorname{div} \nabla f$. f is said to be *harmonic* if its Laplacian is identically zero.
Let M be a connected, compact 3-dimensional manifold with boundary in $\mathbb{R}^3$, $f : M \to \mathbb{R}$ a harmonic function.

(a) Show that $\int_{\partial M} \langle \nabla f, \boldsymbol{N} \rangle \, \eta = 0$.

(b) Prove that

$$\int_{\partial M} \langle f \nabla f, \boldsymbol{N} \rangle \, \eta = \int_M |\nabla f|^2 \, \mathrm{d}u^1 \wedge \mathrm{d}u^2 \wedge \mathrm{d}u^3.$$

Conclude that the only harmonic functions f on M that satisfy

$$\int_{\partial M} \langle f \nabla f, \boldsymbol{N} \rangle \, \eta = 0$$

are the constant ones.

5.31. Let $f, g : \mathbb{R}^3 \to \mathbb{R}$ be differentiable, and $M \subset \mathbb{R}^3$ a compact 3-dimensional manifold with boundary. Prove *Green's identities*:

$$\int_M (\langle \nabla f, \nabla g \rangle + f \Delta g) \, \mathrm{d}u^1 \wedge \mathrm{d}u^2 \wedge \mathrm{d}u^3 = \int_{\partial M} f \langle \nabla g, \boldsymbol{N} \rangle \, \eta;$$

$$\int_M (f \Delta g - g \Delta f) \, \mathrm{d}u^1 \wedge \mathrm{d}u^2 \wedge \mathrm{d}u^3 = \int_{\partial M} \langle f \nabla g - g \nabla f, \boldsymbol{N} \rangle \, \eta.$$

5.32. (a) Prove that $g : \mathbb{R}^3 \setminus \{\mathbf{0}\} \to \mathbb{R}$, where $g(\boldsymbol{p}) = 1/|\boldsymbol{p}|$, is harmonic.

(b) Suppose $f : \mathbb{R}^3 \to \mathbb{R}$ is harmonic on some open set U, and consider $\boldsymbol{p}_0$, $r > 0$ such that $B_r(\boldsymbol{p}_0) \subset U$. Show that for any $0 < r' < r$,

$$\frac{1}{4\pi r'^2} \int_{\partial B_{r'}(\boldsymbol{p}_0)} f \, \eta = \frac{1}{4\pi r^2} \int_{\partial B_r(\boldsymbol{p}_0)} f \, \eta,$$

by applying Green's second identity (Exercise 5.31) to f and g on the manifold with boundary $\overline{B_r}(\boldsymbol{p}_0) \setminus B_{r'}(\boldsymbol{p}_0)$. Conclude that

$$f(\boldsymbol{p}_0) = \frac{1}{4\pi r^2} \int_{\partial B_r(\boldsymbol{p}_0)} f \, \eta.$$

This identity is known as the *mean value theorem for harmonic functions*.

(c) Prove the *maximum principle for harmonic functions*: If f is harmonic on a closed and bounded region $K \subset \mathbb{R}^3$, then the maximum and minimum of f occur on the (topological) boundary of K. *Hint:* We may assume f is not constant. If, say, the maximum were to occur at an interior point $\boldsymbol{p}_0$, this would contradict the mean value theorem.

5.33. Let $\boldsymbol{X}$ be a vector field on $\mathbb{R}^3$, and $\omega = \boldsymbol{X}^\flat$ the dual 1-form. Given $\boldsymbol{a} \in \mathbb{R}^3$, consider any plane P through $\boldsymbol{a}$ and a unit vector $\boldsymbol{n}$ perpendicular to P. As usual, we identify $\boldsymbol{n}$ with a vector in the tangent space at $\boldsymbol{a}$. Show that if S_r denotes the circle of radius r centered at $\boldsymbol{a}$ that lies in P, then

$$\langle \operatorname{curl} \boldsymbol{X}(\boldsymbol{a}), \boldsymbol{n} \rangle = \lim_{r \to 0} \frac{1}{\pi r^2} \int_{S_r} \imath^* \omega .$$

Here $\imath$ denotes the inclusion map of the circle in the disk of radius r in P, and the disk is oriented so that $\boldsymbol{n}$ is the positive unit normal.
If $\boldsymbol{X}$ represents the velocity field of a fluid in $\mathbb{R}^3$, interpret the above identity in terms of rotation of the fluid about an axis. Which axis direction measures maximal rotation? A fluid with zero curl everywhere is said to be *irrotational*.

6 Manifolds as metric spaces

In this chapter, we endow any connected manifold M with a distance function (see Section 1.8). This distance is not the restriction of the standard distance in the ambient Euclidean space, but rather the length of the shortest path in the space between two points, provided such a path exists. It tuns out that the open sets this distance generates coincide with the open sets of M induced by the ambient space: i.e., those of the form $U \cap M$, where U is open in the ambient space. We begin by taking a more detailed look at geodesics, since they represent the kind of curves having this property.

6.1 Extremal properties of geodesics

One aim of this section is to establish the fact that geodesics are locally length-minimizing; i.e., if $\boldsymbol{c}$ is a geodesic that is short enough, then it is (modulo reparametrization) the shortest curve between its endpoints. This is clearly not true in general for arbitrarily long geodesics, since for example on a sphere, any geodesic only minimizes up to the antipodal point. In order to state this more precisely, we need to introduce the following concept: If $\boldsymbol{c} : [0, a] \to M$ is a curve in $M \subset \mathbb{R}^n$, a *variation of* $\boldsymbol{c}$ is a map $\boldsymbol{V} : [0, a] \times [-\varepsilon, \varepsilon] \to M$ such that if $\boldsymbol{V}_s : [0, a] \to M$ denotes the curve $\boldsymbol{V}_s(t) = V(t, s)$ for $|s| \leq \varepsilon$, then $\boldsymbol{V}_0 = \boldsymbol{c}$.

Notice that $\nabla_{\boldsymbol{D}} \dot{\boldsymbol{V}}_s(t) = \nabla_{\boldsymbol{D}_1} \boldsymbol{V}_* \boldsymbol{D}_1(t, s)$. In fact, if $\imath_s : \mathbb{R} \to \mathbb{R}^2$ is given by $\imath_s(t) = (t, s)$, then $\boldsymbol{V}_s = \boldsymbol{V} \circ \imath_s$, so that

$$\dot{\boldsymbol{V}}_s = \boldsymbol{V}_* \imath_{s*} \boldsymbol{D} = \boldsymbol{V}_* \boldsymbol{D}_1 \circ \imath_s,$$

and

$$\nabla_{\boldsymbol{D}} \dot{\boldsymbol{V}}_s = \nabla_{\boldsymbol{D}} (\boldsymbol{V}_* \boldsymbol{D}_1 \circ \imath_s) = \nabla_{\imath_{s*} \boldsymbol{D}} \boldsymbol{V}_* \boldsymbol{D}_1 = \nabla_{\boldsymbol{D}_1 \circ \imath_s} \boldsymbol{V}_* \boldsymbol{D}_1 = \left(\nabla_{\boldsymbol{D}_1} \boldsymbol{V}_* \boldsymbol{D}_1 \right) \circ \imath_s.$$

Lemma 6.1.1. *If $\boldsymbol{V}$ is a variation, then $\nabla_{\boldsymbol{D}_1} \boldsymbol{V}_* \boldsymbol{D}_2 = \nabla_{\boldsymbol{D}_2} \boldsymbol{V}_* \boldsymbol{D}_1$.*

Proof. This is, after deciphering notation, just the fact that mixed partial derivatives are equal. First of all, it is enough to show that $D_{\boldsymbol{D}_1} \boldsymbol{V}_* \boldsymbol{D}_2 = D_{\boldsymbol{D}_2} \boldsymbol{V}_* \boldsymbol{D}_1$. Now,

$$\boldsymbol{V}_* \boldsymbol{D}_2(t_0, s_0) = \mathcal{I}_{\boldsymbol{V}(t_0, s_0)} D\boldsymbol{V}(t_0, s_0) \boldsymbol{e}_2 = \mathcal{I}_{\boldsymbol{V}(t_0, s_0)} \begin{bmatrix} D_2 \boldsymbol{V}^1 \\ \vdots \\ D_2 \boldsymbol{V}^n \end{bmatrix} (t_0, s_0),$$

so that

$$D_{\boldsymbol{D}_1} \boldsymbol{V}_* \boldsymbol{D}_2 = \mathcal{I}_{\boldsymbol{V}} \begin{bmatrix} D_1 D_2 \boldsymbol{V}^1 \\ \vdots \\ D_1 D_2 \boldsymbol{V}^n \end{bmatrix} = \mathcal{I}_{\boldsymbol{V}} \begin{bmatrix} D_2 D_1 \boldsymbol{V}^1 \\ \vdots \\ D_2 D_1 \boldsymbol{V}^n \end{bmatrix} = D_{\boldsymbol{D}_2} \boldsymbol{V}_* \boldsymbol{D}_1. \qquad \square$$

Recall that we defined the inner product on the tangent space of $\mathbb{R}^n$ at a point $\boldsymbol{p}$ to be the one for which the canonical isomorphism $\mathcal{I}_{\boldsymbol{p}} : \mathbb{R}^n \to \mathbb{R}^n_{\boldsymbol{p}}$ becomes a linear isometry.

Given a manifold M and $\boldsymbol{p} \in M$, the tangent space $M_{\boldsymbol{p}}$ is also an inner product space, and we can, in the same way, define for any $\boldsymbol{v} \in M_{\boldsymbol{p}}$ a *canonical inner product on* $(M_{\boldsymbol{p}})_{\boldsymbol{v}}$ by requiring that $\mathcal{I}_{\boldsymbol{v}} : M_{\boldsymbol{p}} \to (M_{\boldsymbol{p}})_{\boldsymbol{v}}$ be a linear isometry. The following result may then be interpreted as saying that the derivative of the exponential map at a point is "radially" isometric. Its kernel is orthogonal to the radial direction.

Lemma 6.1.2 (The Gauss lemma). *Let $\boldsymbol{p} \in M$, $\boldsymbol{v} \in \tilde{M}_{\boldsymbol{p}} \subset M_{\boldsymbol{p}}$ in the domain $\tilde{M}_{\boldsymbol{p}}$ of $\exp_{\boldsymbol{p}}$. Consider the ray $\boldsymbol{\varphi}_{\boldsymbol{v}} : \mathbb{R} \to M_{\boldsymbol{p}}$ in direction $\boldsymbol{v}$, $\boldsymbol{\varphi}_{\boldsymbol{v}}(t) = t\boldsymbol{v}$, and let $\boldsymbol{u} = \dot{\boldsymbol{\varphi}}_{\boldsymbol{v}}(1) \in (M_{\boldsymbol{p}})_{\boldsymbol{v}}$. If $\boldsymbol{w} \in (M_{\boldsymbol{p}})_{\boldsymbol{v}}$, then*

$$\langle \exp_{\boldsymbol{p}*} \boldsymbol{u}, \exp_{\boldsymbol{p}*} \boldsymbol{w} \rangle = \langle \boldsymbol{u}, \boldsymbol{w} \rangle.$$

Proof. We may assume $\boldsymbol{v} \neq \boldsymbol{0}$, for otherwise $\boldsymbol{u} = \boldsymbol{0}$, and the statement is trivial. Consider the variation

$$\begin{aligned} \boldsymbol{V} : [0,1] \times [-\varepsilon, \varepsilon] &\to M, \\ (t,s) &\mapsto \exp_{\boldsymbol{p}} \left(t(\boldsymbol{v} + s\mathcal{I}_{\boldsymbol{v}}^{-1}\boldsymbol{w}) \right) \end{aligned}$$

of $\boldsymbol{c}_{\boldsymbol{v}}$, where ε is chosen small enough that $\boldsymbol{V}$ is defined. Each $\boldsymbol{V}_s$ is a geodesic, so that

$$\left(\nabla_{\boldsymbol{D}_1} \boldsymbol{V}_* \boldsymbol{D}_1 \right)(t,s) = \left(\nabla_{\boldsymbol{D}} \dot{\boldsymbol{V}}_s \right)(t) = 0.$$

Furthermore, the tangent vector field of $\boldsymbol{V}_s$ has constant norm equal to

$$|\dot{\boldsymbol{V}}_s(0)| = |\boldsymbol{V}_* \boldsymbol{D}_1(0,s)| = |\boldsymbol{v} + s\mathcal{I}_{\boldsymbol{v}}^{-1}\boldsymbol{w}|.$$

Since the partial derivative of $\langle \boldsymbol{V}_* \boldsymbol{D}_1, \boldsymbol{V}_* \boldsymbol{D}_2 \rangle$ with respect to t equals

$$\begin{aligned} D_1 \langle \boldsymbol{V}_* \boldsymbol{D}_1, \boldsymbol{V}_* \boldsymbol{D}_2 \rangle &= \langle \nabla_{\boldsymbol{D}_1} \boldsymbol{V}_* \boldsymbol{D}_1, \boldsymbol{V}_* \boldsymbol{V}_2 \rangle + \langle \boldsymbol{V}_* \boldsymbol{D}_1, \nabla_{\boldsymbol{D}_1} \boldsymbol{V}_* \boldsymbol{D}_2 \rangle \\ &= \langle \boldsymbol{V}_* \boldsymbol{D}_1, \nabla_{\boldsymbol{D}_1} \boldsymbol{V}_* \boldsymbol{D}_2 \rangle = \langle \boldsymbol{V}_* \boldsymbol{D}_1, \nabla_{\boldsymbol{D}_2} \boldsymbol{V}_* \boldsymbol{D}_1 \rangle \\ &= \frac{1}{2} D_2 \langle \boldsymbol{V}_* \boldsymbol{D}_1, \boldsymbol{V}_* \boldsymbol{D}_1 \rangle, \end{aligned}$$

we deduce that

$$\begin{aligned} D_1 \langle \boldsymbol{V}_* \boldsymbol{D}_1, \boldsymbol{V}_* \boldsymbol{D}_2 \rangle(t,s) &= \frac{1}{2} D_2 (|\boldsymbol{v}|^2 + 2s\langle \boldsymbol{v}, \mathcal{I}_{\boldsymbol{v}}^{-1}\boldsymbol{w} \rangle + s^2 |\mathcal{I}_{\boldsymbol{v}}^{-1}\boldsymbol{w}|) \\ &= \langle \boldsymbol{v}, \mathcal{I}_{\boldsymbol{v}}^{-1}\boldsymbol{w} \rangle + 2s|\mathcal{I}_{\boldsymbol{v}}^{-1}\boldsymbol{w}|, \end{aligned}$$

which becomes $\langle \boldsymbol{v}, \mathcal{I}_{\boldsymbol{v}}^{-1}\boldsymbol{w} \rangle$ when $s = 0$. Now,

$$\langle \boldsymbol{v}, \mathcal{I}_{\boldsymbol{v}}^{-1}\boldsymbol{w} \rangle = \langle \mathcal{I}_{\boldsymbol{v}} \boldsymbol{v}, \boldsymbol{w} \rangle = \langle \boldsymbol{u}, \boldsymbol{w} \rangle,$$

so that the function

$$t \mapsto \langle \boldsymbol{V}_* \boldsymbol{D}_1, \boldsymbol{V}_* \boldsymbol{D}_2 \rangle(t,0)$$

has constant derivative $\langle \boldsymbol{u}, \boldsymbol{w} \rangle$ on $[0,1]$. But it vanishes at the origin (because $\boldsymbol{V}_* \boldsymbol{D}_2(0,0) = \boldsymbol{0}$), so that

$$\langle \boldsymbol{V}_* \boldsymbol{D}_1, \boldsymbol{V}_* \boldsymbol{D}_2 \rangle(t,0) = t\langle \boldsymbol{u}, \boldsymbol{w} \rangle. \tag{6.1.1}$$

On the other hand, the very definition of $\boldsymbol{V}$ implies that

$$\langle \boldsymbol{V}_* \boldsymbol{D}_1, \boldsymbol{V}_* \boldsymbol{D}_2 \rangle (1,0) = \langle \exp_{\boldsymbol{p}*} \boldsymbol{u}, \exp_{\boldsymbol{p}*} \boldsymbol{w} \rangle.$$

Comparing this expression with (6.1.1) now yields the statement. □

Consider a curve $\boldsymbol{\gamma} : I \to M_{\boldsymbol{p}}$. If $\boldsymbol{\gamma}(t) \neq \boldsymbol{0}$, we may write $\dot{\boldsymbol{\gamma}}(t) = \dot{\boldsymbol{\gamma}}_r(t) + \dot{\boldsymbol{\gamma}}_\theta(t)$, with

$$\dot{\boldsymbol{\gamma}}_r = \frac{1}{|\boldsymbol{\gamma}|^2} \langle \dot{\boldsymbol{\gamma}}, \mathcal{I}_{\boldsymbol{\gamma}} \boldsymbol{\gamma} \rangle \mathcal{I}_{\boldsymbol{\gamma}} \boldsymbol{\gamma} \quad \text{and} \quad \dot{\boldsymbol{\gamma}}_\theta = \dot{\boldsymbol{\gamma}} - \dot{\boldsymbol{\gamma}}_r.$$

$\dot{\boldsymbol{\gamma}}_r$ and $\dot{\boldsymbol{\gamma}}_\theta$ are called the *radial* and *polar* components of $\dot{\boldsymbol{\gamma}}$ respectively. Notice that they are mutually orthogonal, because

$$\begin{aligned} \langle \dot{\boldsymbol{\gamma}}_\theta, \dot{\boldsymbol{\gamma}}_r \rangle &= \langle \dot{\boldsymbol{\gamma}}, \dot{\boldsymbol{\gamma}}_r \rangle - \langle \dot{\boldsymbol{\gamma}}_r, \dot{\boldsymbol{\gamma}}_r \rangle \\ &= \frac{1}{|\boldsymbol{\gamma}|^2} \langle \dot{\boldsymbol{\gamma}}, \mathcal{I}_{\boldsymbol{\gamma}} \boldsymbol{\gamma} \rangle^2 - \frac{1}{|\boldsymbol{\gamma}|^4} \langle \dot{\boldsymbol{\gamma}}, \mathcal{I}_{\boldsymbol{\gamma}} \boldsymbol{\gamma} \rangle^2 \langle \mathcal{I}_{\boldsymbol{\gamma}} \boldsymbol{\gamma}, \mathcal{I}_{\boldsymbol{\gamma}} \boldsymbol{\gamma} \rangle \\ &= 0, \end{aligned}$$

since $\langle \mathcal{I}_{\boldsymbol{\gamma}} \boldsymbol{\gamma}, \mathcal{I}_{\boldsymbol{\gamma}} \boldsymbol{\gamma} \rangle = |\boldsymbol{\gamma}|^2$ by definition of the inner product on the "double" tangent space $(M_{\boldsymbol{p}})_{\boldsymbol{\gamma}(t)}$. When $\boldsymbol{\gamma}$ is a ray $\boldsymbol{\varphi}_{\boldsymbol{v}}$, i.e., $\boldsymbol{\gamma}(t) = t\boldsymbol{v}$ for some $\boldsymbol{v} \in M_{\boldsymbol{p}}$, then $\dot{\boldsymbol{\gamma}} = \dot{\boldsymbol{\gamma}}_r$, and $\boldsymbol{\gamma}$ is length-minimizing. The following lemma says that this property is preserved under the exponential map:

Lemma 6.1.3. *Let $\boldsymbol{p} \in M$, and consider a vector $\boldsymbol{v}$ in the domain $\tilde{M}_{\boldsymbol{p}}$ of $\exp_{\boldsymbol{p}}$. Denote by $\boldsymbol{\varphi}_{\boldsymbol{v}} : [0,1] \to \tilde{M}_{\boldsymbol{p}}$ the ray from $\boldsymbol{0}$ to $\boldsymbol{v}$, $\boldsymbol{\varphi}_{\boldsymbol{v}}(t) = t\boldsymbol{v}$. If $\boldsymbol{\gamma} : [0,1] \to \tilde{M}_p$ is any (piecewise-smooth) curve with $\boldsymbol{\gamma}(0) = \boldsymbol{\varphi}_{\boldsymbol{v}}(0) = \boldsymbol{0}$ and $\boldsymbol{\gamma}(1) = \boldsymbol{\varphi}_{\boldsymbol{v}}(1) = \boldsymbol{v}$, then*

$$L(\exp \circ \boldsymbol{\gamma}) \geq L(\exp \circ \boldsymbol{\varphi}_{\boldsymbol{v}}).$$

Inequality is strict if there is some $t_0 \in [0,1]$ for which the polar component of $\dot{\boldsymbol{\gamma}}$ at t_0 does not vanish under the exponential map; i.e., if $\exp_{\boldsymbol{p}*} \dot{\boldsymbol{\gamma}}_\theta(t_0) \neq \boldsymbol{0}$.

Proof. We may assume, by Proposition 2.4.1, that $\boldsymbol{\gamma}$ is differentiable, and that both $\boldsymbol{v}$ and $\boldsymbol{\gamma}(t)$ are nonzero for $t \in [0,1]$. By the Gauss Lemma,

$$|\exp_{\boldsymbol{p}*} \dot{\boldsymbol{\gamma}}|^2 = |\exp_{\boldsymbol{p}*} \dot{\boldsymbol{\gamma}}_r|^2 + |\exp_{\boldsymbol{p}*} \dot{\boldsymbol{\gamma}}_\theta|^2 \geq |\exp_{\boldsymbol{p}*} \dot{\boldsymbol{\gamma}}_r|^2 = |\dot{\boldsymbol{\gamma}}_r|^2. \tag{6.1.2}$$

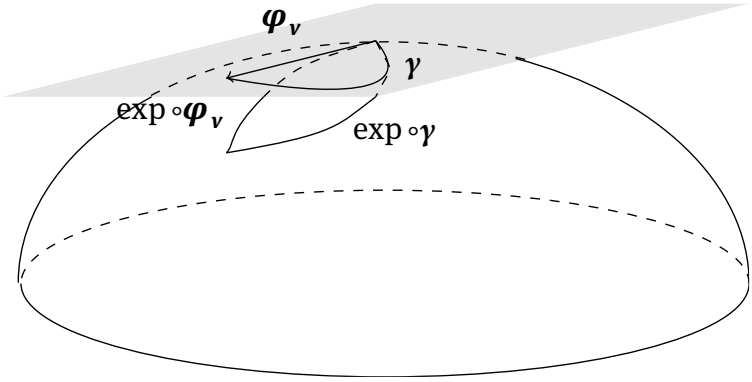
Now, $|\boldsymbol{\gamma}|' = |\dot{\boldsymbol{\gamma}}_r|$, because

$$|\boldsymbol{\gamma}|' = \langle \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle^{1/2\prime} = \frac{\langle \boldsymbol{\gamma}, \boldsymbol{\gamma}' \rangle}{|\boldsymbol{\gamma}|} = \frac{\langle \mathcal{I}_{\boldsymbol{\gamma}} \boldsymbol{\gamma}, \dot{\boldsymbol{\gamma}} \rangle}{|\boldsymbol{\gamma}|} = |\dot{\boldsymbol{\gamma}}_r|.$$

Thus,

$$\begin{aligned} L(\exp_{\boldsymbol{p}} \circ \boldsymbol{\gamma}) &= \int_0^1 |\exp_{\boldsymbol{p}*} \dot{\boldsymbol{\gamma}}| \geq \int_0^1 |\dot{\boldsymbol{\gamma}}_r| = \int_0^1 |\boldsymbol{\gamma}|' = |\boldsymbol{\gamma}(1)| = |\boldsymbol{v}| \\ &= L(\exp_{\boldsymbol{p}} \circ \boldsymbol{\varphi}_{\boldsymbol{v}}). \end{aligned}$$

The last assertion of the lemma is clear, since the inequality in (6.1.2) is strict on some interval around t_0 if $\exp_{\boldsymbol{p}*} \dot{\boldsymbol{\gamma}}_\theta(t_0) \neq 0$. □



The following theorem is a mathematical formulation of the first sentence in this section:

Theorem 6.1.1. *Let $p \in M$, and choose $\varepsilon > 0$ so that $\exp_p : U_\varepsilon \to \exp_p(U_\varepsilon)$ is a diffeomorphism, where $U_\varepsilon = \{v \in M_p \mid |v| < \varepsilon\}$ is the open ball of radius ε centered at $0 \in M_p$. For $v \in U_\varepsilon$, denote as usual by $c_v : [0,1] \to M$ the geodesic in direction v, $c_v(t) = \exp_p(tv)$. Then for any piecewise-smooth curve $c : [0,1] \to M$ with $c(0) = c_v(0) = p$ and $c(1) = c_v(1) = q$, the length of c is at least as great as that of c_v, and is strictly greater unless c equals c_v up to reparametrization.*

Proof. For $u \in U_\varepsilon$, let φ_u be the ray $t \mapsto tu$. Suppose first that the image of c is contained inside $\exp_p(U_\varepsilon)$, so that there exists a lift γ of c in U_ε; i.e., $c = \exp_p \circ \gamma$, $\gamma(0) = 0$, $\gamma(1) = v$. By Lemma 6.1.3, $L(c) \geq L(c_v)$. We claim that if c is not a reparametrization of c_v, then for some $t \in [0,1]$, $\dot{\gamma}$ is not radial (and therefore $L(c) > L(c_v)$ by the same lemma): Otherwise,

$$L(\gamma) = \int_0^1 |\dot{\gamma}| = \int_0^1 |\dot{\gamma}_r| = \int_0^1 |\gamma|' = |\gamma(1)| \tag{6.1.3}$$

as in the proof of Lemma 6.1.3. On the other hand, there exists $t_0 \in (0,1)$ such that $\gamma(t_0) \notin \{sv \mid s \in [0,1]\}$. Then

$$L(\gamma) = L(\gamma|_{[0,t_0]}) + L(\gamma|_{[t_0,1]}) \geq |\gamma(t_0)| + |\gamma(1) - \gamma(t_0)| > |\gamma(1)|,$$

which contradicts (6.1.3). This establishes the result if the image of c lies in $\exp_p(U_\varepsilon)$. Next, suppose c is not entirely contained inside $\exp_p(U_\varepsilon)$, and let $b = \sup\{t \mid c[0,t] \subset \exp_p(U_\varepsilon)\}$. There must exist some $t_0 \in (0,b)$ such that $v_0 := (\exp_p|_{U_\varepsilon})^{-1} c(t_0)$ has norm greater than that of v. Then

$$L(c) \geq L(c|_{[0,t_0]}) \geq |v_0| > |v| = L(c_v). \qquad \square$$

A geodesic c is said to be *minimal* if its length is equal to the distance between its endpoints. Theorem 6.1.1 asserts that for each $p \in M$, there exists an $\varepsilon > 0$ such that all geodesics of length less than ε emanating from p are minimal.

It is also worth noting that any curve c that is minimal in the above sense is necessarily a geodesic: if t_0 is a point in the domain of c, then the restriction of c to $[t_0, t_0 + \varepsilon]$

is also minimal whenever $t_0 + \varepsilon$ lies in the domain of $\boldsymbol{c}$, and by Theorem 6.1.1, it is a geodesic. Since t_0 is arbitrary, the claim follows.

6.2 Jacobi fields

The previous section indicates that the exponential map $\exp_{\boldsymbol{p}}$ at $\boldsymbol{p} \in M$ plays a fundamental role in the geometry of M. It turns out that its derivative can be conveniently expressed in terms of certain vector fields along geodesics emanating from $\boldsymbol{p}$. But first a word on notation: if $\boldsymbol{c}$ is a curve in M and $\boldsymbol{Y}$ a vector field along $\boldsymbol{c}$, we have in the past used the notation $\boldsymbol{Y}'$ to denote the covariant derivative of $\boldsymbol{Y}$ along $\boldsymbol{c}$ *in the ambient Euclidean space*, and $\boldsymbol{Y}'^{\top} = \nabla_{\boldsymbol{D}} Y$ to denote the corresponding covariant derivative in M. This notation becomes cumbersome when taking second derivatives. Furthermore, we also wish to use the superscripts $^{\top}$ and $^{\perp}$ to describe components of vector fields tangent and orthogonal to $\boldsymbol{c}$ respectively rather than to M. With this in mind, we introduce the following

Notation: Until further notice, the covariant derivative $\nabla_{\boldsymbol{D}}\boldsymbol{Y}$ of a vector field $\boldsymbol{Y}$ along a curve $\boldsymbol{c}$ in M will also be denoted by $\boldsymbol{Y}'$.

Definition 6.2.1. Let $\boldsymbol{c}$ be a geodesic in M. A vector field $\boldsymbol{Y}$ along $\boldsymbol{c}$ is called a *Jacobi field* along $\boldsymbol{c}$ if

$$\boldsymbol{Y}'' + R(\boldsymbol{Y}, \dot{\boldsymbol{c}})\dot{\boldsymbol{c}} = 0.$$

Notice that the collection $\mathcal{J}_{\boldsymbol{c}}$ of Jacobi fields along $\boldsymbol{c}$ is a vector space that contains $\dot{\boldsymbol{c}}$. It turns out that the subspace of Jacobi fields orthogonal to $\dot{\boldsymbol{c}}$ is the one of interest to us: If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are Jacobi, then

$$\langle \boldsymbol{Y}'', \boldsymbol{X}\rangle = -\langle R(\boldsymbol{Y}, \dot{\boldsymbol{c}})\dot{\boldsymbol{c}}, \boldsymbol{X}\rangle = -\langle R(\boldsymbol{X}, \dot{\boldsymbol{c}})\dot{\boldsymbol{c}}, \boldsymbol{Y}\rangle = \langle \boldsymbol{X}'', \boldsymbol{Y}\rangle.$$

Thus, $\langle \boldsymbol{X}', \boldsymbol{Y}\rangle - \langle \boldsymbol{Y}', \boldsymbol{X}\rangle$ must be constant, since its derivative is $\langle \boldsymbol{X}'', \boldsymbol{Y}\rangle - \langle \boldsymbol{Y}'', \boldsymbol{X}\rangle = 0$. In particular, $\langle \boldsymbol{Y}, \dot{\boldsymbol{c}}\rangle' = \langle \boldsymbol{Y}', \dot{\boldsymbol{c}}\rangle = \langle \boldsymbol{Y}', \dot{\boldsymbol{c}}\rangle - \langle \boldsymbol{Y}, \dot{\boldsymbol{c}}'\rangle$ is constant, so that for a normal geodesic, the tangential component $\boldsymbol{Y}^{\top}$ of $\boldsymbol{Y}$ is given by

$$\boldsymbol{Y}^{\top} = \langle \boldsymbol{Y}, \dot{\boldsymbol{c}}\rangle\dot{\boldsymbol{c}} = (a + bt)\dot{\boldsymbol{c}}, \qquad a = \langle \boldsymbol{Y}, \dot{\boldsymbol{c}}\rangle(0), \quad b = \langle \boldsymbol{Y}, \dot{c}\rangle'(0),$$

and satisfies the Jacobi equation. It follows that the component $\boldsymbol{Y}^{\perp} = \boldsymbol{Y} - \boldsymbol{Y}^{\top}$ of $\boldsymbol{Y}$ orthogonal to $\dot{\boldsymbol{c}}$ is also a Jacobi field.

Proposition 6.2.1. *Let $\boldsymbol{c} : I \to M$ be a geodesic, $t_0 \in I$. For any $\boldsymbol{v}, \boldsymbol{w} \in M_{\boldsymbol{c}(t_0)}$ there exists a unique Jacobi field $\boldsymbol{Y}$ along $\boldsymbol{c}$ with $\boldsymbol{Y}(t_0) = \boldsymbol{v}$ and $\boldsymbol{Y}'(t_0) = \boldsymbol{w}$.*

Proof. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be parallel fields along $\boldsymbol{c}$ with $\boldsymbol{X}_n = \dot{\boldsymbol{c}}$, and such that $\boldsymbol{X}_1(t_0), \ldots, \boldsymbol{X}_{n-1}(t_0)$ form an orthonormal basis of $\dot{\boldsymbol{c}}(t_0)^{\perp}$. Any vector field $\boldsymbol{Y}$ along $\boldsymbol{c}$ can then be

expressed as

$$\boldsymbol{Y} = \sum_i f^i \boldsymbol{X}_i, \qquad f^i = \begin{cases} \langle \boldsymbol{Y}, \boldsymbol{X}_i \rangle, & \text{for } i \leq n-1, \\ \langle \boldsymbol{Y}, \frac{\boldsymbol{X}_n}{|\boldsymbol{X}_n|^2} \rangle, & \text{for } i = n. \end{cases}$$

Since $\boldsymbol{X}_i$ is parallel, $\boldsymbol{Y}'' = \sum f^{i\prime\prime} \boldsymbol{X}_i$. Furthermore, $R(\boldsymbol{X}_i, \dot{\boldsymbol{c}})\dot{\boldsymbol{c}} = \sum_{j=1}^{n-1} h_i^j \boldsymbol{X}_j$, where $h_i^j = \langle R(\boldsymbol{X}_i, \dot{\boldsymbol{c}})\dot{\boldsymbol{c}}, \boldsymbol{X}_j \rangle$, so that $R(\boldsymbol{Y}, \dot{\boldsymbol{c}})\dot{\boldsymbol{c}} = \sum_{i,j=1}^{n-1} f^i h_i^j \boldsymbol{X}_j$. The Jacobi equation then translates into

$$\sum_{j=1}^{n-1} \left(f^{j\prime\prime} + \sum_{i=1}^{n-1} f^i h_i^j \right) \boldsymbol{X}_j = 0, \qquad f^{n\prime\prime} = 0,$$

or equivalently, since the $\boldsymbol{X}_i$ are linearly independent everywhere and $h_i^n = \langle R(\boldsymbol{X}_i, \dot{\boldsymbol{c}})\dot{\boldsymbol{c}}, \dot{\boldsymbol{c}} \rangle = 0$,

$$f^{j\prime\prime} + \sum_{i=1}^{n-1} f^i h_i^j = 0, \qquad j = 1, \ldots, n.$$

This is a homogeneous system of n linear second-order equations, which has a unique solution given initial values for f^j and $f^{j\prime}$ at t_0; in our case, we have $f^j(t_0) = \langle \boldsymbol{v}, \boldsymbol{X}_j(t_0) \rangle$, $f^{j\prime}(t_0) = \langle \boldsymbol{w}, \boldsymbol{X}_j(t_0) \rangle$ $(j < n)$, $f^n(t_0) = \langle \boldsymbol{v}, (\dot{\boldsymbol{c}}/|\dot{\boldsymbol{c}}|^2)(t_0) \rangle$, and $f^{n\prime}(t_0) = \langle \boldsymbol{w}, (\dot{\boldsymbol{c}}/|\dot{\boldsymbol{c}}|^2)(t_0) \rangle$, thereby establishing the result. □

The existence part of Proposition 6.2.1 implies that the linear map

$$\begin{aligned} \mathcal{J}_{\boldsymbol{c}} &\to M_{\boldsymbol{c}(t_0)} \times M_{\boldsymbol{c}(t_0)}, \\ \boldsymbol{Y} &\mapsto (\boldsymbol{Y}(t_0), \boldsymbol{Y}'(t_0)) \end{aligned}$$

is onto. The uniqueness part implies it has trivial kernel. In other words, it is an isomorphism, and the space $\mathcal{J}_{\boldsymbol{c}}$ of Jacobi fields along $\boldsymbol{c}$ has dimension $2n$.

Example 6.2.1. Let M^n be a space of constant curvature κ, and let c_κ, s_κ denote the solutions of the differential equation

$$f'' + \kappa f = 0$$

with $c_\kappa(0) = 1$, $c_\kappa'(0) = 0$, $s_\kappa(0) = 0$, $s_\kappa'(0) = 1$.Thus, $c_\kappa(t) = \cos \sqrt{\kappa} t$ if $\kappa > 0$, 1 if $\kappa = 0$, $\cosh \sqrt{-\kappa} t$ if κ is negative, whereas s_κ is obtained by replacing cos, 1, and cosh by sin, $1_{\mathbb{R}}$, and sinh respectively. Consider a normal geodesic $\boldsymbol{c} : [0, b] \to M$. Given $\boldsymbol{v}, \boldsymbol{w} \in M_{\boldsymbol{c}(0)}$ orthogonal to $\dot{\boldsymbol{c}}(0)$, the Jacobi field $\boldsymbol{Y}$ along $\boldsymbol{c}$ with $\boldsymbol{Y}(0) = \boldsymbol{v}$ and $\boldsymbol{Y}'(0) = \boldsymbol{w}$ is given by

$$\boldsymbol{Y} = c_\kappa \boldsymbol{E} + s_\kappa \boldsymbol{F},$$

where $\boldsymbol{E}$ and $\boldsymbol{F}$ are the parallel fields along $\boldsymbol{c}$ with $\boldsymbol{E}(0) = \boldsymbol{v}$ and $\boldsymbol{F}(0) = \boldsymbol{w}$: Indeed, $\boldsymbol{Y}'' = c_\kappa'' \boldsymbol{E} + s_\kappa'' \boldsymbol{F} = -\kappa \boldsymbol{Y} = -R(\boldsymbol{Y}, \dot{\boldsymbol{c}})\dot{\boldsymbol{c}}$, so that $\boldsymbol{Y}$ is a Jacobi field, and clearly satisfies the initial conditions at 0.

Jacobi fields essentially arise out of variations where all curves are geodesics: If $\boldsymbol{V} : [0, a] \times I \to M$ is a variation of $\boldsymbol{c}$ with $\boldsymbol{V}(t, 0) = \boldsymbol{c}(t)$ for $t \in [0, a]$, the *variational vector field* of V is the vector field $\boldsymbol{Y}$ along $\boldsymbol{c}$ given by $\boldsymbol{Y}(t) = \boldsymbol{V}_* \boldsymbol{D}_2(t, 0)$.

Proposition 6.2.2. *Let $\boldsymbol{c} : [0,a] \to M$ be a geodesic. If $\boldsymbol{V}$ is a variation of $\boldsymbol{c}$ through geodesics – meaning that $\boldsymbol{V}_s$ is a geodesic for each s, then the variational vector field of $\boldsymbol{V}$ is Jacobi along $\boldsymbol{c}$. Conversely, let $\boldsymbol{Y}$ be a Jacobi field along $\boldsymbol{c}$. Then there exists a variation $\boldsymbol{V}$ of $\boldsymbol{c}$ through geodesics whose variational vector field equals $\boldsymbol{Y}$.*

Proof. Given a variation $\boldsymbol{V}$ of $\boldsymbol{c}$ through geodesics, define vector fields $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ along $\boldsymbol{V}$ by $\tilde{\boldsymbol{X}} = \boldsymbol{V}_*\boldsymbol{D}_1$, $\tilde{\boldsymbol{Y}} = \boldsymbol{V}_*\boldsymbol{D}_2$. By assumption, $\nabla_{\boldsymbol{D}_1}\tilde{\boldsymbol{X}} = \boldsymbol{0}$, so that

$$\begin{aligned} R(\tilde{\boldsymbol{Y}},\tilde{\boldsymbol{X}})\tilde{\boldsymbol{X}} &= \nabla_{\boldsymbol{D}_2}\nabla_{\boldsymbol{D}_1}\tilde{\boldsymbol{X}} - \nabla_{\boldsymbol{D}_1}\nabla_{\boldsymbol{D}_2}\tilde{\boldsymbol{X}} = -\nabla_{\boldsymbol{D}_1}\nabla_{\boldsymbol{D}_2}\tilde{\boldsymbol{X}} \\ &= -\nabla_{\boldsymbol{D}_1}\nabla_{\boldsymbol{D}_2}V_*\boldsymbol{D}_1 = -\nabla_{\boldsymbol{D}_1}\nabla_{\boldsymbol{D}_1}V_*\boldsymbol{D}_2 \\ &= -\nabla_{\boldsymbol{D}_1}\nabla_{\boldsymbol{D}_1}\tilde{\boldsymbol{Y}}. \end{aligned}$$

When $s = 0$, the above expression becomes $R(\boldsymbol{Y},\dot{\boldsymbol{c}})\dot{\boldsymbol{c}} = -\boldsymbol{Y}''$, and $\boldsymbol{Y}$ is Jacobi.

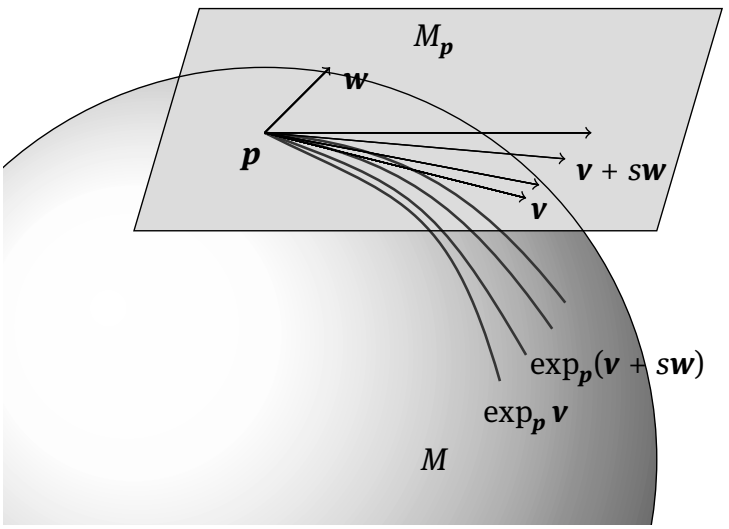
Conversely, suppose $\boldsymbol{Y}$ is a Jacobi field along $\boldsymbol{c}$, and set $\boldsymbol{v} := \boldsymbol{Y}(0)$, $\boldsymbol{w} := \boldsymbol{Y}'(0)$. Let $\boldsymbol{\gamma}$ be a curve with $\dot{\boldsymbol{\gamma}}(0) = \boldsymbol{v}$, and $\boldsymbol{X}$, $\boldsymbol{W}$ parallel fields along $\boldsymbol{\gamma}$ with $\boldsymbol{X}(0) = \dot{\boldsymbol{c}}(0)$, $\boldsymbol{W}(0) = \boldsymbol{w}$. Since $t(\boldsymbol{X}(0) + s\boldsymbol{W}(0))$ belongs to the domain of $\exp_{\boldsymbol{\gamma}(0)}$ for $0 \le t \le a$ and small enough s, there must exist $\varepsilon > 0$ small enough so that $t(\boldsymbol{X}(s) + s\boldsymbol{W}(s))$ belongs to the domain of $\exp_{\boldsymbol{\gamma}(s)}$ for $(t,s) \in [0,a] \times (-\varepsilon,\varepsilon)$. Consider the variation

$$\begin{aligned} \boldsymbol{V} : [0,a] \times (-\varepsilon,\varepsilon) &\to M, \\ (t,s) &\mapsto \exp_{\boldsymbol{\gamma}(s)} t(\boldsymbol{X}(s) + s\boldsymbol{W}(s)) \end{aligned}$$

of $\boldsymbol{c}$. Now, the curves $t \mapsto \boldsymbol{V}(t,s)$ are geodesics, so the variational vector field $\boldsymbol{Z}$ is Jacobi along $\boldsymbol{c}$. Moreover, $\boldsymbol{V}(0,s) = \boldsymbol{\gamma}(s)$, so that $\boldsymbol{Z}(0) = \dot{\boldsymbol{\gamma}}(0) = \boldsymbol{v}$. Finally,

$$\boldsymbol{Z}'(0) = \nabla_{\boldsymbol{D}_1(0,0)}\boldsymbol{V}_*\boldsymbol{D}_2 = \nabla_{\boldsymbol{D}_2(0,0)}\boldsymbol{V}_*\boldsymbol{D}_1 = \boldsymbol{W}(0) = \boldsymbol{w},$$

because $\boldsymbol{V}_*\boldsymbol{D}_1(0,s) = \boldsymbol{X}(s) + s\boldsymbol{W}(s)$, and $\boldsymbol{X}$, $\boldsymbol{W}$ are parallel along $\boldsymbol{\gamma}$. By Proposition 6.2.1, $\boldsymbol{Z} = \boldsymbol{Y}$. □



A geodesic variation of $t \mapsto \exp_{\boldsymbol{p}} t\boldsymbol{v}$ with fixed initial point. The corresponding Jacobi field $\boldsymbol{Y}$ vanishes at $t = 0$

In the special case when $\boldsymbol{Y}(0) = 0$, the variation from Proposition 6.2.2 becomes $\boldsymbol{V}(t,s) = \exp_{\boldsymbol{c}(0)} t(\dot{\boldsymbol{c}}(0) + s\boldsymbol{w})$. It is the image via the exponential map of a variation of the line segment $t \mapsto t\dot{\boldsymbol{c}}(0)$ in $M_{\boldsymbol{p}}$ by rays from the origin. The Jacobi field $\boldsymbol{Y}$ with initial conditions $\boldsymbol{Y}(0) = 0$, $\boldsymbol{Y}'(0) = \boldsymbol{w}$ is given by

$$\boldsymbol{Y}(t) = \exp_{\boldsymbol{c}(0)*}(t\mathcal{I}_{t\dot{\boldsymbol{c}}(0)}\boldsymbol{w}). \tag{6.2.1}$$

Definition 6.2.2. If $\boldsymbol{c} : [a,b] \to M$ is a geodesic, $t_0 \in (a,b]$ is said to be a *conjugate point* of $\boldsymbol{c}$ if there exists a nontrivial Jacobi field $\boldsymbol{Y}$ along $\boldsymbol{c}$ that vanishes at a and t_0.

Conjugate points correspond to critical points of the exponential map: our next result implies that for $\boldsymbol{p} \in M$ and $\boldsymbol{u} \in M_{\boldsymbol{p}}$, $\boldsymbol{u}$ is a critical point of $\exp_{\boldsymbol{p}}$ if and only if 1 is a conjugate point of the geodesic $t \mapsto \exp_{\boldsymbol{p}}(t\boldsymbol{u})$.

Theorem 6.2.1. *Let $\boldsymbol{p} \in M$, $\boldsymbol{u} \in M_{\boldsymbol{p}}$, $\boldsymbol{c} : [0,a] \to M$ the geodesic $t \mapsto \exp_{\boldsymbol{p}}(t\boldsymbol{u})$, and $t_0 \in (0,a]$. The vector space $\mathcal{J}_{\boldsymbol{c}}^{t_0}$ of all Jacobi fields $\boldsymbol{Y}$ along $\boldsymbol{c}$ that vanish at 0 and t_0 has the same dimension as the kernel of the derivative of $\exp_{\boldsymbol{p}}$ at $t_0\boldsymbol{u}$.*

Proof. Any Jacobi field $\boldsymbol{Y}$ along $\boldsymbol{c}$ with $\boldsymbol{Y}(0) = \boldsymbol{0}$ has the form

$$\boldsymbol{Y}(t) = \exp_{\boldsymbol{p}*(t\boldsymbol{u})}(t\mathcal{I}_{t\boldsymbol{u}}\boldsymbol{Y}'(0))$$

by (6.2.1). Thus, if $\boldsymbol{Y} \in \mathcal{J}_{\boldsymbol{c}}^{t_0}$, then

$$\boldsymbol{Y}(t_0) = \exp_{\boldsymbol{p}*}(t_0\mathcal{I}_{t_0\boldsymbol{u}}\boldsymbol{Y}'(0)) = \boldsymbol{0},$$

and $\mathcal{I}_{t_0\boldsymbol{u}}\boldsymbol{Y}'(0) \in \ker \exp_{\boldsymbol{p}*(t_0\boldsymbol{u})}$. The map

$$\begin{aligned} L : \mathcal{J}_{\boldsymbol{c}}^{t_0} &\longrightarrow \ker \exp_{\boldsymbol{p}*(t_0\boldsymbol{u})}, \\ \boldsymbol{Y} &\longmapsto \mathcal{I}_{t_0\boldsymbol{u}}\boldsymbol{Y}'(0) \end{aligned}$$

is linear and has trivial kernel since any Jacobi field $\boldsymbol{Y}$ with $\boldsymbol{Y}(0) = \boldsymbol{Y}'(0) = \boldsymbol{0}$ is trivial. It is also onto, for if $\boldsymbol{v}$ lies in the kernel of the derivative of $\exp_{\boldsymbol{p}}$ at $t_0\boldsymbol{u}$, then the Jacobi field $\boldsymbol{Y}$ along $\boldsymbol{c}$ with $\boldsymbol{Y}(0) = \boldsymbol{0}$ and $\boldsymbol{Y}'(0) = \mathcal{I}_{t_0\boldsymbol{u}}^{-1}\boldsymbol{v}$ vanishes at t_0 by (6.2.1), so that $\boldsymbol{Y} \in J_{\boldsymbol{c}}^{t_0}$ and $L\boldsymbol{Y} = \boldsymbol{v}$. L is therefore an isomorphism, and the theorem follows. □

Example 6.2.2. If M^n is a space of constant curvature κ and $\boldsymbol{c} : [0,\infty) \to M$ is a normal geodesic in M, then by Example 6.2.1, any Jacobi field $\boldsymbol{Y}$ along $\boldsymbol{c}$ with $\boldsymbol{Y}(0) = \boldsymbol{0}$ can be written as $\boldsymbol{Y} = s_\kappa \boldsymbol{E}$, where $\boldsymbol{E}$ is a parallel field and $s_\kappa(t) = t$ if $\kappa = 0$, $\sin\sqrt{\kappa}t$ if $\kappa > 0$, and $\sinh\sqrt{-\kappa}t$ if $\kappa < 0$. Normal geodesics therefore have no conjugate points if $\kappa \leq 0$, and have $k\pi/\sqrt{\kappa}$, $k \in \mathbb{N}$, as conjugate points when $\kappa > 0$. In the latter case, the images via $\boldsymbol{c}$ of the critical points alternate between $\boldsymbol{c}(0)$ and its antipode.

More generally, we have the following:

Theorem 6.2.2. *In a manifold M with nonpositive curvature, geodesics have no conjugate points.*

Proof. Let $\boldsymbol{Y}$ denote a Jacobi field along a geodesic in M. We will show that the norm of $\boldsymbol{Y}$ is a convex function. Since a convex nonnegative function that vanishes at two points vanishes everywhere, the statement will follow.
Specifically, we claim that if $\boldsymbol{Y}(t) \neq \boldsymbol{0}$, then $|\boldsymbol{Y}|''(t) \geq 0$. To see this, observe that $|\boldsymbol{Y}|' = \langle \boldsymbol{Y}, \boldsymbol{Y} \rangle^{1/2\prime} = \langle \boldsymbol{Y}, \boldsymbol{Y}' \rangle / |\boldsymbol{Y}|$. If $K(t)$ denotes the curvature of the plane spanned by the tangent vector of the geodesic at t and by $\boldsymbol{Y}(t)$, then $K \leq 0$ so that

$$\begin{aligned} |\boldsymbol{Y}|'' &= \frac{|\boldsymbol{Y}|(|\boldsymbol{Y}'|^2 + \langle \boldsymbol{Y}, \boldsymbol{Y}'' \rangle) - \langle \boldsymbol{Y}, \boldsymbol{Y}' \rangle^2 / |\boldsymbol{Y}|}{|\boldsymbol{Y}|^2} \\ &= \frac{1}{|\boldsymbol{Y}|^3}(|\boldsymbol{Y}|^2 |\boldsymbol{Y}'|^2 - K|\boldsymbol{Y}|^4 - \langle \boldsymbol{Y}, \boldsymbol{Y}' \rangle^2) \\ &\geq \frac{1}{|\boldsymbol{Y}|^3}(|\boldsymbol{Y}|^2 |\boldsymbol{Y}'|^2 - \langle \boldsymbol{Y}, \boldsymbol{Y}' \rangle^2), \end{aligned}$$

and the last expression is nonnegative by the Cauchy-Schwartz inequality. □

6.3 The length function of a variation

Let $\boldsymbol{c} : [0, a] \to M$ be a normal geodesic, and $V : [0, a] \times I \to M$ a variation of $\boldsymbol{c}$, where I is an open interval containing 0. In this section, we discuss the length function $L : I \to \mathbb{R}$ of the variation, with $L(s)$ denoting the length of the curve V_s, $V_s(t) = V(t, s)$. It turns out that when the variation has fixed endpoints, then $\boldsymbol{c}$ is shorter than nearby curves in the variation, provided it has no conjugate points. $\boldsymbol{Y} : [0, a] \to TM$ will denote the variational vector field of V, $\boldsymbol{Y}(t) = V_* \boldsymbol{D}_2(t, 0)$, and $\boldsymbol{Y}_\perp = \boldsymbol{Y} - \langle \boldsymbol{Y}, \dot{\boldsymbol{c}} \rangle \dot{\boldsymbol{c}}$ its component orthogonal to the geodesic.

Lemma 6.3.1. *With notation as above,*

$$L'(0) = \langle \boldsymbol{Y}, \dot{\boldsymbol{c}} \rangle |_0^a, \tag{6.3.1}$$

$$L''(0) = \left(\int_0^a |\boldsymbol{Y}_\perp'|^2 - \langle R(\boldsymbol{Y}_\perp, \dot{\boldsymbol{c}}) \dot{\boldsymbol{c}}, \boldsymbol{Y}_\perp \rangle \right) + \langle \nabla_{\boldsymbol{D}_2} V_* \boldsymbol{D}_2(t, 0), \dot{\boldsymbol{c}}(t) \rangle |_0^a. \tag{6.3.2}$$

In particular, if V has fixed endpoints (meaning $V(0, s) = \boldsymbol{c}(0)$, $V(a, s) = \boldsymbol{c}(a)$ for all s), then

$$L'(0) = 0, \tag{6.3.3}$$

$$L''(0) = \int_0^a |\boldsymbol{Y}_\perp'|^2 - \langle R(\boldsymbol{Y}_\perp, \dot{\boldsymbol{c}}) \dot{\boldsymbol{c}}, \boldsymbol{Y}_\perp \rangle. \tag{6.3.4}$$

Proof. By definition, $L(s) = \int_0^a |V_* \boldsymbol{D}_1|(t, s)\, \mathrm{d}t$. Since $V_* \boldsymbol{D}_1$ is continuous and has constant norm 1 when $s = 0$, compactness of $[0, a]$ implies that $V_* \boldsymbol{D}_1$ is nonzero on $[0, a] \times$

$(-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$. This means that L is differentiable at 0, and by Theorem 2.3.1,

$$L'(s) = \int_0^a \boldsymbol{D}_2 |V_* \boldsymbol{D}_1|(t,s)\,\mathrm{d}t = \int_0^a \frac{\boldsymbol{D}_2 \langle V_* \boldsymbol{D}_1, V_* \boldsymbol{D}_1 \rangle}{2|V_* \boldsymbol{D}_1|}(t,s)\,\mathrm{d}t$$
$$= \int_0^a \frac{\langle \nabla_{\boldsymbol{D}_2} V_* \boldsymbol{D}_1, V_* \boldsymbol{D}_1 \rangle}{|V_* \boldsymbol{D}_1|}(t,s)\,\mathrm{d}t = \int_0^a \frac{\langle \nabla_{\boldsymbol{D}_1} V_* \boldsymbol{D}_2, V_* \boldsymbol{D}_1 \rangle}{|V_* \boldsymbol{D}_1|}(t,s)\,\mathrm{d}t.$$

Note that for the second equality above, we used the fact $|\boldsymbol{a}| = \langle \boldsymbol{a}, \boldsymbol{a} \rangle^{1/2}$ and the chain rule. When $s = 0$, $|V_* \boldsymbol{D}_1| \equiv 1$, and

$$\langle \nabla_{\boldsymbol{D}_1} V_* \boldsymbol{D}_2, V_* \boldsymbol{D}_1 \rangle(t,0) = \langle \boldsymbol{Y}', \dot{\boldsymbol{c}} \rangle(t) = \langle \boldsymbol{Y}, \dot{\boldsymbol{c}} \rangle'(t),$$

so that $L'(0) = \langle \boldsymbol{Y}, \dot{\boldsymbol{c}} \rangle|_0^a$, which establishes (6.3.1). For (6.3.2), we have

$$L''(s) = \int_0^a \boldsymbol{D}_2 \left(\frac{\langle \nabla_{\boldsymbol{D}_1} V_* \boldsymbol{D}_2, V_* \boldsymbol{D}_1 \rangle}{|V_* \boldsymbol{D}_1|} \right)(t,s)\,\mathrm{d}t,$$

and the integrand equals

$$\frac{\langle \nabla_{\boldsymbol{D}_2} \nabla_{\boldsymbol{D}_1} V_* \boldsymbol{D}_2, V_* \boldsymbol{D}_1 \rangle + \langle \nabla_{\boldsymbol{D}_1} V_* \boldsymbol{D}_2, \nabla_{\boldsymbol{D}_2} V_* \boldsymbol{D}_1 \rangle}{|V_* \boldsymbol{D}_1|} - \frac{\langle \nabla_{\boldsymbol{D}_1} V_* \boldsymbol{D}_2, \boldsymbol{D}_1 \rangle^2}{|V_* \boldsymbol{D}_1|^3} =$$
$$\frac{\langle \nabla_{\boldsymbol{D}_2} \nabla_{\boldsymbol{D}_1} V_* \boldsymbol{D}_2, V_* \boldsymbol{D}_1 \rangle + |\nabla_{\boldsymbol{D}_1} V_* \boldsymbol{D}_2|^2}{|V_* \boldsymbol{D}_1|} - \frac{\langle \nabla_{\boldsymbol{D}_1} V_* \boldsymbol{D}_2, \boldsymbol{D}_1 \rangle^2}{|V_* \boldsymbol{D}_1|^3}.$$

When $s = 0$, this expression becomes

$$\langle \nabla_{\boldsymbol{D}_2} \nabla_{\boldsymbol{D}_1} V_* \boldsymbol{D}_2(t,0), \dot{\boldsymbol{c}}(t) \rangle + |\nabla_{\boldsymbol{D}_1} \boldsymbol{Y}|^2(t) - \langle \nabla_{\boldsymbol{D}_1} \boldsymbol{Y}, \dot{\boldsymbol{c}} \rangle^2(t).$$

Now,

$$|\nabla_{\boldsymbol{D}_1} \boldsymbol{Y}|^2 = |\nabla_{\boldsymbol{D}_1} (\boldsymbol{Y}_\perp + \langle \boldsymbol{Y}, \dot{\boldsymbol{c}} \rangle \dot{\boldsymbol{c}})|^2 = |\nabla_{\boldsymbol{D}_1} \boldsymbol{Y}_\perp + \boldsymbol{D}_1(\langle \boldsymbol{Y}, \dot{\boldsymbol{c}} \rangle)\dot{\boldsymbol{c}}|^2$$
$$= |\nabla_{\boldsymbol{D}_1} \boldsymbol{Y}_\perp|^2 + (\boldsymbol{D}_1 \langle \boldsymbol{Y}, \dot{\boldsymbol{c}} \rangle)^2 \text{ (since } \langle \nabla_{\boldsymbol{D}_1} \boldsymbol{Y}_\perp, \dot{\boldsymbol{c}} \rangle = \boldsymbol{D}_1 \langle \boldsymbol{Y}_\perp, \dot{\boldsymbol{c}} \rangle = 0)$$
$$= |\nabla_{\boldsymbol{D}_1} \boldsymbol{Y}_\perp|^2 + \langle \nabla_{\boldsymbol{D}_1} \boldsymbol{Y}, \dot{\boldsymbol{c}} \rangle^2,$$

so that

$$L''(0) = \int_0^a \left(\langle \nabla_{\boldsymbol{D}_2} \nabla_{\boldsymbol{D}_1} V_* \boldsymbol{D}_2, \dot{\boldsymbol{c}} \rangle + |\boldsymbol{Y}_\perp'|^2 \right).$$

Furthermore,

$$\nabla_{\boldsymbol{D}_2} \nabla_{\boldsymbol{D}_1} V_* \boldsymbol{D}_2(t,0) = \nabla_{\boldsymbol{D}_1} \nabla_{\boldsymbol{D}_2} V_* \boldsymbol{D}_2(t,0) - (R(\dot{\boldsymbol{c}}, \boldsymbol{Y})\boldsymbol{Y})(t),$$

and since $\langle \nabla_{\boldsymbol{D}_1} \nabla_{\boldsymbol{D}_2} V_* \boldsymbol{D}_2(t,0), \dot{\boldsymbol{c}}(t) \rangle = \boldsymbol{D}_1 \langle \nabla_{\boldsymbol{D}_2} V_* \boldsymbol{D}_2(t,0), \dot{\boldsymbol{c}}(t) \rangle$, we obtain

$$L''(0) = \left(\int_0^a |\boldsymbol{Y}_\perp'|^2 - \langle R(\dot{\boldsymbol{c}}, \boldsymbol{Y})\boldsymbol{Y}, \dot{\boldsymbol{c}} \rangle \right) + \langle \nabla_{\boldsymbol{D}_2} V_* \boldsymbol{D}_2(t,0), \dot{\boldsymbol{c}}(t) \rangle|_0^a.$$

The symmetries of the curvature tensor and the fact that $\langle R(\dot{\boldsymbol{c}}, \boldsymbol{Y})\boldsymbol{Y}, \dot{\boldsymbol{c}}\rangle = \langle R(\dot{\boldsymbol{c}}, \boldsymbol{Y}_\perp)\boldsymbol{Y}_\perp, \dot{\boldsymbol{c}}\rangle$ now yield (6.3.2).
If V has fixed endpoints, then $V_*\boldsymbol{D}_2(0,s)$, $V_*\boldsymbol{D}_2(a,s)$, $\boldsymbol{Y}(0)$ and $\boldsymbol{Y}(a)$ all vanish, thereby establishing (6.3.3) and (6.3.4). □

We will need a more general version of the lemma. Let I be an open interval containing 0. A *piecewise smooth variation* of a smooth curve $\boldsymbol{c} : [0,a] \to M$ is a continuous map $V : [0,a] \times I \to M$ for which there exists a partition $0 = t_0 < t_1 < \cdots < t_l = a$ of $[0,a]$ such that each $V_i := V_{|[t_{i-1},t_i]\times I}$ is a variation of $\boldsymbol{c}_{|[t_{i-1},t_i]}$, $i = 1,\ldots,l$. A *piecewise smooth vector field* along $\boldsymbol{c}$ is a piecewise smooth curve $\boldsymbol{Y} : [0,a] \to TM$ such that $\boldsymbol{Y}(t) \in M_{\boldsymbol{c}(t)}$ for all t. If V is a piecewise smooth variation of $\boldsymbol{c}$, then by assumption, the map $V_*\boldsymbol{D}_2 : [0,a]\times I \to TM$, where $V_*\boldsymbol{D}_{2|[t_{i-1},t_i]\times I} = V_{i*}\boldsymbol{D}_2$, is continuous. It then induces a piecewise smooth vector field $\boldsymbol{Y}$ along $\boldsymbol{c}$, with $\boldsymbol{Y}(t) = V_*\boldsymbol{D}_2(t,0)$. Even though $\boldsymbol{Y}$ is not, in general, differentiable at t_i, $\boldsymbol{Y}'$ may be extended to a not necessarily continuous vector field on $[0,a]$ by defining $\boldsymbol{Y}'(t_i) = \lim_{t\to t_i^+} \boldsymbol{Y}_i'(t)$. Notice that even though $\nabla_{\boldsymbol{D}_1} V_*\boldsymbol{D}_2(t_i,s)$ may not exist, $\nabla_{\boldsymbol{D}_2} V_*\boldsymbol{D}_2(t_i,s)$ does. As before, set $\boldsymbol{Y}_\perp = \boldsymbol{Y} - \langle \boldsymbol{Y}, \dot{\boldsymbol{c}}\rangle\dot{\boldsymbol{c}}$, $L(s) = L(V_s)$, where $V_s(t) = V(t,s)$.

Proposition 6.3.1. *Let $\boldsymbol{c} : [0,a] \to M$ be a normal geodesic, and $V : [0,a] \times I \to M$ a piecewise smooth variation of $\boldsymbol{c}$. Then equations* (6.3.1) *through* (6.3.4) *from Lemma 6.3.1 hold.*

Proof. With notation as above, if $L_i : I \to \mathbb{R}$ denotes the length function of V_i, $i = 1,\ldots,l$, then $L = \sum_i L_i$. The lemma then implies that

$$L'(0) = \sum_{i=1}^{l} L_i'(0) = \sum_{i=1}^{l} \langle \boldsymbol{Y}, \dot{\boldsymbol{c}}\rangle|_{t_{i-1}}^{t_i} = \langle \boldsymbol{Y}, \dot{\boldsymbol{c}}\rangle|_0^a,$$

and

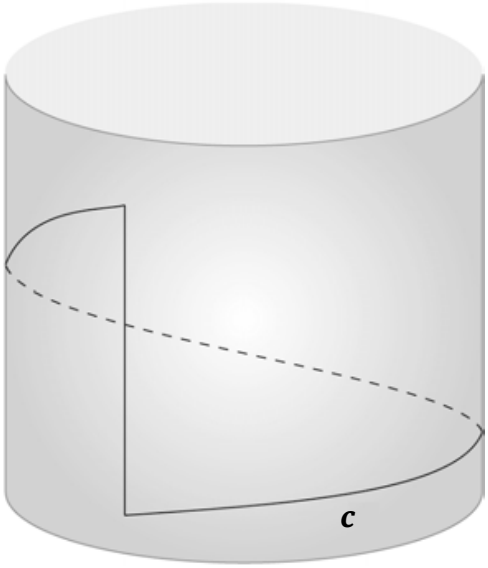
$$\begin{aligned} L''(0) &= \sum_{i=1}^{l} L_i''(0) \\ &= \sum_{i=1}^{l} \left[\int_{t_{i-1}}^{t_i} |\boldsymbol{Y}_\perp'|^2 - \langle R(\boldsymbol{Y}_\perp, \dot{\boldsymbol{c}})\dot{\boldsymbol{c}}, \boldsymbol{Y}_\perp\rangle + \langle \nabla_{\boldsymbol{D}_2} V_*\boldsymbol{D}_2(t,0), \dot{\boldsymbol{c}}(t)\rangle|_{t_{i-1}}^{t_i} \right] \\ &= \left(\int_0^a |\boldsymbol{Y}_\perp'|^2 - \langle R(\boldsymbol{Y}_\perp, \dot{\boldsymbol{c}})\dot{\boldsymbol{c}}, \boldsymbol{Y}_\perp\rangle \right) + \langle \nabla_{\boldsymbol{D}_2} V_*\boldsymbol{D}_2(t,0), \dot{\boldsymbol{c}}(t)\rangle|_0^a. \end{aligned}$$

The last two identities in the statement follow just as in the Lemma. □

Proposition 6.3.2. *Let $\boldsymbol{c} : [0,a] \to M$ be a normal geodesic without conjugate points. If $V : [0,a] \times I \to M$ is a variation of $\boldsymbol{c}$ with fixed endpoints, then for sufficiently small s, $L(V_s) \geq L(V_0) = L(\boldsymbol{c})$, and strict inequality holds provided V_s is not a reparametrization of $\boldsymbol{c}$.*

Proof. Let $\boldsymbol{\gamma} : \mathbb{R} \to M_{\boldsymbol{p}}$ denote the ray $t \mapsto t\dot{\boldsymbol{c}}(0)$ in the tangent space at $\boldsymbol{p} = \boldsymbol{c}(0)$, so that $\boldsymbol{c} = \exp \circ \gamma_{|[0,a]}$. Since $\boldsymbol{c}$ has no conjugate points, $\exp_{\boldsymbol{p}}$ has maximal rank at each $\boldsymbol{\gamma}(t)$ and is therefore a diffeomorphism in a neighborhood of $\boldsymbol{\gamma}(t)$. By compactness of $[a, b]$, there exists a partition $0 = t_0 < t_1 < \cdots < t_l = a$ of the interval and open sets $U_1, \ldots, U_l$ in $M_{\boldsymbol{p}}$ such that $\boldsymbol{\gamma}[t_{i-1}, t_i] \subset U_i$ and $\exp_{\boldsymbol{p}} : U_i \to \exp(U_i)$ is a diffeomorphism for each i from 1 to l. By Lemma 1.7.1, there exists for each i some $\varepsilon_i > 0$ such that $V([t_{i-1}, t_i] \times (-\varepsilon_i, \varepsilon_i)) \subset \exp(U_i)$. Let $\varepsilon = \min\{\varepsilon_1, \ldots, \varepsilon_l\} > 0$, and define $\tilde{V} : [0, a] \times (-\varepsilon, \varepsilon) \to M_{\boldsymbol{p}}$ by $\tilde{V}(t, s) = (\exp_{\boldsymbol{p}|U_i})^{-1}(V(t, s))$ if $t \in [t_{i-1}, t_i]$. Each $\tilde{V}_s$, where $\tilde{V}(s) = \tilde{V}(t, s)$, is then a curve in $M_{\boldsymbol{p}}$ from $\mathbf{0}$ to $\dot{\boldsymbol{c}}(0)$. Since $V_s = \exp \circ \tilde{V}_s$, the claim follows from Lemma 6.1.3. □

It is in general not true, though, that a geodesic without conjugate points is minimal. The cylinder $M = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\}$ provides many examples: If $a, b > 0$ and $a^2 + b^2 = 1$, then by Example 3.11.1, the curve $\boldsymbol{c} : [0, 2\pi/a] \to M$, $\boldsymbol{c}(t) = (\cos(at), \sin(at), bt)$, is a normal geodesic joining $(1, 0, 0)$ to $(1, 0, 2\pi b/a)$, and since M is flat, $\boldsymbol{c}$ has no conjugate points by Theorem 6.2.2. The meridian $t \mapsto (1, 0, t)$ also joins the endpoints of $\boldsymbol{c}$ and has length $2\pi b/a < 2\pi/a$. The proposition only guarantees that $\boldsymbol{c}$ will be shorter than sufficiently nearby curves in a variation with fixed endpoints.



The geodesic $\boldsymbol{c}$ has no conjugate points, but is not minimal

6.4 The index form of a geodesic

Let $\boldsymbol{c} : [0, a] \to M$ denote a normal geodesic. In this section, we introduce a symmetric bilinear form on the space $\mathfrak{V}_{\boldsymbol{c}}$ of piecewise smooth vector fields along $\boldsymbol{c}$ that are orthogonal to $\boldsymbol{c}$ and vanish at its endpoints. This form is closely related to the existence of conjugate points of $\boldsymbol{c}$. As a first application, we will show that a geodesic does not minimize length past its first conjugate point; specifically, if $\boldsymbol{c} : [0, a] \to M$ has a conjugate point in $(0, a)$, then there exist curves in M joining $\boldsymbol{c}(0)$ to $\boldsymbol{c}(a)$ that are shorter than $\boldsymbol{c}$ and arbitrarily close to it.

Proposition 6.4.1. *Let $\boldsymbol{c} : [0,a] \to M$ be a normal geodesic, and $\mathfrak{V}_{\boldsymbol{c}}$ the (infinite-dimensional) vector space of all piecewise smooth vector fields $\boldsymbol{Y}$ along $\boldsymbol{c}$ such that $\langle \boldsymbol{Y}, \dot{\boldsymbol{c}} \rangle = 0$ and $\boldsymbol{Y}(0) = \boldsymbol{0}$, $\boldsymbol{Y}(a) = \boldsymbol{0}$. Then there exists a unique symmetric bilinear form I on $\mathfrak{V}_{\boldsymbol{c}}$ such that if V is a variation of $\boldsymbol{c}$ with fixed endpoints that has $\boldsymbol{Y} \in \mathfrak{V}_{\boldsymbol{c}}$ as variational vector field, then*

$$I(\boldsymbol{Y}, \boldsymbol{Y}) = L''(0),$$

where L denotes the length function of V. I is called the index form *of $\boldsymbol{c}$.*

Proof. Uniqueness follows from the fact that I is determined by its restriction to the diagonal $\Delta = \{(\boldsymbol{Y}, \boldsymbol{Y}) \mid \boldsymbol{Y} \in \mathfrak{V}_{\boldsymbol{c}}\}$ in $\mathfrak{V}_{\boldsymbol{c}} \times \mathfrak{V}_{\boldsymbol{c}}$, because $I(\boldsymbol{Y}, \boldsymbol{Z}) = 1/2(I(\boldsymbol{Y} + \boldsymbol{Z}, \boldsymbol{Y} + \boldsymbol{Z}) - I(\boldsymbol{Y}, \boldsymbol{Y}) - I(\boldsymbol{Z}, \boldsymbol{Z}))$ for any $\boldsymbol{Y}, \boldsymbol{Z} \in \mathfrak{V}_{\boldsymbol{c}}$. Existence follows from (6.3.4), which implies that

$$I(\boldsymbol{Y}, \boldsymbol{Z}) = \int_0^a \Big(\langle \boldsymbol{Y}', \boldsymbol{Z}' \rangle - \langle R(\boldsymbol{Y}, \dot{\boldsymbol{c}})\dot{\boldsymbol{c}}, \boldsymbol{Z} \rangle \Big). \tag{6.4.1}$$

This formula also shows that I is indeed symmetric and bilinear. □

Notice that any piecewise smooth vector field $\boldsymbol{Y}$ along $\boldsymbol{c}$ is the variational vector field of some piecewise smooth variation V of $\boldsymbol{c}$, namely

$$V(t,s) = \exp(s\boldsymbol{Y}(t)), \tag{6.4.2}$$

where of course, s is chosen small enough so that the exponential map is defined.

Proposition 6.4.2. *Let $\boldsymbol{c} : [0,a] \to M$ be a normal geodesic, $\boldsymbol{Y}, \boldsymbol{Z} \in \mathfrak{V}_{\boldsymbol{c}}$. If $\boldsymbol{Z}$ is differentiable, then*

$$I(\boldsymbol{Y}, \boldsymbol{Z}) = -\int_0^a \langle \boldsymbol{Y}, \boldsymbol{Z}'' + R(\boldsymbol{Z}, \dot{\boldsymbol{c}})\dot{\boldsymbol{c}} \rangle. \tag{6.4.3}$$

In particular, if $\boldsymbol{Z}$ is a Jacobi field, then

$$I(\boldsymbol{Y}, \boldsymbol{Z}) = 0. \tag{6.4.4}$$

Proof. By assumption, there exists a partition $0 = t_0 < t_1 < \cdots < t_l = a$ of $[0,a]$ such that $\boldsymbol{Y}$ is differentiable on $[t_{i-1}, t_i]$, $1 \le i \le l$. Since $\langle \boldsymbol{Y}, \boldsymbol{Z}' \rangle'(t) = \langle \boldsymbol{Y}', \boldsymbol{Z}' \rangle(t) + \langle \boldsymbol{Y}, \boldsymbol{Z}'' \rangle(t)$ if $t \neq t_i$,

$$\begin{aligned}
\int_0^a \langle \boldsymbol{Y}', \boldsymbol{Z}' \rangle &= \sum_{i=1}^{l} \Big(\int_{t_{i-1}}^{t_i} \langle \boldsymbol{Y}, \boldsymbol{Z}' \rangle' \Big) - \int_0^a \langle \boldsymbol{Y}, \boldsymbol{Z}'' \rangle \\
&= \sum_{i=1}^{l} \langle \boldsymbol{Y}, \boldsymbol{Z}' \rangle|_{t_{i-1}}^{t_i} - \int_0^a \langle \boldsymbol{Y}, \boldsymbol{Z}'' \rangle = \langle \boldsymbol{Y}, \boldsymbol{Z}' \rangle|_0^b - \int_0^a \langle \boldsymbol{Y}, \boldsymbol{Z}'' \rangle \\
&= -\int_0^a \langle \boldsymbol{Y}, \boldsymbol{Z}'' \rangle,
\end{aligned}$$

because $\boldsymbol{Y}$ vanishes at the endpoints. Substituting this expression in (6.4.1) and using the identity $\langle R(\boldsymbol{Y},\dot{\boldsymbol{c}})\dot{\boldsymbol{c}},\boldsymbol{Z}\rangle = \langle R(\boldsymbol{Z},\dot{\boldsymbol{c}})\dot{\boldsymbol{c}},\boldsymbol{Y}\rangle$ now yields (6.4.3); the latter immediately implies (6.4.4). □

Remark 6.4.1. The index form extends to the larger space of all piecewise smooth vector fields along $\boldsymbol{c}$ (not necessarily zero at the endpoints) by the same formula (6.4.1). The proof of Proposition 6.4.2 shows that for vector fields $\boldsymbol{Y}$, $\boldsymbol{Z}$ in this larger space,

$$I(\boldsymbol{Y},\boldsymbol{Z}) = \langle \boldsymbol{Y},\boldsymbol{Z}'\rangle|_0^a - \int_0^a \langle \boldsymbol{Y},\boldsymbol{Z}'' + R(\boldsymbol{Z},\dot{\boldsymbol{c}})\dot{\boldsymbol{c}}\rangle, \tag{6.4.5}$$

if $\boldsymbol{Z}$ is differentiable, and

$$I(\boldsymbol{Y},\boldsymbol{Z}) = \langle \boldsymbol{Y},\boldsymbol{Z}'\rangle|_0^a \tag{6.4.6}$$

if in addition $\boldsymbol{Z}$ is Jacobi. We will refer to it as the *extended index form* of $\boldsymbol{c}$.

Theorem 6.4.1. *If $\boldsymbol{c} : [0,a] \to M$ is a normal geodesic with a conjugate point $t_0 < a$, then there exists a variation V of $\boldsymbol{c}$ with fixed endpoints such that the curves V_s, where $V_s(t) = V(t,s)$, are shorter than $\boldsymbol{c}$ for sufficiently small s. In particular, $\boldsymbol{c}$ is not a minimal geodesic.*

Proof. By assumption, there exists a Jacobi field $\boldsymbol{Z}$ along $\boldsymbol{c}$ with $\boldsymbol{Z}(0) = \boldsymbol{0}$, $\boldsymbol{Z}(t_0) = \boldsymbol{0}$. Let $\boldsymbol{E}$ denote the parallel vector field along $\boldsymbol{c}$ that equals $-\boldsymbol{Z}'(t_0)$ at t_0, and define $\boldsymbol{Y} := \varphi\boldsymbol{E} \in \mathfrak{V}_{\boldsymbol{c}}$, where $\varphi : [0,a] \to [0,1]$ is some function satisfying $\varphi(0) = \varphi(a) = 0$, $\varphi(t_0) = 1$, cf. Lemma 2.2.1. For each $r > 0$, define $\boldsymbol{Z}_r \in \mathfrak{V}_{\boldsymbol{c}}$ by

$$\boldsymbol{Z}_r(t) = \begin{cases} \boldsymbol{Z}(t) + r\boldsymbol{Y}(t) & \text{if } t \le t_0, \\ r\boldsymbol{Y}(t) & \text{if } t \ge t_0. \end{cases}$$

Let I_0 denote the extended index form (in the sense of Remark 6.4.1) of the restriction of $\boldsymbol{c}$ to $[0,t_0]$. Then

$$I(\boldsymbol{Z}_r,\boldsymbol{Z}_r) = I_0(\boldsymbol{Z}_r,\boldsymbol{Z}_r) + \int_{t_0}^a \left(\langle r\boldsymbol{Y}', r\boldsymbol{Y}'\rangle - \langle R(r\boldsymbol{Y},\dot{\boldsymbol{c}})\dot{\boldsymbol{c}}, r\boldsymbol{Y}\rangle\right),$$

and recalling that $\boldsymbol{Y}(t_0) = -\boldsymbol{Z}'(t_0)$,

$$\begin{aligned} I_0(\boldsymbol{Z}_r,\boldsymbol{Z}_r) &= I_0(\boldsymbol{Z},\boldsymbol{Z}) + 2I_0(r\boldsymbol{Y},\boldsymbol{Z}) + I_0(r\boldsymbol{Y},r\boldsymbol{Y}) \\ &= \langle \boldsymbol{Z},\boldsymbol{Z}'\rangle|_0^{t_0} + 2r\langle \boldsymbol{Y},\boldsymbol{Z}'\rangle|_0^{t_0} + r^2 I_0(\boldsymbol{Y},\boldsymbol{Y}) \\ &= -2r|\boldsymbol{Z}'|^2(t_0) + r^2 I_0(\boldsymbol{Y},\boldsymbol{Y}), \end{aligned}$$

so that

$$I(\boldsymbol{Z}_r,\boldsymbol{Z}_r) = -2r|\boldsymbol{Z}'|^2(t_0) + r^2 I(\boldsymbol{Y},\boldsymbol{Y}) = r(-2|\boldsymbol{Z}'|^2(t_0) + rI(\boldsymbol{Y},\boldsymbol{Y})).$$

$\boldsymbol{Z}'(t_0) \neq 0$ since $\boldsymbol{Z}$ is not identically zero, so the above expression is negative for sufficiently small $r > 0$. By Proposition 6.4.1 and (6.4.2), the length function L of the

corresponding variation $(t,s) \mapsto \exp(s\boldsymbol{Z}_r(t))$ satisfies $L''(0) < 0$, and the length of $\boldsymbol{c} = V_0$ is then a strict local maximum. □

The *kernel* of a bilinear form b on a vector space E is defined to be the set

$$\ker b = \{\boldsymbol{u} \in E \mid b(\boldsymbol{u},\boldsymbol{v}) = 0 \text{ for all } \boldsymbol{v} \in E\}.$$

It is clearly a subspace of E. Even though the vector space $\mathfrak{V}_{\boldsymbol{c}}$ is infinite-dimensional, the kernel of the index form has finite dimension:

Lemma 6.4.1. *If $\boldsymbol{c} : [0,a] \to M$ is a normal geodesic, then the kernel of the index form I of $\boldsymbol{c}$ consists of all Jacobi fields $\boldsymbol{Y} \in \mathfrak{V}_{\boldsymbol{c}}$.*

Proof. By (6.4.4), every Jacobi field $\boldsymbol{Y} \in \mathfrak{V}_{\boldsymbol{c}}$ belongs to the kernel of I. Conversely, suppose that $\boldsymbol{Z}$ lies in the kernel, and consider a partition $0 = t_0 < t_1 < \cdots < t_l = a$ of $[0,a]$ such that each $\boldsymbol{Z}_i := \boldsymbol{Z}_{|[t_{i-1},t_i]}$ is differentiable. If $\varphi_i : [t_{i-1},t_i] \to \mathbb{R}$ is a smooth function that vanishes at the endpoints and is positive elsewhere, $i = 1,\dots,l$, define a vector field $\boldsymbol{Y}_i = \varphi_i(\boldsymbol{Z}_i'' + R(\boldsymbol{Z}_i,\dot{\boldsymbol{c}})\dot{\boldsymbol{c}})$ along $\boldsymbol{c}_i := \boldsymbol{c}_{|[t_{i-1},t_i]}$, and $\boldsymbol{Y} \in \mathfrak{V}_{\boldsymbol{c}}$ by $\boldsymbol{Y}_{|[t_{i-1},t_i]} = \boldsymbol{Y}_i$. Let I_i denote the extended index form of $\boldsymbol{c}_i$. Since each $\boldsymbol{Z}_i$ is differentiable, (6.4.3) implies that

$$\begin{aligned} 0 = I(\boldsymbol{Y},\boldsymbol{Z}) &= \sum_{i=1}^{l} I_i(\boldsymbol{Y}_i,\boldsymbol{Z}_i) = -\sum_{i=1}^{l}\int_{t_{i-1}}^{t_i} \langle \boldsymbol{Y}_i, \boldsymbol{Z}_i'' + R(\boldsymbol{Z}_i,\dot{\boldsymbol{c}})\dot{\boldsymbol{c}}\rangle \\ &= -\sum_{i=1}^{l}\int_{t_{i-1}}^{t_i} \varphi_i |\boldsymbol{Z}_i'' + R(\boldsymbol{Z}_i,\dot{\boldsymbol{c}})\dot{\boldsymbol{c}}|^2, \end{aligned}$$

so that each $\boldsymbol{Z}_i$ is Jacobi. To see that $\boldsymbol{Z}$ itself is a Jacobi field, it suffices to show that it is differentiable at t_i, $i = 1,\dots,l-1$. This in turn will follow once we establish that $\boldsymbol{Z}_i'(t_i) = \boldsymbol{Z}_{i+1}'(t_i)$, since a Jacobi field is uniquely determined by its value and the value of its derivative at any one point. To do so, fix some i, denote by $\boldsymbol{E}$ the parallel field along $\boldsymbol{c}$ with $\boldsymbol{E}(t_i) = \boldsymbol{Z}_{i+1}'(t_i) - \boldsymbol{Z}_i'(t_i)$, and let $\varphi : [0,b] \to \mathbb{R}$ be a smooth nonnegative function that equals 1 at t_i and has its support inside (t_{i-1},t_{i+1}). If $\boldsymbol{X} := \varphi\boldsymbol{E} \in \mathfrak{V}_{\boldsymbol{c}}$, then

$$\begin{aligned} 0 = I(\boldsymbol{X},\boldsymbol{Z}) &= I_i(\boldsymbol{X}_{|[t_{i-1},t_i]},\boldsymbol{Z}_i) + I_{i+1}(\boldsymbol{X}_{|[t_i,t_{i+1}]},\boldsymbol{Z}_{i+1}) \\ &= \langle \boldsymbol{Z}_{i+1}'(t_i) - \boldsymbol{Z}_i'(t_i), \boldsymbol{Z}_i'(t_i)\rangle - \langle \boldsymbol{Z}_{i+1}'(t_i) - \boldsymbol{Z}_i'(t_i), \boldsymbol{Z}_{i+1}'(t_i)\rangle \\ &= -|\boldsymbol{Z}_{i+1}'(t_i) - \boldsymbol{Z}_i'(t_i)|^2, \end{aligned}$$

so that $\boldsymbol{Z}$ is indeed Jacobi. □

We are now able to characterize geodesics without conjugate points in terms of the index form:

Theorem 6.4.2. *A normal geodesic $\boldsymbol{c}$ has no conjugate points if and only if its index form is positive definite on $\mathfrak{V}_{\boldsymbol{c}}$.*

Proof. Notice that if $\boldsymbol{c}$ has no conjugate points, then $I(\boldsymbol{X},\boldsymbol{X}) \geq 0$ for any $\boldsymbol{X} \in \mathfrak{V}_{\boldsymbol{c}}$: indeed, if $I(\boldsymbol{X},\boldsymbol{X}) < 0$, then the curves V_s in the variation $(t,s) \mapsto V(t,s) := \exp(s\boldsymbol{X}(t))$ would be shorter than $\boldsymbol{c}$ for small s, contradicting Proposition 6.3.2. Next, suppose $I(\boldsymbol{X},\boldsymbol{X}) = 0$. We must show that $\boldsymbol{X}$ is identically zero. But for any $\boldsymbol{Y} \in \mathfrak{V}_{\boldsymbol{c}}$,

$$0 \leq I(\boldsymbol{X} + t\boldsymbol{Y}, \boldsymbol{X} + t\boldsymbol{Y}) = t(tI(\boldsymbol{Y},\boldsymbol{Y}) + 2I(\boldsymbol{X},\boldsymbol{Y})),$$

which is only possible if $I(\boldsymbol{X},\boldsymbol{Y}) = 0$. Thus, $\boldsymbol{X}$ is Jacobi by the Lemma, and since $\boldsymbol{c}$ has no conjugate points, $\boldsymbol{X} \equiv \boldsymbol{0}$.
Conversely, assume I is positive definite on $\mathfrak{V}_{\boldsymbol{c}}$. $\boldsymbol{c}$ cannot have a conjugate point $t_0 < b$ by Theorem 6.4.1. Nor can it have b as conjugate point, for otherwise there would exist a nonzero Jacobi field $\boldsymbol{Y} \in \mathfrak{V}_{\boldsymbol{c}}$, and $I(\boldsymbol{Y},\boldsymbol{Y}) = 0$ by the Lemma. □

Remarks 6.4.2. (i) Given $\boldsymbol{p},\boldsymbol{q} \in M$, it is tempting to view the collection $\Omega_{\boldsymbol{p},\boldsymbol{q}}$ of all piecewise smooth curves from $\boldsymbol{p}$ to $\boldsymbol{q}$ as a manifold. We have not defined infinite-dimensional manifolds, and are therefore not in a position to formalize this, but the analogy is nevertheless suggestive: a 'point' in this manifold is a curve $\boldsymbol{c}$ from $\boldsymbol{p}$ to $\boldsymbol{q}$, a 'curve' through the point is a variation of $\boldsymbol{c}$, and its 'tangent vector' at the point is the variational vector field. Thus, the tangent space of the point is $\mathfrak{V}_{\boldsymbol{c}}$. The length function $L : \Omega_{\boldsymbol{p},\boldsymbol{q}} \to \mathbb{R}$ assigns to each point $\boldsymbol{c}$ the length of $\boldsymbol{c}$, and (6.3.3) says that geodesics are 'critical points' of L. The index form then corresponds to the 'Hessian' of L.

(ii) Recall the extended index form of a geodesic $\boldsymbol{c}$ introduced in Remark 6.4.1. In the space of all piecewise smooth vector fields along $\boldsymbol{c}$, Jacobi fields minimize this extended index form in the following sense: suppose $\boldsymbol{c} : [0,a] \to M$ has no conjugate points. Notice that for any $\boldsymbol{u} \in M_{\boldsymbol{c}(0)}$ and $\boldsymbol{v} \in M_{\boldsymbol{c}(a)}$ there exists a unique Jacobi field $\boldsymbol{Y}$ along $\boldsymbol{c}$ with $\boldsymbol{Y}(0) = \boldsymbol{u}$ and $\boldsymbol{Y}(a) = \boldsymbol{v}$, because the linear map

$$\begin{aligned} \mathcal{J}_{\boldsymbol{c}} &\longrightarrow M_{\boldsymbol{c}(0)} \times M_{\boldsymbol{c}(a)}, \\ \boldsymbol{Y} &\longmapsto (\boldsymbol{Y}(0), \boldsymbol{Y}(a)) \end{aligned}$$

has trivial kernel and is then an isomorphism (since both spaces have the same dimension). We claim that if $\boldsymbol{X}$ is a vector field along $\boldsymbol{c}$ with $\boldsymbol{X}(0) = \boldsymbol{Y}(0)$ and $\boldsymbol{X}(a) = \boldsymbol{Y}(a)$, then

$$I(\boldsymbol{X},\boldsymbol{X}) \geq I(\boldsymbol{Y},\boldsymbol{Y}),$$

and inequality is strict provided $\boldsymbol{X} \neq \boldsymbol{Y}$. In fact, if $\boldsymbol{X} - \boldsymbol{Y} \neq \boldsymbol{0}$, then it is a nontrivial element of $\mathfrak{V}_{\boldsymbol{c}}$, and by Theorem 6.4.2,

$$0 < I(\boldsymbol{X} - \boldsymbol{Y}, \boldsymbol{X} - \boldsymbol{Y}) = I(\boldsymbol{X},\boldsymbol{X}) + I(\boldsymbol{Y},\boldsymbol{Y}) - 2I(\boldsymbol{X},\boldsymbol{Y}).$$

The claim follows from this inequality once we observe that $I(\boldsymbol{X},\boldsymbol{Y}) = I(\boldsymbol{Y},\boldsymbol{Y})$. The latter identity, in turn, holds because by (6.4.6),

$$I(\boldsymbol{X},\boldsymbol{Y}) = \langle \boldsymbol{X}, \boldsymbol{Y}' \rangle|_0^a = \langle \boldsymbol{Y}, \boldsymbol{Y}' \rangle|_0^a = I(\boldsymbol{Y},\boldsymbol{Y}).$$

6.5 The distance function

A manifold M is said to be *connected* if any two points of M can be joined by a curve lying in the manifold. Notice that regardless of whether or not it is connected, a manifold is always locally connected, by which we mean that any point has a connected neighborhood, namely the domain of an appropriate chart about the point.

Unless specified otherwise, all spaces will be assumed to be connected. Given $\boldsymbol{p}$, $\boldsymbol{q} \in M$, let $\Omega_{\boldsymbol{p},\boldsymbol{q}}$ denote the collection of all curves $\boldsymbol{c} : [0,1] \to M$ with $\boldsymbol{c}(0) = \boldsymbol{p}$ and $\boldsymbol{c}(1) = \boldsymbol{q}$. As usual, $L(\boldsymbol{c})$ is the length of $\boldsymbol{c}$.

Definition 6.5.1. The *distance* between $\boldsymbol{p}$ and $\boldsymbol{q}$ is the number

$$\mathrm{d}(\boldsymbol{p},\boldsymbol{q}) = \inf\{L(\boldsymbol{c}) \mid \boldsymbol{c} \in \Omega_{\boldsymbol{p},\boldsymbol{q}}\}.$$

The distance is well defined, since it is the infimum of a nonempty set (by connectedness of M) bounded below (by zero). It is not true, in general, that the infimum is a minimum; i.e., there need not exist a curve from $\boldsymbol{p}$ to $\boldsymbol{q}$ whose length equals the distance between them (consider for example $M = \mathbb{R}^n \setminus \{\mathbf{0}\}$, any nonzero $\boldsymbol{p}$, and $\boldsymbol{q} = -\boldsymbol{p}$). If such a curve exists, however, then the (length of the) restriction of $\boldsymbol{c}$ to any subinterval $[t_1, t_2] \subset [0,1]$ also realizes the distance between the corresponding endpoints, so that by Theorem 6.1.1, $\boldsymbol{c}$ is a geodesic.

Theorem 6.5.1. (M, d) *is a metric space. Furthermore, the open sets in the metric space coincide with the usual open sets; i.e.,* $U \subset M$ *is open in* (M, d) *if and only if* $U = V \cap M$ *for some open set* V *in Euclidean space.*

Proof. The first axiom for a metric space, $\mathrm{d}(\boldsymbol{p},\boldsymbol{q}) = \mathrm{d}(\boldsymbol{q},\boldsymbol{p})$ for all $\boldsymbol{p}$ and $\boldsymbol{q}$ is clear, since $\boldsymbol{c}$ is a curve from $\boldsymbol{p}$ to $\boldsymbol{q}$ if and only if $-\boldsymbol{c}$, where $-\boldsymbol{c}(t) = \boldsymbol{c}(1-t)$, is a curve from $\boldsymbol{q}$ to $\boldsymbol{p}$, and the two have the same length. The triangle inequality $\mathrm{d}(\boldsymbol{p},\boldsymbol{r}) \le \mathrm{d}(\boldsymbol{p},\boldsymbol{q}) + \mathrm{d}(\boldsymbol{q},\boldsymbol{r})$ for any three points $\boldsymbol{p}$, $\boldsymbol{q}$, and $\boldsymbol{r}$ is also easy to see: given any $\varepsilon > 0$, there exists a curve $\boldsymbol{c}_1$ from $\boldsymbol{p}$ to $\boldsymbol{q}$ with length smaller than $\mathrm{d}(\boldsymbol{p},\boldsymbol{q}) + \varepsilon/2$, and similarly a curve $\boldsymbol{c}_2$ from $\boldsymbol{q}$ to $\boldsymbol{r}$ with length less than $\mathrm{d}(\boldsymbol{q},\boldsymbol{r}) + \varepsilon/2$. Then $\boldsymbol{c}$, where $\boldsymbol{c}(t) = \boldsymbol{c}_1(2t)$ for $0 \le t \le 1/2$ and $\boldsymbol{c}(t) = \boldsymbol{c}_2(2t-1)$ for $1/2 \le t \le 1$, is a curve from $\boldsymbol{p}$ to $\boldsymbol{r}$ of length less than $\mathrm{d}(\boldsymbol{p},\boldsymbol{q}) + \mathrm{d}(\boldsymbol{q},\boldsymbol{r}) + \varepsilon$. This implies that $\mathrm{d}(\boldsymbol{p},\boldsymbol{r}) \le \mathrm{d}(\boldsymbol{p},\boldsymbol{q}) + \mathrm{d}(\boldsymbol{q},\boldsymbol{r}) + \varepsilon$ for any $\varepsilon > 0$, and the triangle inequality follows.
The last condition that must be satisfied is $\mathrm{d}(\boldsymbol{p},\boldsymbol{q}) \ge 0$, and $\mathrm{d}(\boldsymbol{p},\boldsymbol{q}) = 0$ if and only $\boldsymbol{p} = \boldsymbol{q}$. The first part is clear, as is the fact that $\mathrm{d}(\boldsymbol{p},\boldsymbol{p}) = 0$. So assume $\mathrm{d}(\boldsymbol{p},\boldsymbol{q}) = 0$. By Theorem 6.1.1, there exists $\varepsilon > 0$ such that $\exp_{\boldsymbol{p}}$ maps the open ball $U_\varepsilon = \{\boldsymbol{v} \in M_{\boldsymbol{p}} \mid |\boldsymbol{v}| < \varepsilon\}$ of radius ε centered at the origin in the tangent space at $\boldsymbol{p}$ diffeomorphically onto $V = \exp_{\boldsymbol{p}}(U_\varepsilon)$. Furthermore, any curve originating at $\boldsymbol{p}$ that leaves V has length greater than ε, so that $\boldsymbol{q} \in V$. But then there exists a shortest curve from $\boldsymbol{p}$ to $\boldsymbol{q}$ which is a geodesic of length $|(\exp_{\boldsymbol{p}|U_\varepsilon})^{-1}\boldsymbol{q}|$. Thus, $(\exp_{\boldsymbol{p}|U_\varepsilon})^{-1}\boldsymbol{q} = \mathbf{0} \in M_{\boldsymbol{p}}$, and $\boldsymbol{q} = \boldsymbol{p}$ as claimed.
It remains to establish that $U \subset M$ is open in the usual sense if and only if it is open in the metric space (M, d). The latter, we recall, means that for any $\boldsymbol{p} \in U$ there exists

$\varepsilon > 0$ such that $B^M_\varepsilon(\boldsymbol{p}) \subset U$, where

$$B^M_\varepsilon(\boldsymbol{p}) = \{\boldsymbol{q} \in M \mid \mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) < \varepsilon\}.$$

We denote by $B_\varepsilon(\boldsymbol{p}) = \{\boldsymbol{q} \in \mathbb{R}^{n+k} \mid |\boldsymbol{q} - \boldsymbol{p}| < \varepsilon\}$ the corresponding distance ball in Euclidean space. Suppose first that $U \subset M$ is open in the usual sense. Given $\boldsymbol{p} \in U$, there exists $r > 0$ such that $B_r(\boldsymbol{p}) \cap M \subset U$. But for $\boldsymbol{q} \in M$, $\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) \geq |\boldsymbol{p} - \boldsymbol{q}|$, so that $B^M_r(\boldsymbol{p}) \subset B_r(\boldsymbol{p}) \cap M \subset U$, and U is open in (M, d). For the converse, it is enough to show that $B^M_r(\boldsymbol{p})$ is open in the usual sense for any $\boldsymbol{p} \in M$ and $r > 0$. This will in turn follow from the fact that the distance function $\mathrm{d}_{\boldsymbol{p}} : M \to [0, \infty)$ from $\boldsymbol{p}$, $\mathrm{d}_{\boldsymbol{p}}(\boldsymbol{q}) := \mathrm{d}(\boldsymbol{p}, \boldsymbol{q})$, is continuous, since $B^M_r(\boldsymbol{p}) = \mathrm{d}^{-1}_{\boldsymbol{p}}[0, r)$. To show this, we must establish that if $\{\boldsymbol{q}_i\}$ is a sequence in M converging to $\boldsymbol{q}$, then $\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}_i) \to \mathrm{d}(\boldsymbol{p}, \boldsymbol{q})$. Now,

$$\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) - \mathrm{d}(\boldsymbol{q}, \boldsymbol{q}_i) \leq \mathrm{d}(\boldsymbol{p}, \boldsymbol{q}_i) \leq \mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) + \mathrm{d}(\boldsymbol{q}, \boldsymbol{q}_i)$$

by the triangle inequality, so it is actually enough to show that $\mathrm{d}(\boldsymbol{p}, \boldsymbol{p}_i) \to 0$ if $\boldsymbol{p}_i \to \boldsymbol{p}$. If U_ε is as above, and $\boldsymbol{p}_i \to \boldsymbol{p}$, then $\boldsymbol{p}_i \in \exp_{\boldsymbol{p}}(U_\varepsilon)$ for large enough i, and $\mathrm{d}(\boldsymbol{p}, \boldsymbol{p}_i) = |(\exp_{\boldsymbol{p}|U_\varepsilon})^{-1}\boldsymbol{p}_i|$. The claim then follows from the continuity of $|(\exp_{\boldsymbol{p}|U_\varepsilon})^{-1}|$. □

The second statement in the above theorem essentially says that even though the distance on M is not the same as that of the ambient space, any map $\boldsymbol{f} : M \to \mathbb{R}^k$ is continuous in the usual sense if and only if it is continuous as a map from the metric space (M, d).

Corollary 6.5.1. *If $U_\varepsilon = \{\boldsymbol{v} \in M_{\boldsymbol{p}} \mid |\boldsymbol{v}| < \varepsilon\}$ is a neighborhood of the origin in $M_{\boldsymbol{p}}$ on which $\exp_{\boldsymbol{p}}$ is a diffeomorphism, then $\exp_{\boldsymbol{p}}(U_\varepsilon) = B^M_\varepsilon(\boldsymbol{p})$.*

Proof. It is clear that the left side is contained in the right one, since any point in $\exp_{\boldsymbol{p}}(U_\varepsilon)$ can be joined to $\boldsymbol{p}$ by a geodesic of length smaller than ε. To show that they are equal, we only need to establish that $\exp_{\boldsymbol{p}}(U_\varepsilon)$ is both open and closed, and apply Theorem 2.4.4, since $B^M_\varepsilon(\boldsymbol{p})$ is connected (recall that any $\boldsymbol{q} \in B^M_\varepsilon(\boldsymbol{p})$ can be joined to $\boldsymbol{p}$ by means of a curve of length smaller than ε, so that this curve lies entirely in $B^M_\varepsilon(\boldsymbol{p})$). By assumption, $(\exp_{\boldsymbol{p}|U_\varepsilon})^{-1}$ is continuous, so that $\exp_{\boldsymbol{p}}(U_\varepsilon)$ is open. To see that it is closed, consider a boundary point $\boldsymbol{q}$ of $\exp_{\boldsymbol{p}}(U_\varepsilon)$ in $B^M_\varepsilon(\boldsymbol{p})$ and a sequence $\{\boldsymbol{q}_i\}$ contained in $\exp_{\boldsymbol{p}}(U_\varepsilon)$ that converges to $\boldsymbol{q}$, see also Exercise 1.27. The sequence $\boldsymbol{v}_i := (\exp_{\boldsymbol{p}|U_\varepsilon})^{-1}(\boldsymbol{q}_i)$, being bounded, contains a subsequence $\{\boldsymbol{v}_{i_j}\}$ that converges to some $\boldsymbol{v} \in \bar{U}_\varepsilon$. Now, $|\boldsymbol{v}_i| = \mathrm{d}(\boldsymbol{p}, \boldsymbol{q}_i) \to \mathrm{d}(\boldsymbol{p}, \boldsymbol{q})$, so the subsequence also converges in norm to $\mathrm{d}(\boldsymbol{p}, \boldsymbol{q})$. This means that $|\boldsymbol{v}| = \mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) < \varepsilon$, and $\boldsymbol{v} \in U_\varepsilon$. We then have that $\boldsymbol{q}_{i_j} = \exp_{\boldsymbol{p}}(\boldsymbol{v}_{i_j}) \to \exp_{\boldsymbol{p}}(\boldsymbol{v})$, and since $\boldsymbol{q}_i \to \boldsymbol{q}$, $\exp_{\boldsymbol{p}}(\boldsymbol{v}) = \boldsymbol{q}$. This shows that $\exp_{\boldsymbol{p}}(U_\varepsilon)$ is closed in $B^M_\varepsilon(\boldsymbol{p})$ and concludes the argument. □

We will shortly see that the conclusion of the above corollary is still true under the much weaker assumption that $\exp_{\boldsymbol{p}}$ is defined on U_ε. When there is no danger of confusing M with the ambient Euclidean space, we will denote $B^M_\varepsilon(\boldsymbol{p})$ by $B_\varepsilon(\boldsymbol{p})$.

Proposition 6.5.1. *Let W denote a neighborhood of the zero section in TM such that the restriction of $(\pi, \exp)$ to W is a diffeomorphism onto its image, cf. Remark 3.7.1. Given $\boldsymbol{p} \in M$, choose $\varepsilon > 0$ small enough that $B_\varepsilon(\boldsymbol{p}) \times B_\varepsilon(\boldsymbol{p}) \subset (\pi, \exp)(W)$. Then*

$$\mathrm{d}(\boldsymbol{q}_1, \boldsymbol{q}_2) = |\left((\pi, \exp)_{|W}\right)^{-1}(\boldsymbol{q}_1, \boldsymbol{q}_2)|, \qquad \boldsymbol{q}_1, \boldsymbol{q}_2 \in B_{\varepsilon/3}(\boldsymbol{p}) \times B_{\varepsilon/3}(\boldsymbol{p}).$$

In particular, $\mathrm{d}^2 : M \times M \to \mathbb{R}$ is differentiable on a neighborhood U of the diagonal $\Delta = \{(\boldsymbol{p}, \boldsymbol{p}) \mid \boldsymbol{p} \in M\}$ in $M \times M$, and the distance function d itself is differentiable on $U \setminus \Delta$.

Proof. Given $\boldsymbol{p} \in M$, $U_\varepsilon^{\boldsymbol{p}}$ will denote the open ball of radius ε centered at the origin in $M_{\boldsymbol{p}}$. Notice that if $\boldsymbol{q}_1 \in B_{\varepsilon/3}(\boldsymbol{p})$, then $\exp_{\boldsymbol{q}_1}(U_{2\varepsilon/3}^{\boldsymbol{q}_1}) \subset B_{2\varepsilon/3}(\boldsymbol{q}_1) \subset B_\varepsilon(\boldsymbol{p})$, so that the restriction of $\exp_{\boldsymbol{q}_1}$ to $U_{2\varepsilon/3}^{\boldsymbol{q}_1}$ is a diffeomorphism onto its image $B_{2\varepsilon/3}(\boldsymbol{q}_1)$ by the corollary above. But if $\boldsymbol{q}_2 \in B_{\varepsilon/3}(\boldsymbol{p})$, then it also belongs to $B_{2\varepsilon/3}(\boldsymbol{q}_1)$, so that

$$\mathrm{d}(\boldsymbol{q}_1, \boldsymbol{q}_2) = |\left(\exp_{\boldsymbol{q}_1|U_{2\varepsilon/3}^{\boldsymbol{q}_1}}\right)^{-1} \boldsymbol{q}_2| = |\left((\pi, \exp)_{|W}\right)^{-1}(\boldsymbol{q}_1, \boldsymbol{q}_2)|.$$

The remaining assertions then follow from the fact that in an inner product space, the norm function is differentiable away from the origin, see Exercise 2.8. □

It is in general not true, however, that the distance function is smooth everywhere outside the diagonal: consider for example $M = S^1$. One of the possible two normal geodesics emanating from $\boldsymbol{p} = (1, 0)$ is $\boldsymbol{c}$, where $\boldsymbol{c}(t) = (\cos t, \sin t)$. Then $\boldsymbol{c}$ minimizes up to time π, so that

$$\mathrm{d}(\boldsymbol{p}, \boldsymbol{c}(t)) = \begin{cases} L(\boldsymbol{c}_{|[0,t]} = t \text{ if } t < \pi, \\ L(\boldsymbol{c}_{|[t,2\pi]} = 2\pi - t \text{ if } t > \pi. \end{cases}$$

We leave it as an exercise to show that this implies that the distance function squared $\mathrm{d}^2 : S^1 \times S^1 \to \mathbb{R}$ is not differentiable exactly on the anti-diagonal $\{(\boldsymbol{q}, -\boldsymbol{q}) \mid \boldsymbol{q} \in S^1\}$.

6.6 The Hopf-Rinow theorem

Recall that a metric space is said to be complete if every Cauchy sequence in the space converges. One of the most striking results that relates metric and differential geometric properties of a manifold is the Hopf-Rinow theorem, which roughly says that completeness as a metric space is equivalent to geodesics being defined for all time; the latter property is of course equivalent to completeness of the geodesic spray vector field introduced in Section 3.7.

Fix a point $\boldsymbol{p}$ in a manifold M, and for $\varepsilon > 0$, denote by $U_\varepsilon \subset M_{\boldsymbol{p}}$ the open ball of radius ε centered at the origin in the tangent space at $\boldsymbol{p}$.

Definition 6.6.1. The *injectivity radius at $\boldsymbol{p}$* is defined as

$$\mathrm{inj}_{\boldsymbol{p}} = \sup\{r > 0 \mid \exp_p : U_r \to B_r(\boldsymbol{p}) \text{ is a diffeomorphism}\},$$

provided the set on the right is bounded above, and as ∞ if it is unbounded. The *injectivity radius of a subset* $A \subset M$ is defined as

$$\mathrm{inj}_A = \inf\{\mathrm{inj}_{\boldsymbol{p}} \mid \boldsymbol{p} \in A\}.$$

Examples and Remarks 6.6.1. (i) When $M = \mathbb{R}^n$, $\mathrm{inj}_{\boldsymbol{p}} = \infty$ for every $\boldsymbol{p}$.
(ii) On $M = \mathbb{R}^n \setminus \{\mathbf{0}\}$, $\mathrm{inj}_{\boldsymbol{p}} = |\boldsymbol{p}|$. The injectivity radius of M itself is zero.
(iii) On the sphere of radius r, the injectivity radius is πr at any point, and therefore so is the injectivity radius of the whole sphere.
(iv) If $\boldsymbol{q} \in B_r(\boldsymbol{p})$, where $r \leq \mathrm{inj}_{\boldsymbol{p}}$, then there exists a unique normal geodesic from $\boldsymbol{p}$ to $\boldsymbol{q}$ of length equal to $\mathrm{d}(\boldsymbol{p}, \boldsymbol{q})$: in fact, $\boldsymbol{q} \in B_{r'}(\boldsymbol{p})$, where $r' = (\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) + \mathrm{inj}_{\boldsymbol{p}})/2 < r$, because $\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) \leq (\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) + \mathrm{inj}_{\boldsymbol{p}})/2$. Furthermore, $r' < r$, so that $\exp_{\boldsymbol{p}} : U_{r'} \to B_{r'}(\boldsymbol{p})$ is a diffeomorphism.

Lemma 6.6.1. *The injectivity radius of a (nonempty) compact set is positive.*

Proof. Suppose, to the contrary, that A is a compact set with zero injectivity radius. Then there exists a sequence $\{\boldsymbol{p}_k\} \subset A$ such that $\mathrm{inj}_{\boldsymbol{p}_k} \to 0$. This in turn means that for each natural number k, there exists some $\boldsymbol{v}_k \in M_{\boldsymbol{p}_k}$ with $|\boldsymbol{v}_k| \to 0$ but $\boldsymbol{v}_k$ lies outside the open set W on which $(\pi, \exp) : W \to \pi(W) \times \exp(W)$ is a diffeomorphism. Now, A is compact, so after passing to a subsequence if necessary, it may be assumed that $\{\boldsymbol{p}_k\}$ converges to some $\boldsymbol{p} \in A$. Then $\{\boldsymbol{v}_k\}$ converges to the zero vector in $M_{\boldsymbol{p}}$, and since $\boldsymbol{v}_k$ lies in the closed set $TM \setminus W$, so does the zero vector. This contradicts the fact that W is a neighborhood of the zero section. □

Most of the work required in proving the Hopf-Rinow theorem is contained in the following:

Proposition 6.6.1. *Let* $\boldsymbol{p} \in M$. *If* $\exp_{\boldsymbol{p}}$ *is defined on* U_ε, *then there exists a minimal geodesic joining* $\boldsymbol{p}$ *to any* $\boldsymbol{q} \in B_\varepsilon(\boldsymbol{p})$.

Proof. Denote by I the set of all $r \in (0, \varepsilon)$ for which there exists a minimal geodesic from $\boldsymbol{p}$ to any point in the closure of $B_r(\boldsymbol{p})$. I is an interval by definition, and is nonempty because the injectivity radius at $\boldsymbol{p}$ is positive. We will show that $I = (0, \varepsilon)$ by arguing that it is both open and closed in $(0, \varepsilon)$. This will prove the proposition, since $B_\varepsilon(\boldsymbol{p}) = \cup_{r \in (0,\varepsilon)} \overline{B_r(\boldsymbol{p})}$.
To see that it is closed, assume $(0, \delta) \subset I$ for some $\delta > 0$. We must show that $\delta \in I$; i.e., that for any $\boldsymbol{q}$ at distance $\leq \delta$ from $\boldsymbol{p}$, there exists a geodesic $\boldsymbol{c} : [0, 1] \to M$ from $\boldsymbol{p}$ to $\boldsymbol{q}$ with length $L(\boldsymbol{c}) = \mathrm{d}(\boldsymbol{p}, \boldsymbol{q})$. So choose some sequence $\{\boldsymbol{q}_k\} \subset B_\delta(\boldsymbol{p})$ that converges to $\boldsymbol{q}$. By assumption, there exists for each k a minimal geodesic $\boldsymbol{c}_k : [0, 1] \to M$ from $\boldsymbol{p}$ to $\boldsymbol{q}_k$. Now, $|\dot{\boldsymbol{c}}_k(0)| = L(\boldsymbol{c}_k) = \mathrm{d}(\boldsymbol{p}, \boldsymbol{q}_k) < \delta$, so that $\{\dot{\boldsymbol{c}}_k(0)\}$ is contained inside the compact set $\overline{U_\delta}$, and some subsequence $\{\dot{\boldsymbol{c}}_{k_i}(0)\}$ converges to, say, $\boldsymbol{v} \in \overline{U_\delta}$. The geodesic $\boldsymbol{c} : [0, 1] \to M$, where $\boldsymbol{c}(t) = \exp_{\boldsymbol{p}}(t\boldsymbol{v})$, joins $\boldsymbol{p}$ to $\boldsymbol{q}$ because

$$\boldsymbol{q} = \lim_{i \to \infty} \boldsymbol{q}_{k_i} = \lim_{i \to \infty} \exp(\dot{\boldsymbol{c}}_{k_i}(0)) = \exp(\lim_{i \to \infty} \dot{\boldsymbol{c}}_{k_i}(0)) = \exp(\boldsymbol{v}).$$

Furthermore, its length satisfies

$$L(\boldsymbol{c}) = |\boldsymbol{v}| = \lim_{i\to\infty} |\dot{\boldsymbol{c}}_{k_i}(0)| = \lim_{i\to\infty} \mathrm{d}(\boldsymbol{p}, \boldsymbol{q}_{k_i}) = \mathrm{d}(\boldsymbol{p}, \boldsymbol{q})$$

by continuity of the distance function. This shows that I is closed.
To see that it is open, assume $(0, \delta] \subset I$. Notice that the closure of $B_\delta(\boldsymbol{p})$ must then be compact: indeed, by continuity of $\exp_{\boldsymbol{p}}$, $\exp_{\boldsymbol{p}}(\overline{U_\delta})$ is contained in $\overline{B_\delta(\boldsymbol{p})}$, and being compact, is closed. But it also contains $B_\delta(\boldsymbol{p})$, and must therefore also contain its closure. Thus, $\overline{B_\delta(\boldsymbol{p})} = \exp_{\boldsymbol{q}}(\overline{U_\delta})$ is compact. By Lemma 6.6.1, the injectivity radius α of $\overline{B_\delta(\boldsymbol{p})}$ is positive. We claim that if $0 < \beta < \min\{\alpha, \varepsilon - \delta\}$, then $\delta + \beta \in I$, thereby implying that I is open. To establish the claim, consider a point $\boldsymbol{q} \in \overline{B_{\delta+\beta}(\boldsymbol{p})} \setminus \overline{B_\delta(\boldsymbol{p})}$, and a sequence $\boldsymbol{c}_k : [0, 1] \to M$ of curves from $\boldsymbol{p}$ to $\boldsymbol{q}$ with length $\leq \mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) + 1/k$. The intermediate value theorem guarantees the existence of a parameter value t_k for which $\mathrm{d}(\boldsymbol{p}, \boldsymbol{c}_k(t_k)) = \delta$. By compactness of $\overline{B_\delta(\boldsymbol{p})}$, we may assume after passing to a subsequence if necessary that $\boldsymbol{r}_k := \boldsymbol{c}_k(t_k)$ converges to some $\boldsymbol{r}$, which, by continuity of distance, lies at distance δ from $\boldsymbol{p}$. Now,

$$\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) + \frac{1}{k} \geq L(\boldsymbol{c}_k) = L(\boldsymbol{c}_{k|[0,t_k]}) + L(\boldsymbol{c}_{k|[t_k,1]}) \geq \mathrm{d}(\boldsymbol{p}, \boldsymbol{r}_k) + \mathrm{d}(\boldsymbol{r}_k, \boldsymbol{q}), \tag{6.6.1}$$

so that

$$\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) \geq \mathrm{d}(\boldsymbol{p}, \boldsymbol{r}) + \mathrm{d}(\boldsymbol{r}, \boldsymbol{q}). \tag{6.6.2}$$

Indeed, if $\mathrm{d}(\boldsymbol{p}, \boldsymbol{q})$ were smaller than $\mathrm{d}(\boldsymbol{p}, \boldsymbol{r}) + \mathrm{d}(\boldsymbol{r}, \boldsymbol{q})$, then $\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) < \mathrm{d}(\boldsymbol{p}, \boldsymbol{r}) + \mathrm{d}(\boldsymbol{r}, \boldsymbol{q}) - 2\varepsilon$ for some $\varepsilon > 0$, and thus, for all large enough k,

$$\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) + \frac{1}{k} < \mathrm{d}(\boldsymbol{p}, \boldsymbol{r}) + \mathrm{d}(\boldsymbol{r}, \boldsymbol{q}) - \varepsilon. \tag{6.6.3}$$

But then (6.6.1) and (6.6.3) would imply that

$$\mathrm{d}(\boldsymbol{p}, \boldsymbol{r}_k) + \mathrm{d}(\boldsymbol{r}_k, \boldsymbol{q}) < \mathrm{d}(\boldsymbol{p}, \boldsymbol{r}) + \mathrm{d}(\boldsymbol{r}, \boldsymbol{q}) - \varepsilon,$$

which contradicts the fact that $\{\boldsymbol{r}_k\}$ converges to $\boldsymbol{r}$. Thus, (6.6.2) holds, and together with the triangle inequality, we obtain

$$\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) = \mathrm{d}(\boldsymbol{p}, \boldsymbol{r}) + \mathrm{d}(\boldsymbol{r}, \boldsymbol{q}). \tag{6.6.4}$$

By assumption, there exists a minimal geodesic $\boldsymbol{c}_1 : [0, \delta] \to M$ from $\boldsymbol{p}$ to $\boldsymbol{r}$. Furthermore, $\mathrm{d}(\boldsymbol{r}, \boldsymbol{q}) = \mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) - \mathrm{d}(\boldsymbol{p}, \boldsymbol{r}) \leq \delta + \beta - \delta = \beta$ is less than the injectivity radius of $\overline{B_\delta(\boldsymbol{p})}$, and since the latter set contains $\boldsymbol{r}$, there exists a minimal geodesic $\boldsymbol{c}_2 : [\delta, \delta + \mathrm{d}(\boldsymbol{r}, \boldsymbol{q})] \to M$ from $\boldsymbol{r}$ to $\boldsymbol{q}$. The *a priori* only piecewise geodesic $\boldsymbol{c} : [0, \delta + \mathrm{d}(\boldsymbol{r}, \boldsymbol{q})] \to M$, where $\boldsymbol{c}(t) = \boldsymbol{c}_1(t)$ when $t \leq \delta$, and $\boldsymbol{c}(t) = \boldsymbol{c}_2(t)$ when $t \geq \delta$, is a curve from $\boldsymbol{p}$ to $\boldsymbol{q}$ whose length realizes the distance between its endpoints by (6.6.4). It must therefore be a geodesic; i.e., $\dot{\boldsymbol{c}}_1(\delta) = \dot{\boldsymbol{c}}_2(\delta)$. This completes the proof. □

Remark 6.6.1. We reiterate a fact observed in the proof of the proposition: if $\exp_{\boldsymbol{p}}$ is defined on U_ε, then the closure $\overline{B_r(\boldsymbol{p})}$ of $B_r(\boldsymbol{p})$ is compact for any $r \in (0, \varepsilon)$: indeed, continuity of the exponential map implies that $\exp_{\boldsymbol{p}}(\overline{U_r})$ is compact, hence closed, and must therefore contain $\overline{B_r(\boldsymbol{p})}$. But it must also be contained inside $\overline{B_r(\boldsymbol{p})}$ by continuity of $\exp_{\boldsymbol{p}}$ again, so the two sets are equal.

Theorem 6.6.1. *The following statements are equivalent in a connected Riemannian manifold M:*

(1) *M is complete as a metric space.*
(2) *For <u>all</u> $\boldsymbol{p} \in M$, $\exp_{\boldsymbol{p}}$ is defined on all of $M_{\boldsymbol{p}}$.*
(3) *For <u>some</u> $\boldsymbol{p} \in M$, $\exp_{\boldsymbol{p}}$ is defined on all of $M_{\boldsymbol{p}}$.*
(4) *Any bounded set of M (with respect to the distance d) has compact closure.*

Furthermore, completeness of M implies that any two points $\boldsymbol{p}$ and $\boldsymbol{q}$ of M can be joined by a geodesic of length $\mathrm{d}(\boldsymbol{p}, \boldsymbol{q})$.

Proof. *1 $\Rightarrow$ 2*: Let S denote the geodesic spray on TM (see Theorem 3.7.2), and $\boldsymbol{\gamma} : I \to TM$ a maximal integral curve of S, so that $\boldsymbol{c} := \pi \circ \boldsymbol{\gamma}$ is a geodesic with $\dot{\boldsymbol{c}} = \boldsymbol{\gamma}$. To show that $I = \mathbb{R}$, it is enough to show that it is closed, since it is already open. So consider a sequence $\{t_k\} \subset I$, $t_k \to t_0 \in \mathbb{R}$. The sequence $\{\boldsymbol{c}(t_k)\}$ is then a Cauchy sequence in M, because $|\dot{\boldsymbol{c}}|$ is constant equal to some $a > 0$, so that

$$\mathrm{d}(\boldsymbol{c}(t_k), \boldsymbol{c}(t_l)) \leq \left| \int_{t_k}^{t_l} |\dot{\boldsymbol{c}}| \right| = a|t_k - t_l|,$$

and the Cauchy property for $\{\boldsymbol{c}(t_k)\}$ follows from that for $\{t_k\}$. By assumption, $\{\boldsymbol{c}(t_k)\}$ converges and is therefore contained in some compact set $K \subset M$. But then $\{\boldsymbol{\gamma}(t_k)\}$ lies in the compact set $\{\boldsymbol{v} \in TM \mid \pi(\boldsymbol{v}) \in K \text{ and } |\boldsymbol{v}| = a\}$, and has therefore a convergent subsequence. By Theorem 3.3.3, $t_0 \in I$.

2 $\Rightarrow$ 3: Immediate.

3 $\Rightarrow$ 4: If A is bounded, then it is contained inside some closed metric ball of sufficiently large radius centered at $\boldsymbol{p}$. The latter is compact by Remark 6.6.1, so that $\bar{A}$ is also compact.

4 $\Rightarrow$ 1: We already observed in Chapter 1 that any Cauchy sequence is bounded. It therefore lies inside a compact set by hypothesis, and thus admits a convergent subsequence. Then the sequence itself converges as shown in the proof of Theorem 1.8.4. To complete the proof of the theorem, it suffices to show that the second statement implies that any two points of M can be joined by a minimal geodesic. This is an immediate consequence of Proposition 6.6.1. □

6.7 Curvature comparison

One of the most active areas of research in Riemannian geometry is the interaction between the shape of a space and its curvature. What can one say about the manifold if its curvature is known? Although this topic is too vast and complex to allow for a comprehensive account here, we wish to establish a couple of results that hint at the beauty of this subject.

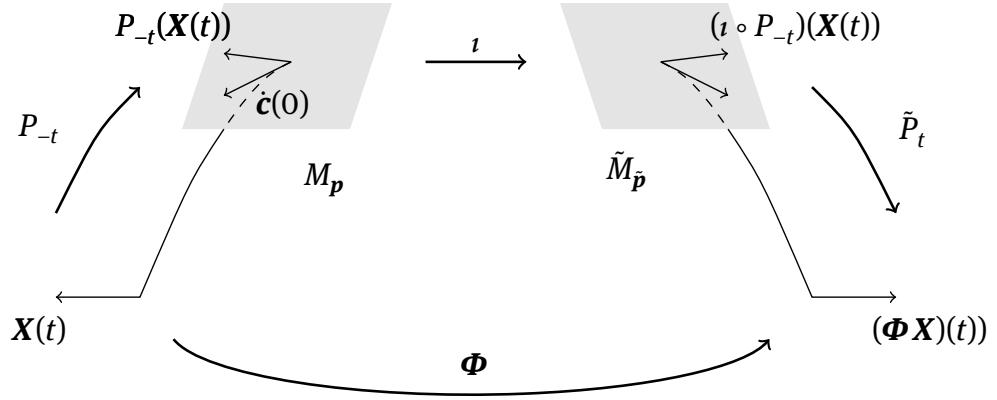
In this section we will see that the larger the curvature of a space, the earlier conjugate points appear along geodesics. This fact will be derived by comparing the index form in spaces where the curvature of one is larger than that of the other. As an application, a complete space with curvature bounded below by a positive constant must be compact. The construction used in the proof also shows how the metric is determined by the curvature.

Denote by $\mathfrak{X}_{\boldsymbol{c}}$ the vector space of all piecewise-smooth vector fields along a curve $\boldsymbol{c}$. Let $M, \tilde{M}$ denote manifolds of the same dimension n, $\boldsymbol{c} : [0, b] \to M$ (resp. $\tilde{\boldsymbol{c}} : [0, b] \to \tilde{M}$) a normal geodesic in M (resp. $\tilde{M}$), and $\boldsymbol{p} = \boldsymbol{c}(0)$, $\tilde{\boldsymbol{p}} = \tilde{\boldsymbol{c}}(0)$ their respective starting points. Suppose $\imath : M_{\boldsymbol{p}} \to M_{\tilde{\boldsymbol{p}}}$ is a linear isometry mapping $\dot{\boldsymbol{c}}(0)$ to $\dot{\tilde{\boldsymbol{c}}}(0)$. If $\boldsymbol{P}_t : M_{\boldsymbol{p}} \to M_{\boldsymbol{c}(t)}$ and $\tilde{\boldsymbol{P}}_t : \tilde{M}_{\tilde{\boldsymbol{p}}} \to \tilde{M}_{\tilde{\boldsymbol{c}}(t)}$ denote parallel translation along $\boldsymbol{c}$ and $\tilde{\boldsymbol{c}}$ respectively, define a map $\boldsymbol{\Phi} : \mathfrak{X}_{\boldsymbol{c}} \to \mathfrak{X}_{\tilde{\boldsymbol{c}}}$ by setting

$$(\boldsymbol{\Phi}\boldsymbol{X})(t) = (\tilde{\boldsymbol{P}}_t \circ \imath \circ \boldsymbol{P}_t^{-1})\boldsymbol{X}(t) = (\tilde{\boldsymbol{P}}_t \circ \imath \circ \boldsymbol{P}_{-t})\boldsymbol{X}(t), \qquad \boldsymbol{X} \in \mathfrak{X}_{\boldsymbol{c}},$$

with $P_{-t} = P_t^{-1}$ denoting parallel translation along $-\boldsymbol{c}$ from $M_{\boldsymbol{c}(t)}$ to $M_{\boldsymbol{p}}$. Notice that $\boldsymbol{\Phi}\dot{\boldsymbol{c}} = \dot{\tilde{\boldsymbol{c}}}$ and that if $\boldsymbol{X}$ is parallel along $\boldsymbol{c}$, then $\boldsymbol{\Phi}\boldsymbol{X}$ is the parallel field along $\tilde{c}$ where $\boldsymbol{\Phi}\boldsymbol{X}(0) = \imath\boldsymbol{X}(0)$.

Let $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n = \dot{\boldsymbol{c}}$ denote an orthonormal basis of parallel vector fields along $\boldsymbol{c}$, so that $\tilde{\boldsymbol{Z}}_1, \dots, \tilde{\boldsymbol{Z}}_n$, with $\tilde{\boldsymbol{Z}}_i = \boldsymbol{\Phi}\boldsymbol{Z}_i$, is one along $\tilde{\boldsymbol{c}}$. Any piecewise smooth vector field $\boldsymbol{X}$ along $\boldsymbol{c}$ may then be written as $\boldsymbol{X} = \sum_i \varphi_i \boldsymbol{Z}_i$, with $\varphi_i = \langle \boldsymbol{X}, \boldsymbol{Z}_i \rangle$ piecewise smooth. By construction, $\boldsymbol{\Phi}\boldsymbol{X} = \sum_i \varphi_i \tilde{\boldsymbol{Z}}_i$.



It follows immediately that if $\mathfrak{X}_{\boldsymbol{c}}^{\perp}$ denotes the subspace of $\mathfrak{X}_{\boldsymbol{c}}$ consisting of all those orthogonal to $\boldsymbol{c}$, and $\mathfrak{V}_{\boldsymbol{c}}$ the subspace consisting of all the elements in $\mathfrak{X}_{\boldsymbol{c}}^{\perp}$ that vanish at the endpoints, then $\boldsymbol{\Phi}$ maps $\mathfrak{X}_{\boldsymbol{c}}$ isomorphically onto $\mathfrak{X}_{\tilde{\boldsymbol{c}}}$, and the same is true for the corresponding subspaces. Furthermore, $\boldsymbol{\Phi}$ commutes with covariant differentiation, since

$$(\boldsymbol{\Phi}\boldsymbol{X})' = \sum_i \varphi_i' \tilde{\boldsymbol{Z}}_i = \boldsymbol{\Phi}(\boldsymbol{X}'). \tag{6.7.1}$$

In order to state our next result, we introduce some terminology: for each $t \in [0, b]$, let $\boldsymbol{\Phi}_t : M_{\boldsymbol{c}(t)} \to \tilde{M}_{\tilde{\boldsymbol{c}}(t)}$ be the linear isometry $\tilde{P}_t \circ \imath \circ P_t^{-1}$. Thus, $(\boldsymbol{\Phi}\boldsymbol{X})(t) = \boldsymbol{\Phi}_t(\boldsymbol{X}(t))$ for any vector field $\boldsymbol{X}$ along $\boldsymbol{c}$.

Proposition 6.7.1. *Suppose that for any $t \in [0, b]$ and for any 2-plane $E_t \subset M_{\boldsymbol{c}(t)}$ containing $\dot{\boldsymbol{c}}(t)$, the sectional curvature K_{E_t} of E_t is greater than or equal to that of $\boldsymbol{\Phi}_t(E_t)$. Then*

$$I(\boldsymbol{X}, \boldsymbol{X}) \le I(\boldsymbol{\Phi}\boldsymbol{X}, \boldsymbol{\Phi}\boldsymbol{X}), \qquad \boldsymbol{X} \in \mathfrak{X}_{\boldsymbol{c}}^{\perp}.$$

In particular, if $\tilde{\boldsymbol{c}}$ has a conjugate point, then so does $\boldsymbol{c}$.

Proof. Given $\boldsymbol{X} \in \mathfrak{X}_{\boldsymbol{c}}$, define functions $K, \tilde{K} : [0, b] \to \mathbb{R}$ by letting $K(t)$ denote the sectional curvature of the plane spanned by $\dot{\boldsymbol{c}}(t)$ and $\boldsymbol{X}(t)$, and similarly, letting $\tilde{K}(t)$ denote the curvature of the image of that plane via $\boldsymbol{\Phi}_t$. We then have

$$\begin{aligned} I(\boldsymbol{X}, \boldsymbol{X}) &= \int_0^b |\boldsymbol{X}'|^2 - K|\boldsymbol{X}| \le \int_0^b |\boldsymbol{X}'|^2 - \tilde{K}|\boldsymbol{X}| = \int_0^b |(\boldsymbol{\Phi}\boldsymbol{X})'|^2 - \tilde{K}|\boldsymbol{\Phi}\boldsymbol{X}| \\ &= I(\boldsymbol{\Phi}\boldsymbol{X}, \boldsymbol{\Phi}\boldsymbol{X}), \end{aligned}$$

as claimed. Notice that we used both (6.7.1) and the fact that $\boldsymbol{\Phi}$ is a linear isometry. □

Remarks 6.7.1. (i) The same argument shows that if the curvature of each plane E_t in the statement of Proposition is less than or equal to that of $\boldsymbol{\Phi}(E_t)$, then $I(\boldsymbol{X}, \boldsymbol{X}) \ge I(\boldsymbol{\Phi}\boldsymbol{X}, \boldsymbol{\Phi}\boldsymbol{X})$ for every $\boldsymbol{X} \in \mathfrak{X}_{\boldsymbol{c}}^{\perp}$.

(ii) The hypothesis of the Proposition is clearly satisfied if there exists a constant κ such that $K_M \ge \kappa \ge K_{\tilde{M}}$.

The *diameter* of a bounded metric space M is defined to be

$$\operatorname{diam} M = \sup\{\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) \mid \boldsymbol{p}, \boldsymbol{q} \in M\}.$$

For example, a sphere of constant curvature $\kappa > 0$ has diameter $\pi/\sqrt{\kappa}$.

Theorem 6.7.1. *If M is complete and has sectional curvature $K \ge \kappa > 0$, then its diameter is no larger than $\pi/\sqrt{\kappa}$. In particular, M is compact.*

Proof. Recall from Example 6.2.2 that on the sphere $\tilde{M}$ of constant curvature κ, any normal geodesic has $\pi/\sqrt{\kappa}$ as conjugate point. Proposition 6.7.1 and Theorem 6.4.1 then imply that any geodesic of M with length greater than $\pi/\sqrt{\kappa}$ is not minimal, so that

$\operatorname{diam} M \le \pi/\sqrt{\kappa}$. Compactness of M follows from completeness and the Hopf-Rinow theorem. □

As a final application, we adapt the construction used at the beginning of the section to prove a result of E. Cartan that illustrates how the curvature locally determines the metric. Let $\iota : M_{\boldsymbol{p}} \to \tilde{M}_{\tilde{\boldsymbol{p}}}$ be a linear isometry, and $U \subset M_{\boldsymbol{p}}$ a neighborhood of $\mathbf{0} \in M_{\boldsymbol{p}}$ on which $\exp_{\boldsymbol{p}} : U \to V := \exp_{\boldsymbol{p}}(U)$ is a diffeomorphism. By restricting U further if necessary, we may assume that the exponential map of $\tilde{M}$ at $\tilde{\boldsymbol{p}}$ maps $\iota(U)$ diffeomorphically onto its image. Set

$$\boldsymbol{f} = \exp_{\tilde{\boldsymbol{p}}} \circ \iota \circ (\exp_{\boldsymbol{p}|U})^{-1} : V \to \tilde{M}.$$

Next, we define a map $\boldsymbol{\Phi} : TV \to T\boldsymbol{f}(V)$ that sends each $M_{\boldsymbol{q}}$ to $\tilde{M}_{\boldsymbol{f}(\boldsymbol{q})}$, $\boldsymbol{q} \in V$, as follows: given any $\boldsymbol{q} \in V$, denote by $\boldsymbol{c}$ the unique minimal geodesic from $\boldsymbol{p}$ to $\boldsymbol{q}$, and by $\tilde{\boldsymbol{c}}$ the geodesic $t \mapsto \exp_{\tilde{\boldsymbol{p}}}(t \cdot \iota\dot{\boldsymbol{c}}(0))$. If P and $\tilde{P}$ are parallel translation along $\boldsymbol{c}$ and $\tilde{\boldsymbol{c}}$ respectively, then the restriction $\boldsymbol{\Phi} : M_{\boldsymbol{q}} \to \tilde{M}_{\boldsymbol{f}(\boldsymbol{q})}$ of $\boldsymbol{\Phi}$ to $M_{\boldsymbol{q}}$ is given by $\tilde{P} \circ \iota \circ P^{-1}$. This is essentially the map $\boldsymbol{\Phi}$ used earlier, but acting on individual vectors rather than vector fields. Notice that $\boldsymbol{\Phi}$ covers $\boldsymbol{f}$, in the sense that the diagram

$$\begin{array}{ccc} TV & \xrightarrow{\boldsymbol{\Phi}} & T\boldsymbol{f}(V) \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\tilde{\pi}} \\ V & \xrightarrow[\boldsymbol{f}]{} & \boldsymbol{f}(V) \end{array}$$

commutes, with π and $\tilde{\pi}$ denoting the respective tangent bundle projections.

Theorem 6.7.2. *With notation as above, suppose that*

$$\langle R(\boldsymbol{x},\boldsymbol{y})\boldsymbol{u},\boldsymbol{v}\rangle = \langle \tilde{R}(\boldsymbol{\Phi}\boldsymbol{x},\boldsymbol{\Phi}\boldsymbol{y})\boldsymbol{\Phi}\boldsymbol{u},\boldsymbol{\Phi}\boldsymbol{v}\rangle, \qquad \boldsymbol{x},\boldsymbol{y},\boldsymbol{u},\boldsymbol{v} \in M_{\boldsymbol{q}}, \quad \boldsymbol{q} \in V.$$

Then $\boldsymbol{f}$ is an isometry, and $\boldsymbol{f}_{*\boldsymbol{p}} = \iota$.

Proof. By construction, $\boldsymbol{f} : V \to \boldsymbol{f}(V)$ is a diffeomorphism and $\boldsymbol{f}_{*\boldsymbol{p}} = \iota$. To see that $\boldsymbol{f}$ is isometric, fix an arbitrary $\boldsymbol{q} \in V$ and $\boldsymbol{u} \in M_{\boldsymbol{q}}$. It must be shown that $|\boldsymbol{f}_*\boldsymbol{u}| = |\boldsymbol{u}|$. So consider the minimal geodesic $\boldsymbol{c} : [0, t_0] \to V$ from $\boldsymbol{p}$ to $\boldsymbol{q}$, and the unique Jacobi field $\boldsymbol{Y}$ along $\boldsymbol{c}$ with $\boldsymbol{Y}(0) = \mathbf{0}$, $\boldsymbol{Y}(t_0) = \boldsymbol{u}$, cf. Exercise 6.2. Let $\boldsymbol{E}_1, \dots, \boldsymbol{E}_n$ denote a basis of parallel orthonormal fields along $\boldsymbol{c}$ with $\boldsymbol{E}_n = \dot{\boldsymbol{c}}$. The fields $\tilde{\boldsymbol{E}}_i$, where $\tilde{\boldsymbol{E}}_i(t) = \boldsymbol{\Phi}\boldsymbol{E}_i(t)$, then form a basis of parallel orthonormal fields along $\tilde{\boldsymbol{c}}$ by definition of $\boldsymbol{\Phi}$, and $\tilde{\boldsymbol{E}}_n = \dot{\tilde{\boldsymbol{c}}}$. If $f_i = \langle \boldsymbol{Y}, \boldsymbol{E}_i\rangle$, $i = 1, \dots, n$, so that $\boldsymbol{Y} = \sum_i f_i \boldsymbol{E}_i$, then the Jacobi condition for $\boldsymbol{Y}$ reads

$$f_j'' + \sum_i f_i \langle R(\boldsymbol{E}_i, \boldsymbol{E}_n)\boldsymbol{E}_n, \boldsymbol{E}_j\rangle = 0, \qquad j = 1, \dots, n.$$

This implies that the vector field $\tilde{\boldsymbol{Y}} := \boldsymbol{\Phi}\boldsymbol{Y}$ is then also Jacobi, because $\tilde{\boldsymbol{Y}} = \sum_i f_i \tilde{\boldsymbol{E}}_i$, and

$$\langle R(\boldsymbol{E}_i, \boldsymbol{E}_n)\boldsymbol{E}_n, \boldsymbol{E}_j\rangle = \langle R(\tilde{\boldsymbol{E}}_i, \tilde{\boldsymbol{E}}_n)\tilde{\boldsymbol{E}}_n, \tilde{\boldsymbol{E}}_j\rangle.$$

By definition of $\boldsymbol{\Phi}$, $|\tilde{\boldsymbol{Y}}(t)| = |\boldsymbol{Y}(t)|$ for all t, so the theorem is proved once we establish that $\tilde{\boldsymbol{Y}}(t_0) = \boldsymbol{f}_{*\boldsymbol{q}}\boldsymbol{u} = \boldsymbol{f}_{*\boldsymbol{q}}\boldsymbol{Y}(t_0)$. But both $\boldsymbol{Y}$ and $\tilde{\boldsymbol{Y}}$ vanish at 0, so that by (6.2.1),

$$\boldsymbol{Y}(t) = \exp_{\boldsymbol{p}*}\left(t\mathcal{I}_{t\dot{\boldsymbol{c}}(0)}\boldsymbol{Y}'(0)\right),$$

and similarly,

$$\tilde{\boldsymbol{Y}}(t) = \exp_{\tilde{\boldsymbol{p}}*}\left(t\mathcal{I}_{t\dot{\tilde{\boldsymbol{c}}}(0)}\tilde{\boldsymbol{Y}}'(0)\right) = \exp_{\tilde{\boldsymbol{p}}*}\left(t\mathcal{I}_{t\dot{\tilde{\boldsymbol{c}}}(0)}\imath\boldsymbol{Y}'(0)\right).$$

Thus,

$$\begin{aligned}\tilde{\boldsymbol{Y}}(t) &= \exp_{\tilde{\boldsymbol{p}}*}\left(\mathcal{I}_{t\dot{\tilde{\boldsymbol{c}}}(0)}t\imath\boldsymbol{Y}'(0)\right) = \left(\exp_{\tilde{\boldsymbol{p}}*}\circ\mathcal{I}_{t\dot{\tilde{\boldsymbol{c}}}(0)}\circ\imath\circ\mathcal{I}^{-1}_{t\dot{\boldsymbol{c}}(0)}\circ\exp^{-1}_{\boldsymbol{p}*}\right)\boldsymbol{Y}(t)\\ &= \left(\exp_{\tilde{\boldsymbol{p}}*}\circ\imath_*\circ\exp^{-1}_{\boldsymbol{p}*}\right)\boldsymbol{Y}(t)\\ &= \boldsymbol{f}_{*\boldsymbol{c}(t)}\boldsymbol{Y}(t),\end{aligned}$$

as claimed. This concludes the proof. Notice that we have actually shown that $\boldsymbol{\Phi}$ is the derivative of $\boldsymbol{f}$, since for every Jacobi field $\boldsymbol{Y}$ as above, $\tilde{\boldsymbol{Y}} = \boldsymbol{\Phi}\boldsymbol{Y}$ and $\tilde{\boldsymbol{Y}} = \boldsymbol{f}_*\boldsymbol{Y}$. □

Remark 6.7.2. The above theorem shows that any two spaces M and $\tilde{M}$ with the same constant curvature κ are locally isometric; i.e., for any $\boldsymbol{p}$ in M, there exists a neighborhood U of $\boldsymbol{p}$ and an isometry $\boldsymbol{f} : U \to \boldsymbol{f}(U) \subset \tilde{M}$. Such an $\boldsymbol{f}$ need not be extendable to all of M, though. For example, $\mathbb{R}^2$ and $S^1 \times S^1 \subset \mathbb{R}^4$ are both flat and therefore locally isometric, but not globally, since one is compact and the other is not.

6.8 Exercises

6.1. (a) Let $\boldsymbol{Y}$ be a Jacobi field along a geodesic $\boldsymbol{c}$, and f an affine function; i.e., a function of the form $f(t) = at + b$, $a, b \in \mathbb{R}$. Show that $\boldsymbol{c} \circ f$ is a geodesic, and that $\boldsymbol{Y} \circ f$ is a Jacobi field along $\boldsymbol{c} \circ f$.

(b) Notice that for any function f, not necessarily affine, $\boldsymbol{c} \circ f$ has (part of) the same image as $\boldsymbol{c}$. Explain why $\boldsymbol{c} \circ f$ is no longer a geodesic if f is not affine.

6.2. Prove that if t_0 is not a conjugate point of a geodesic $\boldsymbol{c} : [a, b] \to M$, then for any $\boldsymbol{v} \in M_{\boldsymbol{c}(a)}$ and $\boldsymbol{w} \in M_{\boldsymbol{c}(t_0)}$, there exists exactly one Jacobi field $\boldsymbol{Y}$ along $\boldsymbol{c}$ with $\boldsymbol{Y}(a) = \boldsymbol{v}$ and $\boldsymbol{Y}(t_0) = \boldsymbol{w}$.

6.3. Suppose V is a variation of $\boldsymbol{c} : [0, b] \to M$ with fixed endpoints, such that all the curves V_s are geodesics. Show that the length function of V is constant and that b is a conjugate point of $\boldsymbol{c}$.

6.4. Recall from Chapter 3 that a Killing field on M is a vector field whose flow consists of isometries of M. Prove that the restriction of a Killing field to any geodesic is a Jacobi field along that geodesic. *Hint:* Perhaps the easiest way to see this is to consider the flow $\boldsymbol{\Phi}_s$ of the Killing field, and consider the variation $(t, s) \mapsto \boldsymbol{\Phi}_s(\boldsymbol{c}(t))$ of the geodesic $\boldsymbol{c}$.

6.5. Let $\boldsymbol{c} : [0, b] \to M$ be a geodesic in a manifold with sectional curvature K, and V a (nonconstant) variation of $\boldsymbol{c}$ with fixed endpoints. Prove that if $K \leq 0$, then the length of V_s is larger than that of $\boldsymbol{c}$ if s is small enough. Show by means of an example that this does not necessarily imply that $\boldsymbol{c}$ is a minimal geodesic.

6.6. Let $\boldsymbol{E}$ denote a parallel vector field along a geodesic $\boldsymbol{c} : [0, b] \to M$ that is orthogonal to $\dot{\boldsymbol{c}}$. Choose an interval I around 0 small enough that the geodesic $s \mapsto \exp(s\boldsymbol{E}(t))$ is defined on I for all $t \in [0, b]$, and consider the variation $V : [0, b] \times I \to M$ of $\boldsymbol{c}$ given by $V(t, s) = \exp(s\boldsymbol{E}(t))$. Prove that if M has positive curvature, then $L(V_s) < L(\boldsymbol{c})$ for small s and if M has negative curvature, then $L(V_s) > L(\boldsymbol{c})$. Interpret this fact in terms of the lengths of circles of latitude compared to that of the equator on a sphere.

6.7. Let M denote the n-dimensional sphere of radius r, $\boldsymbol{c} : [0, a] \to M$ a normal geodesic, and $\mathfrak{E}$ the vector space of Jacobi fields along $\boldsymbol{c}$ that vanish at 0.
(a) Give an explicit formula for $I(Y_1, Y_2)$ in terms of $\langle \boldsymbol{u}_1, \boldsymbol{u}_2 \rangle$ if $Y_i \in \mathfrak{E}$ with $Y_i(0) = \boldsymbol{u}_i$.
(b) Show that the index form is negative definite on $\mathfrak{E}$ if $a \in (\pi r/2, \pi r)$.

6.8. Show that the distance function $\mathrm{d} : M \times M \to \mathbb{R}$ is continuous; i.e., show that if $\boldsymbol{p}_i \to \boldsymbol{p}$ and $\boldsymbol{q}_i \to \boldsymbol{q}$, then $\mathrm{d}(\boldsymbol{p}_i, \boldsymbol{q}_i) \to \mathrm{d}(\boldsymbol{p}, \boldsymbol{q})$.

6.9. Prove the claim made in the proof of Proposition 6.6.1: given $\boldsymbol{p} \in M$,

$$B_\varepsilon(\boldsymbol{p}) = \cup_{r \in (0,\varepsilon)} \overline{B_r(\boldsymbol{p})}.$$

6.10. Let $\tilde{M}$, M denote manifolds with distance functions $\tilde{\mathrm{d}}$ and d respectively. Show that if $\boldsymbol{f} : \tilde{M} \to M$ is isometric, then $\tilde{\mathrm{d}}(\boldsymbol{p}, \boldsymbol{q}) \geq \mathrm{d}(\boldsymbol{f}(\boldsymbol{p}), \boldsymbol{f}(\boldsymbol{q}))$ for all $\boldsymbol{p}, \boldsymbol{q} \in \tilde{M}$. Deduce that if $\boldsymbol{f}$ is an isometry, then it is distance-preserving; i.e., $\tilde{\mathrm{d}}(\boldsymbol{p}, \boldsymbol{q}) = \mathrm{d}(\boldsymbol{f}(\boldsymbol{p}), \boldsymbol{f}(\boldsymbol{q}))$ for all $\boldsymbol{p}, \boldsymbol{q} \in \tilde{M}$. R. Palais has shown that the converse is also true: any distance-preserving map between manifolds is an isometry.

6.11. (a) According to Palais' result mentioned in the exercise above, any distance-preserving map $\boldsymbol{f} : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ in Euclidean space is a Euclidean motion. Prove this directly.
Hint: The map $\boldsymbol{x} \mapsto \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{0})$ preserves norms. Adapt the proof of Proposition 3.11.1 to conclude it is an orthogonal transformation.
(b) Let $\boldsymbol{f} : S^n \to S^n$ be distance-preserving. Show that $\boldsymbol{f}$ is the restriction of an orthogonal transformation.
Hint: Extend $\boldsymbol{f}$ radially to $\mathbb{R}^{n+1}$.

6.12. Show that for $M = S^1$, the distance function squared $\mathrm{d}^2 : M \times M \to \mathbb{R}$ is not differentiable exactly on the circle $\{(\boldsymbol{p}, -\boldsymbol{p}) \mid \boldsymbol{p} \in S^1\}$.

6.13. Give examples showing that a manifold in which any two points can be joined by a minimal geodesic is not necessarily complete.

6.14. Let M^n be a manifold in $\mathbb{R}^{n+k}$.
(a) Show that a Cauchy sequence in M is Cauchy as a sequence in $\mathbb{R}^{n+k}$ (i.e., with respect to the Euclidean distance).
(b) Prove that if M is closed in $\mathbb{R}^{n+k}$, then it is complete.

6.15. (a) Give examples showing that the converse to part (b) in the previous problem does not hold; i.e, there exist complete manifolds that are not closed as a subset of the ambient Euclidean space.
(b) Show that if M is a subset of $\mathbb{R}^n$ that is complete with the restriction of the distance function from $\mathbb{R}^n$ (i.e., $\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) = |\boldsymbol{p} - \boldsymbol{q}|$ for $\boldsymbol{p}, \boldsymbol{q} \in M$), then M is closed in the ambient Euclidean space.

6.16. Suppose $\boldsymbol{f} : M \to N$ is a diffeomorphism. If M is complete, is it always true that N is also complete? What if, in addition, $\boldsymbol{f}$ is an isometry?

6.17. M is said to be a *Riemannian homogeneous space* if its group of isometries acts transitively on M; i.e., given $\boldsymbol{p}, \boldsymbol{q} \in M$, there exists an isometry of M that maps $\boldsymbol{p}$ to $\boldsymbol{q}$.
(a) Give several examples of such spaces.
(b) Prove that a Riemannian homogeneous space is necessarily complete.

6.18. A *ray* in M is a normal geodesic $\boldsymbol{c} : [0, \infty) \to M$ that minimizes distance for all time; i.e., $\mathrm{d}(\boldsymbol{c}(0), \boldsymbol{c}(t)) = t$ for all $t \geq 0$.
(a) Prove that if M is complete and noncompact, then for any $\boldsymbol{p} \in M$, then there exists a ray $\boldsymbol{c}$ with $\boldsymbol{c}(0) = \boldsymbol{p}$. *Hint:* Consider a sequence $\boldsymbol{q}_n \in M$ with $\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}_n) > n$, $n \in \mathbb{N}$, and minimal normal geodesics $\boldsymbol{c}_n$ from $\boldsymbol{p}$ to $\boldsymbol{q}_n$. The sequence $\dot{\boldsymbol{c}}_n(0)$ has some convergent subsequence. If $\boldsymbol{v}$ is the limit, consider the geodesic in direction $\boldsymbol{v}$.
(b) Show by means of an example that completeness is necessary in part (a).

6.19. A point $\boldsymbol{p} \in M$ is said to be a *pole* of M if every normal geodesic emanating from $\boldsymbol{p}$ is a ray (see previous exercise). Let $f : \mathbb{R}^{n+1} \to \mathbb{R}$ be the function given by

$$f = u^{n+1} - \sum_{i=1}^{n} (u^i)^2,$$

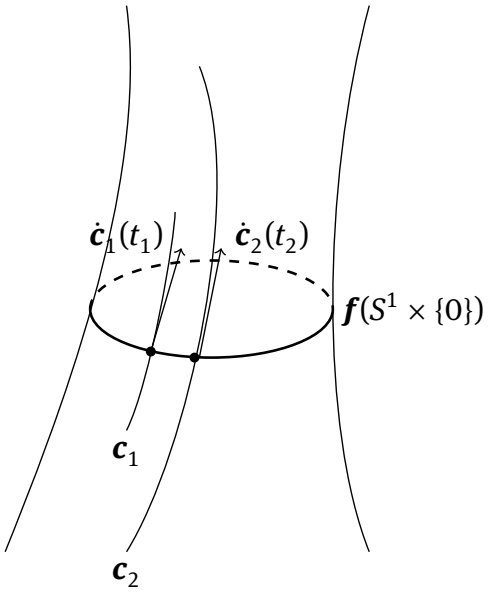
and consider the paraboloid $M = f^{-1}(0)$. Show that the origin is a pole of M. Are there any other poles?

6.20. A *line* in M is a normal geodesic $\boldsymbol{c} : (-\infty, \infty) \to M$ that minimizes distance for all time; i.e., $\mathrm{d}(\boldsymbol{c}(t_0)\boldsymbol{c}(t_1)) = |t_0 - t_1|$ for all t_0, t_1.
(a) If M admits a line, it must of course be noncompact. Show by means of examples, however, that there exist noncompact complete manifolds that admit no lines. This contrasts with rays, which exist in any noncompact, complete manifold, cf. Exercise 6.18.
(b) Recall from Chapter 5 that a given space can have different metrics, in the sense that there are diffeomorphisms which are not isometries. For example, the right

circular cylinder

$$S^1 \times \mathbb{R} = \{(x, y, z) \mid x^2 + y^2 = 1,\ z \in \mathbb{R}\}$$

is diffeomorphic to the hyperboloid $N = \{(x, y, z) \mid x^2 + y^2 - z^2 = 1\}$, but not isometric. Show however, that *any complete* manifold M diffeomorphic to $S^1 \times \mathbb{R}$ admits a line. *Hint:* Let $\boldsymbol{f} : S^1 \times \mathbb{R} \to M$ be a diffeomorphism, and consider a sequence of normal minimal geodesics $\boldsymbol{c}_n$ in M from $\boldsymbol{f}(1, 0, -n)$ to $\boldsymbol{f}(1, 0, n)$, $n \in \mathbb{N}$. Each $\boldsymbol{c}_n$ must intersect $\boldsymbol{f}(S^1 \times \{\boldsymbol{0}\})$ for a unique value $t_n \in \mathbb{R}$ of its parameter. Show that the sequence $\dot{\boldsymbol{c}}_n(t_n)$ has a convergent subsequence. If $\boldsymbol{u}$ is the limit of this subsequence, consider $t \mapsto \exp(t\boldsymbol{u})$.



6.21. Recall that the diameter of M is $\operatorname{diam}(M) = \sup\{\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) \mid \boldsymbol{p}, \boldsymbol{q} \in M\}$ if this set is bounded above, and ∞ otherwise.

(a) Show that if M is complete, then $\operatorname{diam}(M) < \infty$ if and only if M is compact.

(b) Prove that if M is compact, then there exist $\boldsymbol{p}, \boldsymbol{q} \in M$ such that $\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) = \operatorname{diam}(M)$.

6.22. Suppose M is a complete manifold, and denote by $T_1M = \{\boldsymbol{u} \in TM \mid |u| = 1\}$ the *unit tangent sphere bundle* of M. If $\pi : TM \to M$ is the tangent bundle projection, define

$$\begin{aligned} s : T_1M &\longrightarrow \mathbb{R}_+ \cup \{\infty\}, \\ \boldsymbol{v} &\longmapsto \sup\{t > 0 \mid \mathrm{d}(\pi(\boldsymbol{v}), \exp(t\boldsymbol{v})) = t\}. \end{aligned}$$

Roughly speaking, $s(\boldsymbol{v})$ is the largest parameter value for which the geodesic $\boldsymbol{c}_{\boldsymbol{v}}$ in direction $\boldsymbol{v}$ is minimal.

The *tangential cut locus* of $\boldsymbol{p} \in M$ is

$$C_{\boldsymbol{p}} = \{s(\boldsymbol{v})\boldsymbol{v} \mid \boldsymbol{v} \in M_{\boldsymbol{p}} \cap T_1M, \quad s(\boldsymbol{v}) < \infty\},$$

and the *cut locus* of $\boldsymbol{p} \in M$ is $C(\boldsymbol{p}) = \exp_{\boldsymbol{p}}(C_{\boldsymbol{p}})$. Determine the tangential cut locus and the cut locus at a generic point $\boldsymbol{p} \in M$, if

(a) $M = \mathbb{R}^n$;
(b) $M = S^n$;
(c) $M = S^1 \times \mathbb{R}$.

6.23. With the notation and terminology from Exercise 6.22, show that for $\boldsymbol{p}, \boldsymbol{q} \in M$, $\boldsymbol{q}$ lies in the cut locus of $\boldsymbol{p}$ if and only if $\boldsymbol{p}$ lies in the cut locus of $\boldsymbol{q}$.

6.24. With the notation and terminology from Exercise 6.22, show that the distance $\mathrm{d}(\boldsymbol{p}, C(\boldsymbol{p}))$ from a point to its cut locus equals the injectivity radius $\mathrm{inj}_{\boldsymbol{p}}$ at that point. *Hint:* Notice that

$$\mathrm{d}(\boldsymbol{p}, C(\boldsymbol{p})) = \inf\{s(\boldsymbol{v}) \mid \boldsymbol{v} \in T_1 M \cap M_{\boldsymbol{p}}\}.$$

Argue by contradiction to rule out the possibilities $\mathrm{d}(\boldsymbol{p}, C(\boldsymbol{p})) > \mathrm{inj}_{\boldsymbol{p}}$ and $\mathrm{d}(\boldsymbol{p}, C(\boldsymbol{p})) < \mathrm{inj}_{\boldsymbol{p}}$.
This fact can be used to prove that in general, the function $\mathrm{inj} : M \to \mathbb{R}_+ \cup \{\infty\}$ is continuous.

6.25. A subset A of a manifold M is said to be *convex* if any $\boldsymbol{p}, \boldsymbol{q} \in A$ can be joined by some geodesic $\boldsymbol{c}$ contained in A, with length $L(\boldsymbol{c}) = \mathrm{d}(\boldsymbol{p}, \boldsymbol{q})$. If, in addition, this geodesic is unique, then A is said to be *strongly convex*.
(a) Show that any convex subset of S^n is either all of S^n or else is contained in some hemisphere.
(b) Prove that for any manifold M and $\boldsymbol{p} \in M$ there exists $\varepsilon > 0$ such that the metric ball $B_\varepsilon(\boldsymbol{p})$ of radius ε about $\boldsymbol{p}$ is strongly convex. *Hint*: see Proposition 6.5.1.

6.26. This exercise examines the gradient and Hessian of the distance function from a point.
(a) Prove that for $\boldsymbol{p} \in M$, there exists $\varepsilon > 0$ such that the distance function $\boldsymbol{q} \mapsto \mathrm{d}(\boldsymbol{p}, \boldsymbol{q})$ from $\boldsymbol{p}$ is smooth on $W := B_\varepsilon(\boldsymbol{p}) \setminus \{\boldsymbol{p}\}$. Is it differentiable on all of $B_\varepsilon(\boldsymbol{p})$?
(b) Let f denote the restriction of this function to W, $f(\boldsymbol{q}) = \mathrm{d}(\boldsymbol{p}, \boldsymbol{q})$, $\boldsymbol{q} \in W$, and $\boldsymbol{c} : [0, 1] \to W$ a (necessarily minimal) geodesic from $\boldsymbol{p}$ to $\boldsymbol{q}$. If $\boldsymbol{u} \in M_{\boldsymbol{q}}$ and $\boldsymbol{c}_{\boldsymbol{u}}$ is the geodesic $\boldsymbol{u} \mapsto \exp(t\boldsymbol{u})$, prove that there exists a variation $V : [0, 1] \times (-\delta, \delta) \to W$ of $\boldsymbol{c}$ such that V_s is the minimal geodesic from $\boldsymbol{p}$ to $\boldsymbol{c}_{\boldsymbol{u}}(s)$ if δ is small enough.
(c) Let L denote the length function of V, $L(s)$ = length of V_s. Show that

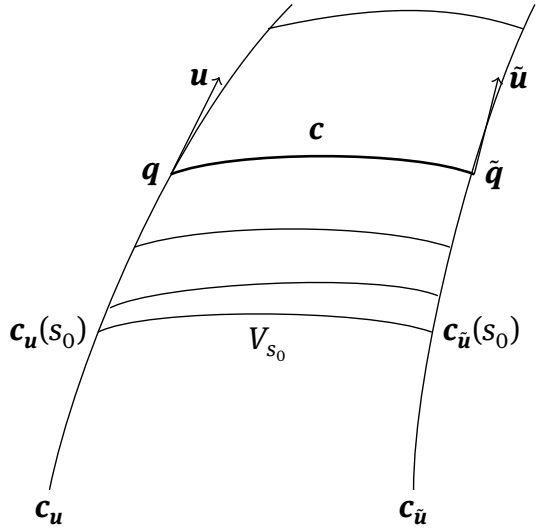
$$\langle \nabla f(\boldsymbol{q}), \boldsymbol{u} \rangle = (f \circ \boldsymbol{c}_{\boldsymbol{u}})'(0) = L'(0) = \langle \dot{\boldsymbol{c}}(1), \boldsymbol{u} \rangle,$$

so that $\nabla f(\boldsymbol{q}) = \dot{\boldsymbol{c}}(1)$, and

$$h_f(\boldsymbol{q})(\boldsymbol{u}, \boldsymbol{u}) = (f \circ \boldsymbol{c}_{\boldsymbol{u}})''(0) = L''(0) = \langle \boldsymbol{Y}, \boldsymbol{Y}' \rangle(1),$$

where $\boldsymbol{Y}$ is the Jacobi field along $\boldsymbol{c}$ with $\boldsymbol{Y}(0) = \boldsymbol{0}$, and $\boldsymbol{Y}(1) = \boldsymbol{u}$.

6.27. Let $\boldsymbol{p} \in M$, $\varepsilon > 0$ such that the square of the distance function $\mathrm{d}^2 : M \times M \to \mathbb{R}$, when restricted to $B_\varepsilon(\boldsymbol{p}) \times B_\varepsilon(\boldsymbol{p})$, is differentiable. Denote this restriction by f. The goal of this problem is to determine the gradient ∇f and the Hessian H_f of f.



Let $\boldsymbol{q}, \tilde{\boldsymbol{q}} \in B_\varepsilon(\boldsymbol{p})$, $\boldsymbol{c} : [0,1] \to B_\varepsilon(\boldsymbol{p})$ the geodesic from $\boldsymbol{q}$ to $\tilde{\boldsymbol{q}}$ with length $\mathrm{d}(\boldsymbol{q}, \tilde{\boldsymbol{q}})$, and $\boldsymbol{u} \in M_{\boldsymbol{q}}$, $\tilde{\boldsymbol{u}} \in M_{\tilde{\boldsymbol{q}}}$. If $\boldsymbol{c}_{\boldsymbol{u}}$, $\boldsymbol{c}_{\tilde{\boldsymbol{u}}}$ denote the geodesics in direction $\boldsymbol{u}$ and $\tilde{\boldsymbol{u}}$ respectively, strong convexity (see Exercise 6.25) guarantees the existence of a variation $V : [0,1] \times (-\delta, \delta) \to B_\varepsilon(\boldsymbol{p})$ of $\boldsymbol{c}$ by geodesics, with V_s minimal from $\boldsymbol{c}_{\boldsymbol{u}}(s)$ to $\boldsymbol{c}_{\tilde{\boldsymbol{u}}}(s)$, $|s| < \delta$, provided δ is small enough. Then

$$\langle \nabla f(\boldsymbol{q}, \tilde{\boldsymbol{q}}), (\boldsymbol{u}, \tilde{\boldsymbol{u}}) \rangle = D_{(\boldsymbol{u}, \tilde{\boldsymbol{u}})} f = h'(0),$$

where $h(s) = L(V_s)$.

(a) Use this to show that $\nabla f(\boldsymbol{q}, \tilde{\boldsymbol{q}}) = 2(-\dot{\boldsymbol{c}}(0), \dot{\boldsymbol{c}}(1))$.

(b) Use the variation from above to show that for $\boldsymbol{q}, \tilde{\boldsymbol{q}} \in \mathcal{B}_\varepsilon(\boldsymbol{p})$,

$$H_f(\boldsymbol{q}, \tilde{\boldsymbol{q}})(\boldsymbol{u}, \tilde{\boldsymbol{u}}) = 2(-\boldsymbol{Y}'(0), \boldsymbol{Y}'(1)), \qquad \boldsymbol{u} \in M_{\boldsymbol{q}}, \quad \tilde{\boldsymbol{u}} \in M_{\tilde{\boldsymbol{q}}},$$

where $\boldsymbol{Y}$ is the Jacobi field along $\boldsymbol{c}$ with $\boldsymbol{Y}(0) = \boldsymbol{u}$, $\boldsymbol{Y}(1) = \tilde{\boldsymbol{u}}$.
Deduce that the Hessian form h_f of f satisfies

$$h_f(\boldsymbol{q}, \tilde{\boldsymbol{q}})((\boldsymbol{u}, \tilde{\boldsymbol{u}}), (\boldsymbol{u}, \tilde{\boldsymbol{u}})) = 2\langle \boldsymbol{Y}, \boldsymbol{Y}' \rangle|_0^1.$$

(c) Compute the gradient, Hessian, and Hessian form of g, where g is the square of the distance function from $\boldsymbol{p}$, $g(\boldsymbol{q}) = \mathrm{d}^2(\boldsymbol{p}, \boldsymbol{q})$, on $B_\varepsilon(\boldsymbol{p})$. *Hint:* An easy way to do this is to notice that $g = f \circ \iota_{\boldsymbol{p}}$, where $\iota_{\boldsymbol{p}} : M \to M \times M$ maps $\boldsymbol{q}$ to $(\boldsymbol{p}, \boldsymbol{q})$.

6.28. This problem, which uses concepts and results from Exercises 6.25 and 6.27, explores a result of J. H. C. Whitehead; namely, for any point $\boldsymbol{p}$ in a manifold M, there exists some $\varepsilon > 0$ such that every metric ball $B_\delta(\boldsymbol{q})$ contained in $B_\varepsilon(\boldsymbol{p})$ is strongly convex.

(a) Let $\varepsilon_0 > 0$ such that the closure K of the ball of radius ε_0 centered at $\boldsymbol{p}$ is compact, and consider the compact set

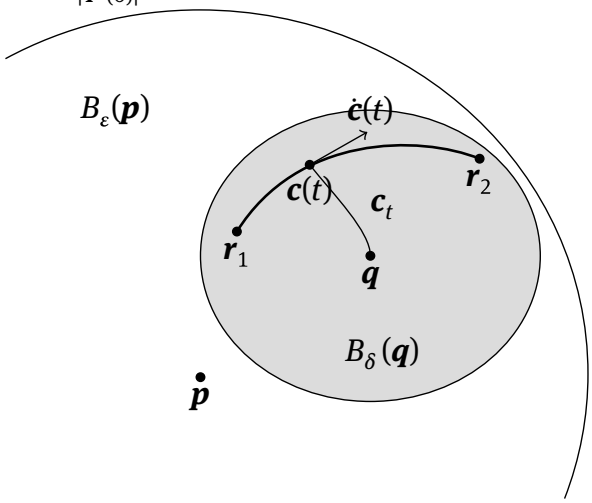
$$C = \{(\boldsymbol{u}, \boldsymbol{v}) \in TM \times TM \mid |\boldsymbol{u}| = |\boldsymbol{v}| = 1, \quad \pi(\boldsymbol{u}) = \pi(\boldsymbol{v}) \in K\},$$

where $\pi : TM \to M$ is the bundle projection. For $(\boldsymbol{u}, \boldsymbol{v}) \in C$, denote by $\boldsymbol{Y_{uv}}$ the Jacobi field along $t \mapsto \exp(t\boldsymbol{v})$ with $\boldsymbol{Y_{uv}}(0) = \boldsymbol{0}$, $\boldsymbol{Y'_{uv}}(0) = \boldsymbol{u}$. Use the fact that

$$\langle \boldsymbol{Y_{uv}}, \boldsymbol{Y'_{uv}} \rangle'(0) = |\boldsymbol{u}|^2 = 1$$

to show that there exists $\varepsilon \in (0, \varepsilon_0)$ such that $\langle \boldsymbol{Y_{uv}}, \boldsymbol{Y'_{uv}} \rangle|_0^t > 0$ for any $0 < t \leq \varepsilon$ and $(\boldsymbol{u}, \boldsymbol{v}) \in C$.

(b) Prove that for any $\boldsymbol{p} \in M$, there exists some positive ε such that
   (1) $B_\varepsilon(\boldsymbol{p})$ is strongly convex, and the square of the distance function on M, when restricted to $B_\varepsilon(\boldsymbol{p}) \times B_\varepsilon(\boldsymbol{p})$, is differentiable.
   (2) For any $\boldsymbol{q} \in B_\varepsilon(\boldsymbol{p})$ and any geodesic $\boldsymbol{c} : [0,1] \to B_\varepsilon(\boldsymbol{p})$ with $\boldsymbol{c}(0) = \boldsymbol{q}$, if $\boldsymbol{Y}$ is a Jacobi field along $\boldsymbol{c}$ with $\boldsymbol{Y}(0) = \boldsymbol{0}$, $\boldsymbol{Y}'(0) \neq \boldsymbol{0}$, then $\langle \boldsymbol{Y}, \boldsymbol{Y}' \rangle|_0^t > 0$ for all $t \in (0,1]$. *Hint*: $\boldsymbol{Z} = \frac{1}{|\boldsymbol{Y}'(0)|}\boldsymbol{Y}$ is a Jacobi field satisfying $\boldsymbol{Z}(0) = \boldsymbol{0}$, $|\boldsymbol{Z}'(0)| = 1$.



(c) Show that Whitehead's result holds for the ε obtained in (b), as follows: let $\boldsymbol{q} \in B_\varepsilon(\boldsymbol{p})$ and $\delta > 0$ such that $B_\delta(\boldsymbol{q}) \subset B_\varepsilon(\boldsymbol{p})$. Since a convex subset of a strongly convex set is strongly convex, it suffices to show that the ball of radius δ about $\boldsymbol{q}$ is convex. Denote by $h : B_\varepsilon(\boldsymbol{p}) \to \mathbb{R}$ the distance function from $\boldsymbol{q}$ squared, $h(\boldsymbol{r}) = \mathrm{d}^2(\boldsymbol{q}, \boldsymbol{r})$. Given $\boldsymbol{r}_1, \boldsymbol{r}_2 \in B_\delta(\boldsymbol{q})$, there is a unique geodesic $\boldsymbol{c} : [0,1] \to B_\varepsilon(\boldsymbol{p})$ with $\boldsymbol{c}(0) = \boldsymbol{r}_1$, $\boldsymbol{c}(1) = \boldsymbol{r}_2$. Prove that

$$(h \circ \boldsymbol{c})''(t) = 2b_h(\boldsymbol{c}(t))(\dot{\boldsymbol{c}}(t), \dot{\boldsymbol{c}}(t)) = 2\langle \boldsymbol{Y}_t, \boldsymbol{Y}'_t \rangle|_0^1,$$

where b_h is the Hessian form of h, and $\boldsymbol{Y}$ is the Jacobi field along the minimal geodesic $\boldsymbol{c}_t : [0,1] \to B_\varepsilon(\boldsymbol{p})$ joining $\boldsymbol{q}$ to $\boldsymbol{c}(t)$, with $\boldsymbol{Y}_t(0) = \boldsymbol{0}$, $\boldsymbol{Y}_t(1) = \dot{\boldsymbol{c}}(t)$. In particular, $h \circ \boldsymbol{c}$ is a convex function. Deduce that $(h \circ \boldsymbol{c})(t) < \delta^2$ for all t, and conclude that $B_\delta(\boldsymbol{q})$ is convex.

6.29. A manifold is said to be *locally symmetric* if its curvature tensor is parallel; i.e., if for any geodesic $\boldsymbol{c}$ and parallel fields $\boldsymbol{X}_i$ along $\boldsymbol{c}$, $1 \leq i \leq 3$, $R(\boldsymbol{X}_1, \boldsymbol{X}_2)\boldsymbol{X}_3$ is a parallel field along $\boldsymbol{c}$.

(a) Prove that a space of constant curvature is locally symmetric.

(b) Let M be a locally symmetric space, $\boldsymbol{c}$ a normal geodesic in M. Show that there exists an orthonormal basis of parallel vector fields $\boldsymbol{E}_i$ along $\boldsymbol{c}$, $1 \leq i \leq n$, with

$\boldsymbol{E}_n = \dot{\boldsymbol{c}}$, and constants λ_i such that $R(\boldsymbol{E}_i, \dot{\boldsymbol{c}})\dot{\boldsymbol{c}} = \lambda_i \boldsymbol{E}_i$. Notice that λ_i is the sectional curvature of the plane spanned by $\dot{\boldsymbol{c}}(t)$ and $\boldsymbol{E}_i(t)$.

(c) Prove that $\boldsymbol{Y}$ is a Jacobi field along $\boldsymbol{c}$ if and only if

$$\boldsymbol{Y} = \sum_{i=0}^{n} f_i \boldsymbol{E}_i, \text{ where } f_i \text{ satisfies } f_i'' + \lambda_i f_i = 0.$$

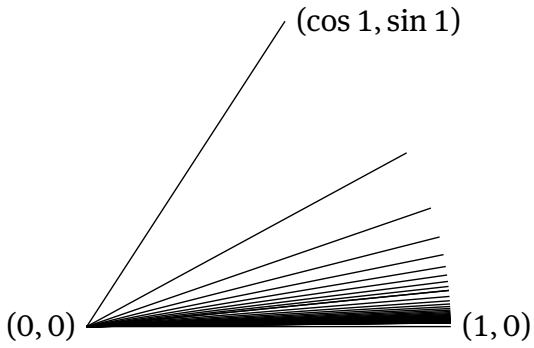
6.30. The definition of distance that was used for a manifold M,

$$\mathrm{d}(\boldsymbol{p}, \boldsymbol{q}) = \inf\{L(\boldsymbol{c}) \mid \boldsymbol{c} : [0,1] \to M,\ \boldsymbol{c}(0) = \boldsymbol{p},\ \boldsymbol{c}(1) = \boldsymbol{q}\},$$

makes sense for more general connected subsets of Euclidean space that are not manifolds, see for example [6].

(a) Show that the boundary of any square in $\mathbb{R}^2$ is a well-defined metric space with the distance above, and that the open sets in the metric space coincide with the usual open sets as in Theorem 6.5.1.

(b) Denote by L_0 the line segment in $\mathbb{R}^2$ connecting the origin to $(1,0)$, by L_n the line segment connecting the origin to $(\cos(1/n), \sin(1/n))$, $n \in \mathbb{N}$, and let $M = \cup_{n=0}^{\infty} L_n$, together with the above distance. Prove that the open sets in M are not the usual ones. *Hint:* consider metric balls centered about, say, $(1/2, 0)$.



7 Hypersurfaces

7.1 Hypersurfaces and orientation

A *hypersurface* is an n-dimensional submanifold of $\mathbb{R}^{n+1}$ for some n. The case $n = 2$, which corresponds to surfaces in 3-dimensional space, was historically the first to be studied. Most of the concepts introduced in Chapter 3 are easier to understand and work with when the manifold's dimension is one less than that of the ambient Euclidean space.

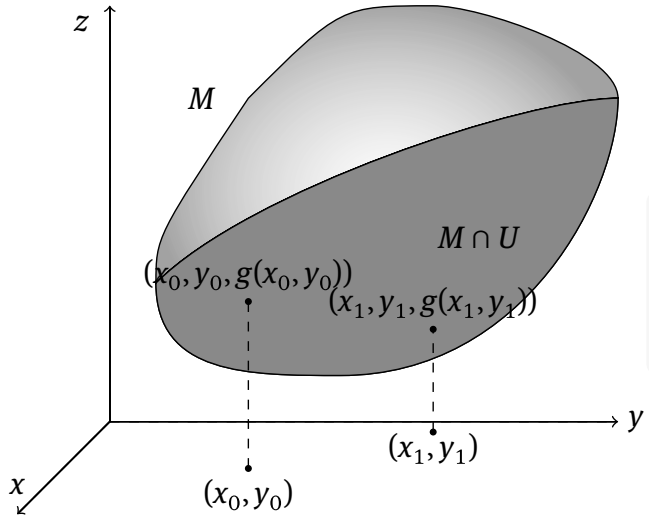
The most useful way of describing a hypersurface is by means of Corollary 3.2.1, which we recall in the present context:

Proposition 7.1.1. *$M^n \subset \mathbb{R}^{n+1}$ is a hypersurface if and only if any $\boldsymbol{p} \in M$ admits a neighborhood $U \subset \mathbb{R}^{n+1}$ and a function $f : U \to \mathbb{R}$ having 0 as a regular value, such that $M \cap U = f^{-1}(0)$.*

For example, in the above situation, the vector field ∇f is normal – i.e., orthogonal – to M on $M \cap U$: by Proposition 3.1.1, given $\boldsymbol{q} \in M \cap U$, $M_{\boldsymbol{q}} = \ker f_{*\boldsymbol{q}}$, and the latter is just $\nabla f(\boldsymbol{q})^{\perp}$.

Proposition 7.1.2. *A hypersurface is orientable if and only if it admits a unit length normal vector field.*

Proof. If the hypersurface is orientable, it admits an orientation. Let ω denote the standard volume form on $\mathbb{R}^n$. If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n-1}$ is a local positive basis of vector fields on $U \subset M$, there is a unique unit normal field $\boldsymbol{N}$ on U such that $(i(\boldsymbol{N})\omega)(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n-1}) = \omega(\boldsymbol{N}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n-1}) > 0$. Since this can be done in a neighborhood of any point, $\boldsymbol{N}$ is globally defined. Conversely, if $\boldsymbol{N}$ is a unit normal field on M, then the map which assigns to $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1} \in M_{\boldsymbol{p}}$ the value $\omega(\boldsymbol{p})(\boldsymbol{N}(\boldsymbol{p}), \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1})$ is a nowhere-zero $(n-1)$-form on M; i.e., an orientation. □



The darker half $M \cap U$ of M is the graph of a function g of two variables; it equals $f^{-1}(0)$, where $f(x, y, z) = z - g(x, y)$.

In light of the above, we may call the normal vector field an *orientation* of M. Clearly, if M is orientable, then it has two possible orientations, namely $\boldsymbol{N}$ and $-\boldsymbol{N}$, where $\boldsymbol{N}$ is any unit normal field. A large class of orientable hypersurfaces is given by the collection of functions $f : \mathbb{R}^{n+1} \to \mathbb{R}$ that have zero as regular value, since each $M = f^{-1}(0)$ is a hypersurface with orientation $\boldsymbol{N} = \nabla f/|\nabla f|$.

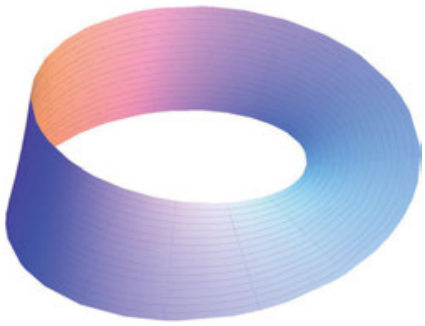


Fig. 7.1: A Möbius strip

One example of a non orientable surface is the Möbius strip. It can be described as the image of the map

$$\begin{aligned} \boldsymbol{h} : (-1/2, 1/2) \times [0, 2\pi] &\to \mathbb{R}^3 \\ (r, \theta) &\mapsto \left(2\cos\theta + r\cos\frac{\theta}{2}, 2\sin\theta + r\cos\frac{\theta}{2}, r\sin\frac{\theta}{2}\right). \end{aligned}$$

$\boldsymbol{h}$ is not, strictly speaking, a parametrization, but it is one locally, since it has maximal rank everywhere. More precisely, restrict $\boldsymbol{h}$ to $(-1/2, 1/2) \times (0, 2\pi)$, and define $\boldsymbol{k} : (-1/2, 1/2) \times (-\pi, \pi) \to \mathbb{R}^3$ by the same formula. $\boldsymbol{h}$ and $\boldsymbol{k}$ are both parametrizations whose images cover the strip M. One obtains normal vector fields along these parametrizations by setting

$$\boldsymbol{N}_{\boldsymbol{h}} = \frac{\partial}{\partial x^1} \times \frac{\partial}{\partial x^2},$$

where $\boldsymbol{x} = \boldsymbol{h}^{-1}$, and using a similar formula for $\boldsymbol{N}_{\boldsymbol{k}}$. Notice that both fields agree on the intersection $(-1/2, 1/2) \times (0, \pi)$ of their domains, since they are given by the same formula. They do not, however, combine to give us a well-defined normal field on M, because, for example,

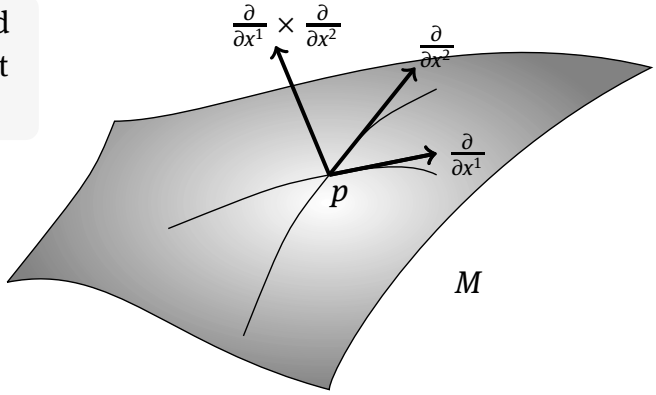
$$\boldsymbol{h}\left(-r, \frac{3\pi}{2}\right) = \boldsymbol{k}\left(r, -\frac{\pi}{2}\right), \text{ but } \boldsymbol{N}_{\boldsymbol{h}}\left(-r, \frac{3\pi}{2}\right) = -\boldsymbol{N}_{\boldsymbol{k}}\left(r, -\frac{\pi}{2}\right),$$

as is easily checked.

The Möbius strip is often realized by taking a long rectangular strip of paper. If one glues the shorter sides together so that the top vertices (and similarly the bottom ones) coincide, one obtains a cylinder. To obtain the Möbius strip instead, the shorter sides are glued in the opposite direction, so that the top left vertex is identified with the bottom right one, and similarly the bottom left is glued to the top right vertex. From a differential geometric perspective, though, this surface is different from the one we

described earlier: the paper strip was not stretched or bent, so it is flat, whereas the parametrized strip has negative curvature, as we shall soon see.

Normal vector induced by a chart $\boldsymbol{x}$ at a point $p \in M$.



The cross product can be generalized to Euclidean spaces of arbitrary dimension, and provides further insight into the concept of orientation. Given any n vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n \in \mathbb{R}^{n+1}$, the map

$$\boldsymbol{u} \mapsto \det(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n, \boldsymbol{u})$$

is a one-form on $\mathbb{R}^{n+1}$, and by Corollary 1.4.2, there exists a unique $\boldsymbol{w} \in \mathbb{R}^{n+1}$ such that

$$\langle \boldsymbol{w}, \boldsymbol{u} \rangle = \det(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n, \boldsymbol{u}).$$

This vector $\boldsymbol{w}$ is denoted $\boldsymbol{v}_1 \times \cdots \times \boldsymbol{v}_n$, and is called the *cross product* of $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$, cf. also Exercise 5.12. By the properties of the determinant, the cross product is nonzero if and only if the vectors are linearly independent, it changes sign whenever $\boldsymbol{v}_i$ is interchanged with $\boldsymbol{v}_j$ $(i \neq j)$, and $\boldsymbol{v}_1 \times \cdots \times \boldsymbol{v}_n \perp \boldsymbol{v}_i$ for $1 \leq i \leq n$. Given $\boldsymbol{p} \in \mathbb{R}^{n+1}$, the canonical isomorphism $\mathcal{I}_{\boldsymbol{p}} : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}_{\boldsymbol{p}}$ extends this cross product to any tangent space.

Now, if M is a hypersurface, then any local parametrization $\boldsymbol{h} = \boldsymbol{x}^{-1} : U \to M$ of M induces an orientation of the manifold $\boldsymbol{h}(U)$, since the vector field

$$\boldsymbol{n}_{\boldsymbol{h}} := \left(\frac{1}{|\frac{\partial}{\partial x^1} \times \cdots \frac{\partial}{\partial x^n}|} \right) \frac{\partial}{\partial x^1} \times \cdots \times \frac{\partial}{\partial x^n}$$

is a unit normal field on $\boldsymbol{h}(U)$. If M is oriented, the parametrization $\boldsymbol{h}$ is said to be *consistent with the orientation* if $\boldsymbol{n}_{\boldsymbol{h}}$ represents the given orientation. Notice that if $\boldsymbol{h}$ is not consistent with the orientation, then interchanging any two component functions of $\boldsymbol{h}$ yields a consistent parametrization.

Evidently, M is orientable if and only if it admits an atlas such that any two parametrizations $\boldsymbol{h}$ and $\boldsymbol{k}$ in the atlas satisfy $\boldsymbol{n}_{\boldsymbol{h}} = \boldsymbol{n}_{\boldsymbol{k}}$ on the intersection of their domains. In the exercises, the reader is asked to show directly that this is equivalent to our previous criterion for orientability, namely requiring that the Jacobian matrix $D(\boldsymbol{h}^{-1} \circ \boldsymbol{k})$ has positive determinant.

7.2 The Gauss map and the second fundamental form

We now associate to each oriented hypersurface M^n a map $\boldsymbol{\gamma} : M \to S^n$ which keeps track of how curved M is in $\mathbb{R}^{n+1}$. Let M^n be an oriented hypersurface with unit normal field $\boldsymbol{n}$, and denote by $\pi_2 : T\mathbb{R}^{n+1} = \mathbb{R}^{n+1} \times \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ the projection onto the second factor. Thus, π_2 is the left inverse of $\mathcal{I}_{\boldsymbol{p}}$ for any $\boldsymbol{p} \in \mathbb{R}^{n+1}$: $\pi_2 \circ \mathcal{I}_{\boldsymbol{p}} = 1_{\mathbb{R}^{n+1}}$. The *Gauss map* of the oriented hypersurface is defined to be

$$\boldsymbol{\gamma} = \pi_2 \circ \boldsymbol{n} : M \to S^n. \tag{7.2.1}$$

Loosely speaking, the Gauss map assigns to each point of M the unit normal vector at that point, parallel translated back to the origin. The image of the Gauss map therefore measures how much the hypersurface differs from a hyperplane:

Examples 7.2.1. (i) A *hyperplane* in $\mathbb{R}^{n+1}$ is a subset of the form $M = \{\boldsymbol{p} \in \mathbb{R}^{n+1} \mid \langle \boldsymbol{p}, \boldsymbol{u} \rangle = a\}$, where $\boldsymbol{u}$ is any fixed nonzero element of $\mathbb{R}^{n+1}$, and a is some real number. When $a = 0$, M is the n-dimensional subspace $\boldsymbol{u}^\perp$. When $a \neq 0$, it is the subspace $\boldsymbol{u}^\perp$ parallel translated to any $\boldsymbol{p}_0 \in M$. Indeed, $\langle \boldsymbol{p} - \boldsymbol{p}_0, \boldsymbol{u} \rangle = 0$ if and only if $\langle \boldsymbol{p}, \boldsymbol{u} \rangle = \langle \boldsymbol{p}_0, \boldsymbol{u} \rangle$, and the latter equals a if $\boldsymbol{p}_0 \in M$. The Gauss map of such a hyperplane has a single point as image, namely $\boldsymbol{u}/|\boldsymbol{u}|$ (or its negative, depending on the orientation).

(ii) If $M^2 \subset \mathbb{R}^3$ denotes the cylinder $\{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\}$ with the outward orientation, then the Gauss map of M is given by $\boldsymbol{\gamma}(x, y, z) = (x, y)$. In particular, its image consists of the equator in S^2.

(iii) Let $M^2 \subset \mathbb{R}^3$ be the paraboloid consisting of all (x, y, z) satisfying $z = x^2 + y^2$. $M = g^{-1}(0)$, where $g : \mathbb{R}^3 \to \mathbb{R}$, $g(x, y, z) = x^2 + y^2 - z$, has zero as regular value. By the discussion from the previous section, the normalized gradient of g yields a unit normal field

$$\boldsymbol{n} = \frac{1}{\sqrt{1 + 4((u^1)^2 + (u^2)^2)}} \left(2u^1 D_1 + 2u^2 D_2 - D_3\right).$$

Since $M = \{(\boldsymbol{p}, |\boldsymbol{p}|^2) \in \mathbb{R}^2 \times \mathbb{R} \mid \boldsymbol{p} \in \mathbb{R}^2\}$, the Gauss map of M for this orientation may be described as

$$\begin{aligned} \boldsymbol{\gamma} : M \subset \mathbb{R}^2 \times \mathbb{R} &\to S^2 \subset \mathbb{R}^2 \times \mathbb{R}, \\ (\boldsymbol{p}, |\boldsymbol{p}|^2) &\mapsto \frac{1}{\sqrt{1 + 4|\boldsymbol{p}|^2}} (2\boldsymbol{p}, -1). \end{aligned}$$

It is clear from the formula that the image of $\boldsymbol{\gamma}$ is contained in the open southern hemisphere of S^2. The image is actually the whole open hemisphere: indeed, the function $t \mapsto 2t/(\sqrt{1 + 4t^2})$ is easily seen to be a diffeomorphism from $\mathbb{R}$ onto $(-1, 1)$: in fact, it has positive derivative everywhere so that it is strictly increasing. By the inverse function theorem, it is a diffeomorphism onto its image; finally, it

approaches 1 as $t \to \infty$, and is an odd function. This establishes the claim. In particular, the map

$$\begin{aligned} \boldsymbol{f} : \mathbb{R}^2 &\to B_1(0) \subset \mathbb{R}^2, \\ \boldsymbol{p} &\mapsto \frac{2}{\sqrt{1+4|\boldsymbol{p}|^2}}\boldsymbol{p} \end{aligned}$$

sends the plane onto the open disk of radius 1 centered at the origin (it is, in fact, a diffeomorphism). The claim follows, since the projection of $\boldsymbol{\gamma}(\boldsymbol{p}, |\boldsymbol{p}|^2)$ onto the xy-plane is precisely $\boldsymbol{f}(\boldsymbol{p})$.

(iv) For the sphere $S^n(r) = \{\boldsymbol{p} \in \mathbb{R}^{n+1} \mid |\boldsymbol{p}| = r\}$ of radius $r > 0$, the Gauss map is given by

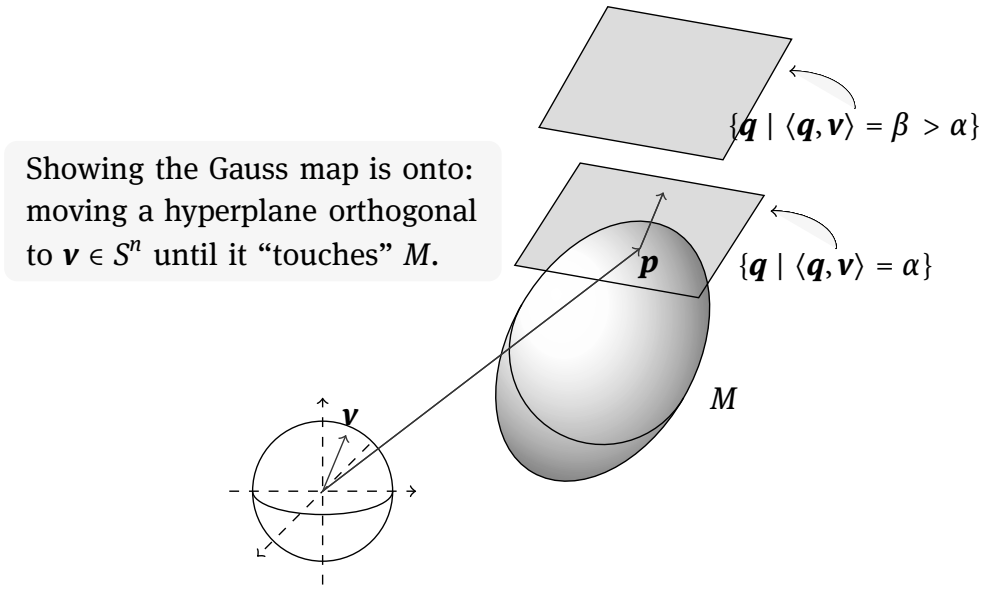
$$\begin{aligned} \boldsymbol{\gamma} : S^n(r) &\to S^n(1), \\ \boldsymbol{p} &\mapsto \frac{1}{r}\boldsymbol{p}, \end{aligned}$$

or its negative, depending on the orientation. In particular, it is onto.

The trivial fact that the Gauss map of a sphere is surjective generalizes as follows:

Theorem 7.2.1. *Let $f : \mathbb{R}^{n+1} \to \mathbb{R}$ be a function that has $\alpha \in \mathbb{R}$ as a regular value, and suppose that $M = f^{-1}(\alpha)$ is nonempty. If M is compact, then the image of its Gauss map is the whole sphere.*

Proof. Although there are some technical details that must be dealt with, the idea of the proof is simple: orient M so that its unit normal field points "outward". Given $\boldsymbol{v} \in S^n$, consider a hyperplane $\{\boldsymbol{q} \mid \langle \boldsymbol{q}, \boldsymbol{v} \rangle = \beta\}$ orthogonal to $\boldsymbol{v}$ that is sufficiently far away (i.e., β large) that it does not intersect the hypersurface. Parallel translate it until it hits M. At the point $\boldsymbol{p}$ of intersection, this hyperplane will coincide with the tangent plane at $\boldsymbol{p}$, so that $\boldsymbol{\gamma}(\boldsymbol{p}) = \boldsymbol{v}$.



Showing the Gauss map is onto: moving a hyperplane orthogonal to $\boldsymbol{v} \in S^n$ until it "touches" M.

Now, for the proof proper: If M_α denotes the set of all $\boldsymbol{p}$ such that $f(\boldsymbol{p}) < \alpha$, and M^α the set of those satisfying $f(\boldsymbol{p}) > \alpha$, then the complement $\mathbb{R}^{n+1} \setminus M$ of M equals the disjoint union of M_α and M^α. Notice that one of these must be compact, and the other noncompact. They cannot both be compact, since the ambient space isn't. If they were both noncompact, we could choose some ball B containing M. $\mathbb{R}^{n+1} \setminus B$ would then contain at least one point from M_α and one from M^α; being connected, it would also contain some continuous curve joining the two. This contradicts the intermediate value theorem, which guarantees that any such curve must intersect M. We may, without loss of generality, assume that M_α is compact. Given $\boldsymbol{v} \in S^n$, let $\boldsymbol{q} \in M$ be a point where the function g, $g(\boldsymbol{p}) = \langle \boldsymbol{p}, \boldsymbol{v} \rangle$, takes on its maximum value when restricted to M. By the method of Lagrange multipliers,

$$\nabla f(\boldsymbol{q}) = \lambda \nabla g(\boldsymbol{q}) = \lambda \mathcal{I}_{\boldsymbol{q}} \boldsymbol{v}$$

for some $\lambda \in \mathbb{R}$.
In order to conclude that $(\nabla f / |\nabla f|)\boldsymbol{q} = \mathcal{I}_{\boldsymbol{q}} \boldsymbol{v}$, we only need to show that $\lambda \geq 0$ (notice that λ is nonzero by assumption). To see this, observe that for any $t > 0$, $\langle \boldsymbol{q} + t\boldsymbol{v}, \boldsymbol{v} \rangle > \langle \boldsymbol{q}, \boldsymbol{v} \rangle$, so $\boldsymbol{q} + t\boldsymbol{v} \notin M$, and it either belongs to M_α or else to M^α for all $t > 0$. Since M_α is compact, it must belong to the latter. Therefore,

$$t \overset{h}{\mapsto} \frac{f(\boldsymbol{q} + t\boldsymbol{v}) - f(\boldsymbol{q})}{t} > 0, \qquad t \neq 0,$$

and

$$0 \leq \lim_{t \to 0} h(t) = \langle \nabla f(\boldsymbol{q}), \mathcal{I}_{\boldsymbol{q}} \boldsymbol{v} \rangle = \lambda,$$

which establishes the claim. □

For a hypersurface, the second fundamental tensor is unique up to sign, since there are exactly two unit normal vectors at any point, with one being the negative of the other. When the hypersurface is oriented by a unit normal vector field $\boldsymbol{n}$, any sign ambiguity disappears, and we define the second fundamental tensor to be $S = S_{\boldsymbol{n}}$. The next result says that up to parallel translation, the second fundamental tensor equals minus the derivative of the Gauss map:

Theorem 7.2.2. *Let $\boldsymbol{n}$ denote the unit normal field on M determined by some orientation of the hypersurface M, and $\gamma : M \to S^n$ the corresponding Gauss map. If S denotes the second fundamental tensor of M, then*

$$\pi_2 \circ S = -\pi_2 \circ \boldsymbol{\gamma}_*.$$

Proof. The argument is essentially an exercise in notation: for $\boldsymbol{p} \in M$, $\boldsymbol{x} \in M_{\boldsymbol{p}}$, let $\boldsymbol{c}$ be a curve in M with $\dot{\boldsymbol{c}}(0) = \boldsymbol{x}$. By Definitions 2.8.6 and 2.8.7,

$$S\boldsymbol{x} = -D_{\boldsymbol{x}} \boldsymbol{n} = -\mathcal{I}_{\boldsymbol{p}} (\pi_2 \circ \boldsymbol{n} \circ \boldsymbol{c})'(0) = -\mathcal{I}_{\boldsymbol{p}} (\boldsymbol{\gamma} \circ \boldsymbol{c})'(0),$$

so that

$$(\pi_2 \circ S)\boldsymbol{x} = -(\boldsymbol{\gamma} \circ \boldsymbol{c})'(0) = -\pi_2(\boldsymbol{\gamma}_* \dot{\boldsymbol{c}}(0)) = -(\pi_2 \circ \boldsymbol{\gamma}_*)\boldsymbol{x}.$$

□

The fact that the second fundamental tensor equals, up to sign, the derivative of the Gauss map, suggests that it provides a measure of how much the surface curves. For example, a manifold is said to be *totally geodesic* if its second fundamental tensor vanishes. For a hypersurface M, this means that the Gauss map has zero derivative, so that, if M is connected, the map is constant by Theorem 2.4.3. in other words, the unit normal field $\boldsymbol{n}$ is parallel, and M is contained in a hyperplane: given $\boldsymbol{p} \in M$, $M \subset \{\boldsymbol{p} + \boldsymbol{q} \mid \boldsymbol{q} \perp \pi_2(\boldsymbol{n}(\boldsymbol{p}))\}$.

One can also consider a weaker condition: a point $\boldsymbol{p} \in M$ is said to be *umbilical* if the second fundamental tensor at $\boldsymbol{p}$ is a multiple of the identity, and M itself is said to be *totally umbilic* if every point of M has that property; i.e., if there is a function $f : M \to \mathbb{R}$ such that $S\boldsymbol{v} = f(\boldsymbol{p})\boldsymbol{v}$ for all $\boldsymbol{v} \in M_{\boldsymbol{p}}$ and $\boldsymbol{p} \in M$. Thus, a totally umbilic manifold is one that "curves equally in all directions". What is remarkable is that in this case, the function f is actually constant. This is implicit in the proof of the following:

Theorem 7.2.3. *A totally umbilic connected hypersurface is part of a sphere or a hyperplane.*

Proof. We first show that the function $f : M \to \mathbb{R}$ in the definition of umbilic manifold is constant. Denote, as usual, by $\boldsymbol{n}$ the unit normal field of M. Thus,

$$D_{\boldsymbol{X}}\boldsymbol{n} = -S\boldsymbol{X} = -f\boldsymbol{X}$$

for any field $\boldsymbol{X}$ on M; given fields $\boldsymbol{Y}$ and $\boldsymbol{Z}$,

$$\begin{aligned}\langle D_{\boldsymbol{X}}D_{\boldsymbol{Y}}\boldsymbol{n}, \boldsymbol{Z}\rangle &= \boldsymbol{X}\langle D_{\boldsymbol{Y}}\boldsymbol{n}, \boldsymbol{Z}\rangle - \langle D_{\boldsymbol{Y}}\boldsymbol{n}, D_{\boldsymbol{X}}\boldsymbol{Z}\rangle = \boldsymbol{X}\langle -f\boldsymbol{Y}, \boldsymbol{Z}\rangle + f\langle \boldsymbol{Y}, D_{\boldsymbol{X}}\boldsymbol{Z}\rangle \\ &= -(\boldsymbol{X}f)\langle \boldsymbol{Y}, \boldsymbol{Z}\rangle - f\langle D_{\boldsymbol{X}}\boldsymbol{Y}, \boldsymbol{Z}\rangle - f\langle \boldsymbol{Y}, D_{\boldsymbol{X}}\boldsymbol{Z}\rangle + f\langle \boldsymbol{Y}, D_{\boldsymbol{X}}\boldsymbol{Z}\rangle \\ &= -(\boldsymbol{X}f)\langle \boldsymbol{Y}, \boldsymbol{Z}\rangle - f\langle D_{\boldsymbol{X}}\boldsymbol{Y}, \boldsymbol{Z}\rangle.\end{aligned}$$

Using (3.9.1), we obtain

$$\begin{aligned}0 &= \langle D_{\boldsymbol{X}}D_{\boldsymbol{Y}}\boldsymbol{n} - D_{\boldsymbol{Y}}D_{\boldsymbol{X}}\boldsymbol{n} - D_{[\boldsymbol{X},\boldsymbol{Y}]}\boldsymbol{n}, \boldsymbol{Z}\rangle \\ &= -(\boldsymbol{X}f)\langle \boldsymbol{Y}, \boldsymbol{Z}\rangle - f\langle D_{\boldsymbol{X}}\boldsymbol{Y}, \boldsymbol{Z}\rangle + (\boldsymbol{Y}f)\langle \boldsymbol{X}, \boldsymbol{Z}\rangle + f\langle D_{\boldsymbol{Y}}\boldsymbol{X}, \boldsymbol{Z}\rangle \\ &\quad + f\langle [\boldsymbol{X}, \boldsymbol{Y}], \boldsymbol{Z}\rangle \\ &= \langle (\boldsymbol{Y}f)\boldsymbol{X} - (\boldsymbol{X}f)\boldsymbol{Y}, \boldsymbol{Z}\rangle.\end{aligned}$$

Since $\boldsymbol{Z}$ is arbitrary, $(\boldsymbol{Y}f)\boldsymbol{X} - (\boldsymbol{X}f)\boldsymbol{Y}$ must vanish. Choosing linearly independent $\boldsymbol{X}$ and $\boldsymbol{Y}$ then implies that $\boldsymbol{X}f \equiv 0$ for any vector field $\boldsymbol{X}$ on M, and since M is connected, f must equal a constant $\lambda \in \mathbb{R}$.

Recall from Examples 2.8.2 (ii) that the position vector field $\boldsymbol{P}$ on $\mathbb{R}^{n+1}$, $\boldsymbol{P}(\boldsymbol{a}) = \mathcal{I}_{\boldsymbol{a}}\boldsymbol{a}$, satisfies $D_{\boldsymbol{X}}\boldsymbol{P} = \boldsymbol{X}$. Thus, if $\boldsymbol{X}$ is a vector field on M, then

$$D_{\boldsymbol{X}}(\boldsymbol{n} + \lambda\boldsymbol{P}) = -\lambda\boldsymbol{X} + \lambda\boldsymbol{X} = 0,$$

and the vector field $\boldsymbol{n} + \lambda\boldsymbol{P}$ is parallel. Parallel translating this field back to the origin, we conclude that the map $\pi_2 \circ \boldsymbol{n} + \lambda\pi_2 \circ \boldsymbol{P}$ is a constant map $\boldsymbol{a}$ for some $\boldsymbol{a} \in \mathbb{R}^{n+1}$.

In other words, $\gamma(\boldsymbol{p}) + \lambda \boldsymbol{p} = \boldsymbol{a}$ for all $\boldsymbol{p} \in M$, where γ is the Gauss map of M. The case when λ is zero was discussed earlier, and corresponds to a hyperplane. If $\lambda \neq 0$, dividing by it and taking norms in the last identity yields

$$|\boldsymbol{p} - \frac{1}{\lambda}\boldsymbol{a}| = \frac{1}{\lambda}, \qquad \boldsymbol{p} \in M,$$

so that M is contained in the sphere of radius $1/\lambda$ centered at $(1/\lambda)\boldsymbol{a}$. □

7.3 Curvature of hypersurfaces

Let M denote an oriented hypersurface in $\mathbb{R}^{n+1}$ with unit normal field $\boldsymbol{n}$ and second fundamental tensor $S = S_{\boldsymbol{n}}$. In addition to those that we have already encountered, there are several other notions of curvature associated to M at each point of the hypersurface: The *principal curvatures* at $\boldsymbol{p} \in M$ are the eigenvalues of $S(\boldsymbol{p})$. The *mean curvature* is defined to be $(1/n)$ times the trace of S. The determinant of $S(\boldsymbol{p})$ is called the *Gaussian curvature* at $\boldsymbol{p}$. Notice that unlike the sectional curvature, the first two (and also the third one if M is odd-dimensional) change sign if one chooses the opposite orientation. Unless otherwise specified, the word curvature by itself will always refer to the sectional curvature.

The definition of the curvature tensor R implies that

$$R(\boldsymbol{x},\boldsymbol{y})\boldsymbol{z} = \langle S\boldsymbol{y},\boldsymbol{z}\rangle S\boldsymbol{x} - \langle S\boldsymbol{x},\boldsymbol{z}\rangle S\boldsymbol{y}, \qquad \boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in M_{\boldsymbol{p}}, \quad \boldsymbol{p} \in M,$$

so that the curvature form k from Chapter 3 is given by

$$k(\boldsymbol{x},\boldsymbol{y}) = \langle R(\boldsymbol{x},\boldsymbol{y})\boldsymbol{y},\boldsymbol{x}\rangle = \langle S\boldsymbol{x},\boldsymbol{x}\rangle \cdot \langle S\boldsymbol{y},\boldsymbol{y}\rangle - \langle S\boldsymbol{x},\boldsymbol{y}\rangle^2. \tag{7.3.1}$$

We say *M has positive curvature at $\boldsymbol{p} \in M$* if every plane in $M_{\boldsymbol{p}}$ has positive sectional curvature. When $n = 2$, this is the same as saying that M has positive Gaussian curvature at $\boldsymbol{p}$, since the sectional curvature of the only tangent plane equals the Gaussian curvature. M is said to be *positively curved* or to have *positive curvature* if this holds for every $\boldsymbol{p} \in M$. Obvious modifications yield the notions of *nonnegative curvature*, *negative curvature*, and *nonpositive curvature*.

One consequence of (7.3.1) is the following:

Theorem 7.3.1. *M has positive curvature at $\boldsymbol{p}$ if and only if the second fundamental form is definite at $\boldsymbol{p}$.*

Proof. Recall from Section 2.6 that the second fundamental form is definite if and only if the eigenvalues of S are all positive or all negative; equivalently, the map

$$\begin{aligned} M_{\boldsymbol{p}} \times M_{\boldsymbol{p}} &\to \mathbb{R} \\ (\boldsymbol{x},\boldsymbol{y}) &\mapsto \langle S\boldsymbol{x},\boldsymbol{y}\rangle \end{aligned}$$

or its negative (that is, replacing S by $-S$) is an inner product on $M_{\boldsymbol{p}}$. So assume that M has positive curvature at $\boldsymbol{p}$. If S has only one eigenvalue λ at $\boldsymbol{p}$, then $\lambda \neq 0$, for otherwise $k = 0$. Then $S = \lambda 1_{M_{\boldsymbol{p}}}$ is definite. If S has more than one eigenvalue, choose unit eigenvectors $\boldsymbol{x}_i$ corresponding to distinct eigenvalues λ_i, $i = 1, 2$. (7.3.1) implies that

$$0 < k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \lambda_1 \cdot \lambda_2.$$

Thus the product of any two eigenvalues is positive, which means that all eigenvalues have the same sign.

Conversely, suppose the second fundamental form is definite. Then the map $(\boldsymbol{x}, \boldsymbol{y}) \mapsto \langle S\boldsymbol{x}, \boldsymbol{y}\rangle$ or its negative is an inner product. The Cauchy-Schwarz inequality for either case implies that the right side of (7.3.1) is positive for linearly independent $\boldsymbol{x}$ and $\boldsymbol{y}$, and the claim follows. □

Theorem 7.3.2. *There are no compact hypersurfaces of nonpositive curvature.*

Proof. Let M be a compact hypersurface in $\mathbb{R}^{n+1}$. The square of the distance to the origin function f, $f(\boldsymbol{p}) = |\boldsymbol{p}|^2$, must assume a maximum at some $\boldsymbol{p}_0$ in the compact set M. We will show that M has positive curvature at $\boldsymbol{p}_0$. In view of Theorem 7.3.1, it is enough to show that the second fundamental form is definite at that point. Now, the vector $\boldsymbol{n} = (1/|\boldsymbol{p}_0|)\mathcal{I}_{\boldsymbol{p}_0}\boldsymbol{p}_0$ is a unit normal to M, because if $\boldsymbol{u} \in M_{\boldsymbol{p}_0}$, and $\boldsymbol{c}$ is any curve in M with $\boldsymbol{c}(0) = \boldsymbol{p}_0$ and $\dot{\boldsymbol{c}}(0) = \boldsymbol{u}$, then

$$0 = (f \circ \boldsymbol{c})'(0) = 2\langle \boldsymbol{c}'(0), \boldsymbol{c}(0)\rangle = 2\langle \dot{\boldsymbol{c}}(0), \mathcal{I}_{\boldsymbol{p}_0}\boldsymbol{p}_0\rangle = 2|\boldsymbol{p}_0|\langle \boldsymbol{u}, \boldsymbol{n}\rangle.$$

Furthermore,

$$0 \geq (f \circ \boldsymbol{c})''(0) = 2(\langle \boldsymbol{c}'', \boldsymbol{c}\rangle(0) + |\boldsymbol{c}'(0)|^2),$$

so that $\langle \boldsymbol{c}'', \boldsymbol{c}\rangle(0) \leq -|\boldsymbol{u}|^2$. But

$$\langle \boldsymbol{c}'', \boldsymbol{c}\rangle(0) = \langle \nabla_{\boldsymbol{D}(0)}\dot{\boldsymbol{c}}, \mathcal{I}_{\boldsymbol{p}_0}\boldsymbol{p}_0\rangle = |\boldsymbol{p}_0|\langle \nabla_{\boldsymbol{D}(0)}\boldsymbol{c}, \boldsymbol{n}\rangle = -|\boldsymbol{p}_0|\langle S_{\boldsymbol{n}}\boldsymbol{u}, \boldsymbol{u}\rangle,$$

and consequently $\langle S_{\boldsymbol{n}}\boldsymbol{u}, \boldsymbol{u}\rangle \geq |\boldsymbol{u}|^2/|\boldsymbol{p}_0|^2$ as claimed. □

A symmetric bilinear form b on an inner product space V is said to be *semi-definite* if either $b(\boldsymbol{v}, \boldsymbol{v}) \geq 0$ for all $\boldsymbol{v} \in V$ or $b(\boldsymbol{v}, \boldsymbol{v}) \leq 0$ for all $\boldsymbol{v} \in V$. In the exercises, the reader is asked to show that M has nonnegative curvature at a point if and only if the second fundamental form is semi-definite at that point.

In order to formulate our next result more concisely, we introduce the following concept: the *Laplacian* of a function f is the function $\Delta f = \operatorname{tr} H_f$, where H_f is the Hessian of f, and tr is the trace. To avoid the notation from becoming cumbersome, we will identify the tangent space of Euclidean space at a point with Euclidean space itself via the usual projection $\pi_2 : T\mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$. Thus, the gradient of f becomes the vector-valued map $[Df]$, and the Hessian of f satisfies

$$H_f(\boldsymbol{p})\boldsymbol{x} = D_{\boldsymbol{x}}\nabla f, \qquad \boldsymbol{x} \in \mathbb{R}^{n+1},$$

where of course the $\boldsymbol{x}$ on the right side is to be interpreted as $\mathcal{I}_{\boldsymbol{p}}\boldsymbol{x}$. To see this, it suffices to verify the identity for $\boldsymbol{x} = \boldsymbol{e}_i$, since both sides are linear in $\boldsymbol{x}$. Now,

$$D_{\boldsymbol{e}_i}\nabla f = D_i[Df](\boldsymbol{p}) = \boldsymbol{c}'(0),$$

where

$$\boldsymbol{c}(t) = [Df(\boldsymbol{p} + t\boldsymbol{e}_i)] = [D_1 f(\boldsymbol{p} + t\boldsymbol{e}_i), \ldots, D_{n+1} f(\boldsymbol{p} + t\boldsymbol{e}_i)].$$

Consequently,

$$\boldsymbol{c}'(0) = [D_{i1}f, \ldots, D_{i(n+1)}f](\boldsymbol{p}) = H_f(\boldsymbol{p})\boldsymbol{e}_i,$$

as claimed.

Theorem 7.3.3. *Let $M^n = f^{-1}(0) \subset \mathbb{R}^{n+1}$, where 0 is a regular value of f, and $\boldsymbol{p} \in M$, $\boldsymbol{x}, \boldsymbol{y} \in M_{\boldsymbol{p}}$. Denote by S, s, H, G the second fundamental tensor, second fundamental form, mean curvature, and Gaussian curvature respectively of M with respect to the standard orientation of M. Then*

- $S\boldsymbol{x} = -\frac{1}{|\nabla f|(\boldsymbol{p})}(H_f\boldsymbol{x})^\top$;
- $s(\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{|\nabla f|(\boldsymbol{p})}h_f(\boldsymbol{x}, \boldsymbol{y})$;
- $H(\boldsymbol{p}) = \frac{1}{n|\nabla f|(\boldsymbol{p})}\left(\frac{h_f(\nabla f, \nabla f)}{|\nabla f|^2}(\boldsymbol{p}) - \Delta f(\boldsymbol{p})\right)$;
- $G(\boldsymbol{p}) = \frac{1}{|\nabla f|^n(\boldsymbol{p})}\det H_f^\top(\boldsymbol{p})$;
- *If $\boldsymbol{x}$ and $\boldsymbol{y}$ form an orthonormal basis of a plane $P \subset M_{\boldsymbol{p}}$, then the sectional curvature of P is*

$$K_P = \frac{1}{|\nabla f|^2(\boldsymbol{p})}\det\begin{bmatrix} h_f(\boldsymbol{x}, \boldsymbol{x}) & h_f(\boldsymbol{x}, \boldsymbol{y}) \\ h_f(\boldsymbol{x}, \boldsymbol{y}) & h_f(\boldsymbol{y}, \boldsymbol{y}) \end{bmatrix}.$$

Proof.

$$\begin{aligned} S\boldsymbol{x} &= -D_{\boldsymbol{x}}\left(\frac{1}{|\nabla f|}\nabla f\right) = -\boldsymbol{x}\left(\frac{1}{|\nabla f|}\right)(\nabla f)^\top(\boldsymbol{p}) - \frac{1}{|\nabla f|(\boldsymbol{p})}(D_{\boldsymbol{x}}\nabla f)^\top \\ &= -\frac{1}{|\nabla f|(\boldsymbol{p})}(D_{\boldsymbol{x}}\nabla f)^\top = -\frac{1}{|\nabla f|(\boldsymbol{p})}(H_f\boldsymbol{x})^\top. \end{aligned}$$

This establishes the first identity. The others follow from the definitions of the different types of curvature, together with (7.3.1) and the identity just proved. □

There is a useful alternative formula for the Gaussian curvature in the above proposition. In order to derive it, we need the following:

Lemma 7.3.1. *Let $L : V \to V$ be a linear transformation on an $(n+1)$-dimensional inner product space V. Let W be an n-dimensional subspace of V and $\pi : V \to W$ the orthogonal projection onto W; i.e., if $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2 \in W \oplus W^\perp$, then $\pi\boldsymbol{x} = \boldsymbol{x}_1$. The restriction $\pi \circ L_{|W}$ of $\pi \circ L$ to W is then an operator on W. Its determinant is given by*

$$\det(\pi \circ L_{|W}) = \langle \tilde{L}\boldsymbol{x}, \boldsymbol{x}\rangle,$$

where $\boldsymbol{x}$ is a unit vector in $W^\perp$, and $\tilde{L}$ is the linear map adjugate to L (see Theorem 1.3.6).

Proof. Extend $\boldsymbol{x}$ to an ordered orthonormal basis $\mathcal{B} = \{\boldsymbol{x}, \boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ of V. Then the last n vectors form an orthonormal basis $\mathcal{B}_1$ of W, and the (i,j)-th entry of the matrix of $\pi \circ L_{|W}$ with respect to this basis is

$$\langle \pi L\boldsymbol{v}_j, \boldsymbol{v}_i\rangle = \langle L\boldsymbol{v}_j - \langle L\boldsymbol{v}_j, \boldsymbol{x}\rangle \boldsymbol{x}, \boldsymbol{v}_i\rangle = \langle L\boldsymbol{v}_j, \boldsymbol{v}_i\rangle,$$

so that the determinant of $\pi \circ L_{|W}$ equals the determinant of the matrix obtained by deleting the first row and column of $[L]_{\mathcal{B}}$. On the other hand, $\langle \tilde{L}\boldsymbol{x}, \boldsymbol{x}\rangle$ is the $(1,1)$ entry of $[\tilde{L}]_{\mathcal{B}}$, which by definition of $\tilde{L}$, is that same determinant. □

The lemma, together with the formula for G from Theorem 7.3.3 now immediately imply the following:

Proposition 7.3.1. *With the hypotheses of Theorem 7.3.3, the Gaussian curvature of M is given by*

$$G = \frac{1}{|\nabla f|^{n+2}} \langle \tilde{H}_f \nabla f, \nabla f\rangle.$$

An important special case occurs when the function f is a quadratic form. As before, we have identified in the next result Euclidean space with its tangent space at a given point.

Proposition 7.3.2. *Suppose L is a self-adjoint operator on $\mathbb{R}^{n+1}$. Denote by b the associated scalar product, $b(\boldsymbol{x}, \boldsymbol{y}) = \langle L\boldsymbol{x}, \boldsymbol{y}\rangle$, and by f the corresponding quadratic form, $f(\boldsymbol{x}) = b(\boldsymbol{x}, \boldsymbol{x})$. Consider a regular value a of f, and the corresponding hypersurface $M = f^{-1}(a)$. Given $\boldsymbol{p} \in M$,*

(1) $\nabla f(\boldsymbol{p}) = 2L\boldsymbol{p}$;
(2) $H_f = 2L$, $h_f = 2b$;
(3) $G(\boldsymbol{p}) = \frac{\det L}{|L\boldsymbol{p}|^{n+2}}\, a$;
(4) *If $\boldsymbol{x}$ and $\boldsymbol{y}$ form an orthonormal basis of a plane $P \subset M_{\boldsymbol{p}}$, then the sectional curvature of P is*

$$K_P = \frac{1}{|L\boldsymbol{p}|^2} \det \begin{bmatrix} b(\boldsymbol{x},\boldsymbol{x}) & b(\boldsymbol{x},\boldsymbol{y}) \\ b(\boldsymbol{x},\boldsymbol{y}) & b(\boldsymbol{y},\boldsymbol{y}) \end{bmatrix}.$$

Proof. Let $\boldsymbol{x} \in \mathbb{R}^{n+1}$, and consider the ray $\boldsymbol{c}_{\boldsymbol{x}}$, $\boldsymbol{c}_{\boldsymbol{x}}(t) = \boldsymbol{p} + t\boldsymbol{x}$. Then

$$\langle \nabla f(\boldsymbol{p}), \boldsymbol{x}\rangle = \boldsymbol{x}f = (f \circ \boldsymbol{c}_{\boldsymbol{x}})'(0).$$

Now,

$$(f \circ \boldsymbol{c}_{\boldsymbol{x}})(t) = b(\boldsymbol{p} + t\boldsymbol{x}, \boldsymbol{p} + t\boldsymbol{x}) = b(\boldsymbol{p},\boldsymbol{p}) + 2tb(\boldsymbol{p},\boldsymbol{x}) + t^2 b(\boldsymbol{x},\boldsymbol{x}),$$

so that $\langle \nabla f(\boldsymbol{p}), \boldsymbol{x}\rangle = 2b(\boldsymbol{p},\boldsymbol{x}) = 2\langle L\boldsymbol{p}, \boldsymbol{x}\rangle$, which establishes (1).
For (2), observe that (1) implies $D_i f(\boldsymbol{p}) = 2\langle L\boldsymbol{p}, \boldsymbol{e}_i\rangle$, so that $D_{ji} f(\boldsymbol{p}) = \boldsymbol{c}'(0)$, where

$$\boldsymbol{c}(t) = 2\langle L(\boldsymbol{p} + t\boldsymbol{e}_j), \boldsymbol{e}_i\rangle = 2(\langle L\boldsymbol{p}, \boldsymbol{e}_i\rangle + t\langle L\boldsymbol{e}_j, \boldsymbol{e}_i\rangle).$$

Thus, $D_{ji} f(\boldsymbol{p}) = 2\langle L\boldsymbol{e}_j, \boldsymbol{e}_i\rangle$. Since these are the respective entries of the matrices of $H_f(\boldsymbol{p})$ and $2L$ in the standard basis, both transformations coincide. The formula for the

hessian form is immediate, and the one for sectional curvature follows from Theorem 7.3.3. Finally, by Proposition 7.3.1 together with (1) and (2),

$$G(\boldsymbol{p}) = \frac{1}{|\nabla f(\boldsymbol{p})|^{n+2}}\langle \tilde{H}_f \nabla f, \nabla f\rangle(\boldsymbol{p}) = \frac{1}{|2L\boldsymbol{p}|^{n+2}}\langle \widetilde{(2L)}(2L)\boldsymbol{p}, 2L\boldsymbol{p}\rangle$$
$$= \frac{\det(2L)}{2^{n+2}|L\boldsymbol{p}|^{n+2}}\langle 2L\boldsymbol{p}, \boldsymbol{p}\rangle = \frac{\det L}{|L\boldsymbol{p}|^{n+2}}\, a. \qquad \square$$

Notice that if the quadratic form b is definite, then M has positive curvature. If it is merely semi-definite, then M has nonnegative curvature by Exercise 7.17. When $n = 2$, the surfaces of the type considered in Proposition 7.3.2 are called *quadrics*. The reader is invited to classify them in the exercises.

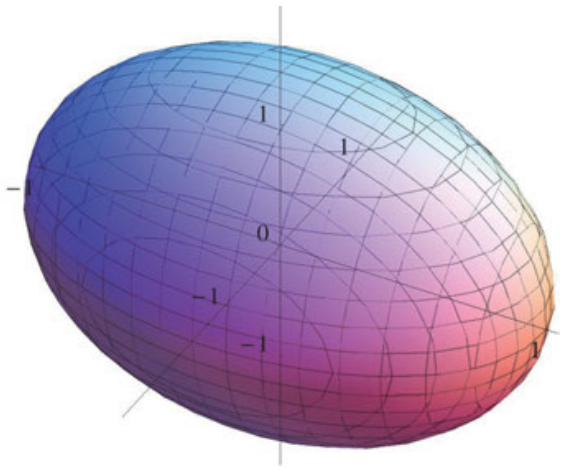


Fig. 7.2: The ellipsoid $x^2 + \frac{2}{3}y^2 + \frac{1}{2}z^2 = 1$.

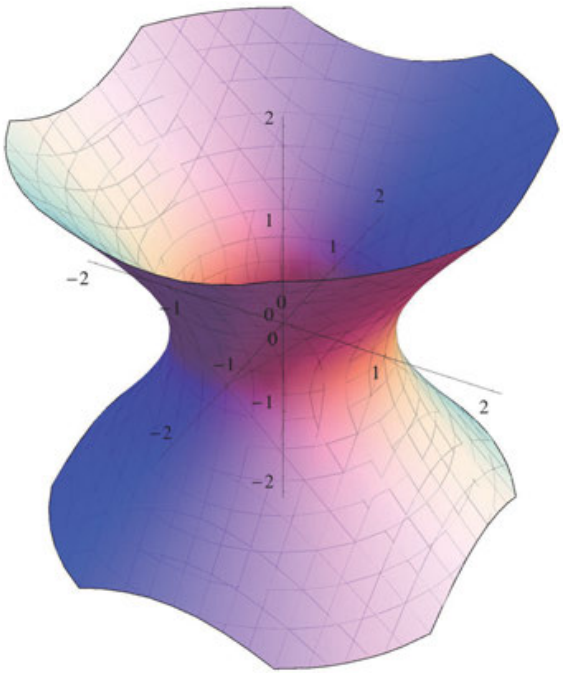


Fig. 7.3: The hyperboloid $x^2 + y^2 - z^2 = 1$.

Example 7.3.1. When $n = 2$, the Gaussian curvature and the sectional curvature coincide. In order to compute the latter, one can therefore use Proposition 7.3.2(3) without having to explicitly find a basis for the tangent space:

– The ellipsoid $(x^2/a^2) + (y^2/b^2) + (z^2/c^2) = 1$ is the surface $\{\boldsymbol{x} \in \mathbb{R}^3 \mid \langle L\boldsymbol{x}, \boldsymbol{x}\rangle = 1\}$, where the matrix of $L : \mathbb{R}^3 \to \mathbb{R}^3$ in the standard basis is

$$[L] = \begin{bmatrix} \frac{1}{a^2} & 0 & 0 \\ 0 & \frac{1}{b^2} & 0 \\ 0 & 0 & \frac{1}{c^2} \end{bmatrix}.$$

The formula for the Gaussian curvature then implies that the curvature at (x, y, z) equals

$$K(x, y, z) = \left[abc\left(\frac{x^2}{a^4} + \frac{y^2}{b^4} + \frac{z^2}{c^4}\right)\right]^{-2}.$$

– The hyperboloid $(x^2/a^2) + (y^2/b^2) - (z^2/c^2) = 1$ differs from the ellipsoid in that the last diagonal entry of $[L]$ has its sign reversed. The curvature is therefore given by

$$K(x, y, z) = -\left[abc\left(\frac{x^2}{a^4} + \frac{y^2}{b^4} + \frac{z^2}{c^4}\right)\right]^{-2}.$$

7.4 The fundamental theorem for hypersurfaces

We have already remarked in Chapter 3 that if $\boldsymbol{F}$ is a rigid motion of Euclidean space, and M is a submanifold, then the restriction of $\boldsymbol{F}$ to M is an isometry with $\boldsymbol{F}(M)$. We also noticed that the converse is not true, that is, an isometry $\boldsymbol{f} : M_1 \to M_2$ between submanifolds need not be the restriction of a rigid motion. In this section, we show it is nevertheless true for hypersurfaces, provided $\boldsymbol{f}$ preserves in addition the second fundamental form (it is also true for more general submanifolds, under additional hypotheses). Specifically:

Theorem 7.4.1. *Let M_i denote oriented hypersurfaces in $\mathbb{R}^{n+1}$ with unit normal fields $\boldsymbol{n}_i$ and second fundamental tensors S_i, $i = 1, 2$. Suppose that $\boldsymbol{f} : M_1 \to M_2$ is an isometry, and that*

$$S_2 \boldsymbol{f}_* \boldsymbol{x} = \boldsymbol{f}_* S_1 \boldsymbol{x}, \qquad \boldsymbol{x} \in M_{1\boldsymbol{p}}, \quad \boldsymbol{p} \in M_1.$$

If M_1 is connected, then there exists a rigid motion $\boldsymbol{F}$ of $\mathbb{R}^{n+1}$ such that $\boldsymbol{f} = \boldsymbol{F} \circ \imath$, where $\imath : M_1 \hookrightarrow \mathbb{R}^{n+1}$ denotes inclusion.

Proof. Notice first that $\boldsymbol{F}$, if it exists, is, up to sign, entirely determined by $\boldsymbol{f}$: indeed, $\boldsymbol{F}(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{b}$ for some orthogonal transformation A and $\boldsymbol{b} \in \mathbb{R}^{n+1}$. Fix any $\boldsymbol{p} \in M_1$. Then $\boldsymbol{F}(\boldsymbol{p}) = A\boldsymbol{p} + \boldsymbol{b} = \boldsymbol{f}(\boldsymbol{p})$, so that $\boldsymbol{b}$ must equal $\boldsymbol{f}(\boldsymbol{p}) - A\boldsymbol{p}$. Furthermore, $D\boldsymbol{F}(\boldsymbol{p}) = A$, so that A must, up to parallel translation, equal $\boldsymbol{f}_{*\boldsymbol{p}}$ on $M_{1\boldsymbol{p}}$. Denoting as usual by $\boldsymbol{\pi}_2$ the projection from $T\mathbb{R}^{n+1}$ to $\mathbb{R}^{n+1}$, it follows that $A(\boldsymbol{\pi}_2 \boldsymbol{n}_1(\boldsymbol{p})) = \pm \boldsymbol{\pi}_2 \boldsymbol{n}_2(\boldsymbol{f}(\boldsymbol{p}))$. This completely determines the operator A on $\mathbb{R}^{n+1}$, and together with the condition on $\boldsymbol{b}$, determines $\boldsymbol{F}$.

With this in mind, define $\boldsymbol{F}$ as above; i.e., consider the orthogonal transformation A determined by

$$A(\boldsymbol{\pi}_2\boldsymbol{n}_1(\boldsymbol{p})) = \boldsymbol{\pi}_2\boldsymbol{n}_2(\boldsymbol{f}(\boldsymbol{p})), \quad A(\boldsymbol{\pi}_2\boldsymbol{x}) = \boldsymbol{\pi}_2\boldsymbol{f}_*\boldsymbol{x} \text{ for } \boldsymbol{x} \in M_{1\boldsymbol{p}},$$

and the rigid motion $\boldsymbol{F}$, where

$$\boldsymbol{F}(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{f}(\boldsymbol{p}) - A\boldsymbol{p}.$$

The result will follow once we show that the map $\boldsymbol{G} := \boldsymbol{F}^{-1} \circ \boldsymbol{f}$ is the identity on M_1. In fact, since M_1 is connected, it suffices to show that $\boldsymbol{G} \circ \boldsymbol{c} = \boldsymbol{c}$ for any curve $\boldsymbol{c}$ in M_1 with $\boldsymbol{c}(0) = \boldsymbol{p}$. By construction, $\boldsymbol{G}(\boldsymbol{p}) = \boldsymbol{p}$ (and $\boldsymbol{G}_{*\boldsymbol{p}}$ is the identity on the tangent space of M_1 at $\boldsymbol{p}$), so $(\boldsymbol{G} \circ \boldsymbol{c})(0) = \boldsymbol{c}(0)$, and we only need to show that $(\boldsymbol{G} \circ \boldsymbol{c})' = \boldsymbol{c}'$.
So choose an orthonormal basis $\boldsymbol{x}_i$ of $M_{1\boldsymbol{p}}$, and let $\boldsymbol{X}_i$ denote the parallel field along $\boldsymbol{c}$ with $\boldsymbol{X}_i(0) = \boldsymbol{x}_i$. By (3.11.6), the fields $\boldsymbol{Y}_i := \boldsymbol{G}_*\boldsymbol{X}_i$ are parallel and orthonormal along $\boldsymbol{G} \circ \boldsymbol{c}$. Moreover, $\boldsymbol{Y}_i(0) = \boldsymbol{X}_i(0)$ because $\boldsymbol{G}_{*\boldsymbol{p}}$ is the identity. Now, $\dot{\boldsymbol{c}} = \sum_i \langle \dot{\boldsymbol{c}}, \boldsymbol{X}_i\rangle \boldsymbol{X}_i$, and

$$\boldsymbol{G}\dot{\circ}\boldsymbol{c} = \sum_i \langle \boldsymbol{G}_*\dot{\boldsymbol{c}}, \boldsymbol{Y}_i\rangle \boldsymbol{Y}_i = \sum_i \langle \boldsymbol{G}_*\dot{\boldsymbol{c}}, \boldsymbol{G}_*\boldsymbol{X}_i\rangle \boldsymbol{Y}_i = \sum_i \langle \dot{\boldsymbol{c}}, \boldsymbol{X}_i\rangle \boldsymbol{Y}_i,$$

so that if $X_i = \boldsymbol{\pi}_2\boldsymbol{X}_i$ and $Y_i = \boldsymbol{\pi}_2\boldsymbol{Y}_i$, then

$$\boldsymbol{c}' = \sum_i \langle \dot{\boldsymbol{c}}, \boldsymbol{X}_i\rangle X_i, \qquad (\boldsymbol{G} \circ \boldsymbol{c})' = \sum_i \langle \dot{\boldsymbol{c}}, \boldsymbol{X}_i\rangle Y_i.$$

It therefore remains to show that $X_i = Y_i$ for $1 \le i \le n$. First, observe that $M := \boldsymbol{F}^{-1}(M_2)$ is a manifold isometric to M_2 with unit normal $\boldsymbol{n} := \boldsymbol{F}_*^{-1} \circ \boldsymbol{n}_2 \circ \boldsymbol{F}$. Set $\boldsymbol{X}_{n+1} = \boldsymbol{n}_1 \circ \boldsymbol{c}$, $\boldsymbol{Y}_{n+1} = \boldsymbol{n} \circ \boldsymbol{G} \circ \boldsymbol{c}$, so that $\{\boldsymbol{X}_i\}_{1\le i\le n+1}$ and $\{\boldsymbol{Y}_i\}_{1\le i\le n+1}$ are orthonormal bases of $\mathbb{R}^{n+1}_{\boldsymbol{c}}$ and $\mathbb{R}^{n+1}_{\boldsymbol{G}\circ\boldsymbol{c}}$ respectively. Now,

$$X_i' = \sum_j g_{ij}X_j, \quad Y_i' = \sum_j h_{ij}Y_j, \quad \text{where } g_{ij} = \langle X_i', X_j\rangle, \quad h_{ij} = \langle Y_i', Y_j\rangle.$$

Since $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ are parallel for $i \le n$, $g_{ij} \equiv h_{ij} \equiv 0$ when $i, j \le n$. In the case that $j = n + 1$, we use the fact that $\boldsymbol{f}$ preserves the second fundamental form: The condition $S_2\boldsymbol{f}_*\boldsymbol{x} = \boldsymbol{f}_*S_1\boldsymbol{x}$ means that $D_{\boldsymbol{f}_*\boldsymbol{x}}\boldsymbol{n}_2 = \boldsymbol{f}_*D_{\boldsymbol{x}}\boldsymbol{n}_1$. Thus,

$$\begin{aligned} D_{\boldsymbol{G}_*\boldsymbol{x}}\boldsymbol{n} &= D_{\boldsymbol{G}_*\boldsymbol{x}}(\boldsymbol{F}_*^{-1} \circ \boldsymbol{n}_2 \circ \boldsymbol{F}) = \boldsymbol{F}_*^{-1}D_{\boldsymbol{G}_*\boldsymbol{x}}(\boldsymbol{n}_2 \circ \boldsymbol{F}) = \boldsymbol{F}_*^{-1}D_{\boldsymbol{F}_*\circ\boldsymbol{G}_*\boldsymbol{x}}\boldsymbol{n}_2 \\ &= \boldsymbol{F}_*^{-1}D_{\boldsymbol{f}_*\boldsymbol{x}}\boldsymbol{n}_2 = \boldsymbol{F}_*^{-1} \circ \boldsymbol{f}_*D_{\boldsymbol{x}}\boldsymbol{n}_1 \\ &= \boldsymbol{G}_*D_{\boldsymbol{x}}\boldsymbol{n}_1. \end{aligned}$$

Now take $\boldsymbol{x} = \dot{\boldsymbol{c}}$ in the above identity to conclude that $\boldsymbol{Y}_{n+1}' = \boldsymbol{G}_*\boldsymbol{X}_{n+1}'$. It follows that

$$\begin{aligned} h_{i\,n+1} &= \langle \boldsymbol{Y}_i', \boldsymbol{Y}_{n+1}\rangle = -\langle \boldsymbol{Y}_i, \boldsymbol{Y}_{n+1}'\rangle = -\langle \boldsymbol{G}_*\boldsymbol{X}_i, \boldsymbol{G}_*\boldsymbol{X}_{n+1}'\rangle \\ &= -\langle \boldsymbol{X}_i, \boldsymbol{X}_{n+1}'\rangle = \langle \boldsymbol{X}_i', \boldsymbol{X}_{n+1}\rangle \\ &= g_{i\,n+1} \end{aligned}$$

for all $i \leq n$. But this also implies equality when $i = n + 1$, because for any i and j,

$$g_{ij} + g_{ji} = \langle \boldsymbol{X}_i', \boldsymbol{X}_j \rangle + \langle \boldsymbol{X}_i, \boldsymbol{X}_j' \rangle = \langle \boldsymbol{X}_i, \boldsymbol{X}_j \rangle' = 0,$$

and similarly for h_{ij}.
Summarizing, we have established that

$$X_i' = \sum_j g_{ij} X_j, \qquad Y_i' = \sum_j g_{ij} Y_j.$$

Let $Z_i = X_i - Y_i$. Then Z_i satisfies the system of ordinary differential equations

$$Z_i' = \sum_j g_{ij} Z_j, \qquad Z_i(0) = \mathbf{0}.$$

The constant vector fields $Z_i \equiv \mathbf{0}$ satisfy this system. By uniqueness of solutions, $Z_i \equiv \mathbf{0}$; i.e. $X_i \equiv Y_i$. □

Example 7.4.1. Let $\boldsymbol{c} : (a, b) \to \mathbb{R}^2$ denote a regular curve with no self-intersections (i.e., $\boldsymbol{c}$ is one-to-one, so that $M = \boldsymbol{c}(a, b)$ is a – perhaps only immersed – submanifold of $\mathbb{R}^2$), which we may assume is parametrized by arclength. Choose the unit normal vector field $\boldsymbol{n}$ along $\boldsymbol{c}$ so that $\det(\boldsymbol{c}', \boldsymbol{\pi}_2 \circ \boldsymbol{n}) \equiv 1$, and denote by S the corresponding second fundamental tensor. The *curvature* of $\boldsymbol{c}$ is

$$\kappa = \langle S\dot{\boldsymbol{c}}, \dot{\boldsymbol{c}} \rangle.$$

The curvature is, therefore, up to sign, equal to the norm of the second fundamental tensor. It is also, up to sign, the norm of the acceleration:

$$\langle S\dot{\boldsymbol{c}}, \dot{\boldsymbol{c}} \rangle = -\langle \boldsymbol{n}', \dot{\boldsymbol{c}} \rangle = \langle \boldsymbol{n}, \dot{\boldsymbol{c}}' \rangle = \pm |\dot{\boldsymbol{c}}'|,$$

since $\dot{\boldsymbol{c}}' \perp \dot{\boldsymbol{c}}$. In the exercises, the reader is asked to show that a circle of radius r has $|\kappa| = 1/r$. This means that going out along the normal line to $\boldsymbol{c}(t)$ at distance $1/|\kappa(t)|$ (in the appropriate direction) and drawing a circle centered there with radius equal to that distance yields the circle that best approximates $\boldsymbol{c}$ at t. It is called the *circle of curvature* at t.

Now, suppose $\boldsymbol{c}_i : (a, b) \to \mathbb{R}^2$, $i = 1, 2$, are two curves parametrized by arc length with no self-intersections. Then the map $\boldsymbol{f} : \boldsymbol{c}_1(a, b) \to \boldsymbol{c}_2(a, b)$, where $\boldsymbol{f} = \boldsymbol{c}_2 \circ \boldsymbol{c}_1^{-1}$, is an isometry. The above theorem says that $\boldsymbol{f}$ extends to a rigid motion of the plane if and only if the curves have the same curvature; i.e., $\kappa_2 \circ \boldsymbol{f} = \kappa_1$.

7.5 Curvature in local coordinates

Although a coordinate-free approach is preferable, it is not always feasible. In this section, we compute the matrix of the second fundamental tensor of a hypersurface M^n with respect to a basis of coordinate vector fields. In the case $n = 2$, the formulas are simple enough to obtain a reasonable expression for the sectional curvature.

Consider a chart $(U, \boldsymbol{x})$ with coordinate vector fields $\partial/\partial x^i = \boldsymbol{x}_*^{-1}\boldsymbol{D}_i \circ \boldsymbol{x}$ and corresponding unit normal field

$$\boldsymbol{n} = \frac{\times_{i=1}^n \frac{\partial}{\partial x^i}}{\left|\times_{i=1}^n \frac{\partial}{\partial x^i}\right|}$$

to M on U. Define $n \times n$ matrix-valued maps s, g, on U by

$$s_{ij} = \left\langle S\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right\rangle, \qquad g_{ij} = \left\langle \frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right\rangle.$$

As usual, let $n = \boldsymbol{\pi}_2 \circ \boldsymbol{n}$. We begin by finding an expression for s:

$$\begin{aligned} s_{ij} &= \left\langle S\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right\rangle = -\left\langle D_{\partial/\partial x^i}\boldsymbol{n}, \frac{\partial}{\partial x^j}\right\rangle = \left\langle D_{\partial/\partial x^i}\frac{\partial}{\partial x^j}, \boldsymbol{n}\right\rangle \\ &= \left\langle \left(D_{\boldsymbol{x}_*^{-1}\boldsymbol{D}_i}\frac{\partial}{\partial x^j}\right)\circ\boldsymbol{x}, \boldsymbol{n}\right\rangle = \left\langle \left(D_{\boldsymbol{D}_i}\left(\frac{\partial}{\partial x^j}\circ\boldsymbol{x}^{-1}\right)\right)\circ\boldsymbol{x}, \boldsymbol{n}\right\rangle \\ &= \left\langle (D_{\boldsymbol{D}_i}\boldsymbol{x}_*^{-1}\boldsymbol{D}_j)\circ\boldsymbol{x}, \boldsymbol{n}\right\rangle = \left\langle (D(D\boldsymbol{x}^{-1}\boldsymbol{e}_j)\boldsymbol{e}_i)\circ\boldsymbol{x}, \boldsymbol{n}\right\rangle \end{aligned}$$

by (2.8.3). By symmetry of s,

$$s_{ij} = \langle (D^2\boldsymbol{x}^{-1}\boldsymbol{e}_i\boldsymbol{e}_j)\circ\boldsymbol{x}, \boldsymbol{n}\rangle. \tag{7.5.1}$$

Alternatively, in terms of a parametrization $\boldsymbol{h} = \boldsymbol{x}^{-1}$,

$$s_{ij}\circ\boldsymbol{h} = \langle D^2\boldsymbol{h}\boldsymbol{e}_i\boldsymbol{e}_j, \boldsymbol{n}\circ\boldsymbol{h}\rangle. \tag{7.5.2}$$

We emphasize that the first term on the right side of the equality sign in (7.5.1) is the map from U to $\mathbb{R}^{n+1}$ that sends $\boldsymbol{p}$ to $(D\boldsymbol{k})((\boldsymbol{x}(\boldsymbol{p}))\boldsymbol{e}_j$, where $\boldsymbol{k} = D\boldsymbol{x}^{-1}\boldsymbol{e}_i : \boldsymbol{x}(U) \to \mathbb{R}^{n+1}$. s is not, in general, the matrix of S in the basis of coordinate vector fields, unless the latter are orthonormal. To find this matrix, we use the following:

Lemma 7.5.1. *Suppose $L : V \to V$ is a linear map on an inner product space V with basis $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$. If A denotes the matrix with (i,j) entry $a_{ij} = \langle L\boldsymbol{v}_i, \boldsymbol{v}_j\rangle$, and B the one with (i,j) entry $b_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{v}_j\rangle$, then the matrix of L with respect to $\mathcal{B}$ is $[L]_{\mathcal{B}} = (AB^{-1})^T$. In particular, if L is self-adjoint, then $[L]_{\mathcal{B}} = B^{-1}A$.*

Proof. If l_{ij} is the (i,j)-th entry of the matrix of L in the given basis, then

$$a_{ij} = \langle L\boldsymbol{v}_i, \boldsymbol{v}_j\rangle = \langle \sum_k l_{ki}\boldsymbol{v}_k, \boldsymbol{v}_j\rangle = \sum_k l_{kj}b_{kj},$$

so that $A = [L]_{\mathcal{B}}^T B$, and the claim follows. If in addition L is self-adjoint, then A is a symmetric matrix. B is always symmetric, and it is easy to see that B^{-1} must then also be symmetric. Thus, $(AB^{-1})^T = B^{-1\,T}A^T = B^{-1}A$. □

As an immediate consequence, we obtain:

Proposition 7.5.1. *The matrix of S in the basis $\{\partial/\partial x^i\}$ equals $g^{-1}s$, where*

$$s_{ij} = \langle (D^2\boldsymbol{x}^{-1}\boldsymbol{e}_i\boldsymbol{e}_j) \circ \boldsymbol{x}, \boldsymbol{n}\rangle, \qquad g_{ij} = \langle \frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\rangle.$$

The proposition may of course be stated alternatively in terms of a parametrization $\boldsymbol{h}$ using (7.5.2). This is in fact the way it is done in the following application:

Corollary 7.5.1. *If $M^2 \subset \mathbb{R}^3$ is locally parametrized by $(U, \boldsymbol{h})$, then the sectional curvature of M at $\boldsymbol{p} = \boldsymbol{h}(\boldsymbol{a})$ is*

$$K(\boldsymbol{p}) = \frac{\det\langle (D^2\boldsymbol{h}\boldsymbol{e}_i\boldsymbol{e}_j)(\boldsymbol{a}), \boldsymbol{n}(\boldsymbol{p})\rangle}{\det\langle D\boldsymbol{h}\boldsymbol{e}_i, D\boldsymbol{h}\boldsymbol{e}_j\rangle(\boldsymbol{a})} = \frac{\det\langle (D^2\boldsymbol{h}\boldsymbol{e}_i\boldsymbol{e}_j)(\boldsymbol{a}), \boldsymbol{n}(\boldsymbol{p})\rangle}{|D\boldsymbol{h}\boldsymbol{e}_1 \times D\boldsymbol{h}\boldsymbol{e}_2|^2(\boldsymbol{a})}.$$

Proof. By (7.3.1), $K = \det S$, and the first part of the identity follows from the proposition, together with (7.5.2). For the second part, if θ denotes the angle between the two coordinate vector fields, then

$$\begin{aligned}\det\langle D\boldsymbol{h}\boldsymbol{e}_i, D\boldsymbol{h}\boldsymbol{e}_j\rangle &= |D\boldsymbol{h}\boldsymbol{e}_1|^2|D\boldsymbol{h}\boldsymbol{e}_2|^2 - \langle D\boldsymbol{h}\boldsymbol{e}_1, D\boldsymbol{h}\boldsymbol{e}_2\rangle^2\\ &= |D\boldsymbol{h}\boldsymbol{e}_1|^2|D\boldsymbol{h}\boldsymbol{e}_2|^2 - |D\boldsymbol{h}\boldsymbol{e}_1|^2|D\boldsymbol{h}\boldsymbol{e}_2|^2\cos^2\theta\\ &= |D\boldsymbol{h}\boldsymbol{e}_1|^2|D\boldsymbol{h}\boldsymbol{e}_2|^2\sin^2\theta\\ &= |D\boldsymbol{h}\boldsymbol{e}_1 \times D\boldsymbol{h}\boldsymbol{e}_2|^2\end{aligned}$$

by (1.6.2). □

Example 7.5.1. Consider the surface $M^2 \subset \mathbb{R}^3$ consisting of the graph of a function $f : U \subset \mathbb{R}^2 \to \mathbb{R}$; i.e, $M = \{(x, y, f(x, y)) \mid (x, y) \in U\}$. There is a natural parametrization $\boldsymbol{h} : U \to \mathbb{R}^3$ of all of M given by $\boldsymbol{h}(\boldsymbol{a}) = (\boldsymbol{a}, f(\boldsymbol{a}))$ for $\boldsymbol{a} \in U$.
$D\boldsymbol{h}\boldsymbol{e}_1$ and $D\boldsymbol{h}\boldsymbol{e}_2$ are the columns of the matrix

$$[D\boldsymbol{h}] = \begin{bmatrix} 1 & 0\\ 0 & 1\\ D_1f & D_2f\end{bmatrix},$$

so that

$$D^2\boldsymbol{h}\boldsymbol{e}_i = \begin{bmatrix} 0 & 0\\ 0 & 0\\ D_{1i}f & D_{2i}f\end{bmatrix}.$$

It follows that

$$(D^2\boldsymbol{h}\boldsymbol{e}_i)\boldsymbol{e}_j = \begin{bmatrix} 0\\ 0\\ D_{ij}f\end{bmatrix}, \qquad \boldsymbol{n} \circ \boldsymbol{h} = \frac{1}{\sqrt{1 + (D_1f)^2 + (D_1f)^2}}\begin{bmatrix} -D_1f\\ -D_2f\\ 1\end{bmatrix}.$$

By Proposition 7.5.1, we have

$$s \circ \boldsymbol{h} = \frac{1}{(1 + |\nabla f|^2)^{1/2}}\begin{bmatrix} D_{11}f & D_{12}f\\ D_{21}f & D_{22}f\end{bmatrix}, \quad g \circ \boldsymbol{h} = \begin{bmatrix} 1 + (D_1f)^2 & D_1fD_2f\\ D_1fD_2f & 1 + (D_2f)^2\end{bmatrix},$$

and by Corollary 7.5.1, the sectional curvature of M is given by

$$K \circ \boldsymbol{h} = \frac{\det H_f}{(1 + |\nabla f|^2)^2}. \tag{7.5.3}$$

In particular, the curvature and the determinant of the Hessian of f have the same sign. It is worth noting that this equation can also be derived from Theorem 7.3.3 if one writes $M = h^{-1}(0)$, where $h(x, y, z) = f(x, y) - z$.

7.6 Convexity and curvature

We have already encountered the notion of convexity in two of its forms: convexity of a function, and convexity of a set. There is a third one, which applies to hypersurfaces and is closely related to curvature. In order to introduce it, consider a hypersurface $M^n \subset \mathbb{R}^{n+1}$, a point $\boldsymbol{p} \in M$, and a unit vector $\boldsymbol{n} \perp M_{\boldsymbol{p}}$. The two *half-spaces at* $\boldsymbol{p}$ are the sets

$$H^+_{\boldsymbol{u}} = \{\boldsymbol{q} \in \mathbb{R}^{n+1} \mid \langle \pi_2(\boldsymbol{n}), \boldsymbol{q} - \boldsymbol{p}\rangle \geq 0\}, \text{ and}$$
$$H^-_{\boldsymbol{n}} = \{\boldsymbol{q} \in \mathbb{R}^{n+1} \mid \langle \pi_2(\boldsymbol{n}), \boldsymbol{q} - \boldsymbol{p}\rangle \leq 0\}.$$

Here, $\pi_2 : T\mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ is the usual projection. Notice that the intersection of the two half-spaces at any $\boldsymbol{p} \in M$ is the "affine space" $\boldsymbol{p} + \pi_2(M_{\boldsymbol{p}})$, which may be construed as a visual representation of the tangent space at $\boldsymbol{p}$. The ambient space itself is the union of these half-spaces. M is said to be *convex* if it is contained in one of the half-spaces at every point of M. This notion of convexity differs from the one given in Definition 2.4.5 which applies to arbitrary sets, not just hypersurfaces. It can, however, be shown that a compact, connected hypersurface M is convex if and only if the region consisting of all points "inside" and on M is convex in the previous sense of the word.

More generally, M is said to be *convex at* $\boldsymbol{p} \in M$ if there exists a neighborhood U of $\boldsymbol{p}$ in $\mathbb{R}^{n+1}$ such that $U \cap M$ is contained in one of the half-spaces at $\boldsymbol{p}$. A convex hypersurface is therefore convex at any of its points, but the converse is not true in general: for example, the hypersurface in $\mathbb{R}^2$ parametrized by $\boldsymbol{h} : (0, 2\pi) \to \mathbb{R}^2$, where $\boldsymbol{h}(t) = (t \cos t, t \sin t)$, is convex at every point, yet fails to be globally so, see Figure 7.4.

Convexity is closely related to curvature; to see how, recall a property of the second fundamental tensor that was touched upon previously: let $\boldsymbol{n}$ be a unit vector orthogonal to $M_{\boldsymbol{p}}$, and S the second fundamental tensor with respect to $\boldsymbol{n}$. Given $\boldsymbol{x} \in M_{\boldsymbol{p}}$,

$$\langle S\boldsymbol{x}, \boldsymbol{x}\rangle = \langle \boldsymbol{n}, \dot{\boldsymbol{c}}'(0)\rangle, \tag{7.6.1}$$

where $\boldsymbol{c}$ denotes any curve in M with $\dot{\boldsymbol{c}}(0) = \boldsymbol{x}$. This is because if $\boldsymbol{n}$ is a local normal unit field to M that equals $\boldsymbol{n}$ at $\boldsymbol{p}$, then

$$\langle S\boldsymbol{x}, \boldsymbol{x}\rangle = -\langle D_{\boldsymbol{x}}\boldsymbol{n}, \dot{\boldsymbol{c}}(0)\rangle = \langle \boldsymbol{n}(\boldsymbol{p}), \dot{\boldsymbol{c}}'(0)\rangle.$$

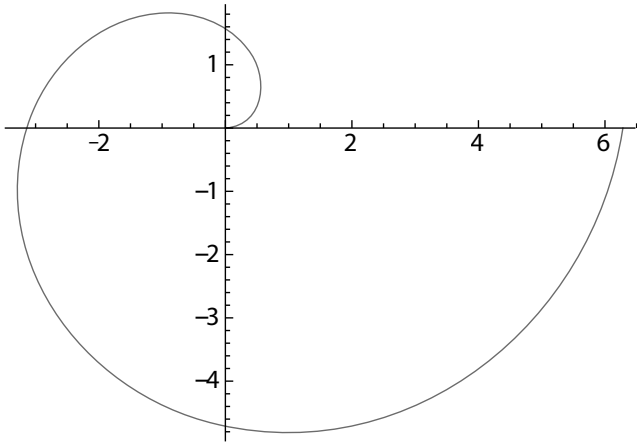


Fig. 7.4: A locally, but not globally, convex curve

Theorem 7.6.1. *If a hypersurface is convex at $\boldsymbol{p}$, then the sectional curvatures are nonnegative at $\boldsymbol{p}$.*

Proof. Let $\boldsymbol{n}$ denote a unit vector orthogonal to $M_{\boldsymbol{p}}$, S the second fundamental tensor with respect to $\boldsymbol{n}$. According to Exercise 7.17, it suffices to show that S is semi-definite. By hypothesis, there exists a neighborhood U of $\boldsymbol{p}$ in Euclidean space such that $M \cap U$ is contained either in $H^+_{\boldsymbol{n}}$ or in $H^-_{\boldsymbol{n}}$. Suppose the former holds. We will show that $\langle S\boldsymbol{x}, \boldsymbol{x}\rangle \geq 0$ for any $\boldsymbol{x} \in M_{\boldsymbol{p}}$. To see this, let $f : \mathbb{R}^{n+1} \to \mathbb{R}$ be given by

$$f(\boldsymbol{q}) = \langle \pi_2(\boldsymbol{n}), \boldsymbol{q} - \boldsymbol{p}\rangle.$$

The gradient of f is the parallel vector field on $\mathbb{R}^{n+1}$ that equals $\boldsymbol{n}$ at $\boldsymbol{p}$. If $\boldsymbol{c}$ is any curve in M with $\dot{\boldsymbol{c}}(0) = \boldsymbol{x} \in M_{\boldsymbol{p}}$, then $f \circ \boldsymbol{c}$ is nonnegative in a neighborhood of 0, and vanishes at 0. In particular, f has a minimum at 0, so that $(f \circ \boldsymbol{c})''(0) \geq 0$. Now, $(f \circ \boldsymbol{c})' = \langle \nabla f \circ \boldsymbol{c}, \dot{\boldsymbol{c}}\rangle$, and

$$(f \circ \boldsymbol{c})'' = \langle \nabla f \circ \boldsymbol{c}, \dot{\boldsymbol{c}}\rangle' = \langle \nabla f \circ \boldsymbol{c}, \dot{\boldsymbol{c}}'\rangle$$

since the gradient of f is parallel. Thus, by (7.6.1),

$$0 \leq (f \circ \boldsymbol{c})''(0) = \langle \boldsymbol{n}, \dot{\boldsymbol{c}}'(0)\rangle = \langle S\boldsymbol{x}, \boldsymbol{x}\rangle,$$

as claimed. The case when $M \cap U$ is contained in $H^-_{\boldsymbol{n}}$ is similar, except that $f \circ \boldsymbol{c}$ now has a maximum at 0, so that $\langle S\boldsymbol{x}, \boldsymbol{x}\rangle = (f \circ \boldsymbol{c})''(0) \leq 0$. □

The converse is, in general, not true: for example, the right cylinder in $\mathbb{R}^3$ over the sine curve in $\mathbb{R}^2$ can be parametrized by $(s, t) \mapsto (s, \sin s, t)$. Corollary 7.5.1 implies that the surface is flat, even though it is clearly not convex. A more striking example can be found in Exercise 7.24. The converse does hold, however, under stronger assumptions:

Theorem 7.6.2. *Let M^n be a hypersurface. If the sectional curvatures are positive at some point, then M is convex at that point.*

Proof. Let $\boldsymbol{n}$ denote a unit normal field to M in a neighborhood of a point $\boldsymbol{p}$ where the curvature is positive. By Theorem 7.3.1, the second fundamental form S with respect

to $\boldsymbol{n}(\boldsymbol{p})$ is definite, and we may assume, without loss of generality, that it is positive definite. Consider the function f defined in some connected neighborhood of $\boldsymbol{p}$ by $f(\boldsymbol{q}) = \langle \boldsymbol{q} - \boldsymbol{p}, \pi_2(\boldsymbol{n}(\boldsymbol{p}))\rangle$. If $\boldsymbol{c}$ is any regular curve in M passing through $\boldsymbol{p}$ at $t = 0$, then $(f \circ \boldsymbol{c})'(0) = \langle \pi_2(\boldsymbol{n}(\boldsymbol{p})), \boldsymbol{c}'(0)\rangle = 0$, and as in the proof of the previous theorem, $(f \circ \boldsymbol{c})'' = \langle S\dot{\boldsymbol{c}}(0), \dot{\boldsymbol{c}}(0)\rangle$ which is positive by assumption. Thus, $f \circ \boldsymbol{c}$ has a minimum at 0, so that all points of M close enough to $\boldsymbol{p}$ will lie in the half-space $H^+_{\boldsymbol{n}(\boldsymbol{p})}$. This completes the argument. □

Example 7.6.1. If we visualize the tangent plane of a hypersurface M in $\mathbb{R}^{n+1}$ at $\boldsymbol{p}$ as the affine subspace $\boldsymbol{p} + \pi_2(M_{\boldsymbol{p}})$ (which is the intuitive interpretation for $n = 2$), then convexity at $\boldsymbol{p}$ is equivalent to saying that in a neighborhood of $\boldsymbol{p}$, M lies on one side of this affine plane. Thus, if a hypersurface does not have curvature ≥ 0 at some point, then it must cross its tangent plane there. This implies in particular something about the shape of minimal hypersurfaces: namely, they must cross their tangent plane at every point where not all curvatures vanish. Indeed, if the hypersurface does not cross its tangent space somewhere, then the second fundamental tensor is semi-definite there. Since the trace is zero, the tensor itself must vanish.

7.7 Ruled surfaces

A *ruled surface* in $\mathbb{R}^3$ is a hypersurface parametrized by a map $\boldsymbol{h}$ of the form $\boldsymbol{h}(s,t) = \boldsymbol{c}_1(s) + t\boldsymbol{c}_2(s)$, where $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are curves in $\mathbb{R}^3$. Notice that the curves $\boldsymbol{h}_s$, where $\boldsymbol{h}_s(t) = \boldsymbol{h}(s,t)$ are straight lines; they are called the *rulings* of the surface. Since

$$D\boldsymbol{h}(s,t)\boldsymbol{e}_1 = \boldsymbol{c}_1'(s) + t\boldsymbol{c}_2'(s), \qquad D\boldsymbol{h}(s,t)\boldsymbol{e}_2 = \boldsymbol{c}_2(s),$$

the two vectors above must be linearly independent for all s, t if $\boldsymbol{h}$ is to be a parametrization. This will be the case (for small t at least) if for example $\boldsymbol{c}_1'(s)$ and $\boldsymbol{c}_2(s)$ are linearly independent. An interesting feature of these surfaces is that they are always nonpositively curved:

Proposition 7.7.1. *The sectional curvature of a ruled surface is given by*

$$(K \circ \boldsymbol{h})(s,t) = -\frac{\langle \boldsymbol{c}_1'(s), \boldsymbol{c}_2(s) \times \boldsymbol{c}_2'(s)\rangle^2}{|(\boldsymbol{c}_1'(s) + t\boldsymbol{c}_2'(s)) \times \boldsymbol{c}_2(s)|^4}.$$

Proof. Notice that if $\boldsymbol{h}_t$ denotes the map $s \mapsto \boldsymbol{h}(s,t)$, then the formula we aim to establish becomes

$$K \circ \boldsymbol{h}_t = -\frac{\langle \boldsymbol{c}_1', \boldsymbol{c}_2 \times \boldsymbol{c}_2'\rangle^2}{|(\boldsymbol{c}_1' + t\boldsymbol{c}_2') \times \boldsymbol{c}_2|^4}.$$

The unit normal field determined by the parametrization is

$$\boldsymbol{n} \circ \boldsymbol{h}_t = \frac{D\boldsymbol{h}\boldsymbol{e}_1 \times D\boldsymbol{h}\boldsymbol{e}_2}{|D\boldsymbol{h}\boldsymbol{e}_1 \times D\boldsymbol{h}\boldsymbol{e}_2|} = \frac{(\boldsymbol{c}_1' + t\boldsymbol{c}_2') \times \boldsymbol{c}_2}{|(\boldsymbol{c}_1' + t\boldsymbol{c}_2') \times \boldsymbol{c}_2|}.$$

Now, $D^2\boldsymbol{h}\boldsymbol{e}_2\boldsymbol{e}_2 = \mathbf{0}$, so the matrix $\langle D^2\boldsymbol{h}\boldsymbol{e}_i\boldsymbol{e}_j, \boldsymbol{n}\circ\boldsymbol{h}\rangle$ has determinant

$$\det\langle D^2\boldsymbol{h}\boldsymbol{e}_i\boldsymbol{e}_j, \boldsymbol{n}\circ\boldsymbol{h}\rangle = -\langle D^2\boldsymbol{h}\boldsymbol{e}_1\boldsymbol{e}_2, \boldsymbol{n}\circ\boldsymbol{h}\rangle^2 = -\frac{\langle \boldsymbol{c}_2', (\boldsymbol{c}_1' + t\boldsymbol{c}_2')\times \boldsymbol{c}_2\rangle^2}{|(\boldsymbol{c}_1' + t\boldsymbol{c}_2')\times \boldsymbol{c}_2|^2},$$

and by Corollary 7.5.1, $K\circ\boldsymbol{h}_t = -\langle \boldsymbol{c}_2', (\boldsymbol{c}_1' + t\boldsymbol{c}_2')\times \boldsymbol{c}_2\rangle^2/|(\boldsymbol{c}_1' + t\boldsymbol{c}_2')\times \boldsymbol{c}_2|^4$. The definition of the cross-product implies that $\langle \boldsymbol{x}\times\boldsymbol{y}, \boldsymbol{z}\rangle = \langle \boldsymbol{y}\times\boldsymbol{z}, \boldsymbol{x}\rangle$ for $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^3$. Thus,

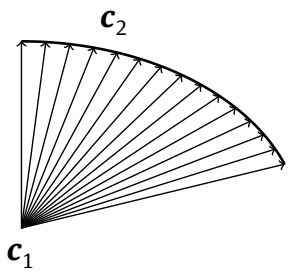
$$\langle (\boldsymbol{c}_1' + t\boldsymbol{c}_2')\times \boldsymbol{c}_2, \boldsymbol{c}_2'\rangle = \langle \boldsymbol{c}_2\times\boldsymbol{c}_2', \boldsymbol{c}_1' + t\boldsymbol{c}_2'\rangle = \langle \boldsymbol{c}_2\times\boldsymbol{c}_2', \boldsymbol{c}_1'\rangle$$

since $\boldsymbol{c}_2' \perp \boldsymbol{c}_2\times\boldsymbol{c}_2'$, and the result follows. □

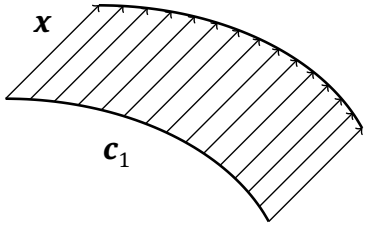
There are two special cases worth mentioning:

(1) $\boldsymbol{c}_1' \equiv 0$. t must then be different from zero, and $\boldsymbol{c}_1$ is a constant curve $\boldsymbol{p}$. M is a generalized cone over $\boldsymbol{c}_2$ with vertex $\boldsymbol{p}$ deleted. The cone is flat.

(2) $\boldsymbol{c}_2' \equiv 0$. Now $\boldsymbol{c}_2$ is a constant $\boldsymbol{x}$ and M is a generalized cylinder over $\boldsymbol{c}_1$ with axis parallel to $\boldsymbol{x}$. Again, M is flat.



Ruled surface with $\boldsymbol{c}_1$ constant



Ruled surface with $\boldsymbol{c}_2$ = constant $\boldsymbol{x}$

These special cases having been dealt with, we will assume that $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are regular and parametrized by arc length. We may also normalize $\boldsymbol{c}_2$ so that it lies on the unit sphere. In this case, $\boldsymbol{c}_2$ is called the *directrix*. The formula for the curvature simplifies substantially if we assume that $\langle \boldsymbol{c}_1', \boldsymbol{c}_2'\rangle \equiv 0$. Although it is not immediately obvious, it turns out this may always be done; i.e., $\boldsymbol{c}_2$ may be replaced, if necessary, by another curve $\boldsymbol{\gamma}_2$ satisfying $\langle \boldsymbol{c}_1', \boldsymbol{\gamma}_2'\rangle \equiv 0$. Under these assumptions, we have that

$$\begin{aligned}|(\boldsymbol{c}_1' + t\boldsymbol{c}_2')\times \boldsymbol{c}_2|^2 &= |\boldsymbol{c}_1' + t\boldsymbol{c}_2'|^2|\boldsymbol{c}_2'|^2 - \langle \boldsymbol{c}_1' + t\boldsymbol{c}_2', \boldsymbol{c}_2\rangle^2\\ &= |\boldsymbol{c}_1'|^2 + t^2|\boldsymbol{c}_2'|^2 - \langle \boldsymbol{c}_1', \boldsymbol{c}_2\rangle^2 = |\boldsymbol{c}_1'|^2|\boldsymbol{c}_2|^2 + t^2 - \langle \boldsymbol{c}_1', \boldsymbol{c}_2\rangle^2\\ &= t^2 + |\boldsymbol{c}_1'\times \boldsymbol{c}_2|^2.\end{aligned}$$

Now, $\boldsymbol{c}_2'$ is orthogonal to $\boldsymbol{c}_1', \boldsymbol{c}_2$, and the same is true for $\boldsymbol{c}_1'\times\boldsymbol{c}_2$. Assuming linear independence of $\boldsymbol{c}_1', \boldsymbol{c}_2$, it follows that $\boldsymbol{c}_1'\times\boldsymbol{c}_2$ and $\boldsymbol{c}_2'$ are orthogonal to a common plane, and therefore linearly dependent. Thus, $\boldsymbol{c}_1'\times\boldsymbol{c}_2 = \langle \boldsymbol{c}_1'\times\boldsymbol{c}_2, \boldsymbol{c}_2'\rangle\boldsymbol{c}_2' = \langle \boldsymbol{c}_2\times\boldsymbol{c}_2', \boldsymbol{c}_1'\rangle\boldsymbol{c}_2'$; i.e.,

$|\boldsymbol{c}_1' \times \boldsymbol{c}_2|^2 = \langle \boldsymbol{c}_1', \boldsymbol{c}_2 \times \boldsymbol{c}_2' \rangle^2$, and the formula for the curvature becomes

$$K \circ \boldsymbol{h} = -\frac{|\boldsymbol{c}_1' \times \boldsymbol{c}_2|^2}{(t^2 + |\boldsymbol{c}_1' \times \boldsymbol{c}_2|^2)^2}. \tag{7.7.1}$$

Notice that since $\boldsymbol{c}_1'$ and $\boldsymbol{c}_2$ are assumed to be linearly independent, the curvature is strictly negative.

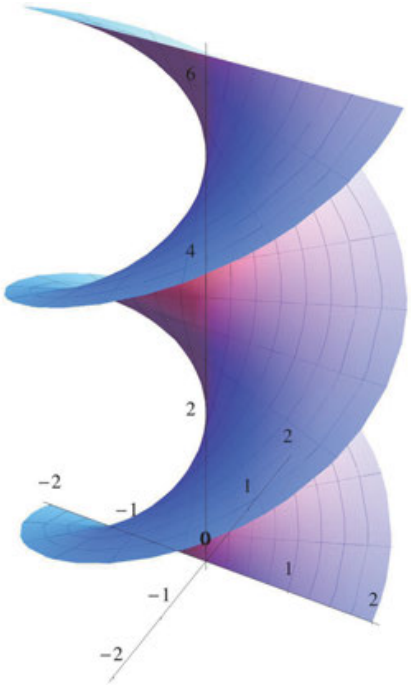


Fig. 7.5: A helicoid

One example of a ruled surface is the Möbius strip from Section 4.1, with $\boldsymbol{h}(s,t) = (\cos s, \sin s, 0) + t(\cos(s/2), \sin(s/2), \sin(s/2))$. Another is the hyperboloid $(x/a)^2 + (y/b)^2 - (z/c)^2 = 1$ from Section 4.3, which may be parametrized by $\boldsymbol{h}(s,t) = (a\cos s, b\sin s, 0) + t(-a\sin s, b\cos s, c)$. One that we have not encountered before is the *helicoid*: For any $s \in \mathbb{R}$, consider the Euclidean motion $\boldsymbol{k}_s$ of $\mathbb{R}^3$, where

$$\boldsymbol{k}_s(x,y,z) = \begin{bmatrix} \cos s & -\sin s & 0 \\ \sin s & \cos s & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ s \end{bmatrix}.$$

$\boldsymbol{k}_s$ is called a *glide rotation* as it rotates a point by angle s around the z-axis and translates it by s along that same axis. The so-called orbit $\{\boldsymbol{k}_s(\boldsymbol{p}) \mid s \in \mathbb{R}\}$ of a point $\boldsymbol{p}$ is a helix. The helicoid itself is just the orbit of the whole x-axis; i.e., $\boldsymbol{h}(s,t) = \boldsymbol{k}_s(t,0,0)$. According to (7.7.1), the curvature of the helicoid is

$$K \circ \boldsymbol{h}(s,t) = -\frac{1}{(1+t^2)^2},$$

which only depends on the distance $|t|$ from the point $\boldsymbol{h}(s,t)$ to the z-axis. The helicoid is actually related to the catenoid introduced in Exercise 3.28. Both are minimal surfaces, and either one can be smoothly and isometrically deformed into the other via $\boldsymbol{F}$, where

$$\begin{aligned}\boldsymbol{F}(\theta,s,t) = (&\cos\theta\,\sinh t\sin s + \sin\theta\,\cosh t\cos s, -\cos\theta\,\sinh t\cos s\\ &+\sin\theta\,\cosh t\sin s, s\cos\theta + t\sin\theta).\end{aligned}$$

When $\theta = 0$, one obtains the parametrization of the helicoid above rotated by $-\pi/2$:

$$\boldsymbol{F}(0,s,t)^T = \begin{bmatrix}\sinh t\sin s\\ \sinh t(-\cos s)\\ s\end{bmatrix} = \begin{bmatrix}0&1&0\\-1&0&0\\0&0&1\end{bmatrix}\boldsymbol{k}_s(\sinh t,0,0)^T.$$

Similarly, $\theta = \pi/2$ yields the catenoid. For each fixed θ_0, the map $(s,t)\mapsto \boldsymbol{F}(\theta_0,s,t)$ parametrizes a minimal surface. Notice that

$$\boldsymbol{F}(\theta,s,t) = \cos\theta\,\boldsymbol{F}(0,s,t) + \sin\theta\,\boldsymbol{F}(\frac{\pi}{2},s,t),$$

see also Exercise 7.25.

7.8 Surfaces of revolution

Surfaces of revolution, where the graph of a function of one variable is rotated about an axis, were introduced in Examples 3.1.1. It is convenient to allow for more general curves than just graphs. For clarity of notation in the sometimes rather complicated formulas that follow, we will use subscripts rather than superscripts to denote the component functions of the curve.

Definition 7.8.1. Let $\boldsymbol{\gamma} = (\gamma_1, 0, \gamma_2) : (a,b)\to\mathbb{R}^3$ denote a curve whose image lies in the x-z plane. The *surface of revolution M with profile curve $\boldsymbol{\gamma}$* is the surface parametrized by $\boldsymbol{h} : (a,b)\times[0,2\pi)\to\mathbb{R}^3$, where

$$\boldsymbol{h}(u,v) = (\gamma_1(u)\cos v, \gamma_1(u)\sin v, \gamma_2(u)).$$

Thus, the surface is obtained by rotating the image of $\boldsymbol{\gamma}$ about the z axis. The curves $u\mapsto\boldsymbol{h}(u,v)$ and $v\mapsto\boldsymbol{h}(u,v)$ are called *meridians* and *parallels* respectively. The meridian $u\mapsto\boldsymbol{h}(u,v_0)$ is the profile curved rotated by angle v_0, and the parallel $v\mapsto\boldsymbol{h}(u_0,v)$ is the circle obtained by rotating the point $\boldsymbol{\gamma}(u_0)$ about the z-axis.

The coordinate fields $\frac{\partial}{\partial x^i} = \boldsymbol{h}_*\circ\boldsymbol{D}_i\circ\boldsymbol{h}^{-1}$ are determined by the vector fields

$$\begin{aligned}\boldsymbol{h}_*\boldsymbol{D}_1(u,v) &= \gamma_1'(u)\cos v(\boldsymbol{D}_1\circ\boldsymbol{h})(u,v) + \gamma_1'(u)\sin v(\boldsymbol{D}_2\circ\boldsymbol{h})(u,v)\\ &\quad+\gamma_2'(u)(\boldsymbol{D}_3\circ\boldsymbol{h})(u,v)\\ \boldsymbol{h}_*\boldsymbol{D}_2(u,v) &= -\gamma_1(u)\sin v(\boldsymbol{D}_1\circ\boldsymbol{h})(u,v) + \gamma_1(u)\cos v(\boldsymbol{D}_2\circ\boldsymbol{h})(u,v)\end{aligned}\tag{7.8.1}$$

along $\boldsymbol{h}$. The latter are just the tangent fields to the meridians and parallels. Notice that the second one is parallel to the x-y plane and that they are mutually orthogonal. Their cross-product $\boldsymbol{h}_*\boldsymbol{D}_1 \times \boldsymbol{h}_*\boldsymbol{D}_2$, yields, after normalizing, a unit normal field $\boldsymbol{N}$ along $\boldsymbol{h}$ given by

$$\begin{aligned}\boldsymbol{N}(u,v) = \frac{1}{(\gamma_1'^2 + \gamma_2'^2)^{1/2}(u)}\Big(\gamma_2'(u)\cos v\boldsymbol{D}_1 + \gamma_2'(u)\sin v\boldsymbol{D}_2 \\ - \gamma_1'(u)\boldsymbol{D}_3\Big) \circ \boldsymbol{h}(u,v).\end{aligned} \tag{7.8.2}$$

Theorem 7.8.1. *The sectional curvature of the surface of revolution parametrized by* $\boldsymbol{h}$ *is*

$$K \circ \boldsymbol{h} = \frac{(\gamma_1'\gamma_2'' - \gamma_2'\gamma_1'')\gamma_2'}{\gamma_1(\gamma_1'^2 + \gamma_2'^2)^2}.$$

In particular, if $\boldsymbol{\gamma}$ *is parametrized by arc-length, then* $K \circ \boldsymbol{h} = -\gamma_1''/\gamma_1$.

Proof. Using (7.8.2), the covariant derivatives of the unit normal field along meridians and parallels are given by

$$D_{\boldsymbol{D}_1}\boldsymbol{N} = \frac{\gamma_1'\gamma_2'' - \gamma_2'\gamma_1''}{(\gamma_1'^2 + \gamma_2'^2)^{3/2}}\boldsymbol{h}_*\boldsymbol{D}_1, \qquad D_{\boldsymbol{D}_2}\boldsymbol{N} = \frac{\gamma_2'}{\gamma_1(\gamma_1'^2 + \gamma_2'^2)^{1/2}}\boldsymbol{h}_*\boldsymbol{D}_2.$$

Thus, the coordinate vector fields are eigenvector fields at each point of M of the second fundamental tensor, with corresponding principal curvatures

$$\lambda_1 = \frac{\gamma_1'\gamma_2'' - \gamma_2'\gamma_1''}{(\gamma_1'^2 + \gamma_2'^2)^{3/2}}, \qquad \lambda_2 = \frac{\gamma_2'}{\gamma_1(\gamma_1'^2 + \gamma_2'^2)^{1/2}}. \tag{7.8.3}$$

Since the sectional (and Gaussian) curvature is the product of the principal curvatures, the result follows. When $\boldsymbol{\gamma}$ is parametrized by arc length, $\sum_i \gamma_i'\gamma_i'' = 0$, so that

$$K \circ \boldsymbol{h} = \frac{\gamma_1'\gamma_2'\gamma_2'' - \gamma_1''\gamma_2'^2}{\gamma_1} = \frac{\gamma_1'(-\gamma_1'\gamma_1'') - \gamma_1''(1 - \gamma_1'^2)}{\gamma_1} = -\frac{\gamma_1''}{\gamma_1}. \tag{7.8.4}$$

□

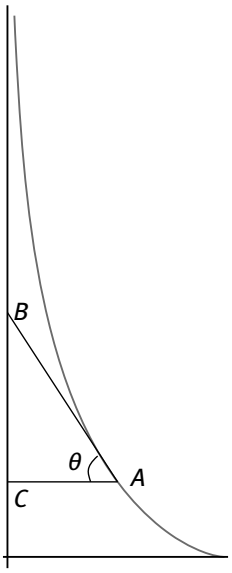


Fig. 7.6: A tractrix

Example 7.8.1. Consider an object located at $(a, 0)$, $a > 0$, in the plane, attached to a rope of length a whose other end is held by someone standing initially at the origin $(0, 0)$. A *tractrix* (from the Latin verb "*trahere*" – to draw) is the path traced by the object as it is being pulled by the person walking along the y-axis.
Denoting by $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ the parametrization by arc-length of the tractrix with $\boldsymbol{\gamma}(0) = (a, 0)$, we see from the figure that

$$a \equiv AB = \frac{AC}{\cos\theta} = -\frac{AC}{\gamma_1'} = -\frac{\gamma_1}{\gamma_1'},$$

with the minus sign being due to the fact that

$$\gamma_1' = \langle \boldsymbol{\gamma}', \boldsymbol{e}_1 \rangle = |\boldsymbol{\gamma}'| \cos(\pi - \theta) = -\cos\theta\,.$$

Thus, $\gamma_1' = -(1/a)\gamma_1$, and $\gamma_1(t) = ae^{-t/a}$. Furthermore,

$$1 = |\boldsymbol{\gamma}'|^2(t) = (ae^{-t/a})^2 + \gamma_2'^{\,2}(t),$$

so that $\gamma_2'(t) = \pm\sqrt{1 - (ae^{-t/a})^2}$. If the person walks along the positive portion of the y-axis, the tractrix is then given by

$$\boldsymbol{\gamma}(t) = \left(ae^{-t/a}, \int_0^t \sqrt{1 - (ae^{-s/a})^2}\, \mathrm{d}s \right), \qquad t \geq 0.$$

The surface of revolution generated by the tractrix is called a *pseudosphere*. The name was coined in 1868 by the Italian mathematician Eugenio Beltrami who proposed it as a model of non-Euclidean geometry. By (7.8.4), this pseudosphere has constant negative curvature $K = -1/a^2$.

One appealing feature of surfaces of revolution is that, in contrast to general surfaces, geodesics are readily described. Notice first that any meridian is the image of a geodesic. This follows from Proposition 3.11.2, since it is (part of) the fixed point set of reflection in the plane that contains the meridian and the z-axis. For parallels, the situation is different: we claim that $\boldsymbol{c}$, where $\boldsymbol{c}(t) = \boldsymbol{h}(u_0, t)$, is a geodesic if and only if u_0 is a critical point of γ_1. To see this, notice that $\dot{\boldsymbol{c}}' = -\alpha \cos \boldsymbol{D}_1 \circ \boldsymbol{c} - \alpha \sin \boldsymbol{D}_2 \circ \boldsymbol{c}$, where $\alpha := \gamma_1(u_0)$. On the other hand, $\boldsymbol{c}$ is a geodesic if and only if $\boldsymbol{N} \circ \boldsymbol{c}$ and $\dot{\boldsymbol{c}}'$ are linearly dependent. Comparing with (7.8.2) now yields the claim.
The other geodesics can be described by means of the following

Theorem 7.8.2 (Clairaut). *Let $\boldsymbol{c} : I \to M$ denote a normal geodesic in M, and $\boldsymbol{\rho} = \boldsymbol{h}^{-1} \circ \boldsymbol{c} : I \to \mathbb{R}^2$. If $\alpha(t) = \sphericalangle(\dot{\boldsymbol{c}}(t), \boldsymbol{h}_*(\boldsymbol{D}_2 \circ \boldsymbol{\rho})(t))$ is the angle between $\dot{\boldsymbol{c}}(t)$ and the parallel through $\boldsymbol{c}(t)$, and if $r = ((u^1)^2 + (u^2)^2)^{1/2}$ denotes the distance from a point to the z-axis, then $(r \circ \boldsymbol{c}) \cos\alpha$ is constant.*
Conversely, if $\boldsymbol{c}$ is a curve parametrized by arc length for which $(r \circ \boldsymbol{c}) \cos\alpha$ is constant, and $\boldsymbol{c}$ is not a parallel, then $\boldsymbol{c}$ is a geodesic.

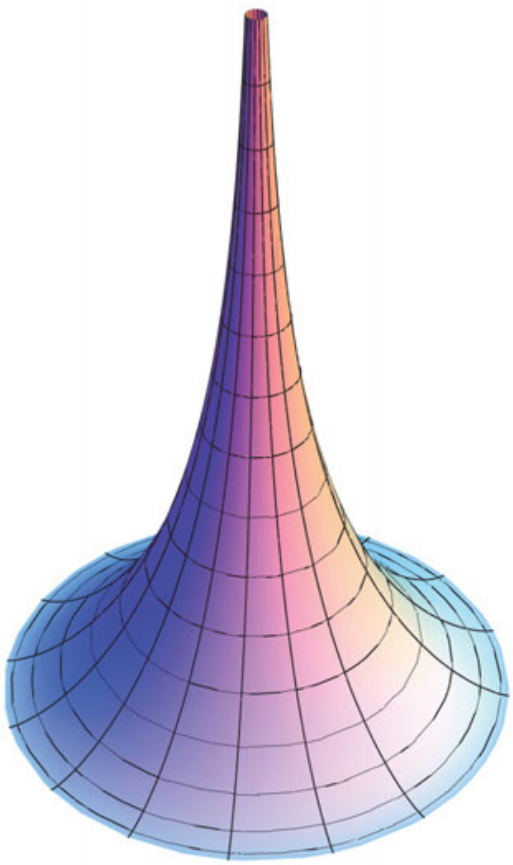


Fig. 7.7: A pseudosphere

Proof. We will, as we did with $\boldsymbol{c}$, denote with a subscript the component functions $\rho_i = u^i \circ \boldsymbol{\rho}$, $i = 1, 2$, of $\boldsymbol{\rho}$. We will also abbreviate $\boldsymbol{h}_*(\boldsymbol{D}_i \circ \boldsymbol{\rho})$ by $\boldsymbol{h}_*\boldsymbol{D}_i \circ \boldsymbol{\rho}$. Then

$$\begin{aligned}
\dot{\boldsymbol{c}} &= \rho_1' \boldsymbol{h}_* \boldsymbol{D}_1 \circ \boldsymbol{\rho} + \rho_2' \boldsymbol{h}_* \boldsymbol{D}_2 \circ \boldsymbol{\rho} \\
&= \Big((\gamma_1 \circ \rho_1)' \cos\rho_2 - (\gamma_1 \circ \rho_1)(\sin\rho_2)\rho_2'\Big)\boldsymbol{D}_1 \circ \boldsymbol{c} \\
&\quad + \Big((\gamma_1 \circ \rho_1)' \sin\rho_2 + (\gamma_1 \circ \rho_1)(\cos\rho_2)\rho_2'\Big)\boldsymbol{D}_2 \circ \boldsymbol{c} + \rho_1'(\gamma_2' \circ \rho_1)\boldsymbol{D}_3 \circ \boldsymbol{c} \\
&= \Big((\gamma_1 \circ \rho_1)\cos\rho_2\Big)'\boldsymbol{D}_1 \circ \boldsymbol{c} + \Big((\gamma_1 \circ \rho_1)\sin\rho_2\Big)'\boldsymbol{D}_2 \circ \boldsymbol{c} + (\gamma_2 \circ \rho_1)'\boldsymbol{D}_3 \circ \boldsymbol{c}.
\end{aligned}$$

Now, (7.8.1) implies

$$(\boldsymbol{h}_*\boldsymbol{D}_2 \circ \boldsymbol{\rho})' = -\Big((\gamma_1 \circ \rho_1)\sin\rho_2\Big)'\boldsymbol{D}_1 \circ \boldsymbol{c} + \Big((\gamma_1 \circ \rho_1)\cos\rho_2\Big)'\boldsymbol{D}_2 \circ \boldsymbol{c},$$

so that

$$\langle \dot{\boldsymbol{c}}, \boldsymbol{h}_*\boldsymbol{D}_2 \circ \boldsymbol{\rho}\rangle' = \langle \dot{\boldsymbol{c}}, (\boldsymbol{h}_*\boldsymbol{D}_2 \circ \boldsymbol{\rho})'\rangle = 0,$$

and $\langle \dot{\boldsymbol{c}}, \boldsymbol{h}_*\boldsymbol{D}_2 \circ \boldsymbol{\rho}\rangle$ is constant. But the latter also equals

$$|\dot{\boldsymbol{c}}||\boldsymbol{h}_*\boldsymbol{D}_2 \circ \boldsymbol{\rho}|\cos\alpha = |\gamma_1 \circ \rho_1|\cos\alpha = (r \circ \boldsymbol{c})\cos\alpha.$$

Conversely, if $(r \circ \boldsymbol{c})\cos\alpha$ is constant, then by the above calculation

$$\langle \dot{\boldsymbol{c}}', \boldsymbol{h}_*\boldsymbol{D}_2 \circ \boldsymbol{\rho}\rangle = \langle \dot{\boldsymbol{c}}, \boldsymbol{h}_*\boldsymbol{D}_2 \circ \boldsymbol{\rho}\rangle' - \langle \dot{\boldsymbol{c}}, (\boldsymbol{h}_*\boldsymbol{D}_2 \circ \boldsymbol{\rho})'\rangle = -\langle \dot{\boldsymbol{c}}, (\boldsymbol{h}_*\boldsymbol{D}_2 \circ \boldsymbol{\rho})'\rangle = 0.$$

Now, $\dot{\boldsymbol{c}}'$ is orthogonal to $\dot{\boldsymbol{c}}$ because $\boldsymbol{c}$ is parametrized by arc length. It follows that at any t_0 where $\dot{\boldsymbol{c}}(t_0)$ and $\boldsymbol{h}_*\boldsymbol{D}_2 \circ \boldsymbol{\rho}(t_0)$ are linearly independent, $\dot{\boldsymbol{c}}'(t_0)$ is orthogonal to the surface; i.e., $(\nabla_{\boldsymbol{D}}\dot{\boldsymbol{c}})(t_0) = 0$. On the other hand, if $\dot{\boldsymbol{c}}(t_0)$ and $\boldsymbol{h}_*\boldsymbol{D}_2 \circ \boldsymbol{\rho}(t_0)$ are linearly dependent, then there exists a sequence $t_n \to t_0$ such that $\dot{\boldsymbol{c}}(t_n)$ and $\boldsymbol{h}_*\boldsymbol{D}_2 \circ \boldsymbol{\rho}(t_n)$ are linearly independent, for otherwise $\boldsymbol{c}$ would be a meridian. Then $(\nabla_{\boldsymbol{D}}\dot{\boldsymbol{c}})(t_n) = 0$, and by continuity, so is $(\nabla_{\boldsymbol{D}}\dot{\boldsymbol{c}})(t_0)$. □

Clairaut's theorem yields a fairly complete qualitative description of the geodesics on M. Let us illustrate this in the case when M is a paraboloid of revolution; specifically, the image of the curve $\boldsymbol{c}$ is the the set $\{(x, 0, x^2) \mid x \geq 0\}$. All the meridians are the (images of the) geodesics emanating from the vertex $\mathbf{0}$. Since none of these intersect, any such geodesic $\boldsymbol{c} : [0, \infty) \to M$ is a *ray*; i.e., $\boldsymbol{c}$ is the shortest curve from the vertex to $\boldsymbol{c}(t)$ for any $t > 0$. Assume then that $\boldsymbol{c} : (-\infty, \infty) \to M$ is a geodesic that is not a meridian. Consider the lowest point on the image of $\boldsymbol{c}$. The existence of such a point is discussed in Exercise 7.10. Since the distance to the z-axis increases with height, at this lowest point, $r \circ \boldsymbol{c}$ is minimal and the angle α is zero. After reparametrizing $\boldsymbol{c}$ if necessary, we may assume this occurs at $t = 0$, so that

$$(r \circ \boldsymbol{c})(t) \cos \alpha(t) = (r \circ \boldsymbol{c})(0) =: k.$$

Since no parallel is a geodesic, $(r \circ \boldsymbol{c})(t) > (r \circ \boldsymbol{c})(0)$ for $t > 0$, and $\boldsymbol{c}$ rises; i.e., $u^3 \circ \boldsymbol{c}$ increases. We claim $(u^3 \circ \boldsymbol{c})(t) \to \infty$ as $t \to \infty$. To see this, let $\boldsymbol{c}_0(t)$ denote the point on the (image of the) profile curve $\boldsymbol{\gamma}$ that lies on the same parallel as $\boldsymbol{c}(t)$; i.e., $\boldsymbol{c}_0(t)$ is obtained by rotating $\boldsymbol{c}(t)$ about the z-axis by an angle $-\gamma_2(t)$. Thus,

$$\boldsymbol{c}_0 = \boldsymbol{h}(\rho_1, 0) = (\gamma_1 \circ \rho_1, 0, \gamma_2 \circ \rho_1) = \boldsymbol{\gamma} \circ \rho_1,$$

and $\dot{\boldsymbol{c}}_0 = \rho_1' \, \dot{\boldsymbol{\gamma}} \circ \rho_1$. The claim follows once we show that ρ_1 is unbounded. Assuming $\boldsymbol{c}$ is parametrized by arc length, the speed of $\boldsymbol{c}_0$ is

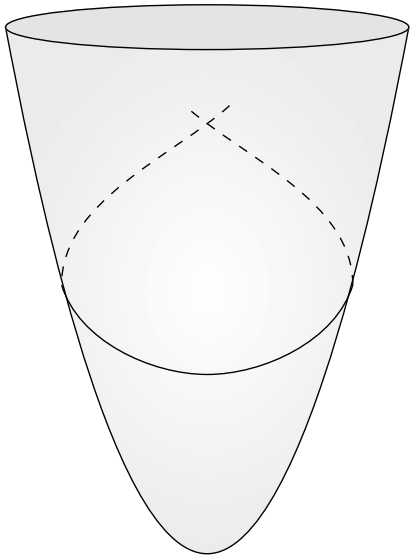
$$\begin{aligned} |\dot{\boldsymbol{c}}_0| = \rho_1' &= \langle \dot{\boldsymbol{c}}, \boldsymbol{h}_*\boldsymbol{D}_1 \circ \boldsymbol{\rho} \rangle = |\dot{\boldsymbol{c}}||\boldsymbol{h}_*\boldsymbol{D}_1 \circ \boldsymbol{\rho}| \sin \alpha = \sin \alpha \\ &= \left(1 - \frac{k^2}{(\gamma_1 \circ \rho_1)^2}\right)^{1/2}, \end{aligned} \tag{7.8.5}$$

since

$$\sin \alpha = (1 - \cos^2 \alpha)^{1/2} = \left(1 - \frac{k^2}{(r \circ \boldsymbol{c})^2}\right)^{1/2} = \left(1 - \frac{k^2}{(r \circ \boldsymbol{c}_0)^2}\right)^{1/2}$$

and $r \circ \boldsymbol{c}_0 = \gamma_1 \circ \rho_1$. Differentiating (7.8.5) yields

$$\rho_1'' = \left(1 - \frac{k^2}{(\gamma_1 \circ \rho_1)^2}\right)^{-1/2} k^2 (\gamma_1 \circ \rho_1)^{-3} (\gamma_1' \circ \rho_1) \rho_1'.$$



Now, γ_1 and γ_1' are both positive in our case, so ρ_1'' and ρ_1' have the same sign. Furthermore, $\rho_1'(0) = 0$ and $\rho_1'(t) > 0$ for small $t > 0$. By the mean value theorem, $\rho_1'(t) > 0$ for all $t > 0$: otherwise, letting t_0 denote the infimum of those $t > 0$ such that $\rho_1(t) \leq 0$, we have that $\rho_1'(t_0) \leq 0$, and $\rho_1'(t) > 0$ for all $t < t_0$. This is impossible, since the mean value theorem guarantees the existence of some $t \in (0, t_0)$ satisfying

$$\rho_1''(t) = \frac{\rho_1'(t_0) - \rho_1'(0)}{t_0} \leq 0,$$

and therefore, $\rho_1'(t)$ is also nonpositive. Thus, $\rho_1', \rho_1'' > 0$ always; i.e., ρ_1 is a convex function, hence unbounded. This establishes the claim.

Notice also that reflection R in the plane containing the meridian through $\boldsymbol{c}(0)$ is an isometry that must leave the image of $\boldsymbol{c}$ invariant: $R \circ \boldsymbol{c}_{|[0,\infty)}$ is a geodesic with initial tangent vector $-\dot{\boldsymbol{c}}(0)$, so that by uniqueness, it must equal $t \mapsto \boldsymbol{c}(-t)$.

Further examples are explored in the exercises.

7.9 Exercises

7.1. Suppose $\boldsymbol{h}$ and $\boldsymbol{k}$ are two overlapping parametrizations – in the sense that their images have nonempty intersection – of $M^n \subset \mathbb{R}^{n+1}$ with inverses $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively, and consider a point $\boldsymbol{p} = \boldsymbol{h}(\boldsymbol{a}) \in M$ in the common image. Show that for any $\boldsymbol{u} \in \mathbb{R}^{n+1}_{\boldsymbol{p}}$,

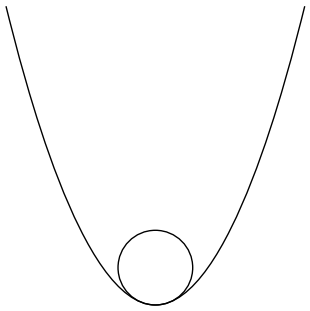
$$\begin{aligned}&\det\left(\frac{\partial}{\partial x^1}(\boldsymbol{p}), \ldots, \frac{\partial}{\partial x^n}(\boldsymbol{p}), \boldsymbol{u}\right)\\ &\quad = \det[D(\boldsymbol{k}^{-1} \circ \boldsymbol{h})](\boldsymbol{a}) \det\left(\frac{\partial}{\partial y^1}(\boldsymbol{p}), \ldots, \frac{\partial}{\partial y^n}(\boldsymbol{p}), \boldsymbol{u}\right).\end{aligned}$$

Conclude that M is orientable if and only if it admits an atlas with the property that $\det(\boldsymbol{k}^{-1} \circ \boldsymbol{h}) > 0$ for any two overlapping parametrizations $\boldsymbol{h}$, $\boldsymbol{k}$ in the atlas.

7.2. Show that an $n \times n$ invertible matrix is symmetric if and only if its inverse is symmetric.

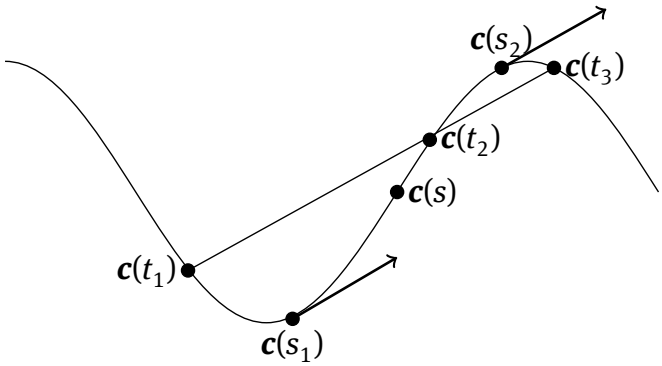
7.3. With the terminology from Example 7.4.1, determine the curvature of
(a) a straight line in $\mathbb{R}^2$;
(b) a circle of radius $r > 0$ in $\mathbb{R}^2$.

7.4. This long exercise is meant to clarify the remark made in Example 7.4.1, which asserts that the circle of curvature of a curve $\boldsymbol{c}$ at a point $\boldsymbol{c}(t_0)$ is the circle that best approximates the curve at that point. The underlying idea is that there is one, and only one, circle that passes through any three non-colinear points. Letting those points approach $\boldsymbol{c}(t_0)$ results in this limit circle.



The circle of curvature at the vertex of a parabola

So let $\boldsymbol{c} : [a, b] \to \mathbb{R}^2$ denote a regular curve parametrized by arc length, and $t_0 \in (a, b)$ be a point where $\boldsymbol{c}''(t_0) \neq \boldsymbol{0}$.
(a) Show that any three points on the curve that are sufficiently close to $\boldsymbol{c}(t_0)$ cannot be colinear; i.e., they cannot lie on a common line.
One outline of a proof: Suppose, to the contrary, that $\boldsymbol{c}(t_i)$ are colinear for $t_1 < t_2 < t_3$. The mean value theorem then implies that there exist $s_1 \in (t_1, t_2)$ and $s_2 \in (t_2, t_3)$ satisfying $\boldsymbol{c}'(s_1) = \boldsymbol{c}'(s_2)$.



Now, $\boldsymbol{c}'$ has its image in S^1 and cannot be onto S^1 if s_1 and s_2 are close enough; i.e., if the t_i are close enough to t_0, $1 \leq i \leq 3$. If $s \in (s_1, s_2)$ is such that $\boldsymbol{c}'(s)$ is furthest away from $\boldsymbol{c}'(s_i)$, show that $\boldsymbol{c}''(s) = \boldsymbol{0}$. This is impossible if the t_i are sufficiently close to t_0.

(b) Let t_i, $1 \le i \le 3$, be as in part (a) so that the $\boldsymbol{c}(t_i)$ are not colinear, and consider the center $\boldsymbol{p}_{123}$ of the unique circle that passes through these three points. Prove that there exist $s_1, s_2 \in (t_1, t_3)$ such that

$$\langle \boldsymbol{c}'(s_1), \boldsymbol{c}(s_1) - \boldsymbol{p}_{123} \rangle = 0, \tag{7.9.1}$$

$$\langle \boldsymbol{c}''(s_2), \boldsymbol{c}(s_2) - \boldsymbol{p}_{123} \rangle = -|\boldsymbol{c}'(s_2)|^2 = -1. \tag{7.9.2}$$

Hint: the function $t \mapsto |\boldsymbol{c}(t) - \boldsymbol{p}_{123}|^2$ has the same value at each t_i, so its derivative must vanish at some $r_1 \in (t_1, t_2)$ and $r_2 \in (t_2, t_3)$. By the same reasoning, its second derivative vanishes at some point in (r_1, r_2).

(c) Show that there exists a unique point $\boldsymbol{p} \in \mathbb{R}^2$ satisfying

$$\langle \boldsymbol{c}'(t_0), \boldsymbol{c}(t_0) - \boldsymbol{p} \rangle = 0,$$
$$\langle \boldsymbol{c}''(t_0), \boldsymbol{c}(t_0) - \boldsymbol{p} \rangle = -1.$$

Compare these equations with (7.9.1) to conclude that as $t_i \to t_0$, $1 \le i \le 3$, the unique circle passing through $\boldsymbol{c}(t_1)$, $\boldsymbol{c}(t_2)$, and $\boldsymbol{c}(t_3)$ approaches a circle passing through $\boldsymbol{c}(t_0)$. Furthermore, the latter circle has curvature equal to the curvature of $\boldsymbol{c}$ at t_0, and its center $\boldsymbol{p}$ lies on the line through $\boldsymbol{c}(t_0)$ perpendicular to the tangent line to $\boldsymbol{c}$ at $\boldsymbol{c}(t_0)$.

7.5. Let $\boldsymbol{c} : [a, b] \to \mathbb{R}^2$ be a regular curve parametrized by arc length. Prove that the curvature κ of $\boldsymbol{c}$ satisfies

$$\kappa = \det \begin{bmatrix} \boldsymbol{c}' & \boldsymbol{c}'' \end{bmatrix}.$$

Hint: The right side of the above equation equals, up to sign, the area of the parallelogram spanned by $\boldsymbol{c}'$ and $\boldsymbol{c}''$.

7.6. Even for simple curves, parametrization by arc length can be difficult to obtain (try using the formula from Exercise 7.5 to compute the curvature of a parabola). This problem generalizes that formula to regular curves that are not parametrized by arc length. If $\boldsymbol{c} : [a, b] \to \mathbb{R}^2$ is such a curve, let $s(t) = \int_a^t |\boldsymbol{c}'|$, so that $\tilde{\boldsymbol{c}} : \boldsymbol{c} \circ s^{-1}$ is the reparametrization of $\boldsymbol{c}$ by arc length. The curvature $\kappa(t)$ of $\boldsymbol{c}$ at $t \in (a, b)$ is defined to be the curvature of $\tilde{\boldsymbol{c}}$ at $s(t)$. Show that

$$\kappa = \frac{1}{|\boldsymbol{c}'|^3} \det \begin{bmatrix} \boldsymbol{c}' & \boldsymbol{c}'' \end{bmatrix}.$$

Compute the curvature of the parabola $y = x^2$.

7.7. The concepts discussed in the previous exercises generalize in part to curves in $\mathbb{R}^3$. Even though these are no longer hypersurfaces, they are important enough in classical differential geometry to warrant at least cursory mention.

So let $\boldsymbol{c} : [a, b] \to \mathbb{R}^3$ be a regular curve parametrized by arc length. Define $\boldsymbol{T} := \boldsymbol{c}'$. Notice that $\boldsymbol{T}$ is actually a map into S^2, even though we often identify it with the velocity vector field $\dot{\boldsymbol{c}}$ of the curve. The *curvature* of $\boldsymbol{c}$ is the function

$$\kappa = |\boldsymbol{T}'|.$$

Thus, unlike plane curves, the curvature is always nonnegative – essentially because we can no longer appeal to the orientation of $\mathbb{R}^2$. When $\kappa(t) \neq 0$, define the *principal normal* $\boldsymbol{N}$ at t by

$$\boldsymbol{N}(t) = \frac{1}{|\boldsymbol{T}'|(t)}\boldsymbol{T}'(t),$$

so that

$$\boldsymbol{T}' = \kappa \boldsymbol{N}.$$

Finally, the *binormal* is by definition

$$\boldsymbol{B}(t) = \boldsymbol{T}(t) \times \boldsymbol{N}(t),$$

so that when $\kappa(t) \neq 0$, the vectors $\boldsymbol{T}(t)$, $\boldsymbol{N}(t)$, and $\boldsymbol{B}(t)$ form a positively oriented orthonormal basis of $\mathbb{R}^3$.

(a) Show that $\boldsymbol{B}'$ is a multiple of $\boldsymbol{N}$. We may therefore define a function τ by the equation

$$\boldsymbol{B}' = -\tau \boldsymbol{N}.$$

τ is called the *torsion* of $\boldsymbol{c}$.

(b) Prove that $\boldsymbol{N}' = -\kappa \boldsymbol{T} + \tau \boldsymbol{B}$. The three identities

$$\begin{aligned} \boldsymbol{T}' &= \kappa \boldsymbol{N} \\ \boldsymbol{N}' &= -\kappa \boldsymbol{T} + \tau \boldsymbol{B} \\ \boldsymbol{B}' &= -\tau \boldsymbol{N} \end{aligned}$$

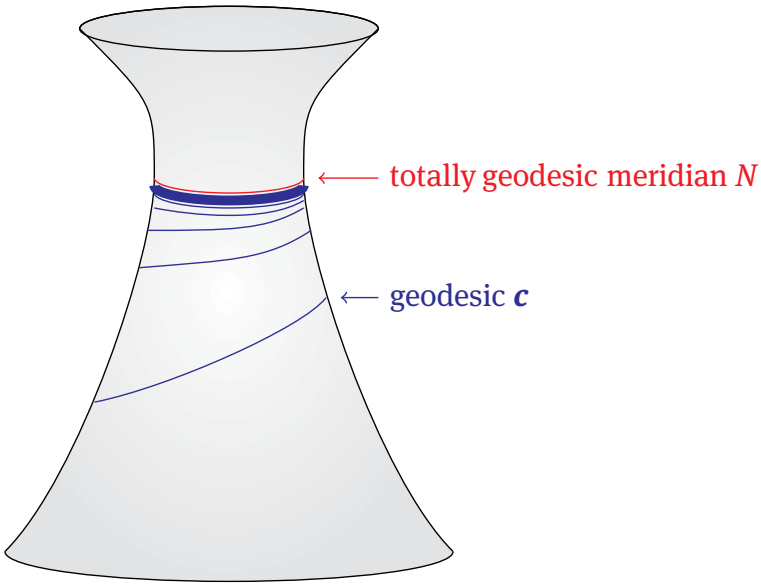
are called the *Serret-Frénet formulas*.

It can be shown that given continuous functions $\kappa, \tau : [a, b] \to \mathbb{R}$ with $\kappa > 0$, there exists one and only one (up to a rigid motion of $\mathbb{R}^3$) curve in 3-space parametrized by arc length that has κ as curvature and τ as torsion. This should be compared with Example 7.4.1 which asserts a similar property for plane curves involving only the curvature.

7.8. (a) Prove that any two surfaces in $\mathbb{R}^3$ with the same constant curvature κ are locally isometric. *Hint:* Use Theorem 6.7.2.

(b) Show that there exist (noncomplete) surfaces of revolution with constant curvature 1 that are not round spheres. These are not totally umbilic, however, and therefore do not contradict the fundamental theorem for submanifolds.

7.9. Consider a surface of revolution, and suppose that there is a point $\boldsymbol{p} = \boldsymbol{\gamma}(t_0)$ on the profile curve $\boldsymbol{\gamma}$ where the distance γ_1 to the z-axis has a strict local minimum. Thus, the parallel N through $\boldsymbol{p}$ is (the image of) a geodesic. Assume without loss of generality that $\boldsymbol{\gamma}$ is pointing upward, and choose $t_1 < t_0$ close enough to t_0 that $\gamma_{1|(t_1,t_0)}$ is strictly decreasing. Set $\boldsymbol{q} = \boldsymbol{\gamma}(t_1)$, and $\alpha \in (0, \pi/2)$ the angle with $\cos\alpha = \gamma_1(t_0)/\gamma_1(t_1)$. Show that the two geodesics emanating from $\boldsymbol{q}$ at angle α with the parallel increase in height for all time, and come arbitrarily close to N without ever reaching it.



7.10. As in the previous exercise, assume $\boldsymbol{\gamma}$ is pointing upward. Let $\boldsymbol{p} \in M$, $\alpha \in (0, \pi/2)$, and set $k := r(\boldsymbol{p}) \cos \alpha$. Denote by $\boldsymbol{c}$ one of the two geodesics from $\boldsymbol{p}$ that make an angle α with the parallel $\frac{\partial}{\partial x^2}$. The previous exercise investigated the behavior of $\boldsymbol{c}$ when k is a local minimum value of γ_1. This problem explores the case when k is not a critical value of γ_1; for instance, M could be the paraboloid discussed earlier, but inverted so that its vertex is the highest point. Let N denote the first parallel above $\boldsymbol{p}$ at distance k from the z-axis. Show that $\boldsymbol{c}$ hits N tangentially and then winds back down. The image of $\boldsymbol{c}$ is invariant under reflection in the plane containing the meridian through the point where $\boldsymbol{c}$ hits N, so that $\boldsymbol{c}$ passes again through $\boldsymbol{p}$ at an angle α with the parallel.

7.11. Let M^2 be a surface in $\mathbb{R}^3$, $\boldsymbol{p} \in M$, and U a neighborhood of the origin in $M_{\boldsymbol{p}}$ such that $\exp_{\boldsymbol{p}} : U \to \exp_{\boldsymbol{p}}(U) \subset M$ is a diffeomorphism. By means of a linear isometry $M_{\boldsymbol{p}} \cong \mathbb{R}^2$, introduce polar coordinates $(\tilde{r}, \tilde{\theta})$ on $M_{\boldsymbol{p}} \setminus L$, where L is some ray from the origin. If V denotes $\exp_{\boldsymbol{p}}(U \setminus L)$, then

$$\boldsymbol{x} = (r, \theta) := (\tilde{r}, \tilde{\theta}) \circ (\exp_{\boldsymbol{p}}^{-1})_{|V}$$

defines a chart on V. The object of this exercise is to derive a formula for the curvature K in terms of the "polar" chart $\boldsymbol{x}$.

For the sake of brevity, we denote the coordinate vector fields $\partial/\partial r$ and $\partial/\partial\theta$ by ∂_r and ∂_θ respectively.

(a) Prove that $\nabla_{\partial_r} \partial_r = 0$, $\langle \partial_r, \partial_r \rangle = 1$, and $\langle \partial_r, \partial_\theta \rangle = 0$.

(b) Let $G = \langle \partial_\theta, \partial_\theta \rangle$. Show that

$$G^2 K = \langle D_{\partial_r} \partial_r, \partial_r \times \partial_\theta \rangle \langle D_{\partial_\theta} \partial_\theta, \partial_r \times \partial_\theta \rangle - \langle D_{\partial_r} \partial_\theta, \partial_r \times \partial_\theta \rangle^2.$$

(c) Identify all vector fields in the identity above with their coordinate vector fields in the standard basis; i.e., identify ∂_r with the column matrix whose transpose is

$\big[\langle \partial_r, D_1\rangle \quad \langle \partial_r, D_2\rangle \quad \langle \partial_r, D_3\rangle\big]$, etc. Use properties of the cross product to show that

$$G^2K = \det\begin{bmatrix}(D_{\partial_r}\partial_r)^T\\ \partial_r^T\\ \partial_\theta^T\end{bmatrix}\cdot\det\begin{bmatrix}(D_{\partial_\theta}\partial_\theta)^T\\ \partial_r^T\\ \partial_\theta^T\end{bmatrix} - \det\left(\begin{bmatrix}(D_{\partial_r}\partial_\theta)^T\\ \partial_r^T\\ \partial_\theta^T\end{bmatrix}\right)^2,$$

and deduce that

$$G^2K = \det\begin{bmatrix}\langle D_{\partial_r}\partial_r, D_{\partial_\theta}\partial_\theta\rangle & \langle D_{\partial_r}\partial_r, \partial_r\rangle & \langle D_{\partial_r}\partial_r, \partial_\theta\rangle\\ \langle \partial_r, D_{\partial_\theta}\partial_\theta\rangle & 1 & 0\\ \langle \partial_\theta, D_{\partial_\theta}\partial_\theta\rangle & 0 & G\end{bmatrix}$$
$$- \det\begin{bmatrix}\langle D_{\partial_r}\partial_\theta, D_{\partial_r}\partial_\theta\rangle & \langle D_{\partial_r}\partial_\theta, \partial_r\rangle & \langle D_{\partial_r}\partial_\theta, \partial_\theta\rangle\\ \langle \partial_r, D_{\partial_r}\partial_\theta\rangle & 1 & 0\\ \langle \partial_\theta, D_{\partial_r}\partial_\theta\rangle & 0 & G\end{bmatrix}.$$

(d) Show that

$$G^2K = G\big(\langle D_{\partial_r}\partial_r, D_{\partial_\theta}\partial_\theta\rangle - \langle D_{\partial_r}\partial_\theta, D_{\partial_r}\partial_\theta\rangle\big) + \frac{1}{4}(D_{\partial_r}G)^2.$$

(e) Prove that

$$\frac{1}{2}D_{\partial_r}D_{\partial_r}G = \langle D_{\partial_r}D_{\partial_\theta}\partial_r, \partial_\theta\rangle + \langle D_{\partial_r}\partial_\theta, D_{\partial_r}\partial_\theta\rangle,$$

and

$$0 = D_{\partial_\theta}\langle D_{\partial_r}\partial_r, \partial_\theta\rangle = \langle D_{\partial_\theta}D_{\partial_r}\partial_r, \partial_\theta\rangle + \langle D_{\partial_r}\partial_r, D_{\partial_\theta}\partial_\theta\rangle.$$

Conclude that

$$\langle D_{\partial_r}\partial_r, D_{\partial_\theta}\partial_\theta\rangle - \langle D_{\partial_r}\partial_\theta, D_{\partial_r}\partial_\theta\rangle = -\frac{1}{2}D_{\partial_r}D_{\partial_r}G.$$

(f) Combine (d) and (e) to obtain

$$4KG^2 = -2GD_{\partial_r}D_{\partial_r}G + (D_{\partial_r}G)^2.$$

This last identity may be rewritten as

$$K = -\frac{D_{\partial_r}D_{\partial_r}\sqrt{G}}{\sqrt{G}}. \tag{7.9.3}$$

Observe that integral curves of ∂_r are the normal geodesics originating from $\boldsymbol{p}$. Let $\boldsymbol{v}$ be a unit vector in $M_{\boldsymbol{p}}$, $\boldsymbol{\gamma_v}$ the ray $t \mapsto t\boldsymbol{v}$ in $M_{\boldsymbol{p}}$, and $\boldsymbol{c_v} = \exp\circ\boldsymbol{\gamma_v}$ the geodesic with initial tangent vector $\boldsymbol{v}$. The restriction $\partial_{\tilde\theta}\circ\boldsymbol{\gamma_v}$ of the polar coordinate vector field $\partial_{\tilde\theta}$ on $M_{\boldsymbol{p}}$ is $t \mapsto t\mathcal{I}_{t\boldsymbol{v}}\boldsymbol{w}$, where $\boldsymbol{w}$ is a unit vector orthogonal to $\boldsymbol{v}$. Since the polar coordinate fields on $M_{\boldsymbol{p}}$ and on M are $\exp_{\boldsymbol{p}}$-related,

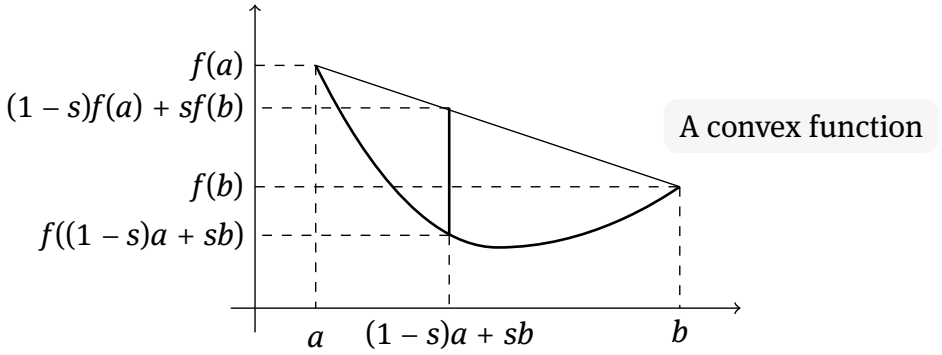
$$\partial_\theta\circ\boldsymbol{c_v}(t) = \exp_{\boldsymbol{p}*}\, t\mathcal{I}_{t\boldsymbol{v}}\boldsymbol{w},$$

cf. also (6.2.1). In other words, $\partial_\theta\circ\boldsymbol{c_v}$ is one of the two Jacobi fields J orthogonal to $\boldsymbol{c_v}$ satisfying $J(0) = \boldsymbol{0}$, $|J'(0)| = 1$. Thus, (7.9.3) says that in nonnegative (resp. nonpositive) curvature, this Jacobi field has concave (resp. convex) norm, see also Exercise 7.12.

7.12. Recall that a function $f : I \to \mathbb{R}$ defined on an interval I is said to be *convex* if its graph lies below any secant line; explicitely, for $a, b \in I$ with $a < b$,

$$f((1-s)a + sb) \leq (1-s)f(a) + sf(b), \qquad s \in [0,1].$$

When the inequality above is strict for all $s \in (0,1)$ (and all $a < b$), f is said to be *strictly convex*. f is said to be *concave* if $-f$ is convex.



(a) Show that a convex function is necessarily continuous. Prove that if f is C^2, then f is convex if and only if $f'' \geq 0$, and that if $f'' > 0$, then f is strictly convex.
(b) A function $f : \mathbb{R}^{n+1} \to \mathbb{R}$ is said to be convex or concave if the restriction $f \circ \boldsymbol{c}$ of f to any geodesic $\boldsymbol{c}$ of $\mathbb{R}^{n+1}$ has that property. Prove that if this is the case, then $M = f^{-1}(a)$ has nonnegative curvature for any regular value a of f.
(c) If $f : \mathbb{R}^{n+1} \to \mathbb{R}$ is convex, show that the set $f^{-1}(-\infty, a)$ is convex for any $a \in \mathbb{R}$.
(d) Prove that any convex, bounded function $f : \mathbb{R}^{n+1} \to \mathbb{R}$ is constant.

7.13. Prove that the graph of $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ has nonnegative curvature if f is either convex or concave (see Exercise 7.12).

7.14. Let $f : (a, b) \to \mathbb{R}$. Show that the cylinder over the graph of f (i.e., the hypersurface parametrized by $(s, t) \mapsto (s, f(s), t)$) is flat.

7.15. Show that if the hypersurface M is positively curved at $\boldsymbol{p}$, then it is strictly convex at that point: i.e., there is a neighborhood U of $\boldsymbol{p}$ in M such that $U \cap \{\boldsymbol{p} + \pi_2(M_{\boldsymbol{p}})\} = \{\boldsymbol{p}\}$.

7.16. Use Exercise 7.15 to show that any hypersurface with positive curvature is orientable. Is this still true for nonnegative curvature?

7.17. Suppose b is a semi-definite symmetric bilinear form on an inner product space $(V, \langle , \rangle)$. The eigenvalues of the associated self-adjoint operator L, where $\langle L\boldsymbol{v}, \boldsymbol{w}\rangle = b(\boldsymbol{v}, \boldsymbol{w})$, are then either all nonnegative or all nonpositive.
(a) Write $V = V_0 \oplus V_1$, where V_0 is the kernel of L, and V_1 its orthogonal complement, and let $\pi : V \to V_1$ denote the projection. Show that $b(\boldsymbol{v}, \boldsymbol{w}) = b(\pi\boldsymbol{v}, \pi\boldsymbol{w})$ for all $\boldsymbol{v}$ and $\boldsymbol{w}$ in V. Conclude that

$$b(\boldsymbol{v}, \boldsymbol{v}) \cdot b(\boldsymbol{w}, \boldsymbol{w}) \geq b^2(\boldsymbol{v}, \boldsymbol{w}), \qquad \boldsymbol{v}, \boldsymbol{w} \in V.$$

(b) Show that M has nonnegative curvature at a point if and only if the second fundamental form is semi-definite at that point.

7.18. A *quadric* in $\mathbb{R}^3$ is a set of the form $p^{-1}(0)$, where p is a polynomial of degree ≤ 2; i.e.,

$$p(x_1, x_2, x_3) = \sum_{i,j=1}^{3} a_{ij} x_i x_j + \sum_{i=1}^{3} b_i x_i + c.$$

(a) Show that there exists a linear isometry L of $\mathbb{R}^3$ and $A_i, B_i, C \in \mathbb{R}$ such that

$$L(M) = \{(x_1, x_2, x_3) \mid \sum_{i=1}^{3} A_i x_i^2 + B_i x_i + C = 0\}.$$

(b) Prove that there exists a translation T of $\mathbb{R}^3$ such that

$$TL(M) = \{(x_1, x_2, x_3) \mid \sum_{i=1}^{3} p_i(x_i) + \gamma = 0\},$$

where each p_i has the form $p_i(x_i) = \alpha_i x_i^2$ or $p_i(x_i) = \beta_i x_i$; i.e., each B_i in (a) may be assumed to be zero if the corresponding A_i is nonzero.

(c) Show that a nonempty quadric is isometric to

- an ellipsoid
$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1,$$
a hyperboloid of one sheet
$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1,$$
or a hyperboloid of two sheets
$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = -1,$$
if none of the A_i vanish;
- an elliptic paraboloid
$$z = \frac{x^2}{a^2} + \frac{y^2}{b^2},$$
or a hyperbolic paraboloid
$$z = \frac{x^2}{a^2} - \frac{y^2}{b^2},$$
if exactly one $A_i = 0$, but the corresponding $B_i \neq 0$;
- a line, or a cylinder over a line, parabola, ellipse, or hyperbola in the x-y plane otherwise.

7.19. Prove that a surface $M^2 \subset \mathbb{R}^3$ with negative curvature has no umbilical points. *Hint:* There is a one sentence proof.

7.20. Let L be a self-adjoint operator on $\mathbb{R}^3$, and suppose that

$$M = \{\boldsymbol{x} \in \mathbb{R}^3 \mid \langle L\boldsymbol{x}, \boldsymbol{x}\rangle = 1\}$$

is nonempty. Assume also that 1 is a regular value of $\boldsymbol{x} \mapsto \langle L\boldsymbol{x}, \boldsymbol{x}\rangle$, so that M is a manifold (and a quadric).

(a) Show that $\boldsymbol{x} \in M$ is umbilical if and only if

$$\det\begin{bmatrix} L\boldsymbol{x} & L\boldsymbol{u} & \boldsymbol{u}\end{bmatrix} = 0 \text{ for all } \boldsymbol{u} \perp L\boldsymbol{x}.$$

(b) Use (a) to prove that a 2-dimensional sphere (which corresponds to L being a multiple of the identity) is totally umbilic.

(c) Suppose L has 2 distinct eigenvalues $\lambda_1 > 0$ and λ_2, where the λ_2-eigenspace is 2-dimensional. Show that there exist at least two points $\pm\boldsymbol{x}$ of M that are umbilical. *Hint:* $\boldsymbol{x}$ will be an eigenvector.

7.21. Let M^2 be an oriented surface in $\mathbb{R}^3$, with unit normal $\boldsymbol{n}$ and corresponding second fundamental tensor S. A curve $\boldsymbol{c}$ in M is called a *line of curvature* if $\dot{\boldsymbol{c}}(t)$ is an eigenvector of S at $\boldsymbol{c}(t)$ for all t.

(a) Suppose P is a plane that intersects M orthogonally (in the sense that if $\boldsymbol{c}$ parametrizes $P \cap M$, then $\boldsymbol{n} \circ \boldsymbol{c}$ is tangent to the plane). Show that $\boldsymbol{c}$ is a line of curvature, and that the corresponding principal curvature is the curvature of $\boldsymbol{c}$. Notice that this again proves that a sphere is totally umbilic.

(b) Part (a) shows that the meridians in a surface of revolution are lines of curvature. Prove that parallels are also lines of curvature. *Hint:* the restriction of $\boldsymbol{n}$ to the parallel makes a constant angle with the axis of revolution.

(c) Let $f : \mathbb{R} \to \mathbb{R}$ be a smooth, even function with $f(0) = 0$, and M the surface of revolution obtained by revolving the curve $t \mapsto (t, 0, f(t))$, $t \geq 0$ about the z-axis. Show that the origin is umbilical.

7.22. Both Exercise 7.20 and 7.21 imply that if M is the ellipsoid of revolution

$$\frac{x^2 + y^2}{a^2} + \frac{z^2}{c^2} = 1,$$

then the points $\pm(0, 0, c)$ on the axis of revolution are umbilical. In order to find any other possible umbilics, it is convenient to work with the description we gave of a surface of revolution.

(a) Suppose the profile curve $\boldsymbol{\gamma} = (\gamma_1, 0, \gamma_2)$ can be expressed as the graph of a function $x = f(z)$ in the x-z plane; i.e., $\gamma_1 = f$, $\gamma_2(s) = s$. Show that the principal curvatures from (7.8.3) become

$$\lambda_1 = \frac{-f''}{(1 + (f')^2)^{3/2}}, \qquad \lambda_2 = \frac{1}{f(1 + (f')^2)^{1/2}},$$

whenever $f \neq 0$. Thus, umbilical points correspond to those values of s for which $(ff'' + (f')^2)(s) + 1 = 0$, and occur in (possibly degenerate) parallels.

(b) Prove that there are no other umbilics on the ellipsoid of revolution. This contrasts with the generic ellipsoid $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$ which admits exactly four umbilics if a, b, and c are all distinct.

7.23. Show that the paraboloid

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

has two umbilical points if $a \neq b$ and only one (the vertex) if $a = b$.

7.24. A *parabolic umbilic* on a surface M in $\mathbb{R}^3$ is a point where both principal curvatures are zero (in [4] such a point is said to be planar). In particular, the point is umbilical, and the sectional curvature and second fundamental form vanish there. Show that on the surface given by the graph of f, where $f(x, y) = xy^2 - x^3$, the origin is a parabolic umbilic. It is often called a *monkey saddle* because it has three downward slopes emanating from it, two for the legs and one for the tail.

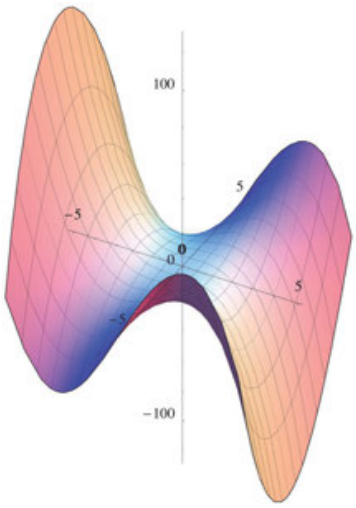


Fig. 7.8: A "monkey saddle"

7.25. Recall that $\boldsymbol{h}_1$, with $\boldsymbol{h}_1(s, t) = (t \cos s, t \sin s, s)$, parametrizes a helicoid, and that $\boldsymbol{h}_2$, where $\boldsymbol{h}_2(u, v) = (\cosh v \cos u, \cosh v \sin u, v)$, yields a catenoid.
Prove that both surfaces are locally isometric. *Hint:* reparametrize the catenoid by $\tilde{\boldsymbol{h}}_2$, where

$$\tilde{\boldsymbol{h}}_2(s, t) = \boldsymbol{h}_2(s, \operatorname{arccosh} \sqrt{1 + t^2}),$$

and consider $\tilde{\boldsymbol{h}}_2 \circ \boldsymbol{h}_1^{-1}$.

7.26. Let $M = \{(x, y, z) \in \mathbb{R}^3 \mid e^z \cos x = \cos y\}$.
(a) Show that M is an orientable 2-dimensional manifold.
(b) Prove that M may be described as follows: given $k, l \in \mathbb{Z}$, consider the open squares

$$S^1_{k,l} = (-\frac{\pi}{2} + 2k\pi, \frac{\pi}{2} + 2k\pi) \times (-\frac{\pi}{2} + 2l\pi, \frac{\pi}{2} + 2l\pi),$$

and

$$S^2_{k,l} = \{(x + \pi, y + \pi) \mid (x, y) \in S^1_{k,l}\}.$$

Then M consists of the graph of f, where $f(x, y) = \ln(\cos y / \cos x)$, over each of these squares, together with the vertical lines that pass through the vertices of the squares. M is called *Scherk's surface*. It is named after the 19th century German mathematician who discovered it.

(c) Show that M is a minimal surface.

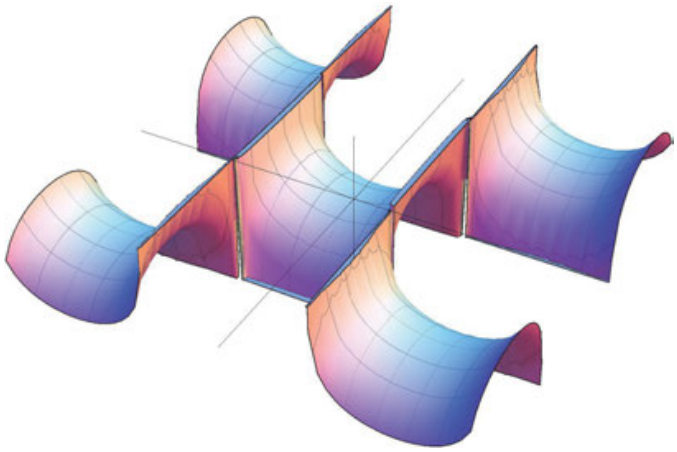


Fig. 7.9: Scherk's surface

7.27. The image of $\boldsymbol{h}$, where

$$\boldsymbol{h}(u, v) = \frac{1}{2}(u(1 - \frac{u^2}{3} + v^2), v(1 + \frac{v^2}{3} - u^2), u^2 - v^2)),$$

is called *Enneper's surface*. $\boldsymbol{h}$ is not strictly speaking a parametrization (and its image M is not a manifold) because it is not one-to-one. Nevertheless, M is locally a manifold.

(a) Show that M is a minimal surface.

(b) Prove that M has strictly negative curvature

$$(K \circ \boldsymbol{h})(u, v) = -\frac{4}{9^3}(1 + u^2 + v^2)^4.$$

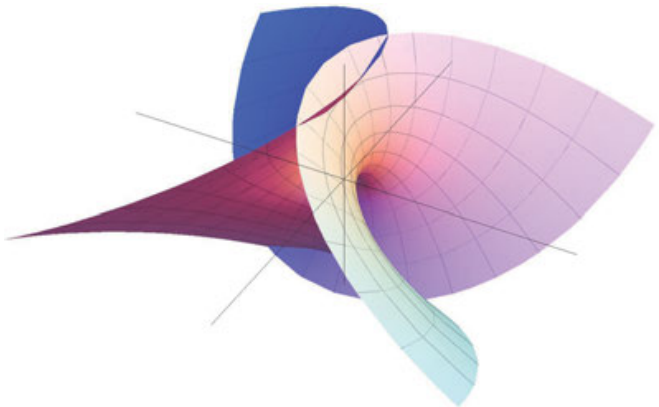


Fig. 7.10: Enneper's surface

Appendix A

In this appendix, we review some basic properties of real numbers that are used throughout the text but are usually not discussed in detail in Calculus or other lower level math courses. This is by no means a comprehensive overview; rather, it emphasizes those features – such as the rationals being countable and dense – that provide interesting examples and counterexamples, particularly in the theory of limits and integration.

Recall that there is a nested sequence $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}$, with $\mathbb{N} = \{1, 2, \dots\}$ denoting the *natural numbers*, $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ the *integers*, and $\mathbb{Q} = \{p/q \mid p \in \mathbb{Z}, q \in \mathbb{N}\}$ the *rationals*. A real number that is not rational is said to be *irrational*.

Definition. $\alpha \in \mathbb{R}$ is said to be an *upper bound* (resp. a *lower bound*) of $A \subset \mathbb{R}$ if for any $a \in A$, $\alpha \geq a$ (resp. $\leq a$). An upper bound is said to be the *least upper bound* or *supremum* of A if it is less than or equal to any other upper bound of A. It is denoted $\sup A$. Similarly, a lower bound of A is called the *greatest lower bound* or *infimum* if it greater or equal to any other lower bound. The infimum of A, if it exists, is denoted $\inf A$.

It is clear from the definition that infimum and supremum, if they exist, are unique. Notice that if we set $-A = \{-a \mid a \in A\}$, then A is bounded above by α if and only if $-A$ is bounded below by $-\alpha$. Furthermore, α is less than or equal to any other upper bound of A if and only if $-\alpha$ is greater than or equal to any other lower bound of $-A$; i.e., $-\sup A = \inf(-A)$, provided one of them exists. The supremum of a set may, or may not, belong to it; for example, the supremum of the interval $(0, 1)$ is 1, which does not lie in the open interval. Indeed, 1 is an upper bound, and by definition, if α is any upper bound of $(0, 1)$, then $\alpha \geq 1$; i.e., $1 = \sup(0, 1)$. When the supremum does belong to the set, it is called the *maximum*. Similar considerations apply to the infimum, which is called the *minimum* when it belongs to the set.

A useful criterion for determining the least upper bound of a set is the following: $\alpha \in \mathbb{R}$ is the least upper bound of A if and only if

(1) α is an upper bound of A, and

(2) for any $\varepsilon > 0$, there exists some $a \in A$ such that $\alpha - \varepsilon < a$.

To see this, suppose first that $\alpha = \sup A$. Condition (1) is then verified by definition. If (2) were not, then there would exist some $\varepsilon > 0$ such that $\alpha - \varepsilon$ is greater than or equal to any element of A. But this would mean that $\alpha - \varepsilon$ is an upper bound of A which is smaller that the least upper bound. Thus, (2) is also verified. Conversely, suppose α satisfies both conditions (1) and (2). To establish that $\alpha = \sup A$, we only need to show that if β is an upper bound of A, then $\alpha \leq \beta$. So suppose that β is an upper bound of A but $\beta < \alpha$. Take $\varepsilon = \alpha - \beta > 0$ in (2) to deduce that there exists some $a \in A$ with $\alpha - \varepsilon = \beta < a$. This contradicts the fact that β is an upper bound of A.

A similar characterization with inequalities reversed applies to the infimum: a lower bound α of A is the greatest lower bound if and only if for any $\varepsilon > 0$, there is some $a \in A$ satisfying $a < \alpha + \varepsilon$.

A fundamental property of real numbers is the following:

Completeness Axiom. *Every nonempty subset of* $\mathbb{R}$ *that is bounded above has a least upper bound.*

Notice that by the above discussion, the completeness axiom can also be phrased as: every nonempty subset of real numbers that is bounded below has a greatest lower bound.

Let us examine some consequences of the completeness axiom:

Theorem (Well-ordering Principle). *Every nonempty subset of the natural numbers has a smallest element (i.e., a minimum).*

Proof. Let A be a nonempty set of natural numbers. Being bounded below (by 0), A has an infimum α. We claim $\alpha \in \mathbb{N}$: suppose α is not a natural number. Using the fact that α is the greatest lower bound of A, there exists $a \in A$ such that $\alpha < a < \alpha + 1/2$. Since $\alpha \neq a$, the number $\varepsilon = a - \alpha$ is positive, and there exists $\tilde{a} \in A$ such that $\alpha < \tilde{a} < \alpha + \varepsilon = a < \alpha + 1/2$. This is impossible, because both a and $\tilde{a}$ are natural numbers, so their distance cannot be less than 1/2. Thus, $\alpha \in \mathbb{N}$, and must therefore also belong to A: otherwise there would exist a natural number in A at distance less than 1/2 from α, which again is impossible. □

The next result is an extremely useful tool for proving properties involving natural numbers:

Theorem (Mathematical Induction). *Suppose* $P(n)$ *is a statement for each* $n \in \mathbb{N}$ *such that*
(1) $P(1)$ *is true, and*
(2) *for any* $k \in \mathbb{N}$, $P(k+1)$ *is true whenever* $P(k)$ *is.*
Then $P(n)$ *is true for all* $n \in \mathbb{N}$.

Proof. We argue by contradiction: suppose $P(n)$ is not true for all n, so that the set A of all $n \in \mathbb{N}$ for which $P(n)$ is not true is nonempty. By the well-ordering principle, A has a smallest element $k \in A \subset \mathbb{N}$. By (1), $k > 1$, so that $k - 1 \in \mathbb{N}$, and $P(k-1)$ is true. This contradicts (2), since $P(k)$ is false. □

Theorem (Archimedean Principle). *If* a *and* b *are real numbers with* $a > 0$, *there exists* $n \in \mathbb{N}$ *such that* $na > b$.

Proof. Let us first rephrase the statement: since $a > 0$, $b/a \in \mathbb{R}$, and we must show that there exists some $n \in \mathbb{N}$ that is larger than b/a. In other words, the Archimedean principle just asserts that the set of natural numbers is not bounded above. Once again, the argument will be by contradiction. Suppose $\mathbb{N}$ is bounded above. By the completeness

axiom, $\mathbb{N}$ has a least upper bound α. Since α is the smallest upper bound, there must exist some natural number n such that $\alpha - 1 < n \leq \alpha$. The first inequality says that α is less than the natural number $n + 1$, which is a contradiction. □

Our next endeavor is to compare the sizes of different sets of numbers. For finite sets, this is straightforward: sets A and B are said to have the same *cardinality* if there exists a bijection (i.e., a one-to-one and onto) map $f : A \to B$. It is therefore natural to extend this definition to infinite sets, and we do so. Strange things happen in the process, however. For example, intuitively, $\mathbb{Z}$ should have at least twice as many elements as $\mathbb{N}$, since $\mathbb{Z} = \mathbb{N} \cup (-\mathbb{N}) \cup \{0\}$. Nevertheless both sets have the same cardinality: it is easy to check that the map $f : \mathbb{Z} \to \mathbb{N}$, where $f(n) = 2n + 1$ if $n \geq 0$, and $f(n) = -2n$ if $n < 0$, is a bijection.

Definition. A set A is said to be *countable* if there exists a one-to-one map $f : A \to \mathbb{N}$.

Thus, any finite set is countable, but so is the set of all integers. A useful alternative characterization of countability is the following: A is countable if and only if there exists a surjective $f : \mathbb{N} \to A$. Indeed, if $f : \mathbb{N} \to A$ is surjective, we may construct a one-to-one map $g : A \to \mathbb{N}$ by defining $g(a)$ to be any one element in $f^{-1}(a) = \{n \in \mathbb{N} \mid f(n) = a\}$. Conversely, if A is countable, and $g : A \to \mathbb{N}$ is one-to-one, we obtain a surjective map $f : \mathbb{N} \to A$ by setting $f(n) = a$ if a is the (necessarily unique) element of A such that $g(a) = n$, and $f(n) = a_0$ for some fixed $a_0 \in A$ if there is no $a \in A$ with $g(a) = n$.

Since by definition, a map $f : \mathbb{N} \to A$ is a sequence, we may write any countable set A as $A = \{a_1, a_2, \dots\}$, where $a_n = f(n)$. More generally, $\mathbb{N}$ may be replaced in the above discussion by any countably infinite set (i.e., any set with the same cardinality as $\mathbb{N}$).

Proposition. *A countable union of countable sets is countable.*

Proof. If E_i, $i \in \mathbb{N}$, is countable, we may write $E_i = \{x_{i1}, x_{i2}, \dots\}$. Define $f : \cup_{i=1}^{\infty} E_i \to \mathbb{N}$ by $f(x_{ij}) = 2^i 3^j$. f is one-to-one, for if $f(x_{ij}) = f(x_{kl})$, then $2^i 3^j = 2^k 3^l$, and $2^{i-k} = 3^{l-j}$. This can only hold if $i = k$ and $j = l$. □

It follows that the Cartesian product $A \times B$ of countable sets A and B is again countable: if $A = \{a_1, a_2, \dots\}$, then $A \times B = \cup_{i=1}^{\infty} \{a_i\} \times B$ is a countable union of countable sets. In particular,

Theorem. *The set $\mathbb{Q}$ of rational numbers is countable.*

Proof. The map $f : \mathbb{Z} \times \mathbb{N} \to \mathbb{Q}$, $f(m, n) = m/n$, is onto. □

A set which is not countable is said to be *uncountable*. There are many such:

Theorem. *The interval $(0, 1)$ is uncountable. In particular, the set of irrational numbers between 0 and 1 is uncountable.*

Proof. Any number $a \in (0,1)$ admits a decimal expansion

$$a = .a_1a_2a_3\cdots = \sum_{i=1}^{\infty} \frac{a_i}{10^i}, \qquad a_i \in \{0,1,2,\dots,9\}.$$

In fact, for any $b > 0$, let $[b]$ denote the unique $n \in \mathbb{N}\cup\{0\}$ such that $n \le b < n+1$; i.e., $[b]$ is the largest integer smaller than or equal to b (specifically, $[b] = \sup\{n \in \mathbb{N} \mid n \le b\}$ – notice that the least upper bound of a set of integers, if it exists, is an integer, so the supremum is in fact a maximum). It is not hard to check that in the formula above, a_n may be taken to be $\left[10(10^{n-1}a - [10^{n-1}a])\right]$. The reason we write "may be taken" is that this decimal expansion need not be unique; e.g., $0.5 = 0.499999\dots$. This, however, is the only way in which the expansion can fail to be unique: more precisely, suppose $a = \sum(a_k/10^k) = \sum(b_k/10^k) = b$, and let i be the smallest integer such that $a_i \neq b_i$; assuming without loss of generality that $a_i > b_i$, we must have $a_i = b_i + 1$, $a_{k+i} = 0$ and $b_{k+i} = 9$ for all $k \ge 1$. To see this, observe that

$$0 = 10^i \sum_{k=1}^{\infty} \frac{a_k - b_k}{10^k} = \sum_{k=i}^{\infty} \frac{a_k - b_k}{10^{k-i}} = a_i - b_i + \sum_{k=1}^{\infty} \frac{a_{k+i} - b_{k+i}}{10^k}. \tag{A.1}$$

Now, $a_i - b_i$ is an integer between 1 and 9, whereas the last series in the above expression satisfies in absolute value

$$\left|\sum_{k=1}^{\infty} \frac{a_{k+i} - b_{k+i}}{10^k}\right| \le \sum_{k=1}^{\infty} \frac{|a_{k+i} - b_{k+i}|}{10^k} \le \sum_{k=1}^{\infty} \frac{9}{10^k} = 1.$$

Thus, if the expression in (A.1) is to equal zero, then $a_i - b_i$ must equal 1 and $a_{k+i} - b_{k+i} = -9$, as claimed.

Now, suppose that $(0,1)$ is countable, so that $(0,1) = \{a_1, a_2, \dots\}$, with $a_i = .a_{i1}a_{i2}\dots$. The contradiction will arise once we exhibit some $b \in (0,1)$ that is not equal to any a_i. So set $b = .b_1b_2\dots$, where $b_i = 1$ if $a_{ii} = 2$ and $b_i = 2$ otherwise. Then b differs in its i-th digit from that of a_i. Since the only numbers that appear in the decimal expansion of b are 1 and 2, this expansion is unique, so that b is not equal to any a_i. This shows that $(0,1)$ is uncountable.

The assertion that the set A of irrationals between 0 and 1 is uncountable is now clear: if it were countable, then $(0,1)$ would be the union of two countable sets, namely A and the set of rationals between 0 and 1, and would then also be countable. □

Theorem. *For any real numbers $a < b$, the interval (a,b) contains infinitely many rational and infinitely many irrational numbers.*

Proof. We first show that the interval contains at least one rational and one irrational, beginning with rational: it may be assumed that $a > 0$, for $k + a > 0$ if k is a sufficiently large natural number, and if r is a rational between $k + a$ and $k + b$, then $r - k$ is one between a and b. We must find natural numbers m, n such that $a < m/n < b$, or equivalently, such that $na < m < nb$. This in turn requires that $nb - na > 1$. So choose

any $n \in \mathbb{N}$ with $n > 1/(b-a), 1/a$ (which is possible by the Archimedean principle). If $na \in \mathbb{N}$, then $r = (na+1)/n$ is a rational between a and b. If $na \notin \mathbb{N}$, consider the set E of all natural numbers no larger than na. Since $na > 1$, E is nonempty, and being bounded above admits a supremum k_0. Furthermore, $k_0 \in \mathbb{N}$, so it is strictly smaller than na, and

$$na < k_0 + 1 < na + 1 < nb.$$

We may then take r to equal $(k_0+1)/n$.
To exhibit an irrational, let $\alpha \in (a,b) \cap \mathbb{Q}$ (the existence of such an α has just been established), and use the Archimedean Principle to find $N \in \mathbb{N}$ such that $\sqrt{2}/N < b-\alpha$. Then $\alpha + \sqrt{2}/N$ is an irrational in (a,b).
To see that there are infinitely many rationals, let $\alpha_1 \in (a,b) \cap \mathbb{Q}$, and construct inductively an infinite sequence of distinct rationals by choosing some $\alpha_{n+1} \in (a, \alpha_n) \cap \mathbb{Q}$. A similar argument shows that there are infinitely many irrationals. □

The Theorem says that the closure of $\mathbb{Q}$ is the whole real line. This is often expressed by saying that $\mathbb{Q}$ is *dense* in $\mathbb{R}$.

Examples and Remarks. (i) Given any positive number a, there exists an increasing sequence $\{b_n\}$ of rationals with $\lim_{n\to\infty} b_n = a$. In fact, writing a as an infinite decimal $a = a_1.a_2\ldots$ with $a_1 \in \mathbb{N} \cup \{0\}$, and $a_i \in \{0,1,\ldots,9\}$ for $i > 1$, define $b_n = a_1.a_2\ldots a_n$. $\{b_n\}$ is clearly an increasing sequence, and it converges to a because $a - b_n < 9/10^{n+1}$. This shows, in particular, that the completeness axiom does not hold for the rational numbers: if a as above is irrational, then the set $\{x \in \mathbb{Q} \mid x < a\}$ is bounded above, but its supremum a does not lie in $\mathbb{Q}$.

(ii) Let $f : (0,1) \to \mathbb{R}$ be given by $f(x) = 0$ if x is irrational, $f(x) = 1/q$ if $x = p/q$ is rational, and p, q have no common factors. Then f is continuous at every irrational point and discontinuous elsewhere; i.e., f has a countable number of discontinuities. In fact, if x_0 is rational, there exists a sequence $\{x_n\}$ of irrationals converging to x_0 (for example, let x_n be an irrational number in $(x_0 - 1/n, x_0) \cap (0, x_0)$), so that $f(x_n) = 0$ does not converge to $f(x_0)$. Next, suppose x_0 is irrational. It is enough to show that if $x_n \to x_0$, $x_n \in \mathbb{Q} \cap (0,1)$, then $f(x_n) \to 0$; i.e., if $x_n = p_n/q_n$, then $q_n \to \infty$. Suppose not. Then there exists $N \in \mathbb{N}$ such that for any $M \in \mathbb{N}$, $q_n \le N$ for some $n \ge M$. In other words, there exists a subsequence $\{q_{n_k}\}$ whose terms are all less than or equal to N. This means that the subsequence $\{p_{n_k}/q_{n_k}\}$ can only take on the values $1/N, 2/N, \ldots, (N-1)/N$. Since x_0 does not equal any of these values, this subsequence does not converge to x_0, and hence neither does the original one.

Appendix B

The basic aim of this section is to show that any invertible linear transformation is a finite composition of certain elementary ones. It has been relegated to an appendix because it is only used in the proof of the change of variables theorem for integrals. Nevertheless, it has several useful applications to rank and systems of linear equations.

Definition B.1. A linear map $L : \mathbb{R}^n \to \mathbb{R}^n$ is said to be an *elementary transformation* if it belongs to one of the following groups:

- type 1: there exist indices $1 \leq i < j \leq n$ such that $L\boldsymbol{e}_i = \boldsymbol{e}_j$, $L\boldsymbol{e}_j = \boldsymbol{e}_i$, and $L\boldsymbol{e}_k = \boldsymbol{e}_k$ for $k \neq i, j$;
- type 2: there exists $1 \leq i \leq n$, $a \neq 0$, such that $L\boldsymbol{e}_i = a\boldsymbol{e}_i$, and $L\boldsymbol{e}_k = \boldsymbol{e}_k$ for $k \neq i$;
- type 3: there exist $1 \leq i, j \leq n$, $i \neq j$, $a \in \mathbb{R}$, such that $L\boldsymbol{e}_i = \boldsymbol{e}_i + a\boldsymbol{e}_j$, and $L\boldsymbol{e}_k = \boldsymbol{e}_k$ for $k \neq i$.

The matrix of an elementary transformation with respect to the standard basis is called an *elementary matrix*.

It is clear from the definition that an elementary transformation is an isomorphism, and that its inverse is an elementary transformation of the same type. It follows that an elementary matrix is invertible, and its inverse is an elementary matrix of the same type. Properties of elementary transformations yield corresponding properties for elementary transformations, and vice-versa. The reader may wish to keep this in mind since we will usually only outline properties of one rather than of both.

Thus, an elementary matrix is of

- type 1 if it is obtained by interchanging two columns of the $n \times n$ identity matrix I_n,
- type 2 if it is obtained by multiplying a column of I_n by a nonzero scalar, and
- type 3 if it is obtained by adding a multiple of one column of I_n to another column.

These three operations are called *elementary column operations*. Replacing columns by rows, one obtains three types of *elementary row operations*. It is straightforward to check that an elementary matrix of a given type is also obtained by performing an elementary row operation of that type on I_n.

Given an $m \times n$ matrix A, we will denote by A_i its i-th row, and by A^j its j-th column. Notice that if $\boldsymbol{e}_i$ is the i-th coordinate vector in the standard basis written as a column, then $A_i = \boldsymbol{e}_i^T A$, and $A^j = A\boldsymbol{e}_j$.

One can also perform elementary row or column operations on any matrix. This is actually equivalent to multiplying that matrix by an elementary one:

Proposition B.1. *Let B denote the matrix obtained by performing an elementary row (resp. column) operation on $A \in M_{m,n}$, and E the elementary matrix obtained by performing that same operation on I_m (resp. I_n). Then $B = EA$ (resp. AE).*

Proof. One has to check the claim for each type of operation. We illustrate this for a row operation of type 2, and leave the others as an exercise: let E denote the elementary matrix obtained by interchanging rows i and j of I_n. Then the i-th row of EA is

$$(EA)_i = E_i A = \boldsymbol{e}_j^T A = A_j = B_i,$$

as claimed, and by symmetry, $(EA)_j = B_j$. If, on the other hand, $k \neq i, j$, then

$$(EA)_k = E_k A = \boldsymbol{e}_k^T A = A_k = B_k,$$

so that $EA = B$, as claimed. □

Proposition B.2. *Multiplying a matrix by an elementary one does not affect its rank.*

Proof. Recall that the rank of A is the rank of the linear transformation L_A. If E is an elementary matrix, then

$$\begin{aligned}\operatorname{rank}(EA) &= \operatorname{rank}(L_E \circ L_A) = \dim L_E(L_A(\mathbb{R}^n)) = \dim L_A(\mathbb{R}^n) = \operatorname{rank}(L_A) \\ &= \operatorname{rank}(A)\end{aligned}$$

since L_E is an isomorphism. Similarly, the nullity of $L_{AE} = L_A \circ L_E$ equals the nullity of L_A, so that by Theorem 1.2.2, AE and A have the same rank. □

Next, we show that any matrix A can be transformed by means of finitely many elementary row and column operations into a diagonal matrix (i.e. $a_{ij} = 0$ if $i \neq j$) with diagonal entries 1 or zero. By the Proposition, the number of 1's equals the rank of the matrix. The proof is best illustrated by means of an example:

Example B.1. Let

$$A = \begin{bmatrix} 1 & 2 & 3 & 1 \\ 2 & 1 & 1 & 1 \\ 1 & -1 & 1 & 0 \end{bmatrix}.$$

Subtract 2 times row 1 from row 2, and subtract row 1 from 3 to obtain

$$\begin{bmatrix} 1 & 2 & 3 & 1 \\ 0 & -3 & -5 & -1 \\ 0 & -3 & -2 & -1 \end{bmatrix}.$$

Subtract row 2 from row 3 to get

$$\begin{bmatrix} 1 & 2 & 3 & 1 \\ 0 & -3 & -5 & -1 \\ 0 & 0 & 3 & 0 \end{bmatrix}.$$

Next, divide row 2 by -3 and row 3 by 3:

$$\begin{bmatrix} 1 & 2 & 3 & 1 \\ 0 & 1 & \frac{5}{3} & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Subtract 2 times the first column from the second, 3 times the first column from the third, and one time the first column from the fourth:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \frac{5}{3} & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

The final step consists in subtracting 5/3 times the second column from the third and 1/3 times the second column from the fourth. The result is

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

Theorem B.1. *Let A be an $m \times n$ matrix of rank r. Then a finite number of elementary row and column operations transforms A into the $m \times n$ matrix B, where $b_{ij} = 0$ if $i \neq j$, $b_{ii} = 1$ if $i \leq r$, and $b_{ii} = 0$ if $i > r$.*

Proof. If A has rank zero, then it is the zero matrix, and it is already in the desired form. So assume $A \neq 0$. The proof will be by induction on the number m of rows of A. Suppose $m = 1$. By assumption, there exists a nonzero element in A. By interchanging columns if necessary, we may assume it appears in the first column, so that $A = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}$, with $a_1 \neq 0$. Multiplying the first column by $1/a_1$, and subtracting a_i times the first column from the i-th one for $i = 2, \dots, n$ establishes the claim. Next, assume the theorem holds for any matrix with k rows, and let A be a $(k+1) \times n$ matrix. Since $A \neq 0$, some $a_{ij} \neq 0$. Interchange rows 1 and i, and columns 1 and j to move a_{ij} into the first row and first column. Divide the first row (or column) by a_{ij} to obtain a matrix with 1 in the first row and column. Now subtract a_{1j} times the first column from the j-th one for $j = 2, \dots, n$, and subtract a_{i1} times the first row from the i-th one for $i = 2, \dots, k+1$. The net result is now a matrix of the form

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & B & \\ 0 & & & \end{bmatrix},$$

where B is $k \times (n-1)$. Finally, applying the induction hypothesis to B, one can transform it by means of finitely many elementary operations into a diagonal matrix of the specified form. □

Corollary B.1. *Every isomorphism $L : \mathbb{R}^n \to \mathbb{R}^n$ is a finite composition of elementary transformations.*

Proof. Let A denote the matrix of L with respect to the standard basis, so that $L = L_A$ (see Example 1.2.3). Then A is invertible and can be transformed into the identity

matrix I_n by means of a finite sequence of elementary row and column operations. Thus, by Proposition B.1, there exist elementary matrices $E_1, \ldots, E_k$ and $E'_1, \ldots, E'_l$ such that

$$E_1 E_2 \cdots E_k A E'_1 E'_2 \cdots E'_l = I_n.$$

Multiply both sides of the above identity on the left by $E_k^{-1} \cdots E_1^{-1}$ and on the right by $E_l'^{-1} \cdots E_1'^{-1}$ to obtain

$$A = E_k^{-1} \cdots E_1^{-1} E_l'^{-1} \cdots E_1'^{-1}.$$

Since the inverse of an elementary matrix is elementary, A equals a product of elementary matrices. But then

$$L_A = L_{E_k^{-1}} \circ \cdots \circ L_{E_1'^{-1}}$$

equals a composition of elementary transformations, as claimed. □

Bibliography

[1] N. Aders, "Amsler-Polarplanimeter-2", Own Work. Licensed under Creative Commons Attribution-Share Alike 3.0 via Wikimedia Commons.

[2] T. Apostol, *Mathematical Analysis*, second edition, Addison-Wesley (1974).

[3] G. Ascoli, "Vedute sintetiche sugli strumenti integratori", *Rend. Sem. Mat. Fis. Milano* **18** (1947).

[4] L. Auslander, *Differential Geometry*, Harper & Row (1967).

[5] M. Berger, B. Gostiaux, *Géométrie Différentielle*, Armand Colin (1972).

[6] D. Burago, Y. Burago, S. Ivanov, *A course in Metric Geometry*, Graduate Studies in Mathematics **33**, American Mathematical Society (2001).

[7] M. do Carmo, *Differential geometry of curves and surfaces* (translated from the Portuguese), Prentice-Hall (1976).

[8] M. do Carmo, *Differential forms and applications* (translated from the 1971 Portuguese original), Springer-Verlag (1994).

[9] M. Hirsch and S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press (1974).

[10] O. Knill, "The planimeter and the theorem of Green", URL: http://www.math.harvard.edu/~knill/teaching/math21a2000/planimeter

[11] J. Milnor, *Topology from the Differentiable Viewpoint*, University Press of Virginia (1965).

[12] M. Spivak, *Calculus on Manifolds*, Addison-Wesley (1965).

[13] M. Spivak, *A Comprehensive Introduction to Differential Geometry*, Volumes I–V, Publish or Perish (1975–1979).

[14] S. Sternberg, *Lectures on Differential Geometry*, Prentice-Hall (1964).

[15] D. Struik, *Differential Geometry*, Addison-Wesley (1950).

[16] J. Thorpe, *Elementary Topics in Differential Geometry*, Springer-Verlag (1979).

Index

This book offers an introduction to differential geometry for the non-specialist. It includes most of the required material from multivariable calculus, linear algebra, and basic analysis. An intuitive approach and a minimum of prerequisites make it a valuable companion for students of mathematics and physics.
The main focus is on manifolds in Euclidean space and the metric properties they inherit from it. Among the topics discussed are curvature and how it affects the shape of space, and the generalization of the fundamental theorem of calculus known as Stokes' theorem.

- A thorough introduction to differential geometry
- Covers all important concepts and theorems
- Many examples including applications to physics

Gerard Walschap
is a professor of mathematics at the University of Oklahoma, and specializes in differential geometry.

www.degruyter.com
ISBN 978-3-11-036949-6