Notes on Differential Geometry and Lie Groups

Jean Gallier
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
e-mail: jean@cis.upenn.edu

December 7, 2012

To my daughter Mia, my wife Anne,

my son Philippe, and my daughter Sylvie.

Preface

The motivations for writing these notes arose while I was coteaching a seminar on Special Topics in Machine Perception with Kostas Daniilidis in the Spring of 2004. In the Spring of 2005, I gave a version of my course *Advanced Geometric Methods in Computer Science* (CIS610), with the main goal of discussing statistics on diffusion tensors and shape statistics in medical imaging. This is when I realized that it was necessary to cover some material on Riemannian geometry but I ran out of time after presenting Lie groups and never got around to doing it! Then, in the Fall of 2006 I went on a wonderful and very productive sabbatical year in Nicholas Ayache's group (ACSEPIOS) at INRIA Sophia Antipolis where I learned about the beautiful and exciting work of Vincent Arsigny, Olivier Clatz, Hervé Delingette, Pierre Fillard, Grégoire Malandin, Xavier Pennec, Maxime Sermesant, and, of course, Nicholas Ayache, on statistics on manifolds and Lie groups applied to medical imaging. This inspired me to write chapters on differential geometry and, after a few additions made during Fall 2007 and Spring 2008, notably on left-invariant metrics on Lie groups, my little set of notes from 2004 had grown into the manuscript found here.

Let me go back to the seminar on Special Topics in Machine Perception given in 2004. The main theme of the seminar was group-theoretical methods in visual perception. In particular, Kostas decided to present some exciting results from Christopher Geyer's Ph.D. thesis [63] on scene reconstruction using two parabolic catadioptric cameras (Chapters 4 and 5). Catadioptric cameras are devices which use both mirrors (catioptric elements) and lenses (dioptric elements) to form images. Catadioptric cameras have been used in computer vision and robotics to obtain a wide field of view, often greater than 180°, unobtainable from perspective cameras. Applications of such devices include navigation, surveillance and vizualization, among others. Technically, certain matrices called *catadioptric fundamental matrices* come up. Geyer was able to give several equivalent characterizations of these matrices (see Chapter 5, Theorem 5.2). To my surprise, the Lorentz group $\mathbf{O}(3,1)$ (of the theory of special relativity) comes up naturally! The set of fundamental matrices turns out to form a manifold, $\mathcal{F}$, and the question then arises: What is the dimension of this manifold? Knowing the answer to this question is not only theoretically important but it is also practically very significant because it tells us what are the "degrees of freedom" of the problem.

Chris Geyer found an elegant and beautiful answer using some rather sophisticated concepts from the theory of group actions and Lie groups (Theorem 5.10): The space $\mathcal{F}$ is

isomorphic to the quotient

$$\mathbf{O}(3,1) \times \mathbf{O}(3,1)/H_F,$$

where H_F is the stabilizer of any element, F, in $\mathcal{F}$. Now, it is easy to determine the dimension of H_F by determining the dimension of its Lie algebra, which is 3. As $\dim \mathbf{O}(3,1) = 6$, we find that $\dim \mathcal{F} = 2 \cdot 6 - 3 = 9$.

Of course, a certain amount of machinery is needed in order to understand how the above results are obtained: group actions, manifolds, Lie groups, homogenous spaces, Lorentz groups, etc. As most computer science students, even those specialized in computer vision or robotics, are not familiar with these concepts, we thought that it would be useful to give a fairly detailed exposition of these theories.

During the seminar, I also used some material from my book, Gallier [60], especially from Chapters 11, 12 and 14. Readers might find it useful to read some of this material beforehand or in parallel with these notes, especially Chapter 14, which gives a more elementary introduction to Lie groups and manifolds. For the reader's convenience, I have incorporated a slightly updated version of chapter 14 from [60] as Chapter 1 of this manuscript. In fact, during the seminar, I lectured on most of Chapter 2, but only on the "gentler" versions of Chapters 3, 5, as in [60] and not at all on Chapter 7, which was written after the course had ended.

One feature worth pointing out is that we give a complete proof of the surjectivity of the exponential map, $\exp\colon \mathfrak{so}(1,3) \to \mathbf{SO}_0(1,3)$, for the Lorentz group $\mathbf{SO}_0(3,1)$ (see Section 5.5, Theorem 5.22). Although we searched the literature quite thoroughly, we did not find a proof of this specific fact (the physics books we looked at, even the most reputable ones, seem to take this fact as obvious and there are also wrong proofs, see the Remark following Theorem 2.6). We are aware of two proofs of the surjectivity of $\exp\colon \mathfrak{so}(1,n) \to \mathbf{SO}_0(1,n)$ in the general case where where n is arbitrary: One due to Nishikawa [119] (1983) and an earlier one due to Marcel Riesz [127] (1957). In both cases, the proof is quite involved (40 pages or so). In the case of $\mathbf{SO}_0(1,3)$, a much simpler argument can be made using the fact that $\varphi\colon \mathbf{SL}(2,\mathbb{C}) \to \mathbf{SO}_0(1,3)$, is surjective and that its kernel is $\{I, -I\}$ (see Proposition 5.21). Actually, a proof of this fact is not easy to find in the literature either (and, beware there are wrong proofs, again, see the Remark following Theorem 2.6). We have made sure to provide all the steps of the proof of the surjectivity of $\exp\colon \mathfrak{so}(1,3) \to \mathbf{SO}_0(1,3)$. For more on this subject, see the discussion in Section 5.5, after Corollary 5.18.

One of the "revelations" I had while on sabbatical in Nicholas' group was that many of the data that radiologists deal with (for instance, "diffusion tensors") do not live in Euclidean spaces, which are flat, but instead in more complicated curved spaces (Riemannian manifolds). As a consequence, even a notion as simple as the average of a set of data does not make sense in such spaces. Similarly, it is not clear how to define the covariance matrix of a random vector.

Pennec [121], among others, introduced a framework based on Riemannian Geometry for defining some basic statistical notions on curved spaces and gave some algorithmic methods

to compute these basic notions. Based on work in Vincent Arsigny's Ph.D. thesis, Arsigny, Fillard, Pennec and Ayache [5] introduced a new Lie group structure on the space of symmetric positive definite matrices, which allowed them to transfer strandard statistical concepts to this space (abusively called "tensors".) One of my goals in writing these notes is to provide a rather thorough background in differential geometry so that one will then be well prepared to read the above papers by Arsigny, Fillard, Pennec, Ayache and others, on statistics on manifolds.

At first, when I was writing these notes, I felt that it was important to supply most proofs. However, when I reached manifolds and differential geometry concepts, such as connections, geodesics and curvature, I realized that how formidable a task it was! Since there are lots of very good book on differential geometry, not without regrets, I decided that it was best to try to "demistify" concepts rather than fill many pages with proofs. However, when omitting a proof, I give precise pointers to the literature. In some cases where the proofs are really beautiful, as in the Theorem of Hopf and Rinow, Myers' Theorem or the Cartan-Hadamard Theorem, I could not resist to supply complete proofs!

Experienced differential geometers may be surprised and perhaps even irritated by my selection of topics. I beg their forgiveness! Primarily, I have included topics that I felt would be useful for my purposes and thus, I have omitted some topics found in all respectable differential geomety book (such as spaces of constant curvature). On the other hand, I have occasionally included topics because I found them particularly beautiful (such as characteristic classes) even though they do not seem to be of any use in medical imaging or computer vision. I also hope that readers with a more modest background will not be put off by the level of abstraction in some of the chapters and instead will be inspired to read more about these concepts, including fibre bundles!

I have also included chapters that present material having significant practical applications. These include

1. Chapter 4, on constructing manifolds from gluing data, has applications to surface reconstruction from 3D meshes,

2. Chapter 16, on spherical harmonics, has applications in computer graphics and computer vision

3. Chapter 19, on the "Log-Euclidean framework", has applications in medical imaging and

4. Chapter 21, on Clifford algebras and spinnors, has applications in robotics and computer graphics.

Of course, as anyone who attempts to write about differential geometry and Lie groups, I faced the dilemma of including or not including a chapter on differential forms. Given that our intented audience probably knows very little about them, I decided to provide a fairly

detailed treatment including a brief treatment of vector-valued differential forms. Of course, this made it necessary to review tensor products, exterior powers, *etc.*, and I have included a rather extensive chapter on this material.

I must aknowledge my debt to two of my main sources of inspiration: Berger's *Panoramic View of Riemannian Geometry* [16] and Milnor's *Morse Theory* [107]. In my opinion, Milnor's book is still one of the best references on basic differential geometry. His exposition is remarkably clear and insightful and his treatment of the variational approach to geodesics is unsurpassed. We borrowed heavily from Milnor [107]. Since Milnor's book is typeset in "ancient" typewritten format (1973!), readers might enjoy reading parts of it typeset in LaTeX. I hope that the readers of these notes will be well prepared to read standard differential geometry texts such as do Carmo [51], Gallot, Hulin, Lafontaine [61] and O'Neill [120] but also more advanced sources such as Sakai [131], Petersen [122], Jost [84], Knapp [90] and of course, Milnor [107].

Acknowledgement: I would like to thank Eugenio Calabi, Chris Croke, Ron Donagi, David Harbater, Herman Gluck, Alexander Kirillov, Steve Shatz and Wolfgand Ziller for their encouragement, advice, inspiration and for what they taught us.

Contents

Chapter 1

Introduction to Manifolds and Lie Groups

> Le rôle prépondérant de la théorie des groupes en mathématiques a été longtemps insoupçonné; il y a quatre-vingts ans, le nom même de groupe était ignoré. C'est Galois qui, le premier, en a eu une notion claire, mais c'est seulement depuis les travaux de Klein et surtout de Lie que l'on a commencé à voir qu'il n'y a presque aucune théorie mathématique où cette notion ne tienne une place importante.
>
> —**Henri Poincaré**

1.1 The Exponential Map

The purpose of this chapter is to give a "gentle" and fairly concrete introduction to manifolds, Lie groups and Lie algebras, our main objects of study.

Most texts on Lie groups and Lie algebras begin with prerequisites in differential geometry that are often formidable to average computer scientists (or average scientists, whatever that means!). We also struggled for a long time, trying to figure out what Lie groups and Lie algebras are all about, but this can be done! A good way to sneak into the wonderful world of Lie groups and Lie algebras is to play with explicit matrix groups such as the group of rotations in $\mathbb{R}^2$ (or $\mathbb{R}^3$) and with the exponential map. After actually computing the exponential $A = e^B$ of a 2×2 skew symmetric matrix B and observing that it is a rotation matrix, and similarly for a 3×3 skew symmetric matrix B, one begins to suspect that there is something deep going on. Similarly, after the discovery that every real invertible $n \times n$ matrix A can be written as $A = RP$, where R is an orthogonal matrix and P is a positive definite symmetric matrix, and that P can be written as $P = e^S$ for some symmetric matrix S, one begins to appreciate the exponential map.

Our goal in this chapter is to give an elementary and concrete introduction to Lie groups and Lie algebras by studying a number of the so-called *classical groups*, such as the general linear group $\mathbf{GL}(n, \mathbb{R})$, the special linear group $\mathbf{SL}(n, \mathbb{R})$, the orthogonal group $\mathbf{O}(n)$, the

special orthogonal group $\mathbf{SO}(n)$, and the group of affine rigid motions $\mathbf{SE}(n)$, and their Lie algebras $\mathfrak{gl}(n,\mathbb{R})$ (all matrices), $\mathfrak{sl}(n,\mathbb{R})$ (matrices with null trace), $\mathfrak{o}(n)$, and $\mathfrak{so}(n)$ (skew symmetric matrices). Now, Lie groups are at the same time, groups, topological spaces and manifolds, so we will also have to introduce the crucial notion of a *manifold*.

The inventors of Lie groups and Lie algebras (starting with Lie!) regarded Lie groups as groups of symmetries of various topological or geometric objects. Lie algebras were viewed as the "infinitesimal transformations" associated with the symmetries in the Lie group. For example, the group $\mathbf{SO}(n)$ of rotations is the group of orientation-preserving isometries of the Euclidean space $\mathbb{E}^n$. The Lie algebra $\mathfrak{so}(n,\mathbb{R})$ consisting of real skew symmetric $n \times n$ matrices is the corresponding set of infinitesimal rotations. The geometric link between a Lie group and its Lie algebra is the fact that the Lie algebra can be viewed as the tangent space to the Lie group at the identity. There is a map from the tangent space to the Lie group, called the *exponential map*. The Lie algebra can be considered as a linearization of the Lie group (near the identity element), and the exponential map provides the "delinearization," i.e., it takes us back to the Lie group. These concepts have a concrete realization in the case of groups of matrices and, for this reason, we begin by studying the behavior of the exponential maps on matrices.

We begin by defining the exponential map on matrices and proving some of its properties. The exponential map allows us to "linearize" certain algebraic properties of matrices. It also plays a crucial role in the theory of linear differential equations with constant coefficients. But most of all, as we mentioned earlier, it is a stepping stone to Lie groups and Lie algebras. On the way to Lie algebras, we derive the classical "Rodrigues-like" formulae for rotations and for rigid motions in $\mathbb{R}^2$ and $\mathbb{R}^3$. We give an elementary proof that the exponential map is surjective for both $\mathbf{SO}(n)$ and $\mathbf{SE}(n)$, not using any topology, just certain normal forms for matrices (see Gallier [60], Chapters 11 and 12).

The last section gives a quick introduction to manifolds, Lie groups and Lie algebras. Rather than defining abstract manifolds in terms of charts, atlases, *etc.*, we consider the special case of embedded submanifolds of $\mathbb{R}^N$. This approach has the pedagogical advantage of being more concrete since it uses parametrizations of subsets of $\mathbb{R}^N$, which should be familiar to the reader in the case of curves and surfaces. The general definition of a manifold will be given in Chapter 3.

Also, rather than defining Lie groups in full generality, we define linear Lie groups using the famous result of Cartan (apparently actually due to Von Neumann) that a closed subgroup of $\mathbf{GL}(n,\mathbb{R})$ is a manifold, and thus a Lie group. This way, Lie algebras can be "computed" using tangent vectors to curves of the form $t \mapsto A(t)$, where $A(t)$ is a matrix. This section is inspired from Artin [7], Chevalley [34], Marsden and Ratiu [103], Curtis [39], Howe [81], and Sattinger and Weaver [135].

Given an $n \times n$ (real or complex) matrix $A = (a_{i\,j})$, we would like to define the exponential

e^A of A as the sum of the series

$$e^A = I_n + \sum_{p \geq 1} \frac{A^p}{p!} = \sum_{p \geq 0} \frac{A^p}{p!},$$

letting $A^0 = I_n$. The problem is, Why is it well-defined? The following lemma shows that the above series is indeed absolutely convergent.

Lemma 1.1. *Let $A = (a_{i\,j})$ be a (real or complex) $n \times n$ matrix, and let*

$$\mu = \max\{|a_{i\,j}| \mid 1 \leq i, j \leq n\}.$$

If $A^p = (a_{i\,j}^{(p)})$, then

$$\left|a_{i\,j}^{(p)}\right| \leq (n\mu)^p$$

for all i, j, $1 \leq i, j \leq n$. As a consequence, the n^2 series

$$\sum_{p \geq 0} \frac{a_{i\,j}^{(p)}}{p!}$$

converge absolutely, and the matrix

$$e^A = \sum_{p \geq 0} \frac{A^p}{p!}$$

is a well-defined matrix.

Proof. The proof is by induction on p. For $p = 0$, we have $A^0 = I_n$, $(n\mu)^0 = 1$, and the lemma is obvious. Assume that

$$|a_{i\,j}^{(p)}| \leq (n\mu)^p$$

for all i, j, $1 \leq i, j \leq n$. Then we have

$$\left|a_{i\,j}^{(p+1)}\right| = \left|\sum_{k=1}^{n} a_{i\,k}^{(p)} a_{k\,j}\right| \leq \sum_{k=1}^{n} \left|a_{i\,k}^{(p)}\right| |a_{k\,j}| \leq \mu \sum_{k=1}^{n} \left|a_{i\,k}^{(p)}\right| \leq n\mu(n\mu)^p = (n\mu)^{p+1},$$

for all i, j, $1 \leq i, j \leq n$. For every pair (i, j) such that $1 \leq i, j \leq n$, since

$$\left|a_{i\,j}^{(p)}\right| \leq (n\mu)^p,$$

the series

$$\sum_{p \geq 0} \frac{\left|a_{i\,j}^{(p)}\right|}{p!}$$

is bounded by the convergent series

$$e^{n\mu} = \sum_{p\geq 0} \frac{(n\mu)^p}{p!},$$

and thus it is absolutely convergent. This shows that

$$e^A = \sum_{k\geq 0} \frac{A^k}{k!}$$

is well defined. □

It is instructive to compute explicitly the exponential of some simple matrices. As an example, let us compute the exponential of the real skew symmetric matrix

$$A = \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix}.$$

We need to find an inductive formula expressing the powers A^n. Let us observe that

$$\begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix} = \theta \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix}^2 = -\theta^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then, letting

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

we have

$$\begin{aligned} A^{4n} &= \theta^{4n} I_2, \\ A^{4n+1} &= \theta^{4n+1} J, \\ A^{4n+2} &= -\theta^{4n+2} I_2, \\ A^{4n+3} &= -\theta^{4n+3} J, \end{aligned}$$

and so

$$e^A = I_2 + \frac{\theta}{1!}J - \frac{\theta^2}{2!}I_2 - \frac{\theta^3}{3!}J + \frac{\theta^4}{4!}I_2 + \frac{\theta^5}{5!}J - \frac{\theta^6}{6!}I_2 - \frac{\theta^7}{7!}J + \cdots.$$

Rearranging the order of the terms, we have

$$e^A = \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \cdots\right) I_2 + \left(\frac{\theta}{1!} - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots\right) J.$$

We recognize the power series for $\cos\theta$ and $\sin\theta$, and thus

$$e^A = \cos\theta I_2 + \sin\theta J,$$

that is

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

Thus, e^A is a rotation matrix! This is a general fact. If A is a skew symmetric matrix, then e^A is an orthogonal matrix of determinant $+1$, i.e., a rotation matrix. Furthermore, every rotation matrix is of this form; i.e., the exponential map from the set of skew symmetric matrices to the set of rotation matrices is surjective. In order to prove these facts, we need to establish some properties of the exponential map. But before that, let us work out another example showing that the exponential map is not always surjective. Let us compute the exponential of a real 2×2 matrix with null trace of the form

$$A = \begin{pmatrix} a & b \\ c & -a \end{pmatrix}.$$

We need to find an inductive formula expressing the powers A^n. Observe that

$$A^2 = (a^2 + bc)I_2 = -\det(A)I_2.$$

If $a^2 + bc = 0$, we have

$$e^A = I_2 + A.$$

If $a^2 + bc < 0$, let $\omega > 0$ be such that $\omega^2 = -(a^2 + bc)$. Then, $A^2 = -\omega^2 I_2$. We get

$$e^A = I_2 + \frac{A}{1!} - \frac{\omega^2}{2!}I_2 - \frac{\omega^2}{3!}A + \frac{\omega^4}{4!}I_2 + \frac{\omega^4}{5!}A - \frac{\omega^6}{6!}I_2 - \frac{\omega^6}{7!}A + \cdots.$$

Rearranging the order of the terms, we have

$$e^A = \left(1 - \frac{\omega^2}{2!} + \frac{\omega^4}{4!} - \frac{\omega^6}{6!} + \cdots\right) I_2 + \frac{1}{\omega}\left(\omega - \frac{\omega^3}{3!} + \frac{\omega^5}{5!} - \frac{\omega^7}{7!} + \cdots\right) A.$$

We recognize the power series for $\cos\omega$ and $\sin\omega$, and thus

$$e^A = \cos\omega\, I_2 + \frac{\sin\omega}{\omega}A.$$

If $a^2 + bc > 0$, let $\omega > 0$ be such that $\omega^2 = (a^2 + bc)$. Then $A^2 = \omega^2 I_2$. We get

$$e^A = I_2 + \frac{A}{1!} + \frac{\omega^2}{2!}I_2 + \frac{\omega^2}{3!}A + \frac{\omega^4}{4!}I_2 + \frac{\omega^4}{5!}A + \frac{\omega^6}{6!}I_2 + \frac{\omega^6}{7!}A + \cdots.$$

Rearranging the order of the terms, we have

$$e^A = \left(1 + \frac{\omega^2}{2!} + \frac{\omega^4}{4!} + \frac{\omega^6}{6!} + \cdots\right) I_2 + \frac{1}{\omega}\left(\omega + \frac{\omega^3}{3!} + \frac{\omega^5}{5!} + \frac{\omega^7}{7!} + \cdots\right) A.$$

If we recall that $\cosh\omega = \left(e^{\omega} + e^{-\omega}\right)/2$ and $\sinh\omega = \left(e^{\omega} - e^{-\omega}\right)/2$, we recognize the power series for $\cosh\omega$ and $\sinh\omega$, and thus

$$e^A = \cosh\omega\, I_2 + \frac{\sinh\omega}{\omega} A.$$

It immediately verified that in all cases,

$$\det\left(e^A\right) = 1.$$

This shows that the exponential map is a function from the set of 2×2 matrices with null trace to the set of 2×2 matrices with determinant 1. This function is not surjective. Indeed, $tr(e^A) = 2\cos\omega$ when $a^2 + bc < 0$, $tr(e^A) = 2\cosh\omega$ when $a^2 + bc > 0$, and $tr(e^A) = 2$ when $a^2 + bc = 0$. As a consequence, for any matrix A with null trace,

$$\operatorname{tr}\left(e^A\right) \geq -2,$$

and any matrix B with determinant 1 and whose trace is less than -2 is not the exponential e^A of any matrix A with null trace. For example,

$$B = \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix},$$

where $a < 0$ and $a \neq -1$, is not the exponential of any matrix A with null trace.

A fundamental property of the exponential map is that if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, then the eigenvalues of e^A are $e^{\lambda_1}, \ldots, e^{\lambda_n}$. For this we need two lemmas.

Lemma 1.2. *Let A and U be (real or complex) matrices, and assume that U is invertible. Then*

$$e^{UAU^{-1}} = Ue^AU^{-1}.$$

Proof. A trivial induction shows that

$$UA^pU^{-1} = (UAU^{-1})^p,$$

and thus

$$\begin{aligned} e^{UAU^{-1}} &= \sum_{p\geq 0} \frac{(UAU^{-1})^p}{p!} = \sum_{p\geq 0} \frac{UA^pU^{-1}}{p!} \\ &= U\left(\sum_{p\geq 0} \frac{A^p}{p!}\right) U^{-1} = Ue^AU^{-1}. \end{aligned}$$

□

Say that a square matrix A is an *upper triangular matrix* if it has the following shape,

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & a_{1\,3} & \dots & a_{1\,n-1} & a_{1\,n} \\ 0 & a_{2\,2} & a_{2\,3} & \dots & a_{2\,n-1} & a_{2\,n} \\ 0 & 0 & a_{3\,3} & \dots & a_{3\,n-1} & a_{3\,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{n-1\,n-1} & a_{n-1\,n} \\ 0 & 0 & 0 & \dots & 0 & a_{n\,n} \end{pmatrix},$$

i.e., $a_{i\,j} = 0$ whenever $j < i$, $1 \leq i, j \leq n$.

Lemma 1.3. *Given any complex $n \times n$ matrix A, there is an invertible matrix P and an upper triangular matrix T such that*

$$A = PTP^{-1}.$$

Proof. We prove by induction on n that if $f\colon \mathbb{C}^n \to \mathbb{C}^n$ is a linear map, then there is a basis $(u_1, \dots, u_n)$ with respect to which f is represented by an upper triangular matrix. For $n = 1$ the result is obvious. If $n > 1$, since $\mathbb{C}$ is algebraically closed, f has some eigenvalue $\lambda_1 \in \mathbb{C}$, and let u_1 be an eigenvector for λ_1. We can find $n-1$ vectors $(v_2, \dots, v_n)$ such that $(u_1, v_2, \dots, v_n)$ is a basis of $\mathbb{C}^n$, and let W be the subspace of dimension $n-1$ spanned by $(v_2, \dots, v_n)$. In the basis $(u_1, v_2 \dots, v_n)$, the matrix of f is of the form

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & \dots & a_{1\,n} \\ 0 & a_{2\,2} & \dots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n\,2} & \dots & a_{n\,n} \end{pmatrix},$$

since its first column contains the coordinates of $\lambda_1 u_1$ over the basis $(u_1, v_2, \dots, v_n)$. Letting $p\colon \mathbb{C}^n \to W$ be the projection defined such that $p(u_1) = 0$ and $p(v_i) = v_i$ when $2 \leq i \leq n$, the linear map $g\colon W \to W$ defined as the restriction of $p \circ f$ to W is represented by the $(n-1) \times (n-1)$ matrix $(a_{i\,j})_{2 \leq i,j \leq n}$ over the basis $(v_2, \dots, v_n)$. By the induction hypothesis, there is a basis $(u_2, \dots, u_n)$ of W such that g is represented by an upper triangular matrix $(b_{i\,j})_{1 \leq i,j \leq n-1}$.

However,

$$\mathbb{C}^n = \mathbb{C}u_1 \oplus W,$$

and thus $(u_1, \dots, u_n)$ is a basis for $\mathbb{C}^n$. Since p is the projection from $\mathbb{C}^n = \mathbb{C}u_1 \oplus W$ onto W and $g\colon W \to W$ is the restriction of $p \circ f$ to W, we have

$$f(u_1) = \lambda_1 u_1$$

and

$$f(u_{i+1}) = a_{1\,i} u_1 + \sum_{j=1}^{n-1} b_{i\,j} u_{j+1}$$

for some $a_{1\,i} \in \mathbb{C}$, when $1 \leq i \leq n-1$. But then the matrix of f with respect to $(u_1, \ldots, u_n)$ is upper triangular. Thus, there is a change of basis matrix P such that $A = PTP^{-1}$ where T is upper triangular. □

Remark: If E is a Hermitian space, the proof of Lemma 1.3 can be easily adapted to prove that there is an *orthonormal* basis $(u_1, \ldots, u_n)$ with respect to which the matrix of f is upper triangular. In terms of matrices, this means that there is a unitary matrix U and an upper triangular matrix T such that $A = UTU^*$. This is usually known as *Schur's lemma*. Using this result, we can immediately rederive the fact that if A is a Hermitian matrix, then there is a unitary matrix U and a real diagonal matrix D such that $A = UDU^*$.

If $A = PTP^{-1}$ where T is upper triangular, note that the diagonal entries on T are the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A. Indeed, A and T have the same characteristic polynomial. This is because if A and B are any two matrices such that $A = PBP^{-1}$, then

$$\begin{aligned}
\det(A - \lambda\, I) &= \det(PBP^{-1} - \lambda\, P\, IP^{-1}),\\
&= \det(P(B - \lambda\, I)P^{-1}),\\
&= \det(P)\det(B - \lambda\, I)\det(P^{-1}),\\
&= \det(P)\det(B - \lambda\, I)\det(P)^{-1},\\
&= \det(B - \lambda\, I).
\end{aligned}$$

Furthermore, it is well known that the determinant of a matrix of the form

$$\begin{pmatrix}
\lambda_1 - \lambda & a_{1\,2} & a_{1\,3} & \ldots & a_{1\,n-1} & a_{1\,n}\\
0 & \lambda_2 - \lambda & a_{2\,3} & \ldots & a_{2\,n-1} & a_{2\,n}\\
0 & 0 & \lambda_3 - \lambda & \ldots & a_{3\,n-1} & a_{3\,n}\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\
0 & 0 & 0 & \ldots & \lambda_{n-1} - \lambda & a_{n-1\,n}\\
0 & 0 & 0 & \ldots & 0 & \lambda_n - \lambda
\end{pmatrix}$$

is $(\lambda_1 - \lambda)\cdots(\lambda_n - \lambda)$, and thus the eigenvalues of $A = PTP^{-1}$ are the diagonal entries of T. We use this property to prove the following lemma.

Lemma 1.4. *Given any complex $n \times n$ matrix A, if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, then $e^{\lambda_1}, \ldots, e^{\lambda_n}$ are the eigenvalues of e^A. Furthermore, if u is an eigenvector of A for λ_i, then u is an eigenvector of e^A for e^{λ_i}.*

Proof. By Lemma 1.3 there is an invertible matrix P and an upper triangular matrix T such that

$$A = PTP^{-1}.$$

By Lemma 1.2,

$$e^{PTP^{-1}} = Pe^TP^{-1}.$$

However, we showed that A and T have the same eigenvalues, which are the diagonal entries $\lambda_1, \ldots, \lambda_n$ of T, and $e^A = e^{PTP^{-1}} = Pe^TP^{-1}$ and e^T have the same eigenvalues, which are the diagonal entries of e^T. Clearly, the diagonal entries of e^T are $e^{\lambda_1}, \ldots, e^{\lambda_n}$. Now, if u is an eigenvector of A for the eigenvalue λ, a simple induction shows that u is an eigenvector of A^n for the eigenvalue λ^n, from which is follows that u is an eigenvector of e^A for e^λ. □

As a consequence, we can show that

$$\det(e^A) = e^{\mathrm{tr}(A)},$$

where $\mathrm{tr}(A)$ is the *trace of* A, i.e., the sum $a_{11} + \cdots + a_{nn}$ of its diagonal entries, which is also equal to the sum of the eigenvalues of A. This is because the determinant of a matrix is equal to the product of its eigenvalues, and if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, then by Lemma 1.4, $e^{\lambda_1}, \ldots, e^{\lambda_n}$ are the eigenvalues of e^A, and thus

$$\det\left(e^A\right) = e^{\lambda_1} \cdots e^{\lambda_n} = e^{\lambda_1 + \cdots + \lambda_n} = e^{\mathrm{tr}(A)}.$$

This shows that e^A is always an invertible matrix, since e^z is never null for every $z \in \mathbb{C}$. In fact, the inverse of e^A is e^{-A}, but we need to prove another lemma. This is because it is generally not true that

$$e^{A+B} = e^Ae^B,$$

unless A and B commute, i.e., $AB = BA$. We need to prove this last fact.

Lemma 1.5. *Given any two complex $n \times n$ matrices A, B, if $AB = BA$, then*

$$e^{A+B} = e^Ae^B.$$

Proof. Since $AB = BA$, we can expand $(A + B)^p$ using the binomial formula:

$$(A+B)^p = \sum_{k=0}^{p} \binom{p}{k} A^kB^{p-k},$$

and thus

$$\frac{1}{p!}(A+B)^p = \sum_{k=0}^{p} \frac{A^kB^{p-k}}{k!(p-k)!}.$$

Note that for any integer $N \geq 0$, we can write

$$\begin{aligned}
\sum_{p=0}^{2N} \frac{1}{p!}(A+B)^p &= \sum_{p=0}^{2N}\sum_{k=0}^{p} \frac{A^kB^{p-k}}{k!(p-k)!} \\
&= \left(\sum_{p=0}^{N} \frac{A^p}{p!}\right)\left(\sum_{p=0}^{N} \frac{B^p}{p!}\right) + \sum_{\substack{\max(k,l) > N \\ k+l \leq 2N}} \frac{A^k}{k!}\frac{B^l}{l!},
\end{aligned}$$

where there are $N(N+1)$ pairs (k,l) in the second term. Letting

$$\|A\| = \max\{|a_{ij}| \mid 1 \leq i,j \leq n\}, \quad \|B\| = \max\{|b_{ij}| \mid 1 \leq i,j \leq n\},$$

and $\mu = \max(\|A\|, \|B\|)$, note that for every entry c_{ij} in $\left(A^k/k!\right)\left(B^l/l!\right)$ we have

$$|c_{ij}| \leq n\frac{(n\mu)^k}{k!}\frac{(n\mu)^l}{l!} \leq \frac{(n^2\mu)^{2N}}{N!}.$$

As a consequence, the absolute value of every entry in

$$\sum_{\substack{\max(k,l) > N \\ k+l \leq 2N}} \frac{A^k}{k!}\frac{B^l}{l!}$$

is bounded by

$$N(N+1)\frac{(n^2\mu)^{2N}}{N!},$$

which goes to 0 as $N \mapsto \infty$. From this, it immediately follows that

$$e^{A+B} = e^A e^B.$$

□

Now, using Lemma 1.5, since A and $-A$ commute, we have

$$e^A e^{-A} = e^{A+-A} = e^{0_n} = I_n,$$

which shows that the inverse of e^A is e^{-A}.

We will now use the properties of the exponential that we have just established to show how various matrices can be represented as exponentials of other matrices.

1.2 The Lie Groups $\mathbf{GL}(n,\mathbb{R})$, $\mathbf{SL}(n,\mathbb{R})$, $\mathbf{O}(n)$, $\mathbf{SO}(n)$, the Lie Algebras $\mathfrak{gl}(n,\mathbb{R})$, $\mathfrak{sl}(n,\mathbb{R})$, $\mathfrak{o}(n)$, $\mathfrak{so}(n)$, and the Exponential Map

First, we recall some basic facts and definitions. The set of real invertible $n \times n$ matrices forms a group under multiplication, denoted by $\mathbf{GL}(n,\mathbb{R})$. The subset of $\mathbf{GL}(n,\mathbb{R})$ consisting of those matrices having determinant $+1$ is a subgroup of $\mathbf{GL}(n,\mathbb{R})$, denoted by $\mathbf{SL}(n,\mathbb{R})$. It is also easy to check that the set of real $n \times n$ orthogonal matrices forms a group under multiplication, denoted by $\mathbf{O}(n)$. The subset of $\mathbf{O}(n)$ consisting of those matrices having determinant $+1$ is a subgroup of $\mathbf{O}(n)$, denoted by $\mathbf{SO}(n)$. We will also call matrices in $\mathbf{SO}(n)$ *rotation matrices*. Staying with easy things, we can check that the set of real $n \times n$ matrices with null trace forms a vector space under addition, and similarly for the set of skew symmetric matrices.

Definition 1.1. The group $\mathbf{GL}(n, \mathbb{R})$ is called the *general linear group*, and its subgroup $\mathbf{SL}(n, \mathbb{R})$ is called the *special linear group*. The group $\mathbf{O}(n)$ of orthogonal matrices is called the *orthogonal group*, and its subgroup $\mathbf{SO}(n)$ is called the *special orthogonal group* (or *group of rotations*). The vector space of real $n \times n$ matrices with null trace is denoted by $\mathfrak{sl}(n, \mathbb{R})$, and the vector space of real $n \times n$ skew symmetric matrices is denoted by $\mathfrak{so}(n)$.

Remark: The notation $\mathfrak{sl}(n, \mathbb{R})$ and $\mathfrak{so}(n)$ is rather strange and deserves some explanation. The groups $\mathbf{GL}(n, \mathbb{R})$, $\mathbf{SL}(n, \mathbb{R})$, $\mathbf{O}(n)$, and $\mathbf{SO}(n)$ are more than just groups. They are also topological groups, which means that they are topological spaces (viewed as subspaces of $\mathbb{R}^{n^2}$) and that the multiplication and the inverse operations are continuous (in fact, smooth). Furthermore, they are smooth real manifolds.[1] Such objects are called *Lie groups*. The real vector spaces $\mathfrak{sl}(n)$ and $\mathfrak{so}(n)$ are what is called *Lie algebras*. However, we have not defined the algebra structure on $\mathfrak{sl}(n, \mathbb{R})$ and $\mathfrak{so}(n)$ yet. The algebra structure is given by what is called the *Lie bracket*, which is defined as

$$[A, B] = AB - BA.$$

Lie algebras are associated with Lie groups. What is going on is that the Lie algebra of a Lie group is its tangent space at the identity, i.e., the space of all tangent vectors at the identity (in this case, I_n). In some sense, the Lie algebra achieves a "linearization" of the Lie group. The exponential map is a map from the Lie algebra to the Lie group, for example,

$$\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$$

and

$$\exp\colon \mathfrak{sl}(n, \mathbb{R}) \to \mathbf{SL}(n, \mathbb{R}).$$

The exponential map often allows a parametrization of the Lie group elements by simpler objects, the Lie algebra elements.

One might ask, What happened to the Lie algebras $\mathfrak{gl}(n, \mathbb{R})$ and $\mathfrak{o}(n)$ associated with the Lie groups $\mathbf{GL}(n, \mathbb{R})$ and $\mathbf{O}(n)$? We will see later that $\mathfrak{gl}(n, \mathbb{R})$ is the set of *all* real $n \times n$ matrices, and that $\mathfrak{o}(n) = \mathfrak{so}(n)$.

The properties of the exponential map play an important role in studying a Lie group. For example, it is clear that the map

$$\exp\colon \mathfrak{gl}(n, \mathbb{R}) \to \mathbf{GL}(n, \mathbb{R})$$

is well-defined, but since every matrix of the form e^A has a positive determinant, exp is not surjective. Similarly, since

$$\det(e^A) = e^{\mathrm{tr}(A)},$$

[1] We refrain from defining manifolds right now, not to interupt the flow of intuitive ideas.

the map

$$\exp\colon \mathfrak{sl}(n,\mathbb{R}) \to \mathbf{SL}(n,\mathbb{R})$$

is well-defined. However, we showed in Section 1.1 that it is not surjective either. As we will see in the next theorem, the map

$$\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$$

is well-defined and surjective. The map

$$\exp\colon \mathfrak{o}(n) \to \mathbf{O}(n)$$

is well-defined, but it is not surjective, since there are matrices in $\mathbf{O}(n)$ with determinant -1.

Remark: The situation for matrices over the field $\mathbb{C}$ of complex numbers is quite different, as we will see later.

We now show the fundamental relationship between $\mathbf{SO}(n)$ and $\mathfrak{so}(n)$.

Theorem 1.6. *The exponential map*

$$\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$$

is well-defined and surjective.

Proof. First, we need to prove that if A is a skew symmetric matrix, then e^A is a rotation matrix. For this, first check that

$$\left(e^A\right)^\top = e^{A^\top}.$$

Then, since $A^\top = -A$, we get

$$\left(e^A\right)^\top = e^{A^\top} = e^{-A},$$

and so

$$\left(e^A\right)^\top e^A = e^{-A}e^A = e^{-A+A} = e^{0_n} = I_n,$$

and similarly,

$$e^A\left(e^A\right)^\top = I_n,$$

showing that e^A is orthogonal. Also,

$$\det\left(e^A\right) = e^{\mathrm{tr}(A)},$$

and since A is real skew symmetric, its diagonal entries are 0, i.e., $\mathrm{tr}(A) = 0$, and so $\det(e^A) = +1$.

For the surjectivity, we will use Theorem 11.4.4 and Theorem 11.4.5, from Chapter 11 of Gallier [60]. Theorem 11.4.4 says that for every skew symmetric matrix A there is an

orthogonal matrix P such that $A = PD\,P^\top$, where D is a block diagonal matrix of the form

$$D = \begin{pmatrix} D_1 & & \cdots & \\ & D_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & D_p \end{pmatrix}$$

such that each block D_i is either 0 or a two-dimensional matrix of the form

$$D_i = \begin{pmatrix} 0 & -\theta_i \\ \theta_i & 0 \end{pmatrix}$$

where $\theta_i \in \mathbb{R}$, with $\theta_i > 0$. Theorem 11.4.5 says that for every orthogonal matrix R there is an orthogonal matrix P such that $R = PE\,P^\top$, where E is a block diagonal matrix of the form

$$E = \begin{pmatrix} E_1 & & \cdots & \\ & E_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & E_p \end{pmatrix}$$

such that each block E_i is either 1, -1, or a two-dimensional matrix of the form

$$E_i = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix}.$$

If R is a rotation matrix, there is an even number of -1's and they can be grouped into blocks of size 2 associated with $\theta = \pi$. Let D be the block matrix associated with E in the obvious way (where an entry 1 in E is associated with a 0 in D). Since by Lemma 1.2

$$e^A = e^{PDP^{-1}} = Pe^D P^{-1},$$

and since D is a block diagonal matrix, we can compute e^D by computing the exponentials of its blocks. If $D_i = 0$, we get $E_i = e^0 = +1$, and if

$$D_i = \begin{pmatrix} 0 & -\theta_i \\ \theta_i & 0 \end{pmatrix},$$

we showed earlier that

$$e^{D_i} = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix},$$

exactly the block E_i. Thus, $E = e^D$, and as a consequence,

$$e^A = e^{PDP^{-1}} = Pe^D P^{-1} = PEP^{-1} = PE\,P^\top = R.$$

This shows the surjectivity of the exponential. □

When $n = 3$ (and A is skew symmetric), it is possible to work out an explicit formula for e^A. For any 3×3 real skew symmetric matrix

$$A = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix},$$

letting $\theta = \sqrt{a^2 + b^2 + c^2}$ and

$$B = \begin{pmatrix} a^2 & ab & ac \\ ab & b^2 & bc \\ ac & bc & c^2 \end{pmatrix},$$

we have the following result known as *Rodrigues's formula* (1840).

Lemma 1.7. *The exponential map* $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ *is given by*

$$e^A = \cos\theta\, I_3 + \frac{\sin\theta}{\theta} A + \frac{(1 - \cos\theta)}{\theta^2} B,$$

or, equivalently, by

$$e^A = I_3 + \frac{\sin\theta}{\theta} A + \frac{(1 - \cos\theta)}{\theta^2} A^2$$

if $\theta \neq 0$*, with* $e^{0_3} = I_3$*.*

Proof sketch. First, prove that

$$\begin{aligned} A^2 &= -\theta^2 I + B, \\ AB &= BA = 0. \end{aligned}$$

From the above, deduce that

$$A^3 = -\theta^2 A,$$

and for any $k \geq 0$,

$$\begin{aligned} A^{4k+1} &= \theta^{4k} A, \\ A^{4k+2} &= \theta^{4k} A^2, \\ A^{4k+3} &= -\theta^{4k+2} A, \\ A^{4k+4} &= -\theta^{4k+2} A^2. \end{aligned}$$

Then prove the desired result by writing the power series for e^A and regrouping terms so that the power series for cos and sin show up. □

The above formulae are the well-known formulae expressing a rotation of axis specified by the vector (a, b, c) and angle θ. Since the exponential is surjective, it is possible to write down an explicit formula for its inverse (but it is a multivalued function!). This has applications in kinematics, robotics, and motion interpolation.

1.3 Symmetric Matrices, Symmetric Positive Definite Matrices, and the Exponential Map

Recall that a real symmetric matrix is called *positive* (or *positive semidefinite*) if its eigenvalues are all positive or null, and *positive definite* if its eigenvalues are all strictly positive. We denote the vector space of real symmetric $n \times n$ matrices by $\mathbf{S}(n)$, the set of symmetric positive matrices by $\mathbf{SP}(n)$, and the set of symmetric positive definite matrices by $\mathbf{SPD}(n)$.

The next lemma shows that every symmetric positive definite matrix A is of the form e^B for some unique symmetric matrix B. The set of symmetric matrices is a vector space, but it is not a Lie algebra because the Lie bracket $[A, B]$ is not symmetric unless A and B commute, and the set of symmetric (positive) definite matrices is not a multiplicative group, so this result is of a different flavor as Theorem 1.6.

Lemma 1.8. *For every symmetric matrix B, the matrix e^B is symmetric positive definite. For every symmetric positive definite matrix A, there is a unique symmetric matrix B such that $A = e^B$.*

Proof. We showed earlier that

$$\left(e^B\right)^\top = e^{B^\top}.$$

If B is a symmetric matrix, then since $B^\top = B$, we get

$$\left(e^B\right)^\top = e^{B^\top} = e^B,$$

and e^B is also symmetric. Since the eigenvalues $\lambda_1, \ldots, \lambda_n$ of the symmetric matrix B are real and the eigenvalues of e^B are $e^{\lambda_1}, \ldots, e^{\lambda_n}$, and since $e^\lambda > 0$ if $\lambda \in \mathbb{R}$, e^B is positive definite.

If A is symmetric positive definite, by Theorem 11.4.3 from Chapter 11 of Gallier [60], there is an orthogonal matrix P such that $A = PD\,P^\top$, where D is a diagonal matrix

$$D = \begin{pmatrix} \lambda_1 & & \cdots & \\ & \lambda_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \lambda_n \end{pmatrix},$$

where $\lambda_i > 0$, since A is positive definite. Letting

$$L = \begin{pmatrix} \log \lambda_1 & & \cdots & \\ & \log \lambda_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \log \lambda_n \end{pmatrix},$$

it is obvious that $e^L = D$, with $\log \lambda_i \in \mathbb{R}$, since $\lambda_i > 0$.

Let

$$B = PL\,P^\top.$$

By Lemma 1.2, we have

$$e^B = e^{PL\,P^\top} = e^{PLP^{-1}} = Pe^L P^{-1} = Pe^L\,P^\top = PD\,P^\top = A.$$

Finally, we prove that if B_1 and B_2 are symmetric and $A = e^{B_1} = e^{B_2}$, then $B_1 = B_2$. Since B_1 is symmetric, there is an orthonormal basis $(u_1, \ldots, u_n)$ of eigenvectors of B_1. Let $\mu_1, \ldots, \mu_n$ be the corresponding eigenvalues. Similarly, there is an orthonormal basis $(v_1, \ldots, v_n)$ of eigenvectors of B_2. We are going to prove that B_1 and B_2 agree on the basis $(v_1, \ldots, v_n)$, thus proving that $B_1 = B_2$.

Let μ be some eigenvalue of B_2, and let $v = v_i$ be some eigenvector of B_2 associated with μ. We can write

$$v = \alpha_1 u_1 + \cdots + \alpha_n u_n.$$

Since v is an eigenvector of B_2 for μ and $A = e^{B_2}$, by Lemma 1.4

$$A(v) = e^\mu v = e^\mu \alpha_1 u_1 + \cdots + e^\mu \alpha_n u_n.$$

On the other hand,

$$A(v) = A(\alpha_1 u_1 + \cdots + \alpha_n u_n) = \alpha_1 A(u_1) + \cdots + \alpha_n A(u_n),$$

and since $A = e^{B_1}$ and $B_1(u_i) = \mu_i u_i$, by Lemma 1.4 we get

$$A(v) = e^{\mu_1} \alpha_1 u_1 + \cdots + e^{\mu_n} \alpha_n u_n.$$

Therefore, $\alpha_i = 0$ if $\mu_i \neq \mu$. Letting

$$I = \{i \mid \mu_i = \mu,\ i \in \{1, \ldots, n\}\},$$

we have

$$v = \sum_{i \in I} \alpha_i u_i.$$

Now,

$$\begin{aligned} B_1(v) &= B_1\left(\sum_{i \in I} \alpha_i u_i\right) = \sum_{i \in I} \alpha_i B_1(u_i) = \sum_{i \in I} \alpha_i \mu_i u_i \\ &= \sum_{i \in I} \alpha_i \mu u_i = \mu\left(\sum_{i \in I} \alpha_i u_i\right) = \mu v, \end{aligned}$$

since $\mu_i = \mu$ when $i \in I$. Since v is an eigenvector of B_2 for μ,

$$B_2(v) = \mu v,$$

which shows that

$$B_1(v) = B_2(v).$$

Since the above holds for every eigenvector v_i, we have $B_1 = B_2$. □

Lemma 1.8 can be reformulated as stating that the map $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$ is a bijection. It can be shown that it is a homeomorphism. In the case of invertible matrices, the polar form theorem can be reformulated as stating that there is a bijection between the topological space $\mathbf{GL}(n,\mathbb{R})$ of real $n \times n$ invertible matrices (also a group) and $\mathbf{O}(n) \times \mathbf{SPD}(n)$.

As a corollary of the polar form theorem (Theorem 12.1.3 in Chapter 12 of Gallier [60]) and Lemma 1.8, we have the following result: For every invertible matrix A there is a unique orthogonal matrix R and a unique symmetric matrix S such that

$$A = R\,e^S.$$

Thus, we have a bijection between $\mathbf{GL}(n,\mathbb{R})$ and $\mathbf{O}(n) \times \mathbf{S}(n)$. But $\mathbf{S}(n)$ itself is isomorphic to $\mathbb{R}^{n(n+1)/2}$. Thus, there is a bijection between $\mathbf{GL}(n,\mathbb{R})$ and $\mathbf{O}(n) \times \mathbb{R}^{n(n+1)/2}$. It can also be shown that this bijection is a homeomorphism. This is an interesting fact. Indeed, this homeomorphism essentially reduces the study of the topology of $\mathbf{GL}(n,\mathbb{R})$ to the study of the topology of $\mathbf{O}(n)$. This is nice, since it can be shown that $\mathbf{O}(n)$ is compact.

In $A = R\,e^S$, if $\det(A) > 0$, then R must be a rotation matrix (i.e., $\det(R) = +1$), since $\det\left(e^S\right) > 0$. In particular, if $A \in \mathbf{SL}(n,\mathbb{R})$, since $\det(A) = \det(R) = +1$, the symmetric matrix S must have a null trace, i.e., $S \in \mathbf{S}(n) \cap \mathfrak{sl}(n,\mathbb{R})$. Thus, we have a bijection between $\mathbf{SL}(n,\mathbb{R})$ and $\mathbf{SO}(n) \times (\mathbf{S}(n) \cap \mathfrak{sl}(n,\mathbb{R}))$.

We can also show that the exponential map is a surjective map from the skew Hermitian matrices to the unitary matrices (use Theorem 11.4.7 from Chapter 11 in Gallier [60]).

1.4 The Lie Groups $\mathbf{GL}(n,\mathbb{C})$, $\mathbf{SL}(n,\mathbb{C})$, $\mathbf{U}(n)$, $\mathbf{SU}(n)$, the Lie Algebras $\mathfrak{gl}(n,\mathbb{C})$, $\mathfrak{sl}(n,\mathbb{C})$, $\mathfrak{u}(n)$, $\mathfrak{su}(n)$, and the Exponential Map

The set of complex invertible $n \times n$ matrices forms a group under multiplication, denoted by $\mathbf{GL}(n,\mathbb{C})$. The subset of $\mathbf{GL}(n,\mathbb{C})$ consisting of those matrices having determinant $+1$ is a subgroup of $\mathbf{GL}(n,\mathbb{C})$, denoted by $\mathbf{SL}(n,\mathbb{C})$. It is also easy to check that the set of complex $n \times n$ unitary matrices forms a group under multiplication, denoted by $\mathbf{U}(n)$. The subset of $\mathbf{U}(n)$ consisting of those matrices having determinant $+1$ is a subgroup of $\mathbf{U}(n)$, denoted by $\mathbf{SU}(n)$. We can also check that the set of complex $n \times n$ matrices with null trace forms a real vector space under addition, and similarly for the set of skew Hermitian matrices and the set of skew Hermitian matrices with null trace.

Definition 1.2. The group $\mathbf{GL}(n,\mathbb{C})$ is called the *general linear group*, and its subgroup $\mathbf{SL}(n,\mathbb{C})$ is called the *special linear group*. The group $\mathbf{U}(n)$ of unitary matrices is called the *unitary group*, and its subgroup $\mathbf{SU}(n)$ is called the *special unitary group*. The real vector space of complex $n \times n$ matrices with null trace is denoted by $\mathfrak{sl}(n,\mathbb{C})$, the real vector space of skew Hermitian matrices is denoted by $\mathfrak{u}(n)$, and the real vector space $\mathfrak{u}(n) \cap \mathfrak{sl}(n,\mathbb{C})$ is denoted by $\mathfrak{su}(n)$.

Remarks:

(1) As in the real case, the groups $\mathbf{GL}(n, \mathbb{C})$, $\mathbf{SL}(n, \mathbb{C})$, $\mathbf{U}(n)$, and $\mathbf{SU}(n)$ are also topological groups (viewed as subspaces of $\mathbb{R}^{2n^2}$), and in fact, smooth real manifolds. Such objects are called *(real) Lie groups*. The real vector spaces $\mathfrak{sl}(n, \mathbb{C})$, $\mathfrak{u}(n)$, and $\mathfrak{su}(n)$ are *Lie algebras* associated with $\mathbf{SL}(n, \mathbb{C})$, $\mathbf{U}(n)$, and $\mathbf{SU}(n)$. The algebra structure is given by the *Lie bracket*, which is defined as

$$[A,\, B] = AB - BA.$$

(2) It is also possible to define complex Lie groups, which means that they are topological groups and smooth *complex* manifolds. It turns out that $\mathbf{GL}(n, \mathbb{C})$ and $\mathbf{SL}(n, \mathbb{C})$ are complex manifolds, but not $\mathbf{U}(n)$ and $\mathbf{SU}(n)$.

One should be very careful to observe that even though the Lie algebras $\mathfrak{sl}(n, \mathbb{C})$, $\mathfrak{u}(n)$, and $\mathfrak{su}(n)$ consist of matrices with complex coefficients, we view them as *real* vector spaces. The Lie algebra $\mathfrak{sl}(n, \mathbb{C})$ is also a complex vector space, but $\mathfrak{u}(n)$ and $\mathfrak{su}(n)$ are not! Indeed, if A is a skew Hermitian matrix, iA is *not* skew Hermitian, but Hermitian!

Again the Lie algebra achieves a "linearization" of the Lie group. In the complex case, the Lie algebras $\mathfrak{gl}(n, \mathbb{C})$ is the set of *all* complex $n \times n$ matrices, but $\mathfrak{u}(n) \neq \mathfrak{su}(n)$, because a skew Hermitian matrix does not necessarily have a null trace.

The properties of the exponential map also play an important role in studying complex Lie groups. For example, it is clear that the map

$$\exp\colon \mathfrak{gl}(n, \mathbb{C}) \to \mathbf{GL}(n, \mathbb{C})$$

is well-defined, but this time, it is surjective! One way to prove this is to use the Jordan normal form. Similarly, since

$$\det\left(e^A\right) = e^{\mathrm{tr}(A)},$$

the map

$$\exp\colon \mathfrak{sl}(n, \mathbb{C}) \to \mathbf{SL}(n, \mathbb{C})$$

is well-defined, but it is not surjective! As we will see in the next theorem, the maps

$$\exp\colon \mathfrak{u}(n) \to \mathbf{U}(n)$$

and

$$\exp\colon \mathfrak{su}(n) \to \mathbf{SU}(n)$$

are well-defined and surjective.

Theorem 1.9. *The exponential maps*

$$\exp\colon \mathfrak{u}(n) \to \mathbf{U}(n) \quad \textit{and} \quad \exp\colon \mathfrak{su}(n) \to \mathbf{SU}(n)$$

are well-defined and surjective.

Proof. First, we need to prove that if A is a skew Hermitian matrix, then e^A is a unitary matrix. For this, first check that

$$\left(e^A\right)^* = e^{A^*}.$$

Then, since $A^* = -A$, we get

$$\left(e^A\right)^* = e^{A^*} = e^{-A},$$

and so

$$\left(e^A\right)^* e^A = e^{-A}e^A = e^{-A+A} = e^{0_n} = I_n,$$

and similarly, $e^A \left(e^A\right)^* = I_n$, showing that e^A is unitary. Since

$$\det\left(e^A\right) = e^{\mathrm{tr}(A)},$$

if A is skew Hermitian and has null trace, then $\det(e^A) = +1$.

For the surjectivity we will use Theorem 11.4.7 in Chapter 11 of Gallier [60]. First, assume that A is a unitary matrix. By Theorem 11.4.7, there is a unitary matrix U and a diagonal matrix D such that $A = UDU^*$. Furthermore, since A is unitary, the entries $\lambda_1, \ldots, \lambda_n$ in D (the eigenvalues of A) have absolute value $+1$. Thus, the entries in D are of the form $\cos\theta + i\sin\theta = e^{i\theta}$. Thus, we can assume that D is a diagonal matrix of the form

$$D = \begin{pmatrix} e^{i\theta_1} & & \cdots & \\ & e^{i\theta_2} & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & e^{i\theta_p} \end{pmatrix}.$$

If we let E be the diagonal matrix

$$E = \begin{pmatrix} i\theta_1 & & \cdots & \\ & i\theta_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & i\theta_p \end{pmatrix}$$

it is obvious that E is skew Hermitian and that

$$e^E = D.$$

Then, letting $B = UEU^*$, we have

$$e^B = A,$$

and it is immediately verified that B is skew Hermitian, since E is.

If A is a unitary matrix with determinant $+1$, since the eigenvalues of A are $e^{i\theta_1}, \ldots, e^{i\theta_p}$ and the determinant of A is the product

$$e^{i\theta_1} \cdots e^{i\theta_p} = e^{i(\theta_1 + \cdots + \theta_p)}$$

of these eigenvalues, we must have

$$\theta_1 + \cdots + \theta_p = 0,$$

and so, E is skew Hermitian and has zero trace. As above, letting

$$B = UEU^*,$$

we have

$$e^B = A,$$

where B is skew Hermitian and has null trace. □

We now extend the result of Section 1.3 to Hermitian matrices.

1.5 Hermitian Matrices, Hermitian Positive Definite Matrices, and the Exponential Map

Recall that a Hermitian matrix is called *positive* (or *positive semidefinite*) if its eigenvalues are all positive or null, and *positive definite* if its eigenvalues are all strictly positive. We denote the real vector space of Hermitian $n \times n$ matrices by $\mathbf{H}(n)$, the set of Hermitian positive matrices by $\mathbf{HP}(n)$, and the set of Hermitian positive definite matrices by $\mathbf{HPD}(n)$.

The next lemma shows that every Hermitian positive definite matrix A is of the form e^B for some unique Hermitian matrix B. As in the real case, the set of Hermitian matrices is a real vector space, but it is not a Lie algebra because the Lie bracket $[A, B]$ is not Hermitian unless A and B commute, and the set of Hermitian (positive) definite matrices is not a multiplicative group.

Lemma 1.10. *For every Hermitian matrix B, the matrix e^B is Hermitian positive definite. For every Hermitian positive definite matrix A, there is a unique Hermitian matrix B such that $A = e^B$.*

Proof. It is basically the same as the proof of Theorem 1.10, except that a Hermitian matrix can be written as $A = UDU^*$, where D is a real diagonal matrix and U is unitary instead of orthogonal. □

Lemma 1.10 can be reformulated as stating that the map $\exp\colon \mathbf{H}(n) \to \mathbf{HPD}(n)$ is a bijection. In fact, it can be shown that it is a homeomorphism. In the case of complex invertible matrices, the polar form theorem can be reformulated as stating that there is a bijection between the topological space $\mathbf{GL}(n, \mathbb{C})$ of complex $n \times n$ invertible matrices (also a group) and $\mathbf{U}(n) \times \mathbf{HPD}(n)$. As a corollary of the polar form theorem and Lemma 1.10, we

have the following result: For every complex invertible matrix A, there is a unique unitary matrix U and a unique Hermitian matrix S such that

$$A = U\, e^S.$$

Thus, we have a bijection between $\mathbf{GL}(n, \mathbb{C})$ and $\mathbf{U}(n) \times \mathbf{H}(n)$. But $\mathbf{H}(n)$ itself is isomorphic to $\mathbb{R}^{n^2}$, and so there is a bijection between $\mathbf{GL}(n, \mathbb{C})$ and $\mathbf{U}(n) \times \mathbb{R}^{n^2}$. It can also be shown that this bijection is a homeomorphism. This is an interesting fact. Indeed, this homeomorphism essentially reduces the study of the topology of $\mathbf{GL}(n, \mathbb{C})$ to the study of the topology of $\mathbf{U}(n)$. This is nice, since it can be shown that $\mathbf{U}(n)$ is compact (as a real manifold).

In the polar decomposition $A = Ue^S$, we have $|\det(U)| = 1$, since U is unitary, and $\mathrm{tr}(S)$ is real, since S is Hermitian (since it is the sum of the eigenvalues of S, which are real), so that $\det\left(e^S\right) > 0$. Thus, if $\det(A) = 1$, we must have $\det\left(e^S\right) = 1$, which implies that $S \in \mathbf{H}(n) \cap \mathfrak{sl}(n, \mathbb{C})$. Thus, we have a bijection between $\mathbf{SL}(n, \mathbb{C})$ and $\mathbf{SU}(n) \times (\mathbf{H}(n) \cap \mathfrak{sl}(n, \mathbb{C}))$.

In the next section we study the group $\mathbf{SE}(n)$ of affine maps induced by orthogonal transformations, also called rigid motions, and its Lie algebra. We will show that the exponential map is surjective. The groups $\mathbf{SE}(2)$ and $\mathbf{SE}(3)$ play play a fundamental role in robotics, dynamics, and motion planning.

1.6 The Lie Group $\mathbf{SE}(n)$ and the Lie Algebra $\mathfrak{se}(n)$

First, we review the usual way of representing affine maps of $\mathbb{R}^n$ in terms of $(n+1) \times (n+1)$ matrices.

Definition 1.3. The set of affine maps ρ of $\mathbb{R}^n$, defined such that

$$\rho(X) = RX + U,$$

where R is a rotation matrix ($R \in \mathbf{SO}(n)$) and U is some vector in $\mathbb{R}^n$, is a group under composition called the group of *direct affine isometries, or rigid motions*, denoted by $\mathbf{SE}(n)$.

Every rigid motion can be represented by the $(n+1) \times (n+1)$ matrix

$$\begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix}$$

in the sense that

$$\begin{pmatrix} \rho(X) \\ 1 \end{pmatrix} = \begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix}$$

iff

$$\rho(X) = RX + U.$$

Definition 1.4. The vector space of real $(n+1)\times(n+1)$ matrices of the form

$$A = \begin{pmatrix} \Omega & U \\ 0 & 0 \end{pmatrix},$$

where Ω is a skew symmetric matrix and U is a vector in $\mathbb{R}^n$, is denoted by $\mathfrak{se}(n)$.

Remark: The group $\mathbf{SE}(n)$ is a Lie group, and its Lie algebra turns out to be $\mathfrak{se}(n)$.

We will show that the exponential map $\exp\colon \mathfrak{se}(n) \to \mathbf{SE}(n)$ is surjective. First, we prove the following key lemma.

Lemma 1.11. *Given any $(n+1)\times(n+1)$ matrix of the form*

$$A = \begin{pmatrix} \Omega & U \\ 0 & 0 \end{pmatrix}$$

where Ω is any matrix and $U \in \mathbb{R}^n$,

$$A^k = \begin{pmatrix} \Omega^k & \Omega^{k-1}U \\ 0 & 0 \end{pmatrix},$$

where $\Omega^0 = I_n$. As a consequence,

$$e^A = \begin{pmatrix} e^\Omega & VU \\ 0 & 1 \end{pmatrix},$$

where

$$V = I_n + \sum_{k\geq 1} \frac{\Omega^k}{(k+1)!}.$$

Proof. A trivial induction on k shows that

$$A^k = \begin{pmatrix} \Omega^k & \Omega^{k-1}U \\ 0 & 0 \end{pmatrix}.$$

Then we have

$$\begin{aligned} e^A &= \sum_{k\geq 0} \frac{A^k}{k!}, \\ &= I_{n+1} + \sum_{k\geq 1} \frac{1}{k!} \begin{pmatrix} \Omega^k & \Omega^{k-1}U \\ 0 & 0 \end{pmatrix}, \\ &= \begin{pmatrix} I_n + \sum_{k\geq 0} \frac{\Omega^k}{k!} & \sum_{k\geq 1} \frac{\Omega^{k-1}}{k!}U \\ 0 & 1 \end{pmatrix}, \\ &= \begin{pmatrix} e^\Omega & VU \\ 0 & 1 \end{pmatrix}. \end{aligned}$$

□

We can now prove our main theorem. We will need to prove that V is invertible when Ω is a skew symmetric matrix. It would be tempting to write V as

$$V = \Omega^{-1}(e^{\Omega} - I).$$

Unfortunately, for odd n, a skew symmetric matrix of order n is not invertible! Thus, we have to find another way of proving that V is invertible. However, observe that we have the following useful fact:

$$V = I_n + \sum_{k \geq 1} \frac{\Omega^k}{(k+1)!} = \int_0^1 e^{\Omega t} dt.$$

This is what we will use in Theorem 1.12 to prove surjectivity.

Theorem 1.12. *The exponential map*

$$\exp \colon \mathfrak{se}(n) \to \mathbf{SE}(n)$$

is well-defined and surjective.

Proof. Since Ω is skew symmetric, e^{Ω} is a rotation matrix, and by Theorem 1.6, the exponential map

$$\exp \colon \mathfrak{so}(n) \to \mathbf{SO}(n)$$

is surjective. Thus, it remains to prove that for every rotation matrix R, there is some skew symmetric matrix Ω such that $R = e^{\Omega}$ and

$$V = I_n + \sum_{k \geq 1} \frac{\Omega^k}{(k+1)!}$$

is invertible. By Theorem 11.4.4 in Chapter 11 of Gallier [60], for every skew symmetric matrix Ω there is an orthogonal matrix P such that $\Omega = PDP^{\top}$, where D is a block diagonal matrix of the form

$$D = \begin{pmatrix} D_1 & & \cdots & \\ & D_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & D_p \end{pmatrix}$$

such that each block D_i is either 0 or a two-dimensional matrix of the form

$$D_i = \begin{pmatrix} 0 & -\theta_i \\ \theta_i & 0 \end{pmatrix}$$

where $\theta_i \in \mathbb{R}$, with $\theta_i > 0$. Actually, we can assume that $\theta_i \neq k2\pi$ for all $k \in \mathbb{Z}$, since when $\theta_i = k2\pi$ we have $e^{D_i} = I_2$, and D_i can be replaced by two one-dimensional blocks each

consisting of a single zero. To compute V, since $\Omega = PD\,P^\top = PDP^{-1}$, observe that

$$\begin{aligned} V &= I_n + \sum_{k\geq 1} \frac{\Omega^k}{(k+1)!} \\ &= I_n + \sum_{k\geq 1} \frac{PD^kP^{-1}}{(k+1)!} \\ &= P\left(I_n + \sum_{k\geq 1} \frac{D^k}{(k+1)!}\right)P^{-1} \\ &= PWP^{-1}, \end{aligned}$$

where

$$W = I_n + \sum_{k\geq 1} \frac{D^k}{(k+1)!}.$$

We can compute

$$W = I_n + \sum_{k\geq 1} \frac{D^k}{(k+1)!} = \int_0^1 e^{Dt}dt,$$

by computing

$$W = \begin{pmatrix} W_1 & & \cdots & \\ & W_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & W_p \end{pmatrix}$$

by blocks. Since

$$e^{D_i} = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix}$$

when D_i is a 2×2 skew symmetric matrix and $W_i = \int_0^1 e^{D_it}dt$, we get

$$W_i = \begin{pmatrix} \int_0^1 \cos(\theta_i t)dt & \int_0^1 -\sin(\theta_i t)dt \\ \int_0^1 \sin(\theta_i t)dt & \int_0^1 \cos(\theta_i t)dt \end{pmatrix} = \frac{1}{\theta_i}\begin{pmatrix} \sin(\theta_i t)\,|_0^1 & \cos(\theta_i t)\,|_0^1 \\ -\cos(\theta_i t)\,|_0^1 & \sin(\theta_i t)\,|_0^1 \end{pmatrix},$$

that is,

$$W_i = \frac{1}{\theta_i}\begin{pmatrix} \sin\theta_i & -(1-\cos\theta_i) \\ 1-\cos\theta_i & \sin\theta_i \end{pmatrix},$$

and $W_i = 1$ when $D_i = 0$. Now, in the first case, the determinant is

$$\frac{1}{\theta_i^2}\left((\sin\theta_i)^2 + (1-\cos\theta_i)^2\right) = \frac{2}{\theta_i^2}(1-\cos\theta_i),$$

which is nonzero, since $\theta_i \neq k2\pi$ for all $k\in\mathbb{Z}$. Thus, each W_i is invertible, and so is W, and thus, $V = PWP^{-1}$ is invertible. □

In the case $n = 3$, given a skew symmetric matrix

$$\Omega = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix},$$

letting $\theta = \sqrt{a^2 + b^2 + c^2}$, it it easy to prove that if $\theta = 0$, then

$$e^A = \begin{pmatrix} I_3 & U \\ 0 & 1 \end{pmatrix},$$

and that if $\theta \neq 0$ (using the fact that $\Omega^3 = -\theta^2\Omega$), then

$$e^\Omega = I_3 + \frac{\sin\theta}{\theta}\Omega + \frac{(1 - \cos\theta)}{\theta^2}\Omega^2$$

and

$$V = I_3 + \frac{(1 - \cos\theta)}{\theta^2}\Omega + \frac{(\theta - \sin\theta)}{\theta^3}\Omega^2.$$

Our next goal is to define *embedded submanifolds* and (linear) Lie groups. Before doing this, we believe that some readers might appreciate a review of the notion of the *derivative* of a function between two normed vector spaces.

1.7 The Derivative of a Function Between Normed Vector Spaces, a Review

In this brief section, we review some basic notions of differential calculus, in particular, the *derivative* of a function, $f\colon E \to F$, where E and F are normed vector spaces. In most cases, $E = \mathbb{R}^n$ and $F = \mathbb{R}^m$. However, if we need to deal with infinite dimensional manifolds, then it is necessary to allow E and F to be infinite dimensional. This section can be omitted by readers already familiar with this standard material. We omit all proofs and refer the reader to standard analysis textbooks such as Lang [95, 94], Munkres [117], Choquet-Bruhat [37] or Schwartz [136], for a complete exposition.

Let E and F be two *normed vector spaces*, let $A \subseteq E$ be some open subset of A, and let $a \in A$ be some element of A. Even though a is a vector, we may also call it a point.

The idea behind the derivative of the function f at a is that it is a *linear approximation* of f in a small open set around a. The difficulty is to make sense of the quotient

$$\frac{f(a + h) - f(a)}{h}$$

where h is a vector. We circumvent this difficulty in two stages.

A first possibility is to consider the *directional derivative* with respect to a vector $u \neq 0$ in E.

We can consider the vector $f(a+tu) - f(a)$, where $t \in \mathbb{R}$ (or $t \in \mathbb{C}$). Now,

$$\frac{f(a+tu) - f(a)}{t}$$

makes sense.

The idea is that in E, the points of the form $a+tu$ for t in some small interval $[-\epsilon, +\epsilon]$ in $\mathbb{R}$ form a line segment $[r, s]$ in A containing a, and that the image of this line segment defines a small curve segment on $f(A)$. This curve segment is defined by the map $t \mapsto f(a+tu)$, from $[r, s]$ to F, and the directional derivative $\mathrm{D}_u f(a)$ defines the direction of the tangent line at a to this curve.

Definition 1.5. Let E and F be two normed spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, for any $u \neq 0$ in E, the *directional derivative of f at a w.r.t. the vector u*, denoted by $\mathrm{D}_u f(a)$, is the limit (if it exists)

$$\lim_{t \to 0,\, t \in U} \frac{f(a+tu) - f(a)}{t},$$

where $U = \{t \in \mathbb{R} \mid a + tu \in A,\, t \neq 0\}$ (or $U = \{t \in \mathbb{C} \mid a + tu \in A,\, t \neq 0\}$).

Since the map $t \mapsto a + tu$ is continuous, and since $A - \{a\}$ is open, the inverse image U of $A - \{a\}$ under the above map is open, and the definition of the limit in Definition 1.5 makes sense.

Remark: Since the notion of limit is purely topological, the existence and value of a directional derivative is independent of the choice of norms in E and F, as long as they are equivalent norms.

The directional derivative is sometimes called the *Gâteaux derivative*.

In the special case where $E = \mathbb{R}$, $F = \mathbb{R}$ and we let $u = 1$ (i.e., the real number 1, viewed as a vector), it is immediately verified that $\mathrm{D}_1 f(a) = f'(a)$. When $E = \mathbb{R}$ (or $E = \mathbb{C}$) and F is any normed vector space, the derivative $\mathrm{D}_1 f(a)$, also denoted by $f'(a)$, provides a suitable generalization of the notion of derivative.

However, when E has dimension ≥ 2, directional derivatives present a serious problem, which is that their definition is not sufficiently uniform. Indeed, there is no reason to believe that the directional derivatives w.r.t. all nonzero vectors u share something in common. As a consequence, a function can have all directional derivatives at a, and yet not be continuous at a. Two functions may have all directional derivatives in some open sets, and yet their composition may not. Thus, we introduce a more uniform notion.

Definition 1.6. Let E and F be two normed spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, we say that f is *differentiable at* $a \in A$ if there is a linear continuous map, $L\colon E \to F$, and a function, $\epsilon(h)$, such that

$$f(a+h) = f(a) + L(h) + \epsilon(h)\|h\|$$

for every $a + h \in A$, where

$$\lim_{h \to 0,\, h \in U} \epsilon(h) = 0,$$

with $U = \{h \in E \mid a + h \in A,\, h \neq 0\}$. The linear map L is denoted by $\mathrm{D}f(a)$, or $\mathrm{D}f_a$, or $df(a)$, or df_a, or $f'(a)$, and it is called the *Fréchet derivative*, or *total derivative*, or *derivative*, or *total differential*, or *differential*, of f at a.

Since the map $h \mapsto a+h$ from E to E is continuous, and since A is open in E, the inverse image U of $A - \{a\}$ under the above map is open in E, and it makes sense to say that

$$\lim_{h \to 0,\, h \in U} \epsilon(h) = 0.$$

Note that for every $h \in U$, since $h \neq 0$, $\epsilon(h)$ is uniquely determined since

$$\epsilon(h) = \frac{f(a+h) - f(a) - L(h)}{\|h\|},$$

and the value $\epsilon(0)$ plays absolutely no role in this definition. It does no harm to assume that $\epsilon(0) = 0$, and we will assume this from now on.

Remark: Since the notion of limit is purely topological, the existence and value of a derivative is independent of the choice of norms in E and F, as long as they are equivalent norms.

Note that *the continuous linear map L is unique*, if it exists.

The following proposition shows that our new definition is consistent with the definition of the directional derivative and that *the continuous linear map L is unique*, if it exists.

Proposition 1.13. *Let E and F be two normed spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, if $\mathrm{D}f(a)$ is defined, then f is continuous at a and f has a directional derivative $\mathrm{D}_u f(a)$ for every $u \neq 0$ in E. Furthermore,*

$$\mathrm{D}_u f(a) = \mathrm{D}f(a)(u)$$

and thus, $\mathrm{D}f(a)$ is uniquely defined.

Proof. If $L = \mathrm{D}f(a)$ exists, then for any nonzero vector $u \in E$, because A is open, for any $t \in \mathbb{R} - \{0\}$ (or $t \in \mathbb{C} - \{0\}$) small enough, $a + tu \in A$, so

$$\begin{aligned} f(a+tu) &= f(a) + L(tu) + \epsilon(tu)\|tu\| \\ &= f(a) + tL(u) + |t|\epsilon(tu)\|u\| \end{aligned}$$

which implies that

$$L(u) = \frac{f(a+tu) - f(a)}{t} - \frac{|t|}{t}\epsilon(tu)\|u\|,$$

and since $\lim_{t \mapsto 0} \epsilon(tu) = 0$, we deduce that

$$L(u) = \mathrm{D}f(a)(u) = \mathrm{D}_u f(a).$$

Because

$$f(a+h) = f(a) + L(h) + \epsilon(h)\|h\|$$

for all h such that $\|h\|$ is small enough, L is continuous, and $\lim_{h \mapsto 0} \epsilon(h)\|h\| = 0$, we have $\lim_{h \mapsto 0} f(a+h) = f(a)$, that is, f is continuous at a. □

Observe that the uniqueness of $\mathrm{D}f(a)$ follows from Proposition 1.13. Also, when E is of finite dimension, it is easily shown that every linear map is continuous and this assumption is then redundant.

If $\mathrm{D}f(a)$ exists for every $a \in A$, we get a map $\mathrm{D}f \colon A \to \mathcal{L}(E; F)$, called the *derivative of f on A*, and also denoted by *df*. Here, $\mathcal{L}(E; F)$ denotes the vector space of continuous linear maps from E to F.

When E is of finite dimension n, for any basis, $(u_1, \ldots, u_n)$, of E, we can define the directional derivatives with respect to the vectors in the basis $(u_1, \ldots, u_n)$ (actually, we can also do it for an infinite basis). This way, we obtain the definition of partial derivatives, as follows:

Definition 1.7. For any two normed spaces E and F, if E is of finite dimension n, for every basis $(u_1, \ldots, u_n)$ for E, for every $a \in E$, for every function $f \colon E \to F$, the directional derivatives $\mathrm{D}_{u_j} f(a)$ (if they exist) are called the *partial derivatives of f with respect to the basis* $(u_1, \ldots, u_n)$. The partial derivative $\mathrm{D}_{u_j} f(a)$ is also denoted by $\partial_j f(a)$, or $\dfrac{\partial f}{\partial x_j}(a)$.

The notation $\dfrac{\partial f}{\partial x_j}(a)$ for a partial derivative, although customary and going back to Leibniz, is a "logical obscenity." Indeed, the variable x_j really has nothing to do with the formal definition. This is just another of these situations where tradition is just too hard to overthrow!

We now consider a number of standard results about derivatives. A function $f \colon E \to F$ is said to be *affine* if there is some linear map $\overrightarrow{f} : E \to F$ and some fixed vector $c \in F$, such that

$$f(u) = \overrightarrow{f}(u) + c$$

for all $u \in E$. We call f the *linear map associated with f*.

Proposition 1.14. *Given two normed spaces E and F, if $f\colon E \to F$ is a constant function, then $\mathrm{D}f(a) = 0$, for every $a \in E$. If $f\colon E \to F$ is a continuous affine map, then $\mathrm{D}f(a) = \overrightarrow{f}$, for every $a \in E$, where $\overrightarrow{f}$ denotes the linear map associated with f.*

Proposition 1.15. *Given a normed space E and a normed vector space F, for any two functions $f, g\colon E \to F$, for every $a \in E$, if $\mathrm{D}f(a)$ and $\mathrm{D}g(a)$ exist, then $\mathrm{D}(f+g)(a)$ and $\mathrm{D}(\lambda f)(a)$ exist, and*

$$\mathrm{D}(f+g)(a) = \mathrm{D}f(a) + \mathrm{D}g(a),$$
$$\mathrm{D}(\lambda f)(a) = \lambda \mathrm{D}f(a).$$

Proposition 1.16. *Given three normed vector spaces E_1, E_2, and F, for any continuous bilinear map $f\colon E_1 \times E_2 \to F$, for every $(a, b) \in E_1 \times E_2$, $\mathrm{D}f(a, b)$ exists, and for every $u \in E_1$ and $v \in E_2$,*

$$\mathrm{D}f(a, b)(u, v) = f(u, b) + f(a, v).$$

We now state the very useful *chain rule*.

Theorem 1.17. *Given three normed spaces E, F, and G, let A be an open set in E, and let B an open set in F. For any functions $f\colon A \to F$ and $g\colon B \to G$, such that $f(A) \subseteq B$, for any $a \in A$, if $\mathrm{D}f(a)$ exists and $\mathrm{D}g(f(a))$ exists, then $\mathrm{D}(g \circ f)(a)$ exists, and*

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

Theorem 1.17 has many interesting consequences. We mention one corollary.

Proposition 1.18. *Given two normed spaces E and F, let A be some open subset in E, let B be some open subset in F, let $f\colon A \to B$ be a bijection from A to B, and assume that $\mathrm{D}f$ exists on A and that $\mathrm{D}f^{-1}$ exists on B. Then, for every $a \in A$,*

$$\mathrm{D}f^{-1}(f(a)) = (\mathrm{D}f(a))^{-1}.$$

Proposition 1.18 has the remarkable consequence that the two vector spaces E and F have the same dimension. In other words, a local property, the existence of a bijection f between an open set A of E and an open set B of F, such that f is differentiable on A and f^{-1} is differentiable on B, implies a global property, that the two vector spaces E and F have the same dimension.

If both E and F are of finite dimension, for any basis $(u_1, \ldots, u_n)$ of E and any basis $(v_1, \ldots, v_m)$ of F, every function $f\colon E \to F$ is determined by m functions $f_i\colon E \to \mathbb{R}$ (or $f_i\colon E \to \mathbb{C}$), where

$$f(x) = f_1(x)v_1 + \cdots + f_m(x)v_m,$$

for every $x \in E$. Then, we get

$$\mathrm{D}f(a)(u_j) = \mathrm{D}f_1(a)(u_j)v_1 + \cdots + \mathrm{D}f_i(a)(u_j)v_i + \cdots + \mathrm{D}f_m(a)(u_j)v_m,$$

that is,

$$\mathrm{D}f(a)(u_j) = \partial_j f_1(a)v_1 + \cdots + \partial_j f_i(a)v_i + \cdots + \partial_j f_m(a)v_m.$$

Since the j-th column of the $m \times n$-matrix representing $\mathrm{D}f(a)$ w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ is equal to the components of the vector $\mathrm{D}f(a)(u_j)$ over the basis $(v_1, \ldots, v_m)$, the linear map $\mathrm{D}f(a)$ is determined by the $m \times n$-matrix $J(f)(a) = (\partial_j f_i(a))$, or $J(f)(a) = \left(\dfrac{\partial f_i}{\partial x_j}(a)\right)$:

$$J(f)(a) = \begin{pmatrix} \partial_1 f_1(a) & \partial_2 f_1(a) & \ldots & \partial_n f_1(a) \\ \partial_1 f_2(a) & \partial_2 f_2(a) & \ldots & \partial_n f_2(a) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 f_m(a) & \partial_2 f_m(a) & \ldots & \partial_n f_m(a) \end{pmatrix}$$

or

$$J(f)(a) = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1}(a) & \dfrac{\partial f_1}{\partial x_2}(a) & \ldots & \dfrac{\partial f_1}{\partial x_n}(a) \\ \dfrac{\partial f_2}{\partial x_1}(a) & \dfrac{\partial f_2}{\partial x_2}(a) & \ldots & \dfrac{\partial f_2}{\partial x_n}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1}(a) & \dfrac{\partial f_m}{\partial x_2}(a) & \ldots & \dfrac{\partial f_m}{\partial x_n}(a) \end{pmatrix}.$$

This matrix is called the *Jacobian matrix* of $\mathrm{D}f$ at a. When $m = n$, the determinant, $\det(J(f)(a))$, of $J(f)(a)$ is called the *Jacobian* of $\mathrm{D}f(a)$.

We know that this determinant only depends on $\mathrm{D}f(a)$, and not on specific bases. However, partial derivatives give a means for computing it.

When $E = \mathbb{R}^n$ and $F = \mathbb{R}^m$, for any function $f\colon \mathbb{R}^n \to \mathbb{R}^m$, it is easy to compute the partial derivatives $\dfrac{\partial f_i}{\partial x_j}(a)$. We simply treat the function $f_i\colon \mathbb{R}^n \to \mathbb{R}$ as a function of its j-th argument, leaving the others fixed, and compute the derivative as the usual derivative.

Example 1.1. For example, consider the function $f\colon \mathbb{R}^2 \to \mathbb{R}^2$, defined by

$$f(r, \theta) = (r\cos\theta, r\sin\theta).$$

Then, we have

$$J(f)(r, \theta) = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

and the Jacobian (determinant) has value $\det(J(f)(r, \theta)) = r$.

In the case where $E = \mathbb{R}$ (or $E = \mathbb{C}$), for any function $f\colon \mathbb{R} \to F$ (or $f\colon \mathbb{C} \to F$), the Jacobian matrix of $\mathrm{D}f(a)$ is a column vector. In fact, this column vector is just $\mathrm{D}_1 f(a)$. Then, for every $\lambda \in \mathbb{R}$ (or $\lambda \in \mathbb{C}$), $\mathrm{D}f(a)(\lambda) = \lambda \mathrm{D}_1 f(a)$. This case is sufficiently important to warrant a definition.

Definition 1.8. Given a function $f\colon \mathbb{R} \to F$ (or $f\colon \mathbb{C} \to F$), where F is a normed space, the vector

$$\mathrm{D}f(a)(1) = \mathrm{D}_1 f(a)$$

is called the *vector derivative or velocity vector (in the real case)* at a. We usually identify $\mathrm{D}f(a)$ with its Jacobian matrix $\mathrm{D}_1 f(a)$, which is the column vector corresponding to $\mathrm{D}_1 f(a)$. By abuse of notation, we also let $\mathrm{D}f(a)$ denote the vector $\mathrm{D}f(a)(1) = \mathrm{D}_1 f(a)$.

When $E = \mathbb{R}$, the physical interpretation is that f defines a (parametric) curve that is the trajectory of some particle moving in $\mathbb{R}^m$ as a function of time, and the vector $\mathrm{D}_1 f(a)$ is the *velocity* of the moving particle $f(t)$ at $t = a$.

Example 1.2.

1. When $A = (0, 1)$, and $F = \mathbb{R}^3$, a function $f\colon (0, 1) \to \mathbb{R}^3$ defines a (parametric) curve in $\mathbb{R}^3$. If $f = (f_1, f_2, f_3)$, its Jacobian matrix at $a \in \mathbb{R}$ is
$$J(f)(a) = \begin{pmatrix} \dfrac{\partial f_1}{\partial t}(a) \\ \dfrac{\partial f_2}{\partial t}(a) \\ \dfrac{\partial f_3}{\partial t}(a) \end{pmatrix}$$

2. When $E = \mathbb{R}^2$, and $F = \mathbb{R}^3$, a function $\varphi\colon \mathbb{R}^2 \to \mathbb{R}^3$ defines a parametric surface. Letting $\varphi = (f, g, h)$, its Jacobian matrix at $a \in \mathbb{R}^2$ is
$$J(\varphi)(a) = \begin{pmatrix} \dfrac{\partial f}{\partial u}(a) & \dfrac{\partial f}{\partial v}(a) \\ \dfrac{\partial g}{\partial u}(a) & \dfrac{\partial g}{\partial v}(a) \\ \dfrac{\partial h}{\partial u}(a) & \dfrac{\partial h}{\partial v}(a) \end{pmatrix}$$

3. When $E = \mathbb{R}^3$, and $F = \mathbb{R}$, for a function $f\colon \mathbb{R}^3 \to \mathbb{R}$, the Jacobian matrix at $a \in \mathbb{R}^3$ is
$$J(f)(a) = \begin{pmatrix} \dfrac{\partial f}{\partial x}(a) & \dfrac{\partial f}{\partial y}(a) & \dfrac{\partial f}{\partial z}(a) \end{pmatrix}.$$

More generally, when $f\colon \mathbb{R}^n \to \mathbb{R}$, the Jacobian matrix at $a \in \mathbb{R}^n$ is the row vector

$$J(f)(a) = \left(\frac{\partial f}{\partial x_1}(a) \ \cdots \ \frac{\partial f}{\partial x_n}(a)\right).$$

Its transpose is a column vector called the *gradient* of f at a, denoted by $\mathrm{grad} f(a)$ or $\nabla f(a)$. Then, given any $v \in \mathbb{R}^n$, note that

$$\mathrm{D}f(a)(v) = \frac{\partial f}{\partial x_1}(a)\, v_1 + \cdots + \frac{\partial f}{\partial x_n}(a)\, v_n = \mathrm{grad} f(a) \cdot v,$$

the scalar product of $\mathrm{grad} f(a)$ and v.

When E, F, and G have finite dimensions, $(u_1, \ldots, u_p)$ is a basis for E, $(v_1, \ldots, v_n)$ is a basis for F, and $(w_1, \ldots, w_m)$ is a basis for G, if A is an open subset of E, B is an open subset of F, for any functions $f\colon A \to F$ and $g\colon B \to G$, such that $f(A) \subseteq B$, for any $a \in A$, letting $b = f(a)$, and $h = g \circ f$, if $\mathrm{D}f(a)$ exists and $\mathrm{D}g(b)$ exists, by Theorem 1.17, the Jacobian matrix $J(h)(a) = J(g \circ f)(a)$ w.r.t. the bases $(u_1, \ldots, u_p)$ and $(w_1, \ldots, w_m)$ is the product of the Jacobian matrices $J(g)(b)$ w.r.t. the bases $(v_1, \ldots, v_n)$ and $(w_1, \ldots, w_m)$, and $J(f)(a)$ w.r.t. the bases $(u_1, \ldots, u_p)$ and $(v_1, \ldots, v_n)$:

$$J(h)(a) = \begin{pmatrix} \dfrac{\partial g_1}{\partial y_1}(b) & \dfrac{\partial g_1}{\partial y_2}(b) & \cdots & \dfrac{\partial g_1}{\partial y_n}(b) \\ \dfrac{\partial g_2}{\partial y_1}(b) & \dfrac{\partial g_2}{\partial y_2}(b) & \cdots & \dfrac{\partial g_2}{\partial y_n}(b) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial g_m}{\partial y_1}(b) & \dfrac{\partial g_m}{\partial y_2}(b) & \cdots & \dfrac{\partial g_m}{\partial y_n}(b) \end{pmatrix} \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1}(a) & \dfrac{\partial f_1}{\partial x_2}(a) & \cdots & \dfrac{\partial f_1}{\partial x_p}(a) \\ \dfrac{\partial f_2}{\partial x_1}(a) & \dfrac{\partial f_2}{\partial x_2}(a) & \cdots & \dfrac{\partial f_2}{\partial x_p}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_n}{\partial x_1}(a) & \dfrac{\partial f_n}{\partial x_2}(a) & \cdots & \dfrac{\partial f_n}{\partial x_p}(a) \end{pmatrix}.$$

Thus, we have the familiar formula

$$\frac{\partial h_i}{\partial x_j}(a) = \sum_{k=1}^{k=n} \frac{\partial g_i}{\partial y_k}(b) \frac{\partial f_k}{\partial x_j}(a).$$

Given two normed spaces E and F of finite dimension, given an open subset A of E, if a function $f\colon A \to F$ is differentiable at $a \in A$, then its Jacobian matrix is well defined.

One should be warned that the converse is false. There are functions such that all the partial derivatives exist at some $a \in A$, but yet, the function is not differentiable at a, and not even continuous at a.

However, there are sufficient conditions on the partial derivatives for $\mathrm{D}f(a)$ to exist, namely, continuity of the partial derivatives. If f is differentiable on A, then f defines a function $\mathrm{D}f\colon A \to \mathcal{L}(E; F)$. It turns out that the continuity of the partial derivatives on A is a necessary and sufficient condition for $\mathrm{D}f$ to exist and to be continuous on A.

Theorem 1.19. *Given two normed spaces E and F, where E is of finite dimension n and where $(u_1, \ldots, u_n)$ is a basis of E, given any open subset A of E, given any function $f\colon A \to F$, the derivative $\mathrm{D}f\colon A \to \mathcal{L}(E;F)$ is defined and continuous on A iff every partial derivative $\partial_j f$ (or $\dfrac{\partial f}{\partial x_j}$) is defined and continuous on A, for all j, $1 \leq j \leq n$. As a corollary, if F is of finite dimension m, and $(v_1, \ldots, v_m)$ is a basis of F, the derivative $\mathrm{D}f\colon A \to \mathcal{L}(E;F)$ is defined and continuous on A iff every partial derivative $\partial_j f_i$ $\left(\text{or } \dfrac{\partial f_i}{\partial x_j}\right)$ is defined and continuous on A, for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$.*

Definition 1.9. Given two normed spaces E and F, and an open subset A of E, we say that a function $f\colon A \to F$ is a *C^0-function on A* if f is continuous on A. We say that $f\colon A \to F$ is a *C^1-function on A* if $\mathrm{D}f$ exists and is continuous on A.

Let E and F be two normed spaces, let $U \subseteq E$ be an open subset of E and let $f\colon E \to F$ be a function such that $Df(a)$ exists for all $a \in U$. If $Df(a)$ is injective for all $a \in U$, we say that f is an *immersion* (on U) and if $Df(a)$ is surjective for all $a \in U$, we say that f is a *submersion* (on U).

When E and F are finite dimensional with $\dim(E) = n$ and $\dim(F) = m$, if $m \geq n$, then f is an immersion iff the Jacobian matrix, $J(f)(a)$, has full rank (n) for all $a \in E$ and if $n \geq m$, then f is a submersion iff the Jacobian matrix, $J(f)(a)$, has full rank (m) for all $a \in E$.

A very important theorem is the inverse function theorem. In order for this theorem to hold for infinite dimensional spaces, it is necessary to assume that our normed spaces are complete.

Given a normed vector space, E, we say that a sequence, $(u_n)_n$, with $u_n \in E$, is a *Cauchy sequence* iff for every $\epsilon > 0$, there is some $N > 0$ so that for all $m, n \geq N$,

$$\|u_n - u_m\| < \epsilon.$$

A normed vector space, E, is *complete* iff every Cauchy sequence converges. A complete normed vector space is also called a *Banach space*, after Stefan Banach (1892-1945).

Fortunately, $\mathbb{R}, \mathbb{C}$, and every finite dimensional (real or complex) normed vector space is complete. A real (resp. complex) vector space, E, is a real (resp. complex) *Hilbert space* if it is complete as a normed space with the norm $\|u\| = \sqrt{\langle u, u\rangle}$ induced by its Euclidean (resp. Hermitian) inner product (of course, positive, definite).

Definition 1.10. Given two topological spaces E and F and an open subset A of E, we say that a function $f\colon A \to F$ is a *local homeomorphism from A to F* if for every $a \in A$, there is an open set $U \subseteq A$ containing a and an open set V containing $f(a)$ such that f is a homeomorphism from U to $V = f(U)$. If B is an open subset of F, we say that $f\colon A \to F$ is a *(global) homeomorphism from A to B* if f is a homeomorphism from A to $B = f(A)$.

If E and F are normed spaces, we say that $f\colon A \to F$ is a *local diffeomorphism from A to F* if for every $a \in A$, there is an open set $U \subseteq A$ containing a and an open set V containing $f(a)$ such that f is a bijection from U to V, f is a C^1-function on U, and f^{-1} is a C^1-function on $V = f(U)$. We say that $f\colon A \to F$ is a *(global) diffeomorphism from A to B* if f is a homeomorphism from A to $B = f(A)$, f is a C^1-function on A, and f^{-1} is a C^1-function on B.

Note that a local diffeomorphism is a local homeomorphism. Also, as a consequence of Proposition 1.18, if f is a diffeomorphism on A, then $\mathrm{D}f(a)$ is a bijection for every $a \in A$.

Theorem 1.20. *(Inverse Function Theorem) Let E and F be complete normed spaces, let A be an open subset of E, and let $f\colon A \to F$ be a C^1-function on A. The following properties hold:*

(1) For every $a \in A$, if $\mathrm{D}f(a)$ is invertible, then there exist some open subset $U \subseteq A$ containing a, and some open subset V of F containing $f(a)$, such that f is a diffeomorphism from U to $V = f(U)$. Furthermore,

$$\mathrm{D}f^{-1}(f(a)) = (\mathrm{D}f(a))^{-1}.$$

For every neighborhood N of a, the image $f(N)$ of N is a neighborhood of $f(a)$, and for every open ball $U \subseteq A$ of center a, the image $f(U)$ of U contains some open ball of center $f(a)$.

(2) If $\mathrm{D}f(a)$ is invertible for every $a \in A$, then $B = f(A)$ is an open subset of F, and f is a local diffeomorphism from A to B. Furthermore, if f is injective, then f is a diffeomorphism from A to B.

Part (1) of Theorem 1.20 is often referred to as the "(local) inverse function theorem." It plays an important role in the study of manifolds and (ordinary) differential equations.

If E and F are both of finite dimension, the case where $\mathrm{D}f(a)$ is just injective or just surjective is also important for defining manifolds, using implicit definitions.

1.8 Manifolds, Lie Groups and Lie Algebras

In this section we define precisely manifolds, Lie groups and Lie algebras. One of the reasons that Lie groups are nice is that they have a differential structure, which means that the notion of tangent space makes sense at any point of the group. Furthermore, the tangent space at the identity happens to have some algebraic structure, that of a Lie algebra. Roughly, the tangent space at the identity provides a "linearization" of the Lie group, and it turns out that many properties of a Lie group are reflected in its Lie algebra, and that the loss of information is not too severe. The challenge that we are facing is that unless our readers are

already familiar with manifolds, the amount of basic differential geometry required to define Lie groups and Lie algebras in full generality is overwhelming.

Fortunately, most of the Lie groups that we will consider are subspaces of $\mathbb{R}^N$ for some sufficiently large N. In fact, most of them are isomorphic to subgroups of $\mathbf{GL}(N, \mathbb{R})$ for some suitable N, even $\mathbf{SE}(n)$, which is isomorphic to a subgroup of $\mathbf{SL}(n+1)$. Such groups are called *linear Lie groups* (or *matrix groups*). Since these groups are subspaces of $\mathbb{R}^N$, in a first stage, we do not need the definition of an abstract manifold. We just have to define embedded submanifolds (also called submanifolds) of $\mathbb{R}^N$ (in the case of $\mathbf{GL}(n, \mathbb{R})$, $N = n^2$). This is the path that we will follow. The general definition of manifold will be given in Chapter 3.

In general, the difficult part in proving that a subgroup of $\mathbf{GL}(n, \mathbb{R})$ is a Lie group is to prove that it is a manifold. Fortunately, there is a characterization of the linear groups that obviates much of the work. This characterization rests on two theorems. First, a Lie subgroup H of a Lie group G (where H is an embedded submanifold of G) is closed in G (see Warner [148], Chapter 3, Theorem 3.21, page 97). Second, a theorem of Von Neumann and Cartan asserts that a closed subgroup of $\mathbf{GL}(n, \mathbb{R})$ is an embedded submanifold, and thus, a Lie group (see Warner [148], Chapter 3, Theorem 3.42, page 110). Thus, a linear Lie group is a closed subgroup of $\mathbf{GL}(n, \mathbb{R})$.

Since our Lie groups are subgroups (or isomorphic to subgroups) of $\mathbf{GL}(n, \mathbb{R})$ for some suitable n, it is easy to define the Lie algebra of a Lie group using curves. This approach to define the Lie algebra of a matrix group is followed by a number of authors, such as Curtis [39]. However, Curtis is rather cavalier, since he does not explain why the required curves actually exist, and thus, according to his definition, Lie algebras could be the trivial vector space! Although we will not prove the theorem of Von Neumann and Cartan, we feel that it is important to make clear why the definitions make sense, i.e., why we are not dealing with trivial objects.

A small annoying technical problem will arise in our approach, the problem with discrete subgroups. If A is a subset of $\mathbb{R}^N$, recall that A inherits a topology from $\mathbb{R}^N$ called the *subspace topology*, and defined such that a subset V of A is open if

$$V = A \cap U$$

for some open subset U of $\mathbb{R}^N$. A point $a \in A$ is said to be *isolated* if there is there is some open subset U of $\mathbb{R}^N$ such that

$$\{a\} = A \cap U,$$

in other words, if $\{a\}$ is an open set in A.

The group $\mathbf{GL}(n, \mathbb{R})$ of real invertible $n \times n$ matrices can be viewed as a subset of $\mathbb{R}^{n^2}$, and as such, it is a topological space under the subspace topology (in fact, a dense open subset of $\mathbb{R}^{n^2}$). One can easily check that multiplication and the inverse operation are continuous, and in fact smooth (i.e., C^∞-continuously differentiable). This makes $\mathbf{GL}(n, \mathbb{R})$ a *topological*

group. Any subgroup G of $\mathbf{GL}(n, \mathbb{R})$ is also a topological space under the subspace topology. A subgroup G is called a *discrete subgroup* if it has some isolated point. This turns out to be equivalent to the fact that every point of G is isolated, and thus, G has the discrete topology (every subset of G is open). Now, because $\mathbf{GL}(n, \mathbb{R})$ is Hausdorff, it can be shown that every discrete subgroup of $\mathbf{GL}(n, \mathbb{R})$ is closed (which means that its complement is open). Thus, discrete subgroups of $\mathbf{GL}(n, \mathbb{R})$ are Lie groups! But these are not very interesting Lie groups, and so we will consider only closed subgroups of $\mathbf{GL}(n, \mathbb{R})$ that are not discrete.

Let us now review the definition of an embedded submanifold. For simplicity, we restrict our attention to smooth manifolds. For detailed presentations, see DoCarmo [50, 51], Milnor [109], Marsden and Ratiu [103], Berger and Gostiaux [17], or Warner [148]. For the sake of brevity, we use the terminology *manifold* (but other authors would say *embedded submanifolds*, or something like that).

The intuition behind the notion of a smooth manifold in $\mathbb{R}^N$ is that a subspace M is a manifold of dimension m if every point $p \in M$ is contained in some open subset set U of M (in the subspace topology) that can be parametrized by some function $\varphi\colon \Omega \to U$ from some open subset Ω of the origin in $\mathbb{R}^m$, and that φ has some nice properties that allow the definition of smooth functions on M and of the tangent space at p. For this, φ has to be at least a homeomorphism, but more is needed: φ must be smooth, and the derivative $\varphi'(0_m)$ at the origin must be injective (letting $0_m = \underbrace{(0, \ldots, 0)}_{m}$).

Definition 1.11. Given any integers N, m, with $N \geq m \geq 1$, an *m-dimensional smooth manifold in $\mathbb{R}^N$, for short a manifold*, is a nonempty subset M of $\mathbb{R}^N$ such that for every point $p \in M$ there are two open subsets $\Omega \subseteq \mathbb{R}^m$ and $U \subseteq M$, with $p \in U$, and a smooth function $\varphi\colon \Omega \to \mathbb{R}^N$ such that φ is a homeomorphism between Ω and $U = \varphi(\Omega)$, and $\varphi'(t_0)$ is injective, where $t_0 = \varphi^{-1}(p)$. The function $\varphi\colon \Omega \to U$ is called a *(local) parametrization of M at p*. If $0_m \in \Omega$ and $\varphi(0_m) = p$, we say that $\varphi\colon \Omega \to U$ is *centered at p*.

Recall that $M \subseteq \mathbb{R}^N$ is a topological space under the subspace topology, and U is some open subset of M in the subspace topology, which means that $U = M \cap W$ for some open subset W of $\mathbb{R}^N$. Since $\varphi\colon \Omega \to U$ is a homeomorphism, it has an inverse $\varphi^{-1}\colon U \to \Omega$ that is also a homeomorphism, called a *(local) chart*. Since $\Omega \subseteq \mathbb{R}^m$, for every point $p \in M$ and every parametrization $\varphi\colon \Omega \to U$ of M at p, we have $\varphi^{-1}(p) = (z_1, \ldots, z_m)$ for some $z_i \in \mathbb{R}$, and we call $z_1, \ldots, z_m$ the *local coordinates of p (w.r.t. φ^{-1})*. We often refer to a manifold M without explicitly specifying its dimension (the integer m).

Intuitively, a chart provides a "flattened" local map of a region on a manifold. For instance, in the case of surfaces (2-dimensional manifolds), a chart is analogous to a planar map of a region on the surface. For a concrete example, consider a map giving a planar representation of a country, a region on the earth, a curved surface.

Remark: We could allow $m = 0$ in definition 1.11. If so, a manifold of dimension 0 is just a set of isolated points, and thus it has the discrete topology. In fact, it can be shown that

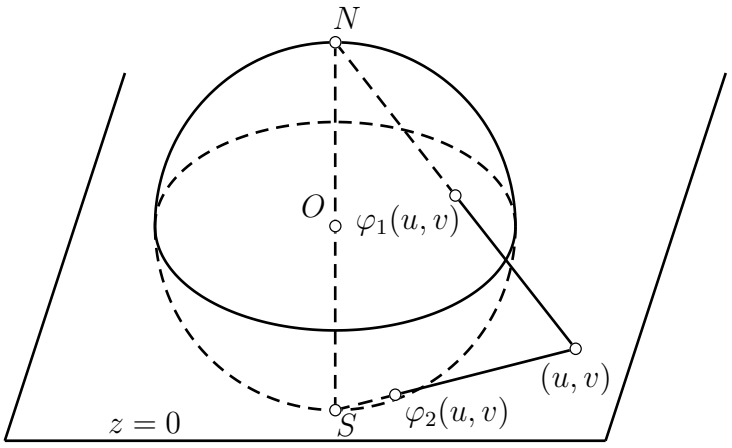


Figure 1.1: Inverse stereographic projections

a discrete subset of $\mathbb{R}^N$ is countable. Such manifolds are not very exciting, but they do correspond to discrete subgroups.

Example 1.3. The unit sphere S^2 in $\mathbb{R}^3$ defined such that

$$S^2 = \left\{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\right\}$$

is a smooth 2-manifold, because it can be parametrized using the following two maps φ_1 and φ_2:

$$\varphi_1 \colon (u, v) \mapsto \left(\frac{2u}{u^2 + v^2 + 1}, \frac{2v}{u^2 + v^2 + 1}, \frac{u^2 + v^2 - 1}{u^2 + v^2 + 1}\right)$$

and

$$\varphi_2 \colon (u, v) \mapsto \left(\frac{2u}{u^2 + v^2 + 1}, \frac{2v}{u^2 + v^2 + 1}, \frac{1 - u^2 - v^2}{u^2 + v^2 + 1}\right).$$

The map φ_1 corresponds to the inverse of the stereographic projection from the north pole $N = (0, 0, 1)$ onto the plane $z = 0$, and the map φ_2 corresponds to the inverse of the stereographic projection from the south pole $S = (0, 0, -1)$ onto the plane $z = 0$, as illustrated in Figure 1.1. We leave as an exercise to check that the map φ_1 parametrizes $S^2 - \{N\}$ and that the map φ_2 parametrizes $S^2 - \{S\}$ (and that they are smooth, homeomorphisms, etc.). Using φ_1, the open lower hemisphere is parametrized by the open disk of center O and radius 1 contained in the plane $z = 0$.

The chart φ_1^{-1} assigns local coordinates to the points in the open lower hemisphere. If we draw a grid of coordinate lines parallel to the x and y axes inside the open unit disk and map these lines onto the lower hemisphere using φ_1, we get curved lines on the lower hemisphere.

These "coordinate lines" on the lower hemisphere provide local coordinates for every point on the lower hemisphere. For this reason, older books often talk about *curvilinear coordinate systems* to mean the coordinate lines on a surface induced by a chart. We urge our readers to define a manifold structure on a torus. This can be done using four charts.

Every open subset of $\mathbb{R}^N$ is a manifold in a trivial way. Indeed, we can use the inclusion map as a parametrization. In particular, $\mathbf{GL}(n,\mathbb{R})$ is an open subset of $\mathbb{R}^{n^2}$, since its complement is closed (the set of invertible matrices is the inverse image of the determinant function, which is continuous). Thus, $\mathbf{GL}(n,\mathbb{R})$ is a manifold. We can view $\mathbf{GL}(n,\mathbb{C})$ as a subset of $\mathbb{R}^{(2n)^2}$ using the embedding defined as follows: For every complex $n \times n$ matrix A, construct the real $2n \times 2n$ matrix such that every entry $a + ib$ in A is replaced by the 2×2 block

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

where $a, b \in \mathbb{R}$. It is immediately verified that this map is in fact a group isomorphism. Thus, we can view $\mathbf{GL}(n,\mathbb{C})$ as a subgroup of $\mathbf{GL}(2n,\mathbb{R})$, and as a manifold in $\mathbb{R}^{(2n)^2}$.

A 1-manifold is called a *(smooth) curve*, and a 2-manifold is called a *(smooth) surface* (although some authors require that they also be connected).

The following two lemmas provide the link with the definition of an abstract manifold. The first lemma is easily shown using the inverse function theorem.

Lemma 1.21. *Given an m-dimensional manifold M in $\mathbb{R}^N$, for every $p \in M$ there are two open sets $O, W \subseteq \mathbb{R}^N$ with $0_N \in O$ and $p \in M \cap W$, and a smooth diffeomorphism $\varphi\colon O \to W$, such that $\varphi(0_N) = p$ and*

$$\varphi(O \cap (\mathbb{R}^m \times \{0_{N-m}\})) = M \cap W.$$

The next lemma is easily shown from Lemma 1.21 (see Berger and Gostiaux [17], Theorem 2.1.9 or DoCarmo [51], Chapter 0, Section 4). It is a key technical result used to show that interesting properties of maps between manifolds do not depend on parametrizations.

Lemma 1.22. *Given an m-dimensional manifold M in $\mathbb{R}^N$, for every $p \in M$ and any two parametrizations $\varphi_1\colon \Omega_1 \to U_1$ and $\varphi_2\colon \Omega_2 \to U_2$ of M at p, if $U_1 \cap U_2 \neq \emptyset$, the map $\varphi_2^{-1} \circ \varphi_1\colon \varphi_1^{-1}(U_1 \cap U_2) \to \varphi_2^{-1}(U_1 \cap U_2)$ is a smooth diffeomorphism.*

The maps $\varphi_2^{-1} \circ \varphi_1\colon \varphi_1^{-1}(U_1 \cap U_2) \to \varphi_2^{-1}(U_1 \cap U_2)$ are called *transition maps*. Lemma 1.22 is illustrated in Figure 1.2.

Using Definition 1.11, it may be quite hard to prove that a space is a manifold. Therefore, it is handy to have alternate characterizations such as those given in the next Proposition:

Proposition 1.23. *A subset, $M \subseteq \mathbb{R}^{m+k}$, is an m-dimensional manifold iff either*

(1) For every $p \in M$, there is some open subset, $W \subseteq \mathbb{R}^{m+k}$, with $p \in W$ and a (smooth) submersion, $f\colon W \to \mathbb{R}^k$, so that $W \cap M = f^{-1}(0)$,
or

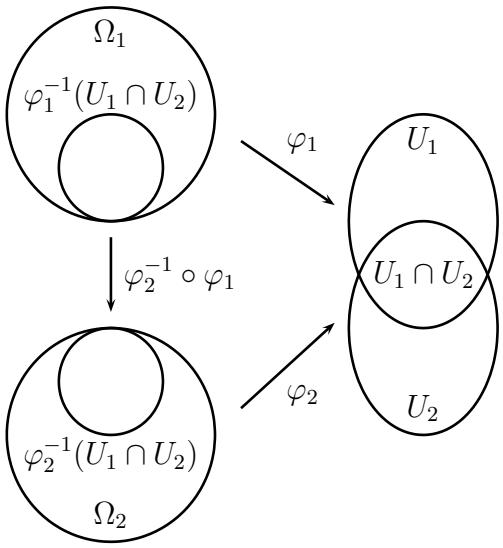


Figure 1.2: Parametrizations and transition functions

(2) For every $p \in M$, there is some open subset, $W \subseteq \mathbb{R}^{m+k}$, with $p \in W$ and a (smooth) map, $f\colon W \to \mathbb{R}^k$, so that $f'(p)$ is surjective and $W \cap M = f^{-1}(0)$.

Observe that condition (2), although apparently weaker than condition (1), is in fact equivalent to it, but more convenient in practice. This is because to say that $f'(p)$ is surjective means that the Jacobian matrix of $f'(p)$ has rank k, which means that some determinant is nonzero, and because the determinant function is continuous this must hold in some open subset $W_1 \subseteq W$ containing p. Consequently, the restriction, f_1, of f to W_1 is indeed a submersion and $f_1^{-1}(0) = W_1 \cap f^{-1}(0) = W_1 \cap W \cap M = W_1 \cap M$.

A proof of Proposition 1.23 can be found in Lafontaine [93] or Berger and Gostiaux [17]. Lemma 1.21 and Proposition 1.23 are actually *equivalent* to Definition 1.11. This equivalence is also proved in Lafontaine [93] and Berger and Gostiaux [17].

The proof, which is somewhat illuminating, is based on two technical lemmas that are proved using the inverse function theorem (for example, see Guillemin and Pollack [70], Chapter 1, Sections 3 and 4).

Lemma 1.24. *Let $U \subseteq \mathbb{R}^m$ be an open subset of $\mathbb{R}^m$ and pick some $a \in U$. If $f\colon U \to \mathbb{R}^n$ is a smooth immersion at a, i.e., df_a is injective (so, $m \leq n$), then there is an open set, $V \subseteq \mathbb{R}^n$, with $f(a) \in V$, an open subset, $U' \subseteq U$, with $a \in U'$ and $f(U') \subseteq V$, an open subset $O \subseteq \mathbb{R}^{n-m}$, and a diffeomorphism, $\theta\colon V \to U' \times O$, so that*

$$\theta(f(x_1, \ldots, x_m)) = (x_1, \ldots, x_m, 0, \ldots, 0),$$

for all $(x_1, \ldots, x_m) \in U'$ *as illustrated in the diagram below*

$$\begin{array}{ccc} U' \subseteq U & \xrightarrow{f} & f(U') \subseteq V \\ & \searrow^{in_1} & \downarrow \theta \\ & & U' \times O \end{array}$$

where $in_1(x_1, \ldots, x_m) = (x_1, \ldots, x_m, 0, \ldots, 0)$.

Proof. Since f is an immersion, its Jacobian matrix, $J(f)$, (an $n \times m$ matrix) has rank m and by permuting coordinates if needed, we may assume that the first m rows of $J(f)$ are linearly independent and we let

$$A = \left(\frac{\partial f_i}{\partial x_j}(a)\right)$$

be this invertible $m \times m$ matrix. Define the map, $g\colon U \times \mathbb{R}^{n-m} \to \mathbb{R}^n$, by

$$g(x, y) = (f_1(x), \ldots, f_m(x), y_1 + f_{m+1}(x), \ldots, y_{n-m} + f_n(x)),$$

for all $x \in U$ and all $y \in \mathbb{R}^{n-m}$. The Jacobian matrix of g at $(a, 0)$ is of the form

$$J = \begin{pmatrix} A & 0 \\ B & I \end{pmatrix}$$

so $\det(J) = \det(A)\det(I) = \det(A) \neq 0$, since A is invertible. By the inverse function theorem, there are some open subsets $W \subseteq U \times \mathbb{R}^{n-m}$ with $(a, 0) \in W$ and $V \subseteq \mathbb{R}^n$ such that the restriction of g to W is a diffeomorphism between W and V. Since $W \subseteq U \times \mathbb{R}^{n-m}$ is an open set, we can find some open subsets $U' \subseteq U$ and $O \subseteq \mathbb{R}^{n-m}$ so that $U' \times O \subseteq W$, $a \in U'$, and we can replace W by $U' \times O$ and restrict further g to this open set so that we obtain a diffeomorphism from $U' \times O$ to (a smaller) V. If $\theta\colon V \to U' \times O$ is the inverse of this diffeomorphism, then $f(U') \subseteq V$ and since $g(x, 0) = f(x)$,

$$\theta(g(x, 0)) = \theta(f(x_1, \ldots, x_m)) = (x_1, \ldots, x_m, 0, \ldots, 0),$$

for all $x = (x_1, \ldots, x_m) \in U'$. □

Lemma 1.25. *Let* $W \subseteq \mathbb{R}^m$ *be an open subset of* $\mathbb{R}^m$ *and pick some* $a \in W$. *If* $f\colon W \to \mathbb{R}^n$ *is a smooth submersion at* a, *i.e.,* df_a *is surjective (so,* $m \geq n$*), then there is an open set,* $V \subseteq W \subseteq \mathbb{R}^m$, *with* $a \in V$, *and a diffeomorphism* $\psi\colon O \to V$, *with domain* $O \subseteq \mathbb{R}^m$, *so that*

$$f(\psi(x_1, \ldots, x_m)) = (x_1, \ldots, x_n),$$

for all $(x_1, \ldots, x_m) \in O$ *as illustrated in the diagram below*

$$\begin{array}{ccccc} O \subseteq \mathbb{R}^m & \xrightarrow{\psi} & V \subseteq W \subseteq \mathbb{R}^m & & \\ \downarrow \pi & & & \swarrow f & \\ \mathbb{R}^n & & & & \end{array}$$

where $\pi(x_1, \ldots, x_m) = (x_1, \ldots, x_n)$.

Proof. Since f is a submersion, its Jacobian matrix, $J(f)$, (an $n \times m$ matrix) has rank n and by permuting coordinates if needed, we may assume that the first n columns of $J(f)$ are linearly independent and we let

$$A = \left(\frac{\partial f_i}{\partial x_j}(a)\right)$$

be this invertible $n \times n$ matrix. Define the map, $g\colon W \to \mathbb{R}^m$, by

$$g(x) = (f(x), x_{n+1}, \ldots, x_m),$$

for all $x \in W$. The Jacobian matrix of g at a is of the form

$$J = \begin{pmatrix} A & B \\ 0 & I \end{pmatrix}$$

so $\det(J) = \det(A)\det(I) = \det(A) \neq 0$, since A is invertible. By the inverse function theorem, there are some open subsets $V \subseteq W$ with $a \in V$ and $O \subseteq \mathbb{R}^m$ such that the restriction of g to V is a diffeomorphism between V and O. Let $\psi\colon O \to V$ be the inverse of this diffeomorphism. Because $g \circ \psi = \mathrm{id}$, we have

$$(x_1, \ldots, x_m) = g(\psi(x)) = (f(\psi(x)), \psi_{n+1}(x), \ldots, \psi_m(x)),$$

that is,

$$f(\psi(x_1, \ldots, x_m)) = (x_1, \ldots, x_n)$$

for all $(x_1, \ldots, x_m) \in O$, as desired. □

Using Lemmas 1.24 and 1.25, we can prove the following theorem which confirms that all our characterizations of a manifold are equivalent.

Theorem 1.26. *A nonempty subset, $M \subseteq \mathbb{R}^N$, is an m-manifold (with $1 \leq m \leq N$) iff any of the following conditions hold:*

(1) For every $p \in M$, there are two open subsets $\Omega \subseteq \mathbb{R}^m$ and $U \subseteq M$, with $p \in U$, and a smooth function $\varphi\colon \Omega \to \mathbb{R}^N$ such that φ is a homeomorphism between Ω and $U = \varphi(\Omega)$, and $\varphi'(0)$ is injective, where $p = \varphi(0)$.

(2) For every $p \in M$, there are two open sets $O, W \subseteq \mathbb{R}^N$ with $0_N \in O$ and $p \in M \cap W$, and a smooth diffeomorphism $\varphi\colon O \to W$, such that $\varphi(0_N) = p$ and

$$\varphi(O \cap (\mathbb{R}^m \times \{0_{N-m}\})) = M \cap W.$$

(3) For every $p \in M$, there is some open subset, $W \subseteq \mathbb{R}^N$, with $p \in W$ and a smooth submersion, $f\colon W \to \mathbb{R}^{N-m}$, so that $W \cap M = f^{-1}(0)$.

(4) For every $p \in M$, there is some open subset, $W \subseteq \mathbb{R}^N$, with $p \in W$ and $N - m$ smooth functions, $f_i \colon W \to \mathbb{R}$, so that the linear forms $df_1(p), \ldots, df_{N-m}(p)$ are linearly independent and

$$W \cap M = f_1^{-1}(0) \cap \cdots \cap f_{N-m}^{-1}(0).$$

Proof. If (1) holds, then by Lemma 1.24, replacing Ω by a smaller open subset $\Omega' \subseteq \Omega$ if necessary, there is some open subset $V \subseteq \mathbb{R}^N$ with $p \in V$ and $\varphi(\Omega') \subseteq V$, an open subset, $O \subseteq \mathbb{R}^{N-m}$, and some diffeomorphism, $\theta \colon V \to \Omega' \times O$, so that

$$(\theta \circ \varphi)(x_1, \ldots, x_m) = (x_1, \ldots, x_m, 0, \ldots, 0),$$

for all $(x_1, \ldots, x_m) \in \Omega'$. Observe that the above condition implies that

$$(\theta \circ \varphi)(\Omega') = \theta(V) \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\}).$$

Since φ is a homeomorphism between Ω and its image in M and since $\Omega' \subseteq \Omega$ is an open subset, $\varphi(\Omega') = M \cap W'$ for some open subset $W' \subseteq \mathbb{R}^N$, so if we let $W = V \cap W'$, because $\varphi(\Omega') \subseteq V$ it follows that $\varphi(\Omega') = M \cap W$ and

$$\theta(W \cap M) = \theta(\varphi(\Omega')) = \theta(V) \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\}).$$

However, θ is injective and $\theta(W \cap M) \subseteq \theta(W)$ so

$$\begin{aligned} \theta(W \cap M) &= \theta(W) \cap \theta(V) \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\}) \\ &= \theta(W \cap V) \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\}) \\ &= \theta(W) \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\}). \end{aligned}$$

If we let $O = \theta(W)$, we get

$$\theta^{-1}(O \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\})) = M \cap W,$$

which is (2).

If (2) holds, we can write $\varphi^{-1} = (f_1, \ldots, f_N)$ and because $\varphi^{-1} \colon W \to O$ is a diffeomorphism, $df_1(q), \ldots, df_N(q)$ are linearly independent for all $q \in W$, so the map

$$f = (f_{m+1}, \ldots, f_N)$$

is a submersion, $f \colon W \to \mathbb{R}^{N-m}$, and we have $f(x) = 0$ iff $f_{m+1}(x) = \cdots = f_N(x) = 0$ iff

$$\varphi^{-1}(x) = (f_1(x), \ldots, f_m(x), 0, \ldots, 0)$$

iff $\varphi^{-1}(x) \in O \cap (\mathbb{R}^m \times \{0_{N-m}\})$ iff $x \in \varphi(O \cap (\mathbb{R}^m \times \{0_{N-m}\}) = M \cap W$, because

$$\varphi(O \cap (\mathbb{R}^m \times \{0_{N-m}\})) = M \cap W.$$

Thus, $M \cap W = f^{-1}(0)$, which is (3).

The proof that (3) implies (2) uses Lemma 1.25 instead of Lemma 1.24. If $f\colon W \to \mathbb{R}^{N-m}$ is the submersion such that $M \cap W = f^{-1}(0)$ given by (3), then by Lemma 1.25, there are open subsets $V \subseteq W$, $O \subseteq \mathbb{R}^N$ and a diffeomorphism, $\psi\colon O \to V$ so that

$$f(\psi(x_1, \ldots, x_N)) = (x_1, \ldots, x_{N-m})$$

for all $(x_1, \ldots, x_N) \in O$. If σ is the permutation of variables given by

$$\sigma(x_1, \ldots, x_m, x_{m+1}, \ldots, x_N) = (x_{m+1}, \ldots, x_N, x_1, \ldots, x_m),$$

then $\varphi = \psi \circ \sigma$ is a diffeomorphism such that

$$f(\varphi(x_1, \ldots, x_N)) = (x_{m+1}, \ldots, x_N)$$

for all $(x_1, \ldots, x_N) \in O$. If we denote the restriction of f to V by g, it is clear that

$$M \cap V = g^{-1}(0)$$

and because $g(\varphi(x_1, \ldots, x_N)) = 0$ iff $(x_{m+1}, \ldots, x_N) = 0_{N-m}$ and φ is a bijection,

$$\begin{aligned} M \cap V &= \{(y_1, \ldots, y_N) \in V \mid g(y_1, \ldots, y_N) = 0\} \\ &= \{\varphi(x_1, \ldots, x_N) \mid (\exists (x_1, \ldots, x_N) \in O)(g(\varphi(x_1, \ldots, x_N)) = 0)\} \\ &= \varphi(O \cap (\mathbb{R}^m \times \{0_{N-m}\})), \end{aligned}$$

which is (2).

If (2) holds, then $\varphi\colon O \to W$ is a diffeomorphism,

$$O \cap (\mathbb{R}^m \times \{0_{N-m}\}) = \Omega \times \{0_{N-m}\}$$

for some open subset, $\Omega \subseteq \mathbb{R}^m$, and the map $\psi\colon \Omega \to \mathbb{R}^N$ given by

$$\psi(x) = \varphi(x, 0_{N-m})$$

is an immersion on Ω and a homeomorhism onto $U \cap M$, which implies (1).

If (3) holds, then if we write $f = (f_1, \ldots, f_{N-m})$, with $f_i\colon W \to \mathbb{R}$, then the fact that $df(p)$ is a submersion is equivalent to the fact that the linear forms $df_1(p), \ldots, df_{N-m}(p)$ are linearly independent and

$$M \cap W = f^{-1}(0) = f_1^{-1}(0) \cap \cdots \cap f_{N-m}^{-1}(0).$$

Finally, if (4) holds, then if we define $f\colon W \to \mathbb{R}^{N-m}$ by

$$f = (f_1, \ldots, f_{N-m}),$$

because $df_1(p), \ldots, df_{N-m}(p)$ are linearly independent we get a smooth map which is a submersion at p such that

$$M \cap W = f^{-1}(0).$$

Now, f is a submersion at p iff $df(p)$ is surjective, which means that a certain determinant is nonzero and since the determinant function is continuous, this determinant is nonzero on some open subset, $W' \subseteq W$, containing p, so if we restrict f to W', we get an immersion on W' such that $M \cap W' = f^{-1}(0)$. □

Condition (4) says that locally (that is, in a small open set of M containing $p \in M$), M is "cut out" by $N - m$ smooth functions, $f_i \colon W \to \mathbb{R}$, in the sense that the portion of the manifold $M \cap W$ is the intersection of the $N - m$ hypersurfaces, $f_i^{-1}(0)$, (the zero-level sets of the f_i) and that this intersection is "clean", which means that the linear forms $df_1(p), \ldots, df_{N-m}(p)$ are linearly independent.

As an illustration of Theorem 1.26, we can show again that the sphere

$$S^n = \{x \in \mathbb{R}^{n+1} \mid \|x\|_2^2 - 1 = 0\}$$

is an n-dimensional manifold in $\mathbb{R}^{n+1}$. Indeed, the map $f \colon \mathbb{R}^{n+1} \to \mathbb{R}$ given by $f(x) = \|x\|_2^2 - 1$ is a submersion (for $x \neq 0$) since

$$df(x)(y) = 2\sum_{k=1}^{n+1} x_k y_k.$$

We can also show that the rotation group, $\mathbf{SO}(n)$, is an $\frac{n(n-1)}{2}$-dimensional manifold in $\mathbb{R}^{n^2}$.

Indeed, $\mathbf{GL}^+(n)$ is an open subset of $\mathbb{R}^{n^2}$ (recall, $\mathbf{GL}^+(n) = \{A \in \mathbf{GL}(n) \mid \det(A) > 0\}$) and if f is defined by

$$f(A) = A^\top A - I,$$

where $A \in \mathbf{GL}^+(n)$, then $f(A)$ is symmetric, so $f(A) \in \mathbf{S}(n) = \mathbb{R}^{\frac{n(n+1)}{2}}$.

It is easy to show (using directional derivatives) that

$$df(A)(H) = A^\top H + H^\top A.$$

But then, $df(A)$ is surjective for all $A \in \mathbf{SO}(n)$, because if S is any symmetric matrix, we see that

$$df(A)\left(\frac{AS}{2}\right) = S.$$

As $\mathbf{SO}(n) = f^{-1}(0)$, we conclude that $\mathbf{SO}(n)$ is indeed a manifold.

A similar argument proves that $\mathbf{O}(n)$ is an $\frac{n(n-1)}{2}$-dimensional manifold. Using the map, $f \colon \mathbf{GL}(n) \to \mathbb{R}$, given by $A \mapsto \det(A)$, we can prove that $\mathbf{SL}(n)$ is a manifold of dimension $n^2 - 1$.

Remark: We have $df(A)(B) = \det(A)\mathrm{tr}(A^{-1}B)$ for every $A \in \mathbf{GL}(n)$, where $f(A) = \det(A)$.

The third characterization of Theorem 1.26 suggests the following definition.

Definition 1.12. Let $f\colon \mathbb{R}^{m+k} \to \mathbb{R}^k$ be a smooth function. A point, $p \in \mathbb{R}^{m+k}$, is called a *critical point (of f)* iff df_p is *not* surjective and a point $q \in \mathbb{R}^k$ is called a *critical value (of f)* iff $q = f(p)$, for some critical point, $p \in \mathbb{R}^{m+k}$. A point $p \in \mathbb{R}^{m+k}$ is a *regular point (of f)* iff p is not critical, i.e., df_p is surjective, and a point $q \in \mathbb{R}^k$ is a *regular value (of f)* iff it is not a critical value. In particular, any $q \in \mathbb{R}^k - f(\mathbb{R}^{m+k})$ is a regular value and $q \in f(\mathbb{R}^{m+k})$ is a regular value iff *every* $p \in f^{-1}(q)$ is a regular point (but, in contrast, q is a critical value iff *some* $p \in f^{-1}(q)$ is critical).

Part (3) of Theorem 1.26 implies the following useful proposition:

Proposition 1.27. *Given any smooth function, $f\colon \mathbb{R}^{m+k} \to \mathbb{R}^k$, for every regular value, $q \in f(\mathbb{R}^{m+k})$, the preimage, $Z = f^{-1}(q)$, is a manifold of dimension m.*

Definition 1.12 and Proposition 1.27 can be generalized to manifolds. Regular and critical values of smooth maps play an important role in differential topology. Firstly, given a smooth map, $f\colon \mathbb{R}^{m+k} \to \mathbb{R}^k$, almost every point of $\mathbb{R}^k$ is a regular value of f. To make this statement precise, one needs the notion of a *set of measure zero*. Then, *Sard's theorem* says that the set of critical values of a smooth map has measure zero. Secondly, if we consider smooth functions, $f\colon \mathbb{R}^{m+1} \to \mathbb{R}$, a point $p \in \mathbb{R}^{m+1}$ is critical iff $df_p = 0$. Then, we can use second order derivatives to further classify critical points. The *Hessian matrix* of f (at p) is the matrix of second-order partials

$$H_f(p) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(p)\right)$$

and a critical point p is a *nondegenerate critical point* if $H_f(p)$ is a nonsingular matrix. The remarkable fact is that, at a nondegenerate critical point, p, the local behavior of f is completely determined, in the sense that after a suitable change of coordinates (given by a smooth diffeomorphism)

$$f(x) = f(p) - x_1^2 - \cdots - x_\lambda^2 + x_{\lambda+1}^2 + \cdots + x_{m+1}^2$$

near p, where λ called the *index of f at p* is an integer which depends only on p (in fact, λ is the number of negative eigenvalues of $H_f(p)$). This result is known as *Morse lemma* (after Marston Morse, 1892-1977).

Smooth functions whose critical points are all nondegenerate are called *Morse functions*. It turns out that every smooth function, $f\colon \mathbb{R}^{m+1} \to \mathbb{R}$, gives rise to a large supply of Morse functions by adding a linear function to it. More precisely, the set of $a \in \mathbb{R}^{m+1}$ for which the function f_a given by

$$f_a(x) = f(x) + a_1 x_1 + \cdots + a_{m+1} x_{m+1}$$

is not a Morse function has measure zero.

Morse functions can be used to study topological properties of manifolds. In a sense to be made precise and under certain technical conditions, a Morse function can be used to reconstuct a manifold by attaching cells, up to homotopy equivalence. However, these results are way beyond the scope of this book. A fairly elementary exposition of nondegenerate critical points and Morse functions can be found in Guillemin and Pollack [70] (Chapter 1, Section 7). Sard's theorem is proved in Appendix 1 of Guillemin and Pollack [70] and also in Chapter 2 of Milnor [109]. Morse theory (starting with Morse lemma) and much more, is discussed in Milnor [107], widely recognized as a mathematical masterpiece. An excellent and more leisurely introduction to Morse theory is given in Matsumoto [106], where a proof of Morse lemma is also given.

Let us now review the definitions of a smooth curve in a manifold and the tangent vector at a point of a curve.

Definition 1.13. Let M be an m-dimensional manifold in $\mathbb{R}^N$. A *smooth curve γ in M* is any function $\gamma\colon I \to M$ where I is an open interval in $\mathbb{R}$ and such that for every $t \in I$, letting $p = \gamma(t)$, there is some parametrization $\varphi\colon \Omega \to U$ of M at p and some open interval $]t-\epsilon,\, t+\epsilon[\subseteq I$ such that the curve $\varphi^{-1} \circ \gamma\colon]t-\epsilon,\, t+\epsilon[\to \mathbb{R}^m$ is smooth.

Using Lemma 1.22, it is easily shown that Definition 1.13 does not depend on the choice of the parametrization $\varphi\colon \Omega \to U$ at p.

Lemma 1.22 also implies that γ viewed as a curve $\gamma\colon I \to \mathbb{R}^N$ is smooth. Then the *tangent vector to the curve $\gamma\colon I \to \mathbb{R}^N$ at t*, denoted by $\gamma'(t)$, is the value of the derivative of γ at t (a vector in $\mathbb{R}^N$) computed as usual:

$$\gamma'(t) = \lim_{h \mapsto 0} \frac{\gamma(t+h) - \gamma(t)}{h}.$$

Given any point $p \in M$, we will show that the set of tangent vectors to all smooth curves in M through p is a vector space isomorphic to the vector space $\mathbb{R}^m$. The tangent vector at p to a curve γ on a manifold M is illustrated in Figure 1.3.

Given a smooth curve $\gamma\colon I \to M$, for any $t \in I$, letting $p = \gamma(t)$, since M is a manifold, there is a parametrization $\varphi\colon \Omega \to U$ such that $\varphi(0_m) = p \in U$ and some open interval $J \subseteq I$ with $t \in J$ and such that the function

$$\varphi^{-1} \circ \gamma\colon J \to \mathbb{R}^m$$

is a smooth curve, since γ is a smooth curve. Letting $\alpha = \varphi^{-1} \circ \gamma$, the derivative $\alpha'(t)$ is well-defined, and it is a vector in $\mathbb{R}^m$. But $\varphi \circ \alpha\colon J \to M$ is also a smooth curve, which agrees with γ on J, and by the chain rule,

$$\gamma'(t) = \varphi'(0_m)(\alpha'(t)),$$

since $\alpha(t) = 0_m$ (because $\varphi(0_m) = p$ and $\gamma(t) = p$). Observe that $\gamma'(t)$ is a vector in $\mathbb{R}^N$. Now, for every vector $v \in \mathbb{R}^m$, the curve $\alpha\colon J \to \mathbb{R}^m$ defined such that

$$\alpha(u) = (u-t)v$$

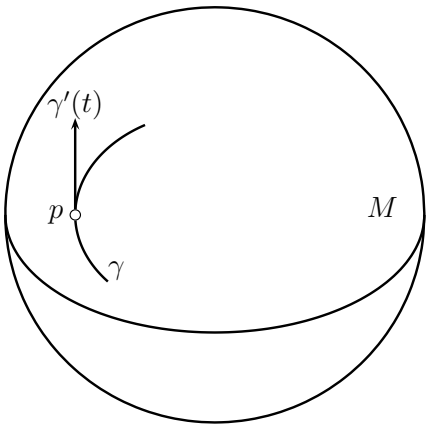


Figure 1.3: Tangent vector to a curve on a manifold

for all $u \in J$ is clearly smooth, and $\alpha'(t) = v$. This shows that the set of tangent vectors at t to all smooth curves (in $\mathbb{R}^m$) passing through 0_m is the entire vector space $\mathbb{R}^m$. Since every smooth curve $\gamma\colon I \to M$ agrees with a curve of the form $\varphi \circ \alpha\colon J \to M$ for some smooth curve $\alpha\colon J \to \mathbb{R}^m$ (with $J \subseteq I$) as explained above, and since it is assumed that $\varphi'(0_m)$ is injective, $\varphi'(0_m)$ maps the vector space $\mathbb{R}^m$ injectively to the set of tangent vectors to γ at p, as claimed. All this is summarized in the following definition.

Definition 1.14. Let M be an m-dimensional manifold in $\mathbb{R}^N$. For every point $p \in M$, the *tangent space* T_pM *at* p is the set of all vectors in $\mathbb{R}^N$ of the form $\gamma'(0)$, where $\gamma\colon I \to M$ is any smooth curve in M such that $p = \gamma(0)$. The set T_pM is a vector space isomorphic to $\mathbb{R}^m$. Every vector $v \in T_pM$ is called a *tangent vector to* M *at* p.

We can now define Lie groups (postponing defining smooth maps).

Definition 1.15. A *Lie group* is a nonempty subset G of $\mathbb{R}^N$ ($N \geq 1$) satisfying the following conditions:

(a) G is a group.

(b) G is a manifold in $\mathbb{R}^N$.

(c) The group operation $\cdot : G \times G \to G$ and the inverse map $^{-1}\colon G \to G$ are smooth.

(Smooth maps are defined in Definition 1.18). It is immediately verified that $\mathbf{GL}(n, \mathbb{R})$ is a Lie group. Since all the Lie groups that we are considering are subgroups of $\mathbf{GL}(n, \mathbb{R})$, the following definition is in order.

Definition 1.16. A *linear Lie group* is a subgroup G of $\mathbf{GL}(n,\mathbb{R})$ (for some $n \geq 1$) which is a smooth manifold in $\mathbb{R}^{n^2}$.

Let $\mathbf{M}(n,\mathbb{R})$ denote the set of all real $n \times n$ matrices (invertible or not). If we recall that the exponential map

$$\exp\colon A \mapsto e^A$$

is well defined on $\mathbf{M}(n,\mathbb{R})$, we have the following crucial theorem due to Von Neumann and Cartan.

Theorem 1.28. *A closed subgroup G of $\mathbf{GL}(n,\mathbb{R})$ is a linear Lie group. Furthermore, the set $\mathfrak{g}$ defined such that*

$$\mathfrak{g} = \{X \in \mathbf{M}(n,\mathbb{R}) \mid e^{tX} \in G \text{ for all } t \in \mathbb{R}\}$$

is a vector space equal to the tangent space $T_I G$ at the identity I, and $\mathfrak{g}$ is closed under the Lie bracket $[-,-]$ defined such that $[A,B] = AB - BA$ for all $A, B \in \mathbf{M}(n,\mathbb{R})$.

Theorem 1.28 applies even when G is a discrete subgroup, but in this case, $\mathfrak{g}$ is trivial (i.e., $\mathfrak{g} = \{0\}$). For example, the set of nonnull reals $\mathbb{R}^* = \mathbb{R} - \{0\} = \mathbf{GL}(1,\mathbb{R})$ is a Lie group under multiplication, and the subgroup

$$H = \{2^n \mid n \in \mathbb{Z}\}$$

is a discrete subgroup of $\mathbb{R}^*$. Thus, H is a Lie group. On the other hand, the set $\mathbb{Q}^* = \mathbb{Q} - \{0\}$ of nonnull rational numbers is a multiplicative subgroup of $\mathbb{R}^*$, but it is not closed, since $\mathbb{Q}$ is dense in $\mathbb{R}$.

The proof of Theorem 1.28 involves proving that when G is not a discrete subgroup, there is an open subset $\Omega \subseteq \mathbf{M}(n,\mathbb{R})$ such that $0_{n,n} \in \Omega$, an open subset $W \subseteq \mathbf{M}(n,\mathbb{R})$ such that $I \in W$, and that $\exp\colon \Omega \to W$ is a diffeomorphism such that

$$\exp(\Omega \cap \mathfrak{g}) = W \cap G.$$

If G is closed and not discrete, we must have $m \geq 1$, and $\mathfrak{g}$ has dimension m.

With the help of Theorem 1.28 it is now very easy to prove that $\mathbf{SL}(n)$, $\mathbf{O}(n)$, $\mathbf{SO}(n)$, $\mathbf{SL}(n,\mathbb{C})$, $\mathbf{U}(n)$, and $\mathbf{SU}(n)$ are Lie groups and to figure out what are their Lie algebras. (Of course, $\mathbf{GL}(n,\mathbb{R})$ is a Lie group, as we already know.)

For example, if $G = \mathbf{GL}(n,\mathbb{R})$, as e^{tA} is invertible for *every* matrix, $A \in \mathbf{M}(n,\mathbb{R})$, we deduce that the Lie algebra, $\mathfrak{gl}(n,\mathbb{R})$, of $\mathbf{GL}(n,\mathbb{R})$ is equal to $\mathbf{M}(n,\mathbb{R})$. We also claim that the Lie algebra, $\mathfrak{sl}(n,\mathbb{R})$, of $\mathbf{SL}(n,\mathbb{R})$ is the set of all matrices with zero trace. Indeed, $\mathfrak{sl}(n,\mathbb{R})$ is the subalgebra of $\mathfrak{gl}(n,\mathbb{R})$ consisting of all matrices $X \in \mathfrak{gl}(n,\mathbb{R})$ such that

$$\det(e^{tX}) = 1$$

for all $t \in \mathbb{R}$, and because $\det(e^{tX}) = e^{\mathrm{tr}(tX)}$, for $t = 1$, we get $\mathrm{tr}(X) = 0$, as claimed.

We can also prove that $\mathbf{SE}(n)$ is a Lie group as follows. Recall that we can view every element of $\mathbf{SE}(n)$ as a real $(n+1) \times (n+1)$ matrix

$$\begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix}$$

where $R \in \mathbf{SO}(n)$ and $U \in \mathbb{R}^n$. In fact, such matrices belong to $\mathbf{SL}(n+1)$. This embedding of $\mathbf{SE}(n)$ into $\mathbf{SL}(n+1)$ is a group homomorphism, since the group operation on $\mathbf{SE}(n)$ corresponds to multiplication in $\mathbf{SL}(n+1)$:

$$\begin{pmatrix} RS & RV + U \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix} \begin{pmatrix} S & V \\ 0 & 1 \end{pmatrix}.$$

Note that the inverse is given by

$$\begin{pmatrix} R^{-1} & -R^{-1}U \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} R^\top & -R^\top U \\ 0 & 1 \end{pmatrix}.$$

Also note that the embedding shows that, as a manifold, $\mathbf{SE}(n)$ is diffeomorphic to $\mathbf{SO}(n) \times \mathbb{R}^n$ (given a manifold M_1 of dimension m_1 and a manifold M_2 of dimension m_2, the product $M_1 \times M_2$ can be given the structure of a manifold of dimension $m_1 + m_2$ in a natural way). Thus, $\mathbf{SE}(n)$ is a Lie group with underlying manifold $\mathbf{SO}(n) \times \mathbb{R}^n$, and in fact, a subgroup of $\mathbf{SL}(n+1)$.

Even though $\mathbf{SE}(n)$ is diffeomorphic to $\mathbf{SO}(n) \times \mathbb{R}^n$ as a manifold, it is *not* isomorphic to $\mathbf{SO}(n) \times \mathbb{R}^n$ as a group, because the group multiplication on $\mathbf{SE}(n)$ is not the multiplication on $\mathbf{SO}(n) \times \mathbb{R}^n$. Instead, $\mathbf{SE}(n)$ is a *semidirect product* of $\mathbf{SO}(n)$ and $\mathbb{R}^n$; see Gallier [60], Chapter 2, Problem 2.19).

Returning to Theorem 1.28, the vector space $\mathfrak{g}$ is called the *Lie algebra* of the Lie group G. Lie algebras are defined as follows.

Definition 1.17. A *(real) Lie algebra* $\mathcal{A}$ is a real vector space together with a bilinear map $[\cdot, \cdot] \colon \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ called the *Lie bracket* on $\mathcal{A}$ such that the following two identities hold for all $a, b, c \in \mathcal{A}$:

$$[a,\, a] = 0,$$

and the so-called *Jacobi identity*

$$[a,\, [b,\, c]] + [c,\, [a,\, b]] + [b,\, [c,\, a]] = 0.$$

It is immediately verified that $[b, a] = -[a, b]$.

In view of Theorem 1.28, the vector space $\mathfrak{g} = T_I G$ associated with a Lie group G is indeed a Lie algebra. Furthermore, the exponential map $\exp\colon \mathfrak{g} \to G$ is well-defined. In general, exp is neither injective nor surjective, as we observed earlier. Theorem 1.28 also provides a kind of recipe for "computing" the Lie algebra $\mathfrak{g} = T_I G$ of a Lie group G. Indeed, $\mathfrak{g}$ is the tangent space to G at I, and thus we can use curves to compute tangent vectors. Actually, for every $X \in T_I G$, the map

$$\gamma_X\colon t \mapsto e^{tX}$$

is a smooth curve in G, and it is easily shown that $\gamma'_X(0) = X$. Thus, we can use these curves. As an illustration, we show that the Lie algebras of $\mathbf{SL}(n)$ and $\mathbf{SO}(n)$ are the matrices with null trace and the skew symmetric matrices.

Let $t \mapsto R(t)$ be a smooth curve in $\mathbf{SL}(n)$ such that $R(0) = I$. We have $\det(R(t)) = 1$ for all $t \in]-\epsilon, \epsilon[$. Using the chain rule, we can compute the derivative of the function

$$t \mapsto \det(R(t))$$

at $t = 0$, and we get

$$\det{}'_I(R'(0)) = 0.$$

It is an easy exercise to prove that

$$\det{}'_I(X) = \mathrm{tr}(X),$$

and thus $\mathrm{tr}(R'(0)) = 0$, which says that the tangent vector $X = R'(0)$ has null trace. Clearly, $\mathfrak{sl}(n, \mathbb{R})$ has dimension $n^2 - 1$.

Let $t \mapsto R(t)$ be a smooth curve in $\mathbf{SO}(n)$ such that $R(0) = I$. Since each $R(t)$ is orthogonal, we have

$$R(t)\, R(t)^\top = I$$

for all $t \in]-\epsilon, \epsilon[$. Taking the derivative at $t = 0$, we get

$$R'(0)\, R(0)^\top + R(0)\, R'(0)^\top = 0,$$

but since $R(0) = I = R(0)^\top$, we get

$$R'(0) + R'(0)^\top = 0,$$

which says that the tangent vector $X = R'(0)$ is skew symmetric. Since the diagonal elements of a skew symmetric matrix are null, the trace is automatically null, and the condition $\det(R) = 1$ yields nothing new. This shows that $\mathfrak{o}(n) = \mathfrak{so}(n)$. It is easily shown that $\mathfrak{so}(n)$ has dimension $n(n-1)/2$.

As a concrete example, the Lie algebra $\mathfrak{so}(3)$ of $\mathbf{SO}(3)$ is the real vector space consisting of all 3×3 real skew symmetric matrices. Every such matrix is of the form

$$\begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix}$$

where $b, c, d \in \mathbb{R}$. The Lie bracket $[A, B]$ in $\mathfrak{so}(3)$ is also given by the usual commutator, $[A, B] = AB - BA$.

We can define an isomorphism of Lie algebras $\psi\colon (\mathbb{R}^3, \times) \to \mathfrak{so}(3)$ by the formula

$$\psi(b, c, d) = \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix}.$$

It is indeed easy to verify that

$$\psi(u \times v) = [\psi(u),\, \psi(v)].$$

It is also easily verified that for any two vectors $u = (b, c, d)$ and $v = (b', c', d')$ in $\mathbb{R}^3$

$$\psi(u)(v) = u \times v.$$

The exponential map $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ is given by Rodrigues's formula (see Lemma 1.7):

$$e^A = \cos\theta\, I_3 + \frac{\sin\theta}{\theta} A + \frac{(1 - \cos\theta)}{\theta^2} B,$$

or equivalently by

$$e^A = I_3 + \frac{\sin\theta}{\theta} A + \frac{(1 - \cos\theta)}{\theta^2} A^2$$

if $\theta \neq 0$, where

$$A = \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix},$$

$\theta = \sqrt{b^2 + c^2 + d^2}$, $B = A^2 + \theta^2 I_3$, and with $e^{0_3} = I_3$.

Using the above methods, it is easy to verify that the Lie algebras $\mathfrak{gl}(n, \mathbb{R})$, $\mathfrak{sl}(n, \mathbb{R})$, $\mathfrak{o}(n)$, and $\mathfrak{so}(n)$, are respectively $\mathbf{M}(n, \mathbb{R})$, the set of matrices with null trace, and the set of skew symmetric matrices (in the last two cases). A similar computation can be done for $\mathfrak{gl}(n, \mathbb{C})$, $\mathfrak{sl}(n, \mathbb{C})$, $\mathfrak{u}(n)$, and $\mathfrak{su}(n)$, confirming the claims of Section 1.4. It is easy to show that $\mathfrak{gl}(n, \mathbb{C})$ has dimension $2n^2$, $\mathfrak{sl}(n, \mathbb{C})$ has dimension $2(n^2 - 1)$, $\mathfrak{u}(n)$ has dimension n^2, and $\mathfrak{su}(n)$ has dimension $n^2 - 1$.

For example, the Lie algebra $\mathfrak{su}(2)$ of $\mathbf{SU}(2)$ (or S^3) is the real vector space consisting of all 2×2 (complex) skew Hermitian matrices of null trace. Every such matrix is of the form

$$i(d\sigma_1 + c\sigma_2 + b\sigma_3) = \begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix},$$

where $b, c, d \in \mathbb{R}$, and $\sigma_1, \sigma_2, \sigma_3$ are the Pauli spin matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

and thus the matrices $i\sigma_1, i\sigma_2, i\sigma_3$ form a basis of the Lie algebra $\mathfrak{su}(2)$. The Lie bracket $[A, B]$ in $\mathfrak{su}(2)$ is given by the usual commutator, $[A, B] = AB - BA$.

It is easily checked that the vector space $\mathbb{R}^3$ is a Lie algebra if we define the Lie bracket on $\mathbb{R}^3$ as the usual cross product $u \times v$ of vectors. Then we can define an isomorphism of Lie algebras $\varphi\colon (\mathbb{R}^3, \times) \to \mathfrak{su}(2)$ by the formula

$$\varphi(b, c, d) = \frac{i}{2}(d\sigma_1 + c\sigma_2 + b\sigma_3) = \frac{1}{2}\begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix}.$$

It is indeed easy to verify that

$$\varphi(u \times v) = [\varphi(u), \varphi(v)].$$

Returning to $\mathfrak{su}(2)$, letting $\theta = \sqrt{b^2 + c^2 + d^2}$, we can write

$$d\sigma_1 + c\sigma_2 + b\sigma_3 = \begin{pmatrix} b & -ic + d \\ ic + d & -b \end{pmatrix} = \theta A,$$

where

$$A = \frac{1}{\theta}(d\sigma_1 + c\sigma_2 + b\sigma_3) = \frac{1}{\theta}\begin{pmatrix} b & -ic + d \\ ic + d & -b \end{pmatrix},$$

so that $A^2 = I$, and it can be shown that the exponential map $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$ is given by

$$\exp(i\theta A) = \cos\theta\, \mathbf{1} + i \sin\theta\, A.$$

In view of the isomorphism $\varphi\colon (\mathbb{R}^3, \times) \to \mathfrak{su}(2)$, where

$$\varphi(b, c, d) = \frac{1}{2}\begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix} = i\frac{\theta}{2}A,$$

the exponential map can be viewed as a map $\exp\colon (\mathbb{R}^3, \times) \to \mathbf{SU}(2)$ given by the formula

$$\exp(\theta v) = \left[\cos\frac{\theta}{2}, \sin\frac{\theta}{2}\, v\right],$$

for every vector θv, where v is a unit vector in $\mathbb{R}^3$ and $\theta \in \mathbb{R}$. In this form, $\exp(\theta v)$ is a quaternion corresponding to a rotation of axis v and angle θ.

As we showed, $\mathbf{SE}(n)$ is a Lie group, and its lie algebra $\mathfrak{se}(n)$ described in Section 1.6 is easily determined as the subalgebra of $\mathfrak{sl}(n + 1)$ consisting of all matrices of the form

$$\begin{pmatrix} B & U \\ 0 & 0 \end{pmatrix}$$

where $B \in \mathfrak{so}(n)$ and $U \in \mathbb{R}^n$. Thus, $\mathfrak{se}(n)$ has dimension $n(n+1)/2$. The Lie bracket is given by

$$\begin{pmatrix} B & U \\ 0 & 0 \end{pmatrix}\begin{pmatrix} C & V \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} C & V \\ 0 & 0 \end{pmatrix}\begin{pmatrix} B & U \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} BC - CB & BV - CU \\ 0 & 0 \end{pmatrix}.$$

We conclude by indicating the relationship between homomorphisms of Lie groups and homomorphisms of Lie algebras. First, we need to explain what is meant by a smooth map between manifolds.

Definition 1.18. Let M_1 (m_1-dimensional) and M_2 (m_2-dimensional) be manifolds in $\mathbb{R}^N$. A function $f\colon M_1 \to M_2$ is *smooth* if for every $p \in M_1$ there are parametrizations $\varphi\colon \Omega_1 \to U_1$ of M_1 at p and $\psi\colon \Omega_2 \to U_2$ of M_2 at $f(p)$ such that $f(U_1) \subseteq U_2$ and

$$\psi^{-1} \circ f \circ \varphi\colon \Omega_1 \to \mathbb{R}^{m_2}$$

is smooth.

Using Lemma 1.22, it is easily shown that Definition 1.18 does not depend on the choice of the parametrizations $\varphi\colon \Omega_1 \to U_1$ and $\psi\colon \Omega_2 \to U_2$. A smooth map f between manifolds is a *smooth diffeomorphism* if f is bijective and both f and f^{-1} are smooth maps.

We now define the derivative of a smooth map between manifolds.

Definition 1.19. Let M_1 (m_1-dimensional) and M_2 (m_2-dimensional) be manifolds in $\mathbb{R}^N$. For any smooth function $f\colon M_1 \to M_2$ and any $p \in M_1$, the function $f'_p\colon T_pM_1 \to T_{f(p)}M_2$, called the *tangent map of f at p, or derivative of f at p, or differential of f at p*, is defined as follows: For every $v \in T_pM_1$ and every smooth curve $\gamma\colon I \to M_1$ such that $\gamma(0) = p$ and $\gamma'(0) = v$,

$$f'_p(v) = (f \circ \gamma)'(0).$$

The map f'_p is also denoted by df_p or T_pf. Doing a few calculations involving the facts that

$$f \circ \gamma = (f \circ \varphi) \circ (\varphi^{-1} \circ \gamma) \quad \text{and} \quad \gamma = \varphi \circ (\varphi^{-1} \circ \gamma)$$

and using Lemma 1.22, it is not hard to show that $f'_p(v)$ does not depend on the choice of the curve γ. It is easily shown that f'_p is a linear map.

Finally, we define homomorphisms of Lie groups and Lie algebras and see how they are related.

Definition 1.20. Given two Lie groups G_1 and G_2, a *homomorphism (or map) of Lie groups* is a function $f\colon G_1 \to G_2$ that is a homomorphism of groups and a smooth map (between the manifolds G_1 and G_2). Given two Lie algebras $\mathcal{A}_1$ and $\mathcal{A}_2$, a *homomorphism (or map) of Lie algebras* is a function $f\colon \mathcal{A}_1 \to \mathcal{A}_2$ that is a linear map between the vector spaces $\mathcal{A}_1$ and $\mathcal{A}_2$ and that preserves Lie brackets, i.e.,

$$f([A, B]) = [f(A), f(B)]$$

for all $A, B \in \mathcal{A}_1$.

An *isomorphism of Lie groups* is a bijective function f such that both f and f^{-1} are maps of Lie groups, and an *isomorphism of Lie algebras* is a bijective function f such that both f and f^{-1} are maps of Lie algebras. It is immediately verified that if $f\colon G_1 \to G_2$ is a homomorphism of Lie groups, then $f'_I\colon \mathfrak{g}_1 \to \mathfrak{g}_2$ is a homomorphism of Lie algebras. If some additional assumptions are made about G_1 and G_2 (for example, connected, simply connected), it can be shown that f is pretty much determined by f'_I.

Alert readers must have noticed that we only defined the Lie algebra of a linear group. In the more general case, we can still define the Lie algebra $\mathfrak{g}$ of a Lie group G as the tangent space $T_I G$ at the identity I. The tangent space $\mathfrak{g} = T_I G$ is a vector space, but we need to define the Lie bracket. This can be done in several ways. We explain briefly how this can be done in terms of so-called adjoint representations. This has the advantage of not requiring the definition of left-invariant vector fields, but it is still a little bizarre!

Given a Lie group G, for every $a \in G$ we define *left translation* as the map $L_a\colon G \to G$ such that $L_a(b) = ab$ for all $b \in G$, and *right translation* as the map $R_a\colon G \to G$ such that $R_a(b) = ba$ for all $b \in G$. The maps L_a and R_a are diffeomorphisms, and their derivatives play an important role. The inner automorphisms $R_{a^{-1}} \circ L_a$ (also written as $R_{a^{-1}} L_a$) also play an important role. Note that

$$R_{a^{-1}} L_a(b) = aba^{-1}.$$

The derivative

$$(R_{a^{-1}} L_a)'_I\colon T_I G \to T_I G$$

of $R_{a^{-1}} L_a\colon G \to G$ at I is an isomorphism of Lie algebras, and since $T_I G = \mathfrak{g}$, we get a map denoted by $\mathrm{Ad}_a\colon \mathfrak{g} \to \mathfrak{g}$. The map $a \mapsto \mathrm{Ad}_a$ is a map of Lie groups

$$\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g}),$$

called the *adjoint representation of* G (where $\mathbf{GL}(\mathfrak{g})$ denotes the Lie group of all bijective linear maps on $\mathfrak{g}$).

In the case of a linear group, one can verify that

$$\mathrm{Ad}(a)(X) = \mathrm{Ad}_a(X) = aXa^{-1}$$

for all $a \in G$ and all $X \in \mathfrak{g}$. The derivative

$$\mathrm{Ad}'_I \colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})$$

of $\mathrm{Ad} \colon G \to \mathbf{GL}(\mathfrak{g})$ at I is map of Lie algebras, denoted by $\mathrm{ad} \colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})$, called the *adjoint representation of* $\mathfrak{g}$. (Recall that Theorem 1.28 immediately implies that the Lie algebra, $\mathfrak{gl}(\mathfrak{g})$, of $\mathbf{GL}(\mathfrak{g})$ is the vector space of all linear maps on $\mathfrak{g}$).

In the case of a linear group, it can be verified that

$$\mathrm{ad}(A)(B) = [A, B]$$

for all $A, B \in \mathfrak{g}$. One can also check that the Jacobi identity on $\mathfrak{g}$ is equivalent to the fact that ad preserves Lie brackets, i.e., ad is a map of Lie algebras:

$$\mathrm{ad}([A, B]) = [\mathrm{ad}(A), \mathrm{ad}(B)]$$

for all $A, B \in \mathfrak{g}$ (where on the right, the Lie bracket is the commutator of linear maps on $\mathfrak{g}$). Thus, we recover the Lie bracket from ad.

This is the key to the definition of the Lie bracket in the case of a general Lie group (not just a linear Lie group). We define the Lie bracket on $\mathfrak{g}$ as

$$[A, B] = \mathrm{ad}(A)(B).$$

To be complete, we have to define the exponential map $\exp \colon \mathfrak{g} \to G$ for a general Lie group. For this we need to introduce some left-invariant vector fields induced by the derivatives of the left translations, and integral curves associated with such vector fields. We will do this in Chapter 5 but for this we will need a deeper study of manifolds (see Chapter 3).

Readers who wish to learn more about Lie groups and Lie algebras should consult (more or less listed in order of difficulty) Curtis [39], Sattinger and Weaver [135], Hall [71] and Marsden and Ratiu [103]. The excellent lecture notes by Carter, Segal, and Macdonald [31] constitute a very efficient (although somewhat terse) introduction to Lie algebras and Lie groups. Classics such as Weyl [152] and Chevalley [34] are definitely worth consulting, although the presentation and the terminology may seem a bit old fashioned. For more advanced texts, one may consult Abraham and Marsden [1], Warner [148], Sternberg [144], Bröcker and tom Dieck [25], and Knapp [90]. For those who read French, Mneimné and Testard [112] is very clear and quite thorough, and uses very little differential geometry, although it is more advanced than Curtis. Chapter 1, by Bryant, in Freed and Uhlenbeck [26] is also worth reading, but the pace is fast.

Chapter 2

Review of Groups and Group Actions

2.1 Groups

Definition 2.1. A *group* is a set G equipped with a binary operation $\cdot\colon G \times G \to G$ that associates an element $a \cdot b \in G$ to every pair of elements $a, b \in G$, and having the following properties: $\cdot$ is associative, has an identity element $e \in G$, and every element in G is invertible (w.r.t. $\cdot$). More explicitly, this means that the following equations hold for all $a, b, c \in G$:

(G1) $a \cdot (b \cdot c) = (a \cdot b) \cdot c$. (associativity);

(G2) $a \cdot e = e \cdot a = a$. (identity);

(G3) For every $a \in G$, there is some $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e$ (inverse).

A group G is *abelian* (or *commutative*) if

$$a \cdot b = b \cdot a$$

for all $a, b \in G$.

A set M together with an operation $\cdot\colon M \times M \to M$ and an element e satisfying only conditions (G1) and (G2) is called a *monoid*. For example, the set $\mathbb{N} = \{0, 1, \ldots, n, \ldots\}$ of natural numbers is a (commutative) monoid under addition. However, it is not a group.

Some examples of groups are given below.

Example 2.1.

1. The set $\mathbb{Z} = \{\ldots, -n, \ldots, -1, 0, 1, \ldots, n, \ldots\}$ of integers is a group under addition, with identity element 0. However, $\mathbb{Z}^* = \mathbb{Z} - \{0\}$ is not a group under multiplication.

2. The set $\mathbb{Q}$ of rational numbers (fractions p/q with $p, q \in \mathbb{Z}$ and $q \neq 0$) is a group under addition, with identity element 0. The set $\mathbb{Q}^* = \mathbb{Q} - \{0\}$ is also a group under multiplication, with identity element 1.

3. Similarly, the sets $\mathbb{R}$ of real numbers and $\mathbb{C}$ of complex numbers are groups under addition (with identity element 0), and $\mathbb{R}^* = \mathbb{R} - \{0\}$ and $\mathbb{C}^* = \mathbb{C} - \{0\}$ are groups under multiplication (with identity element 1).

4. The sets $\mathbb{R}^n$ and $\mathbb{C}^n$ of n-tuples of real or complex numbers are groups under componentwise addition:

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n),$$

with identity element $(0, \ldots, 0)$. All these groups are abelian.

5. Given any nonempty set S, the set of bijections $f\colon S \to S$, also called *permutations of* S, is a group under function composition (i.e., the multiplication of f and g is the composition $g \circ f$), with identity element the identity function id_S. This group is not abelian as soon as S has more than two elements.

6. The set of $n \times n$ matrices with real (or complex) coefficients is a group under addition of matrices, with identity element the null matrix. It is denoted by $\mathrm{M}_n(\mathbb{R})$ (or $\mathrm{M}_n(\mathbb{C})$).

7. The set $\mathbb{R}[X]$ of all polynomials in one variable with real coefficients is a group under addition of polynomials.

8. The set of $n \times n$ invertible matrices with real (or complex) coefficients is a group under matrix multiplication, with identity element the identity matrix I_n. This group is called the *general linear group* and is usually denoted by $\mathbf{GL}(n, \mathbb{R})$ (or $\mathbf{GL}(n, \mathbb{C})$).

9. The set of $n \times n$ invertible matrices with real (or complex) coefficients and determinant $+1$ is a group under matrix multiplication, with identity element the identity matrix I_n. This group is called the *special linear group* and is usually denoted by $\mathbf{SL}(n, \mathbb{R})$ (or $\mathbf{SL}(n, \mathbb{C})$).

10. The set of $n \times n$ invertible matrices with real coefficients such that $RR^\top = I_n$ and of determinant $+1$ is a group called the *orthogonal group* and is usually denoted by $\mathbf{SO}(n)$ (where $R^\top$ is the *transpose* of the matrix R, i.e., the rows of $R^\top$ are the columns of R). It corresponds to the rotations in $\mathbb{R}^n$.

11. Given an open interval $]a, b[$, the set $\mathcal{C}(]a, b[)$ of continuous functions $f\colon]a, b[\to \mathbb{R}$ is a group under the operation $f + g$ defined such that

$$(f + g)(x) = f(x) + g(x)$$

for all $x \in]a, b[$.

It is customary to denote the operation of an abelian group G by $+$, in which case the inverse a^{-1} of an element $a \in G$ is denoted by $-a$.

The identity element of a group is *unique*. In fact, we can prove a more general fact:

Fact 1. If a binary operation $\cdot\colon M \times M \to M$ is associative and if $e' \in M$ is a left identity and $e'' \in M$ is a right identity, which means that

$$e' \cdot a = a \quad \text{for all} \quad a \in M \tag{G2l}$$

and

$$a \cdot e'' = a \quad \text{for all} \quad a \in M, \tag{G2r}$$

then $e' = e''$.

Proof. If we let $a = e''$ in equation (G2l), we get

$$e' \cdot e'' = e'',$$

and if we let $a = e'$ in equation (G2r), we get

$$e' \cdot e'' = e',$$

and thus

$$e' = e' \cdot e'' = e'',$$

as claimed. □

Fact 1 implies that the identity element of a monoid is unique, and since every group is a monoid, the identity element of a group is unique. Furthermore, every element in a group has a *unique inverse*. This is a consequence of a slightly more general fact:

Fact 2. In a monoid M with identity element e, if some element $a \in M$ has some left inverse $a' \in M$ and some right inverse $a'' \in M$, which means that

$$a' \cdot a = e \tag{G3l}$$

and

$$a \cdot a'' = e, \tag{G3r}$$

then $a' = a''$.

Proof. Using (G3l) and the fact that e is an identity element, we have

$$(a' \cdot a) \cdot a'' = e \cdot a'' = a''.$$

Similarly, Using (G3r) and the fact that e is an identity element, we have

$$a' \cdot (a \cdot a'') = a' \cdot e = a'.$$

However, since M is monoid, the operation $\cdot$ is associative, so

$$a' = a' \cdot (a \cdot a'') = (a' \cdot a) \cdot a'' = a'',$$

as claimed. □

Remark: Axioms (G2) and (G3) can be weakened a bit by requiring only (G2r) (the existence of a right identity) and (G3r) (the existence of a right inverse for every element) (or (G2l) and (G3l)). It is a good exercise to prove that the group axioms (G2) and (G3) follow from (G2r) and (G3r).

Given a group, G, for any two subsets $R, S \subseteq G$, we let

$$RS = \{r \cdot s \mid r \in R,\ s \in S\}.$$

In particular, for any $g \in G$, if $R = \{g\}$, we write

$$gS = \{g \cdot s \mid s \in S\}$$

and similarly, if $S = \{g\}$, we write

$$Rg = \{r \cdot g \mid r \in R\}.$$

From now on, we will drop the multiplication sign and write g_1g_2 for $g_1 \cdot g_2$.

Definition 2.2. Given a group, G, a subset, H, of G is a *subgroup of* G iff

(1) The identity element, e, of G also belongs to H ($e \in H$);

(2) For all $h_1, h_2 \in H$, we have $h_1h_2 \in H$;

(3) For all $h \in H$, we have $h^{-1} \in H$.

It is easily checked that a subset, $H \subseteq G$, is a subgroup of G iff H is nonempty and whenever $h_1, h_2 \in H$, then $h_1h_2^{-1} \in H$.

If H is a subgroup of G and $g \in G$ is any element, the sets of the form gH are called *left cosets of H in G* and the sets of the form Hg are called *right cosets of H in G*. The left cosets (resp. right cosets) of H induce an equivalence relation, $\sim$, defined as follows: For all $g_1, g_2 \in G$,

$$g_1 \sim g_2 \quad \text{iff} \quad g_1H = g_2H$$

(resp. $g_1 \sim g_2$ iff $Hg_1 = Hg_2$).

Obviously, $\sim$ is an equivalence relation. Now, it is easy to see that $g_1H = g_2H$ iff $g_2^{-1}g_1 \in H$, so the equivalence class of an element $g \in G$ is the coset gH (resp. Hg). The set of left cosets of H in G (which, in general, is **not** a group) is denoted G/H. The "points" of G/H are obtained by "collapsing" all the elements in a coset into a single element.

It is tempting to define a multiplication operation on left cosets (or right cosets) by setting

$$(g_1H)(g_2H) = (g_1g_2)H,$$

but this operation is not well defined in general, unless the subgroup H possesses a special property. This property is typical of the kernels of group homomorphisms, so we are led to

Definition 2.3. Given any two groups, G, G', a function $\varphi\colon G \to G'$ is a *homomorphism* iff

$$\varphi(g_1 g_2) = \varphi(g_1)\varphi(g_2), \quad \text{for all } g_1, g_2 \in G.$$

Taking $g_1 = g_2 = e$ (in G), we see that

$$\varphi(e) = e',$$

and taking $g_1 = g$ and $g_2 = g^{-1}$, we see that

$$\varphi(g^{-1}) = \varphi(g)^{-1}.$$

If $\varphi\colon G \to G'$ and $\psi\colon G' \to G''$ are group homomorphisms, then $\psi \circ \varphi\colon G \to G''$ is also a homomorphism. If $\varphi\colon G \to G'$ is a homomorphism of groups and $H \subseteq G$ and $H' \subseteq G'$ are two subgroups, then it is easily checked that

$$\text{Im } H = \varphi(H) = \{\varphi(g) \mid g \in H\} \quad \text{is a subgroup of } G'$$

(Im H is called the *image of H by φ*) and

$$\varphi^{-1}(H') = \{g \in G \mid \varphi(g) \in H'\} \quad \text{is a subgroup of } G.$$

In particular, when $H' = \{e'\}$, we obtain the *kernel*, Ker φ, of φ. Thus,

$$\text{Ker } \varphi = \{g \in G \mid \varphi(g) = e'\}.$$

It is immediately verified that $\varphi\colon G \to G'$ is injective iff Ker $\varphi = \{e\}$. (We also write Ker $\varphi = (0)$.) We say that φ is an *isomorphism* if there is a homomorphism, $\psi\colon G' \to G$, so that

$$\psi \circ \varphi = \text{id}_G \quad \text{and} \quad \varphi \circ \psi = \text{id}_{G'}.$$

In this case, ψ is unique and it is denoted φ^{-1}. When φ is an isomorphism we say the the groups G and G' are *isomorphic*. When $G' = G$, a group isomorphism is called an *automorphism*.

We claim that $H = \text{Ker } \varphi$ satisfies the following property:

$$gH = Hg, \quad \text{for all } g \in G. \tag{$*$}$$

First, note that $(*)$ is equivalent to

$$gHg^{-1} = H, \quad \text{for all } g \in G,$$

and the above is equivalent to

$$gHg^{-1} \subseteq H, \quad \text{for all } g \in G. \tag{$**$}$$

This is because $gHg^{-1} \subseteq H$ implies $H \subseteq g^{-1}Hg$, and this for all $g \in G$. But,

$$\varphi(ghg^{-1}) = \varphi(g)\varphi(h)\varphi(g^{-1}) = \varphi(g)e'\varphi(g)^{-1} = \varphi(g)\varphi(g)^{-1} = e',$$

for all $h \in H = \text{Ker } \varphi$ and all $g \in G$. Thus, by definition of $H = \text{Ker } \varphi$, we have $gHg^{-1} \subseteq H$.

Definition 2.4. For any group, G, a subgroup, $N \subseteq G$, is a *normal subgroup* of G iff

$$gNg^{-1} = N, \quad \text{for all } g \in G.$$

This is denoted by $N \lhd G$.

If N is a normal subgroup of G, the equivalence relation induced by left cosets is the same as the equivalence induced by right cosets. Furthermore, this equivalence relation, $\sim$, is a *congruence*, which means that: For all $g_1, g_2, g_1', g_2' \in G$,

(1) If $g_1N = g_1'N$ and $g_2N = g_2'N$, then $g_1g_2N = g_1'g_2'N$, and

(2) If $g_1N = g_2N$, then $g_1^{-1}N = g_2^{-1}N$.

As a consequence, we can define a group structure on the set $G/\sim$ of equivalence classes modulo $\sim$, by setting

$$(g_1N)(g_2N) = (g_1g_2)N.$$

This group is denoted G/N. The equivalence class, gN, of an element $g \in G$ is also denoted $\overline{g}$. The map $\pi\colon G \to G/N$, given by

$$\pi(g) = \overline{g} = gN,$$

is clearly a group homomorphism called the *canonical projection*.

Given a homomorphism of groups, $\varphi\colon G \to G'$, we easily check that the groups $G/\mathrm{Ker}\ \varphi$ and $\mathrm{Im}\ \varphi = \varphi(G)$ are isomorphic.

2.2 Group Actions and Homogeneous Spaces, I

If X is a set (usually, some kind of geometric space, for example, the sphere in $\mathbb{R}^3$, the upper half-plane, etc.), the "symmetries" of X are often captured by the action of a group, G, on X. In fact, if G is a Lie group and the action satisfies some simple properties, the set X can be given a manifold structure which makes it a projection (quotient) of G, a so-called "homogeneous space".

Definition 2.5. Given a set, X, and a group, G, a *left action of G on X* (for short, an *action of G on X*) is a function, $\varphi\colon G \times X \to X$, such that

(1) For all $g, h \in G$ and all $x \in X$,

$$\varphi(g, \varphi(h, x)) = \varphi(gh, x),$$

(2) For all $x \in X$,

$$\varphi(1, x) = x,$$

where $1 \in G$ is the identity element of G.

To alleviate the notation, we usually write $g \cdot x$ or even gx for $\varphi(g, x)$, in which case, the above axioms read:

(1) For all $g, h \in G$ and all $x \in X$,

$$g \cdot (h \cdot x) = gh \cdot x,$$

(2) For all $x \in X$,

$$1 \cdot x = x.$$

The set X is called a *(left) G-set*. The action φ is *faithful* or *effective* iff for every g, if $g \cdot x = x$ for all $x \in X$, then $g = 1$; the action φ is *transitive* iff for any two elements $x, y \in X$, there is some $g \in G$ so that $g \cdot x = y$.

Given an action, $\varphi\colon G \times X \to X$, for every $g \in G$, we have a function, $\varphi_g\colon X \to X$, defined by

$$\varphi_g(x) = g \cdot x, \quad \text{for all } x \in X.$$

Observe that φ_g has $\varphi_{g^{-1}}$ as inverse, since

$$\varphi_{g^{-1}}(\varphi_g(x)) = \varphi_{g^{-1}}(g \cdot x) = g^{-1} \cdot (g \cdot x) = (g^{-1}g) \cdot x = 1 \cdot x = x,$$

and similarly, $\varphi_g \circ \varphi_{g^{-1}} = \mathrm{id}$. Therefore, φ_g is a bijection of X, i.e., a permutation of X. Moreover, we check immediately that

$$\varphi_g \circ \varphi_h = \varphi_{gh},$$

so, the map $g \mapsto \varphi_g$ is a group homomorphism from G to $\mathfrak{S}_X$, the group of permutations of X. With a slight abuse of notation, this group homomorphism $G \longrightarrow \mathfrak{S}_X$ is also denoted φ.

Conversely, it is easy to see that any group homomorphism, $\varphi\colon G \to \mathfrak{S}_X$, yields a group action, $\cdot\colon G \times X \longrightarrow X$, by setting

$$g \cdot x = \varphi(g)(x).$$

Observe that an action, φ, is faithful iff the group homomorphism, $\varphi\colon G \to \mathfrak{S}_X$, is injective. Also, we have $g \cdot x = y$ iff $g^{-1} \cdot y = x$, since $(gh) \cdot x = g \cdot (h \cdot x)$ and $1 \cdot x = x$, for all $g, h \in G$ and all $x \in X$.

Definition 2.6. Given two G-sets, X and Y, a function, $f\colon X \to Y$, is said to be *equivariant*, or a *G-map* iff for all $x \in X$ and all $g \in G$, we have

$$f(g \cdot x) = g \cdot f(x).$$

Remark: We can also define a *right action*, $\cdot\colon X \times G \to X$, of a group G on a set X, as a map satisfying the conditions

(1) For all $g, h \in G$ and all $x \in X$,

$$(x \cdot g) \cdot h = x \cdot gh,$$

(2) For all $x \in X$,

$$x \cdot 1 = x.$$

Every notion defined for left actions is also defined for right actions, in the obvious way.

Here are some examples of (left) group actions.

Example 1: The unit sphere S^2 (more generally, S^{n-1}).

Recall that for any $n \geq 1$, the *(real) unit sphere*, S^{n-1}, is the set of points in $\mathbb{R}^n$ given by

$$S^{n-1} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + \cdots + x_n^2 = 1\}.$$

In particular, S^2 is the usual sphere in $\mathbb{R}^3$. Since the group $\mathbf{SO}(3) = \mathbf{SO}(3, \mathbb{R})$ consists of (orientation preserving) linear isometries, i.e., *linear* maps that are distance preserving (and of determinant $+1$), and every linear map leaves the origin fixed, we see that any rotation maps S^2 into itself.

Beware that this would be false if we considered the group of *affine* isometries, $\mathbf{SE}(3)$, of $\mathbb{E}^3$. For example, a screw motion does *not* map S^2 into itself, even though it is distance preserving, because the origin is translated.

Thus, we have an action, $\cdot \colon \mathbf{SO}(3) \times S^2 \to S^2$, given by

$$R \cdot x = Rx.$$

The verification that the above is indeed an action is trivial. This action is transitive. This is because, for any two points x, y on the sphere S^2, there is a rotation whose axis is perpendicular to the plane containing x, y and the center, O, of the sphere (this plane is not unique when x and y are antipodal, i.e., on a diameter) mapping x to y.

Similarly, for any $n \geq 1$, we get an action, $\cdot \colon \mathbf{SO}(n) \times S^{n-1} \to S^{n-1}$. It is easy to show that this action is transitive.

Analogously, we can define the *(complex) unit sphere*, Σ^{n-1}, as the set of points in $\mathbb{C}^n$ given by

$$\Sigma^{n-1} = \{(z_1, \ldots, z_n) \in \mathbb{C}^n \mid z_1\overline{z}_1 + \cdots + z_n\overline{z}_n = 1\}.$$

If we write $z_j = x_j + iy_j$, with $x_j, y_j \in \mathbb{R}$, then

$$\Sigma^{n-1} = \{(x_1, \ldots, x_n, y_1, \ldots, y_n) \in \mathbb{R}^{2n} \mid x_1^2 + \cdots + x_n^2 + y_1^2 + \cdots + y_n^2 = 1\}.$$

Therefore, we can view the complex sphere, Σ^{n-1} (in $\mathbb{C}^n$), as the real sphere, S^{2n-1} (in $\mathbb{R}^{2n}$). By analogy with the real case, we can define an action, $\cdot \colon \mathbf{SU}(n) \times \Sigma^{n-1} \to \Sigma^{n-1}$, of the group, $\mathbf{SU}(n)$, of *linear* maps of $\mathbb{C}^n$ preserving the hermitian inner product (and the origin, as all linear maps do) and this action is transitive.

One should not confuse the unit sphere, Σ^{n-1}, with the hypersurface, $S^{n-1}_{\mathbb{C}}$, given by

$$S^{n-1}_{\mathbb{C}} = \{(z_1, \ldots, z_n) \in \mathbb{C}^n \mid z_1^2 + \cdots + z_n^2 = 1\}.$$

For instance, one should check that a line, L, through the origin intersects Σ^{n-1} in a circle, whereas it intersects $S^{n-1}_{\mathbb{C}}$ in exactly two points!

Example 2: The upper half-plane.

The *upper half-plane*, H, is the open subset of $\mathbb{R}^2$ consisting of all points, $(x, y) \in \mathbb{R}^2$, with $y > 0$. It is convenient to identify H with the set of complex numbers, $z \in \mathbb{C}$, such that $\Im z > 0$. Then, we can define an action, $\cdot\colon \mathbf{SL}(2, \mathbb{R}) \times H \to H$, of the group $\mathbf{SL}(2, \mathbb{R})$ on H, as follows: For any $z \in H$, for any $A \in \mathbf{SL}(2, \mathbb{R})$,

$$A \cdot z = \frac{az + b}{cz + d},$$

where

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

with $ad - bc = 1$. It is easily verified that $A \cdot z$ is indeed always well defined and in H when $z \in H$. This action is transitive (check this).

Maps of the form

$$z \mapsto \frac{az + b}{cz + d},$$

where $z \in \mathbb{C}$ and $ad - bc = 1$, are called *Möbius transformations*. Here, $a, b, c, d \in \mathbb{R}$, but in general, we allow $a, b, c, d \in \mathbb{C}$. Actually, these transformations are not necessarily defined everywhere on $\mathbb{C}$, for example, for $z = -d/c$ if $c \neq 0$. To fix this problem, we add a "point at infinity", ∞, to $\mathbb{C}$ and define Möbius transformations as functions $\mathbb{C} \cup \{\infty\} \longrightarrow \mathbb{C} \cup \{\infty\}$. If $c = 0$, the Möbius transformation sends ∞ to itself, otherwise, $-d/c \mapsto \infty$ and $\infty \mapsto a/c$. The space $\mathbb{C} \cup \{\infty\}$ can be viewed as the plane, $\mathbb{R}^2$, extended with a point at infinity. Using a stereographic projection from the sphere S^2 to the plane, (say from the north pole to the equatorial plane), we see that there is a bijection between the sphere, S^2, and $\mathbb{C} \cup \{\infty\}$. More precisely, the *stereographic projection* of the sphere S^2 from the north pole, $N = (0, 0, 1)$, to the plane $z = 0$ (extended with the point at infinity, ∞) is given by

$$(x, y, z) \in S^2 - \{(0, 0, 1)\} \mapsto \left(\frac{x}{1 - z}, \frac{y}{1 - z}\right) = \frac{x + iy}{1 - z} \in \mathbb{C}, \quad \text{with} \quad (0, 0, 1) \mapsto \infty.$$

The inverse stereographic projection is given by

$$(x, y) \mapsto \left(\frac{2x}{x^2 + y^2 + 1}, \frac{2y}{x^2 + y^2 + 1}, \frac{x^2 + y^2 - 1}{x^2 + y^2 + 1}\right), \quad \text{with} \quad \infty \mapsto (0, 0, 1).$$

Intuitively, the inverse stereographic projection "wraps" the equatorial plane around the sphere. The space $\mathbb{C} \cup \{\infty\}$ is known as the *Riemann sphere*. We will see shortly that

$\mathbb{C} \cup \{\infty\} \cong S^2$ is also the complex projective line, $\mathbb{CP}^1$. In summary, Möbius transformations are bijections of the Riemann sphere. It is easy to check that these transformations form a group under composition for all $a, b, c, d \in \mathbb{C}$, with $ad - bc = 1$. This is the *Möbius group*, denoted $\mathbf{Möb}^+$. The Möbius transformations corresponding to the case $a, b, c, d \in \mathbb{R}$, with $ad - bc = 1$ form a subgroup of $\mathbf{Möb}^+$ denoted $\mathbf{Möb}^+_{\mathbb{R}}$. The map from $\mathbf{SL}(2, \mathbb{C})$ to $\mathbf{Möb}^+$ that sends $A \in \mathbf{SL}(2, \mathbb{C})$ to the corresponding Möbius transformation is a surjective group homomorphism and one checks easily that its kernel is $\{-I, I\}$ (where I is the 2×2 identity matrix). Therefore, the Möbius group $\mathbf{Möb}^+$ is isomorphic to the quotient group $\mathbf{SL}(2, \mathbb{C})/\{-I, I\}$, denoted $\mathbf{PSL}(2, \mathbb{C})$. This latter group turns out to be the group of projective transformations of the projective space $\mathbb{CP}^1$. The same reasoning shows that the subgroup $\mathbf{Möb}^+_{\mathbb{R}}$ is isomorphic to $\mathbf{SL}(2, \mathbb{R})/\{-I, I\}$, denoted $\mathbf{PSL}(2, \mathbb{R})$.

The group $\mathbf{SL}(2, \mathbb{C})$ acts on $\mathbb{C} \cup \{\infty\} \cong S^2$ the same way that $\mathbf{SL}(2, \mathbb{R})$ acts on H, namely: For any $A \in \mathbf{SL}(2, \mathbb{C})$, for any $z \in \mathbb{C} \cup \{\infty\}$,

$$A \cdot z = \frac{az + b}{cz + d},$$

where

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{with} \quad ad - bc = 1.$$

This action is clearly transitive.

One may recall from complex analysis that the (complex) Möbius transformation

$$z \mapsto \frac{z - i}{z + i}$$

is a biholomorphic isomorphism between the upper half plane, H, and the open unit disk,

$$D = \{z \in \mathbb{C} \mid |z| < 1\}.$$

As a consequence, it is possible to define a transitive action of $\mathbf{SL}(2, \mathbb{R})$ on D. This can be done in a more direct fashion, using a group isomorphic to $\mathbf{SL}(2, \mathbb{R})$, namely, $\mathbf{SU}(1, 1)$ (a group of complex matrices), but we don't want to do this right now.

Example 3: The set of $n \times n$ symmetric, positive, definite matrices, $\mathbf{SPD}(n)$.

The group $\mathbf{GL}(n) = \mathbf{GL}(n, \mathbb{R})$ acts on $\mathbf{SPD}(n)$ as follows: For all $A \in \mathbf{GL}(n)$ and all $S \in \mathbf{SPD}(n)$,

$$A \cdot S = ASA^\top.$$

It is easily checked that $ASA^\top$ is in $\mathbf{SPD}(n)$ if S is in $\mathbf{SPD}(n)$. This action is transitive because every SPD matrix, S, can be written as $S = AA^\top$, for some invertible matrix, A (prove this as an exercise).

Example 4: The projective spaces $\mathbb{RP}^n$ and $\mathbb{CP}^n$.

The *(real) projective space*, $\mathbb{RP}^n$, is the set of all lines through the origin in $\mathbb{R}^{n+1}$, i.e., the set of one-dimensional subspaces of $\mathbb{R}^{n+1}$ (where $n \geq 0$). Since a one-dimensional subspace, $L \subseteq \mathbb{R}^{n+1}$, is spanned by any nonzero vector, $u \in L$, we can view $\mathbb{RP}^n$ as the set of equivalence classes of nonzero vectors in $\mathbb{R}^{n+1} - \{0\}$ modulo the equivalence relation,

$$u \sim v \quad \text{iff} \quad v = \lambda u, \quad \text{for some} \quad \lambda \in \mathbb{R},\ \lambda \neq 0.$$

In terms of this definition, there is a projection, $pr\colon (\mathbb{R}^{n+1} - \{0\}) \to \mathbb{RP}^n$, given by $pr(u) = [u]_\sim$, the equivalence class of u modulo $\sim$. Write $[u]$ for the line defined by the nonzero vector, u. Since every line, L, in $\mathbb{R}^{n+1}$ intersects the sphere S^n in two antipodal points, we can view $\mathbb{RP}^n$ as the quotient of the sphere S^n by identification of antipodal points. We write

$$S^n/\{I, -I\} \cong \mathbb{RP}^n.$$

We define an action of $\mathbf{SO}(n+1)$ on $\mathbb{RP}^n$ as follows: For any line, $L = [u]$, for any $R \in \mathbf{SO}(n+1)$,

$$R \cdot L = [Ru].$$

Since R is linear, the line $[Ru]$ is well defined, i.e., does not depend on the choice of $u \in L$. It is clear that this action is transitive.

The *(complex) projective space*, $\mathbb{CP}^n$, is defined analogously as the set of all lines through the origin in $\mathbb{C}^{n+1}$, i.e., the set of one-dimensional subspaces of $\mathbb{C}^{n+1}$ (where $n \geq 0$). This time, we can view $\mathbb{CP}^n$ as the set of equivalence classes of vectors in $\mathbb{C}^{n+1} - \{0\}$ modulo the equivalence relation,

$$u \sim v \quad \text{iff} \quad v = \lambda u, \quad \text{for some} \quad \lambda\ \neq 0 \in \mathbb{C}.$$

We have the projection, $pr\colon \mathbb{C}^{n+1} - \{0\} \to \mathbb{CP}^n$, given by $pr(u) = [u]_\sim$, the equivalence class of u modulo $\sim$. Again, write $[u]$ for the line defined by the nonzero vector, u.

Remark: Algebraic geometers write $\mathbb{P}^n_{\mathbb{R}}$ for $\mathbb{RP}^n$ and $\mathbb{P}^n_{\mathbb{C}}$ (or even $\mathbb{P}^n$) for $\mathbb{CP}^n$.

Recall that $\Sigma^n \subseteq \mathbb{C}^{n+1}$, the unit sphere in $\mathbb{C}^{n+1}$, is defined by

$$\Sigma^n = \{(z_1, \ldots, z_{n+1}) \in \mathbb{C}^{n+1} \mid z_1\overline{z}_1 + \cdots + z_{n+1}\overline{z}_{n+1} = 1\}.$$

For any line, $L = [u]$, where $u \in \mathbb{C}^{n+1}$ is a nonzero vector, writing $u = (u_1, \ldots, u_{n+1})$, a point $z \in \mathbb{C}^{n+1}$ belongs to L iff $z = \lambda(u_1, \ldots, u_{n+1})$, for some $\lambda \in \mathbb{C}$. Therefore, the intersection, $L \cap \Sigma^n$, of the line L and the sphere Σ^n is given by

$$L \cap \Sigma^n = \{\lambda(u_1, \ldots, u_{n+1}) \in \mathbb{C}^{n+1} \mid \lambda \in \mathbb{C},\ \lambda\overline{\lambda}(u_1\overline{u}_1 + \cdots + u_{n+1}\overline{u}_{n+1}) = 1\},$$

i.e.,

$$L \cap \Sigma^n = \left\{ \lambda(u_1, \ldots, u_{n+1}) \in \mathbb{C}^{n+1} \;\middle|\; \lambda \in \mathbb{C},\ |\lambda| = \frac{1}{\sqrt{|u_1|^2 + \cdots + |u_{n+1}|^2}} \right\}.$$

Thus, we see that there is a bijection between $L \cap \Sigma^n$ and the circle, S^1, i.e., geometrically, $L \cap \Sigma^n$ is a circle. Moreover, since any line, L, through the origin is determined by just one other point, we see that for any two lines L_1 and L_2 through the origin,

$$L_1 \neq L_2 \quad \text{iff} \quad (L_1 \cap \Sigma^n) \cap (L_2 \cap \Sigma^n) = \emptyset.$$

However, Σ^n is the sphere S^{2n+1} in $\mathbb{R}^{2n+2}$. It follows that $\mathbb{CP}^n$ is the quotient of S^{2n+1} by the equivalence relation, $\sim$, defined such that

$$y \sim z \quad \text{iff} \quad y, z \in L \cap \Sigma^n, \quad \text{for some line, } L, \text{ through the origin.}$$

Therefore, we can write

$$S^{2n+1}/S^1 \cong \mathbb{CP}^n.$$

Observe that $\mathbb{CP}^n$ can also be viewed as the orbit space of the action, $\cdot\colon S^1 \times S^{2n+1} \to S^{2n+1}$, given by

$$\lambda \cdot (z_1, \ldots, z_{n+1}) = (\lambda z_1, \ldots, \lambda z_{n+1}),$$

where $S^1 = \mathbf{U}(1)$ (the group of complex numbers of modulus 1) and S^{2n+1} is identified with Σ^n. The case $n = 1$ is particularly interesting, as it turns out that

$$S^3/S^1 \cong S^2.$$

This is the famous *Hopf fibration*. To show this, proceed as follows: As

$$S^3 \cong \Sigma^1 = \{(z, z') \in \mathbb{C}^2 \mid |z|^2 + |z'|^2 = 1\},$$

define a map, $\mathrm{HF}\colon S^3 \to S^2$, by

$$\mathrm{HF}((z, z')) = (2z\overline{z'}, |z|^2 - |z'|^2).$$

We leave as a homework exercise to prove that this map has range S^2 and that

$$\mathrm{HF}((z_1, z_1')) = \mathrm{HF}((z_2, z_2')) \quad \text{iff} \quad (z_1, z_1') = \lambda(z_2, z_2'), \quad \text{for some } \lambda \text{ with } |\lambda| = 1.$$

In other words, for any point, $p \in S^2$, the inverse image, $\mathrm{HF}^{-1}(p)$ (also called *fibre* over p), is a circle on S^3. Consequently, S^3 can be viewed as the union of a family of disjoint circles. This is the *Hopf fibration*. It is possible to visualize the Hopf fibration using the stereographic projection from S^3 onto $\mathbb{R}^3$. This is a beautiful and puzzling picture. For example, see Berger [15]. Therefore, HF induces a bijection from $\mathbb{CP}^1$ to S^2, and it is a homeomorphism.

We define an action of $\mathbf{SU}(n+1)$ on $\mathbb{CP}^n$ as follows: For any line, $L = [u]$, for any $R \in \mathbf{SU}(n+1)$,

$$R \cdot L = [Ru].$$

Again, this action is well defined and it is transitive.

Example 5: Affine spaces.

If E is any (real) vector space and X is any set, a transitive and faithful action, $\cdot\colon E \times X \to X$, of the additive group of E on X makes X into an *affine space*. The intuition is that the members of E are translations.

Those familiar with affine spaces as in Gallier [60] (Chapter 2) or Berger [15] will point out that if X is an affine space, then, not only is the action of E on X transitive, but more is true: For any two points, $a, b \in E$, there is a *unique* vector, $u \in E$, such that $u \cdot a = b$. By the way, the action of E on X is usually considered to be a right action and is written additively, so $u \cdot a$ is written $a + u$ (the result of translating a by u). Thus, it would seem that we have to require more of our action. However, this is not necessary because E (under addition) is *abelian*. More precisely, we have the proposition

Proposition 2.1. *If G is an abelian group acting on a set X and the action $\cdot\colon G \times X \to X$ is transitive and faithful, then for any two elements $x, y \in X$, there is a unique $g \in G$ so that $g \cdot x = y$ (the action is simply transitive).*

Proof. Since our action is transitive, there is at least some $g \in G$ so that $g \cdot x = y$. Assume that we have $g_1, g_2 \in G$ with

$$g_1 \cdot x = g_2 \cdot x = y.$$

We shall prove that, actually,

$$g_1 \cdot z = g_2 \cdot z, \quad \text{for all } z \in X.$$

As our action is faithful we must have $g_1 = g_2$, and this proves our proposition.

Pick any $z \in X$. As our action is transitive, there is some $h \in G$ so that $z = h \cdot x$. Then, we have

$$\begin{aligned}
g_1 \cdot z &= g_1 \cdot (h \cdot x) \\
&= (g_1 h) \cdot x \\
&= (h g_1) \cdot x \qquad \text{(since } G \text{ is abelian)} \\
&= h \cdot (g_1 \cdot x) \\
&= h \cdot (g_2 \cdot x) \qquad \text{(since } \quad g_1 \cdot x = g_2 \cdot x) \\
&= (h g_2) \cdot x \\
&= (g_2 h) \cdot x \qquad \text{(since } G \text{ is abelian)} \\
&= g_2 \cdot (h \cdot x) \\
&= g_2 \cdot z.
\end{aligned}$$

Therefore, $g_1 \cdot z = g_2 \cdot z$, for all $z \in X$, as claimed. □

More examples will be considered later.

The subset of group elements that leave some given element $x \in X$ fixed plays an important role.

Definition 2.7. Given an action, $\cdot\colon G \times X \to X$, of a group G on a set X, for any $x \in X$, the group G_x (also denoted $\mathrm{Stab}_G(x)$), called the *stabilizer* of x or *isotropy group at* x is given by

$$G_x = \{g \in G \mid g \cdot x = x\}.$$

We have to verify that G_x is indeed a subgroup of G, but this is easy. Indeed, if $g \cdot x = x$ and $h \cdot x = x$, then we also have $h^{-1} \cdot x = x$ and so, we get $gh^{-1} \cdot x = x$, proving that G_x is a subgroup of G. In general, G_x is **not** a normal subgroup.

Observe that

$$G_{g\cdot x} = gG_xg^{-1},$$

for all $g \in G$ and all $x \in X$.

Indeed,

$$\begin{aligned} G_{g\cdot x} &= \{h \in G \mid h \cdot (g \cdot x) = g \cdot x\} \\ &= \{h \in G \mid hg \cdot x = g \cdot x\} \\ &= \{h \in G \mid g^{-1}hg \cdot x = x\} \\ &= gG_xg^{-1}. \end{aligned}$$

Therefore, the stabilizers of x and $g \cdot x$ are conjugate of each other.

When the action of G on X is transitive, for any fixed $x \in G$, the set X is a quotient (as set, not as group) of G by G_x. Indeed, we can define the map, $\pi_x\colon G \to X$, by

$$\pi_x(g) = g \cdot x, \quad \text{for all } g \in G.$$

Observe that

$$\pi_x(gG_x) = (gG_x) \cdot x = g \cdot (G_x \cdot x) = g \cdot x = \pi_x(g).$$

This shows that $\pi_x\colon G \to X$ induces a quotient map, $\overline{\pi}_x\colon G/G_x \to X$, from the set, G/G_x, of (left) cosets of G_x to X, defined by

$$\overline{\pi}_x(gG_x) = g \cdot x.$$

Since

$$\pi_x(g) = \pi_x(h) \quad \text{iff} \quad g \cdot x = h \cdot x \quad \text{iff} \quad g^{-1}h \cdot x = x \quad \text{iff} \quad g^{-1}h \in G_x \quad \text{iff} \quad gG_x = hG_x,$$

we deduce that $\overline{\pi}_x\colon G/G_x \to X$ is injective. However, since our action is transitive, for every $y \in X$, there is some $g \in G$ so that $g \cdot x = y$ and so, $\overline{\pi}_x(gG_x) = g \cdot x = y$, i.e., the map $\overline{\pi}_x$ is also surjective. Therefore, the map $\overline{\pi}_x\colon G/G_x \to X$ is a bijection (of sets, not groups). The map $\pi_x\colon G \to X$ is also surjective. Let us record this important fact as

Proposition 2.2. *If $\cdot\colon G \times X \to X$ is a transitive action of a group G on a set X, for every fixed $x \in X$, the surjection, $\pi\colon G \to X$, given by*

$$\pi(g) = g \cdot x$$

induces a bijection

$$\overline{\pi}\colon G/G_x \to X,$$

where G_x is the stabilizer of x.

The map $\pi\colon G \to X$ (corresponding to a fixed $x \in X$) is sometimes called a *projection* of G onto X. Proposition 2.2 shows that for every $y \in X$, the subset, $\pi^{-1}(y)$, of G (called the *fibre above* y) is equal to some coset, gG_x, of G and thus, is in bijection with the group G_x itself. We can think of G as a moving family of fibres, G_x, parametrized by X. This point of view of viewing a space as a moving family of simpler spaces is typical in (algebraic) geometry, and underlies the notion of (principal) fibre bundle.

Note that if the action $\cdot\colon G \times X \to X$ is transitive, then the stabilizers G_x and G_y of any two elements $x, y \in X$ are isomorphic, as they as conjugates. Thus, in this case, it is enough to compute one of these stabilizers for a "convenient" x.

As the situation of Proposition 2.2 is of particular interest, we make the following definition:

Definition 2.8. A set, X, is said to be a *homogeneous space* if there is a transitive action, $\cdot\colon G \times X \to X$, of some group, G, on X.

We see that all the spaces of Example 1–5 are homogeneous spaces. Another example that will play an important role when we deal with Lie groups is the situation where we have a group, G, a subgroup, H, of G (not necessarily normal) and where $X = G/H$, the set of left cosets of G modulo H. The group G acts on G/H by left multiplication:

$$a \cdot (gH) = (ag)H,$$

where $a, g \in G$. This action is clearly transitive and one checks that the stabilizer of gH is gHg^{-1}. If G is a topological group and H is a closed subgroup of G (see later for an explanation), it turns out that G/H is Hausdorff (Recall that a topological space, X, is *Hausdorff* iff for any two distinct points $x \neq y \in X$, there exists two disjoint open subsets, U and V, with $x \in U$ and $y \in V$.) If G is a Lie group, we obtain a manifold.

Even if G and X are topological spaces and the action, $\cdot\colon G \times X \to X$, is continuous, the space G/G_x under the quotient topology is, in general, **not** homeomorphic to X.

We will give later sufficient conditions that insure that X is indeed a topological space or even a manifold. In particular, X will be a manifold when G is a Lie group.

In general, an action $\cdot\colon G \times X \to X$ is not transitive on X, but for every $x \in X$, it is transitive on the set

$$O(x) = G \cdot x = \{g \cdot x \mid g \in G\}.$$

Such a set is called the *orbit* of x. The orbits are the equivalence classes of the following equivalence relation:

Definition 2.9. Given an action, $\cdot\colon G \times X \to X$, of some group, G, on X, the equivalence relation, $\sim$, on X is defined so that, for all $x, y \in X$,

$$x \sim y \quad \text{iff} \quad y = g \cdot x, \quad \text{for some } g \in G.$$

For every $x \in X$, the equivalence class of x is the *orbit of* x, denoted $O(x)$ or $\mathrm{Orb}_G(x)$, with

$$O(x) = \{g \cdot x \mid g \in G\}.$$

The set of orbits is denoted X/G.

The orbit space, X/G, is obtained from X by an identification (or merging) process: For every orbit, all points in that orbit are merged into a single point. For example, if $X = S^2$ and G is the group consisting of the restrictions of the two linear maps I and $-I$ of $\mathbb{R}^3$ to S^2 (where $-I(x, y, z) = (-x, -y, -z)$), then

$$X/G = S^2/\{I, -I\} \cong \mathbb{RP}^2.$$

Many manifolds can be obtained in this fashion, including the torus, the Klein bottle, the Möbius band, etc.

Since the action of G is transitive on $O(x)$, by Proposition 2.2, we see that for every $x \in X$, we have a bijection

$$O(x) \cong G/G_x.$$

As a corollary, if both X and G are finite, for any set, $A \subseteq X$, of representatives from every orbit, we have the *orbit formula*:

$$|X| = \sum_{a \in A} [G\colon G_x] = \sum_{a \in A} |G|/|G_x|.$$

Even if a group action, $\cdot\colon G \times X \to X$, is not transitive, when X is a manifold, we can consider the set of orbits, X/G, and if the action of G on X satisfies certain conditions, X/G is actually a manifold. Manifolds arising in this fashion are often called *orbifolds*. In summary, we see that manifolds arise in at least two ways from a group action:

(1) As homogeneous spaces, G/G_x, if the action is transitive.

(2) As orbifolds, X/G.

Of course, in both cases, the action must satisfy some additional properties.

Let us now determine some stabilizers for the actions of Examples 1–4, and for more examples of homogeneous spaces.

(a) Consider the action, $\cdot\colon \mathbf{SO}(n) \times S^{n-1} \to S^{n-1}$, of $\mathbf{SO}(n)$ on the sphere S^{n-1} $(n \geq 1)$ defined in Example 1. Since this action is transitive, we can determine the stabilizer of any convenient element of S^{n-1}, say $e_1 = (1, 0, \ldots, 0)$. In order for any $R \in \mathbf{SO}(n)$ to leave e_1 fixed, the first column of R must be e_1, so R is an orthogonal matrix of the form

$$R = \begin{pmatrix} 1 & U \\ 0 & S \end{pmatrix}, \quad \text{with} \quad \det(S) = 1.$$

As the rows of R must be unit vector, we see that $U = 0$ and $S \in \mathbf{SO}(n-1)$. Therefore, the stabilizer of e_1 is isomorphic to $\mathbf{SO}(n-1)$, and we deduce the bijection

$$\mathbf{SO}(n)/\mathbf{SO}(n-1) \cong S^{n-1}.$$

Strictly speaking, $\mathbf{SO}(n-1)$ is not a subgroup of $\mathbf{SO}(n)$ and in all rigor, we should consider the subgroup, $\widetilde{\mathbf{SO}}(n-1)$, of $\mathbf{SO}(n)$ consisting of all matrices of the form

$$\begin{pmatrix} 1 & 0 \\ 0 & S \end{pmatrix}, \quad \text{with} \quad \det(S) = 1$$

and write

$$\mathbf{SO}(n)/\widetilde{\mathbf{SO}}(n-1) \cong S^{n-1}.$$

However, it is common practice to identify $\mathbf{SO}(n-1)$ with $\widetilde{\mathbf{SO}}(n-1)$.

When $n = 2$, as $\mathbf{SO}(1) = \{1\}$, we find that $\mathbf{SO}(2) \cong S^1$, a circle, a fact that we already knew. When $n = 3$, we find that $\mathbf{SO}(3)/\mathbf{SO}(2) \cong S^2$. This says that $\mathbf{SO}(3)$ is somehow the result of glueing circles to the surface of a sphere (in $\mathbb{R}^3$), in such a way that these circles do not intersect. This is hard to visualize!

A similar argument for the complex unit sphere, Σ^{n-1}, shows that

$$\mathbf{SU}(n)/\mathbf{SU}(n-1) \cong \Sigma^{n-1} \cong S^{2n-1}.$$

Again, we identify $\mathbf{SU}(n-1)$ with a subgroup of $\mathbf{SU}(n)$, as in the real case. In particular, when $n = 2$, as $\mathbf{SU}(1) = \{1\}$, we find that

$$\mathbf{SU}(2) \cong S^3,$$

i.e., the group $\mathbf{SU}(2)$ is topologically the sphere S^3! Actually, this is not surprising if we remember that $\mathbf{SU}(2)$ is in fact the group of unit quaternions.

(b) We saw in Example 2 that the action, $\cdot\colon \mathbf{SL}(2,\mathbb{R}) \times H \to H$, of the group $\mathbf{SL}(2,\mathbb{R})$ on the upper half plane is transitive. Let us find out what the stabilizer of $z = i$ is. We should have

$$\frac{ai+b}{ci+d} = i,$$

that is, $ai + b = -c + di$, i.e.,

$$(d-a)i = b + c.$$

Since a, b, c, d are real, we must have $d = a$ and $b = -c$. Moreover, $ad - bc = 1$, so we get $a^2 + b^2 = 1$. We conclude that a matrix in $\mathbf{SL}(2,\mathbb{R})$ fixes i iff it is of the form

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}, \quad \text{with} \quad a^2 + b^2 = 1.$$

Clearly, these are the rotation matrices in $\mathbf{SO}(2)$ and so, the stabilizer of i is $\mathbf{SO}(2)$. We conclude that

$$\mathbf{SL}(2,\mathbb{R})/\mathbf{SO}(2) \cong H.$$

This time, we can view $\mathbf{SL}(2,\mathbb{R})$ as the result of glueing circles to the upper half plane. This is not so easy to visualize. There is a better way to visualize the topology of $\mathbf{SL}(2,\mathbb{R})$ by making it act on the open disk, D. We will return to this action in a little while.

Now, consider the action of $\mathbf{SL}(2,\mathbb{C})$ on $\mathbb{C} \cup \{\infty\} \cong S^2$. As it is transitive, let us find the stabilizer of $z = 0$. We must have

$$\frac{b}{d} = 0,$$

and as $ad - bc = 1$, we must have $b = 0$ and $ad = 1$. Thus, the stabilizer of 0 is the subgroup, $\mathbf{SL}(2,\mathbb{C})_0$, of $\mathbf{SL}(2,\mathbb{C})$ consisting of all matrices of the form

$$\begin{pmatrix} a & 0 \\ c & a^{-1} \end{pmatrix}, \quad \text{where} \quad a \in \mathbb{C} - \{0\} \quad \text{and} \quad c \in \mathbb{C}.$$

We get

$$\mathbf{SL}(2,\mathbb{C})/\mathbf{SL}(2,\mathbb{C})_0 \cong \mathbb{C} \cup \{\infty\} \cong S^2,$$

but this is not very illuminating.

(c) In Example 3, we considered the action, $\cdot\colon \mathbf{GL}(n) \times \mathbf{SPD}(n) \to \mathbf{SPD}(n)$, of $\mathbf{GL}(n)$ on $\mathbf{SPD}(n)$, the set of symmetric positive definite matrices. As this action is transitive, let us find the stabilizer of I. For any $A \in \mathbf{GL}(n)$, the matrix A stabilizes I iff

$$AIA^\top = AA^\top = I.$$

Therefore, the stabilizer of I is $\mathbf{O}(n)$ and we find that

$$\mathbf{GL}(n)/\mathbf{O}(n) = \mathbf{SPD}(n).$$

Observe that if $\mathbf{GL}^+(n)$ denotes the subgroup of $\mathbf{GL}(n)$ consisting of all matrices with a strictly positive determinant, then we have an action $\cdot\colon \mathbf{GL}^+(n) \times \mathbf{SPD}(n) \to \mathbf{SPD}(n)$ of $\mathbf{GL}^+(n)$ on $\mathbf{SPD}(n)$. This action is transtive and we find that the stabilizer of I is $\mathbf{SO}(n)$; consequently, we get

$$\mathbf{GL}^+(n)/\mathbf{SO}(n) = \mathbf{SPD}(n).$$

(d) In Example 4, we considered the action, $\cdot\colon \mathbf{SO}(n+1) \times \mathbb{RP}^n \to \mathbb{RP}^n$, of $\mathbf{SO}(n+1)$ on the (real) projective space, $\mathbb{RP}^n$. As this action is transitive, let us find the stabilizer of the line, $L = [e_1]$, where $e_1 = (1, 0, \ldots, 0)$. For any $R \in \mathbf{SO}(n+1)$, the line L is fixed iff either $R(e_1) = e_1$ or $R(e_1) = -e_1$, since e_1 and $-e_1$ define the same line. As R is orthogonal with $\det(R) = 1$, this means that R is of the form

$$R = \begin{pmatrix} \alpha & 0 \\ 0 & S \end{pmatrix}, \quad \text{with} \quad \alpha = \pm 1 \quad \text{and} \quad \det(S) = \alpha.$$

But, S must be orthogonal, so we conclude $S \in \mathbf{O}(n)$. Therefore, the stabilizer of $L = [e_1]$ is isomorphic to the group $\mathbf{O}(n)$ and we find that

$$\mathbf{SO}(n+1)/\mathbf{O}(n) \cong \mathbb{RP}^n.$$

Strictly speaking, $\mathbf{O}(n)$ is not a subgroup of $\mathbf{SO}(n+1)$, so the above equation does not make sense. We should write

$$\mathbf{SO}(n+1)/\widetilde{\mathbf{O}}(n) \cong \mathbb{RP}^n,$$

where $\widetilde{\mathbf{O}}(n)$ is the subgroup of $\mathbf{SO}(n+1)$ consisting of all matrices of the form

$$\begin{pmatrix} \alpha & 0 \\ 0 & S \end{pmatrix}, \quad \text{with} \quad S \in \mathbf{O}(n),\ \alpha = \pm 1 \quad \text{and} \quad \det(S) = \alpha.$$

However, the common practice is to write $\mathbf{O}(n)$ instead of $\widetilde{\mathbf{O}}(n)$.

We should mention that $\mathbb{RP}^3$ and $\mathbf{SO}(3)$ are homeomorphic spaces. This is shown using the quaternions, for example, see Gallier [60], Chapter 8.

A similar argument applies to the action, $\cdot\colon \mathbf{SU}(n+1) \times \mathbb{CP}^n \to \mathbb{CP}^n$, of $\mathbf{SU}(n+1)$ on the (complex) projective space, $\mathbb{CP}^n$. We find that

$$\mathbf{SU}(n+1)/\mathbf{U}(n) \cong \mathbb{CP}^n.$$

Again, the above is a bit sloppy as $\mathbf{U}(n)$ is not a subgroup of $\mathbf{SU}(n+1)$. To be rigorous, we should use the subgroup, $\widetilde{\mathbf{U}}(n)$, consisting of all matrices of the form

$$\begin{pmatrix} \alpha & 0 \\ 0 & S \end{pmatrix}, \quad \text{with} \quad S \in \mathbf{U}(n),\ |\alpha| = 1 \quad \text{and} \quad \det(S) = \overline{\alpha}.$$

The common practice is to write $\mathbf{U}(n)$ instead of $\widetilde{\mathbf{U}}(n)$. In particular, when $n = 1$, we find that

$$\mathbf{SU}(2)/\mathbf{U}(1) \cong \mathbb{CP}^1.$$

But, we know that $\mathbf{SU}(2) \cong S^3$ and, clearly, $\mathbf{U}(1) \cong S^1$. So, again, we find that $S^3/S^1 \cong \mathbb{CP}^1$ (but we know, more, namely, $S^3/S^1 \cong S^2 \cong \mathbb{CP}^1$.)

(e) We now consider a generalization of projective spaces (real and complex). First, consider the real case. Given any $n \geq 1$, for any k, with $0 \leq k \leq n$, let $G(k, n)$ be the set of all linear k-dimensional subspaces of $\mathbb{R}^n$ (also called *k-planes*). Any k-dimensional subspace, U, of $\mathbb{R}$ is spanned by k linearly independent vectors, $u_1, \ldots, u_k$, in $\mathbb{R}^n$; write $U = \text{span}(u_1, \ldots, u_k)$. We can define an action, $\cdot\colon \mathbf{O}(n) \times G(k, n) \to G(k, n)$, as follows: For any $R \in \mathbf{O}(n)$, for any $U = \text{span}(u_1, \ldots, u_k)$, let

$$R \cdot U = \text{span}(Ru_1, \ldots, Ru_k).$$

We have to check that the above is well defined. If $U = \text{span}(v_1, \ldots, v_k)$ for any other k linearly independent vectors, $v_1, \ldots, v_k$, we have

$$v_i = \sum_{j=1}^{k} a_{ij} u_j, \quad 1 \leq i \leq k,$$

for some $a_{ij} \in \mathbb{R}$, and so,

$$Rv_i = \sum_{j=1}^{k} a_{ij} Ru_j, \quad 1 \leq i \leq k,$$

which shows that

$$\text{span}(Ru_1, \ldots, Ru_k) = \text{span}(Rv_1, \ldots, Rv_k),$$

i.e., the above action is well defined. This action is transitive. This is because if U and V are any two k-planes, we may assume that $U = \text{span}(u_1, \ldots, u_k)$ and $V = \text{span}(v_1, \ldots, v_k)$, where the u_i's form an orthonormal family and similarly for the v_i's. Then, we can extend these families to orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ or $\mathbb{R}^n$, and w.r.t. the orthonormal basis $(u_1, \ldots, u_n)$, the matrix of the linear map sending u_i to v_i is orthogonal. Thus, it is enough to find the stabilizer of any k-plane. Pick $U = \text{span}(e_1, \ldots, e_k)$, where $(e_1, \ldots, e_n)$ is the canonical basis of $\mathbb{R}^n$ (i.e., $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, with the 1 in the ith position). Now, any $R \in \mathbf{O}(n)$ stabilizes U iff R maps $e_1, \ldots, e_k$ to k linearly independent vectors in the subspace $U = \text{span}(e_1, \ldots, e_k)$, i.e., R is of the form

$$R = \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix},$$

where S is $k \times k$ and T is $(n-k) \times (n-k)$. Moreover, as R is orthogonal, S and T must be orthogonal, i.e., $S \in \mathbf{O}(k)$ and $T \in \mathbf{O}(n-k)$. We deduce that the stabilizer of U is isomorphic to $\mathbf{O}(k) \times \mathbf{O}(n-k)$ and we find that

$$\mathbf{O}(n)/(\mathbf{O}(k) \times \mathbf{O}(n-k)) \cong G(k, n).$$

It turns out that this makes $G(k,n)$ into a smooth manifold of dimension $k(n-k)$ called a *Grassmannian*.

The restriction of the action of $\mathbf{O}(n)$ on $G(k,n)$ to $\mathbf{SO}(n)$ yields an action, $\cdot\colon \mathbf{SO}(n) \times G(k,n) \to G(k,n)$, of $\mathbf{SO}(n)$ on $G(k,n)$. Then, it is easy to see that the stabilizer of the subspace U is isomorphic to the subgroup, $S(\mathbf{O}(k) \times \mathbf{O}(n-k))$, of $\mathbf{SO}(n)$ consisting of the rotations of the form

$$R = \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix},$$

with $S \in \mathbf{O}(k)$, $T \in \mathbf{O}(n-k)$ and $\det(S)\det(T) = 1$. Thus, we also have

$$\mathbf{SO}(n)/S(\mathbf{O}(k) \times \mathbf{O}(n-k)) \cong G(k,n).$$

If we recall the projection $pr\colon \mathbb{R}^{n+1} - \{0\} \to \mathbb{RP}^n$, by definition, a *$k$-plane* in $\mathbb{RP}^n$ is the image under pr of any $(k+1)$-plane in $\mathbb{R}^{n+1}$. So, for example, a line in $\mathbb{RP}^n$ is the image of a 2-plane in $\mathbb{R}^{n+1}$, and a hyperplane in $\mathbb{RP}^n$ is the image of a hyperplane in $\mathbb{R}^{n+1}$. The advantage of this point of view is that the k-planes in $\mathbb{RP}^n$ are arbitrary, i.e., they do not have to go through "the origin" (which does not make sense, anyway!). Then, we see that we can interpret the Grassmannian, $G(k+1,n+1)$, as a space of "parameters" for the k-planes in $\mathbb{RP}^n$. For example, $G(2,n+1)$ parametrizes the lines in $\mathbb{RP}^n$. In this viewpoint, $G(k+1,n+1)$ is usually denoted $\mathbb{G}(k,n)$.

It can be proved (using some exterior algebra) that $G(k,n)$ can be embedded in $\mathbb{RP}^{\binom{n}{k}-1}$. Much more is true. For example, $G(k,n)$ is a projective variety, which means that it can be defined as a subset of $\mathbb{RP}^{\binom{n}{k}-1}$ equal to the zero locus of a set of homogeneous equations. There is even a set of quadratic equations, known as the *Plücker equations*, defining $G(k,n)$. In particular, when $n = 4$ and $k = 2$, we have $G(2,4) \subseteq \mathbb{RP}^5$ and $G(2,4)$ is defined by a single equation of degree 2. The Grassmannian $G(2,4) = \mathbb{G}(1,3)$ is known as the *Klein quadric*. This hypersurface in $\mathbb{RP}^5$ parametrizes the lines in $\mathbb{RP}^3$.

Complex Grassmannians are defined in a similar way, by replacing $\mathbb{R}$ by $\mathbb{C}$ and $\mathbf{O}(n)$ by $\mathbf{U}(n)$ throughout. The complex Grassmannian, $G_{\mathbb{C}}(k,n)$, is a complex manifold as well as a real manifold and we have

$$\mathbf{U}(n)/(\mathbf{U}(k) \times \mathbf{U}(n-k)) \cong G_{\mathbb{C}}(k,n).$$

As in the case of the real Grassmannians, the action of $\mathbf{U}(n)$ on $G_{\mathbb{C}}(k,n)$ yields an action of $\mathbf{SU}(n)$ on $G_{\mathbb{C}}(k,n)$ and we get

$$\mathbf{SU}(n)/S(\mathbf{U}(k) \times \mathbf{U}(n-k)) \cong G_{\mathbb{C}}(k,n),$$

where $S(\mathbf{U}(k) \times \mathbf{U}(n-k))$ is the subgroup of $\mathbf{SU}(n)$ consisting of all matrices, $R \in \mathbf{SU}(n)$, of the form

$$R = \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix},$$

with $S \in \mathbf{U}(k)$, $T \in \mathbf{U}(n-k)$ and $\det(S)\det(T) = 1$.

We now return to case (b) to give a better picture of $\mathbf{SL}(2,\mathbb{R})$. Instead of having $\mathbf{SL}(2,\mathbb{R})$ act on the upper half plane we define an action of $\mathbf{SL}(2,\mathbb{R})$ on the open unit disk, D. Technically, it is easier to consider the group, $\mathbf{SU}(1,1)$, which is isomorphic to $\mathbf{SL}(2,\mathbb{R})$, and to make $\mathbf{SU}(1,1)$ act on D. The group $\mathbf{SU}(1,1)$ is the group of 2×2 complex matrices of the form

$$\begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix}, \quad \text{with} \quad a\overline{a} - b\overline{b} = 1.$$

The reader should check that if we let

$$g = \begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix},$$

then the map from $\mathbf{SL}(2,\mathbb{R})$ to $\mathbf{SU}(1,1)$ given by

$$A \mapsto gAg^{-1}$$

is an isomorphism. Observe that the Möbius transformation associated with g is

$$z \mapsto \frac{z-i}{z+i},$$

which is the holomorphic isomorphism mapping H to D mentionned earlier! Now, we can define a bijection between $\mathbf{SU}(1,1)$ and $S^1 \times D$ given by

$$\begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix} \mapsto (a/|a|, b/a).$$

We conclude that $\mathbf{SL}(2,\mathbb{R}) \cong \mathbf{SU}(1,1)$ is topologically an open solid torus (i.e., with the surface of the torus removed). It is possible to further classify the elements of $\mathbf{SL}(2,\mathbb{R})$ into three categories and to have geometric interpretations of these as certain regions of the torus. For details, the reader should consult Carter, Segal and Macdonald [31] or Duistermatt and Kolk [54] (Chapter 1, Section 1.2).

The group $\mathbf{SU}(1,1)$ acts on D by interpreting any matrix in $\mathbf{SU}(1,1)$ as a Möbius tranformation, i.e.,

$$\begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix} \mapsto \left(z \mapsto \frac{az+b}{\overline{b}z + \overline{a}} \right).$$

The reader should check that these transformations preserve D. Both the upper half-plane and the open disk are models of Lobachevsky's non-Euclidean geometry (where the parallel postulate fails). They are also models of hyperbolic spaces (Riemannian manifolds with constant negative curvature, see Gallot, Hulin and Lafontaine [61], Chapter III). According to Dubrovin, Fomenko, and Novikov [52] (Chapter 2, Section 13.2), the open disk model is due to Poincaré and the upper half-plane model to Klein, although Poincaré was the first to realize that the upper half-plane is a hyperbolic space.

2.3 The Lorentz Groups $\mathbf{O}(n,1)$, $\mathbf{SO}(n,1)$ and $\mathbf{SO}_0(n,1)$

The Lorentz group provides another interesting example. Moreover, the Lorentz group $\mathbf{SO}(3,1)$ shows up in an interesting way in computer vision.

Denote the $p \times p$-identity matrix by I_p, for $p, q, \geq 1$, and define

$$I_{p,q} = \begin{pmatrix} I_p & 0 \\ 0 & -I_q \end{pmatrix}.$$

If $n = p + q$, the matrix $I_{p,q}$ is associated with the nondegenerate symmetric bilinear form

$$\varphi_{p,q}((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sum_{i=1}^{p} x_i y_i - \sum_{j=p+1}^{n} x_j y_j$$

with associated quadratic form

$$\Phi_{p,q}((x_1, \ldots, x_n)) = \sum_{i=1}^{p} x_i^2 - \sum_{j=p+1}^{n} x_j^2.$$

In particular, when $p = 1$ and $q = 3$, we have the *Lorentz metric*

$$x_1^2 - x_2^2 - x_3^2 - x_4^2.$$

In physics, x_1 is interpreted as time and written t and x_2, x_3, x_4 as coordinates in $\mathbb{R}^3$ and written x, y, z. Thus, the Lozentz metric is usually written a

$$t^2 - x^2 - y^2 - z^2,$$

although it also appears as

$$x^2 + y^2 + z^2 - t^2,$$

which is equivalent but slightly less convenient for certain purposes, as we will see later. The space $\mathbb{R}^4$ with the Lorentz metric is called *Minkowski space*. It plays an important role in Einstein's theory of special relativity.

The group $\mathbf{O}(p, q)$ is the set of all $n \times n$-matrices

$$\mathbf{O}(p,q) = \{A \in \mathbf{GL}(n, \mathbb{R}) \mid A^\top I_{p,q} A = I_{p,q}\}.$$

This is the group of all invertible linear maps of $\mathbb{R}^n$ that preserve the quadratic form, $\Phi_{p,q}$, i.e., the group of isometries of $\Phi_{p,q}$. Clearly, $I_{p,q}^2 = I$, so the condition $A^\top I_{p,q} A = I_{p,q}$ is equivalent to $I_{p,q} A^\top I_{p,q} A = I$, which means that

$$A^{-1} = I_{p,q} A^\top I_{p,q}.$$

Thus, $AI_{p,q}A^\top = I_{p,q}$ also holds, which shows that $\mathbf{O}(p,q)$ is closed under transposition (i.e., if $A \in \mathbf{O}(p,q)$, then $A^\top \in \mathbf{O}(p,q)$). We have the subgroup

$$\mathbf{SO}(p,q) = \{A \in \mathbf{O}(p,q) \mid \det(A) = 1\}$$

consisting of the isometries of $(\mathbb{R}^n, \Phi_{p,q})$ with determinant $+1$. It is clear that $\mathbf{SO}(p,q)$ is also closed under transposition. The condition $A^\top I_{p,q} A = I_{p,q}$ has an interpretation in terms of the inner product $\varphi_{p,q}$ and the columns (and rows) of A. Indeed, if we denote the jth column of A by A_j, then

$$A^\top I_{p,q} A = (\varphi_{p,q}(A_i, A_j)),$$

so $A \in \mathbf{O}(p,q)$ iff the columns of A form an "orthonormal basis" w.r.t. $\varphi_{p,q}$, i.e.,

$$\varphi_{p,q}(A_i, A_j) = \begin{cases} \delta_{ij} & \text{if } 1 \leq i, j \leq p; \\ -\delta_{ij} & \text{if } p+1 \leq i, j \leq p+q. \end{cases}$$

The difference with the usual orthogonal matrices is that $\varphi_{p,q}(A_i, A_i) = -1$, if $p+1 \leq i \leq p+q$. As $\mathbf{O}(p,q)$ is closed under transposition, the rows of A also form an orthonormal basis w.r.t. $\varphi_{p,q}$.

It turns out that $\mathbf{SO}(p,q)$ has two connected components and the component containing the identity is a subgroup of $\mathbf{SO}(p,q)$ denoted $\mathbf{SO}_0(p,q)$. The group $\mathbf{SO}_0(p,q)$ turns out to be homeomorphic to $\mathbf{SO}(p) \times \mathbf{SO}(q) \times \mathbb{R}^{pq}$, but this is not easy to prove. (One way to prove it is to use results on pseudo-algebraic subgroups of $\mathbf{GL}(n, \mathbb{C})$, see Knapp [90] or Gallier's notes on Clifford algebras (on the web)).

We will now determine the polar decomposition and the SVD decomposition of matrices in the Lorentz groups $\mathbf{O}(n,1)$ and $\mathbf{SO}(n,1)$. Write $J = I_{n,1}$ and, given any $A \in \mathbf{O}(n,1)$, write

$$A = \begin{pmatrix} B & u \\ v^\top & c \end{pmatrix},$$

where B is an $n \times n$ matrix, u, v are (column) vectors in $\mathbb{R}^n$ and $c \in \mathbb{R}$. We begin with the polar decomposition of matrices in the Lorentz groups $\mathbf{O}(n,1)$.

Proposition 2.3. *Every matrix $A \in \mathbf{O}(n,1)$ has a polar decomposition of the form*

$$A = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{I + vv^\top} & v \\ v^\top & c \end{pmatrix} \quad \textit{or} \quad A = \begin{pmatrix} Q & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{I + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

where $Q \in \mathbf{O}(n)$ and $c = \sqrt{\|v\|^2 + 1}$.

Proof. Write A in block form as above. As the condition for A to be in $\mathbf{O}(n,1)$ is $A^\top J A = J$, we get

$$\begin{pmatrix} B^\top & v \\ u^\top & c \end{pmatrix} \begin{pmatrix} B & u \\ -v^\top & -c \end{pmatrix} = \begin{pmatrix} I_n & 0 \\ 0 & -1 \end{pmatrix},$$

i.e,.

$$\begin{aligned} B^\top B &= I + vv^\top \\ u^\top u &= c^2 - 1 \\ B^\top u &= cv. \end{aligned}$$

If we remember that we also have $AJA^\top = J$, then

$$Bv = cu,$$

which can also be deduced from the three equations above. From $u^\top u = \|u\|^2 = c^2 - 1$, we deduce that $|c| \geq 1$, and from $B^\top B = I + vv^\top$, we deduce that $B^\top B$ is symmetric, positive definite. Now, geometrically, it is well known that $vv^\top / v^\top v$ is the orthogonal projection onto the line determined by v. Consequently, the kernel of $vv^\top$ is the orthogonal complement of v and $vv^\top$ has the eigenvalue 0 with multiplicity $n-1$ and the eigenvalue $c^2 - 1 = \|v\|^2 = v^\top v$ with multiplicity 1. The eigenvectors associated with 0 are orthogonal to v and the eigenvectors associated with $c^2 - 1$ are proportional with v. It follows that $I + vv^\top$ has the eigenvalue 1 with multiplicity $n-1$ and the eigenvalue c^2 with multiplicity 1, the eigenvectors being as before. Now, B has polar form $B = QS_1$, where Q is orthogonal and S_1 is symmetric positive definite and $S_1^2 = B^\top B = I + vv^\top$. Therefore, if $c > 0$, then $S_1 = \sqrt{I + vv^\top}$ is a symmetric positive definite matrix with eigenvalue 1 with multiplicity $n-1$ and eigenvalue c with multiplicity 1, the eigenvectors being as before. If $c < 0$, then change c to $-c$.

Case 1: $c > 0$. Then, v is an eigenvector of S_1 for c and we must also have $Bv = cu$, which implies

$$Bv = QS_1 v = Q(cv) = cQv = cu,$$

so

$$Qv = u.$$

It follows that

$$A = \begin{pmatrix} B & u \\ v^\top & c \end{pmatrix} = \begin{pmatrix} QS_1 & Qv \\ v^\top & c \end{pmatrix} = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{I + vv^\top} & v \\ v^\top & c \end{pmatrix}.$$

Therefore, the polar decomposition of $A \in \mathbf{O}(n,1)$ is

$$A = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{I + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

where $Q \in \mathbf{O}(n)$ and $c = \sqrt{\|v\|^2 + 1}$.

Case 2: $c < 0$. Then, v is an eigenvector of S_1 for $-c$ and we must also have $Bv = cu$, which implies

$$Bv = QS_1 v = Q(-cv) = cQ(-v) = cu,$$

so

$$Q(-v) = u.$$

It follows that

$$A = \begin{pmatrix} B & u \\ v^\top & c \end{pmatrix} = \begin{pmatrix} QS_1 & Q(-v) \\ v^\top & c \end{pmatrix} = \begin{pmatrix} Q & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{I + vv^\top} & -v \\ -v^\top & -c \end{pmatrix}.$$

In this case, the polar decomposition of $A \in \mathbf{O}(n,1)$ is

$$A = \begin{pmatrix} Q & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{I + vv^\top} & -v \\ -v^\top & -c \end{pmatrix},$$

where $Q \in \mathbf{O}(n)$ and $c = -\sqrt{\|v\|^2 + 1}$. Therefore, we conclude that any $A \in \mathbf{O}(n,1)$ has a polar decomposition of the form

$$A = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{I + vv^\top} & v \\ v^\top & c \end{pmatrix} \quad \text{or} \quad A = \begin{pmatrix} Q & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{I + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

where $Q \in \mathbf{O}(n)$ and $c = \sqrt{\|v\|^2 + 1}$. □

Thus, we see that $\mathbf{O}(n,1)$ has four components corresponding to the cases:

(1) $Q \in \mathbf{O}(n)$; $\det(Q) < 0$; $+1$ as the lower right entry of the orthogonal matrix;

(2) $Q \in \mathbf{SO}(n)$; -1 as the lower right entry of the orthogonal matrix;

(3) $Q \in \mathbf{O}(n)$; $\det(Q) < 0$; -1 as the lower right entry of the orthogonal matrix;

(4) $Q \in \mathbf{SO}(n)$; $+1$ as the lower right entry of the orthogonal matrix.

Observe that $\det(A) = -1$ in cases (1) and (2) and that $\det(A) = +1$ in cases (3) and (4). Thus, (3) and (4) correspond to the group $\mathbf{SO}(n,1)$, in which case the polar decomposition is of the form

$$A = \begin{pmatrix} Q & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{I + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

where $Q \in \mathbf{O}(n)$, with $\det(Q) = -1$ and $c = \sqrt{\|v\|^2 + 1}$ or

$$A = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{I + vv^\top} & v \\ v^\top & c \end{pmatrix}$$

where $Q \in \mathbf{SO}(n)$ and $c = \sqrt{\|v\|^2 + 1}$. The components in (1) and (2) are not groups. We will show later that all four components are connected and that case (4) corresponds to a group (Proposition 2.8). This group is the connected component of the identity and it is

denoted $\mathbf{SO}_0(n,1)$ (see Corollary 2.27). For the time being, note that $A \in \mathbf{SO}_0(n,1)$ iff $A \in \mathbf{SO}(n,1)$ and $a_{n+1\,n+1}$ $(= c) > 0$ (here, $A = (a_{i\,j})$.) In fact, we proved above that if $a_{n+1\,n+1} > 0$, then $a_{n+1\,n+1} \geq 1$.

Remark: If we let

$$\Lambda_P = \begin{pmatrix} I_{n-1,1} & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \Lambda_T = I_{n,1}, \quad \text{where} \quad I_{n,1} = \begin{pmatrix} I_n & 0 \\ 0 & -1 \end{pmatrix},$$

then we have the disjoint union

$$\mathbf{O}(n,1) = \mathbf{SO}_0(n,1) \cup \Lambda_P\mathbf{SO}_0(n,1) \cup \Lambda_T\mathbf{SO}_0(n,1) \cup \Lambda_P\Lambda_T\mathbf{SO}_0(n,1).$$

In order to determine the SVD of matrices in $\mathbf{SO}_0(n,1)$, we analyze the eigenvectors and the eigenvalues of the positive definite symmetric matrix

$$S = \begin{pmatrix} \sqrt{I + vv^\top} & v \\ v^\top & c \end{pmatrix}$$

involved in Proposition 2.3. Such a matrix is called a *Lorentz boost*. Observe that if $v = 0$, then $c = 1$ and $S = I_{n+1}$.

Proposition 2.4. *Assume $v \neq 0$. The eigenvalues of the symmetric positive definite matrix*

$$S = \begin{pmatrix} \sqrt{I + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

where $c = \sqrt{\|v\|^2 + 1}$, are 1 with multiplicity $n-1$, and e^α and $e^{-\alpha}$ each with multiplicity 1 (for some $\alpha \geq 0$). An orthonormal basis of eigenvectors of S consists of vectors of the form

$$\begin{pmatrix} u_1 \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} u_{n-1} \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{v}{\sqrt{2}\|v\|} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} \frac{v}{\sqrt{2}\|v\|} \\ -\frac{1}{\sqrt{2}} \end{pmatrix},$$

where the $u_i \in \mathbb{R}^n$ are all orthogonal to v and pairwise orthogonal.

Proof. Let us solve the linear system

$$\begin{pmatrix} \sqrt{I + vv^\top} & v \\ v^\top & c \end{pmatrix} \begin{pmatrix} v \\ d \end{pmatrix} = \lambda \begin{pmatrix} v \\ d \end{pmatrix}.$$

We get

$$\begin{aligned} \sqrt{I + vv^\top}(v) + dv &= \lambda v \\ v^\top v + cd &= \lambda d, \end{aligned}$$

that is (since $c = \sqrt{\|v\|^2 + 1}$ and $\sqrt{I + vv^\top}(v) = cv$),

$$\begin{aligned}(c+d)v &= \lambda v\\ c^2 - 1 + cd &= \lambda d.\end{aligned}$$

Since $v \neq 0$, we get $\lambda = c + d$. Substituting in the second equation, we get

$$c^2 - 1 + cd = (c+d)d,$$

that is,

$$d^2 = c^2 - 1.$$

Thus, either $\lambda_1 = c + \sqrt{c^2-1}$ and $d = \sqrt{c^2-1}$, or $\lambda_2 = c - \sqrt{c^2-1}$ and $d = -\sqrt{c^2-1}$. Since $c \geq 1$ and $\lambda_1\lambda_2 = 1$, set $\alpha = \log(c + \sqrt{c^2-1}) \geq 0$, so that $-\alpha = \log(c - \sqrt{c^2-1})$ and then, $\lambda_1 = e^\alpha$ and $\lambda_2 = e^{-\alpha}$. On the other hand, if u is orthogonal to v, observe that

$$\begin{pmatrix} \sqrt{I + vv^\top} & v \\ v^\top & c \end{pmatrix} \begin{pmatrix} u \\ 0 \end{pmatrix} = \begin{pmatrix} u \\ 0 \end{pmatrix},$$

since the kernel of $vv^\top$ is the orthogonal complement of v. The rest is clear. □

Corollary 2.5. *The singular values of any matrix $A \in \mathbf{O}(n,1)$ are 1 with multiplicity $n-1$, e^α, and $e^{-\alpha}$, for some $\alpha \geq 0$.*

Note that the case $\alpha = 0$ is possible, in which case, A is an orthogonal matrix of the form

$$\begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} Q & 0 \\ 0 & -1 \end{pmatrix},$$

with $Q \in \mathbf{O}(n)$. The two singular values e^α and $e^{-\alpha}$ tell us how much A deviates from being orthogonal.

We can now determine a convenient form for the SVD of matrices in $\mathbf{O}(n,1)$.

Theorem 2.6. *Every matrix $A \in \mathbf{O}(n,1)$ can be written as*

$$A = \begin{pmatrix} P & 0 \\ 0 & \epsilon \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \cosh\alpha & \sinh\alpha \\ 0 & \cdots & 0 & \sinh\alpha & \cosh\alpha \end{pmatrix} \begin{pmatrix} Q^\top & 0 \\ 0 & 1 \end{pmatrix}$$

with $\epsilon = \pm 1$, $P \in \mathbf{O}(n)$ and $Q \in \mathbf{SO}(n)$. When $A \in \mathbf{SO}(n,1)$, we have $\det(P)\epsilon = +1$, and when $A \in \mathbf{SO}_0(n,1)$, we have $\epsilon = +1$ and $P \in \mathbf{SO}(n)$, that is,

$$A = \begin{pmatrix} P & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \cosh\alpha & \sinh\alpha \\ 0 & \cdots & 0 & \sinh\alpha & \cosh\alpha \end{pmatrix} \begin{pmatrix} Q^\top & 0 \\ 0 & 1 \end{pmatrix}$$

with $P \in \mathbf{SO}(n)$ and $Q \in \mathbf{SO}(n)$.

Proof. By Proposition 2.3, any matrix $A \in \mathbf{O}(n)$ can be written as

$$A = \begin{pmatrix} R & 0 \\ 0 & \epsilon \end{pmatrix} \begin{pmatrix} \sqrt{I + vv^\top} & v \\ v^\top & c \end{pmatrix}$$

where $\epsilon = \pm 1$, $R \in \mathbf{O}(n)$ and $c = \sqrt{\|v\|^2 + 1}$. The case where $c = 1$ is trivial, so assume $c > 1$, which means that α from Proposition 2.4 is such that $\alpha > 0$. The key fact is that the eigenvalues of the matrix

$$\begin{pmatrix} \cosh\alpha & \sinh\alpha \\ \sinh\alpha & \cosh\alpha \end{pmatrix}$$

are e^α and $e^{-\alpha}$ and that

$$\begin{pmatrix} e^\alpha & 0 \\ 0 & e^{-\alpha} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \cosh\alpha & \sinh\alpha \\ \sinh\alpha & \cosh\alpha \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

From this fact, we see that the diagonal matrix

$$D = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & e^\alpha & 0 \\ 0 & \cdots & 0 & 0 & e^{-\alpha} \end{pmatrix}$$

of eigenvalues of S is given by

$$D = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \cosh\alpha & \sinh\alpha \\ 0 & \cdots & 0 & \sinh\alpha & \cosh\alpha \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

By Proposition 2.4, an orthonormal basis of eigenvectors of S consists of vectors of the form

$$\begin{pmatrix} u_1 \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} u_{n-1} \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{v}{\sqrt{2}\|v\|} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} \frac{v}{\sqrt{2}\|v\|} \\ -\frac{1}{\sqrt{2}} \end{pmatrix},$$

where the $u_i \in \mathbb{R}^n$ are all orthogonal to v and pairwise orthogonal. Now, if we multiply the matrices

$$\begin{pmatrix} u_1 & \cdots & u_{n-1} & \frac{v}{\sqrt{2}\|v\|} & \frac{v}{\sqrt{2}\|v\|} \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix},$$

we get an orthogonal matrix of the form

$$\begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix}$$

where the columns of Q are the vectors

$$u_1, \cdots, u_{n-1}, \frac{v}{\|v\|}.$$

By flipping u_1 to $-u_1$ if necessary, we can make sure that this matrix has determinant $+1$. Consequently,

$$S = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \cosh\alpha & \sinh\alpha \\ 0 & \cdots & 0 & \sinh\alpha & \cosh\alpha \end{pmatrix} \begin{pmatrix} Q^\top & 0 \\ 0 & 1 \end{pmatrix},$$

so

$$A = \begin{pmatrix} R & 0 \\ 0 & \epsilon \end{pmatrix} \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \cosh\alpha & \sinh\alpha \\ 0 & \cdots & 0 & \sinh\alpha & \cosh\alpha \end{pmatrix} \begin{pmatrix} Q^\top & 0 \\ 0 & 1 \end{pmatrix},$$

and if we let $P = RQ$, we get the desired decomposition. □

Remark: We warn our readers about Chapter 6 of Baker's book [13]. Indeed, this chapter is seriously flawed. The main two Theorems (Theorem 6.9 and Theorem 6.10) are false and as consequence, the proof of Theorem 6.11 is wrong too. Theorem 6.11 states that the exponential map $\exp\colon \mathfrak{so}(n,1) \to \mathbf{SO}_0(n,1)$ is surjective, which is correct, but known proofs are nontrivial and quite lengthy (see Section 5.5). The proof of Theorem 6.12 is also false, although the theorem itself is correct (this is our Theorem 5.22, see Section 5.5). The main problem with Theorem 6.9 (in Baker) is that the existence of the normal form for matrices in $\mathbf{SO}_0(n,1)$ claimed by this theorem is unfortunately false on several accounts. Firstly, it would imply that every matrix in $\mathbf{SO}_0(n,1)$ can be diagonalized, but this is false for $n \geq 2$. Secondly, even if a matrix $A \in \mathbf{SO}_0(n,1)$ is diagonalizable as $A = PDP^{-1}$, Theorem 6.9 (and Theorem 6.10) miss some possible eigenvalues and the matrix P is not necessarily in $\mathbf{SO}_0(n,1)$ (as the case $n = 1$ already shows). For a thorough analysis of the eigenvalues of Lorentz isometries (and much more), one should consult Riesz [127] (Chapter III).

Clearly, a result similar to Theorem 2.6 also holds for the matrices in the groups $\mathbf{O}(1,n)$,

$\mathbf{SO}(1,n)$ and $\mathbf{SO}_0(1,n)$. For example, every matrix $A \in \mathbf{SO}_0(1,n)$ can be written as

$$A = \begin{pmatrix} 1 & 0 \\ 0 & P \end{pmatrix} \begin{pmatrix} \cosh\alpha & \sinh\alpha & 0 & \cdots & 0 \\ \sinh\alpha & \cosh\alpha & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & Q^\top \end{pmatrix},$$

where $P, Q \in \mathbf{SO}(n)$.

In the case $n = 3$, we obtain the *proper orthochronous Lorentz group*, $\mathbf{SO}_0(1,3)$, also denoted $\mathbf{Lor}(1,3)$. By the way, $\mathbf{O}(1,3)$ is called the *(full) Lorentz group* and $\mathbf{SO}(1,3)$ is the *special Lorentz group*.

Theorem 2.6 (really, the version for $\mathbf{SO}_0(1,n)$) shows that the Lorentz group $\mathbf{SO}_0(1,3)$ is generated by the matrices of the form

$$\begin{pmatrix} 1 & 0 \\ 0 & P \end{pmatrix} \quad \text{with } P \in \mathbf{SO}(3)$$

and the matrices of the form

$$\begin{pmatrix} \cosh\alpha & \sinh\alpha & 0 & 0 \\ \sinh\alpha & \cosh\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

This fact will be useful when we prove that the homomorphism $\varphi\colon \mathbf{SL}(2,\mathbb{C}) \to \mathbf{SO}_0(1,3)$ is surjective.

Remark: Unfortunately, unlike orthogonal matrices which can always be diagonalized over $\mathbb{C}$, **not** every matrix in $\mathbf{SO}(1,n)$ can be diagonalized for $n \geq 2$. This has to do with the fact that the Lie algebra $\mathfrak{so}(1,n)$ has non-zero idempotents (see Section 5.5).

It turns out that the group $\mathbf{SO}_0(1,3)$ admits another interesting characterization involving the hypersurface

$$\mathcal{H} = \{(t,x,y,z) \in \mathbb{R}^4 \mid t^2 - x^2 - y^2 - z^2 = 1\}.$$

This surface has two sheets and it is not hard to show that $\mathbf{SO}_0(1,3)$ is the subgroup of $\mathbf{SO}(1,3)$ that preserves these two sheets (does not swap them). Actually, we will prove this fact for any n. In preparation for this we need some definitions and a few propositions.

Let us switch back to $\mathbf{SO}(n,1)$. First, as a matter of notation, we write every $u \in \mathbb{R}^{n+1}$ as $u = (\mathbf{u}, t)$, where $\mathbf{u} \in \mathbb{R}^n$ and $t \in \mathbb{R}$, so that the Lorentz inner product can be expressed as

$$\langle u, v\rangle = \langle(\mathbf{u},t),(\mathbf{v},s)\rangle = \mathbf{u}\cdot\mathbf{v} - ts,$$

where $\mathbf{u}\cdot\mathbf{v}$ is the standard Euclidean inner product (the Euclidean norm of x is denoted $\|x\|$). Then, we can classify the vectors in $\mathbb{R}^{n+1}$ as follows:

Definition 2.10. A nonzero vector, $u = (\mathbf{u}, t) \in \mathbb{R}^{n+1}$ is called

(a) *spacelike* iff $\langle u, u\rangle > 0$, i.e., iff $\|\mathbf{u}\|^2 > t^2$;

(b) *timelike* iff $\langle u, u\rangle < 0$, i.e., iff $\|\mathbf{u}\|^2 < t^2$;

(c) *lightlike* or *isotropic* iff $\langle u, u\rangle = 0$, i.e., iff $\|\mathbf{u}\|^2 = t^2$.

A spacelike (resp. timelike, resp. lightlike) vector is said to be *positive* iff $t > 0$ and *negative* iff $t < 0$. The set of all isotropic vectors

$$\mathcal{H}_n(0) = \{u = (\mathbf{u}, t) \in \mathbb{R}^{n+1} \mid \|\mathbf{u}\|^2 = t^2\}$$

is called the *light cone*. For every $r > 0$, let

$$\mathcal{H}_n(r) = \{u = (\mathbf{u}, t) \in \mathbb{R}^{n+1} \mid \|\mathbf{u}\|^2 - t^2 = -r\},$$

a hyperboloid of two sheets.

It is easy to check that $\mathcal{H}_n(r)$ has two connected components as follows: First, since $r > 0$ and

$$\|\mathbf{u}\|^2 + r = t^2,$$

we have $|t| \geq \sqrt{r}$. Now, for any $x = (x_1, \ldots, x_n, t) \in \mathcal{H}_n(r)$ with $t \geq \sqrt{r}$, we have the continuous path from $(0, \ldots, 0, \sqrt{r})$ to x given by

$$\lambda \mapsto (\lambda x_1, \ldots, \lambda x_n, \sqrt{r + \lambda^2(t^2 - r)}),$$

where $\lambda \in [0, 1]$, proving that the component of $(0, \ldots, 0, \sqrt{r})$ is connected. Similarly, when $t \leq -\sqrt{r}$, we have the continuous path from $(0, \ldots, 0, -\sqrt{r})$ to x given by

$$\lambda \mapsto (\lambda x_1, \ldots, \lambda x_n, -\sqrt{r + \lambda^2(t^2 - r)}),$$

where $\lambda \in [0, 1]$, proving that the component of $(0, \ldots, 0, -\sqrt{r})$ is connected. We denote the sheet containing $(0, \ldots, 0, \sqrt{r})$ by $\mathcal{H}_n^+(r)$ and sheet containing $(0, \ldots, 0, -\sqrt{r})$ by $\mathcal{H}_n^-(r)$

Since every Lorentz isometry, $A \in \mathbf{SO}(n, 1)$, preserves the Lorentz inner product, we conclude that A globally preserves every hyperboloid, $\mathcal{H}_n(r)$, for $r > 0$. We claim that every $A \in \mathbf{SO}_0(n, 1)$ preserves both $\mathcal{H}_n^+(r)$ and $\mathcal{H}_n^-(r)$. This follows immediately from

Proposition 2.7. *If $a_{n+1\,n+1} > 0$, then every isometry, $A \in \mathbf{O}(n, 1)$, preserves all positive (resp. negative) timelike vectors and all positive (resp. negative) lightlike vectors. Moreover, if $A \in \mathbf{O}(n, 1)$ preserves all positive timelike vectors, then $a_{n+1\,n+1} > 0$.*

Proof. Let $u = (\mathbf{u}, t)$ be a nonzero timelike or lightlike vector. This means that

$$\|\mathbf{u}\|^2 \leq t^2 \quad \text{and} \quad t \neq 0.$$

Since $A \in \mathbf{O}(n,1)$, the matrix A preserves the inner product; if $\langle u, u\rangle = \|\mathbf{u}\|^2 - t^2 < 0$, we get $\langle Au, Au\rangle < 0$, which shows that Au is also timelike. Similarly, if $\langle u, u\rangle = 0$, then $\langle Au, Au\rangle = 0$. As $A \in \mathbf{O}(n,1)$, we know that

$$\langle A_{n+1}, A_{n+1}\rangle = -1,$$

that is,

$$\|\mathbf{A}_{n+1}\|^2 - a^2_{n+1,\,n+1} = -1,$$

where $A_{n+1} = (\mathbf{A}_{n+1}, a_{n+1,\,n+1})$ is the $(n+1)$th row of the matrix A. The $(n+1)$th component of the vector Au is

$$\mathbf{u} \cdot \mathbf{A}_{n+1} + a_{n+1,\,n+1} t.$$

By Cauchy-Schwarz,

$$(\mathbf{u} \cdot \mathbf{A}_{n+1})^2 \leq \|\mathbf{u}\|^2 \, \|\mathbf{A}_{n+1}\|^2,$$

so we get,

$$\begin{aligned}(\mathbf{u} \cdot \mathbf{A}_{n+1})^2 &\leq \|\mathbf{u}\|^2 \, \|\mathbf{A}_{n+1}\|^2 \\ &\leq t^2(a^2_{n+1,\,n+1} - 1) = t^2 a^2_{n+1,\,n+1} - t^2 \\ &< t^2 a^2_{n+1,\,n+1},\end{aligned}$$

since $t \neq 0$. It follows that $\mathbf{u} \cdot \mathbf{A}_{n+1} + a_{n+1,\,n+1} t$ has the same sign as t, since $a_{n+1,\,n+1} > 0$. Consequently, if $a_{n+1,\,n+1} > 0$, we see that A maps positive timelike (resp. lightlike) vectors to positive timelike (resp. lightlike) vectors and similarly with negative timelight (resp. lightlike) vectors.

Conversely, as $e_{n+1} = (0, \ldots, 0, 1)$ is timelike and positive, if A preserves all positive timelike vectors, then Ae_{n+1} is timelike positive, which implies $a_{n+1,\,n+1} > 0$. □

Let $\mathbf{O}^+(n,1)$ denote the subset of $\mathbf{O}(n,1)$ consisting of all matrices, $A = (a_{i\,j})$, such that $a_{n+1\,n+1} > 0$. Using Proposition 2.7, we can now show that $\mathbf{O}^+(n,1)$ is a subgroup of $\mathbf{O}(n,1)$ and that $\mathbf{SO}_0(n,1)$ is a subgroup of $\mathbf{SO}(n,1)$. Recall that

$$\mathbf{SO}_0(n,1) = \{A \in \mathbf{SO}(n,1) \mid a_{n+1\,n+1} > 0\}.$$

Note that $\mathbf{SO}_0(n,1) = \mathbf{O}^+(n,1) \cap \mathbf{SO}(n,1)$.

Proposition 2.8. *The set $\mathbf{O}^+(n,1)$ is a subgroup of $\mathbf{O}(n,1)$ and the set $\mathbf{SO}_0(n,1)$ is a subgroup of $\mathbf{SO}(n,1)$.*

Proof. Let $A \in \mathbf{O}^+(n,1) \subseteq \mathbf{O}(n,1)$, so that $a_{n+1\,n+1} > 0$. The inverse of A in $\mathbf{O}(n,1)$ is $JA^\top J$, where

$$J = \begin{pmatrix} I_n & 0 \\ 0 & -1 \end{pmatrix},$$

which implies that $a^{-1}_{n+1\,n+1} = a_{n+1\,n+1} > 0$ and so, $A^{-1} \in \mathbf{O}^+(n,1)$. If $A, B \in \mathbf{O}^+(n,1)$, then, by Proposition 2.7, both A and B preserve all positive timelike vectors, so AB preserve all positive timelike vectors. By Proposition 2.7, again, $AB \in \mathbf{O}^+(n,1)$. Therefore, $\mathbf{O}^+(n,1)$ is a group. But then, $\mathbf{SO}_0(n,1) = \mathbf{O}^+(n,1) \cap \mathbf{SO}(n,1)$ is also a group. □

Since any matrix, $A \in \mathbf{SO}_0(n,1)$, preserves the Lorentz inner product and all positive timelike vectors and since $\mathcal{H}^+_n(1)$ consists of timelike vectors, we see that every $A \in \mathbf{SO}_0(n,1)$ maps $\mathcal{H}^+_n(1)$ into itself. Similarly, every $A \in \mathbf{SO}_0(n,1)$ maps $\mathcal{H}^-_n(1)$ into itself. Thus, we can define an action $\cdot\colon \mathbf{SO}_0(n,1) \times \mathcal{H}^+_n(1) \longrightarrow \mathcal{H}^+_n(1)$ by

$$A \cdot u = Au$$

and similarly, we have an action $\cdot\colon \mathbf{SO}_0(n,1) \times \mathcal{H}^-_n(1) \longrightarrow \mathcal{H}^-_n(1)$.

Proposition 2.9. *The group $\mathbf{SO}_0(n,1)$ is the subgroup of $\mathbf{SO}(n,1)$ that preserves $\mathcal{H}^+_n(1)$ (and $\mathcal{H}^-_n(1)$) i.e.,*

$$\mathbf{SO}_0(n,1) = \{A \in \mathbf{SO}(n,1) \mid A(\mathcal{H}^+_n(1)) = \mathcal{H}^+_n(1) \quad \textit{and} \quad A(\mathcal{H}^-_n(1)) = \mathcal{H}^-_n(1)\}.$$

Proof. We already observed that $A(\mathcal{H}^+_n(1)) = \mathcal{H}^+_n(1)$ if $A \in \mathbf{SO}_0(n,1)$ (and similarly, $A(\mathcal{H}^-_n(1)) = \mathcal{H}^-_n(1)$). Conversely, for any $A \in \mathbf{SO}(n,1)$ such that $A(\mathcal{H}^+_n(1)) = \mathcal{H}^+_n(1)$, as $e_{n+1} = (0, \ldots, 0, 1) \in \mathcal{H}^+_n(1)$, the vector Ae_{n+1} must be positive timelike, but this says that $a_{n+1,\,n+1} > 0$, i.e., $A \in \mathbf{SO}_0(n,1)$. □

Next, we wish to prove that the action $\mathbf{SO}_0(n,1) \times \mathcal{H}^+_n(1) \longrightarrow \mathcal{H}^+_n(1)$ is transitive. For this, we need the next two propositions.

Proposition 2.10. *Let $u = (\mathbf{u}, t)$ and $v = (\mathbf{v}, s)$ be nonzero vectors in $\mathbb{R}^{n+1}$ with $\langle u, v\rangle = 0$. If u is timelike, then v is spacelike (i.e., $\langle v, v\rangle > 0$).*

Proof. We have $\|\mathbf{u}\|^2 < t^2$, so $t \neq 0$. Since $\mathbf{u} \cdot \mathbf{v} - ts = 0$, we get

$$\langle v, v\rangle = \|\mathbf{v}\|^2 - s^2 = \|\mathbf{v}\|^2 - \frac{(\mathbf{u} \cdot \mathbf{v})^2}{t^2}.$$

But, Cauchy-Schwarz implies that $(\mathbf{u} \cdot \mathbf{v})^2 \leq \|\mathbf{u}\|^2 \|\mathbf{v}\|^2$, so we get

$$\langle v, v\rangle = \|\mathbf{v}\|^2 - \frac{(\mathbf{u} \cdot \mathbf{v})^2}{t^2} > \|\mathbf{v}\|^2 - \frac{(\mathbf{u} \cdot \mathbf{v})^2}{\|\mathbf{u}\|^2} \geq 0,$$

as $\|\mathbf{u}\|^2 < t^2$. □

Lemma 2.10 also holds if $u = (\mathbf{u}, t)$ is a nonzero isotropic vector and $v = (\mathbf{v}, s)$ is a nonzero vector that is not collinear with u: If $\langle u, v\rangle = 0$, then v is spacelike (i.e., $\langle v, v\rangle > 0$). The proof is left as an exercise to the reader.

Proposition 2.11. *The action* $\mathbf{SO}_0(n,1) \times \mathcal{H}_n^+(1) \longrightarrow \mathcal{H}_n^+(1)$ *is transitive.*

Proof. Let $e_{n+1} = (0, \ldots, 0, 1) \in \mathcal{H}_n^+(1)$. It is enough to prove that for every $u = (\mathbf{u}, t) \in \mathcal{H}_n^+(1)$, there is some $A \in \mathbf{SO}_0(n,1)$ such that $Ae_{n+1} = u$. By hypothesis,

$$\langle u, u\rangle = \|\mathbf{u}\|^2 - t^2 = -1.$$

We show that we can construct an orthonormal basis, $e_1, \ldots, e_n, u$, with respect to the Lorentz inner product. Consider the hyperplane

$$H = \{v \in \mathbb{R}^{n+1} \mid \langle u, v\rangle = 0\}.$$

Since u is timelike, by Proposition 2.10, every nonzero vector $v \in H$ is spacelike, i.e., $\langle v, v\rangle > 0$. Let $v_1, \ldots, v_n$ be a basis of H. Since all (nonzero) vectors in H are spacelike, we can apply the Gramm-Schmidt orthonormalization procedure and we get a basis $e_1, \ldots, e_n$, of H, such that

$$\langle e_i, e_j\rangle = \delta_{i\,j}, \quad 1 \leq i, j \leq n.$$

Now, by construction, we also have

$$\langle e_i, u\rangle = 0, \quad 1 \leq i \leq n, \quad \text{and} \quad \langle u, u\rangle = -1.$$

Therefore, $e_1, \ldots, e_n, u$ are the column vectors of a Lorentz matrix, A, such that $Ae_{n+1} = u$, proving our assertion. □

Let us find the stabilizer of $e_{n+1} = (0, \ldots, 0, 1)$. We must have $Ae_{n+1} = e_{n+1}$, and the polar form implies that

$$A = \begin{pmatrix} P & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{with} \quad P \in \mathbf{SO}(n).$$

Therefore, the stabilizer of e_{n+1} is isomorphic to $\mathbf{SO}(n)$ and we conclude that $\mathcal{H}_n^+(1)$, as a homogeneous space, is

$$\mathcal{H}_n^+(1) \cong \mathbf{SO}_0(n,1)/\mathbf{SO}(n).$$

We will show in Section 2.5 that $\mathbf{SO}_0(n,1)$ is connected.

2.4 More on $\mathbf{O}(p,q)$

Recall from Section 2.3 that the group $\mathbf{O}(p,q)$ is the set of all $n \times n$-matrices

$$\mathbf{O}(p,q) = \{A \in \mathbf{GL}(n,\mathbb{R}) \mid A^\top I_{p,q} A = I_{p,q}\}.$$

We deduce immediately that $|\det(A)| = 1$ and we also know that $AI_{p,q}A^\top = I_{p,q}$ holds. Unfortunately, when $p \neq 0, 1$ and $q \neq 0, 1$, it does not seem possible to obtain a formula as nice as that given in Proposition 2.3. Nevertheless, we can obtain a formula for the polar form of matrices in $\mathbf{O}(p, q)$. First, recall (for example, see Gallier [60], Chapter 12) that if S is a symmetric positive definite matrix, then there is a unique symmetric positive definite matrix, T, so that

$$S = T^2.$$

We denote T by $S^{\frac{1}{2}}$ or $\sqrt{S}$. By $S^{-\frac{1}{2}}$, we mean the inverse of $S^{\frac{1}{2}}$. In order to obtain the polar form of a matrix in $\mathbf{O}(p, q)$, we begin with the following proposition:

Proposition 2.12. *Every matrix $X \in \mathbf{O}(p, q)$ can be written as*

$$X = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} \alpha^{\frac{1}{2}} & \alpha^{\frac{1}{2}} Z^\top \\ \delta^{\frac{1}{2}} Z & \delta^{\frac{1}{2}} \end{pmatrix},$$

where $\alpha = (I - Z^\top Z)^{-1}$ and $\delta = (I - ZZ^\top)^{-1}$, for some orthogonal matrices $U \in \mathbf{O}(p)$, $V \in \mathbf{O}(q)$ and for some $q \times p$ matrix, Z, such that $I - Z^\top Z$ and $I - ZZ^\top$ are symmetric positive definite matrices. Moreover, U, V, Z are uniquely determined by X.

Proof. If we write

$$X = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

with A a $p \times p$ matrix, D a $q \times q$ matrix, B a $p \times q$ matrix and C a $q \times p$ matrix, then the equations $A^\top I_{p,q} A = I_{p,q}$ and $AI_{p,q}A^\top = I_{p,q}$ yield the (not independent) conditions

$$\begin{aligned} A^\top A &= I + C^\top C \\ D^\top D &= I + B^\top B \\ A^\top B &= C^\top D \\ AA^\top &= I + BB^\top \\ DD^\top &= I + CC^\top \\ AC^\top &= BD^\top. \end{aligned}$$

Since $C^\top C$ is symmetric and since it is easy to show that $C^\top C$ has nonnegative eigenvalues, we deduce that $A^\top A$ is symmetric positive definite and similarly for $D^\top D$. If we assume that the above decomposition of X holds, we deduce that

$$\begin{aligned} A &= U(I - Z^\top Z)^{-\frac{1}{2}} \\ B &= U(I - Z^\top Z)^{-\frac{1}{2}} Z^\top \\ C &= V(I - ZZ^\top)^{-\frac{1}{2}} Z \\ D &= V(I - ZZ^\top)^{-\frac{1}{2}}, \end{aligned}$$

which implies

$$Z = D^{-1}C \quad \text{and} \quad Z^\top = A^{-1}B.$$

Thus, we must check that

$$(D^{-1}C)^\top = A^{-1}B$$

i.e.,

$$C^\top (D^\top)^{-1} = A^{-1}B,$$

namely,

$$AC^\top = BD^\top,$$

which is indeed the last of our identities. Thus, we must have $Z = D^{-1}C = (A^{-1}B)^\top$. The above expressions for A and D also imply that

$$A^\top A = (I - Z^\top Z)^{-1} \quad \text{and} \quad D^\top D = (I - ZZ^\top)^{-1},$$

so we must check that the choice $Z = D^{-1}C = (A^{-1}B)^\top$ yields the above equations.

Since $Z^\top = A^{-1}B$, we have

$$\begin{aligned} Z^\top Z &= A^{-1}BB^\top (A^\top)^{-1} \\ &= A^{-1}(AA^\top - I)(A^\top)^{-1} \\ &= I - A^{-1}(A^\top)^{-1} \\ &= I - (A^\top A)^{-1}. \end{aligned}$$

Therefore,

$$(A^\top A)^{-1} = I - Z^\top Z,$$

i.e.,

$$A^\top A = (I - Z^\top Z)^{-1},$$

as desired. We also have, this time, with $Z = D^{-1}C$,

$$\begin{aligned} ZZ^\top &= D^{-1}CC^\top (D^\top)^{-1} \\ &= D^{-1}(DD^\top - I)(D^\top)^{-1} \\ &= I - D^{-1}(D^\top)^{-1} \\ &= I - (D^\top D)^{-1}. \end{aligned}$$

Therefore,

$$(D^\top D)^{-1} = I - ZZ^\top,$$

i.e.,

$$D^\top D = (I - ZZ^\top)^{-1},$$

as desired. Now, since $A^\top A$ and $D^\top D$ are positive definite, the polar form implies that

$$A = U(A^\top A)^{\frac{1}{2}} = U(I - Z^\top Z)^{-\frac{1}{2}}$$

and

$$D = V(D^\top D)^{\frac{1}{2}} = V(I - ZZ^\top)^{-\frac{1}{2}},$$

for some unique matrices, $U \in \mathbf{O}(p)$ and $V \in \mathbf{O}(q)$. Since $Z = D^{-1}C$ and $Z^\top = A^{-1}B$, we get $C = DZ$ and $B = AZ^\top$, but this is

$$\begin{aligned} B &= U(I - Z^\top Z)^{-\frac{1}{2}} Z^\top \\ C &= V(I - ZZ^\top)^{-\frac{1}{2}} Z, \end{aligned}$$

as required. Therefore, the unique choice of $Z = D^{-1}C = (A^{-1}B)^\top$, U and V does yield the formula of the proposition. □

It remains to show that the matrix

$$\begin{pmatrix} \alpha^{\frac{1}{2}} & \alpha^{\frac{1}{2}} Z^\top \\ \delta^{\frac{1}{2}} Z & \delta^{\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} (I - Z^\top Z)^{-\frac{1}{2}} & (I - Z^\top Z)^{-\frac{1}{2}} Z^\top \\ (I - ZZ^\top)^{-\frac{1}{2}} Z & (I - ZZ^\top)^{-\frac{1}{2}} \end{pmatrix}$$

is symmetric. To prove this, we will use power series and a continuity argument.

Proposition 2.13. *For any $q \times p$ matrix, Z, such that $I - Z^\top Z$ and $I - ZZ^\top$ are symmetric positive definite, the matrix*

$$S = \begin{pmatrix} \alpha^{\frac{1}{2}} & \alpha^{\frac{1}{2}} Z^\top \\ \delta^{\frac{1}{2}} Z & \delta^{\frac{1}{2}} \end{pmatrix}$$

is symmetric, where $\alpha = (I - Z^\top Z)^{-1}$ and $\delta = (I - ZZ^\top)^{-1}$.

Proof. The matrix S is symmetric iff

$$Z\alpha^{\frac{1}{2}} = \delta^{\frac{1}{2}} Z,$$

i.e., iff

$$Z(I - Z^\top Z)^{-\frac{1}{2}} = (I - ZZ^\top)^{-\frac{1}{2}} Z.$$

Consider the matrices

$$\beta(t) = (I - tZ^\top Z)^{-\frac{1}{2}} \quad \text{and} \quad \gamma(t) = (I - tZZ^\top)^{-\frac{1}{2}},$$

for any t with $0 \leq t \leq 1$. We claim that these matrices make sense. Indeed, since $Z^\top Z$ is symmetric, we can write

$$Z^\top Z = PDP^\top$$

where P is orthogonal and D is a diagonal matrix with nonnegative entries. Moreover, as

$$I - Z^\top Z = P(I - D)P^\top$$

and $I - Z^\top Z$ is positive definite, $0 \leq \lambda < 1$, for every eigenvalue in D. But then, as

$$I - tZ^\top Z = P(I - tD)P^\top,$$

we have $1-t\lambda > 0$ for every λ in D and for all t with $0 \leq t \leq 1$, so that $I - tZ^\top Z$ is positive definite and thus, $(I - tZ^\top Z)^{-\frac{1}{2}}$ is also well defined. A similar argument applies to $(I - tZZ^\top)^{-\frac{1}{2}}$. Observe that

$$\lim_{t\to 1} \beta(t) = \alpha^{\frac{1}{2}}$$

since

$$\beta(t) = (I - tZ^\top Z)^{-\frac{1}{2}} = P(I - tD)^{-\frac{1}{2}} P^\top,$$

where $(I-tD)^{-\frac{1}{2}}$ is a diagonal matrix with entries of the form $(1-t\lambda)^{-\frac{1}{2}}$ and these eigenvalues are continuous functions of t for $t \in [0,1]$. A similar argument shows that

$$\lim_{t\to 1} \gamma(t) = \delta^{\frac{1}{2}}.$$

Therefore, it is enough to show that

$$Z\beta(t) = \gamma(t)Z,$$

with $0 \leq t < 1$ and our result will follow by continuity. However, when $0 \leq t < 1$, the power series for $\beta(t)$ and $\gamma(t)$ converge. Thus, we have

$$\beta(t) = 1 + \frac{1}{2}tZ^\top Z - \frac{1}{8}t^2(Z^\top Z)^2 + \cdots + \frac{\frac{1}{2}\left(\frac{1}{2}-1\right)\cdots\left(\frac{1}{2}-k+1\right)}{k!}t^k(Z^\top Z)^k + \cdots$$

and

$$\gamma(t) = 1 + \frac{1}{2}tZZ^\top - \frac{1}{8}t^2(ZZ^\top)^2 + \cdots + \frac{\frac{1}{2}\left(\frac{1}{2}-1\right)\cdots\left(\frac{1}{2}-k+1\right)}{k!}t^k(ZZ^\top)^k + \cdots$$

and we get

$$Z\beta(t) = Z + \frac{1}{2}tZZ^\top Z - \frac{1}{8}t^2Z(Z^\top Z)^2 + \cdots + \frac{\frac{1}{2}\left(\frac{1}{2}-1\right)\cdots\left(\frac{1}{2}-k+1\right)}{k!}t^kZ(Z^\top Z)^k + \cdots$$

and

$$\gamma(t)Z = Z + \frac{1}{2}tZZ^\top Z - \frac{1}{8}t^2(ZZ^\top)^2Z + \cdots + \frac{\frac{1}{2}\left(\frac{1}{2}-1\right)\cdots\left(\frac{1}{2}-k+1\right)}{k!}t^k(ZZ^\top)^kZ + \cdots.$$

However

$$Z(Z^\top Z)^k = Z\underbrace{Z^\top Z\cdots Z^\top Z}_{k} = \underbrace{ZZ^\top\cdots ZZ^\top}_{k}Z = (ZZ^\top)^kZ,$$

which proves that $Z\beta(t) = \gamma(t)Z$, as required. □

Another proof of Proposition 2.13 can be given using the SVD of Z. Indeed, we can write

$$Z = PDQ^\top$$

where P is a $q \times q$ orthogonal matrix, Q is a $p \times p$ orthogonal matrix and D is a $q \times p$ matrix whose diagonal entries are (strictly) positive and all other entries zero. Then,

$$I - Z^\top Z = I - QD^\top P^\top PDQ^\top = Q(I - D^\top D)Q^\top,$$

a symmetric positive definite matrix. We also have

$$I - ZZ^\top = I - PDQ^\top QD^\top P^\top = P(I - DD^\top)P^\top,$$

another symmetric positive definite matrix. Then,

$$Z(I - Z^\top Z)^{-\frac{1}{2}} = PDQ^\top Q(I - D^\top D)^{-\frac{1}{2}}Q^\top = PD(I - D^\top D)^{-\frac{1}{2}}Q^\top$$

and

$$(I - ZZ^\top)^{-\frac{1}{2}} = P(I - DD^\top)^{-\frac{1}{2}}P^\top PDQ^\top = P(I - DD^\top)^{-\frac{1}{2}}DQ^\top,$$

so it suffices to prove that

$$D(I - D^\top D)^{-\frac{1}{2}} = (I - DD^\top)^{-\frac{1}{2}}D.$$

However, D is essentially a diagonal matrix and the above is easily verified, as the reader should check.

Remark: The polar form can also be obtained *via* the exponential map and the Lie algebra, $\mathfrak{o}(p, q)$, of $\mathbf{O}(p, q)$, see Section 5.6.

We also have the following amusing property of the determinants of A and D:

Proposition 2.14. *For any matrix $X \in \mathbf{O}(p, q)$, if we write*

$$X = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

then

$$\det(X) = \det(A)\det(D)^{-1} \quad \textit{and} \quad |\det(A)| = |\det(D)| \geq 1.$$

Proof. Using the identities $A^\top B = C^\top D$ and $D^\top D = I + B^\top B$ proved earlier, observe that

$$\begin{pmatrix} A^\top & 0 \\ B^\top & -D^\top \end{pmatrix}\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A^\top A & A^\top B \\ B^\top A - D^\top C & B^\top B - D^\top D \end{pmatrix} = \begin{pmatrix} A^\top A & A^\top B \\ 0 & -I_q \end{pmatrix}.$$

If we compute determinants, we get

$$\det(A)(-1)^q \det(D)\det(X) = \det(A)^2(-1)^q.$$

It follows that

$$\det(X) = \det(A)\det(D)^{-1}.$$

From $A^\top A = I + C^\top C$ and $D^\top D = I + B^\top B$, we conclude that $\det(A) \geq 1$ and $\det(D) \geq 1$. Since $|\det(X)| = 1$, we have $|\det(A)| = |\det(D)| \geq 1$. □

Remark: It is easy to see that the equations relating A, B, C, D established in the proof of Proposition 2.12 imply that

$$\det(A) = \pm 1 \quad \text{iff} \quad C = 0 \quad \text{iff} \quad B = 0 \quad \text{iff} \quad \det(D) = \pm 1.$$

2.5 Topological Groups

Since Lie groups are topological groups (and manifolds), it is useful to gather a few basic facts about topological groups.

Definition 2.11. A set, G, is a *topological group* iff

(a) G is a Hausdorff topological space;

(b) G is a group (with identity 1);

(c) Multiplication, $\cdot\colon G \times G \to G$, and the inverse operation, $G \longrightarrow G\colon g \mapsto g^{-1}$, are continuous, where $G \times G$ has the product topology.

It is easy to see that the two requirements of condition (c) are equivalent to

(c′) The map $G \times G \longrightarrow G\colon (g, h) \mapsto gh^{-1}$ is continuous.

Given a topological group G, for every $a \in G$ we define *left translation* as the map, $L_a\colon G \to G$, such that $L_a(b) = ab$, for all $b \in G$, and *right translation* as the map, $R_a\colon G \to G$, such that $R_a(b) = ba$, for all $b \in G$. Observe that $L_{a^{-1}}$ is the inverse of L_a and similarly, $R_{a^{-1}}$ is the inverse of R_a. As multiplication is continuous, we see that L_a and R_a are continuous. Moreover, since they have a continuous inverse, they are homeomorphisms. As a consequence, if U is an open subset of G, then so is $gU = L_g(U)$ (resp. $Ug = R_gU$), for all $g \in G$. Therefore, the topology of a topological group (i.e., its family of open sets) is *determined* by the knowledge of the open subsets containing the identity, 1.

Given any subset, $S \subseteq G$, let $S^{-1} = \{s^{-1} \mid s \in S\}$; let $S^0 = \{1\}$ and $S^{n+1} = S^nS$, for all $n \geq 0$. Property (c) of Definition 2.11 has the following useful consequences:

Proposition 2.15. *If G is a topological group and U is any open subset containing 1, then there is some open subset, $V \subseteq U$, with $1 \in V$, so that $V = V^{-1}$ and $V^2 \subseteq U$. Furthermore, $\overline{V} \subseteq U$.*

Proof. Since multiplication $G \times G \longrightarrow G$ is continuous and $G \times G$ is given the product topology, there are open subsets, U_1 and U_2, with $1 \in U_1$ and $1 \in U_2$, so that $U_1U_2 \subseteq U$. Ley $W = U_1 \cap U_2$ and $V = W \cap W^{-1}$. Then, V is an open set containing 1 and, clearly, $V = V^{-1}$ and $V^2 \subseteq U_1U_2 \subseteq U$. If $g \in \overline{V}$, then gV is an open set containing g (since $1 \in V$) and thus, $gV \cap V \neq \emptyset$. This means that there are some $h_1, h_2 \in V$ so that $gh_1 = h_2$, but then, $g = h_2h_1^{-1} \in VV^{-1} = VV \subseteq U$. □

A subset, U, containing 1 and such that $U = U^{-1}$, is called *symmetric*. Using Proposition 2.15, we can give a very convenient characterization of the Hausdorff separation property in a topological group.

Proposition 2.16. *If G is a topological group, then the following properties are equivalent:*

(1) G is Hausdorff;

(2) The set $\{1\}$ is closed;

(3) The set $\{g\}$ is closed, for every $g \in G$.

Proof. The implication (1) $\longrightarrow$ (2) is true in any Hausdorff topological space. We just have to prove that $G - \{1\}$ is open, which goes as follows: For any $g \neq 1$, since G is Hausdorff, there exists disjoint open subsets U_g and V_g, with $g \in U_g$ and $1 \in V_g$. Thus, $\bigcup U_g = G - \{1\}$, showing that $G - \{1\}$ is open. Since L_g is a homeomorphism, (2) and (3) are equivalent. Let us prove that (3) $\longrightarrow$ (1). Let $g_1, g_2 \in G$ with $g_1 \neq g_2$. Then, $g_1^{-1}g_2 \neq 1$ and if U and V are distinct open subsets such that $1 \in U$ and $g_1^{-1}g_2 \in V$, then $g_1 \in g_1U$ and $g_2 \in g_1V$, where g_1U and g_1V are still open and disjoint. Thus, it is enough to separate 1 and $g \neq 1$. Pick any $g \neq 1$. If every open subset containing 1 also contained g, then 1 would be in the closure of $\{g\}$, which is absurd, since $\{g\}$ is closed and $g \neq 1$. Therefore, there is some open subset, U, such that $1 \in U$ and $g \notin U$. By Proposition 2.15, we can find an open subset, V, containing 1, so that $VV \subseteq U$ and $V = V^{-1}$. We claim that V and Vg are disjoint open sets with $1 \in V$ and $g \in gV$.

Since $1 \in V$, it is clear that $1 \in V$ and $g \in gV$. If we had $V \cap gV \neq \emptyset$, then we would have $g \in VV^{-1} = VV \subseteq U$, a contradiction. □

If H is a subgroup of G (not necessarily normal), we can form the set of left cosets, G/H and we have the projection, $p\colon G \to G/H$, where $p(g) = gH = \overline{g}$. If G is a topological group, then G/H can be given the *quotient topology*, where a subset $U \subseteq G/H$ is open iff $p^{-1}(U)$ is open in G. With this topology, p is continuous. The trouble is that G/H is not necessarily Hausdorff. However, we can neatly characterize when this happens.

Proposition 2.17. *If G is a topological group and H is a subgroup of G then the following properties hold:*

(1) The map $p\colon G \to G/H$ is an open map, which means that $p(V)$ is open in G/H whenever V is open in G.

(2) The space G/H is Hausdorff iff H is closed in G.

(3) If H is open, then H is closed and G/H has the discrete topology (every subset is open).

(4) The subgroup H is open iff $1 \in \overset{\circ}{H}$ (i.e., there is some open subset, U, so that $1 \in U \subseteq H$).

Proof. (1) Observe that if V is open in G, then $VH = \bigcup_{h \in H} Vh$ is open, since each Vh is open (as right translation is a homeomorphism). However, it is clear that

$$p^{-1}(p(V)) = VH,$$

i.e., $p^{-1}(p(V))$ is open, which, by definition, means that $p(V)$ is open.

(2) If G/H is Hausdorff, then by Proposition 2.16, every point of G/H is closed, i.e., each coset gH is closed, so H is closed. Conversely, assume H is closed. Let $\overline{x}$ and $\overline{y}$ be two distinct point in G/H and let $x, y \in G$ be some elements with $p(x) = \overline{x}$ and $p(y) = \overline{y}$. As $\overline{x} \neq \overline{y}$, the elements x and y are not in the same coset, so $x \notin yH$. As H is closed, so is yH, and since $x \notin yH$, there is some open containing x which is disjoint from yH, and we may assume (by translation) that it is of the form Ux, where U is an open containing 1. By Proposition 2.15, there is some open V containing 1 so that $VV \subseteq U$ and $V = V^{-1}$. Thus, we have

$$V^2 x \cap yH = \emptyset$$

and in fact,

$$V^2 xH \cap yH = \emptyset,$$

since H is a group. Since $V = V^{-1}$, we get

$$VxH \cap VyH = \emptyset,$$

and then, since V is open, both VxH and VyH are disjoint, open, so $p(VxH)$ and $p(VyH)$ are open sets (by (1)) containing $\overline{x}$ and $\overline{y}$ respectively and $p(VxH)$ and $p(VyH)$ are disjoint (because $p^{-1}(p(VxH)) = VxHH = VxH$ and $p^{-1}(p(VyH)) = VyHH = VyH$ and $VxH \cap VyH = \emptyset$).

(3) If H is open, then every coset gH is open, so every point of G/H is open and G/H is discrete. Also, $\bigcup_{g \notin H} gH$ is open, i.e., H is closed.

(4) Say U is an open subset such that $1 \in U \subseteq H$. Then, for every $h \in H$, the set hU is an open subset of H with $h \in hU$, which shows that H is open. The converse is trivial. □

Proposition 2.18. *If G is a connected topological group, then G is generated by any symmetric neighborhood, V, of 1. In fact,*

$$G = \bigcup_{n \geq 1} V^n.$$

Proof. Since $V = V^{-1}$, it is immediately checked that $H = \bigcup_{n \geq 1} V^n$ is the group generated by V. As V is a neighborhood of 1, there is some open subset, $U \subseteq V$, with $1 \in U$, and so $1 \in \overset{\circ}{H}$. From Proposition 2.17, the subgroup H is open and closed and since G is connected, $H = G$. □

A subgroup, H, of a topological group G is *discrete* iff the induced topology on H is discrete, i.e., for every $h \in H$, there is some open subset, U, of G so that $U \cap H = \{h\}$.

Proposition 2.19. *If G is a topological group and H is discrete subgroup of G, then H is closed.*

Proof. As H is discrete, there is an open subset, U, of G so that $U \cap H = \{1\}$, and by Proposition 2.15, we may assume that $U = U^{-1}$. If $g \in \overline{H}$, as gU is an open set containing g, we have $gU \cap H \neq \emptyset$. Consequently, there is some $y \in gU \cap H = gU^{-1} \cap H$, so $g \in yU$ with $y \in H$. Thus, we have

$$g \in yU \cap \overline{H} \subseteq \overline{yU \cap H} = \overline{\{y\}} = \{y\},$$

since $U \cap H = \{1\}$, $y \in H$ and G is Hausdorff. Therefore, $g = y \in H$. □

Proposition 2.20. *If G is a topological group and H is any subgroup of G, then the closure, $\overline{H}$, of H is a subgroup of G.*

Proof. This follows easily from the continuity of multiplication and of the inverse operation, the details are left as an exercise to the reader. □

Proposition 2.21. *Let G be a topological group and H be any subgroup of G. If H and G/H are connected, then G is connected.*

Proof. It is a standard fact of topology that a space G is connected iff every continuous function, f, from G to the discrete space $\{0, 1\}$ is constant. Pick any continuous function, f, from G to $\{0, 1\}$. As H is connected and left translations are homeomorphisms, all cosets, gH, are connected. Thus, f is constant on every coset, gH. Thus, the function $f\colon G \to \{0, 1\}$ induces a continuous function, $\overline{f}\colon G/H \to \{0, 1\}$, such that $f = \overline{f} \circ p$ (where $p\colon G \to G/H$; the continuity of $\overline{f}$ follows immediately from the definition of the quotient topology on G/H). As G/H is connected, $\overline{f}$ is constant and so, $f = \overline{f} \circ p$ is constant. □

Proposition 2.22. *Let G be a topological group and let V be any connected symmetric open subset containing 1. Then, if G_0 is the connected component of the identity, we have*

$$G_0 = \bigcup_{n \geq 1} V^n$$

and G_0 is a normal subgroup of G. Moreover, the group G/G_0 is discrete.

Proof. First, as V is open, every V^n is open, so the group $\bigcup_{n \geq 1} V^n$ is open, and thus closed, by Proposition 2.17 (3). For every $n \geq 1$, we have the continuous map

$$\underbrace{V \times \cdots \times V}_{n} \longrightarrow V^n : (g_1, \ldots, g_n) \mapsto g_1 \cdots g_n.$$

As V is connected, $V \times \cdots \times V$ is connected and so, V^n is connected. Since $1 \in V^n$ for all $n \geq 1$, and every V^n is connected, we conclude that $\bigcup_{n \geq 1} V^n$ is connected. Now, $\bigcup_{n \geq 1} V^n$ is connected, open and closed, so it is the connected component of 1. Finally, for every $g \in G$, the group gG_0g^{-1} is connected and contains 1, so it is contained in G_0, which proves that G_0 is normal. Since G_0 is open, the group G/G_0 is discrete. □

A topological space, X, is *locally compact* iff for every point $p \in X$, there is a compact neighborhood, C of p, i.e., there is a compact, C, and an open, U, with $p \in U \subseteq C$. For example, manifolds are locally compact.

Proposition 2.23. *Let G be a topological group and assume that G is connected and locally compact. Then, G is countable at infinity, which means that G is the union of a countable family of compact subsets. In fact, if V is any symmetric compact neighborhood of 1, then*

$$G = \bigcup_{n \geq 1} V^n.$$

Proof. Since G is locally compact, there is some compact neighborhood, K, of 1. Then, $V = K \cap K^{-1}$ is also compact and a symmetric neigborhood of 1. By Proposition 2.18, we have

$$G = \bigcup_{n \geq 1} V^n.$$

An argument similar to the one used in the proof of Proposition 2.22 to show that V^n is connected if V is connected proves that each V^n compact if V is compact. □

If a topological group, G acts on a topological space, X, and the action $\cdot : G \times X \to X$ is continuous, we say that G acts *continuously on* X. Under some mild assumptions on G and X, the quotient space, G/G_x, is homeomorphic to X. For example, this happens if X is a Baire space.

Recall that a *Baire space*, X, is a topological space with the property that if $\{F\}_{i \geq 1}$ is any countable family of closed sets, F_i, such that each F_i has empty interior, then $\bigcup_{i \geq 1} F_i$ also has empty interior. By complementation, this is equivalent to the fact that for every countable family of open sets, U_i, such that each U_i is dense in X (i.e., $\overline{U}_i = X$), then $\bigcap_{i \geq 1} U_i$ is also dense in X.

Remark: A subset, $A \subseteq X$, is *rare* if its closure, $\overline{A}$, has empty interior. A subset, $Y \subseteq X$, is *meager* if it is a countable union of rare sets. Then, it is immediately verified that a space, X, is a Baire space iff every nonempty open subset of X is not meager.

The following theorem shows that there are plenty of Baire spaces:

Theorem 2.24. *(Baire) (1) Every locally compact topological space is a Baire space.*

(2) Every complete metric space is a Baire space.

A proof of Theorem 2.24 can be found in Bourbaki [24], Chapter IX, Section 5, Theorem 1.

We can now greatly improve Proposition 2.2 when G and X are topological spaces having some "nice" properties.

Theorem 2.25. *Let G be a topological group which is locally compact and countable at infinity, X a Hausdorff topological space which is a Baire space and assume that G acts transitively and continuously on X. Then, for any $x \in X$, the map $\varphi\colon G/G_x \to X$ is a homeomorphism.*

Proof. We follow the proof given in Bourbaki [24], Chapter IX, Section 5, Proposition 6 (Essentially the same proof can be found in Mneimné and Testard [112], Chapter 2). First, observe that if a topological group acts continuously and transitively on a Hausdorff topological space, then for every $x \in X$, the stabilizer, G_x, is a closed subgroup of G. This is because, as the action is continuous, the projection $\pi\colon G \longrightarrow X\colon g \mapsto g \cdot x$ is continuous, and $G_x = \pi^{-1}(\{x\})$, with $\{x\}$ closed. Therefore, by Proposition 2.17, the quotient space, G/G_x, is Hausdorff. As the map $\pi\colon G \longrightarrow X$ is continuous, the induced map $\varphi\colon G/G_x \to X$ is continuous and by Proposition 2.2, it is a bijection. Therefore, to prove that φ is a homeomorphism, it is enough to prove that φ is an open map. For this, it suffices to show that π is an open map. Given any open, U, in G, we will prove that for any $g \in U$, the element $\pi(g) = g \cdot x$ is contained in the interior of $U \cdot x$. However, observe that this is equivalent to proving that x belongs to the interior of $(g^{-1} \cdot U) \cdot x$. Therefore, we are reduced to the case: If U is any open subset of G containing 1, then x belongs to the interior of $U \cdot x$.

Since G is locally compact, using Proposition 2.15, we can find a compact neighborhood of the form $W = \overline{V}$, such that $1 \in W$, $W = W^{-1}$ and $W^2 \subseteq U$, where V is open with $1 \in V \subseteq U$. As G is countable at infinity, $G = \bigcup_{i \geq 1} K_i$, where each K_i is compact. Since V is open, all the cosets gV are open, and as each K_i is covered by the gV's, by compactness of K_i, finitely many cosets gV cover each K_i and so,

$$G = \bigcup_{i \geq 1} g_i V = \bigcup_{i \geq 1} g_i W,$$

for countably many $g_i \in G$, where each $g_i W$ is compact. As our action is transitive, we deduce that

$$X = \bigcup_{i \geq 1} g_i W \cdot x,$$

where each $g_i W \cdot x$ is compact, since our action is continuous and the $g_i W$ are compact. As X is Hausdorff, each $g_i W \cdot x$ is closed and as X is a Baire space expressed as a union of closed sets, one of the $g_i W \cdot x$ must have nonempty interior, i.e., there is some $w \in W$, with $g_i w \cdot x$ in the interior of $g_i W \cdot x$, for some i. But then, as the map $y \mapsto g \cdot y$ is a homeomorphism for any given $g \in G$ (where $y \in X$), we see that x is in the interior of

$$w^{-1} g_i^{-1} \cdot (g_i W \cdot x) = w^{-1} W \cdot x \subseteq W^{-1} W \cdot x = W^2 \cdot x \subseteq U \cdot x,$$

as desired. □

By Theorem 2.24, we get the following important corollary:

Theorem 2.26. *Let G be a topological group which is locally compact and countable at infinity, X a Hausdorff locally compact topological space and assume that G acts transitively and continuously on X. Then, for any $x \in X$, the map $\varphi\colon G/G_x \to X$ is a homeomorphism.*

As an application of Theorem 2.26 and Proposition 2.21, we show that the Lorentz group $\mathbf{SO}_0(n,1)$ is connected. Firstly, it is easy to check that $\mathbf{SO}_0(n,1)$ and $\mathcal{H}_n^+(1)$ satisfy the assumptions of Theorem 2.26 because they are both manifolds, although this notion has not been discussed yet (but will be in Chapter 3). Also, we saw at the end of Section 2.3 that the action $\cdot\colon \mathbf{SO}_0(n,1) \times \mathcal{H}_n^+(1) \longrightarrow \mathcal{H}_n^+(1)$ of $\mathbf{SO}_0(n,1)$ on $\mathcal{H}_n^+(1)$ is transitive, so that, as topological spaces

$$\mathbf{SO}_0(n,1)/\mathbf{SO}(n) \cong \mathcal{H}_n^+(1).$$

Now, we already showed that $\mathcal{H}_n^+(1)$ is connected so, by Proposition 2.21, the connectivity of $\mathbf{SO}_0(n,1)$ follows from the connectivity of $\mathbf{SO}(n)$ for $n \geq 1$. The connectivity of $\mathbf{SO}(n)$ is a consequence of the surjectivity of the exponential map (for instance, see Gallier [60], Chapter 14) but we can also give a quick proof using Proposition 2.21. Indeed, $\mathbf{SO}(n+1)$ and S^n are both manifolds and we saw in Section 2.2 that

$$\mathbf{SO}(n+1)/\mathbf{SO}(n) \cong S^n.$$

Now, S^n is connected for $n \geq 1$ and $\mathbf{SO}(1) \cong S^1$ is connected. We finish the proof by induction on n.

Corollary 2.27. *The Lorentz group $\mathbf{SO}_0(n,1)$ is connected; it is the component of the identity in $\mathbf{O}(n,1)$.*

Readers who wish to learn more about topological groups may consult Sagle and Walde [130] and Chevalley [34] for an introductory account, and Bourbaki [23], Weil [150] and Pontryagin [123, 124], for a more comprehensive account (especially the last two references).

Chapter 3

Manifolds

3.1 Charts and Manifolds

In Chapter 1 we defined the notion of a manifold embedded in some ambient space, $\mathbb{R}^N$. In order to maximize the range of applications of the theory of manifolds it is necessary to generalize the concept of a manifold to spaces that are not a priori embedded in some $\mathbb{R}^N$. The basic idea is still that, whatever a manifold is, it is a topological space that can be covered by a collection of open subsets, U_α, where each U_α is isomorphic to some "standard model," *e.g.*, some open subset of Euclidean space, $\mathbb{R}^n$. Of course, manifolds would be very dull without functions defined on them and between them. This is a general fact learned from experience: *Geometry arises not just from spaces but from spaces and interesting classes of functions between them*. In particular, we still would like to "do calculus" on our manifold and have good notions of curves, tangent vectors, differential forms, etc. The small drawback with the more general approach is that the definition of a tangent vector is more abstract. We can still define the notion of a curve on a manifold, but such a curve does not live in any given $\mathbb{R}^n$, so it it not possible to define tangent vectors in a simple-minded way using derivatives. Instead, we have to resort to the notion of chart. This is not such a strange idea. For example, a geography atlas gives a set of maps of various portions of the earth and this provides a very good description of what the earth is, without actually imagining the earth embedded in 3-space.

The material of this chapter borrows from many sources, including Warner [148], Berger and Gostiaux [17], O'Neill [120], Do Carmo [51, 50], Gallot, Hulin and Lafontaine [61], Lang [96], Schwartz [136], Hirsch [77], Sharpe [140], Guillemin and Pollack [70], Lafontaine [93], Dubrovin, Fomenko and Novikov [53] and Boothby [18]. A nice (not very technical) exposition is given in Morita [115] (Chapter 1). The recent book by Tu [146] is also highly recommended for its clarity. Among the many texts on manifolds and differential geometry, the book by Choquet-Bruhat, DeWitt-Morette and Dillard-Bleick [37] stands apart because it is one of the clearest and most comprehensive (many proofs are omitted, but this can be an advantage!) Being written for (theoretical) physicists, it contains more examples and applications than most other sources.

Given $\mathbb{R}^n$, recall that the projection functions, $pr_i\colon \mathbb{R}^n \to \mathbb{R}$, are defined by

$$pr_i(x_1, \ldots, x_n) = x_i, \quad 1 \leq i \leq n.$$

For technical reasons (in particular, to ensure the existence of partitions of unity, a crucial tool in manifold theory; see Sections 3.6 and 3.8) and to avoid "esoteric" manifolds that do not arise in practice, from now on, all topological spaces under consideration will be assumed to be Hausdorff and second-countable (which means that the topology has a countable basis).

The first step in generalizing the notion of a manifold is to define charts, a way to say that locally, a manifold "looks like" an open subset of $\mathbb{R}^n$.

Definition 3.1. Given a topological space, M, a *chart* (or *local coordinate map*) is a pair, (U, φ), where U is an open subset of M and $\varphi\colon U \to \Omega$ is a homeomorphism onto an open subset, $\Omega = \varphi(U)$, of $\mathbb{R}^{n_\varphi}$ (for some $n_\varphi \geq 1$). For any $p \in M$, a chart, (U, φ), is a *chart at* p iff $p \in U$. If (U, φ) is a chart, then the functions $x_i = pr_i \circ \varphi$ are called *local coordinates* and for every $p \in U$, the tuple $(x_1(p), \ldots, x_n(p))$ is the set of *coordinates of* p w.r.t. the chart. The inverse, (Ω, φ^{-1}), of a chart is called a *local parametrization*. Given any two charts, (U_i, φ_i) and (U_j, φ_j), if $U_i \cap U_j \neq \emptyset$, we have the *transition maps*, $\varphi_i^j\colon \varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j)$ and $\varphi_j^i\colon \varphi_j(U_i \cap U_j) \to \varphi_i(U_i \cap U_j)$, defined by

$$\varphi_i^j = \varphi_j \circ \varphi_i^{-1} \quad \text{and} \quad \varphi_j^i = \varphi_i \circ \varphi_j^{-1}.$$

Clearly, $\varphi_j^i = (\varphi_i^j)^{-1}$. Observe that the transition maps φ_i^j (resp. φ_j^i) are maps between *open subsets of* $\mathbb{R}^n$. This is good news! Indeed, the whole arsenal of calculus is available for functions on $\mathbb{R}^n$, and we will be able to promote many of these results to manifolds by imposing suitable conditions on transition functions.

As in Section 1.8, whatever our generalized notion of a manifold is, we would like to define the notion of tangent space at a point of manifold, the notion of smooth function between manifolds, and the notion of derivative of a function (at a point) between manifolds. Unfortunately, even though our parametrizations $\varphi^{-1}\colon \Omega \to U$ are homeomorphisms, since U is a subset of a space M *which is not assumed to be contained in* $\mathbb{R}^N$ *(for any* N*)*, the derivative $d\varphi_{t_0}^{-1}$ does not make sense, unlike in the situation of Definition 1.11. Therefore, some extra conditions on the charts must be imposed in order to recapture the fact that for manifolds embedded in $\mathbb{R}^N$, the parametrizations are immersions. An invaluable hint is provided by Lemma 1.22: we require the transition maps $\varphi_i^j\colon \varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j)$ to be sufficiently differentiable. This makes perfect sense since the φ_i^j are functions between open subsets of $\mathbb{R}^n$. It also turns out that these conditions on transition maps guarantee that notions, such as tangent vectors, whose definition seems to depend on the choice of a chart, are in fact independent of the choice of charts. The above motivations suggest the following requirements on charts.

Definition 3.2. Given a topological space, M, given some integer $n \geq 1$ and given some k such that k is either an integer $k \geq 1$ or $k = \infty$, a C^k *n-atlas* (or *n-atlas of class C^k*), $\mathcal{A}$, is a family of charts, $\{(U_i, \varphi_i)\}$, such that

(1) $\varphi_i(U_i) \subseteq \mathbb{R}^n$ for all i;

(2) The U_i cover M, i.e.,
$$M = \bigcup_i U_i;$$

(3) Whenever $U_i \cap U_j \neq \emptyset$, the transition map φ_i^j (and φ_j^i) is a C^k-diffeomorphism. When $k = \infty$, the φ_i^j are smooth diffeomorphisms.

We must ensure that we have enough charts in order to carry out our program of generalizing calculus on $\mathbb{R}^n$ to manifolds. For this, we must be able to add new charts whenever necessary, provided that they are consistent with the previous charts in an existing atlas. Technically, given a C^k n-atlas, $\mathcal{A}$, on M, for any other chart, (U, φ), we say that (U, φ) is *compatible* with the atlas $\mathcal{A}$ iff every map $\varphi_i \circ \varphi^{-1}$ and $\varphi \circ \varphi_i^{-1}$ is C^k (whenever $U \cap U_i \neq \emptyset$). Two atlases $\mathcal{A}$ and $\mathcal{A}'$ on M are *compatible* iff every chart of one is compatible with the other atlas. This is equivalent to saying that the union of the two atlases is still an atlas. It is immediately verified that compatibility induces an equivalence relation on C^k n-atlases on M. In fact, given an atlas, $\mathcal{A}$, for M, the collection, $\widetilde{\mathcal{A}}$, of all charts compatible with $\mathcal{A}$ is a maximal atlas in the equivalence class of charts compatible with $\mathcal{A}$. Finally, we have our generalized notion of a manifold.

Definition 3.3. Given some integer $n \geq 1$ and given some k such that k is either an integer $k \geq 1$ or $k = \infty$, a *C^k-manifold of dimension n* consists of a topological space, M, together with an equivalence class, $\overline{\mathcal{A}}$, of C^k n-atlases, on M. Any atlas, $\mathcal{A}$, in the equivalence class $\overline{\mathcal{A}}$ is called a *differentiable structure of class C^k (and dimension n) on M*. We say that M is *modeled on* $\mathbb{R}^n$. When $k = \infty$, we say that M is a *smooth manifold*.

Remark: It might have been better to use the terminology *abstract manifold* rather than manifold, to emphasize the fact that the space M is not a priori a subspace of $\mathbb{R}^N$, for some suitable N.

We can allow $k = 0$ in the above definitions. In this case, condition (3) in Definition 3.2 is void, since a C^0-diffeomorphism is just a homeomorphism, but φ_i^j is always a homeomorphism. In this case, M is called a *topological manifold of dimension n*. We do not require a manifold to be connected but we require all the components to have the same dimension, n. Actually, on every connected component of M, it can be shown that the dimension, n_φ, of the range of every chart is the same. This is quite easy to show if $k \geq 1$ but for $k = 0$, this requires a deep theorem of Brouwer. (Brouwer's *Invariance of Domain Theorem* states that if $U \subseteq \mathbb{R}^n$ is an open set and if $f\colon U \to \mathbb{R}^n$ is a continuous and injective map, then $f(U)$ is open in $\mathbb{R}^n$. Using Brouwer's Theorem, we can show the following fact: If $U \subseteq \mathbb{R}^m$ and $V \subseteq \mathbb{R}^n$ are two open subsets and if $f\colon U \to V$ is a homeomorphism between U and V, then $m = n$. If $m > n$, then consider the injection, $i\colon \mathbb{R}^n \to \mathbb{R}^m$, where $i(x) = (x, 0_{m-n})$. Clearly, i is injective and continuous, so $i \circ f\colon U \to i(V)$ is injective and continuous and Brouwer's

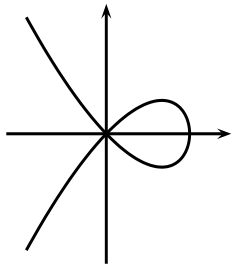


Figure 3.1: A nodal cubic; not a manifold

Theorem implies that $i(V)$ is open in $\mathbb{R}^m$, which is a contradiction, as $i(V) = V \times \{0_{m-n}\}$ is not open in $\mathbb{R}^m$. If $m < n$, consider the homeomorphism $f^{-1}\colon V \to U$.)

What happens if $n = 0$? In this case, every one-point subset of M is open, so every subset of M is open, i.e., M is any (countable if we assume M to be second-countable) set with the discrete topology!

Observe that since $\mathbb{R}^n$ is locally compact and locally connected, so is every manifold (check this!).

In order to get a better grasp of the notion of manifold it is useful to consider examples of non-manifolds. First, consider the curve in $\mathbb{R}^2$ given by the zero locus of the equation

$$y^2 = x^2 - x^3,$$

namely, the set of points

$$M_1 = \{(x, y) \in \mathbb{R}^2 \mid y^2 = x^2 - x^3\}.$$

This curve showed in Figure 3.1 and called a *nodal cubic* is also defined as the parametric curve

$$\begin{aligned} x &= 1 - t^2 \\ y &= t(1 - t^2). \end{aligned}$$

We claim that M_1 is not even a topological manifold. The problem is that the nodal cubic has a self-intersection at the origin. If M_1 was a topological manifold, then there would be a connected open subset, $U \subseteq M_1$, containing the origin, $O = (0, 0)$, namely the intersection of a small enough open disc centered at O with M_1, and a local chart, $\varphi\colon U \to \Omega$, where Ω is some connected open subset of $\mathbb{R}$ (that is, an open interval), since φ is a homeomorphism. However, $U - \{O\}$ consists of four disconnected components and $\Omega - \varphi(O)$ of two disconnected components, contradicting the fact that φ is a homeomorphism.

Let us now consider the curve in $\mathbb{R}^2$ given by the zero locus of the equation

$$y^2 = x^3,$$

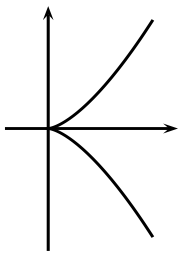


Figure 3.2: A Cuspidal Cubic

namely, the set of points

$$M_2 = \{(x, y) \in \mathbb{R}^2 \mid y^2 = x^3\}.$$

This curve showed in Figure 3.2 and called a *cuspidal cubic* is also defined as the parametric curve

$$\begin{aligned} x &= t^2 \\ y &= t^3. \end{aligned}$$

Consider the map, $\varphi\colon M_2 \to \mathbb{R}$, given by

$$\varphi(x, y) = y^{1/3}.$$

Since $x = y^{2/3}$ on M_2, we see that φ^{-1} is given by

$$\varphi^{-1}(t) = (t^2, t^3)$$

and clearly, φ is a homeomorphism, so M_2 is a topological manifold. However, with the atlas consisting of the single chart $\{\varphi\colon M_2 \to \mathbb{R}\}$, the space M_2 is also a smooth manifold! Indeed, as there is a single chart, condition (3) of Definition 3.2 holds vacuously.

This fact is somewhat unexpected because the cuspidal cubic is not smooth at the origin, since the tangent vector of the parametric curve, $c\colon t \mapsto (t^2, t^3)$, at the origin is the zero vector (the velocity vector at t, is $c'(t) = (2t, 3t^2)$). However, this apparent paradox has to do with the fact that, as a parametric curve, M_2 is not immersed in $\mathbb{R}^2$ since c' is not injective (see Definition 3.24 (a)), whereas as an abstract manifold, with this single chart, M_2 is diffeomorphic to $\mathbb{R}$.

Now, we also have the chart, $\psi\colon M_2 \to \mathbb{R}$, given by

$$\psi(x, y) = y,$$

with ψ^{-1} given by

$$\psi^{-1}(u) = (u^{2/3}, u).$$

With the atlas consisting of the single chart $\{\psi\colon M_2 \to \mathbb{R}\}$, the space M_2 is also a smooth manifold. Observe that

$$\varphi \circ \psi^{-1}(u) = u^{1/3},$$

a map that is *not* differentiable at $u = 0$. Therefore, the atlas $\{\varphi\colon M_2 \to \mathbb{R}, \psi\colon M_2 \to \mathbb{R}\}$ is not C^1 and thus, with respect to that atlas, M_2 is *not* a C^1-manifold. This example also shows that the atlases $\{\varphi\colon M_2 \to \mathbb{R}\}$ and $\{\psi\colon M_2 \to \mathbb{R}\}$ are inequivalent.

The example of the cuspidal cubic reveals one of the subtleties of the definition of a C^k (or C^∞) manifold: whether a topological space is a C^k-manifold or a smooth manifold depends on the choice of atlas. As a consequence, if a space M happens to be a topological manifold because it has an atlas consisting of a single chart, or more generally if it has an atlas whose transition functions "avoid" singularities, then it is automatically a smooth manifold. In particular, if $f\colon U \to \mathbb{R}^m$ is any *continuous* function from some open subset, U, of $\mathbb{R}^n$, to $\mathbb{R}^m$, then the graph $\Gamma(f) \subseteq \mathbb{R}^{n+m}$ of f given by

$$\Gamma(f) = \{(x, f(x)) \in \mathbb{R}^{n+m} \mid x \in U\}$$

is a smooth manifold with respect to the atlas consisting of the single chart, $\varphi\colon \Gamma(f) \to U$, given by

$$\varphi(x, f(x)) = x,$$

with its inverse, $\varphi^{-1}\colon U \to \Gamma(f)$, given by

$$\varphi^{-1}(x) = (x, f(x)).$$

The notion of a submanifold using the concept of "adapted chart" (see Definition 3.23 in Section 3.4) gives a more satisfactory treatment of C^k (or smooth) submanifolds of $\mathbb{R}^n$.

It should also be noted that determining the number of inequivalent differentiable structures on a topological space is a very difficult problem, even for $\mathbb{R}^n$. In the case of $\mathbb{R}^n$, it turns out that any two smooth differentiable structures are diffeomorphic, except for $n = 4$. For $n = 4$, it took some very hard and deep work to show that there are uncountably many distinct diffeomorphism classes of differentiable structures. The case of the spheres S^n is even more mysterious. It is known that there is a single diffeomorphism class for $n = 1, 2, 3$, but for $n = 4$ the answer is unknown! For $n = 15$, there are $16,256$ distinct classes; for more about these issues, see Conlon [38] (Chapter 3). It is also known that every topological manifold admits a smooth structure for $n = 1, 2, 3$. However, for $n = 4$, there exist nonsmoothable manifolds; see Conlon [38] (Chapter 3).

In some cases, M does not come with a topology in an obvious (or natural) way and a slight variation of Definition 3.2 is more convenient in such a situation:

Definition 3.4. Given a set, M, given some integer $n \geq 1$ and given some k such that k is either an integer $k \geq 1$ or $k = \infty$, a C^k *n-atlas* (or *n-atlas of class* C^k), $\mathcal{A}$, is a family of charts, $\{(U_i, \varphi_i)\}$, such that

(1) Each U_i is a subset of M and $\varphi_i \colon U_i \to \varphi_i(U_i)$ is a bijection onto an open subset, $\varphi_i(U_i) \subseteq \mathbb{R}^n$, for all i;

(2) The U_i cover M, i.e.,
$$M = \bigcup_i U_i;$$

(3) Whenever $U_i \cap U_j \neq \emptyset$, the sets $\varphi_i(U_i \cap U_j)$ and $\varphi_j(U_i \cap U_j)$ are open in $\mathbb{R}^n$ and the transition maps φ_i^j and φ_j^i are C^k-diffeomorphisms.

Then, the notion of a chart being compatible with an atlas and of two atlases being compatible is just as before and we get a new definition of a manifold, analogous to Definition 3.3. But, this time, we give M the topology in which the open sets are arbitrary unions of domains of charts, U_i, more precisely, the U_i's of the maximal atlas defining the differentiable structure on M. It is not difficult to verify that the axioms of a topology are verified and M is indeed a topological space with this topology. It can also be shown that when M is equipped with the above topology, then the maps $\varphi_i \colon U_i \to \varphi_i(U_i)$ are homeomorphisms, so M is a manifold according to Definition 3.3. We also require that under this topology, M is Hausdorff and second-countable. A sufficient condition (in fact, also necessary!) for being second-countable is that some atlas be countable. A sufficient condition of M to be Hausdorff is that for all $p, q \in M$ with $p \neq q$, either $p, q \in U_i$ for some U_i or $p \in U_i$ and $q \in U_j$ for some disjoint U_i, U_j. Thus, we are back to the original notion of a manifold where it is assumed that M is already a topological space.

One can also define the topology on M in terms of any of the atlases, $\mathcal{A}$, defining M (not only the maximal one) by requiring $U \subseteq M$ to be open iff $\varphi_i(U \cap U_i)$ is open in $\mathbb{R}^n$, for every chart, (U_i, φ_i), in the atlas $\mathcal{A}$. Then, one can prove that we obtain the same topology as the topology induced by the maximal atlas. For details, see Berger and Gostiaux [17], Chapter 2.

If the underlying topological space of a manifold is compact, then M has some finite atlas. Also, if $\mathcal{A}$ is some atlas for M and (U, φ) is a chart in $\mathcal{A}$, for any (nonempty) open subset, $V \subseteq U$, we get a chart, $(V, \varphi \upharpoonright V)$, and it is obvious that this chart is compatible with $\mathcal{A}$. Thus, $(V, \varphi \upharpoonright V)$ is also a chart for M. This observation shows that if U is any open subset of a C^k-manifold, M, then U is also a C^k-manifold whose charts are the restrictions of charts on M to U.

Interesting manifolds often occur as the result of a quotient construction. For example, real projective spaces and Grassmannians are obtained this way. In this situation, the natural topology on the quotient object is the quotient topology but, unfortunately, even if the original space is Hausdorff, the quotient topology may not be. Therefore, it is useful to have criteria that insure that a quotient topology is Hausdorff (or second-countable). We will present two criteria. First, let us review the notion of quotient topology. For more details, consult Munkres [116], Massey [104, 105], or Tu [146].

Definition 3.5. Given any topological space, X, and any set, Y, for any surjective function, $f\colon X \to Y$, we define the *quotient topology on Y determined by f* (also called the *identification topology on Y determined by f*), by requiring a subset, V, of Y to be open if $f^{-1}(V)$ is an open set in X. Given an equivalence relation R on a topological space X, if $\pi\colon X \to X/R$ is the projection sending every $x \in X$ to its equivalence class $[x]$ in X/R, the space X/R equipped with the quotient topology determined by π is called the *quotient space of X modulo R*. Thus a set, V, of equivalence classes in X/R is open iff $\pi^{-1}(V)$ is open in X, which is equivalent to the fact that $\bigcup_{[x]\in V}[x]$ is open in X.

It is immediately verified that Definition 3.5 defines topologies and that $f\colon X \to Y$ and $\pi\colon X \to X/R$ are continuous when Y and X/R are given these quotient topologies.

One should be careful that if X and Y are topological spaces and $f\colon X \to Y$ is a continuous surjective map, Y *does not* necessarily have the quotient topology determined by f. Indeed, it may not be true that a subset V of Y is open when $f^{-1}(V)$ is open. However, this will be true in two important cases.

Definition 3.6. A continuous map, $f\colon X \to Y$, is an *open map* (or simply *open*) if $f(U)$ is open in Y whenever U is open in X, and similarly, $f\colon X \to Y$, is a *closed map* (or simply *closed*) if $f(F)$ is closed in Y whenever F is closed in X.

Then, Y has the quotient topology induced by the continuous surjective map f if either f is open or f is closed. Indeed, if f is open, then assuming that $f^{-1}(V)$ is open in X, we have $f(f^{-1}(V)) = V$ open in Y. Now, since $f^{-1}(Y - B) = X - f^{-1}(B)$, for any subset, B, of Y, a subset, V, of Y is open in the quotient topology iff $f^{-1}(Y - V)$ is closed in X. From this, we can deduce that if f is a closed map, then V is open in Y iff $f^{-1}(V)$ is open in X.

If $\cdot\colon G \times X \to X$ is an action of a group G on a topological space X and if for every $g \in G$, the map from X to itself given by $x \mapsto g \cdot x$ is continuous, then it can be show that the projection, $\pi\colon X \to X/G$, is an open map. Furthermore, if G is a finite group, then π is a closed map.

Unfortunately, the Hausdorff separation property is not necessarily preserved under quotient. Nevertheless, it is preserved in some special important cases.

Proposition 3.1. *Let X and Y be topological spaces, let $f\colon X \to Y$ be a continuous surjective map, and assume that X is compact and that Y has the quotient topology determined by f. Then Y is Hausdorff iff f is a closed map.*

Proof. If Y is Hausdorff, because X is compact and f is continuous, since every closed set F in X is compact, $f(F)$ is compact, and since Y is Hausdorff, $f(F)$ is closed, and f is a closed map.

For the converse, we use the fact that in a Hausdorff space, E, if A and B are compact disjoint subsets of E, then there exist two disjoint open sets U and V such that $A \subseteq U$ and $B \subseteq V$.

Since X is Hausdorff, every set, $\{a\}$, consisting of a single element, $a \in X$, is closed, and since f is a closed map, $\{f(a)\}$ is also closed in Y. Since f is surjective, every set, $\{b\}$, consisting of a single element, $b \in Y$, is closed. If $b_1, b_2 \in Y$ and $b_1 \neq b_2$, since $\{b_1\}$ and $\{b_2\}$ are closed in Y and f is continuous, the sets $f^{-1}(b_1)$ and $f^{-1}(b_2)$ are closed in X and thus compact and by the fact stated above, there exists some disjoint open sets U_1 and U_2 such that $f^{-1}(b_1) \subseteq U_1$ and $f^{-1}(b_2) \subseteq U_2$. Since f is closed, the sets $f(X - U_1)$ and $f(X - U_2)$ are closed, and thus the sets

$$\begin{aligned} V_1 &= Y - f(X - U_1) \\ V_2 &= Y - f(X - U_2) \end{aligned}$$

are open, and it is immediately verified that $V_1 \cap V_2 = \emptyset$, $b_1 \in V_1$, and $b_2 \in V_2$. This proves that Y is Hausdorff. □

Under the hypotheses of Proposition 3.1, it is easy to show that Y is Hausdorff iff the set

$$\{(x_1, x_2) \in X \times X \mid f(x_1) = f(x_2)\}$$

is closed in $X \times X$.

Another simple criterion uses continuous open maps. The following proposition is proved in Massey [104] (Appendix A, Proposition 5.3).

Proposition 3.2. *Let $f\colon X \to Y$ be a surjective continuous map between topological spaces. If f is an open map then Y is Hausdorff iff the set*

$$\{(x_1, x_2) \in X \times X \mid f(x_1) = f(x_2)\}$$

is closed in $X \times X$.

Note that the hypothesis of Proposition 3.2 implies that Y has the quotient topology determined by f.

A special case of Proposition 3.2 is discussed in Tu [146] (Section 7.5, Theorem 7.8). Given a topological space, X, and an equivalence relation, R, on X, we say that R is *open* if the projection map, $\pi\colon X \to X/R$, is an open map, where X/R is equipped with the quotient topology. Then, if R is an open equivalence relation on X, the topological space X/R is Hausdorff iff R is closed in $X \times X$.

The following proposition, also from Tu [146] (Section 7.5, Theorem 7.9), yields a sufficient condition for second-countability (the proof is really simple):

Proposition 3.3. *If X is a topological space and R is an open equivalence relation on X, then for any basis, $\{B_\alpha\}$, for the topology of X, the family $\{\pi(B_\alpha)\}$ is a basis for the topology of X/R, where $\pi\colon X \to X/R$ is the projection map. Consequently, if X is second-countable, then so is X/R.*

We are now fully prepared to present a variety of examples.

Example 1. The sphere S^n.

Using the stereographic projections (from the north pole and the south pole), we can define two charts on S^n and show that S^n is a smooth manifold. Let $\sigma_N\colon S^n - \{N\} \to \mathbb{R}^n$ and $\sigma_S\colon S^n - \{S\} \to \mathbb{R}^n$, where $N = (0, \cdots, 0, 1) \in \mathbb{R}^{n+1}$ (the north pole) and $S = (0, \cdots, 0, -1) \in \mathbb{R}^{n+1}$ (the south pole) be the maps called respectively *stereographic projection from the north pole* and *stereographic projection from the south pole* given by

$$\sigma_N(x_1, \ldots, x_{n+1}) = \frac{1}{1 - x_{n+1}}\,(x_1, \ldots, x_n) \quad \text{and} \quad \sigma_S(x_1, \ldots, x_{n+1}) = \frac{1}{1 + x_{n+1}}\,(x_1, \ldots, x_n).$$

The inverse stereographic projections are given by

$$\sigma_N^{-1}(x_1, \ldots, x_n) = \frac{1}{\left(\sum_{i=1}^n x_i^2\right) + 1}\left(2x_1, \ldots, 2x_n, \left(\sum_{i=1}^n x_i^2\right) - 1\right)$$

and

$$\sigma_S^{-1}(x_1, \ldots, x_n) = \frac{1}{\left(\sum_{i=1}^n x_i^2\right) + 1}\left(2x_1, \ldots, 2x_n, -\left(\sum_{i=1}^n x_i^2\right) + 1\right).$$

Thus, if we let $U_N = S^n - \{N\}$ and $U_S = S^n - \{S\}$, we see that U_N and U_S are two open subsets covering S^n, both homeomorphic to $\mathbb{R}^n$. Furthermore, it is easily checked that on the overlap, $U_N \cap U_S = S^n - \{N, S\}$, the transition maps

$$\sigma_S \circ \sigma_N^{-1} = \sigma_N \circ \sigma_S^{-1}$$

are given by

$$(x_1, \ldots, x_n) \mapsto \frac{1}{\sum_{i=1}^n x_i^2}\,(x_1, \ldots, x_n),$$

that is, the inversion of center $O = (0, \ldots, 0)$ and power 1. Clearly, this map is smooth on $\mathbb{R}^n - \{O\}$, so we conclude that (U_N, σ_N) and (U_S, σ_S) form a smooth atlas for S^n.

Example 2. The projective space $\mathbb{RP}^n$.

To define an atlas on $\mathbb{RP}^n$ it is convenient to view $\mathbb{RP}^n$ as the set of equivalence classes of vectors in $\mathbb{R}^{n+1} - \{0\}$ modulo the equivalence relation,

$$u \sim v \quad \text{iff} \quad v = \lambda u, \quad \text{for some} \quad \lambda \neq 0 \in \mathbb{R}.$$

Given any $p = [x_1, \ldots, x_{n+1}] \in \mathbb{RP}^n$, we call $(x_1, \ldots, x_{n+1})$ the *homogeneous coordinates* of p. It is customary to write $(x_1\colon \cdots\colon x_{n+1})$ instead of $[x_1, \ldots, x_{n+1}]$. (Actually, in most books, the indexing starts with 0, i.e., homogeneous coordinates for $\mathbb{RP}^n$ are written as $(x_0\colon \cdots\colon x_n)$.) Now, $\mathbb{RP}^n$ can also be viewed as the quotient of the sphere, S^n, under the equivalence relation where any two antipodal points, x and $-x$, are identified. It is not hard to show that the projection $\pi\colon S^n \to \mathbb{RP}^n$ is both open and closed. Since S^n is compact

and second-countable, we can apply our previous results to prove that under the quotient topology, $\mathbb{RP}^n$ is Hausdorff, second-countable, and compact.

We define charts in the following way. For any i, with $1 \leq i \leq n+1$, let

$$U_i = \{(x_1 \colon \cdots \colon x_{n+1}) \in \mathbb{RP}^n \mid x_i \neq 0\}.$$

Observe that U_i is well defined, because if $(y_1 \colon \cdots \colon y_{n+1}) = (x_1 \colon \cdots \colon x_{n+1})$, then there is some $\lambda \neq 0$ so that $y_j = \lambda x_j$, for $j = 1, \ldots, n+1$. We can define a homeomorphism, φ_i, of U_i onto $\mathbb{R}^n$, as follows:

$$\varphi_i(x_1 \colon \cdots \colon x_{n+1}) = \left(\frac{x_1}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_{n+1}}{x_i}\right),$$

where the ith component is omitted. Again, it is clear that this map is well defined since it only involves ratios. We can also define the maps, ψ_i, from $\mathbb{R}^n$ to $U_i \subseteq \mathbb{RP}^n$, given by

$$\psi_i(x_1, \ldots, x_n) = (x_1 \colon \cdots \colon x_{i-1} \colon 1 \colon x_i \colon \cdots \colon x_n),$$

where the 1 goes in the ith slot, for $i = 1, \ldots, n+1$. One easily checks that φ_i and ψ_i are mutual inverses, so the φ_i are homeomorphisms. On the overlap, $U_i \cap U_j$, (where $i \neq j$), as $x_j \neq 0$, we have

$$(\varphi_j \circ \varphi_i^{-1})(x_1, \ldots, x_n) = \left(\frac{x_1}{x_j}, \ldots, \frac{x_{i-1}}{x_j}, \frac{1}{x_j}, \frac{x_i}{x_j}, \ldots, \frac{x_{j-1}}{x_j}, \frac{x_{j+1}}{x_j}, \ldots, \frac{x_n}{x_j}\right).$$

(We assumed that $i < j$; the case $j < i$ is similar.) This is clearly a smooth function from $\varphi_i(U_i \cap U_j)$ to $\varphi_j(U_i \cap U_j)$. As the U_i cover $\mathbb{RP}^n$, we conclude that the (U_i, φ_i) are $n+1$ charts making a smooth atlas for $\mathbb{RP}^n$. Intuitively, the space $\mathbb{RP}^n$ is obtained by gluing the open subsets U_i on their overlaps. Even for $n = 3$, this is not easy to visualize!

Example 3. The Grassmannian $G(k, n)$.

Recall that $G(k, n)$ is the set of all k-dimensional linear subspaces of $\mathbb{R}^n$, also called k-planes. Every k-plane, W, is the linear span of k linearly independent vectors, $u_1, \ldots, u_k$, in $\mathbb{R}^n$; furthermore, $u_1, \ldots, u_k$ and $v_1, \ldots, v_k$ both span W iff there is an invertible $k \times k$-matrix, $\Lambda = (\lambda_{ij})$, such that

$$v_j = \sum_{i=1}^{k} \lambda_{ij} u_i, \quad 1 \leq j \leq k.$$

Obviously, there is a bijection between the collection of k linearly independent vectors, $u_1, \ldots, u_k$, in $\mathbb{R}^n$ and the collection of $n \times k$ matrices of rank k. Furthermore, two $n \times k$ matrices A and B of rank k represent the same k-plane iff

$$B = A\Lambda, \quad \text{for some invertible } k \times k \text{ matrix, } \Lambda.$$

(Note the analogy with projective spaces where two vectors u, v represent the same point iff $v = \lambda u$ for some invertible $\lambda \in \mathbb{R}$.)

The set of $n \times k$ matrices of rank k is a subset of $\mathbb{R}^{n\times k}$, in fact, an open subset. One can show that the equivalence relation on $n \times k$ matrices of rank k given by

$$B = A\Lambda, \quad \text{for some invertible } k \times k \text{ matrix, } \Lambda,$$

is open and that the graph of this equivalence relation is closed. For some help proving these facts, see Problem 7.2 in Tu [146]. By Proposition 3.2, the Grassmannian $G(k, n)$ is Hausdorff and second-countable.

We can define the domain of charts (according to Definition 3.2) on $G(k, n)$ as follows: For every subset, $S = \{i_1, \ldots, i_k\}$ of $\{1, \ldots, n\}$, let U_S be the subset of $n \times k$ matrices, A, of rank k whose rows of index in $S = \{i_1, \ldots, i_k\}$ form an invertible $k \times k$ matrix denoted A_S. Observe that the $k \times k$ matrix consisting of the rows of the matrix AA_S^{-1} whose index belong to S is the identity matrix, I_k. Therefore, we can define a map, $\varphi_S \colon U_S \to \mathbb{R}^{(n-k)\times k}$, where $\varphi_S(A)$ is equal to the $(n - k) \times k$ matrix obtained by deleting the rows of index in S from AA_S^{-1}.

We need to check that this map is well defined, i.e., that it does not depend on the matrix, A, representing W. Let us do this in the case where $S = \{1, \ldots, k\}$, which is notationally simpler. The general case can be reduced to this one using a suitable permutation.

If $B = A\Lambda$, with Λ invertible, if we write

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix},$$

as $B = A\Lambda$, we get $B_1 = A_1\Lambda$ and $B_2 = A_2\Lambda$, from which we deduce that

$$\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} B_1^{-1} = \begin{pmatrix} I_k \\ B_2B_1^{-1} \end{pmatrix} = \begin{pmatrix} I_k \\ A_2\Lambda\Lambda^{-1}A_1^{-1} \end{pmatrix} = \begin{pmatrix} I_k \\ A_2A_1^{-1} \end{pmatrix} = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} A_1^{-1}.$$

Therefore, our map is indeed well-defined. It is clearly injective and we can define its inverse, ψ_S, as follows: Let π_S be the permutation of $\{1, \ldots, n\}$ swaping $\{1, \ldots, k\}$ and S and leaving every other element fixed (i.e., if $S = \{i_1, \ldots, i_k\}$, then $\pi_S(j) = i_j$ and $\pi_S(i_j) = j$, for $j = 1, \ldots, k$). If P_S is the permutation matrix associated with π_S, for any $(n - k) \times k$ matrix, M, let

$$\psi_S(M) = P_S \begin{pmatrix} I_k \\ M \end{pmatrix}.$$

The effect of ψ_S is to "insert into M" the rows of the identity matrix I_k as the rows of index from S. At this stage, we have charts that are bijections from subsets, U_S, of $G(k, n)$ to open subsets, namely, $\mathbb{R}^{(n-k)\times k}$. Then, the reader can check that the transition map $\varphi_T \circ \varphi_S^{-1}$ from $\varphi_S(U_S \cap U_T)$ to $\varphi_T(U_S \cap U_T)$ is given by

$$M \mapsto (C + DM)(A + BM)^{-1},$$

where

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = P_T P_S,$$

is the matrix of the permutation $\pi_T \circ \pi_S$ (this permutation "shuffles" S and T). This map is smooth, as it is given by determinants, and so, the charts (U_S, φ_S) form a smooth atlas for $G(k,n)$. Finally, one can check that the conditions of Definition 3.2 are satisfied, so the atlas just defined makes $G(k,n)$ into a topological space and a smooth manifold.

The Grassmannian $G(k,n)$ is actually compact. To see this, observe that if W is any k-plane, then using the Gram-Schmidt orthonormalization procedure, every basis $B = (b_1, \ldots, b_k)$ for W yields an orthonormal basis $U = (u_1, \ldots, u_k)$ and there is an invertible matrix, Λ, such that

$$U = B\Lambda,$$

where the the columns of B are the b_js and the columns of U are the u_js. The matrices U have orthonormal columns and are characterized by the equation

$$U^\top U = I_k.$$

Consequently, the space of such matrices is closed an clearly bounded in $\mathbb{R}^{n\times k}$ and thus, compact. The Grassmannian $G(k,n)$ is the quotient of this space under our usual equivalence relation and $G(k,n)$ is the image of a compact set under the projection map, which is clearly continuous, so $G(k,n)$ is compact.

Remark: The reader should have no difficulty proving that the collection of k-planes represented by matrices in U_S is precisely the set of k-planes, W, supplementary to the $(n-k)$-plane spanned by the canonical basis vectors $e_{j_{k+1}}, \ldots, e_{j_n}$ (i.e., $\mathrm{span}(W \cup \{e_{j_{k+1}}, \ldots, e_{j_n}\}) = \mathbb{R}^n$, where $S = \{i_1, \ldots, i_k\}$ and $\{j_{k+1}, \ldots, j_n\} = \{1, \ldots, n\} - S$).

Example 4. Product Manifolds.

Let M_1 and M_2 be two C^k-manifolds of dimension n_1 and n_2, respectively. The topological space, $M_1 \times M_2$, with the product topology (the opens of $M_1 \times M_2$ are arbitrary unions of sets of the form $U \times V$, where U is open in M_1 and V is open in M_2) can be given the structure of a C^k-manifold of dimension $n_1 + n_2$ by defining charts as follows: For any two charts, (U_i, φ_i) on M_1 and (V_j, ψ_j) on M_2, we declare that $(U_i \times V_j, \varphi_i \times \psi_j)$ is a chart on $M_1 \times M_2$, where $\varphi_i \times \psi_j \colon U_i \times V_j \to \mathbb{R}^{n_1+n_2}$ is defined so that

$$\varphi_i \times \psi_j(p,q) = (\varphi_i(p), \psi_j(q)), \quad \text{for all } (p,q) \in U_i \times V_j.$$

Example 5. Configuration Spaces.

Interesting classes of manifolds arise in motion planning for mobile robots. The goal is to place several robots in motion, at the same time, in such a way that collisions are avoided. To model such a system, we assume that the location of each robot is a point in some topological space X; for instance the circle (i.e., S^1), $\mathbb{R}^2$, or $\mathbb{R}^3$.

Definition 3.7. The *configuration space* of n distinct points on X denoted by $\mathrm{Conf}^n(X)$, is the space

$$\mathrm{Conf}^n(X) = \left(\prod_{i=1}^{n} X\right) - \Delta,$$

where

$$\Delta = \Big\{(x_1, \ldots, x_n) \in \prod_{i=1}^{n} X \mid x_i = x_j \text{ for some } i \neq j\Big\}.$$

The *unlabeled configuation space* $\mathrm{UConf}^n(X)$, is defined as the quotient of $\mathrm{Conf}^n(X)$ by the equivalence relation defined such that two n-tuples are equivqlent iff one is a permutation of the other. This space is given the quotient topology.

The set Δ, the pairwise diagonal, represents those configuations of n points in X which experience a collision–this is the set of illegal configuations for the robots.

Configuration spaces of points on a manifold M are all (noncompact) manifolds of dimension $n \cdot \dim(M)$. It turns out that the space $\mathrm{Conf}^n(S^1)$ (n points on a circle) is homeomorphic to $(n-1)!$ distinct copies of $S^1 \times \mathbb{R}^{n-1}$, while $\mathrm{UConf}^n(S^1)$ is a connected space. It can also be shown that $\mathrm{Conf}^n(\mathbb{R}^2)$ (two points in the plane) is homeomorphic to $\mathbb{R}^3 \times S^1$. Configuations spaces and their applications to robotic motion planning have been investigated by Robert Ghrist among others.

We define C^k-maps between manifolds as follows:

Definition 3.8. Given any two C^k-manifolds, M and N, of dimension m and n respectively, a *C^k-map* is a continuous function, $h\colon M \to N$, satisfying the following property: For every $p \in M$, there is some chart, (U, φ), at p and some chart, (V, ψ), at $q = h(p)$, with $f(U) \subseteq V$ and

$$\psi \circ h \circ \varphi^{-1}\colon \varphi(U) \longrightarrow \psi(V)$$

a C^k-function.

It is easily shown that Definition 3.8 does not depend on the choice of charts. In particular, if $N = \mathbb{R}$, we obtain a *C^k-function on M*. One checks immediately that a function, $f\colon M \to \mathbb{R}$, is a C^k-map iff for every $p \in M$, there is some chart, (U, φ), at p so that

$$f \circ \varphi^{-1}\colon \varphi(U) \longrightarrow \mathbb{R}$$

is a C^k-function. If U is an open subset of M, the set of C^k-functions on U is denoted by $\mathcal{C}^k(U)$. In particular, $\mathcal{C}^k(M)$ denotes the set of C^k-functions on the manifold, M. Observe that $\mathcal{C}^k(U)$ is a ring.

On the other hand, if M is an open interval of $\mathbb{R}$, say $M =]a, b[$, then $\gamma\colon]a, b[\to N$ is called a *C^k-curve* in N. One checks immediately that a function, $\gamma\colon]a, b[\to N$, is a C^k-map iff for every $q \in N$, there is some chart, (V, ψ), at q so that

$$\psi \circ \gamma\colon]a, b[\longrightarrow \psi(V)$$

is a C^k-function.

It is clear that the composition of C^k-maps is a C^k-map. A C^k-map, $h\colon M \to N$, between two manifolds is a *C^k-diffeomorphism* iff h has an inverse, $h^{-1}\colon N \to M$ (i.e., $h^{-1} \circ h = \mathrm{id}_M$ and $h \circ h^{-1} = \mathrm{id}_N$), and both h and h^{-1} are C^k-maps (in particular, h and h^{-1} are homeomorphisms). Next, we define tangent vectors.

3.2 Tangent Vectors, Tangent Spaces, Cotangent Spaces

Let M be a C^k manifold of dimension n, with $k \geq 1$. The purpose of this section is to define the tangent space, $T_p(M)$, at a point p of a manifold M (and its dual, the cotangent space, $T_p^*(M)$). We provide three definitions of the notion of a tangent vector to a manifold and prove their equivalence. The first definition uses equivalence classes of curves on a manifold and is the most intuitive. The second definition is based on the view that a tangent vector, v, at p induces a differential operator on functions, f, defined locally near M; namely, the map, $f \mapsto v(f)$, is a linear form satisfying an additional property akin to the rule for taking the derivative of a product. Such linear forms are called *point-derivations*. This second definition is more intrinsic than the first but more abstract. However, for any point p on the manifold M and for any chart whose domain contains p, there is a convenient basis of the tangent space $T_p(M)$. The second definition is also the most convenient one to define vector fields. A few technical complications arise when M is not a smooth manifold (when $k \neq \infty$) but these are easily overcome using "stationary germs." As pointed out by Serre in [137] (Chapter III, Section 8), the relationship between the first definition and the second definition of the tangent space at p is best described by a nondegenerate pairing which shows that $T_p(M)$ is the dual of the space of point-derivations at p that vanish on stationay germs. The third definition makes heavy use of the charts and of the transition functions. It is also quite intuitive and it is easy to see that that it is equivalent to the first definition. The third definition is the most convenient one to define the manifold structure of the tangent bundle, $T(M)$ (see Section 3.3).

The most intuitive method to define tangent vectors is to use curves. Let $p \in M$ be any point on M and let $\gamma\colon]-\epsilon, \epsilon[\to M$ be a C^1-curve passing through p, that is, with $\gamma(0) = p$. Unfortunately, if M is not embedded in any $\mathbb{R}^N$, the derivative $\gamma'(0)$ does not make sense. However, for any chart, (U, φ), at p, the map $\varphi \circ \gamma$ is a C^1-curve in $\mathbb{R}^n$ and the tangent vector $v = (\varphi \circ \gamma)'(0)$ is well defined. The trouble is that different curves may yield the same v!

To remedy this problem, we define an equivalence relation on curves through p as follows:

Definition 3.9. Given a C^k manifold, M, of dimension n, for any $p \in M$, two C^1-curves, $\gamma_1\colon]-\epsilon_1, \epsilon_1[\to M$ and $\gamma_2\colon]-\epsilon_2, \epsilon_2[\to M$, through p (i.e., $\gamma_1(0) = \gamma_2(0) = p$) are *equivalent* iff there is some chart, (U, φ), at p so that

$$(\varphi \circ \gamma_1)'(0) = (\varphi \circ \gamma_2)'(0).$$

Now, the problem is that this definition seems to depend on the choice of the chart. Fortunately, this is not the case. For, if (V, ψ) is another chart at p, as p belongs both to U and V, we have $U \cap V \neq 0$, so the transition function $\eta = \psi \circ \varphi^{-1}$ is C^k and, by the chain

rule, we have

$$\begin{aligned}(\psi \circ \gamma_1)'(0) &= (\eta \circ \varphi \circ \gamma_1)'(0)\\ &= \eta'(\varphi(p))((\varphi \circ \gamma_1)'(0))\\ &= \eta'(\varphi(p))((\varphi \circ \gamma_2)'(0))\\ &= (\eta \circ \varphi \circ \gamma_2)'(0)\\ &= (\psi \circ \gamma_2)'(0).\end{aligned}$$

This leads us to the first definition of a tangent vector.

Definition 3.10. (Tangent Vectors, Version 1) Given any C^k-manifold, M, of dimension n, with $k \geq 1$, for any $p \in M$, a *tangent vector to M at p* is any equivalence class of C^1-curves through p on M, modulo the equivalence relation defined in Definition 3.9. The set of all tangent vectors at p is denoted by $T_p(M)$ (or T_pM).

It is obvious that $T_p(M)$ is a vector space. If $u, v \in T_p(M)$ are defined by the curves γ_1 and γ_2, then $u + v$ is defined by the curve $\gamma_1 + \gamma_2$ (we may assume by reparametrization that γ_1 and γ_2 have the same domain.) Similarly, if $u \in T_p(M)$ is defined by a curve γ and $\lambda \in \mathbb{R}$, then λu is defined by the curve $\lambda\gamma$. The reader should check that these definitions do not depend on the choice of the curve in its equivalence class.

Observe that the map that sends a curve, $\gamma\colon]-\epsilon,\epsilon[\to M$, through p (with $\gamma(0) = p$) to its tangent vector, $(\varphi \circ \gamma)'(0) \in \mathbb{R}^n$ (for any chart (U, φ), at p), induces a map, $\overline{\varphi}\colon T_p(M) \to \mathbb{R}^n$. It is easy to check that $\overline{\varphi}$ is a linear bijection (by definition of the equivalence relation on curves through p). This shows that $T_p(M)$ is a vector space of dimension n = dimension of M.

One should observe that unless $M = \mathbb{R}^n$, in which case, for any $p, q \in \mathbb{R}^n$, the tangent space $T_q(M)$ is naturally isomorphic to the tangent space $T_p(M)$ by the translation $q - p$, for an arbitrary manifold, there is no relationship between $T_p(M)$ and $T_q(M)$ when $p \neq q$.

One of the defects of the above definition of a tangent vector is that it has no clear relation to the C^k-differential structure of M. In particular, the definition does not seem to have anything to do with the functions defined locally at p. There is another way to define tangent vectors that reveals this connection more clearly. Moreover, such a definition is more intrinsic, i.e., does not refer explicitly to charts. Our presentation of this second approach is heavily inspired by Schwartz [136] (Chapter 3, Section 9) but also by Warner [148] and Serre [137] (Chapter III, Sections 7 and 8.

As a first step, consider the following: Let (U, φ) be a chart at $p \in M$ (where M is a C^k-manifold of dimension n, with $k \geq 1$) and let $x_i = pr_i \circ \varphi$, the ith local coordinate $(1 \leq i \leq n)$. For any function, f, defined on $U \ni p$, set

$$\left(\frac{\partial}{\partial x_i}\right)_p f = \left.\frac{\partial(f \circ \varphi^{-1})}{\partial X_i}\right|_{\varphi(p)}, \qquad 1 \leq i \leq n.$$

(Here, $(\partial g/\partial X_i)|_y$ denotes the partial derivative of a function $g\colon \mathbb{R}^n \to \mathbb{R}$ with respect to the ith coordinate, evaluated at y.)

We would expect that the function that maps f to the above value is a linear map on the set of functions defined locally at p, but there is technical difficulty: The set of functions defined locally at p is **not** a vector space! To see this, observe that if f is defined on an open $U \ni p$ and g is defined on a different open $V \ni p$, then we do not know how to define $f + g$. The problem is that we need to identify functions that agree on a smaller open subset. This leads to the notion of *germs*.

Definition 3.11. Given any C^k-manifold, M, of dimension n, with $k \geq 1$, for any $p \in M$, a *locally defined function at* p is a pair, (U, f), where U is an open subset of M containing p and f is a function defined on U. Two locally defined functions, (U, f) and (V, g), at p are *equivalent* iff there is some open subset, $W \subseteq U \cap V$, containing p so that

$$f \restriction W = g \restriction W.$$

The equivalence class of a locally defined function at p, denoted $[f]$ or $\mathbf{f}$, is called a *germ at* p.

One should check that the relation of Definition 3.11 is indeed an equivalence relation. Of course, the value at p of all the functions, f, in any germ, $\mathbf{f}$, is $f(p)$. Thus, we set $\mathbf{f}(p) = f(p)$, for any $f \in \mathbf{f}$.

For example, for every $a \in (-1, 1)$, the locally defined functions $(\mathbb{R} - \{1\}, 1/(1-x))$ and $((-1, 1), \sum_{n=0}^{\infty} x^n)$ at a are equivalent.

One should also check that we can define addition of germs, multiplication of a germ by a scalar and multiplication of germs, in the obvious way: If $\mathbf{f} = [f]$ and $\mathbf{g} = [g]$ are two germs at p, and $\lambda \in \mathbb{R}$, then

$$\begin{aligned} [f] + [g] &= [f + g] \\ \lambda[f] &= [\lambda f] \\ [f][g] &= [fg]. \end{aligned}$$

However, we have to check that these definitions make sense, that is, that they don't depend on the choice of representatives chosen in the equivalence classes $[f]$ and $[g]$. Let us give the details of this verification for the sum of two germs, $[f]$ and $[g]$. For any two locally defined functions, (f, U) and (g, V), at p, let $f + g$ be the locally defined function at p with domain $U \cap V$ and such that $(f + g)(x) = f(x) + g(x)$ for all $x \in U \cap V$. We need to check that for any locally defined functions (U_1, f_1), (U_2, f_2), (V_1, g_1), and (V_2, g_2), at p, if (U_1, f_1) and (U_2, f_2) are equivalent and if (V_1, g_1) and (V_2, g_2) are equivalent, then $(U_1 \cap V_1, f_1 + g_1)$ and $(U_2 \cap V_2, f_2 + g_2)$ are equivalent. However, as (U_1, f_1) and (U_2, f_2) are equivalent, there is some $W_1 \subseteq U_1 \cap U_2$ so that $f_1 \restriction W_1 = f_2 \restriction W_1$ and as (V_1, g_1) and (V_2, g_2) are equivalent, there is some $W_2 \subseteq V_1 \cap V_2$ so that $g_1 \restriction W_2 = g_2 \restriction W_2$. Then, observe

that $(f_1 + g_1) \restriction (W_1 \cap W_2) = (f_2 + g_2) \restriction (W_1 \cap W_2)$, which means that $[f_1 + g_1] = [f_2 + g_2]$. Therefore, $[f + g]$ does not depend on the representatives chosen in the equivalence classes $[f]$ and $[g]$ and it makes sense to set

$$[f] + [g] = [f + g].$$

We can proceed in a similar fashion to define $\lambda[f]$ and $[f][g]$. Therefore, the germs at p form a ring.

The ring of germs of C^k-functions at p is denoted $\mathcal{O}^{(k)}_{M,p}$. When $k = \infty$, we usually drop the superscript ∞.

Remark: Most readers will most likely be puzzled by the notation $\mathcal{O}^{(k)}_{M,p}$. In fact, it is standard in algebraic geometry, but it is not as commonly used in differential geometry. For any open subset, U, of a manifold, M, the ring, $\mathcal{C}^k(U)$, of C^k-functions on U is also denoted $\mathcal{O}^{(k)}_M(U)$ (certainly by people with an algebraic geometry bent!). Then, it turns out that the map $U \mapsto \mathcal{O}^{(k)}_M(U)$ is a *sheaf*, denoted $\mathcal{O}^{(k)}_M$, and the ring $\mathcal{O}^{(k)}_{M,p}$ is the *stalk* of the sheaf $\mathcal{O}^{(k)}_M$ at p. Such rings are called *local rings*. Roughly speaking, all the "local" information about M at p is contained in the local ring $\mathcal{O}^{(k)}_{M,p}$. (This is to be taken with a grain of salt. In the C^k-case where $k < \infty$, we also need the "stationary germs," as we will see shortly.)

Now that we have a rigorous way of dealing with functions locally defined at p, observe that the map

$$v_i \colon f \mapsto \left(\frac{\partial}{\partial x_i}\right)_p f$$

yields the same value for all functions f in a germ $\mathbf{f}$ at p. Furthermore, the above map is linear on $\mathcal{O}^{(k)}_{M,p}$. More is true:

(1) For any two functions f, g locally defined at p, we have

$$\left(\frac{\partial}{\partial x_i}\right)_p (fg) = f(p)\left(\frac{\partial}{\partial x_i}\right)_p g + g(p)\left(\frac{\partial}{\partial x_i}\right)_p f.$$

(2) If $(f \circ \varphi^{-1})'(\varphi(p)) = 0$, then

$$\left(\frac{\partial}{\partial x_i}\right)_p f = 0.$$

The first property says that v_i is a *point derivation*. As to the second property, when $(f \circ \varphi^{-1})'(\varphi(p)) = 0$, we say that f *is stationary at* p.

It is easy to check (using the chain rule) that being stationary at p does not depend on the chart, (U, φ), at p or on the function chosen in a germ, $\mathbf{f}$. Therefore, the notion of a stationary germ makes sense.

Definition 3.12. We say that a germ $\mathbf{f}$ at $p \in M$ is a *stationary germ* iff $(f \circ \varphi^{-1})'(\varphi(p)) = 0$ for some chart, (U, φ), at p and some function, f, in the germ, $\mathbf{f}$. The C^k-stationary germs form a subring of $\mathcal{O}^{(k)}_{M,p}$ (but not an ideal!) denoted $\mathcal{S}^{(k)}_{M,p}$.

Remarkably, it turns out that the dual of the vector space, $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$, is isomorphic to the tangent space, $T_p(M)$.

Let us refresh the reader's memory and review quotient vector spaces. If E is a vector space, the set of all linear forms $f\colon E \to \mathbb{R}$ on E is a vector space called the *dual* of E and denoted by E^*. If $H \subseteq E$ is any subspace of E, we define the equivalence relation $\sim$ so that for all $u, v \in E$,

$$u \sim v \quad \text{iff} \quad u - v \in H.$$

Every equivalence class, $[u]$, is equal to the subset $u + H = \{u + h \mid h \in H\}$, called a *coset*, and the set of equivalence classes, E/H, modulo $\sim$ is a vector space under the operations

$$[u] + [v] = [u + v]$$
$$\lambda[u] = [\lambda u].$$

The space E/H is called the *quotient of E by H* or for short, a *quotient space*.

Denote by $\mathcal{L}(E/H)$ the set of linear forms $f\colon E \to \mathbb{R}$ that vanish on H (this means that for every $f \in \mathcal{L}(E/H)$, we have $f(h) = 0$ for all $h \in H$). We claim that there is an isomorphism

$$\mathcal{L}(E/H) \cong (E/H)^*$$

between $\mathcal{L}(E/H)$ and the dual of the quotient space E/H.

To see this, define the map, $f \mapsto \widehat{f}$ from $\mathcal{L}(E/H)$ to $(E/H)^*$ as follows: For any $f \in \mathcal{L}(E/H)$,

$$\widehat{f}([u]) = f(u), \quad [u] \in E/H.$$

This function is well-defined because it does not depend on the representative, u, chosen in the equivalence class $[u]$. Indeed, if $v \sim u$, then $v = u + h$ some $h \in H$ and so

$$f(v) = f(u + h) = f(u) + f(h) = f(u),$$

since $f(h) = 0$ for all $h \in H$. The formula $\widehat{f}([u]) = f(u)$ makes it obvious that $\widehat{f}$ is linear since f is linear. The mapping $f \mapsto \widehat{f}$ is injective. This is beause if $\widehat{f_1} = \widehat{f_2}$, then

$$\widehat{f_1}([u]) = \widehat{f_2}([u])$$

for all $u \in E$, and because $\widehat{f_1}([u]) = f_1(u)$ and $\widehat{f_2}([u]) = f_2(u)$, we get $f_1 u) = f_2(u)$ for all $u \in E$, that is, $f_1 = f_2$. The mapping $f \mapsto \widehat{f}$ is surjective because given any linear form $\varphi \in (E/H)^*$, if we define f by

$$f(u) = \varphi([u])$$

for all $u \in E$, then f is linear, vanishes on H and clearly, $\widehat{f} = \varphi$. Therefore, we have the isomorphism,

$$\mathcal{L}(E/H) \cong (E/H)^*,$$

as claimed.

Let us return to the space of linear forms on $\mathcal{O}^{(k)}_{M,p}$ that vanish on $\mathcal{S}^{(k)}_{M,p}$ (which is isomorphic to $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$). First, we prove that this space has $\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p$ as a basis.

Proposition 3.4. *Given any C^k-manifold, M, of dimension n, with $k \geq 1$, for any $p \in M$ and any chart (U, φ) at p, the n functions, $\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p$, defined on $\mathcal{O}^{(k)}_{M,p}$ by*

$$\left(\frac{\partial}{\partial x_i}\right)_p f = \left.\frac{\partial(f \circ \varphi^{-1})}{\partial X_i}\right|_{\varphi(p)}, \qquad 1 \leq i \leq n$$

are linear forms that vanish on $\mathcal{S}^{(k)}_{M,p}$. Every linear form, L, on $\mathcal{O}^{(k)}_{M,p}$ that vanishes on $\mathcal{S}^{(k)}_{M,p}$ can be expressed in a unique way as

$$L = \sum_{i=1}^{n} \lambda_i \left(\frac{\partial}{\partial x_i}\right)_p,$$

where $\lambda_i \in \mathbb{R}$. Therefore, the

$$\left(\frac{\partial}{\partial x_i}\right)_p, \qquad i = 1, \ldots, n$$

form a basis of the vector space of linear forms on $\mathcal{O}^{(k)}_{M,p}$ that vanish on $\mathcal{S}^{(k)}_{M,p}$.

Proof. The first part of the proposition is trivial, by definition of $(f \circ \varphi^{-1})'(\varphi(p))$ and of $\left(\frac{\partial}{\partial x_i}\right)_p f$.

Next, assume that L is a linear form on $\mathcal{O}^{(k)}_{M,p}$ that vanishes on $\mathcal{S}^{(k)}_{M,p}$. Consider the locally defined function at p given by

$$g(q) = f(q) - \sum_{i=1}^{n} (pr_i \circ \varphi)(q) \left(\frac{\partial}{\partial x_i}\right)_p f.$$

Observe that the germ of g is stationary at p, since

$$\begin{aligned} g(q) = (g \circ \varphi^{-1})(\varphi(q)) &= (f \circ \varphi^{-1})(\varphi(q)) - \sum_{i=1}^{n} (pr_i \circ \varphi)(q) \left(\frac{\partial}{\partial x_i}\right)_p f \\ &= (f \circ \varphi^{-1})(X_1(q) \ldots, X_n(q)) - \sum_{i=1}^{n} X_i(q) \left(\frac{\partial}{\partial x_i}\right)_p f, \end{aligned}$$

with $X_i(q) = (pr_i \circ \varphi)(q)$. It follows that

$$\left.\frac{\partial(g \circ \varphi^{-1})}{\partial X_i}\right|_{\varphi(p)} = \left.\frac{\partial(f \circ \varphi^{-1})}{\partial X_i}\right|_{\varphi(p)} - \left(\frac{\partial}{\partial x_i}\right)_p f = 0.$$

But then, as L vanishes on stationary germs, we get

$$L(f) = \sum_{i=1}^{n} L(pr_i \circ \varphi) \left(\frac{\partial}{\partial x_i}\right)_p f,$$

as desired. We still have to prove linear independence. If

$$\sum_{i=1}^{n} \lambda_i \left(\frac{\partial}{\partial x_i}\right)_p = 0,$$

then, if we apply this relation to the functions $x_i = pr_i \circ \varphi$, as

$$\left(\frac{\partial}{\partial x_i}\right)_p x_j = \delta_{ij},$$

we get $\lambda_i = 0$, for $i = 1, \ldots, n$. □

As the subspace of linear forms on $\mathcal{O}^{(k)}_{M,p}$ that vanish on $\mathcal{S}^{(k)}_{M,p}$ is isomorphic to the dual, $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$, of the space $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$, we see that the

$$\left(\frac{\partial}{\partial x_i}\right)_p, \qquad i = 1, \ldots, n$$

also form a basis of $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$.

To define our second version of tangent vectors, we need to define point-derivations.

Definition 3.13. Given any C^k-manifold, M, of dimension n, with $k \geq 1$, for any $p \in M$, a *derivation at p in M* or *point-derivation on $\mathcal{O}^{(k)}_{M,p}$* is a linear form, v, on $\mathcal{O}^{(k)}_{M,p}$, such that

$$v(\mathbf{fg}) = v(\mathbf{f})\mathbf{g}(p) + \mathbf{f}(p)v(\mathbf{g}),$$

for all germs $\mathbf{f}, \mathbf{g} \in \mathcal{O}^{(k)}_{M,p}$. The above is called the *Leibniz property*.

As expected, point-derivations vanish on constant functions.

Proposition 3.5. *Every point-derivation, v, on $\mathcal{O}^{(k)}_{M,p}$, vanishes on germs of constant functions.*

Proof. If $\mathbf{g}$ is a germ of constant functions at p, then there is some $\lambda \in \mathbb{R}$ so that $g = \lambda$ (a constant function with value λ) for all $g \in \mathbf{g}$. Since v is linear,

$$v(\mathbf{g}) = v(\lambda \mathbf{1}) = \lambda v(\mathbf{1}),$$

where $\mathbf{1}$ is the germ of constant functions with value 1, so we just have to show that $v(\mathbf{1}) = 0$. However, because $\mathbf{1} = \mathbf{1} \cdot \mathbf{1}$ and v is a point-derivation, we get

$$\begin{aligned} v(\mathbf{1}) &= v(\mathbf{1} \cdot \mathbf{1}) \\ &= v(\mathbf{1})\mathbf{1}(p) + \mathbf{1}(p)v(\mathbf{1}) \\ &= v(\mathbf{1})1 + 1v(\mathbf{1}) = 2v(\mathbf{1}) \end{aligned}$$

from which we conclude that $v(\mathbf{1}) = 0$, as claimed. □

Recall that we observed earlier that the $\left(\frac{\partial}{\partial x_i}\right)_p$ are point-derivations at p. Therefore, we have

Proposition 3.6. *Given any C^k-manifold, M, of dimension n, with $k \geq 1$, for any $p \in M$, the linear forms on $\mathcal{O}_{M,p}^{(k)}$ that vanish on $\mathcal{S}_{M,p}^{(k)}$ are exactly the point-derivations on $\mathcal{O}_{M,p}^{(k)}$ that vanish on $\mathcal{S}_{M,p}^{(k)}$.*

Proof. By Proposition 3.4, the

$$\left(\frac{\partial}{\partial x_i}\right)_p, \qquad i = 1, \ldots, n$$

form a basis of the linear forms on $\mathcal{O}_{M,p}^{(k)}$ that vanish on $\mathcal{S}_{M,p}^{(k)}$. Since each $\left(\frac{\partial}{\partial x_i}\right)_p$ is a also a point-derivation at p, the result follows. □

Remarks:

(1) If we let $\mathcal{D}_p^{(k)}(M)$ denote the set of point-derivations on $\mathcal{O}_{M,p}^{(k)}$, then Proposition 3.6 says that any linear form on $\mathcal{O}_{M,p}^{(k)}$ that vanishes on $\mathcal{S}_{M,p}^{(k)}$ belongs to $\mathcal{D}_p^{(k)}(M)$, so we have the inclusion

$$(\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)})^* \subseteq \mathcal{D}_p^{(k)}(M).$$

However, in general, when $k \neq \infty$, a point-derivation on $\mathcal{O}_{M,p}^{(k)}$ does *not* necessarily vanish on $\mathcal{S}_{M,p}^{(k)}$. We will see in Proposition 3.11 that this is true for $k = \infty$.

(2) In the case of smooth manifolds ($k = \infty$) some authors, including Morita [115] (Chapter 1, Definition 1.32) and O'Neil [120] (Chapter 1, Definition 9), define point-derivations as linear derivations with domain $\mathcal{C}^\infty(M)$, the set of all smooth funtions on the entire manifold, M. This definition is simpler in the sense that it does not require the definition of the notion of germ but it is not local, because it is not obvious that if v is a point-derivation at p, then $v(f) = v(g)$ whenever $f, g \in \mathcal{C}^\infty(M)$ agree locally at p. In fact, if two smooth locally defined functions agree near p it may not be possible to extend both of them to the whole of M. However, it can proved that this property is local because on smooth manifolds, "bump functions" exist (see Section 3.6, Proposition 3.30). Unfortunately, this argument breaks down for C^k-manifolds with $k < \infty$ and in this case the ring of germs at p can't be avoided.

Here is now our second definition of a tangent vector.

Definition 3.14. (Tangent Vectors, Version 2) Given any C^k-manifold, M, of dimension n, with $k \geq 1$, for any $p \in M$, a *tangent vector to M at p* is any point-derivation on $\mathcal{O}^{(k)}_{M,p}$ that vanishes on $\mathcal{S}^{(k)}_{M,p}$, the subspace of stationary germs.

Let us consider the simple case where $M = \mathbb{R}$. In this case, for every $x \in \mathbb{R}$, the tangent space, $T_x(\mathbb{R})$, is a one-dimensional vector space isomorphic to $\mathbb{R}$ and $\left(\frac{\partial}{\partial t}\right)_x = \frac{d}{dt}\Big|_x$ is a basis vector of $T_x(\mathbb{R})$. For every C^k-function, f, locally defined at x, we have

$$\left(\frac{\partial}{\partial t}\right)_x f = \left.\frac{df}{dt}\right|_x = f'(x).$$

Thus, $\left(\frac{\partial}{\partial t}\right)_x$ is: compute the derivative of a function at x.

We now prove the equivalence of the two definitions of a tangent vector.

Proposition 3.7. *Let M be any C^k-manifold of dimension n, with $k \geq 1$. For any $p \in M$, let u be any tangent vector (version 1) given by some equivalence class of C^1-curves, $\gamma\colon]-\epsilon, +\epsilon[\to M$, through p (i.e., $p = \gamma(0)$). Then, the map L_u defined on $\mathcal{O}^{(k)}_{M,p}$ by*

$$L_u(\mathbf{f}) = (f \circ \gamma)'(0)$$

is a point-derivation that vanishes on $\mathcal{S}^{(k)}_{M,p}$. Furthermore, the map $u \mapsto L_u$ defined above is an isomorphism between $T_p(M)$ and $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$, the space of linear forms on $\mathcal{O}^{(k)}_{M,p}$ that vanish on $\mathcal{S}^{(k)}_{M,p}$.

Proof. (After L. Schwartz) Clearly, $L_u(\mathbf{f})$ does not depend on the representative, f, chosen in the germ, $\mathbf{f}$. If γ and σ are equivalent curves defining u, then $(\varphi \circ \sigma)'(0) = (\varphi \circ \gamma)'(0)$, so we get

$$(f \circ \sigma)'(0) = (f \circ \varphi^{-1})'(\varphi(p))((\varphi \circ \sigma)'(0)) = (f \circ \varphi^{-1})'(\varphi(p))((\varphi \circ \gamma)'(0)) = (f \circ \gamma)'(0),$$

which shows that $L_u(\mathbf{f})$ does not depend on the curve, γ, defining u. If $\mathbf{f}$ is a stationary germ, then pick any chart, (U, φ), at p and let $\psi = \varphi \circ \gamma$. We have

$$L_u(\mathbf{f}) = (f \circ \gamma)'(0) = ((f \circ \varphi^{-1}) \circ (\varphi \circ \gamma))'(0) = (f \circ \varphi^{-1})'(\varphi(p))(\psi'(0)) = 0,$$

since $(f \circ \varphi^{-1})'(\varphi(p)) = 0$, as $\mathbf{f}$ is a stationary germ. The definition of L_u makes it clear that L_u is a point-derivation at p. If $u \neq v$ are two distinct tangent vectors, then there exist some curves γ and σ through p so that

$$(\varphi \circ \gamma)'(0) \neq (\varphi \circ \sigma)'(0).$$

Thus, there is some i, with $1 \leq i \leq n$, so that if we let $f = pr_i \circ \varphi$, then

$$(f \circ \gamma)'(0) \neq (f \circ \sigma)'(0),$$

and so, $L_u \neq L_v$. This proves that the map $u \mapsto L_u$ is injective.

For surjectivity, recall that every linear map, L, on $\mathcal{O}^{(k)}_{M,p}$ that vanishes on $\mathcal{S}^{(k)}_{M,p}$ can be uniquely expressed as

$$L = \sum_{i=1}^{n} \lambda_i \left(\frac{\partial}{\partial x_i}\right)_p.$$

Define the curve, γ, on M through p by

$$\gamma(t) = \varphi^{-1}(\varphi(p) + t(\lambda_1, \ldots, \lambda_n)),$$

for t in a small open interval containing 0. Then, we have

$$f(\gamma(t)) = (f \circ \varphi^{-1})(\varphi(p) + t(\lambda_1, \ldots, \lambda_n)),$$

and we get

$$(f \circ \gamma)'(0) = (f \circ \varphi^{-1})'(\varphi(p))(\lambda_1, \ldots, \lambda_n) = \sum_{i=1}^{n} \lambda_i \left.\frac{\partial(f \circ \varphi^{-1})}{\partial X_i}\right|_{\varphi(p)} = L(\mathbf{f}).$$

This proves that $T_p(M)$ and $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ are isomorphic. □

There is a conceptually clearer way to define a canonical isomorphism between $T_p(M)$ and the dual of $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$ in terms of a nondegenerate pairing between $T_p(M)$ and $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$ (for the notion of a pairing, see Definition 22.1 and Proposition 22.1). This pairing is described by Serre in [137] (Chapter III, Section 8) for analytic manifolds and can be adapted to our situation.

Define the map, $\omega\colon T_p(M) \times \mathcal{O}^{(k)}_{M,p} \to \mathbb{R}$, so that

$$\omega([\gamma], \mathbf{f}) = (f \circ \gamma)'(0),$$

for all $[\gamma] \in T_p(M)$ and all $\mathbf{f} \in \mathcal{O}^{(k)}_{M,p}$ (with $f \in \mathbf{f}$). It is easy to check that the above expression does not depend on the representatives chosen in the equivalences classes $[\gamma]$ and $\mathbf{f}$ and that ω is bilinear. However, as defined, ω is degenerate because $\omega([\gamma], \mathbf{f}) = 0$ if $\mathbf{f}$ is a stationary germ. Thus, we are led to consider the pairing with domain $T_p(M) \times (\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})$ given by

$$\omega([\gamma], [\mathbf{f}]) = (f \circ \gamma)'(0),$$

where $[\gamma] \in T_p(M)$ and $[\mathbf{f}] \in \mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$, which we also denote $\omega \colon T_p(M) \times (\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}) \to \mathbb{R}$. Then, the following result holds:

Proposition 3.8. *The map $\omega \colon T_p(M) \times (\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}) \to \mathbb{R}$ defined so that*

$$\omega([\gamma], [\mathbf{f}]) = (f \circ \gamma)'(0),$$

for all $[\gamma] \in T_p(M)$ and all $[\mathbf{f}] \in \mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$, is a nondegenerate pairing (with $f \in \mathbf{f}$). Consequently, there is a canonical isomorphism between $T_p(M)$ and $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ and a canonical isomorphism between $T^*_p(M)$ and $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$.

Proof. This is basically a replay of the proof of Proposition 3.7. First, assume that given some $[\gamma] \in T_p(M)$, we have $\omega([\gamma], [\mathbf{f}]) = 0$ for all $[\mathbf{f}] \in \mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$. Pick a chart, (U, φ), with $p \in U$ and let $x_i = pr_i \circ \varphi$. Then, the $\mathbf{x_i}$'s are not stationary germs, since $x_i \circ \varphi^{-1} = pr_i \circ \varphi \circ \varphi^{-1} = pr_i$ and $(pr_i)'(0) = pr_i$ (because pr_i is a linear form). By hypothesis, $\omega([\gamma], [\mathbf{x_i}]) = 0$ for $i = 1, \ldots, n$, which means that

$$(x_i \circ \gamma)'(0) = (pr_i \circ \varphi \circ \gamma)'(0) = 0$$

for $i = 1, \ldots, n$, namely, $pr_i((\varphi \circ \gamma)'(0)) = 0$ for $i = 1, \ldots, n$; that is,

$$(\varphi \circ \gamma)'(0) = 0_n,$$

proving that $[\gamma] = 0$.

Next, assume that given some $[\mathbf{f}] \in \mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$, we have $\omega([\gamma], [\mathbf{f}]) = 0$ for all $[\gamma] \in T_p(M)$. Again, pick a chart, (U, φ). For every $z \in \mathbb{R}^n$, we have the curve γ_z given by

$$\gamma_z(t) = \varphi^{-1}(\varphi(p) + tz)$$

for all t in a small open interval containing 0. Then, by hypothesis,

$$\omega([\gamma_z], [\mathbf{f}]) = (f \circ \gamma_z)'(0) = (f \circ \varphi^{-1})'(\varphi(p))(z) = 0$$

for all $z \in \mathbb{R}^n$, which means that

$$(f \circ \varphi^{-1})'(\varphi(p)) = 0.$$

But then, $\mathbf{f}$ is a stationary germ and so, $[\mathbf{f}] = 0$. Therefore, we proved that ω is a nondegenerate pairing. Since $T_p(M)$ and $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$ have finite dimension, n, it follows by Proposition 22.1 that there is are canonical isomorphisms between $T_p(M)$ and $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ and between $T^*_p(M)$ and $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$. □

In view of Proposition 3.8, we can identify $T_p(M)$ with $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ and $T^*_p(M)$ with $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$.

Remark: Also recall that if E is a finite dimensional space, the map $i_E\colon E \to E^{**}$ defined so that, for any $v \in E$,

$$v \mapsto \widetilde{v}, \quad \text{where} \quad \widetilde{v}(f) = f(v), \quad \text{for all } f \in E^*$$

is a linear isomorphism.

Observe that we can view $\omega(u, \mathbf{f}) = \omega([\gamma], [\mathbf{f}])$ as the result of computing the directional derivative of the locally defined function $f \in \mathbf{f}$ in the direction u (given by a curve γ). Proposition 3.8 also suggests the following definition:

Definition 3.15. Given any C^k-manifold, M, of dimension n, with $k \geq 1$, for any $p \in M$, the *tangent space at* p, denoted $T_p(M)$ is the space of point-derivations on $\mathcal{O}^{(k)}_{M,p}$ that vanish on $\mathcal{S}^{(k)}_{M,p}$. Thus, $T_p(M)$ can be identified with $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$. The space $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$ is called the *cotangent space at* p; it is isomorphic to the dual, $T^*_p(M)$, of $T_p(M)$. (For simplicity of notation we also denote $T_p(M)$ by T_pM and $T^*_p(M)$ by T^*_pM.)

Even though this is just a restatement of Proposition 3.4, we state the following proposition because of its practical usefulness:

Proposition 3.9. *Given any C^k-manifold, M, of dimension n, with $k \geq 1$, for any $p \in M$ and any chart (U, φ) at p, the n tangent vectors,*

$$\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p,$$

form a basis of T_pM.

Observe that if $x_i = pr_i \circ \varphi$, as

$$\left(\frac{\partial}{\partial x_i}\right)_p x_j = \delta_{i,j},$$

the images of $x_1, \ldots, x_n$ in $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$ form the dual basis of the basis $\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p$ of $T_p(M)$.

Given any C^k-function, f, on M, we denote the image of f in $T^*_p(M) = \mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$ by df_p. This is the *differential of f at p*. Using the isomorphism between $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$ and $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^{**}$ described above, df_p corresponds to the linear map in $T^*_p(M)$ defined by

$$df_p(v) = v(\mathbf{f}),$$

for all $v \in T_p(M)$. With this notation, we see that $(dx_1)_p, \ldots, (dx_n)_p$ is a basis of $T_p^*(M)$, and this basis is dual to the basis $\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p$ of $T_p(M)$. For simplicity of notation, we often omit the subscript p unless confusion arises.

Remark: Strictly speaking, a tangent vector, $v \in T_p(M)$, is defined on the space of germs, $\mathcal{O}_{M,p}^{(k)}$, at p. However, it is often convenient to define v on C^k-functions, $f \in \mathcal{C}^k(U)$, where U is some open subset containing p. This is easy: Set

$$v(f) = v(\mathbf{f}).$$

Given any chart, (U, φ), at p, since v can be written in a unique way as

$$v = \sum_{i=1}^{n} \lambda_i \left(\frac{\partial}{\partial x_i}\right)_p,$$

we get

$$v(f) = \sum_{i=1}^{n} \lambda_i \left(\frac{\partial}{\partial x_i}\right)_p f.$$

This shows that $v(f)$ is the *directional derivative of f in the direction v*. The directional derivative, $v(f)$, is also denoted $v[f]$.

When M is a smooth manifold, things get a little simpler. Indeed, it turns out that in this case, every point-derivation vanishes on stationary germs. To prove this, we recall the following result from calculus (see Warner [148]):

Proposition 3.10. *If $g\colon \mathbb{R}^n \to \mathbb{R}$ is a C^k-function ($k \geq 2$) on a convex open, U, about $p \in \mathbb{R}^n$, then for every $q \in U$, we have*

$$g(q) = g(p) + \sum_{i=1}^{n} \left.\frac{\partial g}{\partial X_i}\right|_p (q_i - p_i) + \sum_{i,j=1}^{n} (q_i - p_i)(q_j - p_j) \int_0^1 (1-t) \left.\frac{\partial^2 g}{\partial X_i \partial X_j}\right|_{(1-t)p+tq} dt.$$

In particular, if $g \in C^\infty(U)$, then the integral as a function of q is C^∞.

Proposition 3.11. *Let M be any C^∞-manifold of dimension n. For any $p \in M$, any point-derivation on $\mathcal{O}_{M,p}^{(\infty)}$ vanishes on $\mathcal{S}_{M,p}^{(\infty)}$, the ring of stationary germs. Consequently, $T_p(M) = \mathcal{D}_p^{(\infty)}(M)$.*

Proof. Pick some chart, (U, φ), at p, where U is convex (for instance, an open ball) and let $\mathbf{f}$ be any stationary germ. If we apply Proposition 3.10 to $f \circ \varphi^{-1}$ (for any $f \in \mathbf{f}$) and then compose with φ, we get

$$f = f(p) + \sum_{i=1}^{n} \left.\frac{\partial (f \circ \varphi^{-1})}{\partial X_i}\right|_{\varphi(p)} (x_i - x_i(p)) + \sum_{i,j=1}^{n} (x_i - x_i(p))(x_j - x_j(p))h,$$

near p, where h is C^∞. Since $\mathbf{f}$ is a stationary germ, this yields

$$f = f(p) + \sum_{i,j=1}^{n} (x_i - x_i(p))(x_j - x_j(p))h.$$

If v is any point-derivation, we get

$$\begin{aligned} v(f) = v(f(p)) + \sum_{i,j=1}^{n} \Big[& (x_i - x_i(p))(p)(x_j - x_j(p))(p)v(h) \\ & + (x_i - x_i(p))(p)v(x_j - x_j(p))h(p) + v(x_i - x_i(p))(x_j - x_j(p))(p)h(p)\Big] = 0. \end{aligned}$$

Thus, v vanishes on stationary germs. □

Proposition 3.11 shows that in the case of a smooth manifold, in Definition 3.14, we can omit the requirement that point-derivations vanish on stationary germs, since this is automatic. It is also possible to define $T_p(M)$ just in terms of $\mathcal{O}^{(\infty)}_{M,p}$. Let $\mathfrak{m}_{M,p} \subseteq \mathcal{O}^{(\infty)}_{M,p}$ be the ideal of germs that vanish at p. Then, we also have the ideal $\mathfrak{m}^2_{M,p}$, which consists of all finite sums of products of two elements in $\mathfrak{m}_{M,p}$ and it turns out that $T^*_p(M)$ is isomorphic to $\mathfrak{m}_{M,p}/\mathfrak{m}^2_{M,p}$ (see Warner [148], Lemma 1.16).

Actually, if we let $\mathfrak{m}^{(k)}_{M,p} \subseteq \mathcal{O}^{(k)}_{M,p}$ denote the ideal of C^k-germs that vanish at p and $\mathfrak{s}^{(k)}_{M,p} \subseteq \mathcal{S}^{(k)}_{M,p}$ denote the ideal of stationary C^k-germs that vanish at p, adapting Warner's argument, we can prove the following proposition:

Proposition 3.12. *We have the inclusion, $(\mathfrak{m}^{(k)}_{M,p})^2 \subseteq \mathfrak{s}^{(k)}_{M,p}$ and the isomorphism*

$$(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^* \cong (\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p})^*.$$

As a consequence, $T_p(M) \cong (\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p})^*$ and $T^*_p(M) \cong \mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p}$.

Proof. Given any two germs, $\mathbf{f}, \mathbf{g} \in \mathfrak{m}^{(k)}_{M,p}$, for any two locally defined functions, $f \in \mathbf{f}$ and $g \in \mathbf{g}$, since $f(p) = g(p) = 0$, for any chart, (U, φ), with $p \in U$, by definition of the product fg of two functions, for any $q \in M$ near p, we have

$$\begin{aligned} (fg \circ \varphi^{-1})(q) &= (fg)(\varphi^{-1}(q)) \\ &= f(\varphi^{-1}(q))g(\varphi^{-1}(q)) \\ &= (f \circ \varphi^{-1})(q)(g \circ \varphi^{-1})(q), \end{aligned}$$

so

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

and by the product rule for derivatives, we get

$$(fg \circ \varphi^{-1})'(0) = (f \circ \varphi^{-1})'(0)(g \circ \varphi^{-1})(0) + (f \circ \varphi^{-1})(0)(g \circ \varphi^{-1})'(0) = 0,$$

because $(g \circ \varphi^{-1})(0) = g(\varphi^{-1}(0)) = g(p) = 0$ and $(f \circ \varphi^{-1})(0) = f(\varphi^{-1}(0)) = f(p) = 0$. Therefore, fg is stationary at p and since $fg(p) = 0$, we have $\mathbf{fg} \in \mathfrak{s}^{(k)}_{M,p}$, which implies the inclusion $(\mathfrak{m}^{(k)}_{M,p})^2 \subseteq \mathfrak{s}^{(k)}_{M,p}$.

Now, the key point is that any constant germ is stationary since the derivative of a constant function is zero. Consequently, if v is a linear form on $\mathcal{O}^{(k)}_{M,p}$ vanishing on $\mathcal{S}^{(k)}_{M,p}$, then

$$v(\mathbf{f}) = v(\mathbf{f} - \mathbf{f}(\mathbf{p})),$$

for all $\mathbf{f} \in \mathcal{O}^{(k)}_{M,p}$, where $\mathbf{f}(\mathbf{p})$ denotes the germ of constant functions with value $\mathbf{f}(p)$. We use this fact to define two functions between $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ and $(\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p})^*$ which are mutual inverses.

The map from $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ to $(\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p})^*$ is restriction to $\mathfrak{m}^{(k)}_{M,p}$: every linear form v on $\mathcal{O}^{(k)}_{M,p}$ vanishing on $\mathcal{S}^{(k)}_{M,p}$ yields a linear form on $\mathfrak{m}^{(k)}_{M,p}$ that vanishes on $\mathfrak{s}^{(k)}_{M,p}$.

Conversely, for any linear form ℓ on $\mathfrak{m}^{(k)}_{M,p}$ vanishing on $\mathfrak{s}^{(k)}_{M,p}$, define the function v_ℓ so that

$$v_\ell(\mathbf{f}) = \ell(\mathbf{f} - \mathbf{f}(\mathbf{p})),$$

for any germ $\mathbf{f} \in \mathcal{O}^{(k)}_{M,p}$. Since ℓ is linear, it is clear that v_ℓ is also linear. If f is stationary at p, then $f - f(p)$ is also stationary at p because the derivative of a constant is zero. Obviously, $f - f(p)$ vanishes at p. It follows that v_ℓ vanishes on stationary germs at p.

Using the fact that $v(\mathbf{f}) = v(\mathbf{f} - \mathbf{f}(\mathbf{p}))$, it is easy to check that the above maps between $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ and $(\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p})^*$ are mutual inverses, establishing the desired isomorphism. Because $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ is finite-dimensional, we also have the isomorphism

$$\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p} \cong \mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p}$$

which yields the isomorphims $T_p(M) \cong (\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p})^*$ and $T^*_p(M) \cong \mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p}$. □

When $k = \infty$, Proposition 3.10 shows that every stationary germ that vanishes at p belongs to $\mathfrak{m}^2_{M,p}$. Therefore, when $k = \infty$, we have $\mathfrak{s}^{(\infty)}_{M,p} = \mathfrak{m}^2_{M,p}$ and so, we obtain the result quoted above (from Warner):

$$T^*_p(M) = \mathcal{O}^{(\infty)}_{M,p}/\mathcal{S}^{(\infty)}_{M,p} \cong \mathfrak{m}_{M,p}/\mathfrak{m}^2_{M,p}.$$

Remarks:

(1) The isomorphism

$$(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^* \cong (\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p})^*$$

yields another proof that the linear forms in $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ are point-derivations, using the argument from Warner [148] (Lemma 1.16). It is enough to prove that every linear

form of the form v_ℓ is a point-derivation. Indeed, if ℓ is a linear form on $\mathfrak{m}_{M,p}^{(k)}$ vanishing on $\mathfrak{s}_{M,p}^{(k)}$, we have

$$\begin{aligned} v_\ell(\mathbf{fg}) &= \ell(\mathbf{fg} - \mathbf{f}(\mathbf{p})\mathbf{g}(\mathbf{p})) \\ &= \ell\big((\mathbf{f} - \mathbf{f}(\mathbf{p}))(\mathbf{g} - \mathbf{g}(\mathbf{p})) + (\mathbf{f} - \mathbf{f}(\mathbf{p}))\mathbf{g}(\mathbf{p}) + \mathbf{f}(\mathbf{p})(\mathbf{g} - \mathbf{g}(\mathbf{p}))\big) \\ &= \ell\big((\mathbf{f} - \mathbf{f}(\mathbf{p}))(\mathbf{g} - \mathbf{g}(\mathbf{p}))\big) + \ell(\mathbf{f} - \mathbf{f}(\mathbf{p}))\mathbf{g}(p) + \mathbf{f}(p)\ell(\mathbf{g} - \mathbf{g}(\mathbf{p})) \\ &= v_\ell(\mathbf{f})\mathbf{g}(p) + \mathbf{f}(p)v_\ell(\mathbf{g}), \end{aligned}$$

using the fact that $\ell\big((\mathbf{f} - \mathbf{f}(\mathbf{p}))(\mathbf{g} - \mathbf{g}(\mathbf{p}))\big) = 0$ since $(\mathfrak{m}_{M,p}^{(k)})^2 \subseteq \mathfrak{s}_{M,p}^{(k)}$ and ℓ vanishes on $\mathfrak{s}_{M,p}^{(k)}$, which proves that v_ℓ is a point-derivation.

(2) The ideal $\mathfrak{m}_{M,p}^{(k)}$ is in fact the unique maximal ideal of $\mathcal{O}_{M,p}^{(k)}$. This is because if $\mathbf{f} \in \mathcal{O}_{M,p}^{(k)}$ does not vanish at p, then it is an invertible element of $\mathcal{O}_{M,p}^{(k)}$ and any ideal containing $\mathfrak{m}_{M,p}^{(k)}$ and $\mathbf{f}$ would be equal to $\mathcal{O}_{M,p}^{(k)}$, which it absurd. Thus, $\mathcal{O}_{M,p}^{(k)}$ is a local ring (in the sense of commutative algebra) called the *local ring of germs of C^k-functions at p*. These rings play a crucial role in algebraic geometry.

(3) Using the map $\mathbf{f} \mapsto \mathbf{f} - \mathbf{f}(\mathbf{p})$, it is easy to see that

$$\mathcal{O}_{M,p}^{(k)} \cong \mathbb{R} \oplus \mathfrak{m}_{M,p}^{(k)} \quad \text{and} \quad \mathcal{S}_{M,p}^{(k)} \cong \mathbb{R} \oplus \mathfrak{s}_{M,p}^{(k)}.$$

Yet one more way of defining tangent vectors will make it a little easier to define tangent bundles.

Definition 3.16. (Tangent Vectors, Version 3) Given any C^k-manifold, M, of dimension n, with $k \geq 1$, for any $p \in M$, consider the triples, (U, φ, u), where (U, φ) is any chart at p and u is any vector in $\mathbb{R}^n$. Say that two such triples (U, φ, u) and (V, ψ, v) are *equivalent* iff

$$(\psi \circ \varphi^{-1})'_{\varphi(p)}(u) = v.$$

A *tangent vector* to M at p is an equivalence class of triples, $[(U, \varphi, u)]$, for the above equivalence relation.

The intuition behind Definition 3.16 is quite clear: The vector u is considered as a tangent vector to $\mathbb{R}^n$ at $\varphi(p)$. If (U, φ) is a chart on M at p, we can define a natural isomorphism, $\theta_{U,\varphi,p} \colon \mathbb{R}^n \to T_p(M)$, between $\mathbb{R}^n$ and $T_p(M)$, as follows: For any $u \in \mathbb{R}^n$,

$$\theta_{U,\varphi,p} \colon u \mapsto [(U, \varphi, u)].$$

One immediately checks that the above map is indeed linear and a bijection.

The equivalence of this definition with the definition in terms of curves (Definition 3.10) is easy to prove.

Proposition 3.13. *Let M be any C^k-manifold of dimension n, with $k \geq 1$. For every $p \in M$, for every chart, (U, φ), at p, if x is any tangent vector (version 1) given by some equivalence class of C^1-curves, $\gamma\colon]-\epsilon, +\epsilon[\to M$, through p (i.e., $p = \gamma(0)$), then the map*

$$x \mapsto [(U, \varphi, (\varphi \circ \gamma)'(0))]$$

is an isomorphism between $T_p(M)$-version 1 and $T_p(M)$-version 3.

Proof. If σ is another curve equivalent to γ, then $(\varphi \circ \gamma)'(0) = (\varphi \circ \sigma)'(0)$, so the map is well-defined. It is clearly injective. As for surjectivity, define the curve, γ, on M through p by

$$\gamma(t) = \varphi^{-1}(\varphi(p) + tu).$$

Then, $(\varphi \circ \gamma)(t) = \varphi(p) + tu$ and

$$(\varphi \circ \gamma)'(0) = u.$$

□

After having explored thorougly the notion of tangent vector, we show how a C^k-map, $h\colon M \to N$, between C^k manifolds, induces a linear map, $dh_p\colon T_p(M) \to T_{h(p)}(N)$, for every $p \in M$. We find it convenient to use Version 2 of the definition of a tangent vector. So, let $u \in T_p(M)$ be a point-derivation on $\mathcal{O}^{(k)}_{M,p}$ that vanishes on $\mathcal{S}^{(k)}_{M,p}$. We would like $dh_p(u)$ to be a point-derivation on $\mathcal{O}^{(k)}_{N,h(p)}$ that vanishes on $\mathcal{S}^{(k)}_{N,h(p)}$. Now, for every germ, $\mathbf{g} \in \mathcal{O}^{(k)}_{N,h(p)}$, if $g \in \mathbf{g}$ is any locally defined function at $h(p)$, it is clear that $g \circ h$ is locally defined at p and is C^k and that if $g_1, g_2 \in \mathbf{g}$ then $g_1 \circ h$ and $g_2 \circ h$ are equivalent. The germ of all locally defined functions at p of the form $g \circ h$, with $g \in \mathbf{g}$, will be denoted $\mathbf{g} \circ h$. Then, we set

$$dh_p(u)(\mathbf{g}) = u(\mathbf{g} \circ h).$$

Moreover, if $\mathbf{g}$ is a stationary germ at $h(p)$, then for some chart, (V, ψ) on N at $q = h(p)$, we have $(g \circ \psi^{-1})'(\psi(q)) = 0$ and, for any chart, (U, φ), at p on M, we get

$$(g \circ h \circ \varphi^{-1})'(\varphi(p)) = (g \circ \psi^{-1})'(\psi(q))((\psi \circ h \circ \varphi^{-1})'(\varphi(p))) = 0,$$

which means that $\mathbf{g} \circ h$ is stationary at p. Therefore, $dh_p(u) \in T_{h(p)}(M)$. It is also clear that dh_p is a linear map. We summarize all this in the following definition:

Definition 3.17. Given any two C^k-manifolds, M and N, of dimension m and n, respectively, for any C^k-map, $h\colon M \to N$, and for every $p \in M$, the *differential of h at p* or *tangent map*, $dh_p\colon T_p(M) \to T_{h(p)}(N)$, is the linear map defined so that

$$dh_p(u)(\mathbf{g}) = u(\mathbf{g} \circ h),$$

for every $u \in T_p(M)$ and every germ, $\mathbf{g} \in \mathcal{O}^{(k)}_{N,h(p)}$. The linear map dh_p is also denoted $T_p h$ (and sometimes, h'_p or $D_p h$).

The chain rule is easily generalized to manifolds.

Proposition 3.14. *Given any two C^k-maps $f\colon M \to N$ and $g\colon N \to P$ between smooth C^k-manifolds, for any $p \in M$, we have*

$$d(g \circ f)_p = dg_{f(p)} \circ df_p.$$

In the special case where $N = \mathbb{R}$, a C^k-map between the manifolds M and $\mathbb{R}$ is just a C^k-function on M. It is interesting to see what df_p is explicitly. Since $N = \mathbb{R}$, germs (of functions on $\mathbb{R}$) at $t_0 = f(p)$ are just germs of C^k-functions, $g\colon \mathbb{R} \to \mathbb{R}$, locally defined at t_0. Then, for any $u \in T_p(M)$ and every germ $\mathbf{g}$ at t_0,

$$df_p(u)(\mathbf{g}) = u(\mathbf{g} \circ f).$$

If we pick a chart, (U, φ), on M at p, we know that the $\left(\frac{\partial}{\partial x_i}\right)_p$ form a basis of $T_p(M)$, with $1 \leq i \leq n$. Therefore, it is enough to figure out what $df_p(u)(\mathbf{g})$ is when $u = \left(\frac{\partial}{\partial x_i}\right)_p$. In this case,

$$df_p\left(\left(\frac{\partial}{\partial x_i}\right)_p\right)(\mathbf{g}) = \left.\frac{\partial(g \circ f \circ \varphi^{-1})}{\partial X_i}\right|_{\varphi(p)}.$$

Using the chain rule, we find that

$$df_p\left(\left(\frac{\partial}{\partial x_i}\right)_p\right)(\mathbf{g}) = \left(\frac{\partial}{\partial x_i}\right)_p f \left.\frac{dg}{dt}\right|_{t_0}.$$

Therefore, we have

$$df_p(u) = u(\mathbf{f}) \left.\frac{d}{dt}\right|_{t_0}.$$

This shows that we can identify df_p with the linear form in $T_p^*(M)$ defined by

$$df_p(u) = u(\mathbf{f}), \qquad u \in T_pM,$$

by identifying $T_{t_0}\mathbb{R}$ with $\mathbb{R}$. This is consistent with our previous definition of df_p as the image of f in $T_p^*(M) = \mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)}$ (as $T_p(M)$ is isomorphic to $(\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)})^*$).

Again, even though this is just a restatement of facts we already showed, we state the following proposition because of its practical usefulness:

Proposition 3.15. *Given any C^k-manifold, M, of dimension n, with $k \geq 1$, for any $p \in M$ and any chart (U, φ) at p, the n linear maps,*

$$(dx_1)_p, \ldots, (dx_n)_p,$$

form a basis of T_p^*M, where $(dx_i)_p$, the differential of x_i at p, is identified with the linear form in T_p^*M such that $(dx_i)_p(v) = v(\mathbf{x_i})$, for every $v \in T_pM$ (by identifying $T_\lambda\mathbb{R}$ with $\mathbb{R}$).

In preparation for the definition of the flow of a vector field (which will be needed to define the exponential map in Lie group theory), we need to define the tangent vector to a curve on a manifold. Given a C^k-curve, $\gamma\colon]a,b[\to M$, on a C^k-manifold, M, for any $t_0 \in]a,b[$, we would like to define the tangent vector to the curve γ at t_0 as a tangent vector to M at $p = \gamma(t_0)$. We do this as follows: Recall that $\frac{d}{dt}\big|_{t_0}$ is a basis vector of $T_{t_0}(\mathbb{R}) = \mathbb{R}$. So, define the *tangent vector to the curve γ at t_0*, denoted $\dot{\gamma}(t_0)$ (or $\gamma'(t_0)$, or $\frac{d\gamma}{dt}(t_0)$) by

$$\dot{\gamma}(t_0) = d\gamma_{t_0}\left(\frac{d}{dt}\Big|_{t_0}\right).$$

Sometime, it is necessary to define curves (in a manifold) whose domain is not an open interval. A map, $\gamma\colon [a,b] \to M$, is a *C^k-curve in M* if it is the restriction of some C^k-curve, $\widetilde{\gamma}\colon]a-\epsilon, b+\epsilon[\to M$, for some (small) $\epsilon > 0$. Note that for such a curve (if $k \geq 1$) the tangent vector, $\dot{\gamma}(t)$, is defined for all $t \in [a,b]$. A continuous curve, $\gamma\colon [a,b] \to M$, is *piecewise C^k* iff there a sequence, $a_0 = a, a_1, \ldots, a_m = b$, so that the restriction, γ_i, of γ to each $[a_i, a_{i+1}]$ is a C^k-curve, for $i = 0, \ldots, m-1$. This implies that $\gamma_i'(a_{i+1})$ and $\gamma_{i+1}'(a_{i+1})$ are defined for $i = 0, \ldots, m-1$, but there may be a jump in the tangent vector to γ at a_i, that is, we may have $\gamma_i'(a_{i+1}) \neq \gamma_{i+1}'(a_{i+1})$.

3.3 Tangent and Cotangent Bundles, Vector Fields, Lie Derivative

Let M be a C^k-manifold (with $k \geq 2$). Roughly speaking, a vector field on M is the assignment, $p \mapsto X(p)$, of a tangent vector, $X(p) \in T_p(M)$, to a point $p \in M$. Generally, we would like such assignments to have some smoothness properties when p varies in M, for example, to be C^l, for some l related to k. Now, if the collection, $T(M)$, of all tangent spaces, $T_p(M)$, was a C^l-manifold, then it would be very easy to define what we mean by a C^l-vector field: We would simply require the map, $X\colon M \to T(M)$, to be C^l.

If M is a C^k-manifold of dimension n, then we can indeed make $T(M)$ into a C^{k-1}-manifold of dimension $2n$ and we now sketch this construction.

We find it most convenient to use Version 3 of the definition of tangent vectors, i.e., as equivalence classes of triples (U, φ, x), where (U, φ) is a chart and $x \in \mathbb{R}^n$. First, we let $T(M)$ be the disjoint union of the tangent spaces $T_p(M)$, for all $p \in M$. Formally,

$$T(M) = \{(p, v) \mid p \in M, v \in T_p(M)\}.$$

There is a *natural projection*,

$$\pi\colon T(M) \to M, \quad \text{with} \quad \pi(p, v) = p.$$

We still have to give $T(M)$ a topology and to define a C^{k-1}-atlas. For every chart, (U, φ), of M (with U open in M) we define the function, $\widetilde{\varphi}\colon \pi^{-1}(U) \to \mathbb{R}^{2n}$, by

$$\widetilde{\varphi}(p, v) = (\varphi(p), \theta^{-1}_{U,\varphi,p)}(v)),$$

where $(p, v) \in \pi^{-1}(U)$ and $\theta_{U,\varphi,p}$ is the isomorphism between $\mathbb{R}^n$ and $T_p(M)$ described just after Definition 3.16. It is obvious that $\widetilde{\varphi}$ is a bijection between $\pi^{-1}(U)$ and $\varphi(U) \times \mathbb{R}^n$, an open subset of $\mathbb{R}^{2n}$. We give $T(M)$ the weakest topology that makes all the $\widetilde{\varphi}$ continuous, i.e., we take the collection of subsets of the form $\widetilde{\varphi}^{-1}(W)$, where W is any open subset of $\varphi(U) \times \mathbb{R}^n$, as a basis of the topology of $T(M)$. One easily checks that $T(M)$ is Hausdorff and second-countable in this topology. If (U, φ) and (V, ψ) are two overlapping charts, then the definition of the equivalence relation on triples (U, φ, x) and (V, ψ, y) immediately implies that

$$\theta^{-1}_{(V,\psi,p)} \circ \theta_{(U,\varphi,p)} = (\psi \circ \varphi^{-1})'_z$$

for all $p \in U \cap V$, with $z = \varphi(p) = \psi(p)$, so the transition map,

$$\widetilde{\psi} \circ \widetilde{\varphi}^{-1} \colon \varphi(U \cap V) \times \mathbb{R}^n \longrightarrow \psi(U \cap V) \times \mathbb{R}^n$$

is given by

$$\widetilde{\psi} \circ \widetilde{\varphi}^{-1}(z, x) = (\psi \circ \varphi^{-1}(z), (\psi \circ \varphi^{-1})'_z(x)), \qquad (z, x) \in \varphi(U \cap V) \times \mathbb{R}^n.$$

It is clear that $\widetilde{\psi} \circ \widetilde{\varphi}^{-1}$ is a C^{k-1}-map. Therefore, $T(M)$ is indeed a C^{k-1}-manifold of dimension $2n$, called the *tangent bundle*.

Remark: Even if the manifold M is naturally embedded in $\mathbb{R}^N$ (for some $N \geq n = \dim(M)$), it is not at all obvious how to view the tangent bundle, $T(M)$, as embedded in $\mathbb{R}^{N'}$, for some suitable N'. Hence, we see that the definition of an abtract manifold is unavoidable.

A similar construction can be carried out for the cotangent bundle. In this case, we let $T^*(M)$ be the disjoint union of the cotangent spaces $T^*_p(M)$, that is,

$$T^*(M) = \{(p, \omega) \mid p \in M, \omega \in T^*_p(M)\}.$$

We also have a natural projection $\pi \colon T^*(M) \to M$ with $\pi(p, \omega) = p$, and we can define charts in several ways. One method used by Warner [148] goes as follows: For any chart, (U, φ), on M, we define the function,
$\widetilde{\varphi} \colon \pi^{-1}(U) \to \mathbb{R}^{2n}$, by

$$\widetilde{\varphi}(p, \omega) = \left(\varphi(p), \omega\left(\left(\frac{\partial}{\partial x_1}\right)_p\right), \ldots, \omega\left(\left(\frac{\partial}{\partial x_n}\right)_p\right)\right),$$

where $(p, \omega) \in \pi^{-1}(U)$ and the $\left(\frac{\partial}{\partial x_i}\right)_p$ are the basis of $T_p(M)$ associated with the chart (U, φ). Again, one can make $T^*(M)$ into a C^{k-1}-manifold of dimension $2n$, called the *cotangent bundle*. We leave the details as an exercise to the reader (Or, look at Berger and Gostiaux [17]). Another method using Version 3 of the definition of tangent vectors is presented in Section 7.2. For each chart (U, φ) on M, we obtain a chart

$$\widetilde{\varphi}^* \colon \pi^{-1}(U) \to \varphi(U) \times \mathbb{R}^n \subseteq \mathbb{R}^{2n}$$

on $T^*(M)$ given by

$$\widetilde{\varphi}^*(p,\omega) = (\varphi(p), \theta^*_{U,\varphi,\pi(\omega)}(\omega))$$

for all $(p,\omega) \in \pi^{-1}(U)$, where

$$\theta^*_{U,\varphi,p} = \iota \circ \theta^{\top}_{U,\varphi,p} \colon T^*_p(M) \to \mathbb{R}^n.$$

Here, $\theta^{\top}_{U,\varphi,p} \colon T^*_p(M) \to (\mathbb{R}^n)^*$ is obtained by dualizing the map, $\theta_{U,\varphi,p} \colon \mathbb{R}^n \to T_p(M)$ and $\iota \colon (\mathbb{R}^n)^* \to \mathbb{R}^n$ is the isomorphism induced by the canonical basis $(e_1, \ldots, e_n)$ of $\mathbb{R}^n$ and its dual basis.

For simplicity of notation, we also use the notation TM for $T(M)$ (resp. T^*M for $T^*(M)$).

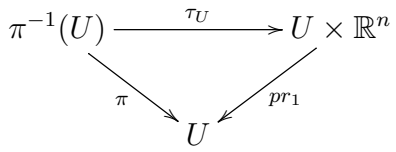
Observe that for every chart, (U,φ), on M, there is a bijection

$$\tau_U \colon \pi^{-1}(U) \to U \times \mathbb{R}^n,$$

given by

$$\tau_U(p,v) = (p, \theta^{-1}_{U,\varphi,p}(v)).$$

Clearly, $pr_1 \circ \tau_U = \pi$, on $\pi^{-1}(U)$ as illustrated by the following commutative diagram:



Thus locally, that is, over U, the bundle $T(M)$ looks like the product manifold $U \times \mathbb{R}^n$. We say that $T(M)$ is *locally trivial* (over U) and we call τ_U a *trivializing map*. For any $p \in M$, the vector space $\pi^{-1}(p) = \{p\} \times T_p(M) \cong T_p(M)$ is called the *fibre above* p. Observe that the restriction of τ_U to $\pi^{-1}(p)$ is a linear isomorphism between $\{p\} \times T_p(M) \cong T_p(M)$ and $\{p\} \times \mathbb{R}^n \cong \mathbb{R}^n$, for any $p \in M$. Furthermore, for any two overlapping charts (U,φ) and (V,ψ), there is a function $g_{UV} \colon U \cap V \to \mathrm{GL}(n,\mathbb{R})$ such that

$$(\tau_U \circ \tau_V^{-1})(p,x) = (p, g_{UV}(p)(x))$$

for all $p \in U \cap V$ and all $x \in \mathbb{R}^n$, with $g_{UV}(p)$ given by

$$g_{UV}(p) = (\varphi \circ \psi^{-1})'_{\varphi(p)}.$$

Obviously, $g_{UV}(p)$ is a linear isomorphism of $\mathbb{R}^n$ for all $p \in U \cap V$. The maps $g_{UV}(p)$ are called the *transition functions* of the tangent bundle.

All these ingredients are part of being a *vector bundle*. For more on bundles, see Chapter 7, in particular, Section 7.2 on vector bundles where the construction of the bundles TM and T^*M is worked out in detail. See also the references in Chapter 7.

When $M = \mathbb{R}^n$, observe that $T(M) = M \times \mathbb{R}^n = \mathbb{R}^n \times \mathbb{R}^n$, i.e., the bundle $T(M)$ is (globally) trivial.

Given a C^k-map, $h\colon M \to N$, between two C^k-manifolds, we can define the function, $dh\colon T(M) \to T(N)$, (also denoted Th, or h_*, or Dh) by setting

$$dh(u) = dh_p(u), \quad \text{iff} \quad u \in T_p(M).$$

We leave the next proposition as an exercise to the reader (A proof can be found in Berger and Gostiaux [17]).

Proposition 3.16. *Given a C^k-map, $h\colon M \to N$, between two C^k-manifolds M and N (with $k \geq 1$), the map $dh\colon T(M) \to T(N)$ is a C^{k-1} map.*

We are now ready to define vector fields.

Definition 3.18. Let M be a C^{k+1} manifold, with $k \geq 1$. For any open subset, U of M, a *vector field on U* is any *section, X, of $T(M)$ over U*, i.e., any function, $X\colon U \to T(M)$, such that $\pi \circ X = \mathrm{id}_U$ (i.e., $X(p) \in T_p(M)$, for every $p \in U$). We also say that X is a *lifting of U into $T(M)$*. We say that X is a *C^k-vector field on U* iff X is a section over U and a C^k-map. The set of C^k-vector fields over U is denoted $\Gamma^{(k)}(U, T(M))$. Given a curve, $\gamma\colon [a,b] \to M$, a *vector field, X, along γ* is any section of $T(M)$ over γ, i.e., a C^k-function, $X\colon [a,b] \to T(M)$, such that $\pi \circ X = \gamma$. We also say that X *lifts γ into $T(M)$*.

The above definition gives a precise meaning to the idea that a C^k-vector field on M is an assignment, $p \mapsto X(p)$, of a tangent vector, $X(p) \in T_p(M)$, to a point, $p \in M$, so that $X(p)$ varies in a C^k-fashion in terms of p.

Clearly, $\Gamma^{(k)}(U, T(M))$ is a real vector space. For short, the space $\Gamma^{(k)}(M, T(M))$ is also denoted by $\Gamma^{(k)}(T(M))$ (or $\mathfrak{X}^{(k)}(M)$ or even $\Gamma(T(M))$ or $\mathfrak{X}(M)$).

Remark: We can also define a *C^j-vector field on U* as a section, X, over U which is a C^j-map, where $0 \leq j \leq k$. Then, we have the vector space, $\Gamma^{(j)}(U, T(M))$, *etc* .

If $M = \mathbb{R}^n$ and U is an open subset of M, then $T(M) = \mathbb{R}^n \times \mathbb{R}^n$ and a section of $T(M)$ over U is simply a function, X, such that

$$X(p) = (p, u), \quad \text{with} \quad u \in \mathbb{R}^n,$$

for all $p \in U$. In other words, X is defined by a function, $f\colon U \to \mathbb{R}^n$ (namely, $f(p) = u$). This corresponds to the "old" definition of a vector field in the more basic case where the manifold, M, is just $\mathbb{R}^n$.

For any vector field $X \in \Gamma^{(k)}(U, T(M))$ and for any $p \in U$, we have $X(p) = (p, v)$ for some $v \in T_p(M)$, and it is convenient to denote the vector v by X_p so that $X(p) = (p, X_p)$. In fact, in most situations it is convenient to identify $X(p)$ with $X_p \in T_p(M)$, and we will do so from now on. This amounts to identifying the isomorphic vector spaces $\{p\} \times T_p(M)$ and $T_p(M)$, which we always do. Let us illustrate the advantage of this convention with the next definition.

Given any C^k-function, $f \in \mathcal{C}^k(U)$, and a vector field, $X \in \Gamma^{(k)}(U, T(M))$, we define the vector field, fX, by

$$(fX)_p = f(p)X_p, \quad p \in U.$$

Obviously, $fX \in \Gamma^{(k)}(U, T(M))$, which shows that $\Gamma^{(k)}(U, T(M))$ is also a $\mathcal{C}^k(U)$-module. For any chart, (U, φ), on M it is easy to check that the map

$$p \mapsto \left(\frac{\partial}{\partial x_i}\right)_p, \quad p \in U,$$

is a C^k-vector field on U (with $1 \leq i \leq n$). This vector field is denoted $\left(\frac{\partial}{\partial x_i}\right)$ or $\frac{\partial}{\partial x_i}$.

Definition 3.19. Let M be a C^{k+1} manifold and let X be a C^k vector field on M. If U is any open subset of M and f is any function in $\mathcal{C}^k(U)$, then the *Lie derivative of f with respect to X*, denoted $X(f)$ or $L_X f$, is the function on U given by

$$X(f)(p) = X_p(f) = X_p(\mathbf{f}), \qquad p \in U.$$

Observe that

$$X(f)(p) = df_p(X_p),$$

where df_p is identified with the linear form in $T_p^*(M)$ defined by

$$df_p(v) = v(\mathbf{f}), \qquad v \in T_pM,$$

by identifying $T_{t_0}\mathbb{R}$ with $\mathbb{R}$ (see the discussion following Proposition 3.14). The Lie derivative, $L_X f$, is also denoted $X[f]$.

As a special case, when (U, φ) is a chart on M, the vector field, $\frac{\partial}{\partial x_i}$, just defined above induces the function

$$p \mapsto \left(\frac{\partial}{\partial x_i}\right)_p f, \quad p \in U,$$

denoted $\frac{\partial}{\partial x_i}(f)$ or $\left(\frac{\partial}{\partial x_i}\right) f$.

It is easy to check that $X(f) \in \mathcal{C}^{k-1}(U)$. As a consequence, every vector field $X \in \Gamma^{(k)}(U, T(M))$ induces a linear map,

$$L_X \colon \mathcal{C}^k(U) \longrightarrow \mathcal{C}^{k-1}(U),$$

given by $f \mapsto X(f)$. It is immediate to check that L_X has the Leibniz property, i.e.,

$$L_X(fg) = L_X(f)g + fL_X(g).$$

Linear maps with this property are called *derivations*. Thus, we see that every vector field induces some kind of differential operator, namely, a linear derivation. Unfortunately, not

every linear derivation of the above type arises from a vector field, although this turns out to be true in the smooth case i.e., when $k = \infty$ (for a proof, see Gallot, Hulin and Lafontaine [61] or Lafontaine [93]).

In the rest of this section, unless stated otherwise, we assume that $k \geq 1$. The following easy proposition holds (c.f. Warner [148]):

Proposition 3.17. *Let X be a vector field on the C^{k+1}-manifold, M, of dimension n. Then, the following are equivalent:*

(a) X is C^k.

(b) If (U, φ) is a chart on M and if $f_1, \ldots, f_n$ are the functions on U uniquely defined by

$$X \restriction U = \sum_{i=1}^{n} f_i \frac{\partial}{\partial x_i},$$

then each f_i is a C^k-map.

(c) Whenever U is open in M and $f \in \mathcal{C}^k(U)$, then $X(f) \in \mathcal{C}^{k-1}(U)$.

Given any two C^k-vector field, X, Y, on M, for any function, $f \in \mathcal{C}^k(M)$, we defined above the function $X(f)$ and $Y(f)$. Thus, we can form $X(Y(f))$ (resp. $Y(X(f))$), which are in $\mathcal{C}^{k-2}(M)$. Unfortunately, even in the smooth case, there is generally no vector field, Z, such that

$$Z(f) = X(Y(f)), \quad \text{for all } f \in \mathcal{C}^k(M).$$

This is because $X(Y(f))$ (and $Y(X(f))$) involve second-order derivatives. However, if we consider $X(Y(f)) - Y(X(f))$, then second-order derivatives cancel out and there is a unique vector field inducing the above differential operator. Intuitively, $XY - YX$ measures the "failure of X and Y to commute."

Proposition 3.18. *Given any C^{k+1}-manifold, M, of dimension n, for any two C^k-vector fields, X, Y, on M, there is a unique C^{k-1}-vector field, $[X, Y]$, such that*

$$[X, Y](f) = X(Y(f)) - Y(X(f)), \quad \textit{for all} \quad f \in \mathcal{C}^{k-1}(M).$$

Proof. First we prove uniqueness. For this it is enough to prove that $[X, Y]$ is uniquely defined on $\mathcal{C}^k(U)$, for any chart, (U, φ). Over U, we know that

$$X = \sum_{i=1}^{n} X_i \frac{\partial}{\partial x_i} \quad \text{and} \quad Y = \sum_{i=1}^{n} Y_i \frac{\partial}{\partial x_i},$$

where $X_i, Y_i \in \mathcal{C}^k(U)$. Then, for any $f \in \mathcal{C}^k(M)$, we have

$$\begin{aligned} X(Y(f)) &= X\left(\sum_{j=1}^{n} Y_j \frac{\partial}{\partial x_j}(f)\right) = \sum_{i,j=1}^{n} X_i \frac{\partial}{\partial x_i}(Y_j) \frac{\partial}{\partial x_j}(f) + \sum_{i,j=1}^{n} X_i Y_j \frac{\partial^2}{\partial x_j \partial x_i}(f) \\ Y(X(f)) &= Y\left(\sum_{i=1}^{n} X_i \frac{\partial}{\partial x_i}(f)\right) = \sum_{i,j=1}^{n} Y_j \frac{\partial}{\partial x_j}(X_i) \frac{\partial}{\partial x_i}(f) + \sum_{i,j=1}^{n} X_i Y_j \frac{\partial^2}{\partial x_i \partial x_j}(f). \end{aligned}$$

However, as $f \in \mathcal{C}^k(M)$, with $k \geq 2$, we have

$$\sum_{i,j=1}^{n} X_i Y_j \frac{\partial^2}{\partial x_j \partial x_i}(f) = \sum_{i,j=1}^{n} X_i Y_j \frac{\partial^2}{\partial x_i \partial x_j}(f),$$

and we deduce that

$$X(Y(f)) - Y(X(f)) = \sum_{i,j=1}^{n} \left(X_i \frac{\partial}{\partial x_i}(Y_j) - Y_i \frac{\partial}{\partial x_i}(X_j) \right) \frac{\partial}{\partial x_j}(f).$$

This proves that $[X, Y] = XY - YX$ is uniquely defined on U and that it is C^{k-1}. Thus, if $[X, Y]$ exists, it is unique.

To prove existence, we use the above expression to define $[X, Y]_U$, locally on U, for every chart, (U, φ). On any overlap, $U \cap V$, by the uniqueness property that we just proved, $[X, Y]_U$ and $[X, Y]_V$ must agree. But then, the $[X, Y]_U$ patch and yield a C^{k-1}-vector field defined on the whole of M. □

Definition 3.20. Given any C^{k+1}-manifold, M, of dimension n, for any two C^k-vector fields, X, Y, on M, the *Lie bracket*, $[X, Y]$, of X and Y, is the C^{k-1} vector field defined so that

$$[X, Y](f) = X(Y(f)) - Y(X(f)), \quad \text{for all} \quad f \in \mathcal{C}^{k-1}(M).$$

An an example, in $\mathbb{R}^3$, if X and Y are the two vector fields,

$$X = \frac{\partial}{\partial x} + y\frac{\partial}{\partial z} \quad \text{and} \quad Y = \frac{\partial}{\partial y},$$

then

$$[X, Y] = -\frac{\partial}{\partial z}.$$

We also have the following simple proposition whose proof is left as an exercise (or, see Do Carmo [51]):

Proposition 3.19. *Given any C^{k+1}-manifold, M, of dimension n, for any C^k-vector fields, X, Y, Z, on M, for all $f, g \in \mathcal{C}^k(M)$, we have:*

(a) $[[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0$ *(Jacobi identity).*

(b) $[X, X] = 0$.

(c) $[fX, gY] = fg[X, Y] + fX(g)Y - gY(f)X$.

(d) $[-, -]$ *is bilinear.*

As a consequence, for smooth manifolds ($k = \infty$), the space of vector fields, $\Gamma^{(\infty)}(T(M))$, is a vector space equipped with a bilinear operation, $[-,-]$, that satisfies the Jacobi identity. This makes $\Gamma^{(\infty)}(T(M))$ a *Lie algebra*.

Let $\varphi\colon M \to N$ be a diffeomorphism between two manifolds. Then, vector fields can be transported from N to M and conversely.

Definition 3.21. Let $\varphi\colon M \to N$ be a diffeomorphism between two C^{k+1} manifolds. For every C^k vector field, Y, on N, the *pull-back of Y along φ* is the vector field, φ^*Y, on M, given by

$$(\varphi^*Y)_p = d\varphi^{-1}_{\varphi(p)}(Y_{\varphi(p)}), \qquad p \in M.$$

For every C^k vector field, X, on M, the *push-forward of X along φ* is the vector field, φ_*X, on N, given by

$$\varphi_*X = (\varphi^{-1})^*X,$$

that is, for every $p \in M$,

$$(\varphi_*X)_{\varphi(p)} = d\varphi_p(X_p),$$

or equivalently,

$$(\varphi_*X)_q = d\varphi_{\varphi^{-1}(q)}(X_{\varphi^{-1}(q)}), \qquad q \in N.$$

It is not hard to check that

$$L_{\varphi_*X} f = L_X(f \circ \varphi) \circ \varphi^{-1},$$

for any function $f \in C^k(N)$.

One more notion will be needed when we deal with Lie algebras.

Definition 3.22. Let $\varphi\colon M \to N$ be a C^{k+1}-map of manifolds. If X is a C^k vector field on M and Y is a C^k vector field on N, we say that *X and Y are φ-related* iff

$$d\varphi \circ X = Y \circ \varphi.$$

The basic result about φ-related vector fields is:

Proposition 3.20. *Let $\varphi\colon M \to N$ be a C^{k+1}-map of manifolds, let X and Y be C^k vector fields on M and let X_1, Y_1 be C^k vector fields on N. If X is φ-related to X_1 and Y is φ-related to Y_1, then $[X,Y]$ is φ-related to $[X_1,Y_1]$.*

Proof. Basically, one needs to unwind the definitions, see Warner [148], Chapter 1. □

3.4 Submanifolds, Immersions, Embeddings

Although the notion of submanifold is intuitively rather clear, technically, it is a bit tricky. In fact, the reader may have noticed that many different definitions appear in books and that it is not obvious at first glance that these definitions are equivalent. What is important is that a submanifold, N, of a given manifold, M, not only have the topology induced M but also that the charts of N be somewhow induced by those of M. (Recall that if X is a topological space and Y is a subset of X, then the *subspace topology on* Y or *topology induced by* X *on* Y has for its open sets all subsets of the form $Y \cap U$, where U is an arbitary open subset of X.).

Given m, n, with $0 \leq m \leq n$, we can view $\mathbb{R}^m$ as a subspace of $\mathbb{R}^n$ using the inclusion

$$\mathbb{R}^m \cong \mathbb{R}^m \times \{\underbrace{(0, \ldots, 0)}_{n-m}\} \hookrightarrow \mathbb{R}^m \times \mathbb{R}^{n-m} = \mathbb{R}^n, \quad (x_1, \ldots, x_m) \mapsto (x_1, \ldots, x_m, \underbrace{0, \ldots, 0}_{n-m}).$$

Definition 3.23. Given a C^k-manifold, M, of dimension n, a subset, N, of M is an *m-dimensional submanifold of M* (where $0 \leq m \leq n$) iff for every point, $p \in N$, there is a chart, (U, φ), of M, with $p \in U$, so that

$$\varphi(U \cap N) = \varphi(U) \cap (\mathbb{R}^m \times \{0_{n-m}\}).$$

(We write $0_{n-m} = \underbrace{(0, \ldots, 0)}_{n-m}$.)

The subset, $U \cap N$, of Definition 3.23 is sometimes called a *slice* of (U, φ) and we say that (U, φ) is *adapted to* N (See O'Neill [120] or Warner [148]).

Other authors, including Warner [148], use the term submanifold in a broader sense than us and they use the word *embedded submanifold* for what is defined in Definition 3.23.

The following proposition has an almost trivial proof but it justifies the use of the word submanifold:

Proposition 3.21. *Given a C^k-manifold, M, of dimension n, for any submanifold, N, of M of dimension $m \leq n$, the family of pairs $(U \cap N, \varphi \restriction U \cap N)$, where (U, φ) ranges over the charts over any atlas for M, is an atlas for N, where N is given the subspace topology. Therefore, N inherits the structure of a C^k-manifold.*

In fact, every chart on N arises from a chart on M in the following precise sense:

Proposition 3.22. *Given a C^k-manifold, M, of dimension n and a submanifold, N, of M of dimension $m \leq n$, for any $p \in N$ and any chart, (W, η), of N at p, there is some chart, (U, φ), of M at p so that*

$$\varphi(U \cap N) = \varphi(U) \cap (\mathbb{R}^m \times \{0_{n-m}\}) \quad \textit{and} \quad \varphi \restriction U \cap N = \eta \restriction U \cap N,$$

where $p \in U \cap N \subseteq W$.

Proof. See Berger and Gostiaux [17] (Chapter 2). □

It is also useful to define more general kinds of "submanifolds."

Definition 3.24. Let $\varphi\colon N \to M$ be a C^k-map of manifolds.

(a) The map φ is an *immersion* of N into M iff $d\varphi_p$ is injective for all $p \in N$.

(b) The set $\varphi(N)$ is an *immersed submanifold* of M iff φ is an injective immersion.

(c) The map φ is an *embedding* of N into M iff φ is an injective immersion such that the induced map, $N \longrightarrow \varphi(N)$, is a homeomorphism, where $\varphi(N)$ is given the subspace topology (equivalently, φ is an open map from N into $\varphi(N)$ with the subspace topology). We say that $\varphi(N)$ (with the subspace topology) is an *embedded submanifold* of M.

(d) The map φ is a *submersion* of N into M iff $d\varphi_p$ is surjective for all $p \in N$.

Again, we warn our readers that certain authors (such as Warner [148]) call $\varphi(N)$, in (b), a submanifold of M! We prefer the terminology *immersed submanifold*.

The notion of immersed submanifold arises naturally in the framework of Lie groups. Indeed, the fundamental correspondence between Lie groups and Lie algebras involves Lie subgroups that are not necessarily closed. But, as we will see later, subgroups of Lie groups that are also submanifolds are always closed. It is thus necessary to have a more inclusive notion of submanifold for Lie groups and the concept of immersed submanifold is just what's needed.

Immersions of $\mathbb{R}$ into $\mathbb{R}^3$ are parametric curves and immersions of $\mathbb{R}^2$ into $\mathbb{R}^3$ are parametric surfaces. These have been extensively studied, for example, see DoCarmo [50], Berger and Gostiaux [17] or Gallier [60].

Immersions (i.e., subsets of the form $\varphi(N)$, where N is an immersion) are generally neither injective immersions (i.e., subsets of the form $\varphi(N)$, where N is an injective immersion) nor embeddings (or submanifolds). For example, immersions can have self-intersections, as the plane curve (nodal cubic): $x = t^2 - 1; y = t(t^2 - 1)$. Note that the cuspidal cubic, $t \mapsto (t^2, t^3)$, is an injective map, but it is not an immersion since its derivative at the origin is zero.

Injective immersions are generally not embeddings (or submanifolds) because $\varphi(N)$ may not be homeomorphic to N. An example is given by the Lemniscate of Bernoulli, an injective immersion of $\mathbb{R}$ into $\mathbb{R}^2$:

$$\begin{aligned} x &= \frac{t(1+t^2)}{1+t^4}, \\ y &= \frac{t(1-t^2)}{1+t^4}. \end{aligned}$$

Another interesting example is the immersion of $\mathbb{R}$ into the 2-torus, $T^2 = S^1 \times S^1 \subseteq \mathbb{R}^4$, given by

$$t \mapsto (\cos t, \sin t, \cos ct, \sin ct),$$

where $c \in \mathbb{R}$. One can show that the image of $\mathbb{R}$ under this immersion is closed in T^2 iff c is rational. Moreover, the image of this immersion is dense in T^2 but not closed iff c is irrational. The above example can be adapted to the torus in $\mathbb{R}^3$: One can show that the immersion given by

$$t \mapsto ((2 + \cos t)\cos(\sqrt{2}\, t), (2 + \cos t)\sin(\sqrt{2}\, t), \sin t),$$

is dense but not closed in the torus (in $\mathbb{R}^3$) given by

$$(s, t) \mapsto ((2 + \cos s)\cos t, (2 + \cos s)\sin t, \sin s),$$

where $s, t \in \mathbb{R}$.

There is, however, a close relationship between submanifolds and embeddings.

Proposition 3.23. *If N is a submanifold of M, then the inclusion map, $j\colon N \to M$, is an embedding. Conversely, if $\varphi\colon N \to M$ is an embedding, then $\varphi(N)$ with the subspace topology is a submanifold of M and φ is a diffeomorphism between N and $\varphi(N)$.*

Proof. See O'Neill [120] (Chapter 1) or Berger and Gostiaux [17] (Chapter 2). □

In summary, embedded submanifolds and (our) submanifolds coincide. Some authors refer to spaces of the form $\varphi(N)$, where φ is an injective immersion, as *immersed submanifolds* and we have adopted this terminology. However, in general, an immersed submanifold is *not* a submanifold. One case where this holds is when N is compact, since then, a bijective continuous map is a homeomorphism. For yet a notion of submanifold intermediate between immersed submanifolds and (our) submanifolds, see Sharpe [140] (Chapter 1).

Our next goal is to review and promote to manifolds some standard results about ordinary differential equations.

3.5 Integral Curves, Flow of a Vector Field, One-Parameter Groups of Diffeomorphisms

We begin with integral curves and (local) flows of vector fields on a manifold.

Definition 3.25. Let X be a C^{k-1} vector field on a C^k-manifold, M, ($k \geq 2$) and let p_0 be a point on M. An *integral curve (or trajectory) for X with initial condition p_0* is a C^{k-1}-curve $\gamma\colon I \to M$, so that

$$\dot{\gamma}(t) = X_{\gamma(t)}{}^{1} \quad \text{for all } t \in I, \quad \text{and} \quad \gamma(0) = p_0,$$

where $I =]a, b[\subseteq \mathbb{R}$ is an open interval containing 0.

[1]Recall our convention: if X is a vector field on M, then for every point $q \in M$ we identify $X(q) = (q, X_q)$ and X_q.

What definition 3.25 says is that an integral curve, γ, with initial condition p_0 is a curve on the manifold M passing through p_0 and such that, for every point $p = \gamma(t)$ on this curve, the tangent vector to this curve at p, i.e., $\dot{\gamma}(t)$, coincides with the value, X_p, of the vector field X at p.

Given a vector field, X, as above, and a point $p_0 \in M$, is there an integral curve through p_0? Is such a curve unique? If so, how large is the open interval I? We provide some answers to the above questions below.

Definition 3.26. Let X be a C^{k-1} vector field on a C^k-manifold, M, $(k \geq 2)$ and let p_0 be a point on M. A *local flow for X at p_0* is a map,

$$\varphi\colon J \times U \to M,$$

where $J \subseteq \mathbb{R}$ is an open interval containing 0 and U is an open subset of M containing p_0, so that for every $p \in U$, the curve $t \mapsto \varphi(t, p)$ is an integral curve of X with initial condition p.

Thus, a local low for X is a family of integral curves for all points in some small open set around p_0 such that these curves all have the same domain, J, independently of the initial condition, $p \in U$.

The following theorem is the main existence theorem of local flows. This is a promoted version of a similar theorem in the classical theory of ODE's in the case where M is an open subset of $\mathbb{R}^n$. For a full account of this theory, see Lang [96] or Berger and Gostiaux [17].

Theorem 3.24. *(Existence of a local flow) Let X be a C^{k-1} vector field on a C^k-manifold, M, $(k \geq 2)$ and let p_0 be a point on M. There is an open interval $J \subseteq \mathbb{R}$ containing 0 and an open subset $U \subseteq M$ containing p_0, so that there is a unique local flow $\varphi\colon J \times U \to M$ for X at p_0. What this means is that if $\varphi_1\colon J \times U \to M$ and $\varphi_2\colon J \times U \to M$ are both local flows with domain $J \times U$, then $\varphi_1 = \varphi_2$. Furthermore, φ is C^{k-1}.*

Theorem 3.24 holds under more general hypotheses, namely, when the vector field satisfies some *Lipschitz* condition, see Lang [96] or Berger and Gostiaux [17].

Now, we know that for any initial condition, p_0, there is some integral curve through p_0. However, there could be two (or more) integral curves $\gamma_1\colon I_1 \to M$ and $\gamma_2\colon I_2 \to M$ with initial condition p_0. This leads to the natural question: How do γ_1 and γ_2 differ on $I_1 \cap I_2$? The next proposition shows they don't!

Proposition 3.25. *Let X be a C^{k-1} vector field on a C^k-manifold, M, $(k \geq 2)$ and let p_0 be a point on M. If $\gamma_1\colon I_1 \to M$ and $\gamma_2\colon I_2 \to M$ are any two integral curves both with initial condition p_0, then $\gamma_1 = \gamma_2$ on $I_1 \cap I_2$.*

Proof. Let $Q = \{t \in I_1 \cap I_2 \mid \gamma_1(t) = \gamma_2(t)\}$. Since $\gamma_1(0) = \gamma_2(0) = p_0$, the set Q is nonempty. If we show that Q is both closed and open in $I_1 \cap I_2$, as $I_1 \cap I_2$ is connected since it is an open interval of $\mathbb{R}$, we will be able to conclude that $Q = I_1 \cap I_2$.

Since by definition, a manifold is Hausdorff, it is a standard fact in topology that the diagonal, $\Delta = \{(p, p) \mid p \in M\} \subseteq M \times M$, is closed, and since

$$Q = I_1 \cap I_2 \cap (\gamma_1, \gamma_2)^{-1}(\Delta)$$

and γ_1 and γ_2 are continuous, we see that Q is closed in $I_1 \cap I_2$.

Pick any $u \in Q$ and consider the curves β_1 and β_2 given by

$$\beta_1(t) = \gamma_1(t + u) \quad \text{and} \quad \beta_2(t) = \gamma_2(t + u),$$

where $t \in I_1 - u$ in the first case and $t \in I_2 - u$ in the second. (Here, if $I =]a, b[$, we have $I - u =]a - u, b - u[$.) Observe that

$$\dot{\beta}_1(t) = \dot{\gamma}_1(t + u) = X(\gamma_1(t + u)) = X(\beta_1(t))$$

and similarly, $\dot{\beta}_2(t) = X(\beta_2(t))$. We also have

$$\beta_1(0) = \gamma_1(u) = \gamma_2(u) = \beta_2(0) = q,$$

since $u \in Q$ (where $\gamma_1(u) = \gamma_2(u)$). Thus, $\beta_1 \colon (I_1 - u) \to M$ and $\beta_2 \colon (I_2 - u) \to M$ are two integral curves with the same initial condition, q. By Theorem 3.24, the uniqueness of local flow implies that there is some open interval, $\widetilde{I} \subseteq I_1 \cap I_2 - u$, such that $\beta_1 = \beta_2$ on $\widetilde{I}$. Consequently, γ_1 and γ_2 agree on $\widetilde{I} + u$, an open subset of Q, proving that Q is indeed open in $I_1 \cap I_2$. □

Proposition 3.25 implies the important fact that there is a *unique maximal* integral curve with initial condition p. Indeed, if $\{\gamma_j \colon I_j \to M\}_{j \in K}$ is the family of all integral curves with initial condition p (for some big index set, K), if we let $I(p) = \bigcup_{j \in K} I_j$, we can define a curve, $\gamma_p \colon I(p) \to M$, so that

$$\gamma_p(t) = \gamma_j(t), \quad \text{if} \quad t \in I_j.$$

Since γ_j and γ_l agree on $I_j \cap I_l$ for all $j, l \in K$, the curve γ_p is indeed well defined and it is clearly an integral curve with initial condition p with the largest possible domain (the open interval, $I(p)$). The curve γ_p is called the *maximal integral curve with initial condition* p and it is also denoted by $\gamma(p, t)$. Note that Proposition 3.25 implies that any two distinct integral curves are disjoint, i.e., do not intersect each other.

Consider the vector field in $\mathbb{R}^2$ given by

$$X = -y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y}.$$

If we write $\gamma(t) = (x(t), y(t))$, the differential equation, $\dot{\gamma}(t) = X(\gamma(t))$, is expressed by

$$\begin{aligned} x'(t) &= -y(t) \\ y'(t) &= x(t), \end{aligned}$$

or, in matrix form,

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

If we write $X = \binom{x}{y}$ and $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, then the above equation is written as

$$X' = AX.$$

Now, as

$$e^{tA} = I + \frac{A}{1!}\, t + \frac{A^2}{2!}\, t^2 + \cdots + \frac{A^n}{n!}\, t^n + \cdots,$$

we get

$$\frac{d}{dt}(e^{tA}) = A + \frac{A^2}{1!}\, t + \frac{A^3}{2!}\, t^2 + \cdots + \frac{A^n}{(n-1)!}\, t^{n-1} + \cdots = Ae^{tA},$$

so we see that $e^{tA}p$ is a solution of the ODE $X' = AX$ with initial condition $X = p$, and by uniqueness, $X = e^{tA}p$ is *the* solution of our ODE starting at $X = p$. Thus, our integral curve, γ_p, through $p = \binom{x_0}{y_0}$ is the circle given by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}.$$

Observe that $I(p) = \mathbb{R}$, for every $p \in \mathbb{R}^2$.

The following interesting question now arises: Given any $p_0 \in M$, if $\gamma_{p_0} \colon I(p_0) \to M$ is the maximal integral curve with initial condition p_0 and, for any $t_1 \in I(p_0)$, if $p_1 = \gamma_{p_0}(t_1) \in M$, then there is a maximal integral curve, $\gamma_{p_1} \colon I(p_1) \to M$, with initial condition p_1; what is the relationship between γ_{p_0} and γ_{p_1}, if any? The answer is given by

Proposition 3.26. *Let X be a C^{k-1} vector field on a C^k-manifold, M, ($k \geq 2$) and let p_0 be a point on M. If $\gamma_{p_0} \colon I(p_0) \to M$ is the maximal integral curve with initial condition p_0, for any $t_1 \in I(p_0)$, if $p_1 = \gamma_{p_0}(t_1) \in M$ and $\gamma_{p_1} \colon I(p_1) \to M$ is the maximal integral curve with initial condition p_1, then*

$$I(p_1) = I(p_0) - t_1 \quad \textit{and} \quad \gamma_{p_1}(t) = \gamma_{\gamma_{p_0}(t_1)}(t) = \gamma_{p_0}(t + t_1), \quad \textit{for all } t \in I(p_0) - t_1.$$

Proof. Let $\gamma(t)$ be the curve given by

$$\gamma(t) = \gamma_{p_0}(t + t_1), \quad \text{for all } \ t \in I(p_0) - t_1.$$

Clearly, γ is defined on $I(p_0) - t_1$ and

$$\dot{\gamma}(t) = \dot{\gamma}_{p_0}(t + t_1) = X(\gamma_{p_0}(t + t_1)) = X(\gamma(t))$$

and $\gamma(0) = \gamma_{p_0}(t_1) = p_1$. Thus, γ is an integal curve defined on $I(p_0) - t_1$ with initial condition p_1. If γ was defined on an interval, $\widetilde{I} \supseteq I(p_0) - t_1$ with $\widetilde{I} \neq I(p_0) - t_1$, then γ_{p_0} would be defined on $\widetilde{I} + t_1 \supset I(p_0)$, an interval strictly bigger than $I(p_0)$, contradicting the maximality of $I(p_0)$. Therefore, $I(p_0) - t_1 = I(p_1)$. □

Proposition 3.26 says that the traces $\gamma_{p_0}(I(p_0))$ and $\gamma_{p_1}(I(p_1))$ in M of the maximal integral curves γ_{p_0} and γ_{p_1} are identical; they only differ by a simple reparametrization $(u = t + t_1)$.

It is useful to restate Proposition 3.26 by changing point of view. So far, we have been focusing on integral curves, i.e., given any $p_0 \in M$, we let t vary in $I(p_0)$ and get an integral curve, γ_{p_0}, with domain $I(p_0)$. Instead of holding $p_0 \in M$ fixed, we can hold $t \in \mathbb{R}$ fixed and consider the set

$$\mathcal{D}_t(X) = \{p \in M \mid t \in I(p)\},$$

i.e., the set of points such that it is possible to "travel for t units of time from p" along the maximal integral curve, γ_p, with initial condition p (It is possible that $\mathcal{D}_t(X) = \emptyset$). By definition, if $\mathcal{D}_t(X) \neq \emptyset$, the point $\gamma_p(t)$ is well defined, and so, we obtain a map,
$\Phi_t^X \colon \mathcal{D}_t(X) \to M$, with domain $\mathcal{D}_t(X)$, given by

$$\Phi_t^X(p) = \gamma_p(t).$$

The above suggests the following definition:

Definition 3.27. Let X be a C^{k-1} vector field on a C^k-manifold, M, $(k \geq 2)$. For any $t \in \mathbb{R}$, let

$$\mathcal{D}_t(X) = \{p \in M \mid t \in I(p)\} \quad \text{and} \quad \mathcal{D}(X) = \{(t, p) \in \mathbb{R} \times M \mid t \in I(p)\}$$

and let $\Phi^X \colon \mathcal{D}(X) \to M$ be the map given by

$$\Phi^X(t, p) = \gamma_p(t).$$

The map Φ^X is called the *(global) flow of* X and $\mathcal{D}(X)$ is called its *domain of definition*. For any $t \in \mathbb{R}$ such that $\mathcal{D}_t(X) \neq \emptyset$, the map, $p \in \mathcal{D}_t(X) \mapsto \Phi^X(t, p) = \gamma_p(t)$, is denoted by Φ_t^X (i.e., $\Phi_t^X(p) = \Phi^X(t, p) = \gamma_p(t)$).

Observe that

$$\mathcal{D}(X) = \bigcup_{p \in M} (I(p) \times \{p\}).$$

Also, using the Φ_t^X notation, the property of Proposition 3.26 reads

$$\Phi_s^X \circ \Phi_t^X = \Phi_{s+t}^X, \tag{$*$}$$

whenever both sides of the equation make sense. Indeed, the above says

$$\Phi_s^X(\Phi_t^X(p)) = \Phi_s^X(\gamma_p(t)) = \gamma_{\gamma_p(t)}(s) = \gamma_p(s+t) = \Phi_{s+t}^X(p).$$

Using the above property, we can easily show that the Φ_t^X are invertible. In fact, the inverse of Φ_t^X is Φ_{-t}^X. First, note that

$$\mathcal{D}_0(X) = M \quad \text{and} \quad \Phi_0^X = \mathrm{id},$$

because, by definition, $\Phi_0^X(p) = \gamma_p(0) = p$, for every $p \in M$. Then, $(*)$ implies that

$$\Phi_t^X \circ \Phi_{-t}^X = \Phi_{t+-t}^X = \Phi_0^X = \mathrm{id},$$

which shows that $\Phi_t^X \colon \mathcal{D}_t(X) \to \mathcal{D}_{-t}(X)$ and $\Phi_{-t}^X \colon \mathcal{D}_{-t}(X) \to \mathcal{D}_t(X)$ are inverse of each other. Moreover, each Φ_t^X is a C^{k-1}-diffeomorphism. We summarize in the following proposition some additional properties of the domains $\mathcal{D}(X)$, $\mathcal{D}_t(X)$ and the maps Φ_t^X (for a proof, see Lang [96] or Warner [148]):

Theorem 3.27. *Let X be a C^{k-1} vector field on a C^k-manifold, M, ($k \geq 2$). The following properties hold:*

(a) For every $t \in \mathbb{R}$, if $\mathcal{D}_t(X) \neq \emptyset$, then $\mathcal{D}_t(X)$ is open (this is trivially true if $\mathcal{D}_t(X) = \emptyset$).

(b) The domain, $\mathcal{D}(X)$, of the flow, Φ^X, is open and the flow is a C^{k-1} map, $\Phi^X \colon \mathcal{D}(X) \to M$.

(c) Each $\Phi_t^X \colon \mathcal{D}_t(X) \to \mathcal{D}_{-t}(X)$ is a C^{k-1}-diffeomorphism with inverse Φ_{-t}^X.

(d) For all $s, t \in \mathbb{R}$, the domain of definition of $\Phi_s^X \circ \Phi_t^X$ is contained but generally not equal to $\mathcal{D}_{s+t}(X)$. However, $\mathrm{dom}(\Phi_s^X \circ \Phi_t^X) = \mathcal{D}_{s+t}(X)$ if s and t have the same sign. Moreover, on $\mathrm{dom}(\Phi_s^X \circ \Phi_t^X)$, we have

$$\Phi_s^X \circ \Phi_t^X = \Phi_{s+t}^X.$$

Remarks:

(1) We may omit the superscript, X, and write Φ instead of Φ^X if no confusion arises.

(2) The reason for using the terminology flow in referring to the map Φ^X can be clarified as follows: For any t such that $\mathcal{D}_t(X) \neq \emptyset$, every integral curve, γ_p, with initial condition $p \in \mathcal{D}_t(X)$, is defined on some open interval containing $[0, t]$, and we can picture these curves as "flow lines" along which the points p flow (travel) for a time interval t. Then, $\Phi^X(t, p)$ is the point reached by "flowing" for the amount of time t on the integral curve γ_p (through p) starting from p. Intuitively, we can imagine the flow of a fluid through M, and the vector field X is the field of velocities of the flowing particles.

Given a vector field, X, as above, it may happen that $\mathcal{D}_t(X) = M$, for all $t \in \mathbb{R}$. In this case, namely, when $\mathcal{D}(X) = \mathbb{R} \times M$, we say that the vector field X is *complete*. Then, the Φ_t^X are diffeomorphisms of M and they form a group. The family $\{\Phi_t^X\}_{t\in\mathbb{R}}$ a called a 1-*parameter group of* X. In this case, Φ^X induces a group homomorphism, $(\mathbb{R}, +) \longrightarrow \mathrm{Diff}(M)$, from the additive group $\mathbb{R}$ to the group of C^{k-1}-diffeomorphisms of M.

By abuse of language, even when it is **not** the case that $\mathcal{D}_t(X) = M$ for all t, the family $\{\Phi_t^X\}_{t\in\mathbb{R}}$ is called a *local* 1-*parameter group generated by* X, even though it is **not** a group, because the composition $\Phi_s^X \circ \Phi_t^X$ may not be defined.

If we go back to the vector field in $\mathbb{R}^2$ given by

$$X = -y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y},$$

since the integral curve, $\gamma_p(t)$, through $p = \binom{x_0}{x_0}$ is given by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix},$$

the global flow associated with X is given by

$$\Phi^X(t, p) = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} p,$$

and each diffeomorphism, Φ_t^X, is the rotation,

$$\Phi_t^X = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}.$$

The 1-parameter group, $\{\Phi_t^X\}_{t\in\mathbb{R}}$, generated by X is the group of rotations in the plane, $\mathbf{SO}(2)$.

More generally, if B is an $n \times n$ invertible matrix that has a real logarithm, A (that is, if $e^A = B$), then the matrix A defines a vector field, X, in $\mathbb{R}^n$, with

$$X = \sum_{i,j=1}^{n} (a_{ij}x_j)\frac{\partial}{\partial x_i},$$

whose integral curves are of the form,

$$\gamma_p(t) = e^{tA}p,$$

and we have

$$\gamma_p(1) = Bp.$$

The one-parameter group, $\{\Phi_t^X\}_{t\in\mathbb{R}}$, generated by X is given by $\{e^{tA}\}_{t\in\mathbb{R}}$.

When M is compact, it turns out that every vector field is complete, a nice and useful fact.

Proposition 3.28. *Let X be a C^{k-1} vector field on a C^k-manifold, M, ($k \geq 2$). If M is compact, then X is complete, i.e., $\mathcal{D}(X) = \mathbb{R} \times M$. Moreover, the map $t \mapsto \Phi_t^X$ is a homomorphism from the additive group $\mathbb{R}$ to the group,* $\mathrm{Diff}(M)$*, of (C^{k-1}) diffeomorphisms of M.*

Proof. Pick any $p \in M$. By Theorem 3.24, there is a local flow, $\varphi_p \colon J(p) \times U(p) \to M$, where $J(p) \subseteq \mathbb{R}$ is an open interval containing 0 and $U(p)$ is an open subset of M containing p, so that for all $q \in U(p)$, the map $t \mapsto \varphi(t, q)$ is an integral curve with initial condition q (where $t \in J(p)$). Thus, we have $J(p) \times U(p) \subseteq \mathcal{D}(X)$. Now, the $U(p)$'s form an open cover of M and since M is compact, we can extract a finite subcover, $\bigcup_{q \in F} U(q) = M$, for some finite subset, $F \subseteq M$. But then, we can find $\epsilon > 0$ so that $]-\epsilon, +\epsilon[\subseteq J(q)$, for all $q \in F$ and for all $t \in]-\epsilon, +\epsilon[$ and, for all $p \in M$, if γ_p is the maximal integral curve with initial condition p, then $]-\epsilon, +\epsilon[\subseteq I(p)$.

For any $t \in]-\epsilon, +\epsilon[$, consider the integral curve, $\gamma_{\gamma_p(t)}$, with initial condition $\gamma_p(t)$. This curve is well defined for all $t \in]-\epsilon, +\epsilon[$, and we have

$$\gamma_{\gamma_p(t)}(t) = \gamma_p(t + t) = \gamma_p(2t),$$

which shows that γ_p is in fact defined for all $t \in]-2\epsilon, +2\epsilon[$. By induction, we see that

$$]-2^n\epsilon, +2^n\epsilon[\subseteq I(p),$$

for all $n \geq 0$, which proves that $I(p) = \mathbb{R}$. As this holds for all $p \in M$, we conclude that $\mathcal{D}(X) = \mathbb{R} \times M$. □

Remarks:

(1) The proof of Proposition 3.28 also applies when X is a vector field with compact support (this means that the closure of the set $\{p \in M \mid X(p) \neq 0\}$ is compact).

(2) If $\varphi \colon M \to N$ is a diffeomorphism and X is a vector field on M, then it can be shown that the local 1-parameter group associated with the vector field, $\varphi_* X$, is

$$\{\varphi \circ \Phi_t^X \circ \varphi^{-1}\}_{t \in \mathbb{R}}.$$

A point $p \in M$ where a vector field vanishes, i.e., $X(p) = 0$, is called a *critical point of* X. Critical points play a major role in the study of vector fields, in differential topology (e.g., the celebrated Poincaré–Hopf index theorem) and especially in Morse theory, but we won't go into this here (curious readers should consult Milnor [107], Guillemin and Pollack [70] or DoCarmo [50], which contains an informal but very clear presentation of the Poincaré–Hopf index theorem). Another famous theorem about vector fields says that every smooth

vector field on a sphere of even dimension (S^{2n}) must vanish in at least one point (the so-called "hairy-ball theorem." On S^2, it says that you can't comb your hair without having a singularity somewhere. Try it, it's true!).

Let us just observe that if an integral curve, γ, passes through a critical point, p, then γ is reduced to the point p, i.e., $\gamma(t) = p$, for all t. Indeed, such a curve is an integral curve with initial condition p. By the uniqueness property, it is the only one. Then, we see that if a maximal integral curve is defined on the whole of $\mathbb{R}$, either it is injective (it has no self-intersection), or it is simply periodic (i.e., there is some $T > 0$ so that $\gamma(t + T) = \gamma(t)$, for all $t \in \mathbb{R}$ and γ is injective on $[0, T[$), or it is reduced to a single point.

We conclude this section with the definition of the Lie derivative of a vector field with respect to another vector field.

Say we have two vector fields X and Y on M. For any $p \in M$, we can flow along the integral curve of X with initial condition p to $\Phi_t(p)$ (for t small enough) and then evaluate Y there, getting $Y(\Phi_t(p))$. Now, this vector belongs to the tangent space $T_{\Phi_t(p)}(M)$, but $Y(p) \in T_p(M)$. So to "compare" $Y(\Phi_t(p))$ and $Y(p)$, we bring back $Y(\Phi_t(p))$ to $T_p(M)$ by applying the tangent map, $d\Phi_{-t}$, at $\Phi_t(p)$, to $Y(\Phi_t(p))$ (Note that to alleviate the notation, we use the slight abuse of notation $d\Phi_{-t}$ instead of $d(\Phi_{-t})_{\Phi_t(p)}$.) Then, we can form the difference $d\Phi_{-t}(Y(\Phi_t(p))) - Y(p)$, divide by t and consider the limit as t goes to 0.

Definition 3.28. Let M be a C^{k+1} manifold. Given any two C^k vector fields, X and Y on M, for every $p \in M$, the *Lie derivative of Y with respect to X at p*, denoted $(L_X\, Y)_p$, is given by

$$(L_X\, Y)_p = \lim_{t \longrightarrow 0} \frac{d\Phi_{-t}(Y(\Phi_t(p))) - Y(p)}{t} = \frac{d}{dt}\left(d\Phi_{-t}(Y(\Phi_t(p)))\right)\Big|_{t=0}.$$

It can be shown that $(L_X\, Y)_p$ is our old friend, the Lie bracket, i.e.,

$$(L_X\, Y)_p = [X, Y]_p.$$

(For a proof, see Warner [148] or O'Neill [120]).

In terms of Definition 3.21, observe that

$$(L_X\, Y)_p = \lim_{t \longrightarrow 0} \frac{\big((\Phi_{-t})_* Y\big)(p) - Y(p)}{t} = \lim_{t \longrightarrow 0} \frac{\big(\Phi_t^* Y\big)(p) - Y(p)}{t} = \frac{d}{dt}\left(\Phi_t^* Y\right)(p)\Big|_{t=0},$$

since $(\Phi_{-t})^{-1} = \Phi_t$.

3.6 Partitions of Unity

To study manifolds, it is often necessary to construct various objects such as functions, vector fields, Riemannian metrics, volume forms, etc., by gluing together items constructed on the domains of charts. Partitions of unity are a crucial technical tool in this gluing process.

The first step is to define "bump functions" (also called plateau functions). For any $r > 0$, we denote by $B(r)$ the open ball

$$B(r) = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + \cdots + x_n^2 < r\},$$

and by $\overline{B(r)} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + \cdots + x_n^2 \leq r\}$, its closure.

Proposition 3.29. *There is a smooth function, $b\colon \mathbb{R}^n \to \mathbb{R}$, so that*

$$b(x) = \begin{cases} 1 & \text{if } x \in \overline{B(1)} \\ 0 & \text{if } x \in \mathbb{R}^n - B(2). \end{cases}$$

Proof. There are many ways to construct such a function. We can proceed as follows: Consider the function, $h\colon \mathbb{R} \to \mathbb{R}$, given by

$$h(x) = \begin{cases} e^{-1/x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0. \end{cases}$$

It is easy to show that h is C^∞ (but **not** analytic!). Then, define $b\colon \mathbb{R}^n \to \mathbb{R}$, by

$$b(x_1, \ldots, x_n) = \frac{h(4 - x_1^2 - \cdots - x_n^2)}{h(4 - x_1^2 - \cdots - x_n^2) + h(x_1^2 + \cdots + x_n^2 - 1)}.$$

It is immediately verified that b satisfies the required conditions. □

Given a topological space, X, for any function, $f\colon X \to \mathbb{R}$, the *support of f*, denoted $\operatorname{supp} f$, is the closed set,

$$\operatorname{supp} f = \overline{\{x \in X \mid f(x) \neq 0\}}.$$

Proposition 3.29 yields the following useful technical result:

Proposition 3.30. *Let M be a smooth manifold. For any open subset, $U \subseteq M$, any $p \in U$ and any smooth function, $f\colon U \to \mathbb{R}$, there exist an open subset, V, with $p \in V$ and a smooth function, $\widetilde{f}\colon M \to \mathbb{R}$, defined on the whole of M, so that $\overline{V}$ is compact,*

$$\overline{V} \subseteq U, \qquad \operatorname{supp} \widetilde{f} \subseteq U$$

and

$$\widetilde{f}(q) = f(q), \qquad \text{for all} \quad q \in \overline{V}.$$

Proof. Using a scaling function, it is easy to find a chart, (W, φ) at p, so that $W \subseteq U$, $B(3) \subseteq \varphi(W)$ and $\varphi(p) = 0$. Let $\widetilde{b} = b \circ \varphi$, where b is the function given by Proposition 3.29. Then, $\widetilde{b}$ is a smooth function on W with support $\varphi^{-1}(\overline{B(2)}) \subseteq W$. We can extend $\widetilde{b}$ outside W, by setting it to be 0 and we get a smooth function on the whole M. If we let $V = \varphi^{-1}(B(1))$, then V is an open subset around p, $\overline{V} = \varphi^{-1}(\overline{B(1)}) \subseteq W$ is compact and, clearly, $\widetilde{b} = 1$ on $\overline{V}$. Therefore, if we set

$$\widetilde{f}(q) = \begin{cases} \widetilde{b}(q) f(q) & \text{if } q \in W \\ 0 & \text{if } q \in M - W, \end{cases}$$

we see that $\widetilde{f}$ satisfies the required properties. □

If X is a (Hausdorff) topological space, a family, $\{U_\alpha\}_{\alpha\in I}$, of subsets U_α of X is a *cover* (or *covering*) of X iff $X = \bigcup_{\alpha\in I} U_\alpha$. A cover, $\{U_\alpha\}_{\alpha\in I}$, such that each U_α is open is an *open cover*. If $\{U_\alpha\}_{\alpha\in I}$ is a cover of X, for any subset, $J \subseteq I$, the subfamily $\{U_\alpha\}_{\alpha\in J}$ is a *subcover* of $\{U_\alpha\}_{\alpha\in I}$ if $X = \bigcup_{\alpha\in J} U_\alpha$, i.e., $\{U_\alpha\}_{\alpha\in J}$ is still a cover of X. Given a cover $\{V_\beta\}_{\beta\in J}$, we say that a family $\{U_\alpha\}_{\alpha\in I}$ is a *refinement* of $\{V_\beta\}_{\beta\in J}$ if it is a cover and if there is a function, $h\colon I \to J$, so that $U_\alpha \subseteq V_{h(\alpha)}$, for all $\alpha \in I$.

A family $\{U_\alpha\}_{\alpha\in I}$ of subsets of X is *locally finite* iff for every point, $p \in X$, there is some open subset, U, with $p \in U$, so that $U \cap U_\alpha \neq \emptyset$ for only finitely many $\alpha \in I$. A space, X, is *paracompact* iff every open cover has an open locally finite refinement.

Remark: Recall that a space, X, is *compact* iff it is Hausdorff and if every open cover has a *finite* subcover. Thus, the notion of paracompactess (due to Jean Dieudonné) is a generalization of the notion of compactness.

Recall that a topological space, X, is *second-countable* if it has a countable basis, i.e., if there is a countable family of open subsets, $\{U_i\}_{i\geq 1}$, so that every open subset of X is the union of some of the U_i's. A topological space, X, if *locally compact* iff it is Hausdorff and for every $a \in X$, there is some compact subset, K, and some open subset, U, with $a \in U$ and $U \subseteq K$. As we will see shortly, every locally compact and second-countable topological space is paracompact.

It is important to observe that every manifold (even not second-countable) is locally compact. Indeed, for every $p \in M$, if we pick a chart, (U, φ), around p, then $\varphi(U) = \Omega$ for some open $\Omega \subseteq \mathbb{R}^n$ ($n = \dim M$). So, we can pick a small closed ball, $\overline{B(q,\epsilon)} \subseteq \Omega$, of center $q = \varphi(p)$ and radius ϵ, and as φ is a homeomorphism, we see that

$$p \in \varphi^{-1}(B(q, \epsilon/2)) \subseteq \varphi^{-1}(\overline{B(q,\epsilon)}),$$

where $\varphi^{-1}(\overline{B(q,\epsilon)})$ is compact and $\varphi^{-1}(B(q,\epsilon/2))$ is open.

Finally, we define partitions of unity.

Definition 3.29. Let M be a (smooth) manifold. A *partition of unity on M* is a family, $\{f_i\}_{i\in I}$, of smooth functions on M (the index set I may be uncountable) such that

(a) The family of supports, $\{\operatorname{supp} f_i\}_{i\in I}$, is locally finite.

(b) For all $i \in I$ and all $p \in M$, we have $0 \leq f_i(p) \leq 1$, and

$$\sum_{i\in I} f_i(p) = 1, \quad \text{for every } p \in M.$$

Note that condition (b) implies that for every $p \in M$ there must be some $i \in I$ such that $f_i(p) > 0$. Thus, $\{\operatorname{supp} f_i\}_{i\in I}$ is a cover of M. If $\{U_\alpha\}_{\alpha\in J}$ is a cover of M, we say that the partition of unity $\{f_i\}_{i\in I}$ is *subordinate* to the cover $\{U_\alpha\}_{\alpha\in J}$ if $\{\operatorname{supp} f_i\}_{i\in I}$ is a refinement of $\{U_\alpha\}_{\alpha\in J}$. When $I = J$ and $\operatorname{supp} f_i \subseteq U_i$, we say that $\{f_i\}_{i\in I}$ is *subordinate* to $\{U_\alpha\}_{\alpha\in I}$ *with the same index set as the partition of unity*.

In Definition 3.29, by (a), for every $p \in M$, there is some open set, U, with $p \in U$ and U meets only finitely many of the supports, $\operatorname{supp} f_i$. So, $f_i(p) \neq 0$ for only finitely many $i \in I$ and the infinite sum $\sum_{i \in I} f_i(p)$ is well defined.

Proposition 3.31. *Let X be a topological space which is second-countable and locally compact (thus, also Hausdorff). Then, X is paracompact. Moreover, every open cover has a countable, locally finite refinement consisting of open sets with compact closures.*

Proof. The proof is quite technical, but since this is an important result, we reproduce Warner's proof for the reader's convenience (Warner [148], Lemma 1.9).

The first step is to construct a sequence of open sets, G_i, such that

1. $\overline{G}_i$ is compact,
2. $\overline{G}_i \subseteq G_{i+1}$,
3. $X = \bigcup_{i=1}^{\infty} G_i$.

As M is second-countable, there is a countable basis of open sets, $\{U_i\}_{i \geq 1}$, for M. Since M is locally compact, we can find a subfamily of $\{U_i\}_{i \geq 1}$ consisting of open sets with compact closures such that this subfamily is also a basis of M. Therefore, we may assume that we start with a countable basis, $\{U_i\}_{i \geq 1}$, of open sets with compact closures. Set $G_1 = U_1$ and assume inductively that

$$G_k = U_1 \cup \cdots \cup U_{j_k}.$$

Since $\overline{G}_k$ is compact, it is covered by finitely many of the U_j's. So, let j_{k+1} be the smallest integer greater than j_k so that

$$\overline{G}_k \subseteq U_1 \cup \cdots \cup U_{j_{k+1}}$$

and set

$$G_{k+1} = U_1 \cup \cdots \cup U_{j_{k+1}}.$$

Obviously, the family $\{G_i\}_{i \geq 1}$ satisfies (1)–(3).

Now, let $\{U_\alpha\}_{\alpha \in I}$ be an arbitrary open cover of M. For any $i \geq 3$, the set $\overline{G}_i - G_{i-1}$ is compact and contained in the open $G_{i+1} - \overline{G}_{i-2}$. For each $i \geq 3$, choose a finite subcover of the open cover $\{U_\alpha \cap (G_{i+1} - \overline{G}_{i-2})\}_{\alpha \in I}$ of $\overline{G}_i - G_{i-1}$, and choose a finite subcover of the open cover $\{U_\alpha \cap G_3\}_{\alpha \in I}$ of the compact set $\overline{G}_2$. We leave it to the reader to check that this family of open sets is indeed a countable, locally finite refinement of the original open cover $\{U_\alpha\}_{\alpha \in I}$ and consists of open sets with compact closures. □

Remarks:

1. Proposition 3.31 implies that a second-countable, locally compact (Hausdorff) topological space is the union of countably many compact subsets. Thus, X is *countable at infinity*, a notion that we already encountered in Proposition 2.23 and Theorem 2.26.

The reason for this odd terminology is that in the Alexandroff one-point compactification of X, the family of open subsets containing the point at infinity (ω) has a countable basis of open sets. (The open subsets containing ω are of the form $(M - K) \cup \{\omega\}$, where K is compact.)

2. A manifold that is countable at infinity has a countable open cover by domains of charts. This is because, if $M = \bigcup_{i\geq 1} K_i$, where the $K_i \subseteq M$ are compact, then for any open cover of M by domains of charts, for every K_i, we can extract a finite subcover, and the union of these finite subcovers is a countable open cover of M by domains of charts. But then, since for every chart, (U_i, φ_i), the map φ_i is a homeomorphism onto some open subset of $\mathbb{R}^n$, which is second-countable, so we deduce easily that M is second-countable. Thus, for manifolds, second-countable is equivalent to countable at infinity.

We can now prove the main theorem stating the existence of partitions of unity. Recall that we are assuming that our manifolds are Hausdorff and second-countable.

Theorem 3.32. *Let M be a smooth manifold and let $\{U_\alpha\}_{\alpha\in I}$ be an open cover for M. Then, there is a countable partition of unity, $\{f_i\}_{i\geq 1}$, subordinate to the cover $\{U_\alpha\}_{\alpha\in I}$ and the support,* $\operatorname{supp} f_i$*, of each f_i is compact. If one does not require compact supports, then there is a partition of unity, $\{f_\alpha\}_{\alpha\in I}$, subordinate to the cover $\{U_\alpha\}_{\alpha\in I}$ with at most countably many of the f_α not identically zero. (In the second case,* $\operatorname{supp} f_\alpha \subseteq U_\alpha$*.)*

Proof. Again, we reproduce Warner's proof (Warner [148], Theorem 1.11). As our manifolds are second-countable, Hausdorff and locally compact, from the proof of Proposition 3.31, we have the sequence of open subsets, $\{G_i\}_{i\geq 1}$ and we set $G_0 = \emptyset$. For any $p \in M$, let i_p be the largest integer such that $p \in M - \overline{G}_{i_p}$. Choose an α_p such that $p \in U_{\alpha_p}$; we can find a chart, (U, φ), centered at p such that $U \subseteq U_{\alpha_p} \cap (G_{i_p+2} - \overline{G}_{i_p})$ and such that $\overline{B(2)} \subseteq \varphi(U)$. Define

$$\psi_p = \begin{cases} b \circ \varphi & \text{on } U \\ 0 & \text{on } M - U, \end{cases}$$

where b is the bump function defined just before Proposition 3.29. Then, ψ_p is a smooth function on M which has value 1 on some open subset, W_p, containing p and has compact support lying in $U \subseteq U_{\alpha_p} \cap (G_{i_p+2} - \overline{G}_{i_p})$. For each $i \geq 1$, choose a finite set of points, $p \in M$, whose corresponding opens, W_p, cover $\overline{G}_i - G_{i-1}$. Order the corresponding ψ_p functions in a sequence, ψ_j, $j = 1, 2, \ldots$. The supports of the ψ_j form a locally finite family of subsets of M. Thus, the function

$$\psi = \sum_{j=1}^{\infty} \psi_j$$

is well-defined on M and smooth. Moreover, $\psi(p) > 0$ for each $p \in M$. For each $i \geq 1$, set

$$f_i = \frac{\psi_i}{\psi}.$$

Then, the family, $\{f_i\}_{i\geq 1}$, is a partition of unity subordinate to the cover $\{U_\alpha\}_{\alpha\in I}$ and supp f_i is compact for all $i \geq 1$.

Now, when we don't require compact support, if we let f_α be identically zero if no f_i has support in U_α and otherwise let f_α be the sum of the f_i with support in U_α, then we obtain a partition of unity subordinate to $\{U_\alpha\}_{\alpha\in I}$ with at most countably many of the f_α not identically zero. We must have $\operatorname{supp} f_\alpha \subseteq U_\alpha$ because for any locally finite family of closed sets, $\{F_\beta\}_{\beta\in J}$, we have $\overline{\bigcup_{\beta\in J} F_\beta} = \bigcup_{\beta\in J} F_\beta$. □

We close this section by stating a famous theorem of Whitney whose proof uses partitions of unity.

Theorem 3.33. *(Whitney, 1935) Any smooth manifold (Hausdorff and second-countable), M, of dimension n is diffeomorphic to a closed submanifold of $\mathbb{R}^{2n+1}$.*

For a proof, see Hirsch [77], Chapter 2, Section 2, Theorem 2.14.

3.7 Manifolds With Boundary

Up to now, we have defined manifolds locally diffeomorphic to an open subset of $\mathbb{R}^m$. This excludes many natural spaces such as a closed disk, whose boundary is a circle, a closed ball, $\overline{B(1)}$, whose boundary is the sphere, S^{m-1}, a compact cylinder, $S^1 \times [0,1]$, whose boundary consist of two circles, a Möbius strip, etc. These spaces fail to be manifolds because they have a boundary, that is, neighborhoods of points on their boundaries are not diffeomorphic to open sets in $\mathbb{R}^m$. Perhaps the simplest example is the (closed) upper half space,

$$\mathbb{H}^m = \{(x_1, \ldots, x_m) \in \mathbb{R}^m \mid x_m \geq 0\}.$$

Under the natural embedding $\mathbb{R}^{m-1} \cong \mathbb{R}^{m-1} \times \{0\} \hookrightarrow \mathbb{R}^m$, the subset $\partial\mathbb{H}^m$ of $\mathbb{H}^m$ defined by

$$\partial\mathbb{H}^m = \{x \in \mathbb{H}^m \mid x_m = 0\}$$

is isomorphic to $\mathbb{R}^{m-1}$ and is called the *boundary of* $\mathbb{H}^m$. We also define the *interior* of $\mathbb{H}^m$ as

$$\operatorname{Int}(\mathbb{H}^m) = \mathbb{H}^m - \partial\mathbb{H}^m.$$

Now, if U and V are open subsets of $\mathbb{H}^m$, where $\mathbb{H}^m \subseteq \mathbb{R}^m$ has the subset topology, and if $f\colon U \to V$ is a continuous function, we need to explain what we mean by f being smooth. We say that $f\colon U \to V$, as above, is *smooth* if it has an extension, $\widetilde{f}\colon \widetilde{U} \to \widetilde{V}$, where $\widetilde{U}$ and $\widetilde{V}$ are open subsets of $\mathbb{R}^m$ with $U \subseteq \widetilde{U}$ and $V \subseteq \widetilde{V}$ and with $\widetilde{f}$ a smooth function. We say that f is a (smooth) *diffeomorphism* iff f^{-1} exists and if both f and f^{-1} are smooth, as just defined.

To define a *manifold with boundary*, we replace everywhere $\mathbb{R}$ by $\mathbb{H}$ in Definition 3.1 and Definition 3.2. So, for instance, given a topological space, M, a *chart* is now pair, (U, φ), where U is an open subset of M and $\varphi\colon U \to \Omega$ is a homeomorphism onto an open subset, $\Omega = \varphi(U)$, of $\mathbb{H}^{n_\varphi}$ (for some $n_\varphi \geq 1$), etc. Thus, we obtain

Definition 3.30. Given some integer $n \geq 1$ and given some k such that k is either an integer $k \geq 1$ or $k = \infty$, a *C^k-manifold of dimension n with boundary* consists of a topological space, M, together with an equivalence class, $\overline{\mathcal{A}}$, of C^k n-atlases, on M (where the charts are now defined in terms of open subsets of $\mathbb{H}^n$). Any atlas, $\mathcal{A}$, in the equivalence class $\overline{\mathcal{A}}$ is called a *differentiable structure of class C^k (and dimension n) on M*. We say that M is *modeled on* $\mathbb{H}^n$. When $k = \infty$, we say that M is a *smooth manifold with boundary*.

It remains to define what is the boundary of a manifold with boundary! By definition, the *boundary*, ∂M, of a manifold (with boundary), M, is the set of all points, $p \in M$, such that there is some chart, $(U_\alpha, \varphi_\alpha)$, with $p \in U_\alpha$ and $\varphi_\alpha(p) \in \partial\mathbb{H}^n$. We also let $\mathrm{Int}(M) = M - \partial M$ and call it the *interior* of M.

Do not confuse the boundary ∂M and the interior $\mathrm{Int}(M)$ of a manifold with boundary embedded in $\mathbb{R}^N$ with the topological notions of boundary and interior of M as a topological space. In general, they are different.

Note that manifolds as defined earlier (In Definition 3.3) are also manifolds with boundary: their boundary is just empty. We shall still reserve the word "manifold" for these, but for emphasis, we will sometimes call them "boundaryless."

The definition of tangent spaces, tangent maps, etc., are easily extended to manifolds with boundary. The reader should note that if M is a manifold with boundary of dimension n, the tangent space, T_pM, is defined for all $p \in M$ and has dimension n, *even* for boundary points, $p \in \partial M$. The only notion that requires more care is that of a submanifold. For more on this, see Hirsch [77], Chapter 1, Section 4. One should also beware that the product of two manifolds with boundary is generally **not** a manifold with boundary (consider the product $[0,1] \times [0,1]$ of two line segments). There is a generalization of the notion of a manifold with boundary called *manifold with corners* and such manifolds are closed under products (see Hirsch [77], Chapter 1, Section 4, Exercise 12).

If M is a manifold with boundary, we see that $\mathrm{Int}(M)$ is a manifold without boundary. What about ∂M? Interestingly, the boundary, ∂M, of a manifold with boundary, M, of dimension n, is a manifold of dimension $n-1$. For this, we need the following Proposition:

Proposition 3.34. *If M is a manifold with boundary of dimension n, for any $p \in \partial M$ on the boundary on M, for any chart, (U, φ), with $p \in M$, we have $\varphi(p) \in \partial\mathbb{H}^n$.*

Proof. Since $p \in \partial M$, by definition, there is some chart, (V, ψ), with $p \in V$ and $\psi(p) \in \partial\mathbb{H}^n$. Let (U, φ) be any other chart, with $p \in M$ and assume that $q = \varphi(p) \in \mathrm{Int}(\mathbb{H}^n)$. The transition map, $\psi \circ \varphi^{-1} \colon \varphi(U \cap V) \to \psi(U \cap V)$, is a diffeomorphism and $q = \varphi(p) \in \mathrm{Int}(\mathbb{H}^n)$. By the inverse function theorem, there is some open, $W \subseteq \varphi(U \cap V) \cap \mathrm{Int}(\mathbb{H}^n) \subseteq \mathbb{R}^n$, with $q \in W$, so that $\psi \circ \varphi^{-1}$ maps W homeomorphically onto some subset, Ω, open in $\mathrm{Int}(\mathbb{H}^n)$, with $\psi(p) \in \Omega$, contradicting the hypothesis, $\psi(p) \in \partial\mathbb{H}^n$. □

Using Proposition 3.34, we immediately derive the fact that ∂M is a manifold of dimension $n-1$. We obtain charts on ∂M by considering the charts $(U \cap \partial M, L \circ \varphi)$, where (U, φ)

is a chart on M such that $U \cap \partial M = \varphi^{-1}(\partial \mathbb{H}^n) \neq \emptyset$ and $L\colon \partial \mathbb{H}^n \to \mathbb{R}^{n-1}$ is the natural isomorphism (see see Hirsch [77], Chapter 1, Section 4).

3.8 Orientation of Manifolds

Although the notion of orientation of a manifold is quite intuitive it is technically rather subtle. We restrict our discussion to smooth manifolds (although the notion of orientation can also be defined for topological manifolds but more work is involved).

Intuitively, a manifold, M, is orientable if it is possible to give a consistent orientation to its tangent space, T_pM, at every point, $p \in M$. So, if we go around a closed curve starting at $p \in M$, when we come back to p, the orientation of T_pM should be the same as when we started. For exampe, if we travel on a Möbius strip (a manifold with boundary) dragging a coin with us, we will come back to our point of departure with the coin flipped. Try it!

To be rigorous, we have to say what it means to orient T_pM (a vector space) and what consistency of orientation means. We begin by quickly reviewing the notion of orientation of a vector space. Let E be a vector space of dimension n. If $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ are two bases of E, a basic and crucial fact of linear algebra says that there is a unique linear map, g, mapping each u_i to the corresponding v_i (i.e., $g(u_i) = v_i$, $i = 1, \ldots, n$). Then, look at the determinant, $\det(g)$, of this map. We know that $\det(g) = \det(P)$, where P is the matrix whose j-th columns consist of the coordinates of v_j over the basis $u_1, \ldots, u_n$. Either $\det(g)$ is negative or it is positive. Thus, we define an equivalence relation on bases by saying that two bases have the *same orientation* iff the determinant of the linear map sending the first basis to the second has positive determinant. An *orientation* of E is the choice of one of the two equivalence classes, which amounts to picking some basis as an orientation frame.

The above definition is perfectly fine but it turns out that it is more convenient, in the long term, to use a definition of orientation in terms of alternate multi-linear maps (in particular, to define the notion of integration on a manifold). Recall that a function, $h\colon E^k \to \mathbb{R}$, is *alternate multi-linear* (or *alternate k-linear*) iff it is linear in each of its arguments (holding the others fixed) and if

$$h(\ldots, x, \ldots, x, \ldots) = 0,$$

that is, h vanishes whenever two of its arguments are identical. Using multi-linearity, we immediately deduce that h vanishes for all k-tuples of arguments, $u_1, \ldots, u_k$, that are linearly dependent and that h is *skew-symmetric*, i.e.,

$$h(\ldots, y, \ldots, x, \ldots) = -h(\ldots, x, \ldots, y, \ldots).$$

In particular, for $k = n$, it is easy to see that if $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ are two bases, then

$$h(v_1, \ldots, v_n) = \det(g) h(u_1, \ldots, u_n),$$

where g is the unique linear map sending each u_i to v_i. This shows that any alternating n-linear function is a multiple of the determinant function and that the space of alternating

n-linear maps is a one-dimensional vector space that we will denote $\bigwedge^n E^*$.[2] We also call an alternating n-linear map an *n-form*. But then, observe that two bases $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ have the same orientation iff

$$\omega(u_1, \ldots, u_n) \quad \text{and} \quad \omega(v_1, \ldots, v_n) \quad \text{have the same sign for all } \omega \in \bigwedge^n E^* - \{0\}$$

(where 0 denotes the zero n-form). As $\bigwedge^n E^*$ is one-dimensional, picking an orientation of E is equivalent to picking a generator (a one-element basis), ω, of $\bigwedge^n E^*$, and to say that $u_1, \ldots, u_n$ has positive orientation iff $\omega(u_1, \ldots, u_n) > 0$.

Given an orientation (say, given by $\omega \in \bigwedge^n E^*$) of E, a linear map, $f\colon E \to E$, is *orientation preserving* iff $\omega(f(u_1), \ldots, f(u_n)) > 0$ whenever $\omega(u_1, \ldots, u_n) > 0$ (or equivalently, iff $\det(f) > 0$).

Now, to define the orientation of an n-dimensional manifold, M, we use charts. Given any $p \in M$, for any chart, (U, φ), at p, the tangent map, $d\varphi^{-1}_{\varphi(p)}\colon \mathbb{R}^n \to T_pM$ makes sense. If $(e_1, \ldots, e_n)$ is the standard basis of $\mathbb{R}^n$, as it gives an orientation to $\mathbb{R}^n$, we can orient T_pM by giving it the orientation induced by the basis $d\varphi^{-1}_{\varphi(p)}(e_1), \ldots, d\varphi^{-1}_{\varphi(p)}(e_n)$. Then, the consistency of orientations of the T_pM's is given by the overlapping of charts. We require that the Jacobian determinants of all $\varphi_j \circ \varphi_i^{-1}$ have the same sign, whenever (U_i, φ_i) and (U_j, φ_j) are any two overlapping charts. Thus, we are led to the definition below. All definitions and results stated in the rest of this section apply to manifolds with or without boundary.

Definition 3.31. Given a smooth manifold, M, of dimension n, an *orientation atlas* of M is any atlas so that the transition maps, $\varphi_i^j = \varphi_j \circ \varphi_i^{-1}$, (from $\varphi_i(U_i \cap U_j)$ to $\varphi_j(U_i \cap U_j)$) all have a positive Jacobian determinant for every point in $\varphi_i(U_i \cap U_j)$. A manifold is *orientable* iff its has some orientation atlas.

Definition 3.31 can be hard to check in practice and there is an equivalent criterion is terms of n-forms which is often more convenient. The idea is that a manifold of dimension n is orientable iff there is a map, $p \mapsto \omega_p$, assigning to every point, $p \in M$, a nonzero n-form, $\omega_p \in \bigwedge^n T_p^*M$, so that this map is smooth. In order to explain rigorously what it means for such a map to be smooth, we can define the *exterior n-bundle*, $\bigwedge^n T^*M$ (also denoted $\bigwedge^*_n M$) in much the same way that we defined the bundles TM and T^*M. There is an obvious smooth projection map, $\pi\colon \bigwedge^n T^*M \to M$. Then, leaving the details of the fact that $\bigwedge^n T^*M$ can be made into a smooth manifold (of dimension n) as an exercise, a smooth map, $p \mapsto \omega_p$, is simply a smooth section of the bundle $\bigwedge^n T^*M$, i.e., a smooth map, $\omega\colon M \to \bigwedge^n T^*M$, so that $\pi \circ \omega = \mathrm{id}$.

[2] We are using the wedge product notation of exterior calculus even though we have not defined alternating tensors and the wedge product yet. This is standard notation and we hope that the reader will not be confused. In fact, in finite dimension, the space of alternating n-linear maps and $\bigwedge^n E^*$ are isomorphic. A thorough treatment of tensor algebra, including exterior algebra, and of differential forms, will be given in Chapters 22 and 8.

Definition 3.32. If M is an n-dimensional manifold, a smooth section, $\omega \in \Gamma(M, \bigwedge^n T^*M)$, is called a (smooth) *$n$-form*. The set of n-forms, $\Gamma(M, \bigwedge^n T^*M)$, is also denoted $\mathcal{A}^n(M)$. An n-form, ω, is a *nowhere-vanishing n-form on M* or *volume form on M* iff ω_p is a nonzero form for every $p \in M$. This is equivalent to saying that $\omega_p(u_1, \ldots, u_n) \neq 0$, for all $p \in M$ and all bases, $u_1, \ldots, u_n$, of T_pM.

The determinant function, $(u_1, \ldots, u_n) \mapsto \det(u_1, \ldots, u_n)$, where the u_i are expressed over the canonical basis $(e_1, \ldots, e_n)$ of $\mathbb{R}^n$, is a volume form on $\mathbb{R}^n$. We will denote this volume form by ω_0. Another standard notation is $dx_1 \wedge \cdots \wedge dx_n$, but this notation may be very puzzling for readers not familiar with exterior algebra. Observe the justification for the term volume form: the quantity $\det(u_1, \ldots, u_n)$ is indeed the (signed) volume of the parallelepiped

$$\{\lambda_1 u_1 + \cdots + \lambda_n u_n \mid 0 \leq \lambda_i \leq 1,\ 1 \leq i \leq n\}.$$

A volume form on the sphere $S^n \subseteq \mathbb{R}^{n+1}$ is obtained as follows:

$$\omega_p(u_1, \ldots u_n) = \det(p, u_1, \ldots u_n),$$

where $p \in S^n$ and $u_1, \ldots u_n \in T_pS^n$. As the u_i are orthogonal to p, this is indeed a volume form.

Observe that if f is a smooth function on M and ω is any n-form, then $f\omega$ is also an n-form.

Definition 3.33. Let $\varphi\colon M \to N$ be a smooth map of manifolds of the same dimension, n, and let $\omega \in \mathcal{A}^n(N)$ be an n-form on N. The *pull-back, $\varphi^*\omega$, of ω to M* is the n-form on M given by

$$\varphi^*\omega_p(u_1, \ldots, u_n) = \omega_{\varphi(p)}(d\varphi_p(u_1), \ldots, d\varphi_p(u_n)),$$

for all $p \in M$ and all $u_1, \ldots, u_n \in T_pM$.

One checks immediately that $\varphi^*\omega$ is indeed an n-form on M. More interesting is the following Proposition:

Proposition 3.35. *(a) If $\varphi\colon M \to N$ is a local diffeomorphism of manifolds, where $\dim M = \dim N = n$, and $\omega \in \mathcal{A}^n(N)$ is a volume form on N, then $\varphi^*\omega$ is a volume form on M. (b) Assume M has a volume form, ω. Then, for every n-form, $\eta \in \mathcal{A}^n(M)$, there is a unique smooth function, $f \in C^\infty(M)$, so that $\eta = f\omega$. If η is a volume form, then $f(p) \neq 0$ for all $p \in M$.*

Proof. (a) By definition,

$$\varphi^*\omega_p(u_1, \ldots, u_n) = \omega_{\varphi(p)}(d\varphi_p(u_1), \ldots, d\varphi_p(u_n)),$$

for all $p \in M$ and all $u_1, \ldots, u_n \in T_pM$. As φ is a local diffeomorphism, $d_p\varphi$ is a bijection for every p. Thus, if $u_1, \ldots, u_n$ is a basis, then so is $d\varphi_p(u_1), \ldots, d\varphi_p(u_n)$, and as ω is nonzero at every point for every basis, $\varphi^*\omega_p(u_1, \ldots, u_n) \neq 0$.

(b) Pick any $p \in M$ and let (U, φ) be any chart at p. As φ is a diffeomorphism, by (a), we see that ${\varphi^{-1}}^*\omega$ is a volume form on $\varphi(U)$. But then, it is easy to see that ${\varphi^{-1}}^*\eta = g{\varphi^{-1}}^*\omega$, for some unique smooth function, g, on $\varphi(U)$ and so, $\eta = f_U\omega$, for some unique smooth function, f_U, on U. For any two overlapping charts, (U_i, φ_i) and (U_j, φ_j), for every $p \in U_i \cap U_j$, for every basis $u_1, \ldots, u_n$ of T_pM, we have

$$\eta_p(u_1, \ldots, u_n) = f_i(p)\omega_p(u_1, \ldots, u_n) = f_j(p)\omega_p(u_1, \ldots, u_n),$$

and as $\omega_p(u_1, \ldots, u_n) \neq 0$, we deduce that f_i and f_j agree on $U_i \cap U_j$. But, then the f_i's patch on the overlaps of the cover, $\{U_i\}$, of M, and so, there is a smooth function, f, defined on the whole of M and such that $f \restriction U_i = f_i$. As the f_i's are unique, so is f. If η is a volume form, then η_p does not vanish for all $p \in M$ and since ω_p is also a volume form, ω_p does not vanish for all $p \in M$, so $f(p) \neq 0$ for all $p \in M$. □

Remark: If φ and ψ are smooth maps of manifolds, it is easy to prove that

$$(\varphi \circ \psi)^* = \psi^* \circ \varphi^*$$

and that

$$\varphi^*(f\omega) = (f \circ \varphi)\varphi^*\omega,$$

where f is any smooth function on M and ω is any n-form.

The connection between Definition 3.31 and volume forms is given by the following important theorem whose proof contains a wonderful use of partitions of unity.

Theorem 3.36. *A smooth manifold (Hausdorff and second-countable) is orientable iff it possesses a volume form.*

Proof. First, assume that a volume form, ω, exists on M, and say $n = \dim M$. For any atlas, $\{(U_i, \varphi_i)\}_i$, of M, by Proposition 3.35, each n-form, ${\varphi_i^{-1}}^*\omega$, is a volume form on $\varphi_i(U_i) \subseteq \mathbb{R}^n$ and

$${\varphi_i^{-1}}^*\omega = f_i\omega_0,$$

for some smooth function, f_i, never zero on $\varphi_i(U_i)$, where ω_0 is a volume form on $\mathbb{R}^n$. By composing φ_i with an orientation-reversing linear map if necessary, we may assume that for this new altlas, $f_i > 0$ on $\varphi_i(U_i)$. We claim that the family $(U_i, \varphi_i)_i$ is an orientation atlas. This is because, on any (nonempty) overlap, $U_i \cap U_j$, as $\omega = \varphi_j^*(f_j\omega_0)$ and $(\varphi_j \circ \varphi_i^{-1})^* = (\varphi_i^{-1})^* \circ \varphi_j^*$, we have

$$(\varphi_j \circ \varphi_i^{-1})^*(f_j\omega_0) = f_i\omega_0,$$

and by the definition of pull-backs, we see that for every $x \in \varphi_i(U_i \cap U_j)$, if we let $y = \varphi_j \circ \varphi_i^{-1}(x)$, then

$$\begin{aligned} f_i(x)(\omega_0)_x(e_1, \ldots, e_n) &= (\varphi_j \circ \varphi_i^{-1})_x^*(f_j\omega_0)(e_1, \ldots, e_n) \\ &= f_j(y)(\omega_0)_y d(\varphi_j \circ \varphi_i^{-1})_x(e_1), \ldots, d(\varphi_j \circ \varphi_i^{-1})_x(e_n)) \\ &= f_j(y)J((\varphi_j \circ \varphi_i^{-1})_x)(\omega_0)_y(e_1, \ldots, e_n), \end{aligned}$$

where $e_1, \ldots, e_n$ is the standard basis of $\mathbb{R}^n$ and $J((\varphi_j \circ \varphi_i^{-1})_x)$ is the Jacobian determinant of $\varphi_j \circ \varphi_i^{-1}$ at x. As both $f_j(y) > 0$ and $f_i(x) > 0$, we have $J((\varphi_j \circ \varphi_i^{-1})_x) > 0$, as desired.

Conversely, assume that $J((\varphi_j \circ \varphi_i^{-1})_x) > 0$, for all $x \in \varphi_i(U_i \cap U_j)$, whenever $U_i \cap U_j \neq \emptyset$. We need to make a volume form on M. For each U_i, let

$$\omega_i = \varphi_i^* \omega_0,$$

where ω_0 is a volume form on $\mathbb{R}^n$. As φ_i is a diffeomorphism, by Proposition 3.35, we see that ω_i is a volume form on U_i. Then, if we apply Theorem 3.32, we can find a partition of unity, $\{f_i\}$, subordinate to the cover $\{U_i\}$, with the same index set. Let,

$$\omega = \sum_i f_i \omega_i.$$

We claim that ω is a volume form on M.

It is clear that ω is an n-form on M. Now, since every $p \in M$ belongs to some U_i, check that on $\varphi_i(U_i)$, we have

$$\varphi_i^{-1*}\omega = \sum_{j \in \text{finite set}} \varphi_i^{-1*}(f_j \omega_j) = \left(\sum_j (f_j \circ \varphi_i^{-1}) J(\varphi_j \circ \varphi_i^{-1}) \right) \omega_0$$

and this sum is strictly positive because the Jacobian determinants are positive and as $\sum_j f_j = 1$ and $f_j \geq 0$, some term must be strictly positive. Therefore, $\varphi_i^{-1*}\omega$ is a volume form on $\varphi_i(U_i)$ and so, $\varphi_i^* \varphi_i^{-1*}\omega = \omega$ is a volume form on U_i. As this holds for all U_i, we conclude that ω is a volume form on M. □

Since we showed that there is a volume form on the sphere, S^n, by Theorem 3.36, the sphere S^n is orientable. It can be shown that the projective spaces, $\mathbb{RP}^n$, are non-orientable iff n is even an thus, orientable iff n is odd. In particular, $\mathbb{RP}^2$ is not orientable. Also, even though M may not be orientable, its tangent bundle, $T(M)$, is always orientable! (Prove it). It is also easy to show that if $f \colon \mathbb{R}^{n+1} \to \mathbb{R}$ is a smooth submersion, then $M = f^{-1}(0)$ is a smooth orientable manifold. Another nice fact is that every Lie group is orientable.

By Proposition 3.35 (b), given any two volume forms, ω_1 and ω_2 on a manifold, M, there is a function, $f \colon M \to \mathbb{R}$, never 0 on M such that $\omega_2 = f\omega_1$. This fact suggests the following definition:

Definition 3.34. Given an orientable manifold, M, two volume forms, ω_1 and ω_2, on M are *equivalent* iff $\omega_2 = f\omega_1$ for some smooth function, $f \colon M \to \mathbb{R}$, such that $f(p) > 0$ for all $p \in M$. An *orientation of M* is the choice of some equivalence class of volume forms on M and an *oriented manifold* is a manifold together with a choice of orientation. If M is a manifold oriented by the volume form, ω, for every $p \in M$, a basis, $(b_1, \ldots, b_n)$ of T_pM is *posively oriented* iff $\omega_p(b_1, \ldots, b_n) > 0$, else it is *negatively oriented* (where $n = \dim(M)$).

If M is an orientable manifold, for any two volume forms ω_1 and ω_2 on M, as $\omega_2 = f\omega_1$ for some function, f, on M which is never zero, f has a constant sign on every connected component of M. Consequently, a connected orientable manifold has two orientations.

We will also need the notion of orientation-preserving diffeomorphism.

Definition 3.35. Let $\varphi\colon M \to N$ be a diffeomorphism of oriented manifolds, M and N, of dimension n and say the orientation on M is given by the volume form ω_1 while the orientation on N is given by the volume form ω_2. We say that φ is *orientation preserving* iff $\varphi^*\omega_2$ determines the same orientation of M as ω_1.

Using Definition 3.35 we can define the notion of a positive atlas.

Definition 3.36. If M is a manifold oriented by the volume form, ω, an atlas for M is *positive* iff for every chart, (U, φ), the diffeomorphism, $\varphi\colon U \to \varphi(U)$, is orientation preserving, where U has the orientation induced by M and $\varphi(U) \subseteq \mathbb{R}^n$ has the orientation induced by the standard orientation on $\mathbb{R}^n$ (with $\dim(M) = n$).

The proof of Theorem 3.36 shows

Proposition 3.37. *If a manifold, M, has an orientation atlas, then there is a uniquely determined orientation on M such that this atlas is positive.*

3.9 Covering Maps and Universal Covering Manifolds

Covering maps are an important technical tool in algebraic topology and more generally in geometry. This brief section only gives some basic definitions and states a few major facts. We apologize for his sketchy nature. Appendix A of O'Neill [120] gives a review of definitions and main results about covering manifolds. Expositions including full details can be found in Hatcher [72], Greenberg [66], Munkres [116], Fulton [57] and Massey [104, 105] (the most extensive).

We begin with covering maps.

Definition 3.37. A map, $\pi\colon M \to N$, between two smooth manifolds is a *covering map* (or *cover*) iff

(1) The map π is smooth and surjective.

(2) For any $q \in N$, there is some open subset, $V \subseteq N$, so that $q \in V$ and

$$\pi^{-1}(V) = \bigcup_{i \in I} U_i,$$

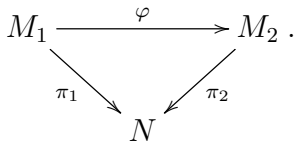
where the U_i are pairwise disjoint open subsets, $U_i \subseteq M$, and $\pi\colon U_i \to V$ is a diffeomorphism for every $i \in I$. We say that V is *evenly covered*.

The manifold, M, is called a *covering manifold* of N.

A *homomorphism* of coverings, $\pi_1\colon M_1 \to N$ and $\pi_2\colon M_2 \to N$, is a smooth map, $\varphi\colon M_1 \to M_2$, so that

$$\pi_1 = \pi_2 \circ \varphi,$$

that is, the following diagram commutes:



We say that the coverings $\pi_1\colon M_1 \to N$ and $\pi_2\colon M_2 \to N$ are *equivalent* iff there is a homomorphism, $\varphi\colon M_1 \to M_2$, between the two coverings and φ is a diffeomorphism.

As usual, the inverse image, $\pi^{-1}(q)$, of any element $q \in N$ is called the *fibre over* q, the space N is called the *base* and M is called the *covering space*. As π is a covering map, each fibre is a discrete space. Note that a homomorphism maps each fibre $\pi_1^{-1}(q)$ in M_1 to the fibre $\pi_2^{-1}(\varphi(q))$ in M_2, for every $q \in M_1$.

Proposition 3.38. *Let $\pi\colon M \to N$ be a covering map. If N is connected, then all fibres, $\pi^{-1}(q)$, have the same cardinality for all $q \in N$. Furthermore, if $\pi^{-1}(q)$ is not finite then it is countably infinite.*

Proof. Pick any point, $p \in N$. We claim that the set

$$S = \{q \in N \mid |\pi^{-1}(q)| = |\pi^{-1}(p)|\}$$

is open and closed.

If $q \in S$, then there is some open subset, V, with $q \in V$, so that $\pi^{-1}(V)$ is evenly covered by some family, $\{U_i\}_{i\in I}$, of disjoint open subsets, U_i, each diffeomorphic to V under π. Then, every $s \in V$ must have a unique preimage in each U_i, so

$$|I| = |\pi^{-1}(s)|, \qquad \text{for all } s \in V.$$

However, as $q \in S$, $|\pi^{-1}(q)| = |\pi^{-1}(p)|$, so

$$|I| = |\pi^{-1}(p)| = |\pi^{-1}(s)|, \qquad \text{for all } s \in V,$$

and thus, $V \subseteq S$. Therefore, S is open. Similary the complement of S is open. As N is connected, $S = N$.

Since M is a manifold, it is second-countable, that is every open subset can be written as some countable union of open subsets. But then, every family, $\{U_i\}_{i\in I}$, of pairwise disjoint open subsets forming an even cover must be countable and since $|I|$ is the common cardinality of all the fibres, every fibre is countable. □

When the common cardinality of fibres is finite it is called the *multiplicity* of the covering (or the number of *sheets*).

For any integer, $n > 0$, the map, $z \mapsto z^n$, from the unit circle $S^1 = \mathbf{U}(1)$ to itself is a covering with n sheets. The map,

$$t\colon \mapsto (\cos(2\pi t), \sin(2\pi t)),$$

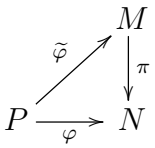
is a covering, $\mathbb{R} \to S^1$, with infinitely many sheets.

It is also useful to note that a covering map, $\pi\colon M \to N$, is a local diffeomorphism (which means that $d\pi_p\colon T_pM \to T_{\pi(p)}N$ is a bijective linear map for every $p \in M$). Indeed, given any $p \in M$, if $q = \pi(p)$, then there is some open subset, $V \subseteq N$, containing q so that V is evenly covered by a family of disjoint open subsets, $\{U_i\}_{i\in I}$, with each $U_i \subseteq M$ diffeomorphic to V under π. As $p \in U_i$ for some i, we have a diffeomorphism, $\pi \restriction U_i\colon U_i \longrightarrow V$, as required.

The crucial property of covering manifolds is that curves in N can be lifted to M, in a unique way. For any map, $\varphi\colon P \to N$, a *lift of φ through π* is a map, $\widetilde{\varphi}\colon P \to M$, so that

$$\varphi = \pi \circ \widetilde{\varphi},$$

as in the following commutative diagram:



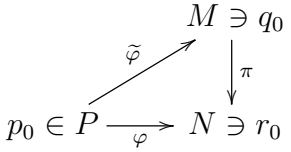
We state without proof the following results:

Proposition 3.39. *If $\pi\colon M \to N$ is a covering map, then for every smooth curve, $\alpha\colon I \to N$, in N (with $0 \in I$) and for any point, $q \in M$, such that $\pi(q) = \alpha(0)$, there is a unique smooth curve, $\widetilde{\alpha}\colon I \to M$, lifting α through π such that $\widetilde{\alpha}(0) = q$.*

Proposition 3.40. *Let $\pi\colon M \to N$ be a covering map and let $\varphi\colon P \to N$ be a smooth map. For any $p_0 \in P$, any $q_0 \in M$ and any $r_0 \in N$ with $\pi(q_0) = \varphi(p_0) = r_0$, the following properties hold:*

(1) If P is connected then there is at most one lift, $\widetilde{\varphi}\colon P \to M$, of φ through π such that $\widetilde{\varphi}(p_0) = q_0$.

(2) If P is simply connected, then such a lift exists.



Theorem 3.41. *Every connected manifold, M, possesses a simply connected covering map, $\pi\colon \widetilde{M} \to M$, that is, with $\widetilde{M}$ simply connected. Any two simply connected coverings of N are equivalent.*

In view of Theorem 3.41, it is legitimate to speak of *the* simply connected cover, $\widetilde{M}$, of M, also called *universal covering* (or *cover*) of M.

Given any point, $p \in M$, let $\pi_1(M, p)$ denote the fundamental group of M with basepoint p (see any of the references listed above, in particular, Massey [104, 105]). If $\varphi\colon M \to N$ is a smooth map, for any $p \in M$, if we write $q = \varphi(p)$, then we have an induced group homomorphism

$$\varphi_*\colon \pi_1(M, p) \to \pi_1(N, q).$$

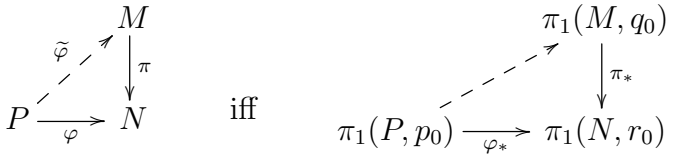
Proposition 3.42. *If $\pi\colon M \to N$ is a covering map, for every $p \in M$, if $q = \pi(p)$, then the induced homomorphism, $\pi_*\colon \pi_1(M, p) \to \pi_1(N, q)$, is injective.*

The next proposition is a stronger version of part (2) of Proposition 3.40:

Proposition 3.43. *Let $\pi\colon M \to N$ be a covering map and let $\varphi\colon P \to N$ be a smooth map. For any $p_0 \in P$, any $q_0 \in M$ and any $r_0 \in N$ with $\pi(q_0) = \varphi(p_0) = r_0$, if P is connected, then a lift, $\widetilde{\varphi}\colon P \to M$, of φ such that $\widetilde{\varphi}(p_0) = q_0$ exists iff*

$$\varphi_*(\pi_1(P, p_0)) \subseteq \pi_*(\pi_1(M, q_0)),$$

as illustrated in the diagram below



Basic Assumption: For any covering, $\pi\colon M \to N$, if N is connected then we also assume that M is connected.

Using Proposition 3.42, we get

Proposition 3.44. *If $\pi\colon M \to N$ is a covering map and N is simply connected, then π is a diffeomorphism (recall that M is connected); thus, M is diffeomorphic to the universal cover, $\widetilde{N}$, of N.*

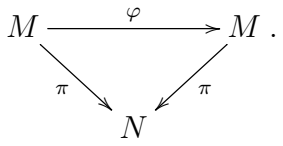
Proof. Pick any $p \in M$ and let $q = \varphi(p)$. As N is simply connected, $\pi_1(N, q) = (0)$. By Proposition 3.42, since $\pi_*\colon \pi_1(M, p) \to \pi_1(N, q)$ is injective, $\pi_1(M, p) = (0)$ so M is simply connected (by hypothesis, M is connected). But then, by Theorem 3.41, M and N are diffeomorphic. □

The following proposition shows that the universal covering of a space covers every other covering of that space. This justifies the terminology "universal covering."

Proposition 3.45. *Say $\pi_1\colon M_1 \to N$ and $\pi_2\colon M_2 \to N$ are two coverings of N, with N connected. Every homomorphism, $\varphi\colon M_1 \to M_2$, between these two coverings is a covering map. As a consequence, if $\pi\colon \widetilde{N} \to N$ is a universal covering of N, then for every covering, $\pi'\colon M \to N$, of N, there is a covering, $\varphi\colon \widetilde{N} \to M$, of M.*

The notion of deck-transformation group of a covering is also useful because it yields a way to compute the fundamental group of the base space.

Definition 3.38. If $\pi\colon M \to N$ is a covering map, a *deck-transformation* is any diffeomorphism, $\varphi\colon M \to M$, such that $\pi = \pi \circ \varphi$, that is, the following diagram commutes:



Note that deck-transformations are just automorphisms of the covering map. The commutative diagram of Definition 3.38 means that a deck transformation permutes every fibre. It is immediately verified that the set of deck transformations of a covering map is a group denoted Γ_π (or simply, Γ), called the *deck-transformation group* of the covering.

Observe that any deck transformation, φ, is a lift of π through π. Consequently, if M is connected, by Proposition 3.40 (1), every deck-transformation is determined by its value at a single point. So, the deck-transformations are determined by their action on each point of any fixed fibre, $\pi^{-1}(q)$, with $q \in N$. Since the fibre $\pi^{-1}(q)$ is countable, Γ is also countable, that is, a discrete Lie group. Moreover, if M is compact, as each fibre, $\pi^{-1}(q)$, is compact and discrete, it must be finite and so, the deck-transformation group is also finite.

The following proposition gives a useful method for determining the fundamental group of a manifold.

Proposition 3.46. *If $\pi\colon \widetilde{M} \to M$ is the universal covering of a connected manifold, M, then the deck-transformation group, $\widetilde{\Gamma}$, is isomorphic to the fundamental group, $\pi_1(M)$, of M.*

Remark: When $\pi\colon \widetilde{M} \to M$ is the universal covering of M, it can be shown that the group $\widetilde{\Gamma}$ acts simply and transitively on every fibre, $\pi^{-1}(q)$. This means that for any two elements, $x, y \in \pi^{-1}(q)$, there is a unique deck-transformation, $\varphi \in \widetilde{\Gamma}$ such that $\varphi(x) = y$. So, there is a bijection between $\pi_1(M) \cong \widetilde{\Gamma}$ and the fibre $\pi^{-1}(q)$.

Proposition 3.41 together with previous observations implies that if the universal cover of a connected (compact) manifold is compact, then M has a finite fundamental group. We will use this fact later, in particular, in the proof of Myers' Theorem.

Chapter 4

Construction of Manifolds From Gluing Data

4.1 Sets of Gluing Data for Manifolds

The definition of a manifold given in Chapter 3 assumes that the underlying set, M, is already known. However, there are situations where we only have some indirect information about the overlap of the domains, U_i, of the local charts defining our manifold, M, in terms of the transition functions,

$$\varphi_{ji}\colon \varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j),$$

but where M itself is not known. For example, this situation happens when trying to construct a surface approximating a 3D-mesh. If we let $\Omega_{ij} = \varphi_i(U_i \cap U_j)$ and $\Omega_{ji} = \varphi_j(U_i \cap U_j)$, then φ_{ji} can be viewed as a "gluing map",

$$\varphi_{ji}\colon \Omega_{ij} \to \Omega_{ji},$$

between two open subets of Ω_i and Ω_j, respectively.

For technical reasons, it is desirable to assume that the images, $\Omega_i = \varphi_i(U_i)$ and $\Omega_j = \varphi_j(U_j)$, of distinct charts are disjoint but this can always be achieved for manifolds. Indeed, the map

$$\beta\colon (x_1, \ldots, x_n) \mapsto \left(\frac{x_1}{\sqrt{1 + \sum_{i=1}^n x_i^2}}, \ldots, \frac{x_n}{\sqrt{1 + \sum_{i=1}^n x_i^2}}\right)$$

is a smooth diffeomorphism from $\mathbb{R}^n$ to the open unit ball $B(0,1)$ with inverse given by

$$\beta^{-1}\colon (x_1, \ldots, x_n) \mapsto \left(\frac{x_1}{\sqrt{1 - \sum_{i=1}^n x_i^2}}, \ldots, \frac{x_n}{\sqrt{1 - \sum_{i=1}^n x_i^2}}\right).$$

Since M has a countable basis, using compositions of β with suitable translations, we can make sure that the Ω_i's are mapped diffeomorphically to disjoint open subsets of $\mathbb{R}^n$.

Remarkably, manifolds can be constructed using the "gluing process" alluded to above from what is often called sets of "gluing data." In this chapter, we are going to describe this construction and prove its correctness in details, provided some mild assumptions on the gluing data. It turns out that this procedure for building manifolds can be made practical. Indeed, it is the basis of a class of new methods for approximating 3D meshes by smooth surfaces, see Siqueira, Xu and Gallier [141].

It turns out that care must be exercised to ensure that the space obtained by gluing the pieces Ω_{ij} and Ω_{ji} is Hausdorff. Some care must also be exercised in formulating the consistency conditions relating the φ_{ji}'s (the so-called "cocycle condition"). This is because the traditional condition (for example, in bundle theory) has to do with triple overlaps of the $U_i = \varphi_i^{-1}(\Omega_i)$ on the manifold, M, (see Chapter 7, especially Theorem 7.4) but in our situation, we do not have M nor the parametrization maps $\theta_i = \varphi_i^{-1}$ and the cocycle condition on the φ_{ji}'s has to be stated in terms of the Ω_i's and the Ω_{ji}'s.

Note that if the Ω_{ij} arise from the charts of a manifold, then nonempty triple intersections $U_i \cap U_j \cap U_k$ of domains of charts have images $\varphi_i(U_i \cap U_j \cap U_k)$ in Ω_i, $\varphi_j(U_i \cap U_j \cap U_k)$ in Ω_j, and $\varphi_k(U_i \cap U_j \cap U_k)$ in Ω_k, and since the φ_i's are bijective maps, we get

$$\varphi_i(U_i \cap U_j \cap U_k) = \varphi_i(U_i \cap U_j \cap U_i \cap U_k) = \varphi_i(U_i \cap U_j) \cap \varphi_i(U_i \cap U_k) = \Omega_{ij} \cap \Omega_{ik},$$

and similarly

$$\varphi_j(U_i \cap U_j \cap U_k) = \Omega_{ji} \cap \Omega_{jk}, \quad \varphi_k(U_i \cap U_j \cap U_k) = \Omega_{ki} \cap \Omega_{kj},$$

and these sets are related. Indeed, we have

$$\begin{aligned} \varphi_{ji}(\Omega_{ij} \cap \Omega_{ik}) &= \varphi_j \circ \varphi_i^{-1}(\varphi_i(U_i \cap U_j) \cap \varphi_i(U_i \cap U_k)) \\ &= \varphi_j(U_i \cap U_j \cap U_k) = \Omega_{ji} \cap \Omega_{jk}, \end{aligned}$$

and similar equations relating the other "triple intersections." In particular,

$$\varphi_{ij}(\Omega_{ji} \cap \Omega_{jk}) = \Omega_{ij} \cap \Omega_{ik},$$

which implies that

$$\varphi_{ji}^{-1}(\Omega_{ji} \cap \Omega_{jk}) = \varphi_{ij}(\Omega_{ji} \cap \Omega_{jk}) \subseteq \Omega_{ik}.$$

This is important, because $\varphi_{ji}^{-1}(\Omega_{ji} \cap \Omega_{jk})$ is the domain of $\varphi_{kj} \circ \varphi_{ji}$ and Ω_{ik} is the domain of φ_{ki}, so the condition $\varphi_{ij}(\Omega_{ji} \cap \Omega_{jk}) = \Omega_{ij} \cap \Omega_{ik}$ implies that the domain of φ_{ki} is a subset of the domain of $\varphi_{kj} \circ \varphi_{ji}$. The definition of gluing data given by Grimm and Hughes [68, 69] misses the above condition.

Finding an easily testable necessary and sufficient criterion for the Hausdorff condition appears to be a very difficult problem. We propose a necessary and sufficient condition, but it is not easily testable in general. If M is a manifold, then observe that difficulties may arise when we want to separate two distinct point, $p, q \in M$, such that p and q neither belong

to the same open, $\theta_i(\Omega_i)$, nor to two disjoint opens, $\theta_i(\Omega_i)$ and $\theta_j(\Omega_j)$, but instead, to the boundary points in $(\partial(\theta_i(\Omega_{ij})) \cap \theta_i(\Omega_i)) \cup (\partial(\theta_j(\Omega_{ji})) \cap \theta_j(\Omega_j))$. In this case, there are some disjoint open subsets, U_p and U_q, of M with $p \in U_p$ and $q \in U_q$, and we get two disjoint open subsets, $V_x = \theta_i^{-1}(U_p) \subseteq \Omega_i$ and $V_y = \theta_j^{-1}(U_q) \subseteq \Omega_j$, with $\theta_i(x) = p$, $\theta_j(y) = q$, and such that $x \in \partial(\Omega_{ij}) \cap \Omega_i$, $y \in \partial(\Omega_{ji}) \cap \Omega_j$, and no point in $V_y \cap \Omega_{ji}$ is the image of any point in $V_x \cap \Omega_{ij}$ by φ_{ji}. Since V_x and V_y are open, we may assume that they are open balls. This necessary condition turns out to be also sufficient.

With the above motivations in mind, here is the definition of sets of gluing data.

Definition 4.1. Let n be an integer with $n \geq 1$ and let k be either an integer with $k \geq 1$ or $k = \infty$. A *set of gluing data* is a triple, $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}, (\varphi_{ji})_{(i,j)\in K})$, satisfying the following properties, where I is a (nonempty) countable set:

(1) For every $i \in I$, the set Ω_i is a nonempty open subset of $\mathbb{R}^n$ called a *parametrization domain*, for short, *p-domain*, and the Ω_i are pairwise disjoint (*i.e.*, $\Omega_i \cap \Omega_j = \emptyset$ for all $i \neq j$).

(2) For every pair $(i, j) \in I \times I$, the set Ω_{ij} is an open subset of Ω_i. Furthermore, $\Omega_{ii} = \Omega_i$ and $\Omega_{ij} \neq \emptyset$ iff $\Omega_{ji} \neq \emptyset$. Each nonempty Ω_{ij} (with $i \neq j$) is called a *gluing domain*.

(3) If we let
$$K = \{(i, j) \in I \times I \mid \Omega_{ij} \neq \emptyset\},$$
then $\varphi_{ji} \colon \Omega_{ij} \to \Omega_{ji}$ is a C^k bijection for every $(i, j) \in K$ called a *transition function* (or *gluing function*) and the following condition holds:

(c) For all i, j, k, if $\Omega_{ji} \cap \Omega_{jk} \neq \emptyset$, then $\varphi_{ij}(\Omega_{ji} \cap \Omega_{jk}) = \Omega_{ij} \cap \Omega_{ik}$, and $\varphi_{ki}(x) = \varphi_{kj} \circ \varphi_{ji}(x)$, for all $x \in \Omega_{ij} \cap \Omega_{ik}$.

Condition (c) is called the *cocycle* condition.

(4) For every pair $(i, j) \in K$, with $i \neq j$, for every $x \in \partial(\Omega_{ij}) \cap \Omega_i$ and every $y \in \partial(\Omega_{ji}) \cap \Omega_j$, there are open balls, V_x and V_y centered at x and y, so that no point of $V_y \cap \Omega_{ji}$ is the image of any point of $V_x \cap \Omega_{ij}$ by φ_{ji}.

Remarks.

(1) In practical applications, the index set, I, is of course finite and the open subsets, Ω_i, may have special properties (for example, connected; open simplicies, *etc.*).

(2) We are only interested in the Ω_{ij}'s that are nonempty but empty Ω_{ij}'s do arise in proofs and constructions and this is why our definition allows them.

(3) Observe that $\Omega_{ij} \subseteq \Omega_i$ and $\Omega_{ji} \subseteq \Omega_j$. If $i \neq j$, as Ω_i and Ω_j are disjoint, so are Ω_{ij} and Ω_{ij}.

(4) The cocycle condition (c) may seem overly complicated but it is actually needed to guarantee the transitivity of the relation, $\sim$, defined in the proof of Proposition 4.1. Flawed versions of condition (c) appear in the literature, see the discussion after the proof of Proposition 4.1. The problem is that $\varphi_{kj} \circ \varphi_{ji}$ is a partial function whose domain, $\varphi_{ji}^{-1}(\Omega_{ji} \cap \Omega_{jk})$, is not necessarily related to the domain, Ω_{ik}, of φ_{ki}. To ensure transitivity of $\sim$, we must assert that whenever the composition $\varphi_{kj} \circ \varphi_{ji}$ has a nonempty domain, this domain is contained in the domain, Ω_{ik}, of φ_{ki}, and that $\varphi_{kj} \circ \varphi_{ji}$ and φ_{ki} agree in $\varphi_{ji}^{-1}(\Omega_{ji} \cap \Omega_{jk})$.

Since the φ_{ji} are bijections, condition (c) implies the following conditions:

(a) $\varphi_{ii} = \mathrm{id}_{\Omega_i}$, for all $i \in I$.

(b) $\varphi_{ij} = \varphi_{ji}^{-1}$, for all $(i, j) \in K$.

To get (a), set $i = j = k$. Then, (b) follows from (a) and (c) by setting $k = i$.

(5) If M is a C^k manifold (including $k = \infty$), then using the notation of our introduction, it is easy to check that the open sets Ω_i, Ω_{ij} and the gluing functions, φ_{ji}, satisfy the conditions of Definition 4.1 (provided that we fix the charts so that the images of distinct charts are disjoint). Proposition 4.1 will show that a manifold can be reconstructed from a set of guing data.

The idea of defining gluing data for manifolds is not new. André Weil introduced this idea to define abstract algebraic varieties by gluing irreducible affine sets in his book [149] published in 1946. The same idea is well-known in bundle theory and can be found in standard texts such as Steenrod [142], Bott and Tu [19], Morita [115] and Wells [151] (the construction of a fibre bundle from a cocycle is given in Chapter 7, see Theorem 7.4).

The beauty of the idea is that it allows the reconstruction of a manifold, M, without having prior knowledge of the topology of this manifold (that is, without having explicitly the underlying topological space M) by gluing open subets of $\mathbb{R}^n$ (the Ω_i's) according to prescribed gluing instructions (namely, glue Ω_i and Ω_j by identifying Ω_{ij} and Ω_{ji} using φ_{ji}). This method of specifying a manifold separates clearly the local structure of the manifold (given by the Ω_i's) from its global structure which is specified by the gluing functions. Furthermore, this method ensures that the resulting manifold is C^k (even for $k = \infty$) with no extra effort since the gluing functions φ_{ji} are assumed to be C^k.

Grimm and Hughes [68, 69] appear to be the first to have realized the power of this latter property for practical applications and we wish to emphasize that this is a very significant discovery. However, Grimm [68] uses a condition stronger than our condition (4) to ensure that the resulting space is Hausdorff. The cocycle condition in Grimm and Hughes [68, 69] is also not strong enough to ensure transitivity of the relation $\sim$. We will come back to these points after the proof of Proposition 4.1.

Working with overlaps of *open subsets* of the parameter domain makes it much easier to enforce smoothness conditions compared to the traditional approach with splines where the

parameter domain is subdivided into *closed* regions and where enforcing smoothness along boundaries is much more difficult.

Let us show that a set of gluing data defines a C^k manifold in a natural way.

Proposition 4.1. *For every set of gluing data, $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}, (\varphi_{ji})_{(i,j)\in K})$, there is an n-dimensional C^k manifold, $M_{\mathcal{G}}$, whose transition functions are the φ_{ji}'s.*

Proof. Define the binary relation, $\sim$, on the disjoint union, $\coprod_{i\in I}\Omega_i$, of the open sets, Ω_i, as follows: For all $x, y \in \coprod_{i\in I}\Omega_i$,

$$x \sim y \quad \text{iff} \quad (\exists (i,j)\in K)(x\in\Omega_{ij}, y\in\Omega_{ji}, y = \varphi_{ji}(x)).$$

Note that if $x \sim y$ and $x \neq y$, then $i \neq j$, as $\varphi_{ii} = \mathrm{id}$. But then, as $x \in \Omega_{ij} \subseteq \Omega_i$, $y \in \Omega_{ji} \subseteq \Omega_j$ and $\Omega_i \cap \Omega_j = \emptyset$ when $i \neq j$, if $x \sim y$ and $x, y \in \Omega_i$, then $x = y$. We claim that $\sim$ is an equivalence relation. This follows easily from the cocycle condition. Clearly, condition 3a of Definition 4.1 ensures reflexivity, while condition 3b ensures symmetry. To check transitivity, assume that $x \sim y$ and $y \sim z$. Then, there are some i, j, k such that (i) $x \in \Omega_{ij}$, $y \in \Omega_{ji} \cap \Omega_{jk}$, $z \in \Omega_{kj}$, and (ii) $y = \varphi_{ji}(x)$ and $z = \varphi_{kj}(y)$. Consequently, $\Omega_{ji} \cap \Omega_{jk} \neq \emptyset$ and $x \in \varphi_{ji}^{-1}(\Omega_{ji} \cap \Omega_{jk})$, so by 3c, we get $\varphi_{ji}^{-1}(\Omega_{ji} \cap \Omega_{jk}) = \Omega_{ij} \cap \Omega_{ik} \subseteq \Omega_{ik}$. So, $\varphi_{ki}(x)$ is defined and by 3c again, $\varphi_{ki}(x) = \varphi_{kj} \circ \varphi_{ji}(x) = z$, *i.e.*, $x \sim z$, as desired.

Since $\sim$ is an equivalence relation, let

$$M_{\mathcal{G}} = \left(\coprod_{i\in I}\Omega_i\right)/\sim$$

be the quotient set and let $p\colon \coprod_{i\in I}\Omega_i \to M_{\mathcal{G}}$ be the quotient map, with $p(x) = [x]$, where $[x]$ denotes the equivalence class of x Also, for every $i \in I$, let $\mathrm{in}_i\colon \Omega_i \to \coprod_{i\in I}\Omega_i$ be the natural injection and let

$$\tau_i = p \circ \mathrm{in}_i\colon \Omega_i \to M_{\mathcal{G}}\,.$$

Since we already noted that if $x \sim y$ and $x, y \in \Omega_i$, then $x = y$, we can conclude that every τ_i is injective. We give $M_{\mathcal{G}}$ the coarsest topology that makes the bijections, $\tau_i\colon \Omega_i \to \tau_i(\Omega_i)$, into homeomorphisms. Then, if we let $U_i = \tau_i(\Omega_i)$ and $\varphi_i = \tau_i^{-1}$, it is immediately verified that the (U_i, φ_i) are charts and that this collection of charts forms a C^k atlas for $M_{\mathcal{G}}$. As there are countably many charts, $M_{\mathcal{G}}$ is second-countable.

To prove that the topology is Hausdorff, we first prove the following:

Claim. For all $(i, j) \in I \times I$, we have $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset$ iff $(i, j) \in K$ and if so,

$$\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji})\,.$$

Assume that $\tau_i(\Omega_i)\cap\tau_j(\Omega_j) \neq \emptyset$ and let $[z] \in \tau_i(\Omega_i)\cap\tau_j(\Omega_j)$. Observe that $[z] \in \tau_i(\Omega_i)\cap\tau_j(\Omega_j)$ iff $z \sim x$ and $z \sim y$, for some $x \in \Omega_i$ and some $y \in \Omega_j$. Consequently, $x \sim y$, which implies that $(i, j) \in K$, $x \in \Omega_{ij}$ and $y \in \Omega_{ji}$. We have $[z] \in \tau_i(\Omega_{ij})$ iff $z \sim x$, for some $x \in \Omega_{ij}$. Then,

either $i = j$ and $z = x$ or $i \neq j$ and $z \in \Omega_{ji}$, which shows that $[z] \in \tau_j(\Omega_{ji})$ and consequently, we get $\tau_i(\Omega_{ij}) \subseteq \tau_j(\Omega_{ji})$. Since the same argument applies by interchanging i and j, we have that $\tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji})$, for all $(i, j) \in K$. Furthermore, because $\Omega_{ij} \subseteq \Omega_i$, $\Omega_{ji} \subseteq \Omega_j$, and $\tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji})$, for all $(i, j) \in K$, we also have that $\tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}) \subseteq \tau_i(\Omega_i) \cap \tau_j(\Omega_j)$, for all $(i, j) \in K$.

For the reverse inclusion, if $[z] \in \tau_i(\Omega_i) \cap \tau_j(\Omega_j)$, then we know that there is some $x \in \Omega_{ij}$ and some $y \in \Omega_{ji}$ such that $z \sim x$ and $z \sim y$, so $[z] = [x] \in \tau_i(\Omega_{ij})$ and $[z] = [y] \in \tau_j(\Omega_{ji})$, and then we get

$$\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \subseteq \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}) .$$

This proves that if $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset$, then $(i, j) \in K$ and

$$\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}) .$$

Finally, assume that $(i, j) \in K$. Then, for any $x \in \Omega_{ij} \subseteq \Omega_i$, we have $y = \varphi_{ji}(x) \in \Omega_{ji} \subseteq \Omega_j$ and $x \sim y$, so that $\tau_i(x) = \tau_j(y)$, which proves that $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset$. So, our claim is true, and we can use it.

We now prove that the topology of $M_{\mathcal{G}}$ is Hausdorff. Pick $[x], [y] \in M_{\mathcal{G}}$ with $[x] \neq [y]$, for some $x \in \Omega_i$ and some $y \in \Omega_j$. Either $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \emptyset$, in which case, as τ_i and τ_j are homeomorphisms, $[x]$ and $[y]$ belong to the two disjoint open sets $\tau_i(\Omega_i)$ and $\tau_j(\Omega_j)$. If not, then by the Claim, $(i, j) \in K$ and

$$\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}) .$$

There are several cases to consider:

1. If $i = j$ then x and y can be separated by disjoint opens, V_x and V_y, and as τ_i is a homeomorphism, $[x]$ and $[y]$ are separated by the disjoint open subsets $\tau_i(V_x)$ and $\tau_j(V_y)$.

2. If $i \neq j$, $x \in \Omega_i - \overline{\Omega_{ij}}$ and $y \in \Omega_j - \overline{\Omega_{ji}}$, then $\tau_i(\Omega_i - \overline{\Omega_{ij}})$ and $\tau_j(\Omega_j - \overline{\Omega_{ji}})$ are disjoint open subsets separating $[x]$ and $[y]$, where $\overline{\Omega_{ij}}$ and $\overline{\Omega_{ji}}$ are the closures of Ω_{ij} and Ω_{ji}, respectively.

3. If $i \neq j$, $x \in \Omega_{ij}$ and $y \in \Omega_{ji}$, as $[x] \neq [y]$ and $y \sim \varphi_{ij}(y)$, then $x \neq \varphi_{ij}(y)$. We can separate x and $\varphi_{ij}(y)$ by disjoint open subsets, V_x and V_y, and $[x]$ and $[y] = [\varphi_{ij}(y)]$ are separated by the disjoint open subsets $\tau_i(V_x)$ and $\tau_i(V_{\varphi_{ij}(y)})$.

4. If $i \neq j$, $x \in \partial(\Omega_{ij}) \cap \Omega_i$ and $y \in \partial(\Omega_{ji}) \cap \Omega_j$, then we use condition 4 of Definition 4.1. This condition yields two disjoint open subsets, V_x and V_y, with $x \in V_x$ and $y \in V_y$, such that no point of $V_x \cap \Omega_{ij}$ is equivalent to any point of $V_y \cap \Omega_{ji}$, and so $\tau_i(V_x)$ and $\tau_j(V_y)$ are disjoint open subsets separating $[x]$ and $[y]$.

Therefore, the topology of $M_{\mathcal{G}}$ is Hausdorff and $M_{\mathcal{G}}$ is indeed a manifold. Finally, it is trivial to verify that the transition maps of $M_{\mathcal{G}}$ are the original gluing functions, φ_{ij}, since $\varphi_i = \tau_i^{-1}$ and $\varphi_{ji} = \varphi_j \circ \varphi_i^{-1}$. □

It should be noted that as nice as it is, Proposition 4.1 is a theoretical construction that yields an "abstract" manifold but does not yield any information as to the geometry of this manifold. Furthermore, the resulting manifold may not be orientable or compact, even if we start with a finite set of p-domains.

Here is an example showing that if condition (4) of Definition 4.1 is omitted then we may get non-Hausdorff spaces. Cindy Grimm uses a similar example in her dissertation [68] (Appendix C2, page 126), but her presentation is somewhat confusing because her Ω_1 and Ω_2 appear to be two disjoint copies of the real line in $\mathbb{R}^2$, but these are not open in $\mathbb{R}^2$!

Let $\Omega_1 = (-3,-1)$, $\Omega_2 = (1,3)$, $\Omega_{12} = (-3,-2)$, $\Omega_{21} = (1,2)$ and $\varphi_{21}(x) = x+4$. The resulting space, M, is a curve looking like a "fork", and the problem is that the images of -2 and 2 in M, which are distinct points of M, cannot be separated. Indeed, the images of any two open intervals $(-2-\epsilon, -2+\epsilon)$ and $(2-\eta, 2+\eta)$ (for $\epsilon, \eta > 0$) always intersect since $(-2-\min(\epsilon,\eta), -2)$ and $(2-\min(\epsilon,\eta), 2)$ are identified. Clearly, condition (4) fails.

Cindy Grimm [68] (page 40) uses a condition stronger than our condition (4) to ensure that the quotient, $M_{\mathcal{G}}$ is Hausdorff, namely, that for all $(i,j) \in K$ with $i \neq j$, the quotient $(\Omega_i \coprod \Omega_j)/\sim$ should be embeddable in $\mathbb{R}^n$. This is a rather strong condition that prevents obtaining a 2-sphere by gluing two open discs in $\mathbb{R}^2$ along an annulus (see Grimm [68], Appendix C2, page 126).

Grimm uses the following cocycle condition in [68] (page 40) and [69] (page 361):

(c') For all $x \in \Omega_{ij} \cap \Omega_{ik}$,

$$\varphi_{ki}(x) = \varphi_{kj} \circ \varphi_{ji}(x).$$

This condition is not strong enough to imply transitivity of the relation $\sim$, as shown by the following counter-example:

Let $\Omega_1 = (0,3)$, $\Omega_2 = (4,5)$, $\Omega_3 = (6,9)$, $\Omega_{12} = (0,1)$, $\Omega_{13} = (2,3)$, $\Omega_{21} = \Omega_{23} = (4,5)$, $\Omega_{32} = (8,9)$, $\Omega_{31} = (6,7)$, $\varphi_{21}(x) = x+4$, $\varphi_{32}(x) = x+4$ and $\varphi_{31}(x) = x+4$.

Note that the pairwise gluings yield Hausdorff spaces. Obviously, $\varphi_{32} \circ \varphi_{21}(x) = x+8$, for all $x \in \Omega_{12}$, but $\Omega_{12} \cap \Omega_{13} = \emptyset$. Thus, $0.5 \sim 4.5 \sim 8.5$, and if the relation $\sim$ was transitive, then we would conclude that $0.5 \sim 8.5$. However, the definition of the relation $\sim$ requires that $\varphi_{31}(0.5)$ be defined, which is not the case. Therefore, the relation $\sim$ is not transitive. The problem is that because $\Omega_{12} \cap \Omega_{13} = \emptyset$, condition ($c'$) holds vacuously, but it is not strong enough to ensure that $\varphi_{31}(0.5)$ is defined.

Here is another counter-example in which $\Omega_{12} \cap \Omega_{13} \neq \emptyset$, using a disconnected open, Ω_2.

Let $\Omega_1 = (0,3)$, $\Omega_2 = (4,5) \cup (6,7)$, $\Omega_3 = (8,11)$, $\Omega_{12} = (0,1) \cup (2,3)$, $\Omega_{13} = (2,3)$, $\Omega_{21} = \Omega_{23} = (4,5) \cup (6,7)$, $\Omega_{32} = (8,9) \cup (10,11)$, $\Omega_{31} = (8,9)$, $\varphi_{21}(x) = x+4$, $\varphi_{32}(x) = x+2$ on $(6,7)$, $\varphi_{32}(x) = x+6$ on $(4,5)$, $\varphi_{31}(x) = x+6$.

Note that the pairwise gluings yield Hausdorff spaces. Obviously, $\varphi_{32} \circ \varphi_{21}(x) = x+6 = \varphi_{31}(x)$ for all $x \in \Omega_{12} \cap \Omega_{13} = (2,3)$. Thus, $0.5 \sim 4.5 \sim 8.5$, but $0.5 \not\sim 8.5$ since $\varphi_{31}(0.5)$ is

undefined. This time, condition (c') holds and is nontrivial since $\Omega_{12} \cap \Omega_{13} = (2, 3)$, but it is not strong enough to ensure that $\varphi_{31}(0.5)$ is defined.

It is possible to give a construction, in the case of a surface, which builds a compact manifold whose geometry is "close" to the geometry of a prescribed 3D-mesh (see Siqueira, Xu and Gallier [141]). Actually, we are not able to guarantee, in general, that the parametrization functions, θ_i, that we obtain are injective, but we are not aware of any algorithm that achieves this.

Given a set of gluing data, $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}, (\varphi_{ji})_{(i,j)\in K})$, it is natural to consider the collection of manifolds, M, parametrized by maps, $\theta_i \colon \Omega_i \to M$, whose domains are the Ω_i's and whose transitions functions are given by the φ_{ji}, that is, such that

$$\varphi_{ji} = \theta_j^{-1} \circ \theta_i.$$

We will say that such manifolds are *induced* by the set of gluing data, $\mathcal{G}$.

The proof of Proposition 4.1 shows that the parametrization maps, τ_i, satisfy the property: $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset$ iff $(i, j) \in K$ and if so,

$$\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}).$$

Furthermore, they also satisfy the consistency condition:

$$\tau_i = \tau_j \circ \varphi_{ji},$$

for all $(i, j) \in K$. If M is a manifold induced by the set of gluing data, $\mathcal{G}$, because the θ_i's are injective and $\varphi_{ji} = \theta_j^{-1} \circ \theta_i$, the two properties stated above for the τ_i's also hold for the θ_i's. We will see in Section 4.2 that the manifold, $M_{\mathcal{G}}$, is a "universal" manifold induced by $\mathcal{G}$ in the sense that every manifold induced by $\mathcal{G}$ is the image of $M_{\mathcal{G}}$ by some C^k map.

Interestingly, it is possible to characterize when two manifolds induced by the same set of gluing data are isomorphic in terms of a condition on their transition functions.

Proposition 4.2. *Given any set of gluing data, $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}, (\varphi_{ji})_{(i,j)\in K})$, for any two manifolds M and M' induced by $\mathcal{G}$ given by families of parametrizations $(\Omega_i, \theta_i)_{i\in I}$ and $(\Omega_i, \theta'_i)_{i\in I}$, respectively, if $f \colon M \to M'$ is a C^k isomorphism, then there are C^k bijections, $\rho_i \colon W_{ij} \to W'_{ij}$, for some open subsets $W_{ij}, W'_{ij} \subseteq \Omega_i$, such that*

$$\varphi'_{ji}(x) = \rho_j \circ \varphi_{ji} \circ \rho_i^{-1}(x), \qquad \textit{for all} \quad x \in W_{ij},$$

with $\varphi_{ji} = \theta_j^{-1} \circ \theta_i$ and $\varphi'_{ji} = \theta_j'^{-1} \circ \theta'_i$. Furthermore, $\rho_i = (\theta_i'^{-1} \circ f \circ \theta_i) \restriction W_{ij}$ and if $\theta_i'^{-1} \circ f \circ \theta_i$ is a bijection from Ω_i to itself and $\theta_i'^{-1} \circ f \circ \theta_i(\Omega_{ij}) = \Omega_{ij}$, for all i, j, then $W_{ij} = W'_{i,j} = \Omega_i$.

Proof. The composition $\theta_i'^{-1} \circ f \circ \theta_i$ is actually a partial function with domain

$$\mathrm{dom}(\theta_i'^{-1} \circ f \circ \theta_i) = \{x \in \Omega_i \mid \theta_i(x) \in f^{-1} \circ \theta'_i(\Omega_i)\}$$

and its "inverse" $\theta_i^{-1} \circ f^{-1} \circ \theta'_i$ is a partial function with domain

$$\mathrm{dom}(\theta_i^{-1} \circ f^{-1} \circ \theta'_i) = \{x \in \Omega_i \mid \theta'_i(x) \in f \circ \theta_i(\Omega_i)\}.$$

The composition ${\theta'_j}^{-1} \circ f \circ \theta_j \circ \varphi_{ji} \circ \theta_i^{-1} \circ f^{-1} \circ \theta'_i$ is also a partial function and we let

$$W_{ij} = \Omega_{ij} \cap \mathrm{dom}({\theta'_j}^{-1} \circ f \circ \theta_j \circ \varphi_{ji} \circ \theta_i^{-1} \circ f^{-1} \circ \theta'_i), \qquad \rho_i = ({\theta'_i}^{-1} \circ f \circ \theta_i) \restriction W_{ij}$$

and $W'_{ij} = \rho_i(W_{ij})$. Observe that $\theta_j \circ \varphi_{ji} = \theta_j \circ \theta_j^{-1} \circ \theta_i = \theta_i$, that is,

$$\theta_i = \theta_j \circ \varphi_{ji}.$$

Using this, on W_{ij}, we get

$$\begin{aligned}
\rho_j \circ \varphi_{ji} \circ \rho_i^{-1} &= {\theta'_j}^{-1} \circ f \circ \theta_j \circ \varphi_{ji} \circ ({\theta'_i}^{-1} \circ f \circ \theta_i)^{-1} \\
&= {\theta'_j}^{-1} \circ f \circ \theta_j \circ \varphi_{ji} \circ \theta_i^{-1} \circ f^{-1} \circ \theta'_i \\
&= {\theta'_j}^{-1} \circ f \circ \theta_i \circ \theta_i^{-1} \circ f^{-1} \circ \theta'_i \\
&= {\theta'_j}^{-1} \circ \theta'_i = \varphi'_{ji},
\end{aligned}$$

as claimed. The last part of the proposition is clear. □

Proposition 4.2 suggests defining a notion of equivalence on sets of gluing data which yields a converse of this proposition.

Definition 4.2. Two sets of gluing data, $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}, (\varphi_{ji})_{(i,j)\in K})$ and $\mathcal{G}' = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}, (\varphi'_{ji})_{(i,j)\in K})$, over the same sets of Ω_i's and Ω_{ij}'s are *equivalent* iff there is a family of C^k bijections, $(\rho_i \colon \Omega_i \to \Omega_i)_{i\in I}$, such that $\rho_i(\Omega_{ij}) = \Omega_{ij}$ and

$$\varphi'_{ji}(x) = \rho_j \circ \varphi_{ji} \circ \rho_i^{-1}(x), \qquad \text{for all} \quad x \in \Omega_{ij},$$

for all i, j.

Here is the converse of Proposition 4.2. It is actually nicer than Proposition 4.2 because we can take $W_{ij} = W'_{ij} = \Omega_i$.

Proposition 4.3. *If two sets of gluing data $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}, (\varphi_{ji})_{(i,j)\in K})$ and $\mathcal{G}' = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}, (\varphi'_{ji})_{(i,j)\in K})$ are equivalent, then there is a C^k isomorphism, $f \colon M_{\mathcal{G}} \to M_{\mathcal{G}'}$, between the manifolds induced by $\mathcal{G}$ and $\mathcal{G}'$. Furthermore, $f \circ \tau_i = \tau'_i \circ \rho_i$, for all $i \in I$.*

Proof. Let $f_i \colon \tau_i(\Omega_i) \to \tau'_i(\Omega_i)$ be the C^k bijection given by

$$f_i = \tau'_i \circ \rho_i \circ \tau_i^{-1},$$

where the $\rho_i \colon \Omega_i \to \Omega_i$'s are the maps giving the equivalence of $\mathcal{G}$ and $\mathcal{G}'$. If we prove that f_i and f_j agree on the overlap, $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji})$, then the f_i patch and yield a C^k isomorphism, $f \colon M_{\mathcal{G}} \to M_{\mathcal{G}'}$. The conditions of Proposition 4.3 imply that

$$\varphi'_{ji} \circ \rho_i = \rho_j \circ \varphi_{ji}$$

and we know that

$$\tau'_i = \tau'_j \circ \varphi'_{ji}.$$

Consequently, for every $[x] \in \tau_j(\Omega_{ji}) = \tau_i(\Omega_{ij})$, with $x \in \Omega_{ij}$, we have

$$\begin{aligned} f_j([x]) &= \tau'_j \circ \rho_j \circ \tau_j^{-1}([x]) \\ &= \tau'_j \circ \rho_j \circ \tau_j^{-1}([\varphi_{ji}(x)]) \\ &= \tau'_j \circ \rho_j \circ \varphi_{ji}(x) \\ &= \tau'_j \circ \varphi'_{ji} \circ \rho_i(x) \\ &= \tau'_i \circ \rho_i(x) \\ &= \tau'_i \circ \rho_i \circ \tau_i^{-1}([x]) \\ &= f_i([x]), \end{aligned}$$

which shows that f_i and f_j agree on $\tau_i(\Omega_i) \cap \tau_j(\Omega_j)$, as claimed. □

In the next section, we describe a class of spaces that can be defined by gluing data and parametrization functions, θ_i, that are not necessarily injective. Roughly speaking, the gluing data specify the topology and the parametrizations define the geometry of the space. Such spaces have more structure than spaces defined parametrically but they are not quite manifolds. Yet, they arise naturally in practice and they are the basis of efficient implementations of very good approximations of 3D meshes.

4.2 Parametric Pseudo-Manifolds

In practice, it is often desirable to specify some n-dimensional geometric shape as a subset of $\mathbb{R}^d$ (usually for $d = 3$) in terms of parametrizations which are functions, θ_i, from some subset of $\mathbb{R}^n$ into $\mathbb{R}^d$ (usually, $n = 2$). For "open" shapes, this is reasonably well understood but dealing with a "closed" shape is a lot more difficult because the parametrized pieces should overlap as smoothly as possible and this is hard to achieve. Furthermore, in practice, the parametrization functions, θ_i, may not be injective. Proposition 4.1 suggests various ways of defining such geometric shapes. For the lack of a better term, we will call these shapes, *parametric pseudo-manifolds*.

Definition 4.3. Let n, k, d be three integers with $d > n \geq 1$ and $k \geq 1$ or $k = \infty$. A *parametric C^k pseudo-manifold of dimension n in $\mathbb{R}^d$* is a pair, $\mathcal{M} = (\mathcal{G}, (\theta_i)_{i \in I})$, where $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j) \in I \times I}, (\varphi_{ji})_{(i,j) \in K})$ is a set of gluing data for some finite set, I, and each θ_i is a C^k function, $\theta_i \colon \Omega_i \to \mathbb{R}^d$, called a *parametrization* such that the following property holds:

(C) For all $(i,j) \in K$, we have

$$\theta_i = \theta_j \circ \varphi_{ji}.$$

For short, we use terminology *parametric pseudo-manifold*. The subset, $M \subseteq \mathbb{R}^d$, given by

$$M = \bigcup_{i \in I} \theta_i(\Omega_i)$$

is called the *image* of the parametric pseudo-manifold, $\mathcal{M}$. When $n = 2$ and $d = 3$, we say that $\mathcal{M}$ is a *parametric pseudo-surface*.

Condition (C) obviously implies that

$$\theta_i(\Omega_{ij}) = \theta_j(\Omega_{ji}),$$

for all $(i,j) \in K$. Consequently, θ_i and θ_j are consistent parametrizations of the overlap, $\theta_i(\Omega_{ij}) = \theta_j(\Omega_{ji})$. Thus, the shape, M, is covered by pieces, $U_i = \theta_i(\Omega_i)$, not necessarily open, with each U_i parametrized by θ_i and where the overlapping pieces, $U_i \cap U_j$, are parametrized consistently. The local structure of M is given by the θ_i's and the global structure is given by the gluing data. We recover a manifold if we require the θ_i to be bijective and to satisfy the following additional conditions:

(C') For all $(i,j) \in K$,

$$\theta_i(\Omega_i) \cap \theta_j(\Omega_j) = \theta_i(\Omega_{ij}) = \theta_j(\Omega_{ji}).$$

(C") For all $(i,j) \notin K$,

$$\theta_i(\Omega_i) \cap \theta_j(\Omega_j) = \emptyset.$$

Even if the θ_i's are not injective, properties (C') and (C") would be desirable since they guarantee that $\theta_i(\Omega_i - \Omega_{ij})$ and $\theta_j(\Omega_j - \Omega_{ji})$ are parametrized uniquely. Unfortunately, these properties are difficult to enforce. Observe that any manifold induced by $\mathcal{G}$ is the image of a parametric pseudo-manifold.

Although this is an abuse of language, it is more convenient to call M a parametric pseudo-manifold, or even a *pseudo-manifold*.

We can also show that the parametric pseudo-manifold, M, is the image in $\mathbb{R}^d$ of the abstract manifold, $M_{\mathcal{G}}$.

Proposition 4.4. *Let $\mathcal{M} = (\mathcal{G}, (\theta_i)_{i \in I})$ be parametric C^k pseudo-manifold of dimension n in $\mathbb{R}^d$, where $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j) \in I \times I}, (\varphi_{ji})_{(i,j) \in K})$ is a set of gluing data for some finite set, I. Then, the parametrization maps, θ_i, induce a surjective map, $\Theta \colon M_{\mathcal{G}} \to M$, from the abstract manifold, $M_{\mathcal{G}}$, specified by $\mathcal{G}$ to the image, $M \subseteq \mathbb{R}^d$, of the parametric pseudo-manifold, $\mathcal{M}$, and the following property holds: For every Ω_i,*

$$\theta_i = \Theta \circ \tau_i,$$

where the $\tau_i \colon \Omega_i \to M_{\mathcal{G}}$ are the parametrization maps of the manifold $M_{\mathcal{G}}$ (see Proposition 4.1). In particular, every manifold, M, induced by the gluing data $\mathcal{G}$ is the image of $M_{\mathcal{G}}$ by a map $\Theta \colon M_{\mathcal{G}} \to M$.

Proof. Recall that

$$M_{\mathcal{G}} = \Big(\coprod_{i\in I} \Omega_i\Big)/\sim,$$

where $\sim$ is the equivalence relation defined so that, for all $x, y \in \coprod_{i\in I} \Omega_i$,

$$x \sim y \quad \text{iff} \quad (\exists (i,j) \in K)(x \in \Omega_{ij},\ y \in \Omega_{ji},\ y = \varphi_{ji}(x)).$$

The proof of Proposition 4.1 also showed that $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset$ iff $(i,j) \in K$ and if so,

$$\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}).$$

In particular,

$$\tau_i(\Omega_i - \Omega_{ij}) \cap \tau_j(\Omega_j - \Omega_{ji}) = \emptyset$$

for all $(i,j) \in I \times I$ ($\Omega_{ij} = \Omega_{ji} = \emptyset$ when $(i,j) \notin K$). These properties with the fact that the τ_i's are injections show that for all $(i,j) \notin K$, we can define $\Theta_i \colon \tau_i(\Omega_i) \to \mathbb{R}^d$ and $\Theta_j \colon \tau_i(\Omega_j) \to \mathbb{R}^d$ by

$$\Theta_i([x]) = \theta_i(x),\ x \in \Omega_i \qquad \Theta_j([y]) = \theta_j(y),\ y \in \Omega_j.$$

For $(i,j) \in K$, as the the τ_i's are injections we can define $\Theta_i \colon \tau_i(\Omega_i - \Omega_{ij}) \to \mathbb{R}^d$ and $\Theta_j \colon \tau_i(\Omega_j - \Omega_{ji}) \to \mathbb{R}^d$ by

$$\Theta_i([x]) = \theta_i(x),\ x \in \Omega_i - \Omega_{ij} \qquad \Theta_j([y]) = \theta_j(y),\ y \in \Omega_j - \Omega_{ji}.$$

It remains to define Θ_i on $\tau_i(\Omega_{ij})$ and Θ_j on $\tau_j(\Omega_{ji})$ in such a way that they agree on $\tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji})$. However, condition (C) in Definition 4.3 says that for all $x \in \Omega_{ij}$,

$$\theta_i(x) = \theta_j(\varphi_{ji}(x)).$$

Consequently, if we define Θ_i on $\tau_i(\Omega_{ij})$ and Θ_j on $\tau_j(\Omega_{ji})$ by

$$\Theta_i([x]) = \theta_i(x),\ x \in \Omega_{ij}, \qquad \Theta_j([y]) = \theta_j(y),\ y \in \Omega_{ji},$$

as $x \sim \varphi_{ji}(x)$, we have

$$\Theta_i([x]) = \theta_i(x) = \theta_j(\varphi_{ji}(x)) = \Theta_j([\varphi_{ji}(x)]) = \Theta_j([x]),$$

which means that Θ_i and Θ_j agree on $\tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji})$. But then, the functions, Θ_i, agree whenever their domains overlap and so, they patch to yield a function, Θ, with domain $M_{\mathcal{G}}$ and image M. By construction, $\theta_i = \Theta \circ \tau_i$ and as a manifold induced by $\mathcal{G}$ is a parametric pseudo-manifold, the last statement is obvious. □

The function, $\Theta \colon M_{\mathcal{G}} \to M$, given by Proposition 4.4 shows how the parametric pseudo-manifold, M, differs from the abstract manifold, $M_{\mathcal{G}}$. As we said before, a practical method for approximating 3D meshes based on parametric pseudo surfaces is described in Siqueira, Xu and Gallier [141].

Chapter 5

Lie Groups, Lie Algebras and the Exponential Map

5.1 Lie Groups and Lie Algebras

In Chapter 1 we defined the notion of a Lie group as a certain type of manifold embedded in $\mathbb{R}^N$, for some $N \geq 1$. Now that we have the general concept of a manifold, we can define Lie groups in more generality. Besides classic references on Lie groups and Lie Algebras, such as Chevalley [34], Knapp [90], Warner [148], Duistermaat and Kolk [54], Bröcker and tom Dieck [25], Sagle and Walde [130], Helgason [74], Serre [138, 137], Kirillov [87], Fulton and Harris [58] and Bourbaki [22], one should be aware of more introductory sources and surveys such as Hall [71], Sattinger and Weaver [135], Carter, Segal and Macdonald [31], Curtis [39], Baker [13], Rossmann [128], Bryant [26], Mneimné and Testard [112] and Arvanitoyeogos [8].

Definition 5.1. A *Lie group* is a nonempty subset, G, satisfying the following conditions:

(a) G is a group (with identity element denoted e or 1).

(b) G is a smooth manifold.

(c) G is a topological group. In particular, the group operation, $\cdot : G \times G \to G$, and the inverse map, $^{-1}\colon G \to G$, are smooth.

We have already met a number of Lie groups: $\mathbf{GL}(n, \mathbb{R})$, $\mathbf{GL}(n, \mathbb{C})$, $\mathbf{SL}(n, \mathbb{R})$, $\mathbf{SL}(n, \mathbb{C})$, $\mathbf{O}(n)$, $\mathbf{SO}(n)$, $\mathbf{U}(n)$, $\mathbf{SU}(n)$, $\mathbf{E}(n, \mathbb{R})$. Also, every linear Lie group (i.e., a closed subgroup of $\mathbf{GL}(n, \mathbb{R})$) is a Lie group.

We saw in the case of linear Lie groups that the tangent space to G at the identity, $\mathfrak{g} = T_1 G$, plays a very important role. In particular, this vector space is equipped with a (non-associative) multiplication operation, the Lie bracket, that makes $\mathfrak{g}$ into a Lie algebra. This is again true in this more general setting.

Recall that Lie algebras are defined as follows:

Definition 5.2. A *(real) Lie algebra*, $\mathcal{A}$, is a real vector space together with a bilinear map, $[\cdot,\cdot]\colon \mathcal{A}\times\mathcal{A}\to\mathcal{A}$, called the *Lie bracket* on $\mathcal{A}$ such that the following two identities hold for all $a, b, c \in \mathcal{A}$:

$$[a,\, a] = 0,$$

and the so-called *Jacobi identity*

$$[a,\, [b,\, c]] + [c,\, [a,\, b]] + [b,\, [c,\, a]] = 0.$$

It is immediately verified that $[b, a] = -[a, b]$.

Let us also recall the definition of homomorphisms of Lie groups and Lie algebras.

Definition 5.3. Given two Lie groups G_1 and G_2, a *homomorphism (or map) of Lie groups* is a function, $f\colon G_1 \to G_2$, that is a homomorphism of groups and a smooth map (between the manifolds G_1 and G_2). Given two Lie algebras $\mathcal{A}_1$ and $\mathcal{A}_2$, a *homomorphism (or map) of Lie algebras* is a function, $f\colon \mathcal{A}_1 \to \mathcal{A}_2$, that is a linear map between the vector spaces $\mathcal{A}_1$ and $\mathcal{A}_2$ and that preserves Lie brackets, i.e.,

$$f([A, B]) = [f(A), f(B)]$$

for all $A, B \in \mathcal{A}_1$.

An *isomorphism of Lie groups* is a bijective function f such that both f and f^{-1} are maps of Lie groups, and an *isomorphism of Lie algebras* is a bijective function f such that both f and f^{-1} are maps of Lie algebras.

The Lie bracket operation on $\mathfrak{g}$ can be defined in terms of the so-called adjoint representation.

Given a Lie group G, for every $a \in G$ we define *left translation* as the map, $L_a\colon G \to G$, such that $L_a(b) = ab$, for all $b \in G$, and *right translation* as the map, $R_a\colon G \to G$, such that $R_a(b) = ba$, for all $b \in G$. Because multiplication and the inverse maps are smooth, the maps L_a and R_a are diffeomorphisms, and their derivatives play an important role. The inner automorphisms $R_{a^{-1}} \circ L_a$ (also written $R_{a^{-1}}L_a$ or $\mathbf{Ad}_a$) also play an important role. Note that

$$R_{a^{-1}}L_a(b) = aba^{-1}.$$

The derivative

$$d(R_{a^{-1}}L_a)_1\colon T_1G \to T_1G$$

of $R_{a^{-1}}L_a\colon G \to G$ at 1 is an isomorphism of Lie algebras, and since $T_1G = \mathfrak{g}$, we get a map denoted

$$\mathrm{Ad}_a\colon \mathfrak{g} \to \mathfrak{g}.$$

The map $a \mapsto \mathrm{Ad}_a$ is a map of Lie groups

$$\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g}),$$

called the *adjoint representation of* G (where $\mathbf{GL}(\mathfrak{g})$ denotes the Lie group of all bijective linear maps on $\mathfrak{g}$).

In the case of a linear group, one can verify that

$$\mathrm{Ad}(a)(X) = \mathrm{Ad}_a(X) = aXa^{-1}$$

for all $a \in G$ and all $X \in \mathfrak{g}$.

The derivative

$$d\mathrm{Ad}_1 \colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})$$

of $\mathrm{Ad} \colon G \to \mathbf{GL}(\mathfrak{g})$ at 1 is map of Lie algebras, denoted by

$$\mathrm{ad} \colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g}),$$

called the *adjoint representation of* $\mathfrak{g}$. (Recall that Theorem 1.28 immediately implies that the Lie algebra, $\mathfrak{gl}(\mathfrak{g})$, of $\mathbf{GL}(\mathfrak{g})$ is the vector space, $\mathrm{End}(\mathfrak{g}, \mathfrak{g})$, of all endomorphisms of $\mathfrak{g}$, that is, the vector space of all linear maps on $\mathfrak{g}$).

In the case of a linear group, it can be verified that

$$\mathrm{ad}(A)(B) = [A,\, B] = AB - BA,$$

for all $A, B \in \mathfrak{g}$.

One can also check (in general) that the Jacobi identity on $\mathfrak{g}$ is equivalent to the fact that ad preserves Lie brackets, i.e., ad is a map of Lie algebras:

$$\mathrm{ad}([u,\, v]) = [\mathrm{ad}(u),\, \mathrm{ad}(v)],$$

for all $u, v \in \mathfrak{g}$ (where on the right, the Lie bracket is the commutator of linear maps on $\mathfrak{g}$).

This is the key to the definition of the Lie bracket in the case of a general Lie group (not just a linear Lie group).

Definition 5.4. Given a Lie group, G, the tangent space, $\mathfrak{g} = T_1 G$, at the identity with the Lie bracket defined by

$$[u,\, v] = \mathrm{ad}(u)(v), \quad \text{for all } u, v \in \mathfrak{g},$$

is the *Lie algebra of the Lie group* G. The Lie algebra, $\mathfrak{g}$, of a Lie group, G, is also denoted by $\mathrm{L}(G)$ (for instance, when the notation $\mathfrak{g}$ is already used for something else).

Actually, we have to justify why $\mathfrak{g}$ really is a Lie algebra. For this, we have

Proposition 5.1. *Given a Lie group, G, the Lie bracket, $[u,\, v] = \mathrm{ad}(u)(v)$, of Definition 5.4 satisfies the axioms of a Lie algebra (given in Definition 5.2). Therefore, $\mathfrak{g}$ with this bracket is a Lie algebra.*

Proof. The proof requires Proposition 5.9, but we prefer to defer the proof of this Proposition until section 5.3. Since

$$\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g})$$

is a Lie group homomorphism, by Proposition 5.9, the map $\mathrm{ad} = d\mathrm{Ad}_1$ is a homomorphism of Lie algebras, $\mathrm{ad}\colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})$, which means that

$$\mathrm{ad}([u,v]) = \mathrm{ad}(u) \circ \mathrm{ad}(v) - \mathrm{ad}(v) \circ \mathrm{ad}(u), \quad \text{for all } u, v \in \mathfrak{g},$$

since the bracket in $\mathfrak{gl}(\mathfrak{g}) = \mathrm{End}(\mathfrak{g}, \mathfrak{g})$, is just the commutator. Applying the above to $w \in \mathfrak{g}$, we get the Jacobi identity. We still have to prove that $[u,u] = 0$, or equivalently, that $[v,u] = -[u,v]$. For this, following Duistermaat and Kolk [54] (Chapter 1, Section 1), consider the map

$$G \times G \longrightarrow G\colon (a,b) \mapsto aba^{-1}b^{-1}.$$

It is easy to see that its differential at $(1,1)$ is the zero map. We can then compute the differential w.r.t. b at $b = 1$ and evaluate at $v \in \mathfrak{g}$, getting $(\mathrm{Ad}_a - \mathrm{id})(v)$. Then, the second derivative w.r.t. a at $a = 1$ evaluated at $u \in \mathfrak{g}$ is $[u,v]$. On the other hand if we differentiate first w.r.t. a and then w.r.t. b, we first get $(\mathrm{id} - \mathrm{Ad}_b)(u)$ and then $-[v,u]$. As our original map is smooth, the second derivative is bilinear symmetric, so $[u,v] = -[v,u]$. □

Remark: After proving that $\mathfrak{g}$ is isomorphic to the vector space of left-invariant vector fields on G, we get another proof of Proposition 5.1.

5.2 Left and Right Invariant Vector Fields, the Exponential Map

A fairly convenient way to define the exponential map is to use left-invariant vector fields.

Definition 5.5. If G is a Lie group, a vector field, X, on G is *left-invariant* (resp. *right-invariant*) iff

$$d(L_a)_b(X(b)) = X(L_a(b)) = X(ab), \quad \text{for all } a, b \in G.$$

(resp.

$$d(R_a)_b(X(b)) = X(R_a(b)) = X(ba), \quad \text{for all } a, b \in G.)$$

Equivalently, a vector field, X, is left-invariant iff the following diagram commutes (and similarly for a right-invariant vector field):

$$\begin{array}{ccc} TG & \xrightarrow{d(L_a)} & TG \\ {\scriptstyle X}\uparrow & & \uparrow{\scriptstyle X} \\ G & \xrightarrow[L_a]{} & G \end{array}$$

If X is a left-invariant vector field, setting $b = 1$, we see that

$$X(a) = d(L_a)_1(X(1)),$$

which shows that X is determined by its value, $X(1) \in \mathfrak{g}$, at the identity (and similarly for right-invariant vector fields).

Conversely, given any $v \in \mathfrak{g}$, we can define the vector field, v^L, by

$$v^L(a) = d(L_a)_1(v), \quad \text{for all } a \in G.$$

We claim that v^L is left-invariant. This follows by an easy application of the chain rule:

$$\begin{aligned} v^L(ab) &= d(L_{ab})_1(v) \\ &= d(L_a \circ L_b)_1(v) \\ &= d(L_a)_b(d(L_b)_1(v)) \\ &= d(L_a)_b(v^L(b)). \end{aligned}$$

Furthermore, $v^L(1) = v$. Therefore, we showed that the map, $X \mapsto X(1)$, establishes an isomorphism between the space of left-invariant vector fields on G and $\mathfrak{g}$. In fact, the map $G \times \mathfrak{g} \longrightarrow TG$ given by $(a, v) \mapsto v^L(a)$ is an isomorphism between $G \times \mathfrak{g}$ and the tangent bundle, TG.

Remark: Given any $v \in \mathfrak{g}$, we can also define the vector field, v^R, by

$$v^R(a) = d(R_a)_1(v), \quad \text{for all } a \in G.$$

It is easily shown that v^R is right-invariant and we also have an isomorphism $G \times \mathfrak{g} \longrightarrow TG$ given by $(a, v) \mapsto v^R(a)$.

Another reason why left-invariant (resp. right-invariant) vector fields on a Lie group are important is that they are complete, i.e., they define a flow whose domain is $\mathbb{R} \times G$. To prove this, we begin with the following easy proposition:

Proposition 5.2. *Given a Lie group, G, if X is a left-invariant (resp. right-invariant) vector field and Φ is its flow, then*

$$\Phi(t, g) = g\Phi(t, 1) \quad (\textit{resp.} \quad \Phi(t, g) = \Phi(t, 1)g), \quad \textit{for all } (t, g) \in \mathcal{D}(X).$$

Proof. Write

$$\gamma(t) = g\Phi(t, 1) = L_g(\Phi(t, 1)).$$

Then, $\gamma(0) = g$ and, by the chain rule

$$\dot{\gamma}(t) = d(L_g)_{\Phi(t,1)}(\dot{\Phi}(t, 1)) = d(L_g)_{\Phi(t,1)}(X(\Phi(t, 1))) = X(L_g(\Phi(t, 1))) = X(\gamma(t)).$$

By the uniqueness of maximal integral curves, $\gamma(t) = \Phi(t, g)$ for all t, and so,

$$\Phi(t, g) = g\Phi(t, 1).$$

A similar argument applies to right-invariant vector fields. □

Proposition 5.3. *Given a Lie group, G, for every $v \in \mathfrak{g}$, there is a unique smooth homomorphism, $h_v\colon (\mathbb{R}, +) \to G$, such that $\dot{h}_v(0) = v$. Furthermore, $h_v(t)$ is the maximal integral curve of both v^L and v^R with initial condition 1 and the flows of v^L and v^R are defined for all $t \in \mathbb{R}$.*

Proof. Let $\Phi^v_t(g)$ denote the flow of v^L. As far as defined, we know that

$$\Phi^v_{s+t}(1) = \Phi^v_s(\Phi^v_t(1)) = \Phi^v_t(1)\Phi^v_s(1),$$

by Proposition 5.2. Now, if $\Phi^v_t(1)$ is defined on $]-\epsilon, \epsilon[$, setting $s = t$, we see that $\Phi^v_t(1)$ is actually defined on $]-2\epsilon, 2\epsilon[$. By induction, we see that $\Phi^v_t(1)$ is defined on $]-2^n\epsilon, 2^n\epsilon[$, for all $n \geq 0$, and so, $\Phi^v_t(1)$ is defined on $\mathbb{R}$ and the map $t \mapsto \Phi^v_t(1)$ is a homomorphism, $h_v\colon (\mathbb{R}, +) \to G$, with $\dot{h}_v(0) = v$. Since $\Phi^v_t(g) = g\Phi^v_t(1)$, the flow, $\Phi^v_t(g)$, is defined for all $(t, g) \in \mathbb{R} \times G$. A similar proof applies to v^R. To show that h_v is smooth, consider the map

$$\mathbb{R} \times G \times \mathfrak{g} \longrightarrow G \times \mathfrak{g}, \quad \text{where} \quad (t, g, v) \mapsto (g\Phi^v_t(1), v).$$

It is immediately seen that the above is the flow of the vector field

$$(g, v) \mapsto (v(g), 0),$$

and thus, it is smooth. Consequently, the restriction of this smooth map to $\mathbb{R} \times \{1\} \times \{v\}$, which is just $t \mapsto \Phi^v_t(1) = h_v(t)$, is also smooth.

Assume $h\colon (\mathbb{R}, +) \to G$ is a smooth homomorphism with $\dot{h}(0) = v$. From

$$h(t + s) = h(t)h(s) = h(s)h(t),$$

if we differentiate with respect to s at $s = 0$, we get

$$\frac{dh}{dt}(t) = d(L_{h(t)})_1(v) = v^L(h(t))$$

and

$$\frac{dh}{dt}(t) = d(R_{h(t)})_1(v) = v^R(h(t)).$$

Therefore, $h(t)$ is an integral curve for v^L and v^R with initial condition $h(0) = 1$ and $h = \Phi^v_t(1)$. □

Since $h_v\colon (\mathbb{R}, +) \to G$ is a homomorphism, the integral curve, h_v, if often referred to as a *one-parameter group*. Proposition 5.3 yields the definition of the exponential map.

Definition 5.6. Given a Lie group, G, the *exponential map*, $\exp\colon \mathfrak{g} \to G$, is given by

$$\exp(v) = h_v(1) = \Phi^v_1(1), \quad \text{for all } v \in \mathfrak{g}.$$

We can see that exp is smooth as follows. As in the proof of Proposition 5.3, we have the smooth map

$$\mathbb{R} \times G \times \mathfrak{g} \longrightarrow G \times \mathfrak{g}, \quad \text{where} \quad (t, g, v) \mapsto (g\Phi_t^v(1), v),$$

which is the flow of the vector field

$$(g, v) \mapsto (v(g), 0).$$

Consequently, the restriction of this smooth map to $\{1\} \times \{1\} \times \mathfrak{g}$, which is just $v \mapsto \Phi_1^v(1) = \exp(v)$, is also smooth.

Observe that for any fixed $t \in \mathbb{R}$, the map

$$s \mapsto h_v(st)$$

is a smooth homomorphism, h, such that $\dot{h}(0) = tv$. By uniqueness, we have

$$h_v(st) = h_{tv}(s).$$

Setting $s = 1$, we find that

$$h_v(t) = \exp(tv), \quad \text{for all } v \in \mathfrak{g} \text{ and all } t \in \mathbb{R}.$$

Then, differentiating with respect to t at $t = 0$, we get

$$v = d\exp_0(v),$$

i.e., $d\exp_0 = \mathrm{id}_{\mathfrak{g}}$. By the inverse function theorem, exp is a local diffeomorphism at 0. This means that there is some open subset, $U \subseteq \mathfrak{g}$, containing 0, such that the restriction of exp to U is a diffeomorphism onto $\exp(U) \subseteq G$, with $1 \in \exp(U)$. In fact, by left-translation, the map $v \mapsto g\exp(v)$ is a local diffeomorphism between some open subset, $U \subseteq \mathfrak{g}$, containing 0 and the open subset, $\exp(U)$, containing g. The exponential map is also natural in the following sense:

Proposition 5.4. *Given any two Lie groups, G and H, for every Lie group homomorphism, $f\colon G \to H$, the following diagram commutes:*

$$\begin{array}{ccc} G & \xrightarrow{f} & H \\ {\scriptstyle \exp}\uparrow & & \uparrow{\scriptstyle \exp} \\ \mathfrak{g} & \xrightarrow[df_1]{} & \mathfrak{h} \end{array}$$

Proof. Observe that the map $h\colon t \mapsto f(\exp(tv))$ is a homomorphism from $(\mathbb{R}, +)$ to G such that $\dot{h}(0) = df_1(v)$. Proposition 5.3 shows that $f(\exp(v)) = \exp(df_1(v))$. □

A useful corollary of Proposition 5.4 is:

Proposition 5.5. *Let G be a connected Lie group and H be any Lie group. For any two homomorphisms, $\varphi_1 \colon G \to H$ and $\varphi_2 \colon G \to H$, if $d(\varphi_1)_1 = d(\varphi_2)_1$, then $\varphi_1 = \varphi_2$.*

Proof. We know that the exponential map is a diffeomorphism on some small open subset, U, containing 0. Now, by Proposition 5.4, for all $a \in \exp_G(U)$, we have

$$\varphi_i(a) = \exp_H(d(\varphi_i)_1(\exp_G^{-1}(a))), \quad i = 1, 2.$$

Since $d(\varphi_1)_1 = d(\varphi_2)_1$, we conclude that $\varphi_1 = \varphi_2$ on $\exp_G(U)$. However, as G is connected, Proposition 2.18 implies that G is generated by $\exp_G(U)$ (we can easily find a symmetric neighborhood of 1 in $\exp_G(U)$). Therefore, $\varphi_1 = \varphi_2$ on G. □

The above proposition shows that if G is connected, then a homomorphism of Lie groups, $\varphi \colon G \to H$, is uniquely determined by the Lie algebra homomorphism, $d\varphi_1 \colon \mathfrak{g} \to \mathfrak{h}$.

We obtain another useful corollary of Proposition 5.4 when we apply it to the adjoint representation of G,

$$\mathrm{Ad} \colon G \to \mathbf{GL}(\mathfrak{g})$$

and to the conjugation map,

$$\mathbf{Ad}_a \colon G \to G,$$

where $\mathbf{Ad}_a(b) = aba^{-1}$. In the first case, $d\mathrm{Ad}_1 = \mathrm{ad}$, with $\mathrm{ad} \colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})$ and in the second case, $d(\mathbf{Ad}_a)_1 = \mathrm{Ad}_a$.

Proposition 5.6. *Given any Lie group, G, the following properties hold:*

(1)

$$\mathrm{Ad}(\exp(u)) = e^{\mathrm{ad}(u)}, \qquad \textit{for all } u \in \mathfrak{g},$$

where $\exp \colon \mathfrak{g} \to G$ is the exponential of the Lie group, G, and $f \mapsto e^f$ is the exponential map given by

$$e^f = \sum_{k=0}^{\infty} \frac{f^k}{k!},$$

for any linear map (matrix), $f \in \mathfrak{gl}(\mathfrak{g})$. Equivalently, the following diagram commutes:

$$\begin{array}{ccc} G & \xrightarrow{\mathrm{Ad}} & \mathbf{GL}(\mathfrak{g}) \\ {\scriptstyle \exp}\uparrow & & \uparrow{\scriptstyle f \mapsto e^f} \\ \mathfrak{g} & \xrightarrow[\mathrm{ad}]{} & \mathfrak{gl}(\mathfrak{g}). \end{array}$$

(2)

$$\exp(t\mathrm{Ad}_g(u)) = g\exp(tu)g^{-1},$$

for all $u \in \mathfrak{g}$, all $g \in G$ and all $t \in \mathbb{R}$. Equivalently, the following diagram commutes:

$$\begin{array}{ccc} G & \xrightarrow{\mathbf{Ad}_g} & G \\ {\scriptstyle\exp}\uparrow & & \uparrow{\scriptstyle\exp} \\ \mathfrak{g} & \xrightarrow[\mathrm{Ad}_g]{} & \mathfrak{g}. \end{array}$$

Since the Lie algebra $\mathfrak{g} = T_1G$ is isomorphic to the vector space of left-invariant vector fields on G and since the Lie bracket of vector fields makes sense (see Definition 3.20), it is natural to ask if there is any relationship between, $[u, v]$, where $[u, v] = \mathrm{ad}(u)(v)$, and the Lie bracket, $[u^L, v^L]$, of the left-invariant vector fields associated with $u, v \in \mathfrak{g}$. The answer is: Yes, they coincide (*via* the correspondence $u \mapsto u^L$). This fact is recorded in the proposition below whose proof involves some rather acrobatic uses of the chain rule found in Warner [148] (Chapter 3), Bröcker and tom Dieck [25] (Chapter 1, Section 2), or Marsden and Ratiu [103] (Chapter 9).

Proposition 5.7. *Given a Lie group, G, we have*

$$[u^L, v^L](1) = \mathrm{ad}(u)(v), \quad \textit{for all } u, v \in \mathfrak{g}.$$

We can apply Proposition 2.22 and use the exponential map to prove a useful result about Lie groups. If G is a Lie group, let G_0 be the connected component of the identity. We know G_0 is a topological normal subgroup of G and it is a submanifold in an obvious way, so it is a Lie group.

Proposition 5.8. *If G is a Lie group and G_0 is the connected component of 1, then G_0 is generated by $\exp(\mathfrak{g})$. Moreover, G_0 is countable at infinity.*

Proof. We can find a symmetric open, U, in $\mathfrak{g}$ in containing 0, on which exp is a diffeomorphism. Then, apply Proposition 2.22 to $V = \exp(U)$. That G_0 is countable at infinity follows from Proposition 2.23. □

5.3 Homomorphisms of Lie Groups and Lie Algebras, Lie Subgroups

If G and H are two Lie groups and $\varphi\colon G \to H$ is a homomorphism of Lie groups, then $d\varphi_1\colon \mathfrak{g} \to \mathfrak{h}$ is a linear map between the Lie algebras $\mathfrak{g}$ and $\mathfrak{h}$ of G and H. In fact, it is a Lie algebra homomorphism, as shown below.

Proposition 5.9. *If G and H are two Lie groups and $\varphi\colon G \to H$ is a homomorphism of Lie groups, then*

$$d\varphi_1 \circ \mathrm{Ad}_g = \mathrm{Ad}_{\varphi(g)} \circ d\varphi_1, \quad \textit{for all } g \in G,$$

that is, the following diagram commutes

$$\begin{array}{ccc} \mathfrak{g} & \xrightarrow{d\varphi_1} & \mathfrak{h} \\ {\scriptstyle \mathrm{Ad}_g}\downarrow & & \downarrow{\scriptstyle \mathrm{Ad}_{\varphi(g)}} \\ \mathfrak{g} & \xrightarrow[d\varphi_1]{} & \mathfrak{h} \end{array}$$

and $d\varphi_1\colon \mathfrak{g} \to \mathfrak{h}$ is a Lie algebra homomorphism.

Proof. Recall that

$$R_{a^{-1}} L_a(b) = aba^{-1}, \quad \text{for all } a, b \in G$$

and that the derivative

$$d(R_{a^{-1}} L_a)_1\colon \mathfrak{g} \to \mathfrak{g}$$

of $R_{a^{-1}} L_a$ at 1 is an isomorphism of Lie algebras, denoted by $\mathrm{Ad}_a\colon \mathfrak{g} \to \mathfrak{g}$. The map $a \mapsto \mathrm{Ad}_a$ is a map of Lie groups

$$\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g}),$$

(where $\mathbf{GL}(\mathfrak{g})$ denotes the Lie group of all bijective linear maps on $\mathfrak{g}$) and the derivative

$$d\mathrm{Ad}_1\colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})$$

of Ad at 1 is map of Lie algebras, denoted by

$$\mathrm{ad}\colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g}),$$

called the adjoint representation of $\mathfrak{g}$ (where $\mathfrak{gl}(\mathfrak{g})$ denotes the Lie algebra of all linear maps on $\mathfrak{g}$). Then, the Lie bracket is defined by

$$[u, v] = \mathrm{ad}(u)(v), \quad \text{for all } u, v \in \mathfrak{g}.$$

Now, as φ is a homomorphism, we have

$$\varphi(R_{a^{-1}} L_a(b)) = \varphi(aba^{-1}) = \varphi(a)\varphi(b)\varphi(a)^{-1} = R_{\varphi(a)^{-1}} L_{\varphi(a)}(\varphi(b)),$$

and by differentiating w.r.t. b at $b = 1$ in the direction, $v \in \mathfrak{g}$, we get

$$d\varphi_1(\mathrm{Ad}_a(v)) = \mathrm{Ad}_{\varphi(a)}(d\varphi_1(v)),$$

proving the first part of the proposition. Differentiating again with respect to a at $a = 1$ in the direction, $u \in \mathfrak{g}$, (and using the chain rule), we get

$$d\varphi_1(\mathrm{ad}(u)(v)) = \mathrm{ad}(d\varphi_1(u))(d\varphi_1(v)),$$

i.e.,

$$d\varphi_1[u, v] = [d\varphi_1(u), d\varphi_1(v)],$$

which proves that $d\varphi_1$ is indeed a Lie algebra homomorphism. □

Remark: If we identify the Lie algebra, $\mathfrak{g}$, of G with the space of left-invariant vector fields on G, the map $d\varphi_1 \colon \mathfrak{g} \to \mathfrak{h}$ is viewed as the map such that, for every left-invariant vector field, X, on G, the vector field $d\varphi_1(X)$ is the unique left-invariant vector field on H such that

$$d\varphi_1(X)(1) = d\varphi_1(X(1)),$$

i.e., $d\varphi_1(X) = d\varphi_1(X(1))^L$. Then, we can give another proof of the fact that $d\varphi_1$ is a Lie algebra homomorphism using the notion of φ-related vector fields.

Proposition 5.10. *If G and H are two Lie groups and $\varphi \colon G \to H$ is a homomorphism of Lie groups, if we identify $\mathfrak{g}$ (resp. $\mathfrak{h}$) with the space of left-invariant vector fields on G (resp. left-invariant vector fields on H), then,*

(a) X and $d\varphi_1(X)$ are φ-related, for every left-invariant vector field, X, on G;

(b) $d\varphi_1 \colon \mathfrak{g} \to \mathfrak{h}$ is a Lie algebra homomorphism.

Proof. The proof uses Proposition 3.20. For details, see Warner [148]. □

We now consider Lie subgroups. As a preliminary result, note that if $\varphi \colon G \to H$ is an injective Lie group homomorphism, then $d\varphi_g \colon T_gG \to T_{\varphi(g)}H$ is injective for all $g \in G$. As $\mathfrak{g} = T_1G$ and T_gG are isomorphic for all $g \in G$ (and similarly for $\mathfrak{h} = T_1H$ and T_hH for all $h \in H$), it is sufficient to check that $d\varphi_1 \colon \mathfrak{g} \to \mathfrak{h}$ is injective. However, by Proposition 5.4, the diagram

$$\begin{array}{ccc} G & \xrightarrow{\varphi} & H \\ {\scriptstyle \exp}\uparrow & & \uparrow{\scriptstyle \exp} \\ \mathfrak{g} & \xrightarrow[d\varphi_1]{} & \mathfrak{h} \end{array}$$

commutes, and since the exponential map is a local diffeomorphism at 0, as φ is injective, then $d\varphi_1$ is injective, too. Therefore, if $\varphi \colon G \to H$ is injective, it is automatically an immersion.

Definition 5.7. Let G be a Lie group. A set, H, is an *immersed (Lie) subgroup* of G iff

(a) H is a Lie group;

(b) There is an injective Lie group homomorphism, $\varphi \colon H \to G$ (and thus, φ is an immersion, as noted above).

We say that H is a *Lie subgroup* (or *closed Lie subgroup*) of G iff H is a Lie group that is a subgroup of G and also a submanifold of G.

Observe that an immersed Lie subgroup, H, is an immersed submanifold, since φ is an injective immersion. However, $\varphi(H)$ may *not* have the subspace topology inherited from G and $\varphi(H)$ may not be closed.

An example of this situation is provided by the 2-torus, $T^2 \cong \mathbf{SO}(2) \times \mathbf{SO}(2)$, which can be identified with the group of 2×2 complex diagonal matrices of the form

$$\begin{pmatrix} e^{i\theta_1} & 0 \\ 0 & e^{i\theta_2} \end{pmatrix}$$

where $\theta_1, \theta_2 \in \mathbb{R}$. For any $c \in \mathbb{R}$, let S_c be the subgroup of T^2 consisting of all matrices of the form

$$\begin{pmatrix} e^{it} & 0 \\ 0 & e^{ict} \end{pmatrix}, \quad t \in \mathbb{R}.$$

It is easily checked that S_c is an immersed Lie subgroup of T^2 iff c is irrational. However, when c is irrational, one can show that S_c is dense in T^2 but not closed.

As we will see below, a Lie subgroup, is always closed. We borrowed the terminology "immersed subgroup" from Fulton and Harris [58] (Chapter 7), but we warn the reader that most books call such subgroups "Lie subgroups" and refer to the second kind of subgroups (that are submanifolds) as "closed subgroups".

Theorem 5.11. *Let G be a Lie group and let (H, φ) be an immersed Lie subgroup of G. Then, φ is an embedding iff $\varphi(H)$ is closed in G. As as consequence, any Lie subgroup of G is closed.*

Proof. The proof can be found in Warner [148] (Chapter 1, Theorem 3.21) and uses a little more machinery than we have introduced. However, we prove that a Lie subgroup, H, of G is closed. The key to the argument is this: Since H is a submanifold of G, there is chart, (U, φ), of G, with $1 \in U$, so that

$$\varphi(U \cap H) = \varphi(U) \cap (R^m \times \{0_{n-m}\}).$$

By Proposition 2.15, we can find some open subset, $V \subseteq U$, with $1 \in V$, so that $V = V^{-1}$ and $\overline{V} \subseteq U$. Observe that

$$\varphi(\overline{V} \cap H) = \varphi(\overline{V}) \cap (R^m \times \{0_{n-m}\})$$

and since $\overline{V}$ is closed and φ is a homeomorphism, it follows that $\overline{V} \cap H$ is closed. Thus, $\overline{V} \cap H = \overline{V} \cap \overline{H}$ (as $\overline{\overline{V} \cap H} = \overline{V} \cap \overline{H}$). Now, pick any $y \in \overline{H}$. As $1 \in V^{-1}$, the open set yV^{-1} contains y and since $y \in \overline{H}$, we must have $yV^{-1} \cap H \neq \emptyset$. Let $x \in yV^{-1} \cap H$, then $x \in H$ and $y \in xV$. Then, $y \in xV \cap \overline{H}$, which implies $x^{-1}y \in V \cap \overline{H} \subseteq \overline{V} \cap \overline{H} = \overline{V} \cap H$. Therefore, $x^{-1}y \in H$ and since $x \in H$, we get $y \in H$ and H is closed. □

We also have the following important and useful theorem: If G is a Lie group, say that a subset, $H \subseteq G$, is an *abstract subgroup* iff it is just a subgroup of the underlying group of G (i.e., we forget the topology and the manifold structure).

Theorem 5.12. *Let G be a Lie group. An abstract subgroup, H, of G is a submanifold (i.e., a Lie subgroup) of G iff H is closed (i.e, H with the induced topology is closed in G).*

Proof. We proved the easy direction of this theorem above. Conversely, we need to prove that if the subgroup, H, with the induced topology is closed in G, then it is a manifold. This can be done using the exponential map, but it is harder. For details, see Bröcker and tom Dieck [25] (Chapter 1, Section 3) or Warner [148], Chapter 3. □

5.4 The Correspondence Lie Groups–Lie Algebras

Historically, Lie was the first to understand that a lot of the structure of a Lie group is captured by its Lie algebra, a simpler object (since it is a vector space). In this short section, we state without proof some of the "Lie theorems", although not in their original form.

Definition 5.8. If $\mathfrak{g}$ is a Lie algebra, a *subalgebra*, $\mathfrak{h}$, of $\mathfrak{g}$ is a (linear) subspace of $\mathfrak{g}$ such that $[u, v] \in \mathfrak{h}$, for all $u, v \in \mathfrak{h}$. If $\mathfrak{h}$ is a (linear) subspace of $\mathfrak{g}$ such that $[u, v] \in \mathfrak{h}$ for all $u \in \mathfrak{h}$ and all $v \in \mathfrak{g}$, we say that $\mathfrak{h}$ is an *ideal* in $\mathfrak{g}$.

For a proof of the theorem below, see Warner [148] (Chapter 3) or Duistermaat and Kolk [54] (Chapter 1, Section 10).

Theorem 5.13. *Let G be a Lie group with Lie algebra, $\mathfrak{g}$, and let (H, φ) be an immersed Lie subgroup of G with Lie algebra $\mathfrak{h}$, then $d\varphi_1\mathfrak{h}$ is a Lie subalgebra of $\mathfrak{g}$. Conversely, for each subalgebra, $\widetilde{\mathfrak{h}}$, of $\mathfrak{g}$, there is a unique connected immersed subgroup, (H, φ), of G so that $d\varphi_1\mathfrak{h} = \widetilde{\mathfrak{h}}$. In fact, as a group, $\varphi(H)$ is the subgroup of G generated by $\exp(\widetilde{\mathfrak{h}})$. Furthermore, normal subgroups correspond to ideals.*

Theorem 5.13 shows that there is a one-to-one correspondence between connected immersed subgroups of a Lie group and subalgebras of its Lie algebra.

Theorem 5.14. *Let G and H be Lie groups with G connected and simply connected and let $\mathfrak{g}$ and $\mathfrak{h}$ be their Lie algebras. For every homomorphism, $\psi\colon \mathfrak{g} \to \mathfrak{h}$, there is a unique Lie group homomorphism, $\varphi\colon G \to H$, so that $d\varphi_1 = \psi$.*

Again a proof of the theorem above is given in Warner [148] (Chapter 3) or Duistermaat and Kolk [54] (Chapter 1, Section 10).

Corollary 5.15. *If G and H are connected and simply connected Lie groups, then G and H are isomorphic iff $\mathfrak{g}$ and $\mathfrak{h}$ are isomorphic.*

It can also be shown that for every finite-dimensional Lie algebra, $\mathfrak{g}$, there is a connected and simply connected Lie group, G, such that $\mathfrak{g}$ is the Lie algebra of G. This is a consequence of deep theorem (whose proof is quite hard) known as *Ado's theorem*. For more on this, see Knapp [90], Fulton and Harris [58], or Bourbaki [22].

In summary, following Fulton and Harris, we have the following two principles of the Lie group/Lie algebra correspondence:

First Principle: If G and H are Lie groups, with G connected, then a homomorphism of Lie groups, $\varphi\colon G \to H$, is uniquely determined by the Lie algebra homomorphism, $d\varphi_1\colon \mathfrak{g} \to \mathfrak{h}$.

Second Principle: Let G and H be Lie groups with G connected and simply connected and let $\mathfrak{g}$ and $\mathfrak{h}$ be their Lie algebras. A linear map, $\psi\colon \mathfrak{g} \to \mathfrak{h}$, is a Lie algebra map iff there is a unique Lie group homomorphism, $\varphi\colon G \to H$, so that $d\varphi_1 = \psi$.

5.5 More on the Lorentz Group $\mathbf{SO}_0(n,1)$

In this section, we take a closer look at the Lorentz group $\mathbf{SO}_0(n,1)$ and, in particular, at the relationship between $\mathbf{SO}_0(n,1)$ and its Lie algebra, $\mathfrak{so}(n,1)$. The Lie algebra of $\mathbf{SO}_0(n,1)$ is easily determined by computing the tangent vectors to curves, $t \mapsto A(t)$, on $\mathbf{SO}_0(n,1)$ through the identity, I. Since $A(t)$ satisfies

$$A^\top J A = J,$$

differentiating and using the fact that $A(0) = I$, we get

$$A'^\top J + J A' = 0.$$

Therefore,

$$\mathfrak{so}(n,1) = \{A \in \mathrm{Mat}_{n+1,n+1}(\mathbb{R}) \mid A^\top J + JA = 0\}.$$

This means that JA is skew-symmetric and so,

$$\mathfrak{so}(n,1) = \left\{ \begin{pmatrix} B & u \\ u^\top & 0 \end{pmatrix} \in \mathrm{Mat}_{n+1,n+1}(\mathbb{R}) \mid u \in \mathbb{R}^n, \quad B^\top = -B \right\}.$$

Observe that every matrix $A \in \mathfrak{so}(n,1)$ can be written uniquely as

$$\begin{pmatrix} B & u \\ u^\top & 0 \end{pmatrix} = \begin{pmatrix} B & 0 \\ 0^\top & 0 \end{pmatrix} + \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix},$$

where the first matrix is skew-symmetric, the second one is symmetric and both belong to $\mathfrak{so}(n,1)$. Thus, it is natural to define

$$\mathfrak{k} = \left\{ \begin{pmatrix} B & 0 \\ 0^\top & 0 \end{pmatrix} \mid B^\top = -B \right\}$$

and

$$\mathfrak{p} = \left\{ \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix} \mid u \in \mathbb{R}^n \right\}.$$

It is immediately verified that both $\mathfrak{k}$ and $\mathfrak{p}$ are subspaces of $\mathfrak{so}(n,1)$ (as vector spaces) and that $\mathfrak{k}$ is a Lie subalgebra isomorphic to $\mathfrak{so}(n)$, but $\mathfrak{p}$ is *not* a Lie subalgebra of $\mathfrak{so}(n,1)$ because it is not closed under the Lie bracket. Still, we have

$$[\mathfrak{k},\mathfrak{k}] \subseteq \mathfrak{k}, \quad [\mathfrak{k},\mathfrak{p}] \subseteq \mathfrak{p}, \quad [\mathfrak{p},\mathfrak{p}] \subseteq \mathfrak{k}.$$

Clearly, we have the direct sum decomposition

$$\mathfrak{so}(n,1) = \mathfrak{k} \oplus \mathfrak{p},$$

known as *Cartan decomposition*. There is also an automorphism of $\mathfrak{so}(n,1)$ known as the *Cartan involution*, namely,

$$\theta(A) = -A^\top,$$

and we see that

$$\mathfrak{k} = \{A \in \mathfrak{so}(n,1) \mid \theta(A) = A\} \quad \text{and} \quad \mathfrak{p} = \{A \in \mathfrak{so}(n,1) \mid \theta(A) = -A\}.$$

Unfortunately, there does not appear to be any simple way of obtaining a formula for $\exp(A)$, where $A \in \mathfrak{so}(n,1)$ (except for small n—there is such a formula for $n = 3$ due to Chris Geyer). However, it is possible to obtain an explicit formula for the matrices in $\mathfrak{p}$. This is because for such matrices, A, if we let $\omega = \|u\| = \sqrt{u^\top u}$, we have

$$A^3 = \omega^2 A.$$

Thus, we get

Proposition 5.16. *For every matrix, $A \in \mathfrak{p}$, of the form*

$$A = \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix},$$

we have

$$e^A = \begin{pmatrix} I + \frac{(\cosh\omega - 1)}{\omega^2} uu^\top & \frac{\sinh\omega}{\omega} u \\ \frac{\sinh\omega}{\omega} u^\top & \cosh\omega \end{pmatrix} = \begin{pmatrix} \sqrt{I + \frac{\sinh^2\omega}{\omega^2} uu^\top} & \frac{\sinh\omega}{\omega} u \\ \frac{\sinh\omega}{\omega} u^\top & \cosh\omega \end{pmatrix}.$$

Proof. Using the fact that $A^3 = \omega^2 A$, we easily prove that

$$e^A = I + \frac{\sinh\omega}{\omega} A + \frac{\cosh\omega - 1}{\omega^2} A^2,$$

which is the first equation of the proposition, since

$$A^2 = \begin{pmatrix} uu^\top & 0 \\ 0 & u^\top u \end{pmatrix} = \begin{pmatrix} uu^\top & 0 \\ 0 & \omega^2 \end{pmatrix}.$$

We leave as an exercise the fact that

$$\left(I + \frac{(\cosh\omega - 1)}{\omega^2} uu^\top\right)^2 = I + \frac{\sinh^2\omega}{\omega^2} uu^\top.$$

□

Now, it clear from the above formula that each e^B, with $B \in \mathfrak{p}$ is a Lorentz boost. Conversely, every Lorentz boost is the exponential of some $B \in \mathfrak{p}$, as shown below.

Proposition 5.17. *Every Lorentz boost,*

$$A = \begin{pmatrix} \sqrt{I + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

with $c = \sqrt{\|v\|^2 + 1}$, *is of the form* $A = e^B$, *for* $B \in \mathfrak{p}$, *i.e., for some* $B \in \mathfrak{so}(n,1)$ *of the form*

$$B = \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}.$$

Proof. We need to find some

$$B = \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}$$

solving the equation

$$\begin{pmatrix} \sqrt{I + \frac{\sinh^2\omega}{\omega^2} uu^\top} & \frac{\sinh\omega}{\omega} u \\ \frac{\sinh\omega}{\omega} u^\top & \cosh\omega \end{pmatrix} = \begin{pmatrix} \sqrt{I + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

with $\omega = \|u\|$ and $c = \sqrt{\|v\|^2 + 1}$. When $v = 0$, we have $A = I$, and the matrix $B = 0$ corresponding to $u = 0$ works. So, assume $v \neq 0$. In this case, $c > 1$. We have to solve the equation $\cosh\omega = c$, that is,

$$e^{2\omega} - 2ce^{\omega} + 1 = 0.$$

The roots of the corresponding algebraic equation $X^2 - 2cX + 1 = 0$ are

$$X = c \pm \sqrt{c^2 - 1}.$$

As $c > 1$, both roots are strictly positive, so we can solve for ω, say $\omega = \log(c + \sqrt{c^2 - 1}) \neq 0$. Then, $\sinh\omega \neq 0$, so we can solve the equation

$$\frac{\sinh\omega}{\omega} u = v,$$

which yields a $B \in \mathfrak{so}(n,1)$ of the right form with $A = e^B$. □

Remarks:

(1) It is easy to show that the eigenvalues of matrices

$$B = \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}$$

are 0, with multiplicity $n-1$, $\|u\|$ and $-\|u\|$. Eigenvectors are also easily determined.

(2) The matrices, $B \in \mathfrak{so}(n,1)$, of the form

$$B = \begin{pmatrix} 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \alpha \\ 0 & \cdots & \alpha & 0 \end{pmatrix}$$

are easily seen to form an abelian Lie subalgebra, $\mathfrak{a}$, of $\mathfrak{so}(n,1)$ (which means that for all $B, C \in \mathfrak{a}$, $[B,C] = 0$, i.e., $BC = CB$). One will easily check that for any $B \in \mathfrak{a}$, as above, we get

$$e^B = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \cosh\alpha & \sinh\alpha \\ 0 & \cdots & 0 & \sinh\alpha & \cosh\alpha \end{pmatrix}$$

The matrices of the form, e^B, with $B \in \mathfrak{a}$, form an abelian subgroup, A, of $\mathbf{SO}_0(n,1)$ isomorphic to $\mathbf{SO}_0(1,1)$. As we already know, the matrices, $B \in \mathfrak{so}(n,1)$, of the form

$$\begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix},$$

where B is skew-symmetric, form a Lie subalgebra, $\mathfrak{k}$, of $\mathfrak{so}(n,1)$. Clearly, $\mathfrak{k}$ is isomorphic to $\mathfrak{so}(n)$ and using the exponential, we get a subgroup, K, of $\mathbf{SO}_0(n,1)$ isomorphic to $\mathbf{SO}(n)$. It is also clear that $\mathfrak{k} \cap \mathfrak{a} = (0)$, but $\mathfrak{k} \oplus \mathfrak{a}$ is *not* equal to $\mathfrak{so}(n,1)$. What is the missing piece? Consider the matrices, $N \in \mathfrak{so}(n,1)$, of the form

$$N = \begin{pmatrix} 0 & -u & u \\ u^\top & 0 & 0 \\ u^\top & 0 & 0 \end{pmatrix},$$

where $u \in \mathbb{R}^{n-1}$. The reader should check that these matrices form an abelian Lie subalgebra, $\mathfrak{n}$, of $\mathfrak{so}(n,1)$ and that

$$\mathfrak{so}(n,1) = \mathfrak{k} \oplus \mathfrak{a} \oplus \mathfrak{n}.$$

This is the *Iwasawa decomposition* of the Lie algebra $\mathfrak{so}(n,1)$. Furthermore, the reader should check that every $N \in \mathfrak{n}$ is nilpotent; in fact, $N^3 = 0$. (It turns out that $\mathfrak{n}$ is a nilpotent Lie algebra, see Knapp [90]). The connected Lie subgroup of $\mathbf{SO}_0(n,1)$ associated with $\mathfrak{n}$ is denoted N and it can be shown that we have the *Iwasawa decomposition* of the Lie group $\mathbf{SO}_0(n,1)$:

$$\mathbf{SO}_0(n,1) = KAN.$$

It is easy to check that $[\mathfrak{a}, \mathfrak{n}] \subseteq \mathfrak{n}$, so $\mathfrak{a} \oplus \mathfrak{n}$ is a Lie subalgebra of $\mathfrak{so}(n,1)$ and $\mathfrak{n}$ is an ideal of $\mathfrak{a} \oplus \mathfrak{n}$. This implies that N is normal in the group corresponding to $\mathfrak{a} \oplus \mathfrak{n}$, so AN is a subgroup (in fact, solvable) of $\mathbf{SO}_0(n,1)$. For more on the Iwasawa decomposition, see Knapp [90]. Observe that the image, $\overline{\mathfrak{n}}$, of $\mathfrak{n}$ under the Cartan involution, θ, is the Lie subalgebra

$$\overline{\mathfrak{n}} = \left\{ \begin{pmatrix} 0 & u & u \\ -u^\top & 0 & 0 \\ u^\top & 0 & 0 \end{pmatrix} \mid u \in \mathbb{R}^{n-1} \right\}.$$

It is easy to see that the centralizer of $\mathfrak{a}$ is the Lie subalgebra

$$\mathfrak{m} = \left\{ \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} \in \mathrm{Mat}_{n+1,n+1}(\mathbb{R}) \mid B \in \mathfrak{so}(n-1) \right\}$$

and the reader should check that

$$\mathfrak{so}(n,1) = \mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n} \oplus \overline{\mathfrak{n}}.$$

We also have

$$[\mathfrak{m}, \mathfrak{n}] \subseteq \mathfrak{n},$$

so $\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$ is a subalgebra of $\mathfrak{so}(n,1)$. The group, M, associated with $\mathfrak{m}$ is isomorphic to $\mathbf{SO}(n-1)$ and it can be shown that $B = MAN$ is a subgroup of $\mathbf{SO}_0(n,1)$. In fact,

$$\mathbf{SO}_0(n,1)/(MAN) = KAN/MAN = K/M = \mathbf{SO}(n)/\mathbf{SO}(n-1) = S^{n-1}.$$

It is customary to denote the subalgebra $\mathfrak{m} \oplus \mathfrak{a}$ by $\mathfrak{g}_0$, the algebra $\mathfrak{n}$ by $\mathfrak{g}_1$ and $\overline{\mathfrak{n}}$ by $\mathfrak{g}_{-1}$, so that $\mathfrak{so}(n,1) = \mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n} \oplus \overline{\mathfrak{n}}$ is also written

$$\mathfrak{so}(n,1) = \mathfrak{g}_0 \oplus \mathfrak{g}_{-1} \oplus \mathfrak{g}_1.$$

By the way, if $N \in \mathfrak{n}$, then

$$e^N = I + N + \frac{1}{2}N^2,$$

and since $N + \frac{1}{2}N^2$ is also nilpotent, e^N can't be diagonalized when $N \neq 0$. This provides a simple example of matrices in $\mathbf{SO}_0(n,1)$ that can't be diagonalized.

Combining Proposition 2.3 and Proposition 5.17, we have the corollary:

Corollary 5.18. *Every matrix, $A \in \mathbf{O}(n,1)$, can be written as*

$$A = \begin{pmatrix} Q & 0 \\ 0 & \epsilon \end{pmatrix} e^{\begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}}$$

where $Q \in \mathbf{O}(n)$, $\epsilon = \pm 1$ and $u \in \mathbb{R}^n$.

Observe that Corollary 5.18 proves that every matrix, $A \in \mathbf{SO}_0(n,1)$, can be written as

$$A = Pe^S, \quad \text{with } P \in K \cong \mathbf{SO}(n) \text{ and } S \in \mathfrak{p},$$

i.e.,

$$\mathbf{SO}_0(n,1) = K\exp(\mathfrak{p}),$$

a version of the polar decomposition for $\mathbf{SO}_0(n,1)$.

Now, it is known that the exponential map, $\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$, is surjective. So, when $A \in \mathbf{SO}_0(n,1)$, since then $Q \in \mathbf{SO}(n)$ and $\epsilon = +1$, the matrix

$$\begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix}$$

is the exponential of some skew symmetric matrix

$$C = \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} \in \mathfrak{so}(n,1),$$

and we can write $A = e^C e^Z$, with $C \in \mathfrak{k}$ and $Z \in \mathfrak{p}$. Unfortunately, C and Z generally don't commute, so it is generally not true that $A = e^{C+Z}$. Thus, we don't get an "easy" proof of the surjectivity of the exponential, $\exp\colon \mathfrak{so}(n,1) \to \mathbf{SO}_0(n,1)$. This is not too surprising because, to the best of our knowledge, proving surjectivity for all n is not a simple matter. One proof is due to Nishikawa [119] (1983). Nishikawa's paper is rather short, but this is misleading. Indeed, Nishikawa relies on a classic paper by Djokovic [49], which itself relies heavily on another fundamental paper by Burgoyne and Cushman [27], published in 1977. Burgoyne and Cushman determine the conjugacy classes for some linear Lie groups and their Lie algebras, where the linear groups arise from an inner product space (real or complex). This inner product is nondegenerate, symmetric, or hermitian or skew-symmetric of skew-hermitian. Altogether, one has to read over 40 pages to fully understand the proof of surjectivity.

In his introduction, Nishikawa states that he is not aware of any other proof of the surjectivity of the exponential for $\mathbf{SO}_0(n,1)$. However, such a proof was also given by Marcel Riesz as early as 1957, in some lectures notes that he gave while visiting the University of Maryland in 1957-1958. These notes were probably not easily available until 1993, when they were published in book form, with commentaries, by Bolinder and Lounesto [127].

Interestingly, these two proofs use very different methods. The Nishikawa–Djokovic–Burgoyne and Cushman proof makes heavy use of methods in Lie groups and Lie algebra, although not far beyond linear algebra. Riesz's proof begins with a deep study of the structure of the minimal polynomial of a Lorentz isometry (Chapter III). This is a beautiful argument that takes about 10 pages. The story is not over, as it takes most of Chapter IV (some 40 pages) to prove the surjectivity of the exponential (actually, Riesz proves other things along the way). In any case, the reader can see that both proofs are quite involved.

It is worth noting that Milnor (1969) also uses techniques very similar to those used by Riesz (in dealing with minimal polynomials of isometries) in his paper on isometries of inner product spaces [108].

What we will do to close this section is to give a relatively simple proof that the exponential map, $\exp\colon \mathfrak{so}(1,3) \to \mathbf{SO}_0(1,3)$, is surjective. In the case of $\mathbf{SO}_0(1,3)$, we can use the fact that $\mathbf{SL}(2,\mathbb{C})$ is a two-sheeted covering space of $\mathbf{SO}_0(1,3)$, which means that there is a homomorphism, $\varphi\colon \mathbf{SL}(2,\mathbb{C}) \to \mathbf{SO}_0(1,3)$, which is surjective and that Ker $\varphi = \{-I, I)$. Then, the small miracle is that, although the exponential, $\exp\colon \mathfrak{sl}(2,\mathbb{C}) \to \mathbf{SL}(2,\mathbb{C})$, is *not* surjective, for every $A \in \mathbf{SL}(2,\mathbb{C})$, *either A or $-A$ is in the image of the exponential!*

Proposition 5.19. *Given any matrix*

$$B = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \mathfrak{sl}(2,\mathbb{C}),$$

let ω be any of the two complex roots of $a^2 + bc$. If $\omega \neq 0$, then

$$e^B = \cosh\omega\, I + \frac{\sinh\,\omega}{\omega}\, B,$$

and $e^B = I + B$ if $a^2 + bc = 0$. Furthermore, every matrix $A \in \mathbf{SL}(2,\mathbb{C})$ is in the image of the exponential map, unless $A = -I + N$, where N is a nonzero nilpotent (i.e., $N^2 = 0$ with $N \neq 0$). Consequently, for any $A \in \mathbf{SL}(2,\mathbb{C})$, either A or $-A$ is of the form e^B, for some $B \in \mathfrak{sl}(2,\mathbb{C})$.

Proof. Observe that

$$A^2 = \begin{pmatrix} a & b \\ c & -a \end{pmatrix}\begin{pmatrix} a & b \\ c & -a \end{pmatrix} = (a^2 + bc)I.$$

Then, it is straighforward to prove that

$$e^B = \cosh\omega\, I + \frac{\sinh\,\omega}{\omega}\, B,$$

where ω is a square root of $a^2 + bc$ is $\omega \neq 0$, otherwise, $e^B = I + B$.

Let

$$A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$$

be any matrix in $\mathbf{SL}(2,\mathbb{C})$. We would like to find a matrix, $B \in \mathfrak{sl}(2,\mathbb{C})$, so that $A = e^B$. In view of the above, we need to solve the system

$$\begin{aligned} \cosh\omega + \frac{\sinh\omega}{\omega}a &= \alpha \\ \cosh\omega - \frac{\sinh\omega}{\omega}a &= \delta \\ \frac{\sinh\omega}{\omega}b &= \beta \\ \frac{\sinh\omega}{\omega}c &= \gamma. \end{aligned}$$

From the first two equations, we get

$$\begin{aligned}\cosh\omega &= \frac{\alpha+\delta}{2}\\ \frac{\sinh\omega}{\omega}a &= \frac{\alpha-\delta}{2}.\end{aligned}$$

Thus, we see that we need to know whether complex cosh is surjective and when complex sinh is zero. We claim:

(1) cosh is surjective.

(2) $\sinh z = 0$ iff $z = n\pi i$, where $n \in \mathbb{Z}$.

Given any $c \in \mathbb{C}$, we have $\cosh\omega = c$ iff

$$e^{2\omega} - 2e^{\omega}c + 1 = 0.$$

The corresponding algebraic equation

$$Z^2 - 2cZ + 1 = 0$$

has discriminant $4(c^2-1)$ and it has two complex roots

$$Z = c \pm \sqrt{c^2-1}$$

where $\sqrt{c^2-1}$ is some square root of c^2-1. Observe that these roots are *never zero*. Therefore, we can find a complex log of $c+\sqrt{c^2-1}$, say ω, so that $e^{\omega} = c+\sqrt{c^2-1}$ is a solution of $e^{2\omega} - 2e^{\omega}c + 1 = 0$. This proves the surjectivity of cosh.

We have $\sinh\omega = 0$ iff $e^{2\omega} = 1$; this holds iff $2\omega = n2\pi i$, i.e., $\omega = n\pi i$.

Observe that

$$\frac{\sinh n\pi i}{n\pi i} = 0 \quad \text{if } n \neq 0, \text{ but} \quad \frac{\sinh n\pi i}{n\pi i} = 1 \quad \text{when } n = 0.$$

We know that

$$\cosh\omega = \frac{\alpha+\delta}{2}$$

can always be solved.

Case 1. If $\omega \neq n\pi i$, with $n \neq 0$, then

$$\frac{\sinh\omega}{\omega} \neq 0$$

and the other equations can also be solved (this includes the case $\omega = 0$). We still have to check that

$$a^2 + bc = \omega^2.$$

This is because, using the fact that $\cosh\omega = \frac{\alpha+\delta}{2}$, $\alpha\delta - \beta\gamma = 1$, and $\cosh^2\omega - \sinh^2\omega = 1$, we have

$$\begin{aligned} a^2 + bc &= \frac{(\alpha-\delta)^2\omega^2}{4\sinh^2\omega} + \frac{\beta\gamma\omega^2}{\sinh^2\omega} \\ &= \frac{\omega^2(\alpha^2+\delta^2-2\alpha\delta+4\beta\gamma)}{4\sinh^2\omega} \\ &= \frac{\omega^2(\alpha^2+\delta^2+2\alpha\delta-4(\alpha\delta-\beta\gamma))}{4\sinh^2\omega} \\ &= \frac{\omega^2((\alpha+\delta)^2-4(\alpha\delta-\beta\gamma))}{4\sinh^2\omega} \\ &= \frac{4\omega^2(\cosh^2\omega-1)}{4\sinh^2\omega} \\ &= \omega^2. \end{aligned}$$

Therefore, in this case, the exponential is surjective. It remains to examine the other case.

Case 2. Assume $\omega = n\pi i$, with $n \neq 0$. If n is even, then $e^\omega = 1$, which implies

$$\alpha + \delta = 2.$$

However, $\alpha\delta - \beta\gamma = 1$ (since $A \in \mathbf{SL}(2,\mathbb{C})$), so we deduce that A has the double eigenvalue, 1. Thus, $N = A - I$ is nilpotent (i.e., $N^2 = 0$) and has zero trace; but then, $N \in \mathfrak{sl}(2,\mathbb{C})$ and

$$e^N = I + N = I + A - I = A.$$

If n is odd, then $e^\omega = -1$, which implies

$$\alpha + \delta = -2.$$

In this case, A has the double eigenvalue -1 and $A + I = N$ is nilpotent. So, $A = -I + N$, where N is nilpotent. If $N \neq 0$, then A cannot be diagonalized. We claim that there is no $B \in \mathfrak{sl}(2,\mathbb{C})$ so that $e^B = A$.

Indeed, any matrix, $B \in \mathfrak{sl}(2,\mathbb{C})$, has zero trace, which means that if λ_1 and λ_2 are the eigenvalues of B, then $\lambda_1 = -\lambda_2$. If $\lambda_1 \neq 0$, then $\lambda_1 \neq \lambda_2$ so B can be diagonalized, but then e^B can also be diagonalized, contradicting the fact that A can't be diagonalized. If $\lambda_1 = \lambda_2 = 0$, then e^B has the double eigenvalue $+1$, but A has eigenvalues -1. Therefore, the only matrices $A \in \mathbf{SL}(2,\mathbb{C})$ that are not in the image of the exponential are those of the form $A = -I + N$, where N is a nonzero nilpotent. However, note that $-A = I - N$ *is* in the image of the exponential. □

Remark: If we restrict our attention to $\mathbf{SL}(2,\mathbb{R})$, then we have the following proposition that can be used to prove that the exponential map, $\exp\colon \mathfrak{so}(1,2) \to \mathbf{SO}_0(1,2)$, is surjective:

Proposition 5.20. *Given any matrix*

$$B = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \mathfrak{sl}(2,\mathbb{R}),$$

if $a^2 + bc > 0$, then let $\omega = \sqrt{a^2+bc} > 0$, and if $a^2 + bc < 0$, then let $\omega = \sqrt{-(a^2+bc)} > 0$ (i.e., $\omega^2 = -(a^2+bc)$). In the first case ($a^2 + bc > 0$), we have

$$e^B = \cosh\omega\, I + \frac{\sinh\,\omega}{\omega}\, B,$$

and in the second case ($a^2 + bc < 0$), we have

$$e^B = \cos\omega\, I + \frac{\sin\,\omega}{\omega}\, B.$$

If $a^2 + bc = 0$, then $e^B = I + B$. Furthermore, every matrix $A \in \mathbf{SL}(2,\mathbb{R})$ whose trace satisfies $\mathrm{tr}(A) \geq -2$ is in the image of the exponential map, unless $A = -I + N$ with $N \neq 0$ nilpotent. Consequently, for any $A \in \mathbf{SL}(2,\mathbb{R})$, either A or $-A$ is of the form e^B, for some $B \in \mathfrak{sl}(2,\mathbb{R})$.

Proof. For any matrix

$$B = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \mathfrak{sl}(2,\mathbb{R}),$$

some simple calcuations show that if $a^2 + bc > 0$, then

$$e^B = \cosh\omega\, I + \frac{\sinh\,\omega}{\omega}\, B$$

with $\omega = \sqrt{a^2+bc} > 0$, and if $a^2 + bc < 0$, then

$$e^B = \cos\omega\, I + \frac{\sin\,\omega}{\omega}\, B$$

with $\omega = \sqrt{-(a^2+bc)} > 0$ (and $e^B = I + B$ when $a^2 + bc = 0$). Let

$$A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$$

be any matrix in $\mathbf{SL}(2,\mathbb{R})$.

First, assume that $\alpha + \delta > 2$. We would like to find a matrix, $B \in \mathfrak{sl}(2,\mathbb{R})$, so that $A = e^B$. In view of the above, we need to solve the system

$$\begin{aligned} \cosh\omega + \frac{\sinh\omega}{\omega} a &= \alpha \\ \cosh\omega - \frac{\sinh\omega}{\omega} a &= \delta \\ \frac{\sinh\omega}{\omega} b &= \beta \\ \frac{\sinh\omega}{\omega} c &= \gamma. \end{aligned}$$

From the first two equations, we get

$$\begin{aligned}\cosh\omega &= \frac{\alpha+\delta}{2}\\ \frac{\sinh\omega}{\omega}a &= \frac{\alpha-\delta}{2}.\end{aligned}$$

As in the proof of Proposition 5.19, $\cosh\omega = c$ iff e^ω is a root of the quadratic equation

$$Z^2 - 2cZ + 1 = 0.$$

This equation has a real roots iff $c^2 \geq 1$, and since $c = \frac{\alpha+\delta}{2}$ and $\alpha+\delta > 2$, our equation has real roots. Furthermore, the root $c+\sqrt{c^2-1}$ is greater than 1, so $\log c$ is a positive real number. Then, as in the proof of Proposition 5.19, we find solutions of our system above. Moreover, these solutions are real and satisfy $a^2 + bc = \omega^2$.

Let us now consider the case where $-2 \leq \alpha+\delta \leq 2$. This time, we try to solve the system

$$\begin{aligned}\cos\omega + \frac{\sin\omega}{\omega}a &= \alpha\\ \cos\omega - \frac{\sin\omega}{\omega}a &= \delta\\ \frac{\sin\omega}{\omega}b &= \beta\\ \frac{\sin\omega}{\omega}c &= \gamma.\end{aligned}$$

We get

$$\begin{aligned}\cos\omega &= \frac{\alpha+\delta}{2}\\ \frac{\sin\omega}{\omega}a &= \frac{\alpha-\delta}{2}.\end{aligned}$$

Because $-2 \leq \alpha+\delta \leq 2$, the first equation has (real) solutions, and we may assume that $0 \leq \omega < 2\pi$.

If $\omega = 0$ is a solution, then $\alpha+\beta = 2$ and we already know that $N = A - I$ is nilpotent and that $e^N = I + N = A$. If $\omega = \pi$, then $\alpha+\beta = -2$ and we know that $N = A + I$ is nilpotent. If $N = 0$, then $A = -I$, and otherwise we already know that $A = -I + N$ is not in the image of the exponential.

If $0 < \omega < \pi$, then $\sin\omega \neq 0$ and the other equations have a solution. We still need to check that

$$a^2 + bc = -\omega^2.$$

Because $\cos\omega = \frac{\alpha+\delta}{2}$, $\alpha\delta - \beta\gamma = 1$, and $\cos^2\omega + \sin^2\omega = 1$, we have

$$\begin{aligned} a^2 + bc &= \frac{(\alpha-\delta)^2\omega^2}{4\sin^2\omega} + \frac{\beta\gamma\omega^2}{\sin^2\omega} \\ &= \frac{\omega^2(\alpha^2+\delta^2-2\alpha\delta+4\beta\gamma)}{4\sinh^2\omega} \\ &= \frac{\omega^2(\alpha^2+\delta^2+2\alpha\delta-4(\alpha\delta-\beta\gamma))}{4\sin^2\omega} \\ &= \frac{\omega^2((\alpha+\delta)^2-4(\alpha\delta-\beta\gamma))}{4\sin^2\omega} \\ &= \frac{4\omega^2(\cos^2\omega-1)}{4\sin^2\omega} \\ &= -\omega^2. \end{aligned}$$

This proves that every matrix $A \in \mathbf{SL}(2,\mathbb{R})$ whose trace satisfies $\mathrm{tr}(A) \geq -2$ is in the image of the exponential map, unless $A = -I + N$ with $N \neq 0$ nilpotent. □

We now return to the relationship between $\mathbf{SL}(2,\mathbb{C})$ and $\mathbf{SO}_0(1,3)$. In order to define a homomorphism, $\varphi\colon \mathbf{SL}(2,\mathbb{C}) \to \mathbf{SO}_0(1,3)$, we begin by defining a linear bijection, h, between $\mathbb{R}^4$ and $\mathbf{H}(2)$, the set of complex 2×2 Hermitian matrices, by

$$(t,x,y,z) \mapsto \begin{pmatrix} t+x & y-iz \\ y+iz & t-x \end{pmatrix}.$$

Those familiar with quantum physics will recognize a linear combination of the Pauli matrices! The inverse map is easily defined and we leave it as an exercise. For instance, given a Hermitian matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

we have

$$t = \frac{a+d}{2},\ x = \frac{a-d}{2}, \quad \text{etc.}$$

Next, for any $A \in \mathbf{SL}(2,\mathbb{C})$, we define a map, $l_A\colon \mathbf{H}(2) \to \mathbf{H}(2)$, *via*

$$S \mapsto ASA^*.$$

(Here, $A^* = \overline{A}^\top$.) Using the linear bijection, $h\colon \mathbb{R}^4 \to \mathbf{H}(2)$, and its inverse, we obtain a map, $\mathrm{lor}_A\colon \mathbb{R}^4 \to \mathbb{R}^4$, where

$$\mathrm{lor}_A = h^{-1} \circ l_A \circ h.$$

As ASA^* is hermitian, we see that l_A is well defined. It is obviously linear and since $\det(A) = 1$ (recall, $A \in \mathbf{SL}(2,\mathbb{C})$) and

$$\det\begin{pmatrix} t+x & y-iz \\ y+iz & t-x \end{pmatrix} = t^2 - x^2 - y^2 - z^2,$$

we see that lor_A preserves the Lorentz metric! Furthermore, it is not hard to prove that $\mathbf{SL}(2,\mathbb{C})$ is connected (use the polar form or analyze the eigenvalues of a matrix in $\mathbf{SL}(2,\mathbb{C})$, for example, as in Duistermatt and Kolk [54] (Chapter 1, Section 1.2)) and that the map

$$\varphi\colon A \mapsto \mathrm{lor}_A$$

is a continuous group homomorphism. Thus, the range of φ is a connected subgroup of $\mathbf{SO}_0(1,3)$. This shows that $\varphi\colon \mathbf{SL}(2,\mathbb{C}) \to \mathbf{SO}_0(1,3)$ is indeed a homomorphism. It remains to prove that it is surjective and that its kernel is $\{I, -I\}$.

Proposition 5.21. *The homomorphism, $\varphi\colon \mathbf{SL}(2,\mathbb{C}) \to \mathbf{SO}_0(1,3)$, is surjective and its kernel is $\{I, -I\}$.*

Proof. Recall that from Theorem 2.6, the Lorentz group $\mathbf{SO}_0(1,3)$ is generated by the matrices of the form

$$\begin{pmatrix} 1 & 0 \\ 0 & P \end{pmatrix} \quad \text{with } P \in \mathbf{SO}(3)$$

and the matrices of the form

$$\begin{pmatrix} \cosh\alpha & \sinh\alpha & 0 & 0 \\ \sinh\alpha & \cosh\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Thus, to prove the surjectivity of φ, it is enough to check that the above matrices are in the range of φ. For matrices of the second kind, the reader should check that

$$A = \begin{pmatrix} e^{\frac{1}{2}\alpha} & 0 \\ 0 & e^{-\frac{1}{2}\alpha} \end{pmatrix}$$

does the job. For matrices of the first kind, we recall that the group of unit quaternions, $q = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$, can be viewed as $\mathbf{SU}(2)$, *via* the correspondence

$$a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} \mapsto \begin{pmatrix} a+ib & c+id \\ -c+id & a-ib \end{pmatrix},$$

where $a, b, c, d \in \mathbb{R}$ and $a^2 + b^2 + c^2 + d^2 = 1$. Moreover, the algebra of quaternions, $\mathbb{H}$, is the real algebra of matrices as above, without the restriction $a^2 + b^2 + c^2 + d^2 = 1$ and $\mathbb{R}^3$ is embedded in $\mathbb{H}$ as the *pure quaternions*, i.e., those for which $a = 0$. Observe that when $a = 0$,

$$\begin{pmatrix} ib & c+id \\ -c+id & -ib \end{pmatrix} = i\begin{pmatrix} b & d-ic \\ d+ic & -b \end{pmatrix} = ih(0,b,d,c).$$

Therefore, we have a bijection between the pure quaternions and the subspace of the hermitian matrices

$$\begin{pmatrix} b & d-ic \\ d+ic & -b \end{pmatrix}$$

for which $a = 0$, the inverse being division by i, i.e., multiplication by $-i$. Also, when q is a unit quaternion, let $\overline{q} = a\mathbf{1} - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}$, and observe that $\overline{q} = q^{-1}$. Using the embedding $\mathbb{R}^3 \hookrightarrow \mathbb{H}$, for every unit quaternion, $q \in \mathbf{SU}(2)$, define the map, $\rho_q \colon \mathbb{R}^3 \to \mathbb{R}^3$, by

$$\rho_q(X) = qX\overline{q} = qXq^{-1},$$

for all $X \in \mathbb{R}^3 \hookrightarrow \mathbb{H}$. Then, it is well known that ρ_q is a rotation (i.e., $\rho_q \in \mathbf{SO}(3)$) and, moreover, the map $q \mapsto \rho_q$, is a surjective homomorphism, $\rho \colon \mathbf{SU}(2) \to \mathbf{SO}(3)$, and Ker $\varphi = \{I, -I\}$ (For example, see Gallier [60], Chapter 8).

Now, consider a matrix, A, of the form

$$\begin{pmatrix} 1 & 0 \\ 0 & P \end{pmatrix} \quad \text{with } P \in \mathbf{SO}(3).$$

We claim that we can find a matrix, $B \in \mathbf{SL}(2,\mathbb{C})$, such that $\varphi(B) = \mathrm{lor}_B = A$. We claim that we can pick $B \in \mathbf{SU}(2) \subseteq \mathbf{SL}(2,\mathbb{C})$. Indeed, if $B \in \mathbf{SU}(2)$, then $B^* = B^{-1}$, so

$$B \begin{pmatrix} t+x & y-iz \\ y+iz & t-x \end{pmatrix} B^* = t \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - iB \begin{pmatrix} ix & z+iy \\ -z+iy & -ix \end{pmatrix} B^{-1}.$$

The above shows that lor_B leaves the coordinate t invariant. The term

$$B \begin{pmatrix} ix & z+iy \\ -z+iy & -ix \end{pmatrix} B^{-1}$$

is a pure quaternion corresponding to the application of the rotation ρ_B induced by the quaternion B to the pure quaternion associated with (x, y, z) and multiplication by $-i$ is just the corresponding hermitian matrix, as explained above. But, we know that for any $P \in \mathbf{SO}(3)$, there is a quaternion, B, so that $\rho_B = P$, so we can find our $B \in \mathbf{SU}(2)$ so that

$$\mathrm{lor}_B = \begin{pmatrix} 1 & 0 \\ 0 & P \end{pmatrix} = A.$$

Finally, assume that $\varphi(A) = I$. This means that

$$ASA^* = S,$$

for all hermitian matrices, S, defined above. In particular, for $S = I$, we get $AA^* = I$, i.e., $A \in \mathbf{SU}(2)$. We have

$$AS = SA$$

for all hermitian matrices, S, defined above, so in particular, this holds for diagonal matrices of the form

$$\begin{pmatrix} t+x & 0 \\ 0 & t-x \end{pmatrix},$$

with $t + x \neq t - x$. We deduce that A is a diagonal matrix, and since it is unitary, we must have $A = \pm I$. Therefore, Ker $\varphi = \{I, -I\}$. □

Remark: The group $\mathbf{SL}(2,\mathbb{C})$ is isomorphic to the group $\mathbf{Spin}(1,3)$, which is a (simply-connected) double-cover of $\mathbf{SO}_0(1,3)$. This is a standard result of Clifford algebra theory, see Bröcker and tom Dieck [25] or Fulton and Harris [58]. What we just did is to provide a direct proof of this fact.

We just proved that there is an isomorphism

$$\mathbf{SL}(2,\mathbb{C})/\{I,-I\} \cong \mathbf{SO}_0(1,3).$$

However, the reader may recall that $\mathbf{SL}(2,\mathbb{C})/\{I,-I\} = \mathbf{PSL}(2,\mathbb{C}) \cong \mathbf{M\ddot{o}b}^+$. Therefore, the Lorentz group is isomorphic to the Möbius group.

We now have all the tools to prove that the exponential map, $\exp\colon \mathfrak{so}(1,3) \to \mathbf{SO}_0(1,3)$, is surjective.

Theorem 5.22. *The exponential map,* $\exp\colon \mathfrak{so}(1,3) \to \mathbf{SO}_0(1,3)$*, is surjective.*

Proof. First, recall from Proposition 5.4 that the following diagram commutes:

$$\begin{array}{ccc} \mathbf{SL}(2,\mathbb{C}) & \xrightarrow{\varphi} & \mathbf{SO}_0(1,3)\,. \\ {\scriptstyle\exp}\uparrow & & \uparrow{\scriptstyle\exp} \\ \mathfrak{sl}(2,\mathbb{C}) & \xrightarrow[d\varphi_1]{} & \mathfrak{so}(1,3) \end{array}$$

Pick any $A \in \mathbf{SO}_0(1,3)$. By Proposition 5.21, the homomorphism φ is surjective and as $\operatorname{Ker}\varphi = \{I,-I\}$, there exists some $B \in \mathbf{SL}(2,\mathbb{C})$ so that

$$\varphi(B) = \varphi(-B) = A.$$

Now, by Proposition 5.19, for any $B \in \mathbf{SL}(2,\mathbb{C})$, either B or $-B$ is of the form e^C, for some $C \in \mathfrak{sl}(2,\mathbb{C})$. By the commutativity of the diagram, if we let $D = d\varphi_1(C) \in \mathfrak{so}(1,3)$, we get

$$A = \varphi(\pm e^C) = e^{d\varphi_1(C)} = e^D,$$

with $D \in \mathfrak{so}(1,3)$, as required. □

Remark: We can restrict the bijection, $h\colon \mathbb{R}^4 \to \mathbf{H}(2)$, defined earlier to a bijection between $\mathbb{R}^3$ and the space of real symmetric matrices of the form

$$\begin{pmatrix} t+x & y \\ y & t-x \end{pmatrix}.$$

Then, if we also restrict ourselves to $\mathbf{SL}(2,\mathbb{R})$, for any $A \in \mathbf{SL}(2,\mathbb{R})$ and any symmetric matrix, S, as above, we get a map

$$S \mapsto ASA^\top.$$

The reader should check that these transformations correspond to isometries in $\mathbf{SO}_0(1,2)$ and we get a homomorphism, $\varphi\colon \mathbf{SL}(2,\mathbb{R}) \to \mathbf{SO}_0(1,2)$. Just as $\mathbf{SL}(2,\mathbb{C})$ is connected, the group $\mathbf{SL}(2,\mathbb{R})$ is also connected (but not simply connected, unlike $\mathbf{SL}(2,\mathbb{C})$). Then, we have a version of Proposition 5.21 for $\mathbf{SL}(2,\mathbb{R})$ and $\mathbf{SO}_0(1,2)$:

Proposition 5.23. *The homomorphism, $\varphi\colon \mathbf{SL}(2,\mathbb{R}) \to \mathbf{SO}_0(1,2)$, is surjective and its kernel is $\{I,-I\}$.*

Using Proposition 5.23, Proposition 5.20, and the commutative diagram

$$\begin{array}{ccc} \mathbf{SL}(2,\mathbb{R}) & \xrightarrow{\varphi} & \mathbf{SO}_0(1,2)\,, \\ {\scriptstyle\exp}\uparrow & & \uparrow{\scriptstyle\exp} \\ \mathfrak{sl}(2,\mathbb{R}) & \xrightarrow[d\varphi_1]{} & \mathfrak{so}(1,2) \end{array}$$

we get a version of Theorem 5.22 for $\mathbf{SO}_0(1,2)$:

Theorem 5.24. *The exponential map, $\exp\colon \mathfrak{so}(1,2) \to \mathbf{SO}_0(1,2)$, is surjective.*

Also observe that $\mathbf{SO}_0(1,1)$ consists of the matrices of the form

$$A = \begin{pmatrix} \cosh\alpha & \sinh\alpha \\ \sinh\alpha & \cosh\alpha \end{pmatrix}$$

and a direct computation shows that

$$e^{\begin{pmatrix} 0 & \alpha \\ \alpha & 0 \end{pmatrix}} = \begin{pmatrix} \cosh\alpha & \sinh\alpha \\ \sinh\alpha & \cosh\alpha \end{pmatrix}.$$

Thus, we see that the map $\exp\colon \mathfrak{so}(1,1) \to \mathbf{SO}_0(1,1)$ is also surjective. Therefore, we have proved that $\exp\colon \mathfrak{so}(1,n) \to \mathbf{SO}_0(1,n)$ is surjective for $n = 1,2,3$. This actually holds for all $n \geq 1$, but the proof is much more involved, as we already discussed earlier.

5.6 More on the Topology of $\mathbf{O}(p,q)$ and $\mathbf{SO}(p,q)$

It turns out that the topology of the group, $\mathbf{O}(p,q)$, is completely determined by the topology of $\mathbf{O}(p)$ and $\mathbf{O}(q)$. This result can be obtained as a simple consequence of some standard Lie group theory. The key notion is that of a pseudo-algebraic group.

Consider the group, $\mathbf{GL}(n,\mathbb{C})$, of invertible $n \times n$ matrices with complex coefficients. If $A = (a_{kl})$ is such a matrix, denote by x_{kl} the real part (resp. y_{kl}, the imaginary part) of a_{kl} (so, $a_{kl} = x_{kl} + iy_{kl}$).

Definition 5.9. A subgroup, G, of $\mathbf{GL}(n,\mathbb{C})$ is *pseudo-algebraic* iff there is a finite set of polynomials in $2n^2$ variables with real coefficients, $\{P_i(X_1,\ldots,X_{n^2},Y_1,\ldots,Y_{n^2})\}_{i=1}^t$, so that

$$A = (x_{kl} + iy_{kl}) \in G \quad \text{iff} \quad P_i(x_{11},\ldots,x_{nn},y_{11},\ldots,y_{nn}) = 0, \quad \text{for } i = 1,\ldots,t.$$

Recall that if A is a complex $n \times n$-matrix, its *adjoint*, A^*, is defined by $A^* = (\overline{A})^\top$. Also, $\mathbf{U}(n)$ denotes the group of unitary matrices, i.e., those matrices, $A \in \mathbf{GL}(n, \mathbb{C})$, so that $AA^* = A^*A = I$, and $\mathbf{H}(n)$ denotes the vector space of Hermitian matrices, i.e., those matrices, A, so that $A^* = A$. Then, we have the following theorem which is essentially a refined version of the polar decomposition of matrices:

Theorem 5.25. *Let G be a pseudo-algebraic subgroup of $\mathbf{GL}(n, \mathbb{C})$ stable under adjunction (i.e., we have $A^* \in G$ whenever $A \in G$). Then, there is some integer, $d \in \mathbb{N}$, so that G is homeomorphic to $(G \cap \mathbf{U}(n)) \times \mathbb{R}^d$. Moreover, if $\mathfrak{g}$ is the Lie algebra of G, the map*

$$(\mathbf{U}(n) \cap G) \times (\mathbf{H}(n) \cap \mathfrak{g}) \longrightarrow G, \quad \text{given by} \quad (U, H) \mapsto Ue^H,$$

is a homeomorphism onto G.

Proof. A proof can be found in Knapp [90], Chapter 1, or Mneimné and Testard [112], Chapter 3. □

We now apply Theorem 5.25 to determine the structure of the space $\mathbf{O}(p, q)$. We know that $\mathbf{O}(p, q)$ consists of the matrices, A, in $\mathbf{GL}(p + q, \mathbb{R})$ such that

$$A^\top I_{p,q} A = I_{p,q},$$

and so, $\mathbf{O}(p, q)$ is clearly pseudo-algebraic. Using the above equation, it is easy to determine the Lie algebra, $\mathfrak{o}(p, q)$, of $\mathbf{O}(p, q)$. We find that $\mathfrak{o}(p, q)$ is given by

$$\mathfrak{o}(p, q) = \left\{ \begin{pmatrix} X_1 & X_2 \\ X_2^\top & X_3 \end{pmatrix} \;\middle|\; X_1^\top = -X_1,\ X_3^\top = -X_3,\ X_2 \text{ arbitrary} \right\}$$

where X_1 is a $p \times p$ matrix, X_3 is a $q \times q$ matrix and X_2 is a $p \times q$ matrix. Consequently, it immediately follows that

$$\mathfrak{o}(p, q) \cap \mathbf{H}(p + q) = \left\{ \begin{pmatrix} 0 & X_2 \\ X_2^\top & 0 \end{pmatrix} \;\middle|\; X_2 \text{ arbitrary} \right\},$$

a vector space of dimension pq.

Some simple calculations also show that

$$\mathbf{O}(p, q) \cap \mathbf{U}(p + q) = \left\{ \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \;\middle|\; X_1 \in \mathbf{O}(p),\ X_2 \in \mathbf{O}(q) \right\} \cong \mathbf{O}(p) \times \mathbf{O}(q).$$

Therefore, we obtain the structure of $\mathbf{O}(p, q)$:

Proposition 5.26. *The topological space $\mathbf{O}(p, q)$ is homeomorphic to $\mathbf{O}(p) \times \mathbf{O}(q) \times \mathbb{R}^{pq}$.*

Since $\mathbf{O}(p)$ has two connected components when $p \geq 1$, we see that $\mathbf{O}(p,q)$ has four connected components when $p, q \geq 1$. It is also obvious that

$$\mathbf{SO}(p,q) \cap \mathbf{U}(p+q) = \left\{ \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \;\middle|\; X_1 \in \mathbf{O}(p),\ X_2 \in \mathbf{O}(q),\ \det(X_1)\det(X_2) = 1 \right\}.$$

This is a subgroup of $\mathbf{O}(p) \times \mathbf{O}(q)$ that we denote $S(\mathbf{O}(p) \times \mathbf{O}(q))$. Furthermore, it is easy to show that $\mathfrak{so}(p,q) = \mathfrak{o}(p,q)$. Thus, we also have

Proposition 5.27. *The topological space $\mathbf{SO}(p,q)$ is homeomorphic to $S(\mathbf{O}(p)\times\mathbf{O}(q))\times\mathbb{R}^{pq}$.*

Observe that the dimension of all these spaces depends only on $p+q$: It is $(p+q)(p+q-1)/2$. Also, $\mathbf{SO}(p,q)$ has two connected components when $p, q \geq 1$. The connected component of I_{p+q} is the group $\mathbf{SO}_0(p,q)$. This latter space is homeomorphic to $\mathbf{SO}(p) \times \mathbf{SO}(q) \times \mathbb{R}^{pq}$.

Theorem 5.25 gives the polar form of a matrix $A \in \mathbf{O}(p,q)$: We have

$$A = Ue^S, \quad \text{with} \quad U \in \mathbf{O}(p) \times \mathbf{O}(q) \quad \text{and} \quad S \in \mathfrak{so}(p,q) \cap \mathbf{S}(p+q),$$

where U is of the form

$$U = \begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix}, \quad \text{with} \quad P \in \mathbf{O}(p) \quad \text{and} \quad Q \in \mathbf{O}(q)$$

and $\mathfrak{so}(p,q) \cap \mathbf{S}(p+q)$ consists of all $(p+q) \times (p+q)$ symmetric matrices of the form

$$S = \begin{pmatrix} 0 & X \\ X^\top & 0 \end{pmatrix},$$

with X an arbitrary $p \times q$ matrix. It turns out that it is not very hard to compute explicitly the exponential, e^S, of such matrices (see Mneimné and Testard [112]). Recall that the functions cosh and sinh also make sense for matrices (since the exponential makes sense) and are given by

$$\cosh(A) = \frac{e^A + e^{-A}}{2} = I + \frac{A^2}{2!} + \cdots + \frac{A^{2k}}{(2k)!} + \cdots$$

and

$$\sinh(A) = \frac{e^A - e^{-A}}{2} = A + \frac{A^3}{3!} + \cdots + \frac{A^{2k+1}}{(2k+1)!} + \cdots.$$

We also set

$$\frac{\sinh(A)}{A} = I + \frac{A^2}{3!} + \cdots + \frac{A^{2k}}{(2k+1)!} + \cdots,$$

which is defined for all matrices, A (even when A is singular). Then, we have

Proposition 5.28. *For any matrix S of the form*

$$S = \begin{pmatrix} 0 & X \\ X^\top & 0 \end{pmatrix},$$

we have

$$e^S = \begin{pmatrix} \cosh((XX^\top)^{\frac{1}{2}}) & \frac{\sinh((XX^\top)^{\frac{1}{2}})X}{(XX^\top)^{\frac{1}{2}}} \\ \frac{\sinh((X^\top X)^{\frac{1}{2}})X^\top}{(X^\top X)^{\frac{1}{2}}} & \cosh((X^\top X)^{\frac{1}{2}}) \end{pmatrix}.$$

Proof. By induction, it is easy to see that

$$S^{2k} = \begin{pmatrix} (XX^\top)^k & 0 \\ 0 & (X^\top X)^k \end{pmatrix}$$

and

$$S^{2k+1} = \begin{pmatrix} 0 & (XX^\top)^k X \\ (X^\top X)^k X^\top & 0 \end{pmatrix}.$$

The rest is left as an exercise. □

Remark: Although at first glance, e^S does not look symmetric, but it is!

As a consequence of Proposition 5.28, every matrix, $A \in \mathbf{O}(p,q)$, has the polar form

$$A = \begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix} \begin{pmatrix} \cosh((XX^\top)^{\frac{1}{2}}) & \frac{\sinh((XX^\top)^{\frac{1}{2}})X}{(XX^\top)^{\frac{1}{2}}} \\ \frac{\sinh((X^\top X)^{\frac{1}{2}})X^\top}{(X^\top X)^{\frac{1}{2}}} & \cosh((X^\top X)^{\frac{1}{2}}) \end{pmatrix},$$

with $P \in \mathbf{O}(p)$, $Q \in \mathbf{O}(q)$ and X an arbitrary $p \times q$ matrix.

5.7 Universal Covering Groups

Every connected Lie group, G, is a manifold and, as such, from results in Section 3.9, it has a universal cover, $\pi\colon \widetilde{G} \to G$, where $\widetilde{G}$ is simply connected. It is possible to make $\widetilde{G}$ into a group so that $\widetilde{G}$ is a Lie group and π is a Lie group homomorphism. We content ourselves with a sketch of the construction whose details can be found in Warner [148], Chapter 3.

Consider the map, $\alpha\colon \widetilde{G} \times \widetilde{G} \to G$, given by

$$\alpha(\widetilde{a}, \widetilde{b}) = \pi(\widetilde{a})\pi(\widetilde{b})^{-1},$$

for all $\widetilde{a}, \widetilde{b} \in \widetilde{G}$, and pick some $\widetilde{e} \in \pi^{-1}(e)$. Since $\widetilde{G} \times \widetilde{G}$ is simply connected, it follows by Proposition 3.40 that there is a unique map, $\widetilde{\alpha}\colon \widetilde{G} \times \widetilde{G} \to \widetilde{G}$, such that

$$\alpha = \pi \circ \widetilde{\alpha} \quad \text{and} \quad \widetilde{e} = \widetilde{\alpha}(\widetilde{e}, \widetilde{e}).$$

For all $\widetilde{a}, \widetilde{b} \in \widetilde{G}$, define

$$\widetilde{b}^{-1} = \widetilde{\alpha}(\widetilde{e}, \widetilde{b}), \qquad \widetilde{a}\widetilde{b} = \widetilde{\alpha}(\widetilde{a}, \widetilde{b}^{-1}). \tag{$*$}$$

Using Proposition 3.40, it can be shown that the above operations make $\widetilde{G}$ into a group and as $\widetilde{\alpha}$ is smooth, into a Lie group. Moreover, π becomes a Lie group homomorphism. We summarize these facts as

Theorem 5.29. *Every connected Lie group has a simply connected covering map, $\pi\colon \widetilde{G} \to G$, where $\widetilde{G}$ is a Lie group and π is a Lie group homomorphism.*

The group, $\widetilde{G}$, is called the *universal covering group* of G. Consider $D = \ker \pi$. Since the fibres of π are countable The group D is a countable closed normal subgroup of $\widetilde{G}$, that is, a discrete normal subgroup of $\widetilde{G}$. It follows that $G \cong \widetilde{G}/D$, where $\widetilde{G}$ is a simply connected Lie group and D is a discrete normal subgroup of $\widetilde{G}$.

We conclude this section by stating the following useful proposition whose proof can be found in Warner [148] (Chapter 3, Proposition 3.26):

Proposition 5.30. *Let $\varphi\colon G \to H$ be a homomorphism of connected Lie groups. Then φ is a covering map iff $d\varphi_e\colon \mathfrak{g} \to \mathfrak{h}$ is an isomorphism of Lie algebras.*

For example, we know that $\mathfrak{su}(2) = \mathfrak{so}(3)$, so the homomorphism from $\mathbf{SU}(2)$ to $\mathbf{SO}(3)$ provided by the representation of 3D rotations by the quaternions is a covering map.

Chapter 6

The Derivative of exp and Dynkin's Formula

6.1 The Derivative of the Exponential Map

We know that if $[X, Y] = 0$, then $\exp(X + Y) = \exp(X)\exp(Y)$, but this generally false if X and Y do not commute. For X and Y in a small enough open subset, U, containing 0, we know that exp is a diffeomorphism from U to its image, so the function, $\mu\colon U \times U \to U$, given by

$$\mu(X, Y) = \log(\exp(X)\exp(Y))$$

is well-defined and it turns out that, for U small enough, it is analytic. Thus, it is natural to seek a formula for the Taylor expansion of μ near the origin. This problem was investigated by Campbell (1897/98), Baker (1905) and in a more rigorous fashion by Hausdorff (1906). These authors gave recursive identities expressing the Taylor expansion of μ at the origin and the corresponding result is often referred to as the *Campbell-Baker-Hausdorff Formula*. F. Schur (1891) and Poincaré (1899) also investigated the exponential map, in particular formulae for its derivative and the problem of expressing the function μ. However, it was Dynkin who finally gave an explicit formula (see Section 6.3) in 1947.

The proof that μ is analytic in a suitable domain can be proved using a formula for the derivative of the exponential map, a formula that was obtained by F. Schur and Poincaré. Thus, we begin by presenting such a formula.

First, we introduce a convenient notation. If A is any real (or complex) $n \times n$ matrix, the following formula is clear:

$$\int_0^1 e^{tA} dt = \sum_{k=0}^{\infty} \frac{A^k}{(k+1)!}.$$

If A is invertible, then the right-hand side can be written explicitly as

$$\sum_{k=0}^{\infty} \frac{A^k}{(k+1)!} = A^{-1}(e^A - I),$$

and we also write the latter as

$$\frac{e^A - I}{A} = \sum_{k=0}^{\infty} \frac{A^k}{(k+1)!}. \tag{$*$}$$

Even if A is not invertible, we use $(*)$ as the definition of $\frac{e^A - I}{A}$.

We can use the following trick to figure out what $(d_X \exp)(Y)$ is:

$$(d_X \exp)(Y) = \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \exp(X + \epsilon Y) = \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} dR_{\exp(X+\epsilon Y)}(1),$$

since by Proposition 5.2, the map, $s \mapsto R_{\exp s(X+\epsilon Y)}$ is the flow of the left-invariant vector field $(X + \epsilon Y)^L$ on G. Now, $(X + \epsilon Y)^L$ is an ϵ-dependent vector field which depends on ϵ in a C^1 fashion. From the theory of ODE's, if $p \mapsto v_\epsilon(p)$ is a smooth vector field depending in a C^1 fashion on a real parameter ϵ and if Φ_t^ϵ denotes its flow (after time), then the map $\epsilon \mapsto \Phi_t^\epsilon$ is differentiable and we have

$$\frac{\partial \Phi_t^\epsilon}{\partial \epsilon}(x) = \int_0^t d_{\Phi_t^\epsilon(x)}(\Phi_{t-s}^\epsilon) \frac{\partial v_\epsilon}{\partial \epsilon}(\Phi_s^\epsilon(x)) ds.$$

See Duistermaat and Kolk [54], Appendix B, Formula (B.10). Using this, the following is proved in Duistermaat and Kolk [54] (Chapter 1, Section 1.5):

Proposition 6.1. *Given any Lie group, G, for any $X \in \mathfrak{g}$, the linear map, $d\exp_X \colon \mathfrak{g} \to T_{\exp(X)}G$, is given by*

$$\begin{aligned} d\exp_X &= (dR_{\exp(X)})_1 \circ \int_0^1 e^{s \operatorname{ad} X} ds = (dR_{\exp(X)})_1 \circ \frac{e^{\operatorname{ad} X} - I}{\operatorname{ad} X} \\ &= (dL_{\exp(X)})_1 \circ \int_0^1 e^{-s \operatorname{ad} X} ds = (dL_{\exp(X)})_1 \circ \frac{I - e^{-\operatorname{ad} X}}{\operatorname{ad} X}. \end{aligned}$$

Remark: If G is a matrix group of $n \times n$ matrices, we see immediately that the derivative of left multiplication $(X \mapsto L_A X = AX)$ is given by

$$(dL_A)_X Y = AY,$$

for all $n \times n$ matrices, X, Y. Consequently, for a matrix group, we get

$$d\exp_X = e^X \left(\frac{I - e^{-\operatorname{ad} X}}{\operatorname{ad} X} \right).$$

Now, if A is a real matrix, it is clear that the (complex) eigenvalues of $\int_0^1 e^{sA} ds$ are of the form

$$\frac{e^\lambda - 1}{\lambda} \quad (= 1 \quad \text{if } \lambda = 0),$$

where λ ranges over the (complex) eigenvalues of A. Consequently, we get

Proposition 6.2. *The singular points of the exponential map,* $\exp\colon \mathfrak{g} \to G$*, that is, the set of* $X \in \mathfrak{g}$ *such that* $d\exp_X$ *is singular (not invertible) are the* $X \in \mathfrak{g}$ *such that the linear map,* $\operatorname{ad} X\colon \mathfrak{g} \to \mathfrak{g}$*, has an eigenvalue of the form* $k2\pi i$*, with* $k \in \mathbb{Z}$ *and* $k \neq 0$*.*

Another way to describe the *singular locus*, Σ, of the exponential map is to say that it is the disjoint union

$$\Sigma = \bigcup_{k \in \mathbb{Z} - \{0\}} k\Sigma_1,$$

where Σ_1 is the algebraic variety in $\mathfrak{g}$ given by

$$\Sigma_1 = \{X \in \mathfrak{g} \mid \det(\operatorname{ad} X - 2\pi i\, I) = 0\}.$$

For example, for $\mathrm{SL}(2, \mathbb{R})$,

$$\Sigma_1 = \left\{ \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \mathfrak{sl}(2) \mid a^2 + bc = -\pi^2 \right\},$$

a two-sheeted hyperboloid mapped to $-I$ by exp.

Let $\mathfrak{g}_e = \mathfrak{g} - \Sigma$ be the set of $X \in \mathfrak{g}$ such that $\frac{e^{\operatorname{ad} X} - I}{\operatorname{ad} X}$ is invertible. This is an open subset of $\mathfrak{g}$ containing 0.

6.2 The Product in Logarithmic Coordinates

Since the map,

$$X \mapsto \frac{e^{\operatorname{ad} X} - I}{\operatorname{ad} X}$$

is invertible for all $X \in \mathfrak{g}_e = \mathfrak{g} - \Sigma$, in view of the chain rule, the inverse of the above map,

$$X \mapsto \frac{\operatorname{ad} X}{e^{\operatorname{ad} X} - I},$$

is an analytic function from $\mathfrak{g}_e$ to $\mathfrak{gl}(\mathfrak{g}, \mathfrak{g})$. Let $\mathfrak{g}_e^2$ be the subset of $\mathfrak{g} \times \mathfrak{g}_e$ consisting of all (X, Y) such that the solution, $t \mapsto Z(t)$, of the differential equation

$$\frac{dZ(t)}{dt} = \frac{\operatorname{ad} Z(t)}{e^{\operatorname{ad} Z(t)} - I}(X)$$

with initial condition $Z(0) = Y (\in \mathfrak{g}_e)$, is defined for all $t \in [0, 1]$. Set

$$\mu(X, Y) = Z(1), \quad (X, Y) \in \mathfrak{g}_e^2.$$

The following theorem is proved in Duistermaat and Kolk [54] (Chapter 1, Section 1.6, Theorem 1.6.1):

Theorem 6.3. *Given any Lie group G with Lie algebra, $\mathfrak{g}$, the set $\mathfrak{g}_e^2$ is an open subset of $\mathfrak{g} \times \mathfrak{g}$ containing $(0,0)$ and the map, $\mu\colon \mathfrak{g}_e^2 \to \mathfrak{g}$, is real-analytic. Furthermore, we have*

$$\exp(X)\exp(Y) = \exp(\mu(X,Y)), \qquad (X,Y) \in \mathfrak{g}_e^2,$$

where $\exp\colon \mathfrak{g} \to G$. If $\mathfrak{g}$ is a complex Lie algebra, then μ is complex-analytic.

We may think of μ as the product in logarithmic coordinates. It is explained in Duistermaat and Kolk [54] (Chapter 1, Section 1.6) how Theorem 6.3 implies that a Lie group can be provided with the structure of a real-analytic Lie group. Rather than going into this, we will state a remarkable formula due to Dynkin expressing the Taylor expansion of μ at the origin.

6.3 Dynkin's Formula

As we said in Section 6.3, the problem of finding the Taylor expansion of μ near the origin was investigated by Campbell (1897/98), Baker (1905) and Hausdorff (1906). However, it was Dynkin who finally gave an explicit formula in 1947. There are actually slightly different versions of Dynkin's formula. One version is given (and proved convergent) in Duistermaat and Kolk [54] (Chapter 1, Section 1.7). Another slightly more explicit version (because it gives a formula for the homogeneous components of $\mu(X,Y)$) is given (and proved convergent) in Bourbaki [22] (Chapter II, §6, Section 4) and Serre [137] (Part I, Chapter IV, Section 8). We present the version in Bourbaki and Serre without proof. The proof uses formal power series and free Lie algebras.

Given $X, Y \in \mathfrak{g}_e^2$, we can write

$$\mu(X,Y) = \sum_{n=1}^{\infty} z_n(X,Y),$$

where $z_n(X,Y)$ is a homogeneous polynomial of degree n in the non-commuting variables X, Y.

Theorem 6.4. *(Dynkin's Formula) If we write $\mu(X,Y) = \sum_{n=1}^{\infty} z_n(X,Y)$, then we have*

$$z_n(X,Y) = \frac{1}{n} \sum_{p+q=n} (z'_{p,q}(X,Y) + z''_{p,q}(X,Y)),$$

with

$$z'_{p,q}(X,Y) = \sum_{\substack{p_1+\cdots+p_m=p \\ q_1+\cdots+q_{m-1}=q-1 \\ p_i+q_i\geq 1,\, p_m\geq 1,\, m\geq 1}} \frac{(-1)^{m+1}}{m} \left(\left(\prod_{i=1}^{m-1} \frac{(\operatorname{ad} X)^{p_i}}{p_i!} \frac{(\operatorname{ad} Y)^{q_i}}{q_i!} \right) \frac{(\operatorname{ad} X)^{p_m}}{p_m!} \right) (Y)$$

and

$$z''_{p,q}(X,Y) = \sum_{\substack{p_1+\cdots+p_{m-1}=p-1\\ q_1+\cdots+q_{m-1}=q\\ p_i+q_i\geq 1,\, m\geq 1}} \frac{(-1)^{m+1}}{m}\left(\prod_{i=1}^{m-1}\frac{(\operatorname{ad} X)^{p_i}}{p_i!}\frac{(\operatorname{ad} Y)^{q_i}}{q_i!}\right)(X).$$

As a concrete illustration of Dynkin's formula, after some labor, the following Taylor expansion up to order 4 is obtained:

$$\begin{aligned}\mu(X,Y) &= X+Y+\frac{1}{2}[X,Y]+\frac{1}{12}[X,[X,Y]]+\frac{1}{12}[Y,[Y,X]]-\frac{1}{24}[X,[Y,[X,Y]]]\\ &\quad + \text{higher order terms.}\end{aligned}$$

Observe that due the lack of associativity of the Lie bracket quite different looking expressions can be obtained using the Jacobi identity. For example,

$$-[X,[Y,[X,Y]]] = [Y,[X,[Y,X]]].$$

There is also an integral version of the Campbell-Baker-Hausdorff formula, see Hall [71] (Chapter 3).

Chapter 7

Bundles, Riemannian Manifolds and Homogeneous Spaces, II

7.1 Fibre Bundles

We saw in Section 2.2 that a transitive action, $\cdot\colon G \times X \to X$, of a group, G, on a set, X, yields a description of X as a quotient G/G_x, where G_x is the stabilizer of any element, $x \in X$. In Theorem 2.26, we saw that if X is a "well-behaved" topological space, G is a "well-behaved" topological group and the action is continuous, then G/G_x is homeomorphic to X. In particular the conditions of Theorem 2.26 are satisfied if G is a Lie group and X is a manifold. Intuitively, the above theorem says that G can be viewed as a family of "fibres", G_x, all isomorphic to G, these fibres being parametrized by the "base space", X, and varying smoothly when x moves in X. We have an example of what is called a fibre bundle, in fact, a principal fibre bundle. Now that we know about manifolds and Lie groups, we can be more precise about this situation.

Although we will not make extensive use of it, we begin by reviewing the definition of a fibre bundle because we believe that it clarifies the notions of vector bundles and principal fibre bundles, the concepts that are our primary concern. The following definition is not the most general but it is sufficient for our needs:

Definition 7.1. A *fibre bundle with (typical) fibre, F, and structure group, G,* is a tuple, $\xi = (E, \pi, B, F, G)$, where E, B, F are smooth manifolds, $\pi\colon E \to B$ is a smooth surjective map, G is a Lie group of diffeomorphisms of F and there is some open cover, $\mathcal{U} = (U_\alpha)_{\alpha \in I}$, of B and a family, $\varphi = (\varphi_\alpha)_{\alpha \in I}$, of diffeomorphisms,

$$\varphi_\alpha\colon \pi^{-1}(U_\alpha) \to U_\alpha \times F.$$

The space, B, is called the *base space*, E is called the *total space*, F is called the *(typical) fibre*, and each φ_α is called a *(local) trivialization*. The pair, $(U_\alpha, \varphi_\alpha)$, is called a *bundle chart* and the family, $\{(U_\alpha, \varphi_\alpha)\}$, is a *trivializing cover*. For each $b \in B$, the space, $\pi^{-1}(b)$,

is called the *fibre above* b; it is also denoted by E_b, and $\pi^{-1}(U_\alpha)$ is also denoted by $E \restriction U_\alpha$. Furthermore, the following properties hold:

(a) The diagram

$$\begin{array}{ccc} \pi^{-1}(U_\alpha) & \xrightarrow{\varphi_\alpha} & U_\alpha \times F \\ & {\scriptstyle \pi} \searrow \quad \swarrow {\scriptstyle p_1} & \\ & U_\alpha & \end{array}$$

commutes for all $\alpha \in I$, where $p_1 \colon U_\alpha \times F \to U_\alpha$ is the first projection. Equivalently, for all $(b, y) \in U_\alpha \times F$,

$$\pi \circ \varphi_\alpha^{-1}(b, y) = b.$$

For every $(U_\alpha, \varphi_\alpha)$ and every $b \in U_\alpha$, because $p_1 \circ \varphi_\alpha = \pi$ (by (a)), the restriction of φ_α to $E_b = \pi^{-1}(b)$ is a diffeomorphism between E_b and $\{b\} \times F$, so we have the diffeomorphism

$$\varphi_{\alpha,b} \colon E_b \to F$$

given by

$$\varphi_{\alpha,b}(Z) = (p_2 \circ \varphi_\alpha)(Z)$$

for all $Z \in E_b$. Furthermore, for all U_α, U_β in $\mathcal{U}$ such that $U_\alpha \cap U_\beta \neq \emptyset$, for every $b \in U_\alpha \cap U_\beta$, there is a relationship between $\varphi_{\alpha,b}$ and $\varphi_{\beta,b}$ which gives the twisting of the bundle:

(b) The diffeomorphism,

$$\varphi_{\alpha,b} \circ \varphi_{\beta,b}^{-1} \colon F \to F,$$

is an element of the group G.

(c) The map, $g_{\alpha\beta} \colon U_\alpha \cap U_\beta \to G$, defined by

$$g_{\alpha\beta}(b) = \varphi_{\alpha,b} \circ \varphi_{\beta,b}^{-1}$$

is smooth. The maps $g_{\alpha\beta}$ are called the *transition maps* of the fibre bundle.

A fibre bundle, $\xi = (E, \pi, B, F, G)$, is also referred to, somewhat loosely, as a *fibre bundle over* B or a G*-bundle* and it is customary to use the notation

$$F \longrightarrow E \longrightarrow B,$$

or

$$\begin{array}{ccc} F & \longrightarrow & E \\ & & \downarrow \\ & & B \end{array}$$

even though it is imprecise (the group G is missing!) and it clashes with the notation for short exact sequences. Observe that the bundle charts, $(U_\alpha, \varphi_\alpha)$, are similar to the charts of a manifold.

Actually, Definition 7.1 is too restrictive because it does not allow for the addition of compatible bundle charts, for example, when considering a refinement of the cover, $\mathcal{U}$. This problem can easily be fixed using a notion of equivalence of trivializing covers analogous to the equivalence of atlases for manifolds (see Remark (2) below). Also Observe that (b) and (c) imply that the isomorphism, $\varphi_\alpha \circ \varphi_\beta^{-1} \colon (U_\alpha \cap U_\beta) \times F \to (U_\alpha \cap U_\beta) \times F$, is related to the smooth map, $g_{\alpha\beta} \colon U_\alpha \cap U_\beta \to G$, by the identity

$$\varphi_\alpha \circ \varphi_\beta^{-1}(b, x) = (b, g_{\alpha\beta}(b)(x)),$$

for all $b \in U_\alpha \cap U_\beta$ and all $x \in F$.

Note that the isomorphism, $\varphi_\alpha \circ \varphi_\beta^{-1} \colon (U_\alpha \cap U_\beta) \times F \to (U_\alpha \cap U_\beta) \times F$, describes how the fibres viewed over U_β are viewed over U_α. Thus, it might have been better to denote $g_{\alpha\beta}$ by g_β^α, so that

$$g_\alpha^\beta = \varphi_{\beta,b} \circ \varphi_{\alpha,b}^{-1},$$

where the subscript, α, indicates the source and the superscript, β, indicates the target.

Intuitively, a fibre bundle over B is a family, $E = (E_b)_{b \in B}$, of spaces, E_b, (fibres) indexed by B and varying smoothly as b moves in B, such that every E_b is diffeomorphic to F. The bundle, $E = B \times F$, where π is the first projection, is called the *trivial bundle* (over B). The trivial bundle, $B \times F$, is often denoted ϵ^F. The local triviality condition (a) says that *locally*, that is, over every subset, U_α, from some open cover of the base space, B, the bundle $\xi \restriction U_\alpha$ is trivial. Note that if G is the trivial one-element group, then the fibre bundle is trivial. In fact, the purpose of the group G is to specify the "twisting" of the bundle, that is, how the fibre, E_b, gets twisted as b moves in the base space, B.

A Möbius strip is an example of a nontrivial fibre bundle where the base space, B, is the circle S^1 and the fibre space, F, is the closed interval $[-1, 1]$ and the structural group is $G = \{1, -1\}$, where -1 is the reflection of the interval $[-1, 1]$ about its midpoint, 0. The total space, E, is the strip obtained by rotating the line segment $[-1, 1]$ around the circle, keeping its midpoint in contact with the circle, and gradually twisting the line segment so that after a full revolution, the segment has been tilted by π. The reader should work out the transition functions for an open cover consisting of two open intervals on the circle.

A Klein bottle is also a fibre bundle for which both the base space and the fibre are the circle, S^1. Again, the reader should work out the details for this example.

Other examples of fibre bundles are:

(1) $\mathbf{SO}(n+1)$, an $\mathbf{SO}(n)$-bundle over the sphere S^n with fibre $\mathbf{SO}(n)$. (for $n \geq 0$).

(2) $\mathbf{SU}(n+1)$, an $\mathbf{SU}(n)$-bundle over the sphere S^{2n+1} with fibre $\mathbf{SU}(n)$ (for $n \geq 0$).

(3) $\mathbf{SL}(2,\mathbb{R})$, an $\mathbf{SO}(2)$-bundle over the upper-half space H, with fibre $\mathbf{SO}(2)$.

(4) $\mathbf{GL}(n,\mathbb{R})$, an $\mathbf{O}(n)$-bundle over the space, $\mathbf{SPD}(n)$, of symmetric, positive definite matrices, with fibre $\mathbf{O}(n)$.

(5) $\mathbf{GL}^+(n,\mathbb{R})$, an $\mathbf{SO}(n)$-bundle over the space, $\mathbf{SPD}(n)$, of symmetric, positive definite matrices, with fibre $\mathbf{SO}(n)$.

(6) $\mathbf{SO}(n+1)$, an $\mathbf{O}(n)$-bundle over the real projective space $\mathbb{RP}^n$ with fibre $\mathbf{O}(n)$ (for $n \geq 0$).

(7) $\mathbf{SU}(n+1)$, an $\mathbf{U}(n)$-bundle over the complex projective space $\mathbb{CP}^n$ with fibre $\mathbf{U}(n)$ (for $n \geq 0$).

(8) $\mathbf{O}(n)$, an $\mathbf{O}(k) \times \mathbf{O}(n-k)$-bundle over the Grassmannian, $G(k,n)$ with fibre $\mathbf{O}(k) \times \mathbf{O}(n-k)$.

(9) $\mathbf{SO}(n)$ an $S(\mathbf{O}(k) \times \mathbf{O}(n-k))$-bundle over the Grassmannian, $G(k,n)$ with fibre $S(\mathbf{O}(k) \times \mathbf{O}(n-k))$.

(10) From Section 2.5, we see that the Lorentz group, $\mathbf{SO}_0(n,1)$, is an $\mathbf{SO}(n)$-bundle over the space, $\mathcal{H}_n^+(1)$, consisting of one sheet of the hyperbolic paraboloid, $\mathcal{H}_n(1)$, with fibre $\mathbf{SO}(n)$.

Observe that in all the examples above, $F = G$, that is, the typical fibre is identical to the group G. Special bundles of this kind are called *principal fibre bundles*.

Remarks:

(1) The above definition is slightly different (but equivalent) to the definition given in Bott and Tu [19], page 47-48. Definition 7.1 is closer to the one given in Hirzebruch [78]. Bott and Tu and Hirzebruch assume that G acts effectively on the left on the fibre, F. This means that there is a smooth action, $\cdot\colon G \times F \to F$, and recall that G acts effectively on F iff for every $g \in G$,

$$\text{if} \quad g \cdot x = x \quad \text{for all } x \in F, \quad \text{then} \quad g = 1.$$

Every $g \in G$ induces a diffeomorphism, $\varphi_g\colon F \to F$, defined by

$$\varphi_g(x) = g \cdot x,$$

for all $x \in F$. The fact that G acts effectively on F means that the map, $g \mapsto \varphi_g$, is injective. This justifies viewing G as a group of diffeomorphisms of F, and from now on, we will denote $\varphi_g(x)$ by $g(x)$.

(2) We observed that Definition 7.1 is too restrictive because it does not allow for the addition of compatible bundle charts. We can fix this problem as follows: Given a trivializing cover, $\{(U_\alpha, \varphi_\alpha)\}$, for any open, U, of B and any diffeomorphism,

$$\varphi\colon \pi^{-1}(U) \to U \times F,$$

we say that (U, φ) is *compatible with the trivializing cover,* $\{(U_\alpha, \varphi_\alpha)\}$, iff whenever $U \cap U_\alpha \neq \emptyset$, there is some smooth map, $g_\alpha\colon U \cap U_\alpha \to G$, so that

$$\varphi \circ \varphi_\alpha^{-1}(b, x) = (b, g_\alpha(b)(x)),$$

for all $b \in U \cap U_\alpha$ and all $x \in F$. Two trivializing covers are *equivalent* iff every bundle chart of one cover is compatible with the other cover. This is equivalent to saying that the union of two trivializing covers is a trivializing cover. Then, we can define a fibre bundle as a tuple, $(E, \pi, B, F, G, \{(U_\alpha, \varphi_\alpha)\})$, where $\{(U_\alpha, \varphi_\alpha)\}$ is an equivalence class of trivializing covers. As for manifolds, given a trivializing cover, $\{(U_\alpha, \varphi_\alpha)\}$, the set of all bundle charts compatible with $\{(U_\alpha, \varphi_\alpha)\}$ is a maximal trivializing cover equivalent to $\{(U_\alpha, \varphi_\alpha)\}$.

A special case of the above occurs when we have a trivializing cover, $\{(U_\alpha, \varphi_\alpha)\}$, with $\mathcal{U} = \{U_\alpha\}$ an open cover of B and another open cover, $\mathcal{V} = (V_\beta)_{\beta \in J}$, of B which is a refinement of $\mathcal{U}$. This means that there is a map, $\tau\colon J \to I$, such that $V_\beta \subseteq U_{\tau(\beta)}$ for all $\beta \in J$. Then, for every $V_\beta \in \mathcal{V}$, since $V_\beta \subseteq U_{\tau(\beta)}$, the restriction of $\varphi_{\tau(\beta)}$ to V_β is a trivialization

$$\varphi'_\beta\colon \pi^{-1}(V_\beta) \to V_\beta \times F$$

and conditions (b) and (c) are still satisfied, so $(V_\beta, \varphi'_\beta)$ is compatible with $\{(U_\alpha, \varphi_\alpha)\}$.

(3) (For readers familiar with sheaves) Hirzebruch defines the sheaf, G_∞, where $\Gamma(U, G_\infty)$ is the group of smooth functions, $g\colon U \to G$, where U is some open subset of B and G is a Lie group acting effectively (on the left) on the fibre F. The group operation on $\Gamma(U, G_\infty)$ is induced by multiplication in G, that is, given two (smooth) functions, $g\colon U \to G$ and $h\colon U \to G$,

$$gh(b) = g(b)h(b),$$

for all $b \in U$.

Beware that gh is **not** function composition, unless G itself is a group of functions, which is the case for vector bundles.

Our conditions (b) and (c) are then replaced by the following equivalent condition: For all U_α, U_β in $\mathcal{U}$ such that $U_\alpha \cap U_\beta \neq \emptyset$, there is some $g_{\alpha\beta} \in \Gamma(U_\alpha \cap U_\beta, G_\infty)$ such that

$$\varphi_\alpha \circ \varphi_\beta^{-1}(b, x) = (b, g_{\alpha\beta}(b)(x)),$$

for all $b \in U_\alpha \cap U_\beta$ and all $x \in F$.

(4) The family of transition functions $(g_{\alpha\beta})$ satisfies the *cocycle condition*,

$$g_{\alpha\beta}(b)g_{\beta\gamma}(b) = g_{\alpha\gamma}(b),$$

for all α, β, γ such that $U_\alpha \cap U_\beta \cap U_\gamma \neq \emptyset$ and all $b \in U_\alpha \cap U_\beta \cap U_\gamma$. Setting $\alpha = \beta = \gamma$, we get

$$g_{\alpha\alpha} = \mathrm{id},$$

and setting $\gamma = \alpha$, we get

$$g_{\beta\alpha} = g_{\alpha\beta}^{-1}.$$

Again, beware that this means that $g_{\beta\alpha}(b) = g_{\alpha\beta}^{-1}(b)$, where $g_{\alpha\beta}^{-1}(b)$ is the inverse of $g_{\beta\alpha}(b)$ in G. In general, $g_{\alpha\beta}^{-1}$ is **not** the functional inverse of $g_{\beta\alpha}$.

The classic source on fibre bundles is Steenrod [142]. The most comprehensive treatment of fibre bundles and vector bundles is probably given in Husemoller [83]. However, we can hardly recommend this book. We find the presentation overly formal and intuitions are absent. A more extensive list of references is given at the end of Section 7.5.

Remark: (The following paragraph is intended for readers familiar with Čech cohomology.) The cocycle condition makes it possible to view a fibre bundle over B as a member of a certain *(Čech) cohomology set*, $\check{H}^1(B, \mathcal{G})$, where $\mathcal{G}$ denotes a certain sheaf of functions from the manifold B into the Lie group G, as explained in Hirzebruch [78], Section 3.2. However, this requires defining a noncommutative version of Čech cohomology (at least, for $\check{H}^1$), and clarifying when two open covers and two trivializations define the same fibre bundle over B, or equivalently, defining when two fibre bundles over B are equivalent. If the bundles under considerations are line bundles (see Definition 7.6), then $\check{H}^1(B, \mathcal{G})$ is actually a group. In this case, $G = \mathrm{GL}(1, \mathbb{R}) \cong \mathbb{R}^*$ in the real case and $G = \mathrm{GL}(1, \mathbb{C}) \cong \mathbb{C}^*$ in the complex case (where $\mathbb{R}^* = \mathbb{R} - \{0\}$ and $\mathbb{C}^* = \mathbb{C} - \{0\}$), and the sheaf $\mathcal{G}$ is the sheaf of smooth (real-valued or complex-valued) functions vanishing nowhere. The group, $\check{H}^1(B, \mathcal{G})$, plays an important role, especially when the bundle is a holomorphic line bundle over a complex manifold. In the latter case, it is called the *Picard group* of B.

The notion of a map between fibre bundles is more subtle than one might think because of the structure group, G. Let us begin with the simpler case where $G = \mathrm{Diff}(F)$, the group of all smooth diffeomorphisms of F.

Definition 7.2. If $\xi_1 = (E_1, \pi_1, B_1, F, \mathrm{Diff}(F))$ and $\xi_2 = (E_2, \pi_2, B_2, F, \mathrm{Diff}(F))$ are two fibre bundles with the same typical fibre, F, and the same structure group, $G = \mathrm{Diff}(F)$, a *bundle map (or bundle morphism)*, $f\colon \xi_1 \to \xi_2$, is a pair, $f = (f_E, f_B)$, of smooth maps, $f_E\colon E_1 \to E_2$ and $f_B\colon B_1 \to B_2$, such that

(a) The following diagram commutes:

$$\begin{array}{ccc} E_1 & \xrightarrow{f_E} & E_2 \\ {\scriptstyle \pi_1}\downarrow & & \downarrow{\scriptstyle \pi_2} \\ B_1 & \xrightarrow[f_B]{} & B_2 \end{array}$$

(b) For every $b \in B_1$, the map of fibres,

$$f_E \restriction \pi_1^{-1}(b)\colon \pi_1^{-1}(b) \to \pi_2^{-1}(f_B(b)),$$

is a diffeomorphism (*preservation of the fibre*).

A bundle map, $f\colon \xi_1 \to \xi_2$, is an *isomorphism* if there is some bundle map, $g\colon \xi_2 \to \xi_1$, called the *inverse of* f such that

$$g_E \circ f_E = \mathrm{id} \quad \text{and} \quad f_E \circ g_E = \mathrm{id}.$$

The bundles ξ_1 and ξ_2 are called *isomorphic*. Given two fibre bundles, $\xi_1 = (E_1, \pi_1, B, F, G)$ and $\xi_2 = (E_2, \pi_2, B, F, G)$, over the same base space, B, a *bundle map (or bundle morphism)*, $f\colon \xi_1 \to \xi_2$, is a pair, $f = (f_E, f_B)$, where $f_B = \mathrm{id}$ (the identity map). Such a bundle map is an *isomorphism* if it has an inverse as defined above. In this case, we say that the bundles ξ_1 and ξ_2 over B are *isomorphic*.

Observe that the commutativity of the diagram in Definition 7.2 implies that f_B is actually determined by f_E. Also, when f is an isomorphism, the surjectivity of π_1 and π_2 implies that

$$g_B \circ f_B = \mathrm{id} \quad \text{and} \quad f_B \circ g_B = \mathrm{id}.$$

Thus, when $f = (f_E, f_B)$ is an isomorphism, both f_E and f_B are diffeomorphisms.

Remark: Some authors do not require the "preservation" of fibres. However, it is automatic for bundle isomorphisms.

When we have a bundle map, $f\colon \xi_1 \to \xi_2$, as above, for every $b \in B$, for any trivializations $\varphi_\alpha\colon \pi_1^{-1}(U_\alpha) \to U_\alpha \times F$ of ξ_1 and $\varphi'_\beta\colon \pi_2^{-1}(V_\beta) \to V_\beta \times F$ of ξ_2, with $b \in U_\alpha$ and $f_B(b) \in V_\beta$, we have the map,

$$\varphi'_\beta \circ f_E \circ \varphi_\alpha^{-1}\colon (U_\alpha \cap f_B^{-1}(V_\beta)) \times F \to V_\beta \times F.$$

Consequently, as φ_α and φ'_α are diffeomorphisms and as f is a diffeomorphism on fibres, we have a map, $\rho_{\alpha,\beta}\colon U_\alpha \cap f_B^{-1}(V_\beta) \to \mathrm{Diff}(F)$, such that

$$\varphi'_\beta \circ f_E \circ \varphi_\alpha^{-1}(b, x) = (f_B(b), \rho_{\alpha,\beta}(b)(x)),$$

for all $b \in U_\alpha \cap f_B^{-1}(V_\beta)$ and all $x \in F$. Unfortunately, in general, there is no garantee that $\rho_{\alpha,\beta}(b) \in G$ or that it be smooth. However, this will be the case when ξ is a vector bundle or a principal bundle.

Since we may always pick U_α and V_β so that $f_B(U_\alpha) \subseteq V_\beta$, we may also write ρ_α instead of $\rho_{\alpha,\beta}$, with $\rho_\alpha\colon U_\alpha \to G$. Then, observe that locally, f_E is given as the composition

$$\pi_1^{-1}(U_\alpha) \xrightarrow{\varphi_\alpha} U_\alpha \times F \xrightarrow{\tilde{f}_\alpha} V_\beta \times F \xrightarrow{{\varphi'_\beta}^{-1}} \pi_2^{-1}(V_\beta)$$

$$z \longrightarrow (b, x) \longrightarrow (f_B(b), \rho_\alpha(b)(x)) \longrightarrow {\varphi'_\beta}^{-1}(f_B(b), \rho_\alpha(b)(x)),$$

with $\widetilde{f}_\alpha(b,x) = (f_B(b), \rho_\alpha(b)(x))$, that is,

$$f_E(z) = {\varphi'_\beta}^{-1}(f_B(b), \rho_\alpha(b)(x)), \qquad \text{with } z \in \pi_1^{-1}(U_\alpha) \text{ and } (b,x) = \varphi_\alpha(z).$$

Conversely, if (f_E, f_B) is a pair of smooth maps satisfying the commutative diagram of Definition 7.2 and the above conditions hold locally, then as φ_α, φ'^{-1}_β and $\rho_\alpha(b)$ are diffeomorphisms on fibres, we see that f_E is a diffeomorphism on fibres.

In the general case where the structure group, G, is not the whole group of diffeomorphisms, $\mathrm{Diff}(F)$, following Hirzebruch [78], we use the local conditions above to define the "right notion" of bundle map, namely Definition 7.3. Another advantage of this definition is that two bundles (with the same fibre, structure group, and base) are isomorphic iff they are equivalent (see Proposition 7.1 and Proposition 7.2).

Definition 7.3. Given two fibre bundles, $\xi_1 = (E_1, \pi_1, B_1, F, G)$ and $\xi_2 = (E_2, \pi_2, B_2, F, G)$, a *bundle map*, $f\colon \xi_1 \to \xi_2$, is a pair, $f = (f_E, f_B)$, of smooth maps, $f_E\colon E_1 \to E_2$ and $f_B\colon B_1 \to B_2$, such that

(a) The diagram

$$\begin{array}{ccc} E_1 & \xrightarrow{f_E} & E_2 \\ {\scriptstyle \pi_1}\downarrow & & \downarrow{\scriptstyle \pi_2} \\ B_1 & \xrightarrow[f_B]{} & B_2 \end{array}$$

commutes.

(b) There is an open cover, $\mathcal{U} = (U_\alpha)_{\alpha\in I}$, for B_1, an open cover, $\mathcal{V} = (V_\beta)_{\beta\in J}$, for B_2, a family, $\varphi = (\varphi_\alpha)_{\alpha\in I}$, of trivializations, $\varphi_\alpha\colon \pi_1^{-1}(U_\alpha) \to U_\alpha \times F$, for ξ_1, a family, $\varphi' = (\varphi'_\beta)_{\beta\in J}$, of trivializations, $\varphi'_\beta\colon \pi_2^{-1}(V_\beta) \to V_\beta \times F$, for ξ_2, such that for every $b \in B$, there are some trivializations, $\varphi_\alpha\colon \pi_1^{-1}(U_\alpha) \to U_\alpha \times F$ and $\varphi'_\beta\colon \pi_2^{-1}(V_\beta) \to V_\beta \times F$, with $f_B(U_\alpha) \subseteq V_\beta$, $b \in U_\alpha$ and some smooth map,

$$\rho_\alpha\colon U_\alpha \to G,$$

such that $\varphi'_\beta \circ f_E \circ \varphi_\alpha^{-1}\colon U_\alpha \times F \to V_\alpha \times F$ is given by

$$\varphi'_\beta \circ f_E \circ \varphi_\alpha^{-1}(b,x) = (f_B(b), \rho_\alpha(b)(x)),$$

for all $b \in U_\alpha$ and all $x \in F$.

A bundle map is an *isomorphism* if it has an inverse as in Definition 7.2. If the bundles ξ_1 and ξ_2 are over the same base, B, then we also require $f_B = \mathrm{id}$.

As we remarked in the discussion before Definition 7.3, condition (b) insures that the maps of fibres,

$$f_E \restriction \pi_1^{-1}(b)\colon \pi_1^{-1}(b) \to \pi_2^{-1}(f_B(b)),$$

are diffeomorphisms. In the special case where ξ_1 and ξ_2 have the same base, $B_1 = B_2 = B$, we require $f_B = \mathrm{id}$ and we can use the same cover (*i.e.*, $\mathcal{U} = \mathcal{V}$) in which case condition (b) becomes: There is some smooth map, $\rho_\alpha \colon U_\alpha \to G$, such that

$$\varphi'_\alpha \circ f \circ {\varphi_\alpha}^{-1}(b, x) = (b, \rho_\alpha(b)(x)),$$

for all $b \in U_\alpha$ and all $x \in F$.

We say that a bundle, ξ, with base B and structure group G is trivial iff ξ is isomorphic to the product bundle, $B \times F$, according to the notion of isomorphism of Definition 7.3.

We can also define the notion of equivalence for fibre bundles over the same base space, B (see Hirzebruch [78], Section 3.2, Chern [33], Section 5, and Husemoller [83], Chapter 5). We will see shortly that two bundles over the same base are equivalent iff they are isomorphic.

Definition 7.4. Given two fibre bundles, $\xi_1 = (E_1, \pi_1, B, F, G)$ and $\xi_2 = (E_2, \pi_2, B, F, G)$, over the same base space, B, we say that ξ_1 and ξ_2 are *equivalent* if there is an open cover, $\mathcal{U} = (U_\alpha)_{\alpha \in I}$, for B, a family, $\varphi = (\varphi_\alpha)_{\alpha \in I}$, of trivializations, $\varphi_\alpha \colon \pi_1^{-1}(U_\alpha) \to U_\alpha \times F$, for ξ_1, a family, $\varphi' = (\varphi'_\alpha)_{\alpha \in I}$, of trivializations, $\varphi'_\alpha \colon \pi_2^{-1}(U_\alpha) \to U_\alpha \times F$, for ξ_2, and a family, $(\rho_\alpha)_{\alpha \in I}$, of smooth maps, $\rho_\alpha \colon U_\alpha \to G$, such that

$$g'_{\alpha\beta}(b) = \rho_\alpha(b) g_{\alpha\beta}(b) \rho_\beta(b)^{-1}, \qquad \text{for all } b \in U_\alpha \cap U_\beta.$$

Since the trivializations are bijections, the family $(\rho_\alpha)_{\alpha \in I}$ is unique. The following proposition shows that isomorphic fibre bundles are equivalent:

Proposition 7.1. *If two fibre bundles, $\xi_1 = (E_1, \pi_1, B, F, G)$ and $\xi_2 = (E_2, \pi_2, B, F, G)$, over the same base space, B, are isomorphic, then they are equivalent.*

Proof. Let $f \colon \xi_1 \to \xi_2$ be a bundle isomorphism. Then, we know that for some suitable open cover of the base, B, and some trivializing families, (φ_α) for ξ_1 and (φ'_α) for ξ_2, there is a family of maps, $\rho_\alpha \colon U_\alpha \to G$, so that

$$\varphi'_\alpha \circ f \circ {\varphi_\alpha}^{-1}(b, x) = (b, \rho_\alpha(b)(x)),$$

for all $b \in U_\alpha$ and all $x \in F$. Recall that

$$\varphi_\alpha \circ \varphi_\beta^{-1}(b, x) = (b, g_{\alpha\beta}(b)(x)),$$

for all $b \in U_\alpha \cap U_\beta$ and all $x \in F$. This is equivalent to

$$\varphi_\beta^{-1}(b, x) = \varphi_\alpha^{-1}(b, g_{\alpha\beta}(b)(x)),$$

so it is notationally advantageous to introduce ψ_α such that $\psi_\alpha = \varphi_\alpha^{-1}$. Then, we have

$$\psi_\beta(b, x) = \psi_\alpha(b, g_{\alpha\beta}(b)(x))$$

and

$$\varphi'_\alpha \circ f \circ \varphi_\alpha^{-1}(b, x) = (b, \rho_\alpha(b)(x))$$

becomes

$$\psi_\alpha(b, x) = f^{-1} \circ \psi'_\alpha(b, \rho_\alpha(b)(x)).$$

We have

$$\psi_\beta(b, x) = \psi_\alpha(b, g_{\alpha\beta}(b)(x)) = f^{-1} \circ \psi'_\alpha(b, \rho_\alpha(b)(g_{\alpha\beta}(b)(x)))$$

and also

$$\psi_\beta(b, x) = f^{-1} \circ \psi'_\beta(b, \rho_\beta(b)(x)) = f^{-1} \circ \psi'_\alpha(b, g'_{\alpha\beta}(b)(\rho_\beta(b)(x)))$$

from which we deduce

$$\rho_\alpha(b)(g_{\alpha\beta}(b)(x)) = g'_{\alpha\beta}(b)(\rho_\beta(b)(x)),$$

that is

$$g'_{\alpha\beta}(b) = \rho_\alpha(b) g_{\alpha\beta}(b) \rho_\beta(b)^{-1}, \qquad \text{for all } b \in U_\alpha \cap U_\beta,$$

as claimed. □

Remark: If $\xi_1 = (E_1, \pi_1, B_1, F, G)$ and $\xi_2 = (E_2, \pi_2, B_2, F, G)$ are two bundles over different bases and $f\colon \xi_1 \to \xi_2$ is a bundle isomorphism, with $f = (f_B, f_E)$, then f_E and f_B are diffeomorphisms and it is easy to see that we get the conditions

$$g'_{\alpha\beta}(f_B(b)) = \rho_\alpha(b) g_{\alpha\beta}(b) \rho_\beta(b)^{-1}, \qquad \text{for all } b \in U_\alpha \cap U_\beta.$$

The converse of Proposition 7.1 also holds.

Proposition 7.2. *If two fibre bundles, $\xi_1 = (E_1, \pi_1, B, F, G)$ and $\xi_2 = (E_2, \pi_2, B, F, G)$, over the same base space, B, are equivalent then they are isomorphic.*

Proof. Assume that ξ_1 and ξ_2 are equivalent. Then, for some suitable open cover of the base, B, and some trivializing families, (φ_α) for ξ_1 and (φ'_α) for ξ_2, there is a family of maps, $\rho_\alpha\colon U_\alpha \to G$, so that

$$g'_{\alpha\beta}(b) = \rho_\alpha(b) g_{\alpha\beta}(b) \rho_\beta(b)^{-1}, \qquad \text{for all } b \in U_\alpha \cap U_\beta,$$

which can be written as

$$g'_{\alpha\beta}(b) \rho_\beta(b) = \rho_\alpha(b) g_{\alpha\beta}(b).$$

For every U_α, define f_α as the composition

$$\pi_1^{-1}(U_\alpha) \xrightarrow{\varphi_\alpha} U_\alpha \times F \xrightarrow{\widetilde{f}_\alpha} U_\alpha \times F \xrightarrow{{\varphi'_\alpha}^{-1}} \pi_2^{-1}(U_\alpha)$$

$$z \longrightarrow (b, x) \longrightarrow (b, \rho_\alpha(b)(x)) \longrightarrow {\varphi'_\alpha}^{-1}(b, \rho_\alpha(b)(x)),$$

that is,

$$f_\alpha(z) = {\varphi'_\alpha}^{-1}(b, \rho_\alpha(b)(x)), \qquad \text{with } z \in \pi_1^{-1}(U_\alpha) \text{ and } (b, x) = \varphi_\alpha(z).$$

Clearly, the definition of f_α implies that

$$\varphi'_\alpha \circ f_\alpha \circ {\varphi_\alpha}^{-1}(b, x) = (b, \rho_\alpha(b)(x)),$$

for all $b \in U_\alpha$ and all $x \in F$ and, locally, f_α is a bundle isomorphism with respect to ρ_α. If we can prove that any two f_α and f_β agree on the overlap, $U_\alpha \cap U_\beta$, then the f_α's patch and yield a bundle map between ξ_1 and ξ_2. Now, on $U_\alpha \cap U_\beta$,

$$\varphi_\alpha \circ \varphi_\beta^{-1}(b, x) = (b, g_{\alpha\beta}(b)(x))$$

yields

$$\varphi_\beta^{-1}(b, x) = \varphi_\alpha^{-1}(b, g_{\alpha\beta}(b)(x)).$$

We need to show that for every $z \in U_\alpha \cap U_\beta$,

$$f_\alpha(z) = {\varphi'_\alpha}^{-1}(b, \rho_\alpha(b)(x)) = {\varphi'_\beta}^{-1}(b, \rho_\beta(b)(x')) = f_\beta(z),$$

where $\varphi_\alpha(z) = (b, x)$ and $\varphi_\beta(z) = (b, x')$.

From $z = \varphi_\beta^{-1}(b, x') = \varphi_\alpha^{-1}(b, g_{\alpha\beta}(b)(x'))$, we get

$$x = g_{\alpha\beta}(b)(x').$$

We also have

$${\varphi'_\beta}^{-1}(b, \rho_\beta(b)(x')) = {\varphi'_\alpha}^{-1}(b, g'_{\alpha\beta}(b)(\rho_\beta(b)(x')))$$

and since $g'_{\alpha\beta}(b)\rho_\beta(b) = \rho_\alpha(b)g_{\alpha\beta}(b)$ and $x = g_{\alpha\beta}(b)(x')$ we get

$${\varphi'_\beta}^{-1}(b, \rho_\beta(b)(x')) = {\varphi'_\alpha}^{-1}(b, \rho_\alpha(b)(g_{\alpha\beta}(b))(x')) = {\varphi'_\alpha}^{-1}(b, \rho_\alpha(b)(x)),$$

as desired. Therefore, the f_α's patch to yield a bundle map, f, with respect to the family of maps, $\rho_\alpha \colon U_\alpha \to G$. The map f is bijective because it is an isomorphism on fibres but it remains to show that it is a diffeomorphism. This is a local matter and as the φ_α and φ'_α are diffeomorphisms, it suffices to show that the map, $\widetilde{f_\alpha} \colon U_\alpha \times F \longrightarrow U_\alpha \times F$, given by

$$(b, x) \mapsto (b, \rho_\alpha(b)(x)).$$

is a diffeomorphism. For this, observe that in local coordinates, the Jacobian matrix of this map is of the form

$$J = \begin{pmatrix} I & 0 \\ C & J(\rho_\alpha(b)) \end{pmatrix},$$

where I is the identity matrix and $J(\rho_\alpha(b))$ is the Jacobian matrix of $\rho_\alpha(b)$. Since $\rho_\alpha(b)$ is a diffeomorphism, $\det(J) \neq 0$ and by the Inverse Function Theorem, the map $\widetilde{f_\alpha}$ is a diffeomorphism, as desired. □

Remark: If in Proposition 7.2, $\xi_1 = (E_1, \pi_1, B_1, F, G)$ and $\xi_2 = (E_2, \pi_2, B_2, F, G)$ are two bundles over different bases and if we have a diffeomorphism, $f_B \colon B_1 \to B_2$, and the conditions

$$g'_{\alpha\beta}(f_B(b)) = \rho_\alpha(b) g_{\alpha\beta}(b) \rho_\beta(b)^{-1}, \qquad \text{for all } b \in U_\alpha \cap U_\beta$$

hold, then there is a bundle isomorphism, (f_B, f_E) between ξ_1 and ξ_2.

It follows from Proposition 7.1 and Proposition 7.2 that two bundles over the same base are equivalent iff they are isomorphic, a very useful fact. Actually, we can use the proof of Proposition 7.2 to show that any bundle morphism, $f \colon \xi_1 \to \xi_2$, between two fibre bundles over the same base, B, is a bundle isomorphism. Because a bundle morphism, f, as above is fibre preserving, f is bijective but it is not obvious that its inverse is smooth.

Proposition 7.3. *Any bundle morphism, $f \colon \xi_1 \to \xi_2$, between two fibre bundles over the same base, B, is an isomorphism.*

Proof. Since f is bijective, this is a local matter and it is enough to prove that each, $\widetilde{f_\alpha} \colon U_\alpha \times F \longrightarrow U_\alpha \times F$, is a diffeomorphism, since f can be written as

$$f = {\varphi'_\alpha}^{-1} \circ \widetilde{f_\alpha} \circ \varphi_\alpha,$$

with

$$\widetilde{f_\alpha}(b, x) = (b, \rho_\alpha(b)(x)).$$

However, the end of the proof of Proposition 7.2 shows that $\widetilde{f_\alpha}$ is a diffeomorphism. □

Given a fibre bundle, $\xi = (E, \pi, B, F, G)$, we observed that the family, $g = (g_{\alpha\beta})$, of transition maps, $g_{\alpha\beta} \colon U_\alpha \cap U_\beta \to G$, induced by a trivializing family, $\varphi = (\varphi_\alpha)_{\alpha \in I}$, relative to the open cover, $\mathcal{U} = (U_\alpha)_{\alpha \in I}$, for B satisfies the *cocycle condition*,

$$g_{\alpha\beta}(b) g_{\beta\gamma}(b) = g_{\alpha\gamma}(b),$$

for all α, β, γ such that $U_\alpha \cap U_\beta \cap U_\gamma \neq \emptyset$ and all $b \in U_\alpha \cap U_\beta \cap U_\gamma$. Without altering anything, we may assume that $g_{\alpha\beta}$ is the (unique) function from $\emptyset$ to G when $U_\alpha \cap U_\beta = \emptyset$. Then, we call a family, $g = (g_{\alpha\beta})_{(\alpha,\beta) \in I \times I}$, as above a *$\mathcal{U}$-cocycle*, or simply, a *cocycle*. Remarkably, given such a cocycle, g, relative to $\mathcal{U}$, a fibre bundle, ξ_g, over B with fibre, F, and structure group, G, having g as family of transition functions, can be constructed. In view of Proposition 7.1, we say that two cocycles, $g = (g_{\alpha\beta})_{(\alpha,\beta) \in I \times I}$ and $g' = (g_{\alpha\beta})_{(\alpha,\beta) \in I \times I}$, are *equivalent* if there is a family, $(\rho_\alpha)_{\alpha \in I}$, of smooth maps, $\rho_\alpha \colon U_\alpha \to G$, such that

$$g'_{\alpha\beta}(b) = \rho_\alpha(b) g_{\alpha\beta}(b) \rho_\beta(b)^{-1}, \qquad \text{for all } b \in U_\alpha \cap U_\beta.$$

Theorem 7.4. *Given two smooth manifolds, B and F, a Lie group, G, acting effectively on F, an open cover, $\mathcal{U} = (U_\alpha)_{\alpha \in I}$, of B, and a cocycle, $g = (g_{\alpha\beta})_{(\alpha,\beta) \in I \times I}$, there is a fibre bundle, $\xi_g = (E, \pi, B, F, G)$, whose transition maps are the maps in the cocycle, g. Furthermore, if g and g' are equivalent cocycles, then ξ_g and $\xi_{g'}$ are isomorphic.*

Proof sketch. First, we define the space, Z, as the disjoint sum

$$Z = \coprod_{\alpha \in I} U_\alpha \times F.$$

We define the relation, $\simeq$, on $Z \times Z$, as follows: For all $(b, x) \in U_\beta \times F$ and $(b, y) \in U_\alpha \times F$, if $U_\alpha \cap U_\beta \neq \emptyset$,

$$(b, x) \simeq (b, y) \quad \text{iff} \quad y = g_{\alpha\beta}(b)(x).$$

We let $E = Z/\simeq$, and we give E the largest topology such that the injections, $\eta_\alpha \colon U_\alpha \times F \to Z$, are smooth. The cocycle condition insures that $\simeq$ is indeed an equivalence relation. We define $\pi \colon E \to B$ by $\pi([b, x]) = b$. If $p \colon Z \to E$ is the the quotient map, observe that the maps, $p \circ \eta_\alpha \colon U_\alpha \times F \to E$, are injective, and that

$$\pi \circ p \circ \eta_\alpha(b, x) = b.$$

Thus,

$$p \circ \eta_\alpha \colon U_\alpha \times F \to \pi^{-1}(U_\alpha)$$

is a bijection, and we define the trivializing maps by setting

$$\varphi_\alpha = (p \circ \eta_\alpha)^{-1}.$$

It is easily verified that the corresponding transition functions are the original $g_{\alpha\beta}$. There are some details to check. A complete proof (the only one we could find!) is given in Steenrod [142], Part I, Section 3, Theorem 3.2. The fact that ξ_g and $\xi_{g'}$ are equivalent when g and g' are equivalent follows from Proposition 7.2 (see Steenrod [142], Part I, Section 2, Lemma 2.10). □

Remark: (The following paragraph is intended for readers familiar with Čech cohomology.) Obviously, if we start with a fibre bundle, $\xi = (E, \pi, B, F, G)$, whose transition maps are the cocycle, $g = (g_{\alpha\beta})$, and form the fibre bundle, ξ_g, the bundles ξ and ξ_g are equivalent. This leads to a characterization of the set of equivalence classes of fibre bundles over a base space, B, as the *cohomology set*, $\check{H}^1(B, \mathcal{G})$. In the present case, the sheaf, $\mathcal{G}$, is defined such that $\Gamma(U, \mathcal{G})$ is the group of smooth maps from the open subset, U, of B to the Lie group, G. Since G is not abelian, the coboundary maps have to be interpreted multiplicatively. If we define the sets of cochains, $C^k(\mathcal{U}, \mathcal{G})$, so that

$$C^0(\mathcal{U}, \mathcal{G}) = \prod_\alpha \mathcal{G}(U_\alpha), \quad C^1(\mathcal{U}, \mathcal{G}) = \prod_{\alpha<\beta} \mathcal{G}(U_\alpha \cap U_\beta), \quad C^2(\mathcal{U}, \mathcal{G}) = \prod_{\alpha<\beta<\gamma} \mathcal{G}(U_\alpha \cap U_\beta \cap U_\gamma),$$

etc., then it is natural to define,

$$\delta_0 \colon C^0(\mathcal{U}, \mathcal{G}) \to C^1(\mathcal{U}, \mathcal{G}),$$

by

$$(\delta_0 g)_{\alpha\beta} = g_\alpha^{-1} g_\beta,$$

for any $g = (g_\alpha)$, with $g_\alpha \in \Gamma(U_\alpha, \mathcal{G})$. As to

$$\delta_1 \colon C^1(\mathcal{U}, \mathcal{G}) \to C^2(\mathcal{U}, \mathcal{G}),$$

since the cocycle condition in the usual case is

$$g_{\alpha\beta} + g_{\beta\gamma} = g_{\alpha\gamma},$$

we set

$$(\delta_1 g)_{\alpha\beta\gamma} = g_{\alpha\beta} g_{\beta\gamma} g_{\alpha\gamma}^{-1},$$

for any $g = (g_{\alpha\beta})$, with $g_{\alpha\beta} \in \Gamma(U_\alpha \cap U_\beta, \mathcal{G})$. Note that a cocycle, $g = (g_{\alpha\beta})$, is indeed an element of $Z^1(\mathcal{U}, \mathcal{G})$, and the condition for being in the kernel of

$$\delta_1 \colon C^1(\mathcal{U}, \mathcal{G}) \to C^2(\mathcal{U}, \mathcal{G})$$

is the cocycle condition,

$$g_{\alpha\beta}(b) g_{\beta\gamma}(b) = g_{\alpha\gamma}(b),$$

for all $b \in U_\alpha \cap U_\beta \cap U_\gamma$. In the commutative case, two cocycles, g and g', are equivalent if their difference is a boundary, which can be stated as

$$g'_{\alpha\beta} + \rho_\beta = g_{\alpha\beta} + \rho_\alpha = \rho_\alpha + g_{\alpha\beta},$$

where $\rho_\alpha \in \Gamma(U_\alpha, \mathcal{G})$, for all $\alpha \in I$. In the present case, two cocycles, g and g', are equivalent iff there is a family, $(\rho_\alpha)_{\alpha \in I}$, with $\rho_\alpha \in \Gamma(U_\alpha, \mathcal{G})$, such that

$$g'_{\alpha\beta}(b) = \rho_\alpha(b) g_{\alpha\beta}(b) \rho_\beta(b)^{-1},$$

for all $b \in U_\alpha \cap U_\beta$. This is the same condition of equivalence defined earlier. Thus, it is easily seen that if $g, h \in Z^1(\mathcal{U}, \mathcal{G})$, then ξ_g and ξ_h are equivalent iff g and h correspond to the same element of the cohomology set, $\check{H}^1(\mathcal{U}, \mathcal{G})$. As usual, $\check{H}^1(B, \mathcal{G})$ is defined as the direct limit of the directed system of sets, $\check{H}^1(\mathcal{U}, \mathcal{G})$, over the preordered directed family of open covers. For details, see Hirzebruch [78], Section 3.1. In summary, there is a bijection between the equivalence classes of fibre bundles over B (with fibre F and structure group G) and the cohomology set, $\check{H}^1(B, \mathcal{G})$. In the case of line bundles, it turns out that $\check{H}^1(B, \mathcal{G})$ is in fact a group.

As an application of Theorem 7.4, we define the notion of *pullback* (or *induced*) bundle. Say $\xi = (E, \pi, B, F, G)$ is a fibre bundle and assume we have a smooth map, $f \colon N \to B$. We seek a bundle, $f^*\xi$, over N, together with a bundle map, $(f^*, f) \colon f^*\xi \to \xi$,

$$\begin{array}{ccc} f^*E & \xrightarrow{f^*} & E \\ {\scriptstyle \pi^*}\downarrow & & \downarrow{\scriptstyle \pi} \\ N & \xrightarrow[f]{} & B \end{array}$$

where, in fact, f^*E is a pullback in the categorical sense. This means that for any other bundle, ξ', over N and any bundle map,

$$\begin{array}{ccc} E' & \xrightarrow{f'} & E \\ {\scriptstyle \pi'}\downarrow & & \downarrow{\scriptstyle \pi} \\ N & \xrightarrow[f]{} & B, \end{array}$$

there is a unique bundle map, $(\widetilde{f}', \mathrm{id})\colon \xi' \to f^*\xi$, so that $(f', f) = (f^*, f) \circ (\widetilde{f}', \mathrm{id})$. Thus, there is an isomorphism (natural),

$$\mathrm{Hom}(\xi', \xi) \cong \mathrm{Hom}(\xi,' f^*\xi).$$

As a consequence, by Proposition 7.3, for any bundle map betwen ξ' and ξ,

$$\begin{array}{ccc} E' & \xrightarrow{f'} & E \\ {\scriptstyle \pi'}\downarrow & & \downarrow{\scriptstyle \pi} \\ N & \xrightarrow[f]{} & B, \end{array}$$

there is an isomorphism, $\xi' \cong f^*\xi$.

The bundle, $f^*\xi$, can be constructed as follows: Pick any open cover, (U_α), of B, then $(f^{-1}(U_\alpha))$ is an open cover of N and check that if $(g_{\alpha\beta})$ is a cocycle for ξ, then the maps, $g_{\alpha\beta} \circ f\colon f^{-1}(U_\alpha) \cap f^{-1}(U_\beta) \to G$, satisfy the cocycle conditions. Then, $f^*\xi$ is the bundle defined by the cocycle, $(g_{\alpha\beta} \circ f)$. We leave as an exercise to show that the pullback bundle, $f^*\xi$, can be defined explicitly if we set

$$f^*E = \{(n, e) \in N \times E \mid f(n) = \pi(e)\},$$

$\pi^* = pr_1$ and $f^* = pr_2$. For any trivialization, $\varphi_\alpha\colon \pi^{-1}(U_\alpha) \to U_\alpha \times F$, of ξ we have

$$(\pi^*)^{-1}(f^{-1}(U_\alpha)) = \{(n, e) \in N \times E \mid n \in f^{-1}(U_\alpha), e \in \pi^{-1}(f(n))\},$$

and so, we have a bijection, $\widetilde{\varphi}_\alpha\colon (\pi^*)^{-1}(f^{-1}(U_\alpha)) \to f^{-1}(U_\alpha) \times F$, given by

$$\widetilde{\varphi}_\alpha(n, e) = (n, pr_2(\varphi_\alpha(e))).$$

By giving f^*E the smallest topology that makes each $\widetilde{\varphi}_\alpha$ a diffeomorphism, $\widetilde{\varphi}_\alpha$, is a trivialization of $f^*\xi$ over $f^{-1}(U_\alpha)$ and $f^*\xi$ is a smooth bundle. Note that the fibre of $f^*\xi$ over a point, $n \in N$, is isomorphic to the fibre, $\pi^{-1}(f(n))$, of ξ over $f(n)$. If $g\colon M \to N$ is another smooth map of manifolds, it is easy to check that

$$(f \circ g)^*\xi = g^*(f^*\xi).$$

Given a bundle, $\xi = (E, \pi, B, F, G)$, and a submanifold, N, of B, we define the *restriction of ξ to N* as the bundle, $\xi \restriction N = (\pi^{-1}(N), \pi \restriction \pi^{-1}(N), B, F, G)$.

Experience shows that most objects of interest in geometry (vector fields, differential forms, *etc.*) arise as sections of certain bundles. Furthermore, deciding whether or not a bundle is trivial often reduces to the existence of a (global) section. Thus, we define the important concept of a section right away.

Definition 7.5. Given a fibre bundle, $\xi = (E, \pi, B, F, G)$, a *smooth section* of ξ is a smooth map, $s\colon B \to E$, so that $\pi \circ s = \mathrm{id}_B$. Given an open subset, U, of B, a *(smooth) section of ξ over U* is a smooth map, $s\colon U \to E$, so that $\pi \circ s(b) = b$, for all $b \in U$; we say that s is a *local section* of ξ. The set of all sections over U is denoted $\Gamma(U, \xi)$ and $\Gamma(B, \xi)$ (for short, $\Gamma(\xi)$) is the set of *global sections* of ξ.

Here is an observation that proves useful for constructing global sections. Let $s\colon B \to E$ be a global section of a bundle, ξ. For every trivialization, $\varphi_\alpha\colon \pi^{-1}(U_\alpha) \to U_\alpha \times F$, let $s_\alpha\colon U_\alpha \to E$ and $\sigma_\alpha\colon U_\alpha \to F$ be given by

$$s_\alpha = s \restriction U_\alpha \quad \text{and} \quad \sigma_\alpha = pr_2 \circ \varphi_\alpha \circ s_\alpha,$$

so that

$$s_\alpha(b) = \varphi_\alpha^{-1}(b, \sigma_\alpha(b)).$$

Obviously, $\pi \circ s_\alpha = \mathrm{id}$, so s_α is a local section of ξ and σ_α is a function, $\sigma_\alpha\colon U_\alpha \to F$. We claim that on overlaps, we have

$$\sigma_\alpha(b) = g_{\alpha\beta}(b)\sigma_\beta(b).$$

Indeed, recall that

$$\varphi_\alpha \circ \varphi_\beta^{-1}(b, x) = (b, g_{\alpha\beta}(b)x),$$

for all $b \in U_\alpha \cap U_\beta$ and all $x \in F$ and as $s_\alpha = s \restriction U_\alpha$ and $s_\beta = s \restriction U_\beta$, s_α and s_β agree on $U_\alpha \cap U_\beta$. Consequently, from

$$s_\alpha(b) = \varphi_\alpha^{-1}(b, \sigma_\alpha(b)) \quad \text{and} \quad s_\beta(b) = \varphi_\beta^{-1}(b, \sigma_\beta(b)),$$

we get

$$\varphi_\alpha^{-1}(b, \sigma_\alpha(b)) = s_\alpha(b) = s_\beta(b) = \varphi_\beta^{-1}(b, \sigma_\beta(b)) = \varphi_\alpha^{-1}(b, g_{\alpha\beta}(b)\sigma_\beta(b)),$$

which implies $\sigma_\alpha(b) = g_{\alpha\beta}(b)\sigma_\beta(b)$, as claimed.

Conversely, assume that we have a collection of functions, $\sigma_\alpha\colon U_\alpha \to F$, satisfying

$$\sigma_\alpha(b) = g_{\alpha\beta}(b)\sigma_\beta(b)$$

on overlaps. Let $s_\alpha\colon U_\alpha \to E$ be given by

$$s_\alpha(b) = \varphi_\alpha^{-1}(b, \sigma_\alpha(b)).$$

Each s_α is a local section and we claim that these sections agree on overlaps, so they patch and define a global section, s. We need to show that

$$s_\alpha(b) = \varphi_\alpha^{-1}(b, \sigma_\alpha(b)) = \varphi_\beta^{-1}(b, \sigma_\beta(b)) = s_\beta(b),$$

for $b \in U_\alpha \cap U_\beta$, that is,

$$(b, \sigma_\alpha(b)) = \varphi_\alpha \circ \varphi_\beta^{-1}(b, \sigma_\beta(b)),$$

and since $\varphi_\alpha \circ \varphi_\beta^{-1}(b, \sigma_\beta(b)) = (b, g_{\alpha\beta}(b)\sigma_\beta(b))$ and by hypothesis, $\sigma_\alpha(b) = g_{\alpha\beta}(b)\sigma_\beta(b)$, our equation $s_\alpha(b) = s_\beta(b)$ is verified.

There are two particularly interesting special cases of fibre bundles:

(1) *Vector bundles*, which are fibre bundles for which the typical fibre is a finite-dimensional vector space, V, and the structure group is a subgroup of the group of linear isomorphisms ($\mathrm{GL}(n, \mathbb{R})$ or $\mathrm{GL}(n, \mathbb{C})$, where $n = \dim V$).

(2) *Principal fibre bundles*, which are fibre bundles for which the fibre, F, is equal to the structure group, G, with G acting on itself by left translation.

First, we discuss vector bundles.

7.2 Vector Bundles

Given a real vector space, V, we denote by $\mathrm{GL}(V)$ (or $\mathrm{Aut}(V)$) the vector space of linear invertible maps from V to V. If V has dimension n, then $\mathrm{GL}(V)$ has dimension n^2. Obviously, $\mathrm{GL}(V)$ is isomorphic to $\mathrm{GL}(n, \mathbb{R})$, so we often write $\mathrm{GL}(n, \mathbb{R})$ instead of $\mathrm{GL}(V)$ but this may be slightly confusing if V is the dual space, W^* of some other space, W. If V is a complex vector space, we also denote by $\mathrm{GL}(V)$ (or $\mathrm{Aut}(V)$) the vector space of linear invertible maps from V to V but this time, $\mathrm{GL}(V)$ is isomorphic to $\mathrm{GL}(n, \mathbb{C})$, so we often write $\mathrm{GL}(n, \mathbb{C})$ instead of $\mathrm{GL}(V)$.

Definition 7.6. A *rank n real smooth vector bundle with fibre V* is a tuple, $\xi = (E, \pi, B, V)$, such that $(E, \pi, B, V, \mathrm{GL}(V))$ is a smooth fibre bundle, the fibre, V, is a real vector space of dimension n and the following conditions hold:

(a) For every $b \in B$, the fibre, $\pi^{-1}(b)$, is an n-dimensional (real) vector space.

(b) For every trivialization, $\varphi_\alpha \colon \pi^{-1}(U_\alpha) \to U_\alpha \times V$, for every $b \in U_\alpha$, the restriction of φ_α to the fibre, $\pi^{-1}(b)$, is a linear isomorphism, $\pi^{-1}(b) \longrightarrow V$.

A *rank n complex smooth vector bundle with fibre V* is a tuple, $\xi = (E, \pi, B, V)$, such that $(E, \pi, B, V, \mathrm{GL}(V))$ is a smooth fibre bundle such that the fibre, V, is an n-dimensional complex vector space (viewed as a real smooth manifold) and conditions (a) and (b) above hold (for complex vector spaces). When $n = 1$, a vector bundle is called a *line bundle*.

The trivial vector bundle, $E = B \times V$, is often denoted ϵ^V. When $V = \mathbb{R}^k$, we also use the notation ϵ^k. Given a (smooth) manifold, M, of dimension n, the tangent bundle, TM, and the cotangent bundle, T^*M, are rank n vector bundles. Indeed, in Section 3.3, we defined trivialization maps (denoted τ_U) for TM. Let us compute the transition functions for the tangent bundle, TM, where M is a smooth manifold of dimension n. Recall from Definition 3.16 that for every $p \in M$, the tangent space, T_pM, consists of all equivalence classes of triples, (U, φ, x), where (U, φ) is a chart with $p \in U$, $x \in \mathbb{R}^n$, and the equivalence relation on triples is given by

$$(U, \varphi, x) \equiv (V, \psi, y) \quad \text{iff} \quad (\psi \circ \varphi^{-1})'_{\varphi(p)}(x) = y.$$

We have a natural isomorphism, $\theta_{U,\varphi,p} \colon \mathbb{R}^n \to T_pM$, between $\mathbb{R}^n$ and T_pM given by

$$\theta_{U,\varphi,p} \colon x \mapsto [(U, \varphi, x)], \qquad x \in \mathbb{R}^n.$$

Observe that for any two overlapping charts, (U, φ) and (V, ψ),

$$\theta^{-1}_{V,\psi,p} \circ \theta_{U,\varphi,p} = (\psi \circ \varphi^{-1})'_z$$

for all $p \in U \cap V$, with $z = \varphi(p) = \psi(p)$. We let TM be the disjoint union,

$$TM = \bigcup_{p \in M} T_pM,$$

define the projection, $\pi \colon TM \to M$, so that $\pi(v) = p$ if $v \in T_pM$, and we give TM the weakest topology that makes the functions, $\widetilde{\varphi} \colon \pi^{-1}(U) \to \mathbb{R}^{2n}$, given by

$$\widetilde{\varphi}(v) = (\varphi \circ \pi(v), \theta^{-1}_{U,\varphi,\pi(v)}(v)),$$

continuous, where (U, φ) is any chart of M. Each function, $\widetilde{\varphi} \colon \pi^{-1}(U) \to \varphi(U) \times \mathbb{R}^n$ is a homeomorphism and given any two overlapping charts, (U, φ) and (V, ψ), since $\theta^{-1}_{V,\psi,p} \circ \theta_{U,\varphi,p} = (\psi \circ \varphi^{-1})'_z$, with $z = \varphi(p) = \psi(p)$, the transition map,

$$\widetilde{\psi} \circ \widetilde{\varphi}^{-1} \colon \varphi(U \cap V) \times \mathbb{R}^n \longrightarrow \psi(U \cap V) \times \mathbb{R}^n,$$

is given by

$$\widetilde{\psi} \circ \widetilde{\varphi}^{-1}(z, x) = (\psi \circ \varphi^{-1}(z), (\psi \circ \varphi^{-1})'_z(x)), \qquad (z, x) \in \varphi(U \cap V) \times \mathbb{R}^n.$$

It is clear that $\widetilde{\psi} \circ \widetilde{\varphi}^{-1}$ is smooth. Moreover, the bijection,

$$\tau_U \colon \pi^{-1}(U) \to U \times \mathbb{R}^n,$$

given by

$$\tau_U(v) = (\pi(v), \theta^{-1}_{U,\varphi,\pi(v)}(v))$$

satisfies $pr_1 \circ \tau_U = \pi$ on $\pi^{-1}(U)$ and is a linear isomorphism restricted to fibres, so it is a trivialization for TM. For any two overlapping charts, $(U_\alpha, \varphi_\alpha)$ and (U_β, φ_β), the transition function, $g_{\alpha\beta}\colon U_\alpha \cap U_\beta \to \mathrm{GL}(n, \mathbb{R})$, is given by

$$g_{\alpha\beta}(p) = (\varphi_\alpha \circ \varphi_\beta^{-1})'_{\varphi(p)}.$$

We can also compute trivialization maps for T^*M. This time, T^*M is the disjoint union,

$$T^*M = \bigcup_{p \in M} T_p^*M,$$

and $\pi\colon T^*M \to M$ is given by $\pi(\omega) = p$ if $\omega \in T_p^*M$, where T_p^*M is the dual of the tangent space, T_pM. For each chart, (U, φ), by dualizing the map, $\theta_{U,\varphi,p}\colon \mathbb{R}^n \to T_pM$, we obtain an isomorphism, $\theta_{U,\varphi,p}^\top\colon T_p^*M \to (\mathbb{R}^n)^*$. Composing $\theta_{U,\varphi,p}^\top$ with the isomorphism, $\iota\colon (\mathbb{R}^n)^* \to \mathbb{R}^n$ (induced by the canonical basis $(e_1, \ldots, e_n)$ of $\mathbb{R}^n$ and its dual basis), we get an isomorphism, $\theta_{U,\varphi,p}^* = \iota \circ \theta_{U,\varphi,p}^\top\colon T_p^*M \to \mathbb{R}^n$. Then, define the bijection,

$$\widetilde{\varphi}^*\colon \pi^{-1}(U) \to \varphi(U) \times \mathbb{R}^n \subseteq \mathbb{R}^{2n},$$

by

$$\widetilde{\varphi}^*(\omega) = (\varphi \circ \pi(\omega), \theta_{U,\varphi,\pi(\omega)}^*(\omega)),$$

with $\omega \in \pi^{-1}(U)$. We give T^*M the weakest topology that makes the functions $\widetilde{\varphi}^*$ continuous and then each function, $\widetilde{\varphi}^*$, is a homeomorphism. Given any two overlapping charts, (U, φ) and (V, ψ), as

$$\theta_{V,\psi,p}^{-1} \circ \theta_{U,\varphi,p} = (\psi \circ \varphi^{-1})'_{\varphi(p)},$$

by dualization we get

$$\theta_{U,\varphi,p}^\top \circ (\theta_{V,\psi,p}^\top)^{-1} = \theta_{U,\varphi,p}^\top \circ (\theta_{V,\psi,p}^{-1})^\top = ((\psi \circ \varphi^{-1})'_{\varphi(p)})^\top,$$

then

$$\theta_{V,\psi,p}^\top \circ (\theta_{U,\varphi,p}^\top)^{-1} = (((\psi \circ \varphi^{-1})'_{\varphi(p)})^\top)^{-1},$$

and so

$$\iota \circ \theta_{V,\psi,p}^\top \circ (\theta_{U,\varphi,p}^\top)^{-1} \circ \iota^{-1} = \iota \circ (((\psi \circ \varphi^{-1})'_{\varphi(p)})^\top)^{-1} \circ \iota^{-1},$$

that is,

$$\theta_{V,\psi,p}^* \circ (\theta_{U,\varphi,p}^*)^{-1} = \iota \circ (((\psi \circ \varphi^{-1})'_{\varphi(p)})^\top)^{-1} \circ \iota^{-1}.$$

Consequently, the transition map,

$$\widetilde{\psi}^* \circ (\widetilde{\varphi}^*)^{-1}\colon \varphi(U \cap V) \times \mathbb{R}^n \longrightarrow \psi(U \cap V) \times \mathbb{R}^n,$$

is given by

$$\widetilde{\psi}^* \circ (\widetilde{\varphi}^*)^{-1}(z, x) = (\psi \circ \varphi^{-1}(z), \iota \circ (((\psi \circ \varphi^{-1})'_z)^\top)^{-1} \circ \iota^{-1}(x)), \qquad (z, x) \in \varphi(U \cap V) \times \mathbb{R}^n.$$

If we view $(\psi \circ \varphi^{-1})'_z$ as a matrix, then we can forget ι and the second component of $\widetilde{\psi}^* \circ (\widetilde{\varphi}^*)^{-1}(z, x)$ is $(((\psi \circ \varphi^{-1})'_z)^\top)^{-1}x$.

We also have trivialization maps, $\tau^*_U \colon \pi^{-1}(U) \to U \times (\mathbb{R}^n)^*$, for T^*M given by

$$\tau^*_U(\omega) = (\pi(\omega), \theta^\top_{U,\varphi,\pi(\omega)}(\omega)),$$

for all $\omega \in \pi^{-1}(U)$. The transition function, $g^*_{\alpha\beta} \colon U_\alpha \cap U_\beta \to \mathrm{GL}(n, \mathbb{R})$, is given by

$$\begin{aligned} g^*_{\alpha\beta}(p)(\eta) &= \tau^*_{U_\alpha,p} \circ (\tau^*_{U_\beta,p})^{-1}(\eta) \\ &= \theta^\top_{U_\alpha,\varphi_\alpha,\pi(\eta)} \circ (\theta^\top_{U_\beta,\varphi_\beta,\pi(\eta)})^{-1}(\eta) \\ &= ((\theta^{-1}_{U_\alpha,\varphi_\alpha,\pi(\eta)} \circ \theta_{U_\beta,\varphi_\beta,\pi(\eta)})^\top)^{-1}(\eta) \\ &= (((\varphi_\alpha \circ \varphi_\beta^{-1})'_{\varphi(p)})^\top)^{-1}(\eta), \end{aligned}$$

with $\eta \in (\mathbb{R}^n)^*$. Also note that $\mathrm{GL}(n, \mathbb{R})$ should really be $\mathrm{GL}((\mathbb{R}^n)^*)$, but $\mathrm{GL}((\mathbb{R}^n)^*)$ is isomorphic to $\mathrm{GL}(n, \mathbb{R})$. We conclude that

$$g^*_{\alpha\beta}(p) = (g_{\alpha\beta}(p)^\top)^{-1}, \qquad \text{for every } p \in M.$$

This is a general property of dual bundles, see Property (f) in Section 7.3.

Maps of vector bundles are maps of fibre bundles such that the isomorphisms between fibres are linear.

Definition 7.7. Given two vector bundles, $\xi_1 = (E_1, \pi_1, B_1, V)$ and $\xi_2 = (E_2, \pi_2, B_2, V)$, with the same typical fibre, V, a *bundle map (or bundle morphism)*, $f \colon \xi_1 \to \xi_2$, is a pair, $f = (f_E, f_B)$, of smooth maps, $f_E \colon E_1 \to E_2$ and $f_B \colon B_1 \to B_2$, such that

(a) The following diagram commutes:

$$\begin{array}{ccc} E_1 & \xrightarrow{f_E} & E_2 \\ {\scriptstyle \pi_1}\downarrow & & \downarrow{\scriptstyle \pi_2} \\ B_1 & \xrightarrow[f_B]{} & B_2 \end{array}$$

(b) For every $b \in B_1$, the map of fibres,

$$f_E \restriction \pi_1^{-1}(b) \colon \pi_1^{-1}(b) \to \pi_2^{-1}(f_B(b)),$$

is a bijective linear map.

A bundle map *isomorphism*, $f \colon \xi_1 \to \xi_2$, is defined as in Definition 7.2. Given two vector bundles, $\xi_1 = (E_1, \pi_1, B, V)$ and $\xi_2 = (E_2, \pi_2, B, V)$, over the same base space, B, we require $f_B = \mathrm{id}$.

Remark: Some authors do not require the preservation of fibres, that is, the map

$$f_E \restriction \pi_1^{-1}(b)\colon \pi_1^{-1}(b) \to \pi_2^{-1}(f_B(b))$$

is simply a linear map. It is automatically bijective for bundle isomorphisms.

Note that Definition 7.7 does not include condition (b) of Definition 7.3. However, because the restrictions of the maps φ_α, φ'_β and f to the fibres are linear isomorphisms, it turns out that condition (b) (of Definition 7.3) does hold. Indeed, if $f_B(U_\alpha) \subseteq V_\beta$, then

$$\varphi'_\beta \circ f \circ \varphi_\alpha^{-1}\colon U_\alpha \times V \longrightarrow V_\beta \times V$$

is a smooth map and, for every $b \in B$, its restriction to $\{b\} \times V$ is a linear isomorphism between $\{b\} \times V$ and $\{f_B(b)\} \times V$. Therefore, there is a smooth map, $\rho_\alpha\colon U_\alpha \to \mathrm{GL}(n, \mathbb{R})$, so that

$$\varphi'_\beta \circ f \circ \varphi_\alpha^{-1}(b, x) = (f_B(b), \rho_\alpha(b)(x))$$

and a vector bundle map is a fibre bundle map.

A *holomorphic vector bundle* is a fibre bundle where E, B are complex manifolds, V is a complex vector space of dimension n, the map π is holomorphic, the φ_α are biholomorphic, and the transition functions, $g_{\alpha\beta}$, are holomorphic. When $n = 1$, a holomorphic vector bundle is called a *holomorphic line bundle*.

Definition 7.4 also applies to vector bundles (just replace G by $\mathrm{GL}(n, \mathbb{R})$ or $\mathrm{GL}(n, \mathbb{C})$) and defines the notion of equivalence of vector bundles over B. Since vector bundle maps are fibre bundle maps, Propositions 7.1 and 7.2 immediately yield

Proposition 7.5. *Two vector bundles, $\xi_1 = (E_1, \pi_1, B, V)$ and $\xi_2 = (E_2, \pi_2, B, V)$, over the same base space, B, are equivalent iff they are isomorphic.*

Since a vector bundle map is a fibre bundle map, Proposition 7.3 also yields the useful fact:

Proposition 7.6. *Any vector bundle map, $f\colon \xi_1 \to \xi_2$, between two vector bundles over the same base, B, is an isomorphism.*

Theorem 7.4 also holds for vector bundles and yields a technique for constructing new vector bundles over some base, B.

Theorem 7.7. *Given a smooth manifold, B, an n-dimensional (real, resp. complex) vector space, V, an open cover, $\mathcal{U} = (U_\alpha)_{\alpha \in I}$ of B, and a cocycle, $g = (g_{\alpha\beta})_{(\alpha,\beta) \in I \times I}$ (with $g_{\alpha\beta}\colon U_\alpha \cap U_\beta \to \mathrm{GL}(n, \mathbb{R})$, resp. $g_{\alpha\beta}\colon U_\alpha \cap U_\beta \to \mathrm{GL}(n, \mathbb{C})$), there is a vector bundle, $\xi_g = (E, \pi, B, V)$, whose transition maps are the maps in the cocycle, g. Furthermore, if g and g' are equivalent cocycles, then ξ_g and $\xi_{g'}$ are equivalent.*

Observe that a coycle, $g = (g_{\alpha\beta})_{(\alpha,\beta)\in I\times I}$, is given by a family of matrices in $\mathrm{GL}(n,\mathbb{R})$ (resp. $\mathrm{GL}(n,\mathbb{C})$).

A vector bundle, ξ, always has a global section, namely the *zero section*, which assigns the element $0 \in \pi^{-1}(b)$, to every $b \in B$. A global section, s, is a *non-zero section* iff $s(b) \neq 0$ for all $b \in B$. It is usually difficult to decide whether a bundle has a nonzero section. This question is related to the nontriviality of the bundle and there is a useful test for triviality. Assume ξ is a trivial rank n vector bundle. Then, there is a bundle isomorphism, $f\colon B \times V \to \xi$. For every $b \in B$, we know that $f(b,-)$ is a linear isomorphism, so for any choice of a basis, $(e_1,\ldots,e_n)$ of V, we get a basis, $(f(b,e_1),\ldots,f(b,e_n))$, of the fibre, $\pi^{-1}(b)$. Thus, we have n global sections, $s_1 = f(-,e_1),\ldots,s_n = f(-,e_n)$, such that $(s_1(b),\ldots,s_n(b))$ forms a basis of the fibre, $\pi^{-1}(b)$, for every $b \in B$.

Definition 7.8. Let $\xi = (E,\pi,B,V)$ be a rank n vector bundle. For any open subset, $U \subseteq B$, an n-tuple of local sections, $(s_1,\ldots,s_n)$, over U is called a *frame over* U iff $(s_1(b),\ldots,s_n(b))$ is a basis of the fibre, $\pi^{-1}(b)$, for every $b \in U$. If $U = B$, then the s_i are global sections and $(s_1,\ldots,s_n)$ is called a *frame* (of ξ).

The notion of a frame is due to Élie Cartan who (after Darboux) made extensive use of them under the name of *moving frame* (and the *moving frame method*). Cartan's terminology is intuitively clear: As a point, b, moves in U, the frame, $(s_1(b),\ldots,s_n(b))$, moves from fibre to fibre. Physicists refer to a frame as a choice of *local gauge*.

The converse of the property established just before Definition 7.8 is also true.

Proposition 7.8. *A rank n vector bundle, ξ, is trivial iff it possesses a frame of global sections.*

Proof. We only need to prove that if ξ has a frame, $(s_1,\ldots,s_n)$, then it is trivial. Pick a basis, $(e_1,\ldots,e_n)$, of V and define the map, $f\colon B \times V \to \xi$, as follows:

$$f(b,v) = \sum_{i=1}^{n} v_i s_i(b),$$

where $v = \sum_{i=1}^{n} v_i e_i$. Clearly, f is bijective on fibres, smooth, and a map of vector bundles. By Proposition 7.6, the bundle map, f, is an isomorphism. □

As an illustration of Proposition 7.8 we can prove that the tangent bundle, TS^1, of the circle, is trivial. Indeed, we can find a section that is everywhere nonzero, *i.e.* a non-vanishing vector field, namely

$$s(\cos\theta,\sin\theta) = (-\sin\theta,\cos\theta).$$

The reader should try proving that TS^3 is also trivial (use the quaternions). However, TS^2 is nontrivial, although this not so easy to prove. More generally, it can be shown that TS^n is nontrivial for all even $n \geq 2$. It can even be shown that S^1, S^3 and S^7 are the only spheres whose tangent bundle is trivial. This is a rather deep theorem and its proof is hard.

Remark: A manifold, M, such that its tangent bundle, TM, is trivial is called *parallelizable*.

The above considerations show that if ξ is any rank n vector bundle, not necessarily trivial, then for any local trivialization, $\varphi_\alpha \colon \pi^{-1}(U_\alpha) \to U_\alpha \times V$, there are always frames over U_α. Indeed, for every choice of a basis, $(e_1, \ldots, e_n)$, of the typical fibre, V, if we set

$$s_i^\alpha(b) = \varphi_\alpha^{-1}(b, e_i), \qquad b \in U_\alpha,\ 1 \leq i \leq n,$$

then $(s_1^\alpha, \ldots, s_n^\alpha)$ is a frame over U_α.

Given any two vector spaces, V and W, both of dimension n, we denote by $\mathrm{Iso}(V, W)$ the space of all linear isomorphisms between V and W. The *space of n-frames*, $F(V)$, is the set of bases of V. Since every basis, $(v_1, \ldots, v_n)$, of V is in one-to-one correspondence with the map from $\mathbb{R}^n$ to V given by $e_i \mapsto v_i$, where $(e_1, \ldots, e_n)$ is the canonical basis of $\mathbb{R}^n$ (so, $e_i = (0, \ldots, 1, \ldots 0)$ with the 1 in the ith slot), we have an isomorphism,

$$F(V) \cong \mathrm{Iso}(\mathbb{R}^n, V).$$

(The choice of a basis in V also yields an isomorphism, $\mathrm{Iso}(\mathbb{R}^n, V) \cong \mathrm{GL}(n, \mathbb{R})$, so $F(V) \cong \mathrm{GL}(n, \mathbb{R})$.)

For any rank n vector bundle, ξ, we can form the *frame bundle*, $F(\xi)$, by replacing the fibre, $\pi^{-1}(b)$, over any $b \in B$ by $F(\pi^{-1}(b))$. In fact, $F(\xi)$ can be constructed using Theorem 7.4. Indeed, identifying $F(V)$ with $\mathrm{Iso}(\mathbb{R}^n, V)$, the group $\mathrm{GL}(n, \mathbb{R})$ acts on $F(V)$ effectively on the left *via*

$$A \cdot v = v \circ A^{-1}.$$

(The only reason for using A^{-1} instead of A is that we want a left action.) The resulting bundle has typical fibre, $F(V) \cong \mathrm{GL}(n, \mathbb{R})$, and turns out to be a principal bundle. We will take a closer look at principal bundles in Section 7.5.

We conclude this section with an example of a bundle that plays an important role in algebraic geometry, the *canonical line bundle on* $\mathbb{RP}^n$. Let $H_n^{\mathbb{R}} \subseteq \mathbb{RP}^n \times \mathbb{R}^{n+1}$ be the subset,

$$H_n^{\mathbb{R}} = \{(L, v) \in \mathbb{RP}^n \times \mathbb{R}^{n+1} \mid v \in L\},$$

where $\mathbb{RP}^n$ is viewed as the set of lines, L, in $\mathbb{R}^{n+1}$ through 0, or more explicitly,

$$H_n^{\mathbb{R}} = \{((x_0 \colon \cdots \colon x_n), \lambda(x_0, \ldots, x_n)) \mid (x_0 \colon \cdots \colon x_n) \in \mathbb{RP}^n,\ \lambda \in \mathbb{R}\}.$$

Geometrically, $H_n^{\mathbb{R}}$ consists of the set of lines, $[(x_0, \ldots, x_n)]$, associated with points, $(x_0 \colon \cdots \colon x_n)$, of $\mathbb{RP}^n$. If we consider the projection, $\pi \colon H_n^{\mathbb{R}} \to \mathbb{RP}^n$, of $H_n^{\mathbb{R}}$ onto $\mathbb{RP}^n$, we see that each fibre is isomorphic to $\mathbb{R}$. We claim that $H_n^{\mathbb{R}}$ is a line bundle. For this, we exhibit trivializations, leaving as an exercise the fact that $H_n^{\mathbb{R}}$ is a manifold.

Recall the open cover, $U_0, \ldots, U_n$, of $\mathbb{RP}^n$, where

$$U_i = \{(x_0 \colon \cdots \colon x_n) \in \mathbb{RP}^n \mid x_i \neq 0\}.$$

Then, the maps, $\varphi_i \colon \pi^{-1}(U_i) \to U_i \times \mathbb{R}$, given by

$$\varphi_i((x_0 \colon \cdots \colon x_n), \lambda(x_0, \ldots, x_n)) = ((x_0 \colon \cdots \colon x_n), \lambda x_i)$$

are trivializations. The transition function, $g_{ij} \colon U_i \cap U_j \to \mathrm{GL}(1, \mathbb{R})$, is given by

$$g_{ij}(x_0 \colon \cdots \colon x_n)(u) = \frac{x_i}{x_j}\, u,$$

where we identify $\mathrm{GL}(1, \mathbb{R})$ and $\mathbb{R}^* = \mathbb{R} - \{0\}$.

Interestingly, the bundle $H_n^{\mathbb{R}}$ is nontrivial for all $n \geq 1$. For this, by Proposition 7.8 and since $H_n^{\mathbb{R}}$ is a line bundle, it suffices to prove that every global section vanishes at some point. So, let σ be any section of $H_n^{\mathbb{R}}$. Composing the projection, $p \colon S^n \longrightarrow \mathbb{RP}^n$, with σ, we get a smooth function, $s = \sigma \circ p \colon S^n \longrightarrow H_n^{\mathbb{R}}$, and we have

$$s(x) = (p(x), f(x)x),$$

for every $x \in S^n$, where $f \colon S^n \to \mathbb{R}$ is a smooth function. Moreover, f satisfies

$$f(-x) = -f(x),$$

since $s(-x) = s(x)$. As S^n is connected and f is continuous, by the intermediate value theorem, there is some x such that $f(x) = 0$, and thus, σ vanishes, as desired.

The reader should look for a geometric representation of $H_1^{\mathbb{R}}$. It turns out that $H_1^{\mathbb{R}}$ is an open Möbius strip, that is, a Möbius strip with its boundary deleted (see Milnor and Stasheff [111], Chapter 2). There is also a complex version of the canonical line bundle on $\mathbb{CP}^n$, with

$$H_n = \{(L, v) \in \mathbb{CP}^n \times \mathbb{C}^{n+1} \mid v \in L\},$$

where $\mathbb{CP}^n$ is viewed as the set of lines, L, in $\mathbb{C}^{n+1}$ through 0. These bundles are also nontrivial. Furthermore, unlike the real case, the dual bundle, H_n^*, is not isomorphic to H_n. Indeed, H_n^* turns out to have nonzero global holomorphic sections!

7.3 Operations on Vector Bundles

Because the fibres of a vector bundle are vector spaces all isomorphic to some given space, V, we can perform operations on vector bundles that extend familiar operations on vector spaces, such as: direct sum, tensor product, (linear) function space, and dual space. Basically, the same operation is applied on fibres. It is usually more convenient to define operations on vector bundles in terms of operations on cocycles, using Theorem 7.7.

(a) (*Whitney Sum* or *Direct Sum*)

If $\xi = (E, \pi, B, V)$ is a rank m vector bundle and $\xi' = (E', \pi', B, W)$ is a rank n vector bundle, both over the same base, B, then their Whitney sum, $\xi \oplus \xi'$, is the rank $(m+n)$

vector bundle whose fibre over any $b \in B$ is the direct sum, $E_b \oplus E'_b$, that is, the vector bundle with typical fibre $V \oplus W$ (given by Theorem 7.7) specified by the cocycle whose matrices are

$$\begin{pmatrix} g_{\alpha\beta}(b) & 0 \\ 0 & g'_{\alpha\beta}(b) \end{pmatrix}, \qquad b \in U_\alpha \cap U_\beta.$$

(b) (*Tensor Product*)

If $\xi = (E, \pi, B, V)$ is a rank m vector bundle and $\xi' = (E', \pi', B, W)$ is a rank n vector bundle, both over the same base, B, then their tensor product, $\xi \otimes \xi'$, is the rank mn vector bundle whose fibre over any $b \in B$ is the tensor product, $E_b \otimes E'_b$, that is, the vector bundle with typical fibre $V \otimes W$ (given by Theorem 7.7) specified by the cocycle whose matrices are

$$g_{\alpha\beta}(b) \otimes g'_{\alpha\beta}(b), \qquad b \in U_\alpha \cap U_\beta.$$

(Here, we identify a matrix with the corresponding linear map.)

(c) (*Tensor Power*)

If $\xi = (E, \pi, B, V)$ is a rank m vector bundle, then for any $k \geq 0$, we can define the tensor power bundle, $\xi^{\otimes k}$, whose fibre over any $b \in \xi$ is the tensor power, $E_b^{\otimes k}$ and with typical fibre $V^{\otimes k}$. (When $k = 0$, the fibre is $\mathbb{R}$ or $\mathbb{C}$). The bundle $\xi^{\otimes k}$ is determined by the cocycle

$$g_{\alpha\beta}^{\otimes k}(b), \qquad b \in U_\alpha \cap U_\beta.$$

(d) (*Exterior Power*)

If $\xi = (E, \pi, B, V)$ is a rank m vector bundle, then for any $k \geq 0$, we can define the exterior power bundle, $\bigwedge^k \xi$, whose fibre over any $b \in \xi$ is the exterior power, $\bigwedge^k E_b$ and with typical fibre $\bigwedge^k V$. The bundle $\bigwedge^k \xi$ is determined by the cocycle

$$\bigwedge^k g_{\alpha\beta}(b), \qquad b \in U_\alpha \cap U_\beta.$$

Using (a), we also have the *exterior algebra bundle*, $\bigwedge \xi = \bigoplus_{k=0}^m \bigwedge^k \xi$. (When $k = 0$, the fibre is $\mathbb{R}$ or $\mathbb{C}$).

(e) (*Symmetric Power*) If $\xi = (E, \pi, B, V)$ is a rank m vector bundle, then for any $k \geq 0$, we can define the symmetric power bundle, $\mathrm{Sym}^k \xi$, whose fibre over any $b \in \xi$ is the exterior power, $\mathrm{Sym}^k E_b$ and with typical fibre $\mathrm{Sym}^k V$. (When $k = 0$, the fibre is $\mathbb{R}$ or $\mathbb{C}$). The bundle $\mathrm{Sym}^k \xi$ is determined by the cocycle

$$\mathrm{Sym}^k g_{\alpha\beta}(b), \qquad b \in U_\alpha \cap U_\beta.$$

(f) (*Dual Bundle*) If $\xi = (E, \pi, B, V)$ is a rank m vector bundle, then its dual bundle, ξ^*, is the rank m vector bundle whose fibre over any $b \in B$ is the dual space, E_b^*, that is,

the vector bundle with typical fibre V^* (given by Theorem 7.7) specified by the cocycle whose matrices are

$$(g_{\alpha\beta}(b)^\top)^{-1}, \qquad b \in U_\alpha \cap U_\beta.$$

The reason for this seemingly complicated formula is this: For any trivialization, $\varphi_\alpha \colon \pi^{-1}(U_\alpha) \to U_\alpha \times V$, for any $b \in B$, recall that the restriction, $\varphi_{\alpha,b} \colon \pi^{-1}(b) \to V$, of φ_α to $\pi^{-1}(b)$ is a linear isomorphism. By dualization we get a map, $\varphi_{\alpha,b}^\top \colon V^* \to (\pi^{-1}(b))^*$, and thus, $\varphi_{\alpha,b}^*$ for ξ^* is given by

$$\varphi_{\alpha,b}^* = (\varphi_{\alpha,b}^\top)^{-1} \colon (\pi^{-1}(b))^* \to V^*.$$

As $g_{\alpha\beta}^*(b) = \varphi_{\alpha,b}^* \circ (\varphi_{\beta,b}^*)^{-1}$, we get

$$\begin{aligned} g_{\alpha\beta}^*(b) &= (\varphi_{\alpha,b}^\top)^{-1} \circ \varphi_{\beta,b}^\top \\ &= ((\varphi_{\beta,b}^\top)^{-1} \circ \varphi_{\alpha,b}^\top)^{-1} \\ &= (\varphi_{\beta,b}^{-1})^\top \circ \varphi_{\alpha,b}^\top)^{-1} \\ &= ((\varphi_{\alpha,b} \circ \varphi_{\beta,b}^{-1})^\top)^{-1} \\ &= (g_{\alpha\beta}(b)^\top)^{-1}, \end{aligned}$$

as claimed.

(g) (*$\mathcal{H}om$ Bundle*)

If $\xi = (E, \pi, B, V)$ is a rank m vector bundle and $\xi' = (E', \pi', B, W)$ is a rank n vector bundle, both over the same base, B, then their $\mathcal{H}om$ bundle, $\mathcal{H}om(\xi, \xi')$, is the rank mn vector bundle whose fibre over any $b \in B$ is $\mathrm{Hom}(E_b, E'_b)$, that is, the vector bundle with typical fibre $\mathrm{Hom}(V, W)$. The transition functions of this bundle are obtained as follows: For any trivializations, $\varphi_\alpha \colon \pi^{-1}(U_\alpha) \to U_\alpha \times V$ and $\varphi'_\alpha \colon (\pi')^{-1}(U_\alpha) \to U_\alpha \times W$, for any $b \in B$, recall that the restrictions, $\varphi_{\alpha,b} \colon \pi^{-1}(b) \to V$ and $\varphi'_{\alpha,b} \colon (\pi')^{-1}(b) \to W$ are linear isomorphisms. Then, we have a linear isomorphism, $\varphi_{\alpha,b}^{\mathrm{Hom}} \colon \mathrm{Hom}(\pi^{-1}(b), (\pi')^{-1}(b)) \longrightarrow \mathrm{Hom}(V, W)$, given by

$$\varphi_{\alpha,b}^{\mathrm{Hom}}(f) = \varphi'_{\alpha,b} \circ f \circ \varphi_{\alpha,b}^{-1}, \qquad f \in \mathrm{Hom}(\pi^{-1}(b), (\pi')^{-1}(b)).$$

Then, $g_{\alpha\beta}^{\mathrm{Hom}}(b) = \varphi_{\alpha,b}^{\mathrm{Hom}} \circ (\varphi_{\beta,b}^{\mathrm{Hom}})^{-1}$.

(h) (*Tensor Bundle of type (r, s)*)

If $\xi = (E, \pi, B, V)$ is a rank m vector bundle, then for any $r, s \geq 0$, we can define the bundle, $T^{r,s}\,\xi$, whose fibre over any $b \in \xi$ is the tensor space $T^{r,s}\,E_b$ and with typical fibre $T^{r,s}\,V$. The bundle $T^{r,s}\xi$ is determined by the cocycle

$$g_{\alpha\beta}^{\otimes^r}(b) \otimes ((g_{\alpha\beta}(b)^\top)^{-1})^{\otimes s}(b), \qquad b \in U_\alpha \cap U_\beta.$$

In view of the canonical isomorphism, $\mathrm{Hom}(V,W) \cong V^* \otimes W$, it is easy to show that $\mathcal{H}om(\xi,\xi')$, is isomorphic to $\xi^* \otimes \xi'$. Similarly, ξ^{**} is isomorphic to ξ. We also have the isomorphism

$$T^{r,s}\xi \cong \xi^{\otimes r} \otimes (\xi^*)^{\otimes s}.$$

Do not confuse the space of bundle morphisms, $\mathrm{Hom}(\xi,\xi')$, with the $\mathcal{H}om$ bundle, $\mathcal{H}om(\xi,\xi')$. However, observe that $\mathrm{Hom}(\xi,\xi')$ is the set of global sections of $\mathcal{H}om(\xi,\xi')$.

As an illustration of (d), consider the exterior power, $\bigwedge^r T^*M$, where M is a manifold of dimension n. We have trivialization maps, $\tau_U^*\colon \pi^{-1}(U) \to U \times \bigwedge^r(\mathbb{R}^n)^*$, for $\bigwedge^r T^*M$ given by

$$\tau_U^*(\omega) = (\pi(\omega), \bigwedge^r \theta^{\top}_{U,\varphi,\pi(\omega)}(\omega)),$$

for all $\omega \in \pi^{-1}(U)$. The transition function, $g_{\alpha\beta}^{\bigwedge^r}\colon U_\alpha \cap U_\beta \to \mathrm{GL}(n,\mathbb{R})$, is given by

$$g_{\alpha\beta}^{\bigwedge^r}(p)(\omega) = (\bigwedge^r(((\varphi_\alpha \circ \varphi_\beta^{-1})'_{\varphi(p)})^{\top})^{-1})(\omega),$$

for all $\omega \in \pi^{-1}(U)$. Consequently,

$$g_{\alpha\beta}^{\bigwedge^r}(p) = \bigwedge^r(g_{\alpha\beta}(p)^{\top})^{-1},$$

for every $p \in M$, a special case of (h).

For rank 1 vector bundles, that is, line bundles, it is easy to show that the set of equivalence classes of line bundles over a base, B, forms a group, where the group operation is $\otimes$, the inverse is $*$ (dual) and the identity element is the trivial bundle. This is the *Picard group* of B.

In general, the dual, E^*, of a bundle is *not* isomorphic to the original bundle, E. This is because, V^* is *not* canonically isomorphic to V and to get a bundle isomorphism between ξ and ξ^*, we need canonical isomorphisms between the fibres. However, if ξ is real, then (using a partition of unity) ξ can be given a Euclidean metric and so, ξ and ξ^* are isomorphic.

It is *not* true in general that a complex vector bundle is isomorphic to its dual because a Hermitian metric only induces a canonical isomorphism between E^* and $\overline{E}$, where $\overline{E}$ is the conjugate of E, with scalar multiplication in $\overline{E}$ given by $(z,w) \mapsto \overline{w}z$.

Remark: Given a real vector bundle, ξ, the *complexification, $\xi_{\mathbb{C}}$, of ξ* is the complex vector bundle defined by

$$\xi_{\mathbb{C}} = \xi \otimes_{\mathbb{R}} \epsilon_{\mathbb{C}},$$

where $\epsilon_{\mathbb{C}} = B \times \mathbb{C}$ is the trivial complex line bundle. Given a complex vector bundle, ξ, by viewing its fibre as a real vector space we obtain the real vector bundle, $\xi_{\mathbb{R}}$. The following facts can be shown:

(1) For every real vector bundle, ξ,

$$(\xi_{\mathbb{C}})_{\mathbb{R}} \cong \xi \oplus \xi.$$

(2) For every complex vector bundle, ξ,

$$(\xi_{\mathbb{R}})_{\mathbb{C}} \cong \xi \oplus \xi^*.$$

The notion of subbundle is defined as follows:

Definition 7.9. Given two vector bundles, $\xi = (E, \pi, B, V)$ and $\xi' = (E', \pi', B, V')$, over the same base, B, we say that ξ is a *subbundle* of ξ' iff E is a submanifold of E', V is a subspace of V' and for every $b \in B$, the fibre, $\pi^{-1}(b)$, is a subspace of the fibre, $(\pi')^{-1}(b)$.

If ξ is a subbundle of ξ', we can form the *quotient bundle*, ξ'/ξ, as the bundle over B whose fibre at $b \in B$ is the quotient space $(\pi')^{-1}(b)/\pi^{-1}(b)$. We leave it as an exercise to define trivializations for ξ'/ξ. In particular, if N is a submanifold of M, then TN is a subbundle of $TM \restriction N$ and the quotient bundle $(TM \restriction N)/TN$ is called the *normal bundle* of N in M.

7.4 Metrics on Bundles, Riemannian Manifolds, Reduction of Structure Groups, Orientation

Fortunately, the rich theory of vector spaces endowed with a Euclidean inner product can, to a great extent, be lifted to vector bundles.

Definition 7.10. Given a (real) rank n vector bundle, $\xi = (E, \pi, B, V)$, we say that ξ is *Euclidean* iff there is a family, $(\langle -, - \rangle_b)_{b \in B}$, of inner products on each fibre, $\pi^{-1}(b)$, such that $\langle -, - \rangle_b$ depends smoothly on b, which means that for every trivializing map, $\varphi_\alpha \colon \pi^{-1}(U_\alpha) \to U_\alpha \times V$, for every frame, $(s_1, \ldots, s_n)$, on U_α, the maps

$$b \mapsto \langle s_i(b), s_j(b) \rangle_b, \qquad b \in U_\alpha,\ 1 \leq i, j \leq n$$

are smooth. We say that $\langle -, - \rangle$ is a *Euclidean metric* (or *Riemannian metric*) on ξ. If ξ is a complex rank n vector bundle, $\xi = (E, \pi, B, V)$, we say that ξ is *Hermitian* iff there is a family, $(\langle -, - \rangle_b)_{b \in B}$, of Hermitian inner products on each fibre, $\pi^{-1}(b)$, such that $\langle -, - \rangle_b$ depends smoothly on b. We say that $\langle -, - \rangle$ is a *Hermitian metric* on ξ. For any smooth manifold, M, if TM is a Euclidean vector bundle, then we say that M is a *Riemannian manifold*.

If M is a Riemannian manifold, the smoothness condition on the metric, $\{\langle -, - \rangle_p\}_{p \in M}$, on TM, can be expressed a little more conveniently. If $\dim(M) = n$, then for every chart,

(U, φ), since $d\varphi^{-1}_{\varphi(p)} \colon \mathbb{R}^n \to T_pM$ is a bijection for every $p \in U$, the n-tuple of vector fields, $(s_1, \ldots, s_n)$, with $s_i(p) = d\varphi^{-1}_{\varphi(p)}(e_i)$, is a frame of TM over U, where $(e_1, \ldots, e_n)$ is the canonical basis of $\mathbb{R}^n$. Since every vector field over U is a linear combination, $\sum_{i=1}^n f_i s_i$, for some smooth functions, $f_i \colon U \to \mathbb{R}$, the condition of Definition 7.10 is equivalent to the fact that the maps,

$$p \mapsto \langle d\varphi^{-1}_{\varphi(p)}(e_i), d\varphi^{-1}_{\varphi(p)}(e_j)\rangle_p, \qquad p \in U,\ 1 \leq i, j \leq n,$$

are smooth. If we let $x = \varphi(p)$, the above condition says that the maps,

$$x \mapsto \langle d\varphi^{-1}_x(e_i), d\varphi^{-1}_x(e_j)\rangle_{\varphi^{-1}(x)}, \qquad x \in \varphi(U),\ 1 \leq i, j \leq n,$$

are smooth.

If M is a Riemannian manifold, the metric on TM is often denoted $g = (g_p)_{p\in M}$. In a chart, (U, φ), using local coordinates, we often use the notation, $g = \sum_{ij} g_{ij} dx_i \otimes dx_j$, or simply, $g = \sum_{ij} g_{ij} dx_i dx_j$, where

$$g_{ij}(p) = \left\langle \left(\frac{\partial}{\partial x_i}\right)_p, \left(\frac{\partial}{\partial x_j}\right)_p \right\rangle_p.$$

For every $p \in U$, the matrix, $(g_{ij}(p))$, is symmetric, positive definite.

The standard Euclidean metric on $\mathbb{R}^n$, namely,

$$g = dx_1^2 + \cdots + dx_n^2,$$

makes $\mathbb{R}^n$ into a Riemannian manifold. Then, every submanifold, M, of $\mathbb{R}^n$ inherits a metric by restricting the Euclidean metric to M. For example, the sphere, S^{n-1}, inherits a metric that makes S^{n-1} into a Riemannian manifold. It is a good exercise to find the local expression of this metric for S^2 in polar coordinates.

A nontrivial example of a Riemannian manifold is the *Poincaré upper half-space*, namely, the set $H = \{(x, y) \in \mathbb{R}^2 \mid y > 0\}$ equipped with the metric

$$g = \frac{dx^2 + dy^2}{y^2}.$$

A way to obtain a metric on a manifold, N, is to pull-back the metric, g, on another manifold, M, along a local diffeomorphism, $\varphi \colon N \to M$. Recall that φ is a local diffeomorphism iff

$$d\varphi_p \colon T_pN \to T_{\varphi(p)}M$$

is a bijective linear map for every $p \in N$. Given any metric g on M, if φ is a local diffeomorphism, we define the *pull-back metric*, φ^*g, on N induced by g as follows: For all $p \in N$, for all $u, v \in T_pN$,

$$(\varphi^*g)_p(u, v) = g_{\varphi(p)}(d\varphi_p(u), d\varphi_p(v)).$$

We need to check that $(\varphi^* g)_p$ is an inner product, which is very easy since $d\varphi_p$ is a linear isomorphism. Our map, φ, between the two Riemannian manifolds $(N, \varphi^* g)$ and (M, g) is a local isometry, as defined below.

Definition 7.11. Given two Riemannian manifolds, (M_1, g_1) and (M_2, g_2), a *local isometry* is a smooth map, $\varphi\colon M_1 \to M_2$, such that $d\varphi_p\colon T_pM_1 \to T_{\varphi(p)}M_2$ is an isometry between the Euclidean spaces $(T_pM_1, (g_1)_p)$ and $(T_{\varphi(p)}M_2, (g_2)_{\varphi(p)})$, for every $p \in M_1$, that is,

$$(g_1)_p(u, v) = (g_2)_{\varphi(p)}(d\varphi_p(u), d\varphi_p(v)),$$

for all $u, v \in T_pM_1$ or, equivalently, $\varphi^* g_2 = g_1$. Moreover, φ is an *isometry* iff it is a local isometry and a diffeomorphism.

The isometries of a Riemannian manifold, (M, g), form a group, $\mathrm{Isom}(M, g)$, called the *isometry group of* (M, g). An important theorem of Myers and Steenrod asserts that the isometry group, $\mathrm{Isom}(M, g)$, is a Lie group.

Given a map, $\varphi\colon M_1 \to M_2$, and a metric g_1 on M_1, in general, φ does not induce any metric on M_2. However, if φ has some extra properties, it does induce a metric on M_2. This is the case when M_2 arises from M_1 as a quotient induced by some group of isometries of M_1. For more on this, see Gallot, Hulin and Lafontaine [61], Chapter 2, Section 2.A.

Now, given a real (resp. complex) vector bundle, ξ, provided that B is a sufficiently nice topological space, namely that B is *paracompact* (see Section 3.6), a Euclidean metric (resp. Hermitian metric) exists on ξ. This is a consequence of the existence of partitions of unity (see Theorem 3.32).

Theorem 7.9. *Every real (resp. complex) vector bundle admits a Euclidean (resp. Hermitian) metric. In particular, every smooth manifold admits a Riemannian metric.*

Proof. Let (U_α) be a trivializing open cover for ξ and pick any frame, $(s_1^\alpha, \ldots, s_n^\alpha)$, over U_α. For every $b \in U_\alpha$, the basis, $(s_1^\alpha(b), \ldots, s_n^\alpha(b))$ defines a Euclidean (resp. Hermitian) inner product, $\langle -, - \rangle_b$, on the fibre $\pi^{-1}(b)$, by declaring $(s_1^\alpha(b), \ldots, s_n^\alpha(b))$ orthonormal w.r.t. this inner product. (For $x = \sum_{i=1}^n x_i s_i^\alpha(b)$ and $y = \sum_{i=1}^n y_i s_i^\alpha(b)$, let $\langle x, y \rangle_b = \sum_{i=1}^n x_i y_i$, resp. $\langle x, y \rangle_b = \sum_{i=1}^n x_i \overline{y}_i$, in the complex case.) The $\langle -, - \rangle_b$ (with $b \in U_\alpha$) define a metric on $\pi^{-1}(U_\alpha)$, denote it $\langle -, - \rangle_\alpha$. Now, using Theorem 3.32, glue these inner products using a partition of unity, (f_α), subordinate to (U_α), by setting

$$\langle x, y \rangle = \sum_\alpha f_\alpha \langle x, y \rangle_\alpha.$$

We verify immediately that $\langle -, - \rangle$ is a Euclidean (resp. Hermitian) metric on ξ. □

The existence of metrics on vector bundles allows the so-called reduction of structure group. Recall that the transition maps of a real (resp. complex) vector bundle, ξ, are functions, $g_{\alpha\beta}\colon U_\alpha \cap U_\beta \to \mathrm{GL}(n, \mathbb{R})$ (resp. $\mathrm{GL}(n, \mathbb{C})$). Let $\mathrm{GL}^+(n, \mathbb{R})$ be the subgroup of $\mathrm{GL}(n, \mathbb{R})$ consisting of those matrices of positive determinant (resp. $\mathrm{GL}^+(n, \mathbb{C})$ be the subgroup of $\mathrm{GL}(n, \mathbb{C})$ consisting of those matrices of positive determinant).

Definition 7.12. For every real (resp. complex) vector bundle, ξ, if it is possible to find a cocycle, $g = (g_{\alpha\beta})$, for ξ with values in a subgroup, H, of $\mathrm{GL}(n, \mathbb{R})$ (resp. of $\mathrm{GL}(n, \mathbb{C})$), then we say that the *structure group of ξ can be reduced to H*. We say that ξ is *orientable* if its structure group can be reduced to $\mathrm{GL}^+(n, \mathbb{R})$ (resp. $\mathrm{GL}^+(n, \mathbb{C})$).

Proposition 7.10. *(a) The structure group of a rank n real vector bundle, ξ, can be reduced to $\mathbf{O}(n)$; it can be reduced to $\mathbf{SO}(n)$ iff ξ is orientable.*

(b) The structure group of a rank n complex vector bundle, ξ, can be reduced to $\mathbf{U}(n)$; it can be reduced to $\mathbf{SU}(n)$ iff ξ is orientable.

Proof. We prove (a), the proof of (b) being similar. Using Theorem 7.9, put a metric on ξ. For every U_α in a trivializing cover for ξ and every $b \in B$, by Gram-Schmidt, orthonormal bases for $\pi^{-1}(b)$ exit. Consider the family of trivializing maps, $\widetilde{\varphi}_\alpha \colon \pi^{-1}(U_\alpha) \to U_\alpha \times V$, such that $\widetilde{\varphi}_{\alpha,b} \colon \pi^{-1}(b) \longrightarrow V$ maps orthonormal bases of the fibre to orthonormal bases of V. Then, it is easy to check that the corresponding cocycle takes values in $\mathbf{O}(n)$ and if ξ is orientable, the determinants being positive, these values are actually in $\mathbf{SO}(n)$. □

Remark: If ξ is a Euclidean rank n vector bundle, then by Proposition 7.10, we may assume that ξ is given by some cocycle, $(g_{\alpha\beta})$, where $g_{\alpha\beta}(b) \in \mathbf{O}(n)$, for all $b \in U_\alpha \cap U_\beta$. We saw in Section 7.3 (f) that the dual bundle, ξ^*, is given by the cocycle

$$(g_{\alpha\beta}(b)^\top)^{-1}, \qquad b \in U_\alpha \cap U_\beta.$$

As $g_{\alpha\beta}(b)$ is an orthogonal matrix, $(g_{\alpha\beta}(b)^\top)^{-1} = g_{\alpha\beta}(b)$, and thus, any Euclidean bundle is isomorphic to its dual. As we noted earlier, this is *false* for Hermitian bundles.

Let $\xi = (E, \pi, B, V)$ be a rank n vector bundle and assume ξ is orientable. A family of trivializing maps, $\varphi_\alpha \colon \pi^{-1}(U_\alpha) \to U_\alpha \times V$, is *oriented* iff for all α, β, the transition function, $g_{\alpha\beta}(b)$ has positive determinant for all $b \in U_\alpha \cap U_\beta$. Two oriented families of trivializing maps, $\varphi_\alpha \colon \pi^{-1}(U_\alpha) \to U_\alpha \times V$ and $\psi_\beta \colon \pi^{-1}(W_\beta) \to W_\alpha \times V$, are *equivalent* iff for every $b \in U_\alpha \cap W_\beta$, the map $pr_2 \circ \varphi_\alpha \circ \psi_\beta^{-1} \restriction \{b\} \times V \colon V \longrightarrow V$ has positive determinant. It is easily checked that this is an equivalence relation and that it partitions all the oriented families of trivializations of ξ into two equivalence classes. Either equivalence class is called an *orientation* of ξ.

If M is a manifold and $\xi = TM$, the tangent bundle of ξ, we know from Section 7.2 that the transition functions of TM are of the form

$$g_{\alpha\beta}(p)(u) = (\varphi_\alpha \circ \varphi_\beta^{-1})'_{\varphi(p)}(u),$$

where each $\varphi_\alpha \colon U_\alpha \to \mathbb{R}^n$ is a chart of M. Consequently, TM is orientable iff the Jacobian of $(\varphi_\alpha \circ \varphi_\beta^{-1})'_{\varphi(p)}$ is positive, for every $p \in M$. This is equivalent to the condition of Definition 3.31 for M to be orientable. Therefore, the tangent bundle, TM, of a manifold, M, is orientable iff M is orientable.

The notion of orientability of a vector bundle, $\xi = (E, \pi, B, V)$, is *not* equivalent to the orientability of its total space, E. Indeed, if we look at the transition functions of the total space of TM given in Section 7.2, we see that TM, *as a manifold*, is always orientable, even if M is not orientable. Yet, *as a bundle*, TM is orientable iff M.

On the positive side, if $\xi = (E, \pi, B, V)$ is an orientable vector bundle and its base, B, is an orientable manifold, then E is orientable too.

To see this, assume that B is a manifold of dimension m, ξ is a rank n vector bundle with fibre V, let $((U_\alpha, \psi_\alpha))_\alpha$ be an atlas for B, let $\varphi_\alpha \colon \pi^{-1}(U_\alpha) \to U_\alpha \times V$ be a collection of trivializing maps for ξ and pick any isomorphism, $\iota \colon V \to \mathbb{R}^n$. Then, we get maps,

$$(\psi_\alpha \times \iota) \circ \varphi_\alpha \colon \pi^{-1}(U_\alpha) \longrightarrow \mathbb{R}^m \times \mathbb{R}^n.$$

It is clear that these maps form an atlas for E. Check that the corresponding transition maps for E are of the form

$$(x, y) \mapsto (\psi_\beta \circ \psi_\alpha^{-1}(x), g_{\alpha\beta}(\psi_\alpha^{-1}(x))y).$$

Moreover, if B and ξ are orientable, check that these transition maps have positive Jacobian.

The fact that every bundle admits a metric allows us to define the notion of *orthogonal complement* of a subbundle. We state the following theorem without proof. The reader is invited to consult Milnor and Stasheff [111] for a proof (Chapter 3).

Proposition 7.11. *Let ξ and η be two vector bundles with ξ a subbundle of η. Then, there exists a subbundle, $\xi^\perp$, of η, such that every fibre of $\xi^\perp$ is the orthogonal complement of the fibre of ξ in the fibre of η, over every $b \in B$ and*

$$\eta \approx \xi \oplus \xi^\perp.$$

In particular, if N is a submanifold of a Riemannian manifold, M, then the orthogonal complement of TN in $TM \restriction N$ is isomorphic to the normal bundle, $(TM \restriction N)/TN$.

Remark: It can be shown (see Madsen and Tornehave [101], Chapter 15) that for every real vector bundle, ξ, there is some integer, k, such that ξ has a complement, η, in ϵ^k, where $\epsilon^k = B \times \mathbb{R}^k$ is the trivial rank k vector bundle, so that

$$\xi \oplus \eta = \epsilon^k.$$

This fact can be used to prove an interesting property of the space of global sections, $\Gamma(\xi)$. First, observe that $\Gamma(\xi)$ is not just a real vector space but also a $C^\infty(B)$-module (see Section 22.19). Indeed, for every smooth function, $f \colon B \to \mathbb{R}$, and every smooth section, $s \colon B \to E$, the map, $fs \colon B \to E$, given by

$$(fs)(b) = f(b)s(b), \qquad b \in B,$$

is a smooth section of ξ. In general, $\Gamma(\xi)$ is not a free $C^\infty(B)$-module unless ξ is trivial. However, the above remark implies that

$$\Gamma(\xi) \oplus \Gamma(\eta) = \Gamma(\epsilon^k),$$

where $\Gamma(\epsilon^k)$ is a free $C^\infty(B)$-module of dimension $\dim(\xi) + \dim(\eta)$. This proves that $\Gamma(\xi)$ is a finitely generated $C^\infty(B)$-module which is a summand of a free $C^\infty(B)$-module. Such modules are *projective modules*, see Definition 22.9 in Section 22.19. Therefore, $\Gamma(\xi)$ is a finitely generated projective $C^\infty(B)$-module. The following isomorphisms can be shown (see Madsen and Tornehave [101], Chapter 16):

Proposition 7.12. *The following isomorphisms hold for vector bundles:*

$$\begin{aligned}
\Gamma(\mathcal{H}om(\xi,\eta)) &\cong \mathrm{Hom}_{C^\infty(B)}(\Gamma(\xi),\Gamma(\eta)) \\
\Gamma(\xi\otimes\eta) &\cong \Gamma(\xi)\otimes_{C^\infty(B)}\Gamma(\eta) \\
\Gamma(\xi^*) &\cong \mathrm{Hom}_{C^\infty(B)}(\Gamma(\xi), C^\infty(B)) = (\Gamma(\xi))^* \\
\Gamma(\textstyle\bigwedge^k \xi) &\cong \bigwedge^k_{C^\infty(B)} (\Gamma(\xi)).
\end{aligned}$$

7.5 Principal Fibre Bundles

We now consider principal bundles. Such bundles arise in terms of Lie groups acting on manifolds.

Definition 7.13. Let G be a Lie group. A *principal fibre bundle*, for short, a *principal bundle*, is a fibre bundle, $\xi = (E, \pi, B, G, G)$, in which the fibre is G and the structure group is also G, viewed as its group of left translations (ie., G acts on itself by multiplication on the left). This means that every transition function, $g_{\alpha\beta}\colon U_\alpha \cap U_\beta \to G$, satisfies

$$g_{\alpha\beta}(b)(h) = g(b)h, \qquad \text{for some } g(b) \in G,$$

for all $b \in U_\alpha \cap U_\beta$ and all $h \in G$. A principal G-bundle is denoted $\xi = (E, \pi, B, G)$.

Note that G in $g_{\alpha\beta}\colon U_\alpha \cap U_\beta \to G$ is viewed as its group of left translations under the isomorphism, $g \mapsto L_g$, and so, $g_{\alpha\beta}(b)$ is some left translation, $L_g(b)$. The inverse of the above isomorphism is given by $L \mapsto L(1)$, so $g(b) = g_{\alpha\beta}(b)(1)$. In view of these isomorphisms, we allow ourself the (convenient) abuse of notation

$$g_{\alpha\beta}(b)(h) = g_{\alpha\beta}(b)h,$$

where, on the left, $g_{\alpha\beta}(b)$ is viewed as a left translation of G and on the right, as an element of G.

When we want to emphasize that a principal bundle has structure group, G, we use the locution *principal G-bundle*.

It turns out that if $\xi = (E, \pi, B, G)$ is a principal bundle, then G acts on the total space, E, on the right. For the next proposition, recall that a right action, $\cdot\colon X \times G \to X$, is *free* iff for every $g \in G$, if $g \neq 1$, then $x \cdot g \neq x$ for all $x \in X$.

Proposition 7.13. *If $\xi = (E, \pi, B, G)$ is a principal bundle, then there is a right action of G on E. This action takes each fibre to itself and is free. Moreover, E/G is diffeomorphic to B.*

Proof. We show how to define the right action and leave the rest as an exercise. Let $\{(U_\alpha, \varphi_\alpha)\}$ be some trivializing cover defining ξ. For every $z \in E$, pick some U_α so that $\pi(z) \in U_\alpha$ and let $\varphi_\alpha(z) = (b, h)$, where $b = \pi(z)$ and $h \in G$. For any $g \in G$, we set

$$z \cdot g = \varphi_\alpha^{-1}(b, hg).$$

If we can show that this action does not depend on the choice of U_α, then it is clear that it is a free action. Suppose that we also have $b = \pi(z) \in U_\beta$ and that $\varphi_\beta(z) = (b, h')$. By definition of the transition functions, we have

$$h' = g_{\beta\alpha}(b)h \quad \text{and} \quad \varphi_\beta(z \cdot g) = (b, g_{\beta\alpha}(b)(hg)).$$

However,

$$g_{\beta\alpha}(b)(hg) = (g_{\beta\alpha}(b)h)g = h'g,$$

hence

$$z \cdot g = \varphi_\beta^{-1}(b, h'g),$$

which proves that our action does not depend on the choice of U_α. □

Observe that the action of Proposition 7.13 is defined by

$$z \cdot g = \varphi_\alpha^{-1}(b, \varphi_{\alpha,b}(z)g), \quad \text{with} \quad b = \pi(z),$$

for all $z \in E$ and all $g \in G$. It is clear that this action satisfies the following two properties: For every $(U_\alpha, \varphi_\alpha)$,

(1) $\pi(z \cdot g) = \pi(z)$ and

(2) $\varphi_\alpha(z \cdot g) = \varphi_\alpha(z) \cdot g$, for all $z \in E$ and all $g \in G$,

where we define the right action of G on $U_\alpha \times G$ so that $(b, h) \cdot g = (b, hg)$. We say that φ_α is *G-equivariant* (or *equivariant*).

The following proposition shows that it is possible to define a principal G-bundle using a suitable right action and equivariant trivializations:

Proposition 7.14. *Let E be a smooth manifold, G a Lie group and let $\cdot\colon E \times G \to E$ be a smooth right action of G on E and assume that*

(a) The right action of G on E is free;

(b) The orbit space, $B = E/G$, is a smooth manifold under the quotient topology and the projection, $\pi\colon E \to E/G$, is smooth;

(c) There is a family of local trivializations, $\{(U_\alpha, \varphi_\alpha)\}$, where $\{U_\alpha\}$ is an open cover for $B = E/G$ and each

$$\varphi_\alpha\colon \pi^{-1}(U_\alpha) \to U_\alpha \times G$$

is an equivariant diffeomorphism, which means that

$$\varphi_\alpha(z \cdot g) = \varphi_\alpha(z) \cdot g,$$

for all $z \in \pi^{-1}(U_\alpha)$ and all $g \in G$, where the right action of G on $U_\alpha \times G$ is $(b, h) \cdot g = (b, hg)$.

Then, $\xi = (E, \pi, E/G, G)$ is a principal G-bundle.

Proof. Since the action of G on E is free, every orbit, $b = z \cdot G$, is isomorphic to G and so, every fibre, $\pi^{-1}(b)$, is isomorphic to G. Thus, given that we have trivializing maps, we just have to prove that G acts by left translation on itself. Pick any (b, h) in $U_\beta \times G$ and let $z \in \pi^{-1}(U_\beta)$ be the unique element such that $\varphi_\beta(z) = (b, h)$. Then, as

$$\varphi_\beta(z \cdot g) = \varphi_\beta(z) \cdot g, \quad \text{for all } g \in G,$$

we have

$$\varphi_\beta(\varphi_\beta^{-1}(b, h) \cdot g) = \varphi_\beta(z \cdot g) = \varphi_\beta(z) \cdot g = (b, h) \cdot g,$$

which implies that

$$\varphi_\beta^{-1}(b, h) \cdot g = \varphi_\beta^{-1}((b, h) \cdot g).$$

Consequently,

$$\varphi_\alpha \circ \varphi_\beta^{-1}(b, h) = \varphi_\alpha \circ \varphi_\beta^{-1}((b, 1) \cdot h) = \varphi_\alpha(\varphi_\beta^{-1}(b, 1) \cdot h) = \varphi_\alpha \circ \varphi_\beta^{-1}(b, 1) \cdot h,$$

and since

$$\varphi_\alpha \circ \varphi_\beta^{-1}(b, h) = (b, g_{\alpha\beta}(b)(h)) \quad \text{and} \quad \varphi_\alpha \circ \varphi_\beta^{-1}(b, 1) = (b, g_{\alpha\beta}(b)(1))$$

we get

$$g_{\alpha\beta}(b)(h) = g_{\alpha\beta}(b)(1)h.$$

The above shows that $g_{\alpha\beta}(b)\colon G \to G$ is the left translation by $g_{\alpha\beta}(b)(1)$ and thus, the transition functions, $g_{\alpha\beta}(b)$, constitute the group of left translations of G and ξ is indeed a principal G-bundle. □

Bröcker and tom Dieck [25] (Chapter I, Section 4) and Duistermaat and Kolk [54] (Appendix A) define principal bundles using the conditions of Proposition 7.14. Propositions 7.13 and 7.14 show that this alternate definition is equivalent to ours (Definition 7.13).

It turns out that when we use the definition of a principal bundle in terms of the conditions of Proposition 7.14, it is convenient to define bundle maps in terms of equivariant maps. As we will see shortly, a map of principal bundles is a fibre bundle map.

Definition 7.14. If $\xi_1 = (E_1, \pi_1, B_1, G)$ and $\xi_2 = (E_2, \pi_2, B_1, G)$ are two principal bundles a *bundle map (or bundle morphism)*, $f\colon \xi_1 \to \xi_2$, is a pair, $f = (f_E, f_B)$, of smooth maps $f_E\colon E_1 \to E_2$ and $f_B\colon B_1 \to B_2$ such that

(a) The following diagram commutes:

$$\begin{array}{ccc} E_1 & \xrightarrow{f_E} & E_2 \\ {\scriptstyle \pi_1}\downarrow & & \downarrow{\scriptstyle \pi_2} \\ B_1 & \xrightarrow[f_B]{} & B_2 \end{array}$$

(b) The map, f_E, is G-*equivariant*, that is,

$$f_E(a \cdot g) = f_E(a) \cdot g, \qquad \text{for all } a \in E_1 \text{ and all } g \in G.$$

A bundle map is an *isomorphism* if it has an inverse as in Definition 7.2. If the bundles ξ_1 and ξ_2 are over the same base, B, then we also require $f_B = \mathrm{id}$.

At first glance, it is not obvious that a map of principal bundles satisfies condition (b) of Definition 7.3. If we define $\widetilde{f}_\alpha\colon U_\alpha \times G \to V_\beta \times G$ by

$$\widetilde{f}_\alpha = \varphi'_\beta \circ f_E \circ \varphi_\alpha^{-1},$$

then locally, f_E is expressed as

$$f_E = {\varphi'_\beta}^{-1} \circ \widetilde{f}_\alpha \circ \varphi_\alpha.$$

Furthermore, it is trivial that if a map is equivariant and invertible then its inverse is equivariant. Consequently, since

$$\widetilde{f}_\alpha = \varphi'_\beta \circ f_E \circ \varphi_\alpha^{-1},$$

as φ_α^{-1}, φ'_β and f_E are equivariant, $\widetilde{f}_\alpha$ is also equivariant and so, $\widetilde{f}_\alpha$ is a map of (trivial) principal bundles. Thus, it it enough to prove that for every map of principal bundles,

$$\varphi\colon U_\alpha \times G \to V_\beta \times G,$$

there is some smooth map, $\rho_\alpha\colon U_\alpha \to G$, so that

$$\varphi(b, g) = (f_B(b), \rho_\alpha(b)(g)), \qquad \text{for all } b \in U_\alpha \text{ and all } g \in G.$$

Indeed, we have the following

Proposition 7.15. *For every map of trivial principal bundles,*

$$\varphi\colon U_\alpha \times G \to V_\beta \times G,$$

there are smooth maps, $f_B\colon U_\alpha \to V_\beta$ and $r_\alpha\colon U_\alpha \to G$, so that

$$\varphi(b, g) = (f_B(b), r_\alpha(b)g), \qquad \textit{for all } b \in U_\alpha \textit{ and all } g \in G.$$

In particular, φ is a diffeomorphism on fibres.

Proof. As φ is a map of principal bundles,

$$\varphi(b, 1) = (f_B(b), r_\alpha(b)), \qquad \text{for all } b \in U_\alpha$$

for some smooth maps, $f_B\colon U_\alpha \to V_\beta$ and $r_\alpha\colon U_\alpha \to G$. Now, using equivariance, we get

$$\varphi(b, g) = \varphi((b, 1)g) = \varphi(g, 1) \cdot g = (f_B(b), r_\alpha(b)) \cdot g = (f_B(b), r_\alpha(b)g),$$

as claimed. □

Consequently, the map, $\rho_\alpha\colon U_\alpha \to G$, given by

$$\rho_\alpha(b)(g) = r_\alpha(b)g \qquad \text{for all } b \in U_\alpha \text{ and all } g \in G$$

satisfies

$$\varphi(b, g) = (f_B(b), \rho_\alpha(b)(g)), \qquad \text{for all } b \in U_\alpha \text{ and all } g \in G$$

and a map of principal bundles is indeed a fibre bundle map (as in Definition 7.3). Since a principal bundle map is a fibre bundle map, Proposition 7.3 also yields the useful fact:

Proposition 7.16. *Any map, $f\colon \xi_1 \to \xi_2$, between two principal bundles over the same base, B, is an isomorphism.*

Even though we are not aware of any practical applications in computer vision, robotics, or medical imaging, we wish to digress briefly on the issue of the triviality of bundles and the existence of sections.

A natural question is to ask whether a fibre bundle, ξ, is isomorphic to a trivial bundle. If so, we say that ξ is trivial. (By the way, the triviality of bundles comes up in physics, in particular, field theory.) Generally, this is a very difficult question, but a first step can be made by showing that it reduces to the question of triviality for principal bundles.

Indeed, if $\xi = (E, \pi, B, F, G)$ is a fibre bundle with fibre, F, using Theorem 7.4, we can construct a principal fibre bundle, $P(\xi)$, using the transition functions, $\{g_{\alpha\beta}\}$, of ξ, but using G itself as the fibre (acting on itself by left translation) instead of F. We obtain the *principal bundle, $P(\xi)$, associated to ξ*. For example, the principal bundle associated with a vector bundle is the *frame bundle*, discussed at the end of Section 7.3. Then, given two

fibre bundles ξ and ξ', we see that ξ and ξ' are isomorphic iff $P(\xi)$ and $P(\xi')$ are isomorphic (Steenrod [142], Part I, Section 8, Theorem 8.2). More is true: The fibre bundle ξ is trivial iff the principal fibre bundle $P(\xi)$ is trivial (this is easy to prove, do it! Otherwise, see Steenrod [142], Part I, Section 8, Corollary 8.4). Morever, there is a test for the triviality of a principal bundle, the existence of a (global) section.

The following proposition, although easy to prove, is crucial:

Proposition 7.17. *If ξ is a principal bundle, then ξ is trivial iff it possesses some global section.*

Proof. If $f\colon B \times G \to \xi$ is an isomorphism of principal bundles over the same base, B, then for every $g \in G$, the map $b \mapsto f(b, g)$ is a section of ξ.

Conversely, let $s\colon B \to E$ be a section of ξ. Then, observe that the map, $f\colon B \times G \to \xi$, given by

$$f(b, g) = s(b)g$$

is a map of principal bundles. By Proposition 7.16, it is an isomorphism, so ξ is trivial. □

Generally, in geometry, many objects of interest arise as global sections of some suitable bundle (or sheaf): vector fields, differential forms, tensor fields, *etc.*

Given a principal bundle, $\xi = (E, \pi, B, G)$, and given a manifold, F, if G acts effectively on F from the left, again, using Theorem 7.4, we can construct a fibre bundle, $\xi[F]$, from ξ, with F as typical fibre and such that $\xi[F]$ has the same transitions functions as ξ. In the case of a principal bundle, there is another slightly more direct construction that takes us from principal bundles to fibre bundles (see Duistermaat and Kolk [54], Chapter 2, and Davis and Kirk [40], Chapter 4, Definition 4.6, where it is called the *Borel construction*). This construction is of independent interest so we describe it briefly (for an application of this construction, see Duistermaat and Kolk [54], Chapter 2).

As ξ is a principal bundle, recall that G acts on E from the right, so we have a right action of G on $E \times F$, *via*

$$(z, f) \cdot g = (z \cdot g, g^{-1} \cdot f).$$

Consequently, we obtain the orbit set, $E \times F/\sim$, denoted $E \times_G F$, where $\sim$ is the equivalence relation

$$(z, f) \sim (z', f') \quad \text{iff} \quad (\exists g \in G)(z' = z \cdot g,\ f' = g^{-1} \cdot f).$$

Note that the composed map,

$$E \times F \xrightarrow{pr_1} E \xrightarrow{\pi} B,$$

factors through $E \times_G F$, since

$$\pi(pr_1(z, f)) = \pi(z) = \pi(z \cdot g) = \pi(pr_1(z \cdot g, g^{-1} \cdot f)).$$

Let $p\colon E \times_G F \to B$ be the corresponding map. The following proposition is not hard to show:

Proposition 7.18. *If $\xi = (E, \pi, B, G)$ is a principal bundle and F is any manifold such that G acts effectively on F from the left, then, $\xi[F] = (E \times_G F, p, B, F, G)$ is a fibre bundle with fibre F and structure group G and $\xi[F]$ and ξ have the same transition functions.*

Let us verify that the charts of ξ yield charts for $\xi[F]$. For any U_α in an open cover for B, we have a diffeomorphism

$$\varphi_\alpha \colon \pi^{-1}(U_\alpha) \to U_\alpha \times G.$$

Observe that we have an isomorphism

$$(U_\alpha \times G) \times_G F \cong U_\alpha \times F,$$

where, as usual, G acts on $U_\alpha \times G$ *via* $(z, h) \cdot g = (z, hg)$, an isomorphism

$$p^{-1}(U_\alpha) \cong \pi^{-1}(U_\alpha) \times_G F,$$

and that φ_α induces an isomorphism

$$\pi^{-1}(U_\alpha) \times_G F \xrightarrow{\varphi_\alpha} (U_\alpha \times G) \times_G F.$$

So, we get the commutative diagram

$$\begin{array}{ccc} p^{-1}(U_\alpha) & \xrightarrow{\sim} & U_\alpha \times F \\ {\scriptstyle p}\downarrow & & \downarrow{\scriptstyle pr_1} \\ U_\alpha & = & U_\alpha, \end{array}$$

which yields a local trivialization for $\xi[F]$. It is easy to see that the transition functions of $\xi[F]$ are the same as the transition functions of ξ.

The fibre bundle, $\xi[F]$, is called the fibre bundle *induced by* ξ. Now, if we start with a fibre bundle, ξ, with fibre, F, and structure group, G, if we make the associated principal bundle, $P(\xi)$, and then the induced fibre bundle, $P(\xi)[F]$, what is the relationship between ξ and $P(\xi)[F]$?

The answer is: ξ and $P(\xi)[F]$ are *equivalent* (this is because the transition functions are the same.)

Now, if we start with a principal G-bundle, ξ, make the fibre bundle, $\xi[F]$, as above, and then the principal bundle, $P(\xi[F])$, we get a principal bundle equivalent to ξ. Therefore, the maps

$$\xi \mapsto \xi[F] \quad \text{and} \quad \xi \mapsto P(\xi),$$

are mutual inverses and they set up a bijection between equivalence classes of principal G-bundles over B and equivalence classes of fibre bundles over B (with structure group, G). Moreover, this map extends to morphisms, so it is functorial (see Steenrod [142], Part I,

Section 2, Lemma 2.6–Lemma 2.10). As a consequence, in order to "classify" equivalence classes of fibre bundles (assuming B and G fixed), it is enough to know how to classify principal G-bundles over B. Given some reasonable conditions on the coverings of B, Milnor solved this classification problem, but this is taking us way beyond the scope of these notes!

The classical reference on fibre bundles, vector bundles and principal bundles, is Steenrod [142]. More recent references include Bott and Tu [19], Madsen and Tornehave [101], Morita [115], Griffith and Harris [67], Wells [151], Hirzebruch [78], Milnor and Stasheff [111], Davis and Kirk [40], Atiyah [10], Chern [33], Choquet-Bruhat, DeWitt-Morette and Dillard-Bleick [37], Hirsh [77], Sato [134], Narasimham [118], Sharpe [140] and also Husemoller [83], which covers more, including characteristic classes.

Proposition 7.14 shows that principal bundles are induced by suitable right actions but we still need sufficient conditions to guarantee conditions (a), (b) and (c). Such conditions are given in the next section.

7.6 Homogeneous Spaces, II

Now that we know about manifolds and Lie groups, we can revisit the notion of homogeneous space given in Definition 2.8, which only applied to groups and sets without any topology or differentiable structure.

Definition 7.15. A *homogeneous space* is a smooth manifold, M, together with a smooth transitive action, $\cdot\colon G \times M \to M$, of a Lie group, G, on M.

In this section, we prove that G is the total space of a principal bundle with base space M and structure group, G_x, the stabilizer of any $x \in M$.

If M is a manifold, G is a Lie group and $\cdot\colon M \times G \to M$ is a smooth right action, in general, M/G is not even Hausdorff. A sufficient condition can be given using the notion of a proper map. If X and Y are two Hausdorff topological spaces,[1] a continuous map, $\varphi\colon X \to Y$, is *proper* iff for every topological space, Z, the map $\varphi \times \mathrm{id}\colon X \times Z \to Y \times Z$ is a *closed map* (A map, f, is a closed map iff the image of any closed set by f is a closed set). If we let Z be a one-point space, we see that a proper map is closed. It can be shown (see Bourbaki, General Topology [23], Chapter 1, Section 10) that a continuous map, $\varphi\colon X \to Y$, is proper iff φ is closed and if $\varphi^{-1}(y)$ is compact for every $y \in Y$. If φ is proper, it is easy to show that $\varphi^{-1}(K)$ is compact in X whenever K is compact in Y. Moreover, if Y is also locally compact, then Y is compactly generated, which means that a subset, C, of Y is closed iff $K \cap C$ is closed in C for every compact subset K of Y (see Munkres [116]). In this case (Y locally compact), φ is a closed map iff $\varphi^{-1}(K)$ is compact in X whenever K is compact

[1] It is not necessary to assume that X and Y are Hausdorff but, if X and/or Y are not Hausdorff, we have to replace "compact" by "quasi-compact." We have no need for this extra generality.

in Y (see Bourbaki, General Topology [23], Chapter 1, Section 10).[2] In particular, this is true if Y is a manifold since manifolds are locally compact. Then, we say that the action, $\cdot\colon M \times G \to M$, is *proper* iff the map,

$$M \times G \longrightarrow M \times M, \quad (x, g) \mapsto (x, x \cdot g),$$

is proper.

If G and M are Hausdorff and G is locally compact, then it can be shown (see Bourbaki, General Topology [23], Chapter 3, Section 4) that the action $\cdot\colon M \times G \to M$ is proper iff for all $x, y \in M$, there exist some open sets, V_x and V_y in M, with $x \in V_x$ and $y \in V_y$, so that the closure, $\overline{K}$, of the set $K = \{g \in G \mid V_x \cdot g \cap V_y \neq \emptyset\}$ is compact in G. In particular, if G has the discrete topology, this conditions holds iff the sets $\{g \in G \mid V_x \cdot g \cap V_y \neq \emptyset\}$ are finite. Also, if G is compact, then $\overline{K}$ is automatically compact, so every compact group acts properly. If the action, $\cdot\colon M \times G \to M$, is proper, then the orbit equivalence relation is closed since it is the image of $M \times G$ in $M \times M$, and so, M/G is Hausdorff. We then have the following theorem proved in Duistermaat and Kolk [54] (Chapter 1, Section 11):

Theorem 7.19. *Let M be a smooth manifold, G be a Lie group and let $\cdot\colon M \times G \to M$ be a right smooth action which is proper and free. Then, M/G is a principal G-bundle of dimension $\dim M - \dim G$.*

Theorem 7.19 has some interesting corollaries. Let G be a Lie group and let H be a closed subgroup of G. Then, there is a right action of H on G, namely

$$G \times H \longrightarrow G, \quad (g, h) \mapsto gh,$$

and this action is clearly free and proper. Because a closed subgroup of a Lie group is a Lie group, we get the following result whose proof can be found in Bröcker and tom Dieck [25] (Chapter I, Section 4) or Duistermaat and Kolk [54] (Chapter 1, Section 11):

Corollary 7.20. *If G is a Lie group and H is a closed subgroup of G, then, the right action of H on G defines a principal H-bundle, $\xi = (G, \pi, G/H, H)$, where $\pi\colon G \to G/H$ is the canonical projection. Moreover, π is a submersion, which means that $d\pi_g$ is surjective for all $g \in G$ (equivalently, the rank of $d\pi_g$ is constant and equal to $\dim G/H$, for all $g \in G$).*

Now, if $\cdot\colon G \times M \to M$ is a smooth transitive action of a Lie group, G, on a manifold, M, we know that the stabilizers, G_x, are all isomorphic and closed (see Section 2.5, Remark after Theorem 2.26). Then, we can let $H = G_x$ and apply Corollary 7.20 to get the following result (mostly proved in in Bröcker and tom Dieck [25] (Chapter I, Section 4):

Proposition 7.21. *Let $\cdot\colon G \times M \to M$ be smooth transitive action of a Lie group, G, on a manifold, M. Then, G/G_x and M are diffeomorphic and G is the total space of a principal bundle, $\xi = (G, \pi, M, G_x)$, where G_x is the stabilizer of any element $x \in M$.*

[2] Duistermaat and Kolk [54] seem to have overlooked the fact that a condition on Y (such as local compactness) is needed in their remark on lines 5-6, page 53, just before Lemma 1.11.3.

Thus, we finally see that homogeneous spaces induce principal bundles. Going back to some of the examples of Section 2.2, we see that

(1) $\mathbf{SO}(n+1)$ is a principal $\mathbf{SO}(n)$-bundle over the sphere S^n (for $n \geq 0$).

(2) $\mathbf{SU}(n+1)$ is a principal $\mathbf{SU}(n)$-bundle over the sphere S^{2n+1} (for $n \geq 0$).

(3) $\mathbf{SL}(2,\mathbb{R})$ is a principal $\mathbf{SO}(2)$-bundle over the upper-half space, H.

(4) $\mathbf{GL}(n,\mathbb{R})$ is a principal $\mathbf{O}(n)$-bundle over the space $\mathbf{SPD}(n)$ of symmetric, positive definite matrices.

(5) $\mathbf{GL}^+(n,\mathbb{R})$, is a principal $\mathbf{SO}(n)$-bundle over the space, $\mathbf{SPD}(n)$, of symmetric, positive definite matrices, with fibre $\mathbf{SO}(n)$.

(6) $\mathbf{SO}(n+1)$ is a principal $\mathbf{O}(n)$-bundle over the real projective space $\mathbb{RP}^n$ (for $n \geq 0$).

(7) $\mathbf{SU}(n+1)$ is a principal $\mathbf{U}(n)$-bundle over the complex projective space $\mathbb{CP}^n$ (for $n \geq 0$).

(8) $\mathbf{O}(n)$ is a principal $\mathbf{O}(k) \times \mathbf{O}(n-k)$-bundle over the Grassmannian, $G(k,n)$.

(9) $\mathbf{SO}(n)$ is a principal $S(\mathbf{O}(k) \times \mathbf{O}(n-k))$-bundle over the Grassmannian, $G(k,n)$.

(10) From Section 2.5, we see that the Lorentz group, $\mathbf{SO}_0(n,1)$, is a principal $\mathbf{SO}(n)$-bundle over the space, $\mathcal{H}_n^+(1)$, consisting of one sheet of the hyperbolic paraboloid $\mathcal{H}_n(1)$.

Thus, we see that both $\mathbf{SO}(n+1)$ and $\mathbf{SO}_0(n,1)$ are principal $\mathbf{SO}(n)$-bundles, the difference being that the base space for $\mathbf{SO}(n+1)$ is the sphere, S^n, which is compact, whereas the base space for $\mathbf{SO}_0(n,1)$ is the (connected) surface, $\mathcal{H}_n^+(1)$, which is not compact. Many more examples can be given, for instance, see Arvanitoyeogos [8].

Chapter 8

Differential Forms

8.1 Differential Forms on Subsets of $\mathbb{R}^n$ and de Rham Cohomology

The theory of differential forms is one of the main tools in geometry and topology. This theory has a surprisingly large range of applications and it also provides a relatively easy access to more advanced theories such as cohomology. For all these reasons, it is really an indispensable theory and anyone with more than a passible interest in geometry should be familiar with it.

The theory of differential forms was initiated by Poincaré and further elaborated by Elie Cartan at the end of the nineteenth century. Differential forms have two main roles:

(1) Describe various systems of partial differential equations on manifolds.

(2) To define various geometric invariants reflecting the global structure of manifolds or bundles. Such invariants are obtained by integrating certain differential forms.

As we will see shortly, as soon as one tries to define integration on higher-dimensional objects, such as manifolds, one realizes that it is not functions that are integrated but instead, differential forms. Furthermore, as by magic, the algebra of differential forms handles changes of variables automatically and yields a neat form of "Stokes formula".

Our goal is to define differential forms on manifolds but we begin with differential forms on open subsets of $\mathbb{R}^n$ in order to build up intuition.

Differential forms are smooth functions on open subset, U, of $\mathbb{R}^n$, taking as values alternating tensors in some exterior power, $\bigwedge^p(\mathbb{R}^n)^*$. Recall from Sections 22.14 and 22.15, in particular, Proposition 22.24, that for every finite-dimensional vector space, E, the isomorphisms, $\mu\colon \bigwedge^n(E^*) \longrightarrow \mathrm{Alt}^n(E;\mathbb{R})$, induced by the linear extensions of the maps given by

$$\mu(v_1^* \wedge \cdots \wedge v_n^*)(u_1, \ldots, u_n) = \det(u_j^*(u_i))$$

yield a canonical isomorphism of algebras, $\mu\colon \bigwedge(E^*) \longrightarrow \mathrm{Alt}(E)$, where

$$\mathrm{Alt}(E) = \bigoplus_{n\geq 0} \mathrm{Alt}^n(E;\mathbb{R})$$

and where $\mathrm{Alt}^n(E;\mathbb{R})$ is the vector space of alternating multilinear maps on $\mathbb{R}^n$. In view of these isomorphisms, we will identify ω and $\mu(\omega)$ for any $\omega \in \bigwedge^n(E^*)$ and we will write $\omega(u_1,\ldots,u_n)$ as an abbrevation for $\mu(\omega)(u_1,\ldots,u_n)$.

Because $\mathrm{Alt}(\mathbb{R}^n)$ is an algebra under the wedge product, differential forms also have a wedge product. However, the power of differential forms stems from the *exterior differential*, d, which is a skew-symmetric version of the usual differentiation operator.

Definition 8.1. Given any open subset, U, of $\mathbb{R}^n$, a smooth *differential p-form on U*, for short, *p-form on U*, is any smooth function, $\omega\colon U \to \bigwedge^p(\mathbb{R}^n)^*$. The vector space of all p-forms on U is denoted $\mathcal{A}^p(U)$. The vector space, $\mathcal{A}^*(U) = \bigoplus_{p\geq 0} \mathcal{A}^p(U)$, is the set of *differential forms on U*.

Observe that $\mathcal{A}^0(U) = C^\infty(U,\mathbb{R})$, the vector space of smooth functions on U and $\mathcal{A}^1(U) = C^\infty(U,(\mathbb{R}^n)^*)$, the set of smooth functions from U to the set of linear forms on $\mathbb{R}^n$. Also, $\mathcal{A}^p(U) = (0)$ for $p > n$.

Remark: The space, $\mathcal{A}^*(U)$, is also denoted $\mathcal{A}^\bullet(U)$. Other authors use $\Omega^p(U)$ instead of $\mathcal{A}^p(U)$ but we prefer to reserve Ω^p for holomorphic forms.

Recall from Section 22.12 that if $(e_1,\ldots,e_n)$ is any basis of $\mathbb{R}^n$ and $(e_1^*,\ldots,e_n^*)$ is its dual basis, then the alternating tensors,

$$e_I^* = e_{i_1}^* \wedge \cdots \wedge e_{i_p}^*,$$

form basis of $\bigwedge^p(\mathbb{R}^n)^*$, where $I = \{i_1,\ldots,i_p\} \subseteq \{1,\ldots,n\}$, with $i_1 < \cdots < i_p$. Thus, with respect to the basis $(e_1,\ldots,e_n)$, every p-form, ω, can be uniquely written

$$\omega(x) = \sum_I f_I(x)\, e_{i_1}^* \wedge \cdots \wedge e_{i_p}^* = \sum_I f_I(x)\, e_I^* \qquad x \in U,$$

where each f_I is a smooth function on U. For example, if $U = \mathbb{R}^2 - \{0\}$, then

$$\omega(x,y) = \frac{-y}{x^2+y^2}\, e_1^* + \frac{x}{x^2+y^2}\, e_2^*$$

is a 2-form on U, (with $e_1 = (1,0)$ and $e_2 = (0,1)$).

We often write ω_x instead of $\omega(x)$. Now, not only is $\mathcal{A}^*(U)$ a vector space, it is also an algebra.

Definition 8.2. The *wedge product* on $\mathcal{A}^*(U)$ is defined as follows: For all $p, q \geq 0$, the wedge product, $\wedge\colon \mathcal{A}^p(U) \times \mathcal{A}^q(U) \to \mathcal{A}^{p+q}(U)$, is given by

$$(\omega \wedge \eta)(x) = \omega(x) \wedge \eta(x), \qquad x \in U.$$

For example, if ω and η are one-forms, then

$$(\omega \wedge \eta)_x(u, v) = \omega_x(u) \wedge \eta_x(v) - \omega_x(v) \wedge \eta_x(u).$$

For $f \in \mathcal{A}^0(U) = C^\infty(U, \mathbb{R})$ and $\omega \in \mathcal{A}^p(U)$, we have $f \wedge \omega = f\omega$. Thus, the algebra, $\mathcal{A}^*(U)$, is also a $C^\infty(U, \mathbb{R})$-module,

Proposition 22.22 immediately yields

Proposition 8.1. *For all forms $\omega \in \mathcal{A}^p(U)$ and $\eta \in \mathcal{A}^q(U)$, we have*

$$\eta \wedge \omega = (-1)^{pq} \omega \wedge \eta.$$

We now come to the crucial operation of exterior differentiation. First, recall that if $f\colon U \to V$ is a smooth function from $U \subseteq \mathbb{R}^n$ to a (finite-dimensional) normed vector space, V, the derivative, $f'\colon U \to \mathrm{Hom}(\mathbb{R}^n, V)$, of f (also denoted, Df) is a function where $f'(x)$ is a linear map, $f'(x) \in \mathrm{Hom}(\mathbb{R}^n, V)$, for every $x \in U$, and such that

$$f'(x)(e_j) = \sum_{i=1}^{m} \frac{\partial f_i}{\partial x_j}(x)\, u_i, \qquad 1 \leq j \leq n,$$

where $(e_1, \ldots, e_n)$ is the canonical basis of $\mathbb{R}^n$ and $(u_1, \ldots, u_m)$ is a basis of V. The $m \times n$ matrix,

$$\left(\frac{\partial f_i}{\partial x_j}\right),$$

is the *Jacobian matrix* of f. We also write $f'_x(u)$ for $f'(x)(u)$. Observe that since a p-form is a smooth map, $\omega\colon U \to \bigwedge^p(\mathbb{R}^n)^*$, its derivative is a map,

$$\omega'\colon U \to \mathrm{Hom}(\mathbb{R}^n, \bigwedge^p(\mathbb{R}^n)^*),$$

such that ω'_x is a linear map from $\mathbb{R}^n$ to $\bigwedge^p(\mathbb{R}^n)^*$, for every $x \in U$. By the isomorphism, $\bigwedge^p(\mathbb{R}^n)^* \cong \mathrm{Alt}^p(\mathbb{R}^n; \mathbb{R})$, we can view ω'_x as a linear map, $\omega_x\colon \mathbb{R}^n \to \mathrm{Alt}^p(\mathbb{R}^n; \mathbb{R})$, or equivalently, as a multilinear form, $\omega'_x\colon (\mathbb{R}^n)^{p+1} \to \mathbb{R}$, which is alternating in its last p arguments. The exterior derivative, $(d\omega)_x$, is obtained by making ω'_x into an alternating map in all of its $p + 1$ arguments.

Definition 8.3. For every $p \geq 0$, the *exterior differential*, $d\colon \mathcal{A}^p(U) \to \mathcal{A}^{p+1}(U)$, is given by

$$(d\omega)_x(u_1, \ldots, u_{p+1}) = \sum_{i=1}^{p+1} (-1)^{i-1} \omega'_x(u_i)(u_1, \ldots, \widehat{u_i}, \ldots, u_{p+1}),$$

for all $\omega \in \mathcal{A}^p(U)$ and all $u_1, \ldots, u_{p+1} \in \mathbb{R}^n$, where the hat over the argument u_i means that it should be omitted.

One should check that $(d\omega)_x$ is indeed alternating but this is easy. If necessary to avoid confusion, we write $d^p\colon \mathcal{A}^p(U) \to \mathcal{A}^{p+1}(U)$ instead of $d\colon \mathcal{A}^p(U) \to \mathcal{A}^{p+1}(U)$.

Remark: Definition 8.3 is the definition adopted by Cartan [29, 30][1] and Madsen and Tornehave [101]. Some authors use a different approach often using Propositions 8.2 and 8.3 as a starting point but we find the approach using Definition 8.3 more direct. Furthermore, this approach extends immediately to the case of vector valued forms.

For any smooth function, $f \in \mathcal{A}^0(U) = C^\infty(U, \mathbb{R})$, we get

$$df_x(u) = f'_x(u).$$

Therefore, for smooth functions, *the exterior differential, df, coincides with the usual derivative, f'* (we identify $\bigwedge^1(\mathbb{R}^n)^*$ and $(\mathbb{R}^n)^*$). For any 1-form, $\omega \in \mathcal{A}^1(U)$, we have

$$d\omega_x(u, v) = \omega'_x(u)(v) - \omega'_x(v)(u).$$

It follows that the map

$$(u, v) \mapsto \omega'_x(u)(v)$$

is symmetric iff $d\omega = 0$.

For a concrete example of exterior differentiation, if

$$\omega(x, y) = \frac{-y}{x^2 + y^2}\, e_1^* + \frac{x}{x^2 + y^2}\, e_2^*,$$

check that $d\omega = 0$.

The following observation is quite trivial but it will simplify notation: On $\mathbb{R}^n$, we have the projection function, $pr_i\colon \mathbb{R}^n \to \mathbb{R}$, with $pr_i(u_1, \ldots, u_n) = u_i$. Note that $pr_i = e_i^*$, where $(e_1, \ldots, e_n)$ is the canonical basis of $\mathbb{R}^n$. Let $x_i\colon U \to \mathbb{R}$ be the restriction of pr_i to U. Then, note that x'_i *is the constant map* given by

$$x'_i(x) = pr_i, \qquad x \in U.$$

[1] We warn the reader that a few typos have crept up in the English translation, Cartan [30], of the orginal version Cartan [29].

It follows that $dx_i = x_i'$ is the constant function with value $pr_i = e_i^*$. Now, since every p-form, ω, can be uniquely expressed as

$$\omega_x = \sum_I f_I(x)\, e_{i_1}^* \wedge \cdots \wedge e_{i_p}^* = \sum_I f_I(x) e_I^*, \qquad x \in U,$$

using Definition 8.2, we see immediately that ω can be uniquely written in the form

$$\omega = \sum_I f_I(x)\, dx_{i_1} \wedge \cdots \wedge dx_{i_p}, \tag{$*$}$$

where the f_I are smooth functions on U.

Observe that for $f \in \mathcal{A}^0(U) = C^\infty(U, \mathbb{R})$, we have

$$df_x = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(x)\, e_i^* \quad \text{and} \quad df = \sum_{i=1}^n \frac{\partial f}{\partial x_i}\, dx_i.$$

Proposition 8.2. *For every p form, $\omega \in \mathcal{A}^p(U)$, with $\omega = f dx_{i_1} \wedge \cdots \wedge dx_{i_p}$, we have*

$$d\omega = df \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_p}.$$

Proof. Recall that $\omega_x = f e_{i_1}^* \wedge \cdots \wedge e_{i_p}^* = f e_I^*$, so

$$\omega_x'(u) = f_x'(u) e_I^* = df_x(u) e_I^*$$

and by Definition 8.3, we get

$$d\omega_x(u_1, \ldots, u_{p+1}) = \sum_{i=1}^{p+1} (-1)^{i-1} df_x(u_i) e_I^*(u_1, \ldots, \widehat{u_i}, \ldots, u_{p+1}) = (df_x \wedge e_I^*)(u_1, \ldots, u_{p+1}),$$

where the last equation is an instance of the equation stated just before Proposition 22.24. □

We can now prove

Proposition 8.3. *For all $\omega \in \mathcal{A}^p(U)$ and all $\eta \in \mathcal{A}^q(U)$,*

$$d(\omega \wedge \eta) = d\omega \wedge \eta + (-1)^p \omega \wedge d\eta.$$

Proof. In view of the unique representation, $(*)$, it is enough to prove the proposition when $\omega = f e_I^*$ and $\eta = g e_J^*$. In this case, as $\omega \wedge \eta = fg\, e_I^* \wedge e_J^*$, by Proposition 8.2, we have

$$\begin{aligned} d(\omega \wedge \eta) &= d(fg) \wedge e_I^* \wedge e_J^* \\ &= ((df)g + f(dg)) \wedge e_I^* \wedge e_J^* \\ &= (df) g e_I^* \wedge e_J^* + f(dg) \wedge e_I^* \wedge e_J^* \\ &= (df) e_I^* \wedge g e_J^* + (-1)^p f \wedge e_I^* \wedge (dg) \wedge e_J^* \\ &= d\omega \wedge \eta + (-1)^p \omega \wedge d\eta, \end{aligned}$$

as claimed. □

We say that d is an *anti-derivation of degree* -1. Finally, here is the crucial and almost magical property of d:

Proposition 8.4. *For every $p \geq 0$, the composition $\mathcal{A}^p(U) \xrightarrow{d} \mathcal{A}^{p+1}(U) \xrightarrow{d} \mathcal{A}^{p+2}(U)$ is identically zero, that is,*

$$d \circ d = 0,$$

or, using superscripts, $d^{p+1} \circ d^p = 0$.

Proof. It is enough to prove the proposition when $\omega = f e_I^*$. We have

$$d\omega_x = df_x \wedge e_I^* = \frac{\partial f}{\partial x_1}(x)\, e_1^* \wedge e_I^* + \cdots + \frac{\partial f}{\partial x_n}(x)\, e_n^* \wedge e_I^*.$$

As $e_i^* \wedge e_j^* = -e_j^* \wedge e_i^*$ and $e_i^* \wedge e_i^* = 0$, we get

$$\begin{aligned}(d \circ d)\omega &= \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(x)\, e_i^* \wedge e_j^* \wedge e_I^* \\ &= \sum_{i<j} \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) - \frac{\partial^2 f}{\partial x_j \partial x_i}(x) \right) e_i^* \wedge e_j^* \wedge e_I^* = 0,\end{aligned}$$

since partial derivatives commute (as f is smooth). □

Propositions 8.2, 8.3 and 8.4 can be summarized by saying that $\mathcal{A}^*(U)$ together with the product, $\wedge$, and the differential, d, is a *differential graded algebra*. As $\mathcal{A}^*(U)) = \bigoplus_{p \geq 0} \mathcal{A}^p(U)$ and $d^p \colon \mathcal{A}^p(U) \to \mathcal{A}^{p+1}(U)$, we can view $d = (d^p)$ as a linear map, $d \colon \mathcal{A}^*(U) \to \mathcal{A}^*(U)$, such that

$$d \circ d = 0.$$

The diagram

$$\mathcal{A}^0(U) \xrightarrow{d} \mathcal{A}^1(U) \longrightarrow \cdots \longrightarrow \mathcal{A}^{p-1}(U) \xrightarrow{d} \mathcal{A}^p(U) \xrightarrow{d} \mathcal{A}^{p+1}(U) \longrightarrow \cdots$$

is called the *de Rham complex* of U. It is a *cochain complex*.

Let us consider one more example. Assume $n = 3$ and consider any function, $f \in \mathcal{A}^0(U)$. We have

$$df = \frac{\partial f}{\partial x}\, dx + \frac{\partial f}{\partial y}\, dy + \frac{\partial f}{\partial z}\, dz$$

and the vector

$$\left(\frac{\partial f}{\partial x}, \quad \frac{\partial f}{\partial y}, \quad \frac{\partial f}{\partial z} \right)$$

is the *gradient* of f. Next, let

$$\omega = P dx + Q dy + R dz$$

be a 1-form on some open, $U \subseteq \mathbb{R}^3$. An easy calculation yields

$$d\omega = \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right) dy \wedge dz + \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right) dz \wedge dx + \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dx \wedge dy.$$

The vector field given by

$$\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}, \quad \frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}, \quad \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}$$

is the *curl* of the vector field given by (P, Q, R). Now, if

$$\eta = A dy \wedge dz + B dz \wedge dx + C dx \wedge dy$$

is a 2-form on $\mathbb{R}^3$, we get

$$d\eta = \left(\frac{\partial A}{\partial x} + \frac{\partial B}{\partial y} + \frac{\partial C}{\partial z}\right) dx \wedge dy \wedge dz.$$

The real number,

$$\frac{\partial A}{\partial x} + \frac{\partial B}{\partial y} + \frac{\partial C}{\partial z}$$

is called the *divergence* of the vector field (A, B, C). When is there a smooth field, (P, Q, R), whose curl is given by a prescribed smooth field, (A, B, C)? Equivalently, when is there a 1-form, $\omega = P dx + Q dy + R dz$, such that

$$d\omega = \eta = A dy \wedge dz + B dz \wedge dx + C dx \wedge dy?$$

By Proposition 8.4, it is necessary that $d\eta = 0$, that is, that (A, B, C) has zero divergence. However, this condition is not sufficient in general; it depends on the topology of U. If U is *star-like*, Poincaré's Lemma (to be considered shortly) says that this condition is sufficient.

Definition 8.4. A differential form, ω, is *closed* iff $d\omega = 0$, *exact* iff $\omega = d\eta$, for some differential form, η. For every $p \geq 0$, let

$$Z^p(U) = \{\omega \in \mathcal{A}^p(U) \mid d\omega = 0\} = \operatorname{Ker} d \colon \mathcal{A}^p(U) \longrightarrow \mathcal{A}^{p+1}(U),$$

be the vector space of closed p-forms, also called *p-cocycles* and for every $p \geq 1$, let

$$B^p(U) = \{\omega \in \mathcal{A}^p(U) \mid \exists \eta \in \mathcal{A}^{p-1}(U),\ \omega = d\eta\} = \operatorname{Im} d \colon \mathcal{A}^{p-1}(U) \longrightarrow \mathcal{A}^p(U),$$

be the vector space of exact p-forms, also called *p-coboundaries*. Set $B^0(U) = (0)$. Forms in $\mathcal{A}^p(U)$ are also called *p-cochains*. As $B^p(U) \subseteq Z^p(U)$ (by Proposition 8.4), for every $p \geq 0$, we define the p^{th} *de Rham cohomology group* of U as the quotient space

$$H^p_{\text{DR}}(U) = Z^p(U)/B^p(U).$$

An element of $H^p_{\text{DR}}(U)$ is called a *cohomology class* and is denoted $[\omega]$, where $\omega \in Z^p(U)$ is a cocycle. The real vector space, $H^\bullet_{\text{DR}}(U) = \bigoplus_{p \geq 0} H^p_{\text{DR}}(U)$, is called the *de Rham cohomology algebra* of U.

We often drop the subscript DR and write $H^p(U)$ for $H^p_{\mathrm{DR}}(U)$ (resp. $H^\bullet(U)$ for $H^\bullet_{\mathrm{DR}}(U)$) when no confusion arises. Proposition 8.4 shows that every exact form is closed but the converse is false in general. Measuring the extent to which closed forms are not exact is the object of *de Rham cohomology*. For example, if we consider the form

$$\omega(x,y) = \frac{-y}{x^2+y^2}\,dx + \frac{x}{x^2+y^2}\,dy,$$

on $U = \mathbb{R}^2 - \{0\}$, we have $d\omega = 0$. Yet, it is not hard to show (using integration, see Madsen and Tornehave [101], Chapter 1) that there is no smooth function, f, on U such that $df = \omega$. Thus, ω is a closed form which is not exact. This is because U is punctured.

Observe that $H^0(U) = Z^0(U) = \{f \in C^\infty(U,\mathbb{R}) \mid df = 0\}$, that is, $H^0(U)$ is the space of locally constant functions on U, equivalently, the space of functions that are constant on the connected components of U. Thus, the cardinality of $H^0(U)$ gives the number of connected components of U. For a large class of open sets (for example, open sets that can be covered by finitely many convex sets), the cohomology groups, $H^p(U)$, are finite dimensional.

Going back to Definition 8.4, we define the vector spaces $Z^*(U)$ and $B^*(U)$ by

$$Z^*(U) = \bigoplus_{p\geq 0} Z^p(U) \quad \text{and} \quad B^*(U) = \bigoplus_{p\geq 0} B^p(U).$$

Now, $\mathcal{A}^*(U)$ is a graded algebra with multiplication, $\wedge$. Observe that $Z^*(U)$ is a subalgebra of $\mathcal{A}^*(U)$, since

$$d(\omega\wedge\eta) = d\omega\wedge\eta + (-1)^p\omega\wedge d\eta,$$

so $d\omega = 0$ and $d\eta = 0$ implies $d(\omega\wedge\eta) = 0$. Furthermore, $B^*(U)$ is an ideal in $Z^*(U)$, because if $\omega = d\eta$ and $d\tau = 0$, then

$$d(\eta\tau) = d\eta\wedge\tau + (-1)^{p-1}\eta\wedge d\tau = \omega\wedge\tau,$$

with $\eta \in \mathcal{A}^{p-1}(U)$. Therefore, $H^\bullet_{\mathrm{DR}} = Z^*(U)/B^*(U)$ inherits a graded algebra structure from $\mathcal{A}^*(U)$. Explicitly, the multiplication in $H^\bullet_{\mathrm{DR}}$ is given by

$$[\omega]\,[\eta] = [\omega\wedge\eta].$$

It turns out that Propositions 8.3 and 8.4 together with the fact that d coincides with the derivative on $\mathcal{A}^0(U)$ characterize the differential, d.

Theorem 8.5. *There is a unique linear map, $d\colon \mathcal{A}^*(U) \to \mathcal{A}^*(U)$, with $d = (d^p)$ and $d^p\colon \mathcal{A}^p(U) \to \mathcal{A}^{p+1}(U)$ for every $p \geq 0$, such that*

(1) $df = f'$, for every $f \in \mathcal{A}^0(U) = C^\infty(U,\mathbb{R})$.

(2) $d \circ d = 0$.

(3) For every $\omega \in \mathcal{A}^p(U)$ and every $\eta \in \mathcal{A}^q(U)$,

$$d(\omega \wedge \eta) = d\omega \wedge \eta + (-1)^p \omega \wedge d\eta.$$

Proof. Existence has already been shown so we only have to prove uniqueness. Let δ be another linear map satisfying (1)–(3). By (1), $df = \delta f = f'$, if $f \in \mathcal{A}^0(U)$. In particular, this hold when $f = x_i$, with $x_i \colon U \to \mathbb{R}$ the restriction of pr_i to U. In this case, we know that $\delta x_i = e_i^*$, the constant function, $e_i^* = pr_i$. By (2), $\delta e_i^* = 0$. Using (3), we get $\delta e_I^* = 0$, for every nonempty subset $I \subseteq \{1, \ldots, n\}$. If $\omega = f e_I^*$, by (3), we get

$$\delta\omega = \delta f \wedge e_I^* + f \wedge \delta e_I^* = \delta f \wedge e_I^* = df \wedge e_I^* = d\omega.$$

Finally, since every differential form is a linear combination of special forms, $f_I e_I^*$, we conclude that $\delta = d$. □

We now consider the action of smooth maps, $\varphi \colon U \to U'$, on differential forms in $\mathcal{A}^*(U')$. We will see that φ induces a map from $\mathcal{A}^*(U')$ to $\mathcal{A}^*(U)$ called a *pull-back map*. This correspond to a change of variables.

Recall Proposition 22.21 which states that if $f \colon E \to F$ is any linear map between two finite-dimensional vector spaces, E and F, then

$$\mu\Big(\Big(\bigwedge^p f^\top\Big)(\omega)\Big)(u_1, \ldots, u_p) = \mu(\omega)(f(u_1), \ldots, f(u_p)), \qquad \omega \in \bigwedge^p F^*,\ \ u_1, \ldots, u_p \in E.$$

We apply this proposition with $E = \mathbb{R}^n$, $F = \mathbb{R}^m$, and $f = \varphi'_x$ $(x \in U)$, and get

$$\mu\Big(\Big(\bigwedge^p (\varphi'_x)^\top\Big)(\omega_{\varphi(x)})\Big)(u_1, \ldots, u_p) = \mu(\omega_{\varphi(x)})(\varphi'_x(u_1), \ldots, \varphi'_x(u_p)), \qquad \omega \in \mathcal{A}^p(V),\ \ u_i \in \mathbb{R}^n.$$

This gives us the behavior of $\bigwedge^p(\varphi'_x)^\top$ under the identification of $\bigwedge^p(\mathbb{R})^*$ and $\mathrm{Alt}^n(\mathbb{R}^n; \mathbb{R})$ *via* the isomorphism μ. Consequently, denoting $\bigwedge^p(\varphi'_x)^\top$ by φ^*, we make the following definition:

Definition 8.5. Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^m$ be two open subsets. For every smooth map, $\varphi \colon U \to V$, for every $p \geq 0$, we define the map, $\varphi^* \colon \mathcal{A}^p(V) \to \mathcal{A}^p(U)$, by

$$\varphi^*(\omega)_x(u_1, \ldots, u_p) = \omega_{\varphi(x)}(\varphi'_x(u_1), \ldots, \varphi'_x(u_p)),$$

for all $\omega \in \mathcal{A}^p(V)$, all $x \in U$ and all $u_1, \ldots, u_p \in \mathbb{R}^n$. We say that $\varphi^*(\omega)$ (for short, $\varphi^*\omega$) is the *pull-back of ω by φ*.

As φ is smooth, $\varphi^*\omega$ is a smooth p-form on U. The maps $\varphi^* \colon \mathcal{A}^p(V) \to \mathcal{A}^p(U)$ induce a map also denoted $\varphi^* \colon \mathcal{A}^*(V) \to \mathcal{A}^*(U)$. Using the chain rule, we check immediately that

$$\begin{aligned} \mathrm{id}^* &= \mathrm{id}, \\ (\psi \circ \varphi)^* &= \varphi^* \circ \psi^*. \end{aligned}$$

As an example, consider the constant form, $\omega = e_i^*$. We claim that $\varphi^* e_i^* = d\varphi_i$, where $\varphi_i = pr_i \circ \varphi$. Indeed,

$$\begin{aligned}(\varphi^* e_i^*)_x(u) &= e_i^*(\varphi'_x(u)) \\ &= e_i^*\left(\sum_{k=1}^m \left(\sum_{l=1}^n \frac{\partial \varphi_k}{\partial x_l}(x)\, u_l\right) e_k\right) \\ &= \sum_{l=1}^n \frac{\partial \varphi_i}{\partial x_l}(x)\, u_l \\ &= \sum_{l=1}^n \frac{\partial \varphi_i}{\partial x_l}(x)\, e_l^*(u) = d(\varphi_i)_x(u).\end{aligned}$$

For another example, assume U and V are open subsets of $\mathbb{R}^n$, $\omega = f dx_1 \wedge \cdots \wedge dx_n$, and write $x = \varphi(y)$, with x coordinates on V and y coordinates on U. Then

$$(\varphi^*\omega)_y = f(\varphi(y)) \det\left(\frac{\partial \varphi_i}{\partial y_j}(y)\right) dy_1 \wedge \cdots \wedge dy_p = f(\varphi(y)) J(\varphi)_y\, dy_1 \wedge \cdots \wedge dy_p,$$

where

$$J(\varphi)_y = \det\left(\frac{\partial \varphi_i}{\partial y_j}(y)\right)$$

is the Jacobian of φ at $y \in U$.

Proposition 8.6. *Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^m$ be two open sets and let $\varphi\colon U \to V$ be a smooth map. Then*

(i) $\varphi^*(\omega \wedge \eta) = \varphi^*\omega \wedge \varphi^*\eta$, *for all* $\omega \in \mathcal{A}^p(V)$ *and all* $\eta \in \mathcal{A}^q(V)$.

(ii) $\varphi^*(f) = f \circ \varphi$, *for all* $f \in \mathcal{A}^0(V)$.

(iii) $d\varphi^*(\omega) = \varphi^*(d\omega)$, *for all* $\omega \in \mathcal{A}^p(V)$, *that is, the following diagram commutes for all* $p \geq 0$*:*

$$\begin{array}{ccc} \mathcal{A}^p(V) & \xrightarrow{\varphi^*} & \mathcal{A}^p(U) \\ {\scriptstyle d}\downarrow & & \downarrow{\scriptstyle d} \\ \mathcal{A}^{p+1}(V) & \xrightarrow{\varphi^*} & \mathcal{A}^{p+1}(U). \end{array}$$

Proof. We leave the proof of (i) and (ii) as an exercise (or see Madsen and Tornehave [101], Chapter 3). First, we prove (iii) in the case $\omega \in \mathcal{A}^0(V)$. Using (i) and (ii) and the calculation

just before Proposition 8.6, we have

$$\begin{aligned}
\varphi^*(df) &= \sum_{k=1}^{m} \varphi^*\left(\frac{\partial f}{\partial x_k}\right) \wedge \varphi^*(e_k^*) \\
&= \sum_{k=1}^{m} \left(\frac{\partial f}{\partial x_k} \circ \varphi\right) \wedge \left(\sum_{l=1}^{n} \frac{\partial \varphi_k}{\partial x_l}\, e_l^*\right) \\
&= \sum_{k=1}^{m}\sum_{l=1}^{n} \left(\frac{\partial f}{\partial x_k} \circ \varphi\right)\left(\frac{\partial \varphi_k}{\partial x_l}\right)\, e_l^* \\
&= \sum_{l=1}^{n} \left(\sum_{k=1}^{m} \left(\frac{\partial f}{\partial x_k} \circ \varphi\right)\frac{\partial \varphi_k}{\partial x_l}\right)\, e_l^* \\
&= \sum_{l=1}^{n} \frac{\partial (f \circ \varphi)}{\partial x_l}\, e_l^* \\
&= d(f \circ \varphi) = d(\varphi^*(f)).
\end{aligned}$$

For the case where $\omega = f e_I^*$, we know that $d\omega = df \wedge e_I^*$. We claim that

$$d\varphi^*(e_I^*) = 0.$$

This is because

$$\begin{aligned}
d\varphi^*(e_I^*) &= d(\varphi^*(e_{i_1}^*) \wedge \cdots \wedge \varphi^*(e_{i_p}^*)) \\
&= \sum (-1)^{k-1} \varphi^*(e_{i_1}^*) \wedge \cdots \wedge d(\varphi^*(e_{i_k}^*)) \wedge \cdots \wedge \varphi^*(e_{i_p}^*) = 0,
\end{aligned}$$

since $\varphi^*(e_{i_k}^*) = d\varphi_{i_k}$ and $d \circ d = 0$. Consequently,

$$d(\varphi^*(f) \wedge \varphi^*(e_I^*)) = d(\varphi^* f) \wedge \varphi^*(e_I^*).$$

Then, we have

$$\varphi^*(d\omega) = \varphi^*(df) \wedge \varphi^*(e_I^*) = d(\varphi^* f) \wedge \varphi^*(e_I^*) = d(\varphi^*(f) \wedge \varphi^*(e_I^*)) = d(\varphi^*(f e_I^*)) = d(\varphi^*\omega).$$

Since every differential form is a linear combination of special forms, $f e_I^*$, we are done. □

The fact that d and pull-back commutes is an important fact: It allows us to show that a map, $\varphi\colon U \to V$, induces a map, $H^\bullet(\varphi)\colon H^\bullet(V) \to H^\bullet(U)$, on cohomology and it is crucial in generalizing the exterior differential to manifolds.

To a smooth map, $\varphi\colon U \to V$, we associate the map, $H^p(\varphi)\colon H^p(V) \to H^p(U)$, given by

$$H^p(\varphi)([\omega]) = [\varphi^*(\omega)].$$

This map is well defined because if we pick any representative, $\omega + d\eta$ in the cohomology class, $[\omega]$, specified by the closed form, ω, then

$$d\varphi^*\omega = \varphi^* d\omega = 0$$

so $\varphi^*\omega$ is closed and

$$\varphi^*(\omega + d\eta) = \varphi^*\omega + \varphi^*(d\eta) = \varphi^*\omega + d\varphi^*\eta,$$

so $H^p(\varphi)([\omega])$ is well defined. It is also clear that

$$H^{p+q}(\varphi)([\omega][\eta]) = H^p(\varphi)([\omega])H^q(\varphi)([\eta]),$$

which means that $H^\bullet(\varphi)$ is a homomorphism of graded algebras. We often denote $H^\bullet(\varphi)$ again by φ^*.

We conclude this section by stating without proof an important result known as the *Poincaré Lemma*. Recall that a subset, $S \subseteq \mathbb{R}^n$ is *star-shaped* iff there is some point, $c \in S$, such that for every point, $x \in S$, the closed line segment, $[c, x]$, joining c and x is entirely contained in S.

Theorem 8.7. *(Poincaré's Lemma)* If $U \subseteq \mathbb{R}^n$ is any star-shaped open set, then we have $H^p(U) = (0)$ for $p > 0$ and $H^0(U) = \mathbb{R}$. Thus, for every $p \geq 1$, every closed form $\omega \in \mathcal{A}^p(U)$ is exact.

Proof. Pick c so that U is star-shaped w.r.t. c and let $g\colon U \to U$ be the constant function with value c. Then, we see that

$$g^*\omega = \begin{cases} 0 & \text{if } \omega \in \mathcal{A}^p(U), \text{ with } p \geq 1, \\ \omega(c) & \text{if } \omega \in \mathcal{A}^0(U), \end{cases}$$

where $\omega(c)$ denotes the constant function with value $\omega(c)$. The trick is to find a family of linear maps, $h^p\colon \mathcal{A}^p(U) \to \mathcal{A}^{p-1}(U)$, for $p \geq 1$, with $h^0 = 0$, such that

$$d \circ h^p + h^{p+1} \circ d = \mathrm{id} - g^*, \qquad p > 0$$

called a *chain homotopy*. Indeed, if $\omega \in \mathcal{A}^p(U)$ is closed and $p \geq 1$, we get $dh^p\omega = \omega$, so ω is exact and if $p = 0$, we get $h^1 d\omega = 0 = \omega - \omega(c)$, so ω is constant. It remains to find the h^p, which is not obvious. A construction of these maps can be found in Madsen and Tornehave [101] (Chapter 3), Warner [148] (Chapter 4), Cartan [30] (Section 2) Morita [115] (Chapter 3). □

In Section 8.2, we promote differential forms to manifolds. As preparation, note that every open subset, $U \subseteq \mathbb{R}^n$, is a manifold and that for every $x \in U$ the tangent space, T_xU, to U at x is canonically isomorphic to $\mathbb{R}^n$. It follows that the tangent bundle, TU, and the

cotangent bundle, T^*U, are trivial, namely, $TU \cong U \times \mathbb{R}^n$ and $T^*U \cong U \times (\mathbb{R}^n)^*$, so the bundle,

$$\bigwedge^k T^*U \cong U \times \bigwedge^k (\mathbb{R}^n)^*,$$

is also trivial. Consequently, we can view $\mathcal{A}^k(U)$ as the set of smooth sections of the vector bundle, $\bigwedge^k T^*(U)$. The generalization to manifolds is then to define the space of differential p-forms on a manifold, M, as the space of smooth sections of the bundle, $\bigwedge^k T^*M$.

8.2 Differential Forms on Manifolds

Let M be any smooth manifold of dimension n. We define the vector bundle, $\bigwedge T^*M$, as the direct sum bundle,

$$\bigwedge T^*M = \bigoplus_{k=0}^{n} \bigwedge^k T^*M,$$

see Section 7.3 for details.

Definition 8.6. Let M be any smooth manifold of dimension n. The set, $\mathcal{A}^k(M)$, of *smooth differential k-forms* on M is the set of smooth sections, $\Gamma(M, \bigwedge^k T^*M)$, of the bundle $\bigwedge^k T^*M$ and the set, $\mathcal{A}^*(M)$, of all *smooth differential forms* on M is the set of smooth sections, $\Gamma(M, \bigwedge T^*M)$, of the bundle $\bigwedge T^*M$.

Recall that a smooth section of the bundle $\bigwedge^k T^*M$ is a smooth funtion $\omega\colon M \to \bigwedge^k T^*M$ such that $\omega(p) \in \bigwedge^k T_p^*M$ for all $p \in M$.

Observe that $\mathcal{A}^0(M) \cong C^\infty(M, \mathbb{R})$, the set of smooth functions on M, since the bundle $\bigwedge^0 T^*M$ is isomorphic to $M \times \mathbb{R}$ and smooth sections of $M \times \mathbb{R}$ are just graphs of smooth functions on M. We also write $C^\infty(M)$ for $C^\infty(M, \mathbb{R})$. If $\omega \in \mathcal{A}^*(M)$, we often write ω_p for $\omega(p)$.

Definition 8.6 is quite abstract and it is important to get a more down-to-earth feeling by taking a local view of differential forms, namely, with respect to a chart. So, let (U, φ) be a local chart on M, with $\varphi\colon U \to \mathbb{R}^n$, and let $x_i = pr_i \circ \varphi$, the ith local coordinate $(1 \leq i \leq n)$ (see Section 3.2). Recall that by Proposition 3.9, for any $p \in U$, the vectors

$$\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_x}\right)_p$$

form a basis of the tangent space, T_pM. Furthermore, by Proposition 3.15 and the discussion following Proposition 3.14, the linear forms, $(dx_1)_p, \ldots, (dx_n)_p$ form a basis of T_p^*M, (where $(dx_i)_p$, the differential of x_i at p, is identified with the linear form such that $df_p(v) = v(\mathbf{f})$, for every smooth function f on U and every $v \in T_pM$). Consequently, locally on U, every k-form, $\omega \in \mathcal{A}^k(M)$, can be written uniquely as

$$\omega = \sum_I f_I dx_{i_1} \wedge \cdots \wedge dx_{i_k} = \sum_I f_I dx_I, \qquad p \in U,$$

where $I = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$, with $i_1 < \ldots < i_k$ and $dx_I = dx_{i_1} \wedge \cdots \wedge dx_{i_k}$. Furthermore, each f_I is a smooth function on U.

Remark: We define the set of smooth (r, s)-*tensor fields* as the set, $\Gamma(M, T^{r,s}(M))$, of smooth sections of the tensor bundle $T^{r,s}(M) = T^{\otimes r}M \otimes (T^*M)^{\otimes s}$. Then, locally in a chart (U, φ), every tensor field $\omega \in \Gamma(M, T^{r,s}(M))$ can be written uniquely as

$$\omega = \sum f^{i_1,\ldots,i_r}_{j_1,\ldots,j_s} \left(\frac{\partial}{\partial x_{i_1}}\right) \otimes \cdots \otimes \left(\frac{\partial}{\partial x_{i_r}}\right) \otimes dx_{j_1} \otimes \cdots \otimes dx_{j_s}.$$

The operations on the algebra, $\bigwedge T^*M$, yield operations on differential forms using pointwise definitions. If $\omega, \eta \in \mathcal{A}^*(M)$ and $\lambda \in \mathbb{R}$, then for every $x \in M$,

$$\begin{aligned} (\omega + \eta)_x &= \omega_x + \eta_x \\ (\lambda\omega)_x &= \lambda\omega_x \\ (\omega \wedge \eta)_x &= \omega_x \wedge \eta_x. \end{aligned}$$

Actually, it is necessary to check that the resulting forms are smooth but this is easily done using charts. When, $f \in \mathcal{A}^0(M)$, we write $f\omega$ instead of $f \wedge \omega$. It follows that $\mathcal{A}^*(M)$ is a graded real algebra and a $C^\infty(M)$-module.

Proposition 8.1 generalizes immediately to manifolds.

Proposition 8.8. *For all forms $\omega \in \mathcal{A}^r(M)$ and $\eta \in \mathcal{A}^s(M)$, we have*

$$\eta \wedge \omega = (-1)^{pq}\omega \wedge \eta.$$

For any smooth map, $\varphi\colon M \to N$, between two manifolds, M and N, we have the differential map, $d\varphi\colon TM \to TN$, also a smooth map and, for every $p \in M$, the map $d\varphi_p\colon T_pM \to T_{\varphi(p)}N$ is linear. As in Section 8.1, Proposition 22.21 gives us the formula

$$\mu\Big(\Big(\bigwedge^k (d\varphi_p)^\top\Big)(\omega_{\varphi(p)})\Big)(u_1, \ldots, u_k) = \mu(\omega_{\varphi(p)})(d\varphi_p(u_1), \ldots, d\varphi_p(u_k)), \qquad \omega \in \mathcal{A}^k(N),$$

for all $u_1, \ldots, u_k \in T_pM$. This gives us the behavior of $\bigwedge^k (d\varphi_p)^\top$ under the identification of $\bigwedge^k T_p^*M$ and $\mathrm{Alt}^k(T_pM; \mathbb{R})$ *via* the isomorphism μ. Here is the extension of Definition 8.5 to differential forms on a manifold.

Definition 8.7. For any smooth map, $\varphi\colon M \to N$, between two smooth manifolds, M and N, for every $k \geq 0$, we define the map, $\varphi^*\colon \mathcal{A}^k(N) \to \mathcal{A}^k(M)$, by

$$\varphi^*(\omega)_p(u_1, \ldots, u_k) = \omega_{\varphi(p)}(d\varphi_p(u_1), \ldots, d\varphi_p(u_k)),$$

for all $\omega \in \mathcal{A}^k(N)$, all $p \in M$, and all $u_1, \ldots, u_k \in T_pM$. We say that $\varphi^*(\omega)$ (for short, $\varphi^*\omega$) is the *pull-back of ω by φ*.

The maps $\varphi^*\colon \mathcal{A}^k(N) \to \mathcal{A}^k(M)$ induce a map also denoted $\varphi^*\colon \mathcal{A}^*(N) \to \mathcal{A}^*(M)$. Using the chain rule, we check immediately that

$$\begin{aligned} \mathrm{id}^* &= \mathrm{id}, \\ (\psi \circ \varphi)^* &= \varphi^* \circ \psi^*. \end{aligned}$$

We need to check that $\varphi^*\omega$ is smooth and for this, it is enough to check it locally on a chart, (U, φ). On U, we know that $\omega \in \mathcal{A}^k(M)$ can be written uniquely as

$$\omega = \sum_I f_I dx_{i_1} \wedge \cdots \wedge dx_{i_k}, \qquad p \in U,$$

with f_I smooth and it is easy to see (using the definition) that

$$\varphi^*\omega = \sum_I (f_I \circ \varphi) d(x_{i_1} \circ \varphi) \wedge \cdots \wedge d(x_{i_k} \circ \varphi),$$

which is smooth.

Remark: The fact that the pull-back of differential forms makes sense for arbitrary smooth maps, $\varphi\colon M \to N$, and not just diffeomorphisms is a major technical superiority of forms over vector fields.

The next step is to define d on $\mathcal{A}^*(M)$. There are several ways to proceed but since we already considered the special case where M is an open subset of $\mathbb{R}^n$, we proceed using charts.

Given a smooth manifold, M, of dimension n, let (U, φ) be any chart on M. For any $\omega \in \mathcal{A}^k(M)$ and any $p \in U$, define $(d\omega)_p$ as follows: If $k = 0$, that is, $\omega \in C^\infty(M)$, let

$$(d\omega)_p = d\omega_p, \qquad \text{the differential of } \omega \text{ at } p$$

and if $k \geq 1$, let

$$(d\omega)_p = \varphi^*\big(d((\varphi^{-1})^*\omega)_{\varphi(p)}\big)_p,$$

where d is the exterior differential on $\mathcal{A}^k(\varphi(U))$. More explicitly, $(d\omega)_p$ is given by

$$(d\omega)_p(u_1, \ldots, u_{k+1}) = d((\varphi^{-1})^*\omega)_{\varphi(p)}(d\varphi_p(u_1), \ldots, d\varphi_p(u_{k+1})),$$

for every $p \in U$ and all $u_1, \ldots, u_{k+1} \in T_pM$. Observe that the above formula is still valid when $k = 0$ if we interpret the symbold d in $d((\varphi^{-1})^*\omega)_{\varphi(p)} = d(\omega \circ \varphi^{-1})_{\varphi(p)}$ as the differential.

Since $\varphi^{-1}\colon \varphi(U) \to U$ is map whose domain is an open subset, $W = \varphi(U)$, of $\mathbb{R}^n$, the form $(\varphi^{-1})^*\omega$ is a differential form in $\mathcal{A}^*(W)$, so $d((\varphi^{-1})^*\omega)$ is well-defined. We need to check that this definition does not depend on the chart, (U, φ). For any other chart, (V, ψ),

with $U \cap V \neq \emptyset$, the map $\theta = \psi \circ \varphi^{-1}$ is a diffeomorphism between the two open subsets, $\varphi(U \cap V)$ and $\psi(U \cap V)$, and $\psi = \theta \circ \varphi$. Let $x = \varphi(p)$. We need to check that

$$d((\varphi^{-1})^*\omega)_x(d\varphi_p(u_1), \ldots, d\varphi_p(u_{k+1})) = d((\psi^{-1})^*\omega)_x(d\psi_p(u_1), \ldots, d\psi_p(u_{k+1})),$$

for every $p \in U \cap V$ and all $u_1, \ldots, u_{k+1} \in T_pM$. However,

$$d((\psi^{-1})^*\omega)_x(d\psi_p(u_1), \ldots, d\psi_p(u_{k+1})) = d((\varphi^{-1} \circ \theta^{-1})^*\omega)_x(d(\theta \circ \varphi)_p(u_1), \ldots, d(\theta \circ \varphi)_p(u_{k+1})),$$

and since

$$(\varphi^{-1} \circ \theta^{-1})^* = (\theta^{-1})^* \circ (\varphi^{-1})^*$$

and, by Proposition 8.6 (iii),

$$d(((\theta^{-1})^* \circ (\varphi^{-1})^*)\omega) = d((\theta^{-1})^*((\varphi^{-1})^*\omega)) = (\theta^{-1})^*(d((\varphi^{-1})^*\omega)),$$

we get

$$\begin{aligned} d((\varphi^{-1} \circ \theta^{-1})^*\omega)_x(d(\theta \circ \varphi)_p(u_1), \ldots, d(\theta \circ \varphi)_p(u_{k+1})) \\ = (\theta^{-1})^*(d((\varphi^{-1})^*\omega))_{\theta(x)}(d(\theta \circ \varphi)_p(u_1), \ldots, d(\theta \circ \varphi)_p(u_{k+1})) \end{aligned}$$

and then

$$\begin{aligned} (\theta^{-1})^*(d((\varphi^{-1})^*\omega))_{\theta(x)}(d(\theta \circ \varphi)_p(u_1), \ldots, d(\theta \circ \varphi)_p(u_{k+1})) \\ = d((\varphi^{-1})^*\omega)_x((d\theta^{-1})_{\theta(x)}(d(\theta \circ \varphi)_p(u_1)), \ldots, (d\theta^{-1})_{\theta(x)}(d(\theta \circ \varphi)_p(u_{k+1}))). \end{aligned}$$

As $(d\theta^{-1})_{\theta(x)}(d(\theta \circ \varphi)_p(u_i)) = d(\theta^{-1} \circ (\theta \circ \varphi))_p(u_i) = d\varphi_p(u_i)$, by the chain rule, we obtain

$$d((\psi^{-1})^*\omega)_x(d\psi_p(u_1), \ldots, d\psi_p(u_{k+1})) = d((\varphi^{-1})^*\omega)_x(d\varphi_p(u_1), \ldots, d\varphi_p(u_{k+1})),$$

as desired.

Observe that $(d\omega)_p$ is smooth on U and as our definition of $(d\omega)_p$ does not depend on the choice of a chart, the forms $(d\omega) \restriction U$ agree on overlaps and yield a differential form, $d\omega$, defined on the whole of M. Thus, we can make the following definition:

Definition 8.8. If M is any smooth manifold, there is a linear map, $d\colon \mathcal{A}^k(M) \to \mathcal{A}^{k+1}(M)$, for every $k \geq 0$, such that, for every $\omega \in \mathcal{A}^k(M)$, for every chart, (U, φ), for every $p \in U$, if $k = 0$, that is, $\omega \in C^\infty(M)$, then

$$(d\omega)_p = d\omega_p, \qquad \text{the differential of } \omega \text{ at } p,$$

else if $k \geq 1$, then

$$(d\omega)_p = \varphi^*\big(d((\varphi^{-1})^*\omega)_{\varphi(p)}\big)_p,$$

where d is the exterior differential on $\mathcal{A}^k(\varphi(U))$ from Definition 8.3. We obtain a linar map, $d\colon \mathcal{A}^*(M) \to \mathcal{A}^*(M)$, called *exterior differentiation*.

Propositions 8.3, 8.4 and 8.6 generalize to manifolds.

Proposition 8.9. *Let M and N be smooth manifolds and let $\varphi\colon M \to N$ be a smooth map.*

(1) For all $\omega \in \mathcal{A}^r(M)$ and all $\eta \in \mathcal{A}^s(M)$,

$$d(\omega \wedge \eta) = d\omega \wedge \eta + (-1)^r \omega \wedge d\eta.$$

(2) For every $k \geq 0$, the composition $\mathcal{A}^k(M) \xrightarrow{d} \mathcal{A}^{k+1}(M) \xrightarrow{d} \mathcal{A}^{k+2}(M)$ is identically zero, that is,

$$d \circ d = 0.$$

(3) $\varphi^*(\omega \wedge \eta) = \varphi^*\omega \wedge \varphi^*\eta$, for all $\omega \in \mathcal{A}^r(N)$ and all $\eta \in \mathcal{A}^s(N)$.

(4) $\varphi^*(f) = f \circ \varphi$, for all $f \in \mathcal{A}^0(N)$.

(5) $d\varphi^*(\omega) = \varphi^*(d\omega)$, for all $\omega \in \mathcal{A}^k(N)$, that is, the following diagram commutes for all $k \geq 0$:

$$\begin{array}{ccc} \mathcal{A}^k(N) & \xrightarrow{\varphi^*} & \mathcal{A}^k(M) \\ {\scriptstyle d}\downarrow & & \downarrow{\scriptstyle d} \\ \mathcal{A}^{k+1}(N) & \xrightarrow{\varphi^*} & \mathcal{A}^{k+1}(M). \end{array}$$

Proof. It is enough to prove these properties in a chart, (U, φ), which is easy. We only check (2). We have

$$\begin{aligned} (d(d\omega))_p &= d\big(\varphi^*\big(d((\varphi^{-1})^*\omega)\big)\big)_p \\ &= \varphi^*\Big[d\Big((\varphi^{-1})^*\Big(\varphi^*\big(d((\varphi^{-1})^*\omega)\big)\Big)\Big)_{\varphi(p)}\Big]_p \\ &= \varphi^*\Big[d\big(d((\varphi^{-1})^*\omega)\big)_{\varphi(p)}\Big]_p \\ &= 0, \end{aligned}$$

as $(\varphi^{-1})^* \circ \varphi^* = (\varphi \circ \varphi^{-1})^* = \mathrm{id}^* = \mathrm{id}$ and $d \circ d = 0$ on forms in $\mathcal{A}^k(\varphi(U))$, with $\varphi(U) \subseteq \mathbb{R}^n$. □

As a consequence, Definition 8.4 of the de Rham cohomology generalizes to manifolds. For every manifold, M, we have the de Rham complex,

$$\mathcal{A}^0(M) \xrightarrow{d} \mathcal{A}^1(M) \longrightarrow \cdots \longrightarrow \mathcal{A}^{k-1}(M) \xrightarrow{d} \mathcal{A}^k(M) \xrightarrow{d} \mathcal{A}^{k+1}(M) \longrightarrow \cdots$$

and we can define the *cohomology groups*, $H^k_{\mathrm{DR}}(M)$, and the graded *cohomology algebra*, $H^\bullet_{\mathrm{DR}}(M)$. For every $k \geq 0$, let

$$Z^k(M) = \{\omega \in \mathcal{A}^k(M) \mid d\omega = 0\} = \operatorname{Ker} d\colon \mathcal{A}^k(M) \longrightarrow \mathcal{A}^{k+1}(M),$$

be the vector space of closed k-forms and for every $k \geq 1$, let

$$B^k(M) = \{\omega \in \mathcal{A}^k(M) \mid \exists \eta \in \mathcal{A}^{k-1}(M),\ \omega = d\eta\} = \operatorname{Im} d\colon \mathcal{A}^{k-1}(M) \longrightarrow \mathcal{A}^k(M),$$

be the vector space of exact k-forms and set $B^0(M) = (0)$. Then, for every $k \geq 0$, we define the k^{th} *de Rham cohomology group* of M as the quotient space

$$H^k_{\text{DR}}(M) = Z^k(M)/B^k(M).$$

The real vector space, $H^\bullet_{\text{DR}}(M) = \bigoplus_{k\geq 0} H^k_{\text{DR}}(M)$, is called the *de Rham cohomology algebra* of M. We often drop the subscript, DR, when no confusion arises. Every smooth map, $\varphi\colon M \to N$, between two manifolds induces an algebra map, $\varphi^*\colon H^\bullet(N) \to H^\bullet(M)$.

Another important property of the exterior differential is that it is a *local operator*, which means that the value of $d\omega$ at p only depends of the values of ω near p. More precisely, we have

Proposition 8.10. *Let M be a smooth manifold. For every open subset, $U \subseteq M$, for any two differential forms, $\omega, \eta \in \mathcal{A}^*(M)$, if $\omega \restriction U = \eta \restriction U$, then $(d\omega) \restriction U = (d\eta) \restriction U$.*

Proof. By linearity, it is enough to show that if $\omega \restriction U = 0$, then $(d\omega) \restriction U = 0$. The crucial ingredient is the existence of "bump functions". By Proposition 3.30 applied to the constant function with value 1, for every $p \in U$, there some open subset, $V \subseteq U$, containing p and a smooth function, $f\colon M \to \mathbb{R}$, such that $\operatorname{supp} f \subseteq U$ and $f \equiv 1$ on V. Consequently, $f\omega$ is a smooth differential form which is identically zero and by Proposition 8.9 (1),

$$d(f\omega) = df \wedge \omega + f d\omega,$$

which, evaluated ap p, yields

$$0 = 0 \wedge \omega_p + 1 d\omega_p,$$

that is, $d\omega_p = 0$, as claimed. □

As in the case of differential forms on $\mathbb{R}^n$, the operator d is uniquely determined by the properties of Theorem 8.5.

Theorem 8.11. *Let M be a smooth manifold. There is a unique linear local operator $d\colon \mathcal{A}^*(M) \to \mathcal{A}^*(M)$, with $d = (d^k)$ and $d^k\colon \mathcal{A}^k(M) \to \mathcal{A}^{k+1}(M)$ for every $k \geq 0$, such that*

(1) $(df)_p = df_p$, where df_p is the differential of f at $p \in M$, for every $f \in \mathcal{A}^0(M) = C^\infty(M)$.

(2) $d \circ d = 0$.

(3) For every $\omega \in \mathcal{A}^r(M)$ and every $\eta \in \mathcal{A}^s(M)$,

$$d(\omega \wedge \eta) = d\omega \wedge \eta + (-1)^r \omega \wedge d\eta.$$

Proof. Existence has already been established. It is enough to prove uniqueness locally. If (U, φ) is any chart and $x_i = pr_i \circ \varphi$ are the corresponding local coordinate maps, we know that every k-form, $\omega \in \mathcal{A}^k(M)$, can be written uniquely as

$$\omega = \sum_I f_I dx_{i_1} \wedge \cdots \wedge dx_{i_k} \qquad p \in U.$$

Consequently, the proof of Theorem 8.5 will go through if we can show that $ddx_{i_j} \restriction U = 0$, since then,

$$d(f_I dx_{i_1} \wedge \cdots \wedge dx_{i_k}) = df_I \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k}.$$

The problem is that dx_{i_j} is only defined on U. However, using Proposition 3.30 again, for every $p \in U$, there some open subset, $V \subseteq U$, containing p and a smooth function, $f\colon M \to \mathbb{R}$, such that $\operatorname{supp} f \subseteq U$ and $f \equiv 1$ on V. Then, fx_{i_j} is a smooth form defined on M such that $fx_{i_j} \restriction V = x_{i_j} \restriction V$, so by Proposition 8.10 (applied twice),

$$0 = dd(fx_{i_j}) \restriction V = ddx_{i_j} \restriction V,$$

which concludes the proof. □

Remark: A closer look at the proof of Theorem 8.11 shows that it is enough to assume $dd\omega = 0$ on forms $\omega \in \mathcal{A}^0(M) = C^\infty(M)$.

Smooth differential forms can also be defined in terms of alternating $C^\infty(M)$-multilinear maps on smooth vector fields. Let $\omega \in \mathcal{A}^p(M)$ be any smooth k-form on M. Then, ω induces an alternating multilinear map

$$\omega\colon \underbrace{\mathfrak{X}(M) \times \cdots \times \mathfrak{X}(M)}_{k} \longrightarrow C^\infty(M)$$

as follows: For any k smooth vector fields, $X_1, \ldots, X_k \in \mathfrak{X}(M)$,

$$\omega(X_1, \ldots, X_k)(p) = \omega_p(X_1(p), \ldots, X_k(p)).$$

This map is obviously alternating and $\mathbb{R}$-linear, but it is also $C^\infty(M)$-linear, since for every $f \in C^\infty(M)$,

$$\begin{aligned} \omega(X_1, \ldots, fX_i, \ldots X_k)(p) &= \omega_p(X_1(p), \ldots, f(p)X_i(p), \ldots, X_k(p)) \\ &= f(p)\omega_p(X_1(p), \ldots, X_i(p), \ldots, X_k(p)) \\ &= (f\omega)_p(X_1(p), \ldots, X_i(p), \ldots, X_k(p)). \end{aligned}$$

(Recall, that the set of smooth vector fields, $\mathfrak{X}(M)$, is a real vector space and a $C^\infty(M)$-module.)

Interestingly, every alternating $C^\infty(M)$-multilinear maps on smooth vector fields determines a differential form. This is because $\omega(X_1, \ldots, X_k)(p)$ only depends on the values of $X_1, \ldots, X_k$ at p.

Proposition 8.12. *Let M be a smooth manifold. For every $k \geq 0$, there is an isomorphism between the space of k-forms, $\mathcal{A}^k(M)$, and the space, $\mathrm{Alt}^k_{C^\infty(M)}(\mathfrak{X}(M))$, of alternating $C^\infty(M)$-multilinear maps on smooth vector fields. That is,*

$$\mathcal{A}^k(M) \cong \mathrm{Alt}^k_{C^\infty(M)}(\mathfrak{X}(M)),$$

viewed as $C^\infty(M)$-modules.

Proof. Let $\Phi\colon \underbrace{\mathfrak{X}(M) \times \cdots \times \mathfrak{X}(M)}_{k} \longrightarrow C^\infty(M)$ be an alternating $C^\infty(M)$-multilinear map. First, we prove that for any vector fields $X_1, \ldots, X_k$ and $Y_1, \ldots, Y_k$, for every $p \in M$, if $X_i(p) = Y_i(p)$, then

$$\Phi(X_1, \ldots, X_k)(p) = \Phi(Y_1, \ldots, Y_k)(p).$$

Observe that

$$\begin{aligned}\Phi(X_1, \ldots, X_k) - \Phi(Y_1, \ldots, Y_k) &= \Phi(X_1 - Y_1, X_2, \ldots, X_k) + \Phi(Y_1, X_2 - Y_2, X_3, \ldots, X_k)\\ &\quad + \Phi(Y_1, Y_2, X_3 - Y_3, \ldots, X_k) + \cdots\\ &\quad + \Phi(Y_1, \ldots, Y_{k-2}, X_{k-1} - Y_{k-1}, X_k)\\ &\quad + \cdots + \Phi(Y_1, \ldots, Y_{k-1}, X_k - Y_k).\end{aligned}$$

As a consequence, it is enough to prove that if $X_i(p) = 0$, for some i, then

$$\Phi(X_1, \ldots, X_k)(p) = 0.$$

Without loss of generality, assume $i = 1$. In any local chart, (U, φ), near p, we can write

$$X_1 = \sum_{i=1}^{n} f_i \frac{\partial}{\partial x_i},$$

and as $X_i(p) = 0$, we have $f_i(p) = 0$, for $i = 1, \ldots, n$. Since the expression on the right-hand side is only defined on U, we extend it using Proposition 3.30, once again. There is some open subset, $V \subseteq U$, containing p and a smooth function, $h\colon M \to \mathbb{R}$, such that $\operatorname{supp} h \subseteq U$ and $h \equiv 1$ on V. Then, we let $h_i = hf_i$, a smooth function on M, $Y_i = h\frac{\partial}{\partial x_i}$, a smooth vector field on M, and we have $h_i \restriction V = f_i \restriction V$ and $Y_i \restriction V = \frac{\partial}{\partial x_i} \restriction V$. Now, it it obvious that

$$X_1 = \sum_{i=1}^{n} h_i Y_i + (1 - h^2)X_1$$

on V, so as Φ is $C^\infty(M)$-multilinear, $h_i(p) = 0$ and $h(p) = 1$, we get

$$\begin{aligned}\Phi(X_1, X_2, \ldots, X_k)(p) &= \Phi(\sum_{i=1}^{n} h_i Y_i + (1 - h^2)X_1, X_2, \ldots, X_k)(p)\\ &= \sum_{i=1}^{n} h_i(p)\Phi(Y_i, X_2, \ldots, X_k)(p) + (1 - h^2(p))\Phi(X_1, X_2, \ldots, X_k)(p) = 0,\end{aligned}$$

as claimed.

Next, we show that Φ induces a smooth differential form. For every $p \in M$, for any $u_1, \ldots, u_k \in T_pM$, we can pick smooth functions, f_i, equal to 1 near p and 0 outside some open near p so that we get smooth vector fields, $X_1, \ldots, X_k$, with $X_k(p) = u_k$. We set

$$\omega_p(u_1, \ldots, u_k) = \Phi(X_1, \ldots, X_k)(p).$$

As we proved that $\Phi(X_1, \ldots, X_k)(p)$ only depends on $X_1(p) = u_1, \ldots, X_k(p) = u_k$, the function ω_p is well defined and it is easy to check that it is smooth. Therefore, the map, $\Phi \mapsto \omega$, just defined is indeed an isomorphism. □

Remarks:

(1) The space, $\mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), C^\infty(M))$, of all $C^\infty(M)$-linear maps, $\mathfrak{X}(M) \longrightarrow C^\infty(M)$, is also a $C^\infty(M)$-module called the *dual* of $\mathfrak{X}(M)$ and sometimes denoted $\mathfrak{X}^*(M)$. Proposition 8.12 shows that as $C^\infty(M)$-modules,

$$\mathcal{A}^1(M) \cong \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), C^\infty(M)) = \mathfrak{X}^*(M).$$

(2) A result analogous to Proposition 8.12 holds for tensor fields. Indeed, there is an isomorphism between the set of tensor fields, $\Gamma(M, T^{r,s}(M))$, and the set of $C^\infty(M)$-multilinear maps,

$$\Phi\colon \underbrace{\mathcal{A}^1(M) \times \cdots \times \mathcal{A}^1(M)}_{r} \times \underbrace{\mathfrak{X}(M) \times \cdots \times \mathfrak{X}(M)}_{s} \longrightarrow C^\infty(M),$$

where $\mathcal{A}^1(M)$ and $\mathfrak{X}(M)$ are $C^\infty(M)$-modules.

Recall from Section 3.3 (Definition 3.19) that for any function, $f \in C^\infty(M)$, and every vector field, $X \in \mathfrak{X}(M)$, the Lie derivative, $X[f]$ (or $X(f)$) of f w.r.t. X is defined so that

$$X[f]_p = df_p(X(p)).$$

Also recall the notion of the *Lie bracket*, $[X, Y]$, of two vector fields (see Definition 3.20). The interpretation of differential forms as $C^\infty(M)$-multilinear forms given by Proposition 8.12 yields the following formula for $(d\omega)(X_1, \ldots, X_{k+1})$, where the X_i are vector fields:

Proposition 8.13. *Let M be a smooth manifold. For every k-form, $\omega \in \mathcal{A}^k(M)$, we have*

$$\begin{aligned}(d\omega)(X_1, \ldots, X_{k+1}) = \sum_{i=1}^{k+1} (-1)^{i-1} X_i[\omega(X_1, \ldots, \widehat{X_i}, \ldots, X_{k+1})] \\ + \sum_{i<j} (-1)^{i+j} \omega([X_i, X_j], X_1, \ldots, \widehat{X_i}, \ldots, \widehat{X_j}, \ldots, X_{k+1})],\end{aligned}$$

for all vector fields, $X_1, \ldots, X_{k+1} \in \mathfrak{X}(M)$:

Proof sketch. First, one checks that the right-hand side of the formula in Proposition 8.13 is alternating and $C^\infty(M)$-multilinear. For this, use Proposition 3.19 (c). Consequently, by Proposition 8.12, this expression defines a $(k+1)$-form. Second, it is enough to check that both sides of the equation agree on charts, (U, φ). Then, we know that $d\omega$ can be written uniquely as

$$\omega = \sum_I f_I dx_{i_1} \wedge \cdots \wedge dx_{i_k} \qquad p \in U.$$

Also, as differential forms are $C^\infty(M)$-multilinear, it is enough to consider vector fields of the form $X_i = \frac{\partial}{\partial x_{j_i}}$. However, for such vector fields, $[X_i, X_j] = 0$, and then it is a simple matter to check that the equation holds. For more details, see Morita [115] (Chapter 2). □

In particular, when $k = 1$, Proposition 8.13 yields the often used formula:

$$d\omega(X, Y) = X[\omega(Y)] - Y[\omega(X)] - \omega([X, Y]).$$

There are other ways of proving the formula of Proposition 8.13, for instance, using Lie derivatives.

Before considering the Lie derivative of differential forms, $L_X\omega$, we define interior multiplication by a vector field, $i(X)(\omega)$. We will see shortly that there is a relationship between L_X, $i(X)$ and d, known as *Cartan's Formula*.

Definition 8.9. Let M be a smooth manifold. For every vector field, $X \in \mathfrak{X}(M)$, for all $k \geq 1$, there is a linear map, $i(X)\colon \mathcal{A}^k(M) \to \mathcal{A}^{k-1}(M)$, defined so that, for all $\omega \in \mathcal{A}^k(M)$, for all $p \in M$, for all $u_1, \ldots, u_{k-1} \in T_pM$,

$$(i(X)\omega)_p(u_1, \ldots, u_{k-1}) = \omega_p(X_p, u_1, \ldots, u_{k-1}).$$

Obviously, $i(X)$ is $C^\infty(M)$-linear in X and it is easy to check that $i(X)\omega$ is indeed a smooth $(k-1)$-form. When $k = 0$, we set $i(X)\omega = 0$. Observe that $i(X)\omega$ is also given by

$$(i(X)\omega)_p = i(X_p)\omega_p, \qquad p \in M,$$

where $i(X_p)$ is the interior product (or insertion operator) defined in Section 22.17 (with $i(X_p)\omega_p$ equal to our right hook, $\omega_p \llcorner X_p$). As a consequence, by Proposition 22.28, the operator $i(X)$ is an anti-derivation of degree -1, that is, we have

$$i(X)(\omega \wedge \eta) = (i(X)\omega) \wedge \eta + (-1)^r \omega \wedge (i(X)\eta),$$

for all $\omega \in \mathcal{A}^r(M)$ and all $\eta \in \mathcal{A}^s(M)$.

Remark: Other authors, including Marsden, use a left hook instead of a right hook and denote $i(X)\omega$ as $X \lrcorner \omega$.

8.3 Lie Derivatives

We just saw in Section 8.2 that for any function, $f \in C^\infty(M)$, and every vector field, $X \in \mathfrak{X}(M)$, the Lie derivative, $X[f]$ (or $X(f)$) of f w.r.t. X is defined so that

$$X[f]_p = df_p(X_p).$$

Recall from Definition 3.28 and the observation immediately following it that for any manifold, M, given any two vector fields, $X, Y \in \mathfrak{X}(M)$, the *Lie derivative of X with respect to Y* is given by

$$(L_X\, Y)_p = \lim_{t \longrightarrow 0} \frac{\left(\Phi_t^* Y\right)_p - Y_p}{t} = \frac{d}{dt}\left(\Phi_t^* Y\right)_p \Big|_{t=0},$$

where Φ_t is the local one-parameter group associated with X (Φ is the global flow associated with X, see Definition 3.27, Theorem 3.27 and the remarks following it) and Φ_t^* is the pull-back of the diffeomorphism Φ_t (see Definition 3.21). Furthermore, recall that

$$L_X Y = [X, Y].$$

We claim that we also have

$$X_p[f] = \lim_{t \longrightarrow 0} \frac{(\Phi_t^* f)(p) - f(p)}{t} = \frac{d}{dt}(\Phi_t^* f)(p) \Big|_{t=0},$$

with $\Phi_t^* f = f \circ \Phi_t$ (as usual for functions).

Recall from Section 3.5 that if Φ is the flow of X, then for every $p \in M$, the map, $t \mapsto \Phi_t(p)$, is an integral curve of X through p, that is

$$\dot{\Phi}_t(p) = X(\Phi_t(p)), \qquad \Phi_0(p) = p,$$

in some open set containing p. In particular, $\dot{\Phi}_0(p) = X_p$. Then, we have

$$\begin{aligned}
\lim_{t \longrightarrow 0} \frac{(\Phi_t^* f)(p) - f(p)}{t} &= \lim_{t \longrightarrow 0} \frac{f(\Phi_t(p)) - f(\Phi_0(p))}{t} \\
&= \frac{d}{dt}(f \circ \Phi_t(p)) \Big|_{t=0} \\
&= df_p(\dot{\Phi}_0(p)) = df_p(X_p) = X_p[f].
\end{aligned}$$

We would like to define the Lie derivative of differential forms (and tensor fields). This can be done algebraically or in terms of flows, the two approaches are equivalent but it seems more natural to give a definition using flows.

Definition 8.10. Let M be a smooth manifold. For every vector field, $X \in \mathfrak{X}(M)$, for every k-form, $\omega \in \mathcal{A}^k(M)$, the *Lie derivative of ω with respect to X*, denoted $L_X\omega$ is given by

$$(L_X\omega)_p = \lim_{t \longrightarrow 0} \frac{\left(\Phi_t^* \omega\right)_p - \omega_p}{t} = \frac{d}{dt}\left(\Phi_t^* \omega\right)_p \Big|_{t=0},$$

where $\Phi_t^*\omega$ is the pull-back of ω along Φ_t (see Definition 8.7).

Obviously, $L_X \colon \mathcal{A}^k(M) \to \mathcal{A}^k(M)$ is a linear map but it has many other interesting properties. We can also define the Lie derivative on tensor fields as a map,
$L_X \colon \Gamma(M, T^{r,s}(M)) \to \Gamma(M, T^{r,s}(M))$, by requiring that for any tensor field,

$$\alpha = X_1 \otimes \cdots \otimes X_r \otimes \omega_1 \otimes \cdots \otimes \omega_s,$$

where $X_i \in \mathfrak{X}(M)$ and $\omega_j \in \mathcal{A}^1(M)$,

$$\Phi_t^* \alpha = \Phi_t^* X_1 \otimes \cdots \otimes \Phi_t^* X_r \otimes \Phi_t^* \omega_1 \otimes \cdots \otimes \Phi_t^* \omega_s,$$

where $\Phi_t^* X_i$ is the pull-back of the vector field, X_i, and $\Phi_t^* \omega_j$ is the pull-back of one-form, ω_j, and then setting

$$(L_X \alpha)_p = \lim_{t \longrightarrow 0} \frac{\left(\Phi_t^* \alpha\right)_p - \alpha_p}{t} = \frac{d}{dt} \left(\Phi_t^* \alpha\right)_p \Big|_{t=0}.$$

So, as long we can define the "right" notion of pull-back, the formula giving the Lie derivative of a function, a vector field, a differential form and more generally, a tensor field, is the same.

The Lie derivative of tensors is used in most areas of mechanics, for example in elasticity (the rate of strain tensor) and in fluid dynamics.

We now state, mostly without proofs, a number of properties of Lie derivatives. Most of these proofs are fairly straightforward computations, often tedious, and can be found in most texts, including Warner [148], Morita [115] and Gallot, Hullin and Lafontaine [61].

Proposition 8.14. *Let M be a smooth manifold. For every vector field, $X \in \mathfrak{X}(M)$, the following properties hold:*

(1) For all $\omega \in \mathcal{A}^r(M)$ and all $\eta \in \mathcal{A}^s(M)$,

$$L_X(\omega \wedge \eta) = (L_X \omega) \wedge \eta + \omega \wedge (L_X \eta),$$

that is, L_X is a derivation.

(2) For all $\omega \in \mathcal{A}^k(M)$, for all $Y_1, \ldots, Y_k \in \mathfrak{X}(M)$,

$$L_X(\omega(Y_1, \ldots, Y_k)) = (L_X \omega)(Y_1, \ldots, Y_k) + \sum_{i=1}^{k} \omega(Y_1, \ldots, Y_{i-1}, L_X Y_i, Y_{i+1}, \ldots, Y_k).$$

(3) The Lie derivative commutes with d:

$$L_X \circ d = d \circ L_X.$$

Proof. We only prove (2). First, we claim that if $\varphi\colon M \to M$ is a diffeomorphism, then for every $\omega \in \mathcal{A}^k(M)$, for all $X_1, \ldots, X_k \in \mathfrak{X}(M)$,

$$(\varphi^*\omega)(X_1, \ldots, X_k) = \varphi^*(\omega((\varphi^{-1})^*X_1, \ldots, (\varphi^{-1})^*X_k)), \tag{$*$}$$

where $(\varphi^{-1})^*X_i$ is the pull-back of the vector field, X_i (also equal to the push-forward, φ_*X_i, of X_i, see Definition 3.21). Recall that

$$((\varphi^{-1})^*Y)_p = d\varphi_{\varphi^{-1}(p)}(Y_{\varphi^{-1}(p)}),$$

for any vector field, Y. Then, for every $p \in M$, we have

$$\begin{aligned}
(\varphi^*\omega(X_1, \ldots, X_k))(p) &= \omega_{\varphi(p)}(d\varphi_p(X_1(p)), \ldots, d\varphi_p(X_k(p))) \\
&= \omega_{\varphi(p)}\big(d\varphi_{\varphi^{-1}(\varphi(p))}(X_1(\varphi^{-1}(\varphi(p)))), \ldots, d\varphi_{\varphi^{-1}(\varphi(p))}(X_k(\varphi^{-1}(\varphi(p))))\big) \\
&= \omega_{\varphi(p)}(((\varphi^{-1})^*X_1)_{\varphi(p)}, \ldots, ((\varphi^{-1})^*X_k)_{\varphi(p)}) \\
&= (\omega((\varphi^{-1})^*X_1, \ldots, (\varphi^{-1})^*X_k)) \circ \varphi)(p) \\
&= \varphi^*(\omega((\varphi^{-1})^*X_1, \ldots, (\varphi^{-1})^*X_k))(p),
\end{aligned}$$

since for any function, $g \in C^\infty(M)$, we have $\varphi^*g = g \circ \varphi$.

We know that

$$X_p[f] = \lim_{t \longrightarrow 0} \frac{(\Phi_t^* f)(p) - f(p)}{t}$$

and for any vector field, Y,

$$[X, Y]_p = (L_X Y)_p = \lim_{t \longrightarrow 0} \frac{\big(\Phi_t^* Y\big)_p - Y_p}{t}.$$

Since the one-parameter group associated with $-X$ is Φ_{-t} (this follows from $\Phi_{-t} \circ \Phi_t = \mathrm{id}$), we have

$$\lim_{t \longrightarrow 0} \frac{\big(\Phi_{-t}^* Y\big)_p - Y_p}{t} = -[X, Y]_p.$$

Now, using $\Phi_t^{-1} = \Phi_{-t}$ and $(*)$, we have

$$\begin{aligned}
(L_X\omega)(Y_1, \ldots, Y_k) &= \lim_{t \longrightarrow 0} \frac{(\Phi_t^*\omega)(Y_1, \ldots, Y_k) - \omega(Y_1, \ldots, Y_k)}{t} \\
&= \lim_{t \longrightarrow 0} \frac{\Phi_t^*(\omega(\Phi_{-t}^*Y_1, \ldots, \Phi_{-t}^*Y_k)) - \omega(Y_1, \ldots, Y_k)}{t} \\
&= \lim_{t \longrightarrow 0} \frac{\Phi_t^*(\omega(\Phi_{-t}^*Y_1, \ldots, \Phi_{-t}^*Y_k)) - \Phi_t^*(\omega(Y_1, \ldots, Y_k))}{t} \\
&\quad + \lim_{t \longrightarrow 0} \frac{\Phi_t^*(\omega(Y_1, \ldots, Y_k)) - \omega(Y_1, \ldots, Y_k)}{t}.
\end{aligned}$$

Call the first term A and the second term B. Then, as

$$X_p[f] = \lim_{t \longrightarrow 0} \frac{(\Phi_t^* f)(p) - f(p)}{t},$$

we have

$$B = X[\omega(Y_1, \ldots, Y_k)].$$

As to A, we have

$$\begin{aligned}
A &= \lim_{t \longrightarrow 0} \frac{\Phi_t^*(\omega(\Phi_{-t}^* Y_1, \ldots, \Phi_{-t}^* Y_k)) - \Phi_t^*(\omega(Y_1, \ldots, Y_k))}{t} \\
&= \lim_{t \longrightarrow 0} \Phi_t^* \left(\frac{\omega(\Phi_{-t}^* Y_1, \ldots, \Phi_{-t}^* Y_k) - \omega(Y_1, \ldots, Y_k)}{t} \right) \\
&= \lim_{t \longrightarrow 0} \Phi_t^* \left(\frac{\omega(\Phi_{-t}^* Y_1, \ldots, \Phi_{-t}^* Y_k) - \omega(Y_1, \Phi_{-t}^* Y_2, \ldots, \Phi_{-t}^* Y_k)}{t} \right) \\
&\quad + \lim_{t \longrightarrow 0} \Phi_t^* \left(\frac{\omega(Y_1, \Phi_{-t}^* Y_2, \ldots, \Phi_{-t}^* Y_k) - \omega(Y_1, Y_2, \Phi_{-t}^* Y_3, \ldots, \Phi_{-t}^* Y_k)}{t} \right) \\
&\quad + \cdots + \lim_{t \longrightarrow 0} \Phi_t^* \left(\frac{\omega(Y_1, \ldots, Y_{k-1}, \Phi_{-t}^* Y_k) - \omega(Y_1, \ldots, Y_k)}{t} \right) \\
&= \sum_{i=1}^{k} \omega(Y_1, \ldots, -[X, Y_i], \ldots, Y_k).
\end{aligned}$$

When we add up A and B, we get

$$\begin{aligned}
A + B &= X[\omega(Y_1, \ldots, Y_k)] - \sum_{i=1}^{k} \omega(Y_1, \ldots, [X, Y_i], \ldots, Y_k) \\
&= (L_X \omega)(Y_1, \ldots, Y_k),
\end{aligned}$$

which finishes the proof. □

Part (2) of Proposition 8.14 shows that the Lie derivative of a differential form can be defined in terms of the Lie derivatives of functions and vector fields:

$$\begin{aligned}
(L_X \omega)(Y_1, \ldots, Y_k) &= L_X(\omega(Y_1, \ldots, Y_k)) - \sum_{i=1}^{k} \omega(Y_1, \ldots, Y_{i-1}, L_X Y_i, Y_{i+1}, \ldots, Y_k) \\
&= X[\omega(Y_1, \ldots, Y_k)] - \sum_{i=1}^{k} \omega(Y_1, \ldots, Y_{i-1}, [X, Y_i], Y_{i+1}, \ldots, Y_k).
\end{aligned}$$

The following proposition is known as *Cartan's Formula*:

Proposition 8.15. *(Cartan's Formula) Let M be a smooth manifold. For every vector field, $X \in \mathfrak{X}(M)$, for every $\omega \in \mathcal{A}^k(M)$, we have*

$$L_X \omega = i(X) d\omega + d(i(X)\omega),$$

that is, $L_X = i(X) \circ d + d \circ i(X)$.

Proof. If $k = 0$, then $L_X f = X[f] = df(X)$ for a function, f, and on the other hand, $i(X)f = 0$ and $i(X)df = df(X)$, so the equation holds. If $k \geq 1$, then by Proposition 8.13, we have

$$\begin{aligned}(i(X)d\omega)(X_1,\ldots,X_k) &= d\omega(X,X_1,\ldots,X_k)\\ &= X[\omega(X_1,\ldots,X_k)] + \sum_{i=1}^{k}(-1)^i X_i[\omega(X,X_1,\ldots,\widehat{X_i},\ldots,X_k)]\\ &\quad + \sum_{j=1}^{k}(-1)^j\omega([X,X_j],X_1,\ldots,\widehat{X_j},\ldots,X_k)\\ &\quad + \sum_{i<j}(-1)^{i+j}\omega([X_i,X_j],X,X_1,\ldots,\widehat{X_i},\ldots,\widehat{X_j},\ldots,X_k).\end{aligned}$$

On the other hand, again by Proposition 8.13, we have

$$\begin{aligned}(di(X)\omega)(X_1,\ldots,X_k) &= \sum_{i=1}^{k}(-1)^{i-1}X_i[\omega(X,X_1,\ldots,\widehat{X_i},\ldots,X_k)]\\ &\quad + \sum_{i<j}(-1)^{i+j}\omega(X,[X_i,X_j],X_1,\ldots,\widehat{X_i},\ldots,\widehat{X_j},\ldots,X_k).\end{aligned}$$

Adding up these two equations, we get

$$\begin{aligned}(i(X)d\omega + di(X))\omega(X_1,\ldots,X_k) &= X[\omega(X_1,\ldots,X_k)]\\ &\quad + \sum_{i=1}^{k}(-1)^i\omega([X,X_i],X_1,\ldots,\widehat{X_i},\ldots,X_k)\\ &= X[\omega(X_1,\ldots,X_k)] - \sum_{i=1}^{k}\omega(X_1,\ldots,[X,X_i],\ldots,X_k) = (L_X\omega)(X_1,\ldots,X_k),\end{aligned}$$

as claimed. □

The following proposition states more useful identities, some of which can be proved using Cartan's formula:

Proposition 8.16. *Let M be a smooth manifold. For all vector fields, $X, Y \in \mathfrak{X}(M)$, for all $\omega \in \mathcal{A}^k(M)$, we have*

(1) $L_X i(Y) - i(Y)L_X = i([X,Y])$.

(2) $L_X L_Y\omega - L_Y L_X\omega = L_{[X,Y]}\omega$.

(3) $L_X i(X)\omega = i(X)L_X\omega$.

(4) $L_{fX}\omega = fL_X\omega + df \wedge i(X)\omega$, *for all* $f \in C^\infty(M)$.

(5) For any diffeomorphism, $\varphi\colon M \to N$, *for all* $Z \in \mathfrak{X}(N)$ *and all* $\beta \in \mathcal{A}^k(N)$,

$$\varphi^* L_Z \beta = L_{\varphi^* Z} \varphi^* \beta.$$

Finally, here is a proposition about the Lie derivative of tensor fields. Obviously, tensor product and contraction of tensor fields are defined pointwise on fibres, that is

$$\begin{aligned} (\alpha \otimes \beta)_p &= \alpha_p \otimes \beta_p \\ (c_{i,j}\alpha)_p &= c_{i,j}\alpha_p, \end{aligned}$$

for all $p \in M$, where $c_{i,j}$ is the contraction operator of Definition 22.5.

Proposition 8.17. *Let* M *be a smooth manifold. For every vector field,* $X \in \mathfrak{X}(M)$, *the Lie derivative,* $L_X\colon \Gamma(M, T^{\bullet,\bullet}(M)) \to \Gamma(M, T^{\bullet,\bullet}(M))$, *is the unique linear local operator satisfying the following properties:*

(1) $L_X f = X[f] = df(X)$, *for all* $f \in C^\infty(M)$.

(2) $L_X Y = [X, Y]$, *for all* $Y \in \mathfrak{X}(M)$.

(3) $L_X(\alpha \otimes \beta) = (L_X\alpha) \otimes \beta + \alpha \otimes (L_X\beta)$, *for all tensor fields,* $\alpha \in \Gamma(M, T^{r_1,s_1}(M))$ *and* $\beta \in \Gamma(M, T^{r_2,s_2}(M))$, *that is,* L_X *is a derivation.*

(4) For all tensor fields $\alpha \in \Gamma(M, T^{r,s}(M))$, *with* $r, s > 0$, *for every contraction operator,* $c_{i,j}$,

$$L_X(c_{i,j}(\alpha)) = c_{i,j}(L_X\alpha).$$

The proof of Proposition 8.17 can be found in Gallot, Hullin and Lafontaine [61] (Chapter 1). The following proposition is also useful:

Proposition 8.18. *For every* $(0, q)$*-tensor,* $S \in \Gamma(M, (T^*)^{\otimes q}(M))$, *we have*

$$(L_X S)(X_1, \ldots, X_q) = X[S(X_1, \ldots, X_q)] - \sum_{i=1}^{q} S(X_1, \ldots, [X, X_i], \ldots, X_q),$$

for all $X_1, \ldots, X_q, X \in \mathfrak{X}(M)$.

There are situations in differential geometry where it is convenient to deal with differential forms taking values in a vector space. This happens when we consider connections and the curvature form on vector bundles and principal bundles and when we study Lie groups, where differential forms valued in a Lie algebra occur naturally.

8.4 Vector-Valued Differential Forms

Let us go back for a moment to differential forms defined on some open subset of $\mathbb{R}^n$. In Section 8.1, a differential form is defined as a smooth map, $\omega\colon U \to \bigwedge^p(\mathbb{R}^n)^*$, and since we have a canonical isomorphism,

$$\mu\colon \bigwedge^p(\mathbb{R}^n)^* \cong \mathrm{Alt}^p(\mathbb{R}^n;\mathbb{R}),$$

such differential forms are real-valued. Now, let F be any normed vector space, possibly infinite dimensional. Then, $\mathrm{Alt}^p(\mathbb{R}^n;F)$ is also a normed vector space and by Proposition 22.33, we have a canonical isomorphism

$$\mu\colon \left(\bigwedge^p(\mathbb{R}^n)^*\right)\otimes F \longrightarrow \mathrm{Alt}^p(\mathbb{R}^n;F).$$

Then, it is natural to define differential forms with values in F as smooth maps,
$\omega\colon U \to \mathrm{Alt}^p(\mathbb{R}^n;F)$. Actually, we can even replace $\mathbb{R}^n$ with any normed vector space, even infinite dimensional, as in Cartan [30], but we do not need such generality for our purposes.

Definition 8.11. Let F by any normed vector space. Given any open subset, U, of $\mathbb{R}^n$, a smooth *differential p-form on U with values in F*, for short, *p-form on U*, is any smooth function, $\omega\colon U \to \mathrm{Alt}^p(\mathbb{R}^n;F)$. The vector space of all p-forms on U is denoted $\mathcal{A}^p(U;F)$. The vector space, $\mathcal{A}^*(U;F) = \bigoplus_{p\geq 0}\mathcal{A}^p(U;F)$, is the set of *differential forms on U with values in F*.

Observe that $\mathcal{A}^0(U;F) = C^\infty(U,F)$, the vector space of smooth functions on U with values in F and $\mathcal{A}^1(U;F) = C^\infty(U,\mathrm{Hom}(\mathbb{R}^n,F))$, the set of smooth functions from U to the set of linear maps from $\mathbb{R}^n$ to F. Also, $\mathcal{A}^p(U;F) = (0)$ for $p > n$.

Of course, we would like to have a "good" notion of exterior differential and we would like as many properties of "ordinary" differential forms as possible to remain valid. As will see in our somewhat sketchy presentation, these goals can be achieved except for some properties of the exterior product.

Using the isomorphism

$$\mu\colon \left(\bigwedge^p(\mathbb{R}^n)^*\right)\otimes F \longrightarrow \mathrm{Alt}^p(\mathbb{R}^n;F)$$

and Proposition 22.34, we obtain a convenient expression for differential forms in $\mathcal{A}^*(U;F)$. If $(e_1,\ldots,e_n)$ is any basis of $\mathbb{R}^n$ and $(e_1^*,\ldots,e_n^*)$ is its dual basis, then every differential p-form, $\omega \in \mathcal{A}^p(U;F)$, can be written uniquely as

$$\omega(x) = \sum_I e_{i_1}^*\wedge\cdots\wedge e_{i_p}^*\otimes f_I(x) = \sum_I e_I^*\otimes f_I(x) \qquad x\in U,$$

where each $f_I \colon U \to F$ is a smooth function on U. By Proposition 22.35, the above property can be restated as the fact every differential p-form, $\omega \in \mathcal{A}^p(U; F)$, can be written uniquely as

$$\omega(x) = \sum_I e^*_{i_1} \wedge \cdots \wedge e^*_{i_p} \cdot f_I(x), \qquad x \in U.$$

where each $f_I \colon U \to F$ is a smooth function on U.

As in Section 22.15 (following H. Cartan [30]) in order to define a multiplication on differential forms we use a bilinear form, $\Phi \colon F \times G \to H$. Then, we can define a multiplication, $\wedge_\Phi$, directly on alternating multilinear maps as follows: For $f \in \mathrm{Alt}^m(\mathbb{R}^n; F)$ and $g \in \mathrm{Alt}^n(\mathbb{R}^n; G)$,

$$(f \wedge_\Phi g)(u_1, \ldots, u_{m+n}) = \sum_{\sigma \in \mathrm{shuffle}(m,n)} \mathrm{sgn}(\sigma)\, \Phi(f(u_{\sigma(1)}, \ldots, u_{\sigma(m)}), g(u_{\sigma(m+1)}, \ldots, u_{\sigma(m+n)})),$$

where $\mathrm{shuffle}(m, n)$ consists of all (m, n)-"shuffles", that is, permutations, σ, of $\{1, \ldots m+n\}$, such that $\sigma(1) < \cdots < \sigma(m)$ and $\sigma(m+1) < \cdots < \sigma(m+n)$.

Then, we obtain a multiplication,

$$\wedge_\Phi \colon \mathcal{A}^p(U; F) \times \mathcal{A}^q(U; G) \to \mathcal{A}^{p+q}(U; H),$$

defined so that, for any differential forms, $\omega \in \mathcal{A}^p(U; F)$ and $\eta \in \mathcal{A}^q(U; G)$,

$$(\omega \wedge_\Phi \eta)_x = \omega_x \wedge_\Phi \eta_x, \qquad x \in U.$$

In general, not much can be said about $\wedge_\Phi$ unless Φ has some additional properties. In particular, $\wedge_\Phi$ is generally not associative. In particular, there is no analog of Proposition 8.1. For simplicity of notation, we write $\wedge$ for $\wedge_\Phi$. Using Φ, we can also define a multiplication,

$$\cdot \colon \mathcal{A}^p(U; F) \times \mathcal{A}^0(U; G) \to \mathcal{A}^p(U; H),$$

given by

$$(\omega \cdot f)_x(u_1, \ldots, u_p) = \Phi(\omega_x(u_1, \ldots, u_p), f(x)),$$

for all $x \in U$ and all $u_1, \ldots, u_p \in \mathbb{R}^n$. This multiplication will be used in the case where $F = \mathbb{R}$ and $G = H$, to obtain a normal form for differential forms.

Generalizing d is no problem. Observe that since a differential p-form is a smooth map, $\omega \colon U \to \mathrm{Alt}^p(\mathbb{R}^n; F)$, its derivative is a map,

$$\omega' \colon U \to \mathrm{Hom}(\mathbb{R}^n, \mathrm{Alt}^p(\mathbb{R}^n; F)),$$

such that ω'_x is a linear map from $\mathbb{R}^n$ to $\mathrm{Alt}^p(\mathbb{R}^n; F)$, for every $x \in U$. We can view ω'_x as a multilinear map, $\omega'_x \colon (\mathbb{R}^n)^{p+1} \to F$, which is alternating in its last p arguments. As in Section 8.1, the exterior derivative, $(d\omega)_x$, is obtained by making ω'_x into an alternating map in all of its $p + 1$ arguments.

Definition 8.12. For every $p \geq 0$, the *exterior differential*, $d\colon \mathcal{A}^p(U;F) \to \mathcal{A}^{p+1}(U;F)$, is given by

$$(d\omega)_x(u_1,\ldots,u_{p+1}) = \sum_{i=1}^{p+1}(-1)^{i-1}\omega'_x(u_i)(u_1,\ldots,\widehat{u_i},\ldots,u_{p+1}),$$

for all $\omega \in \mathcal{A}^p(U;F)$ and all $u_1,\ldots,u_{p+1} \in \mathbb{R}^n$, where the hat over the argument u_i means that it should be omitted.

For any smooth function, $f \in \mathcal{A}^0(U;F) = C^\infty(U,F)$, we get

$$df_x(u) = f'_x(u).$$

Therefore, for smooth functions, *the exterior differential, df, coincides with the usual derivative, f'*. The important observation following Definition 8.3 also applies here. If $x_i\colon U \to \mathbb{R}$ is the restriction of pr_i to U, then x'_i *is the constant map* given by

$$x'_i(x) = pr_i, \qquad x \in U.$$

It follows that $dx_i = x'_i$ is the constant function with value $pr_i = e_i^*$. As a consequence, every p-form, ω, can be uniquely written as

$$\omega_x = \sum_I dx_{i_1} \wedge \cdots \wedge dx_{i_p} \otimes f_I(x)$$

where each $f_I\colon U \to F$ is a smooth function on U. Using the multiplication, $\cdot$, induced by the scalar multiplication in F ($\Phi(\lambda, f) = \lambda f$, with $\lambda \in \mathbb{R}$ and $f \in F$), we see that every p-form, ω, can be uniquely written as

$$\omega = \sum_I dx_{i_1} \wedge \cdots \wedge dx_{i_p} \cdot f_I.$$

As for real-valued functions, for any $f \in \mathcal{A}^0(U;F) = C^\infty(U,F)$, we have

$$df_x(u) = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(x)(u)\, e_i^*,$$

and so,

$$df = \sum_{i=1}^n dx_i \cdot \frac{\partial f}{\partial x_i}.$$

In general, Proposition 8.3 fails unless F is finite-dimensional (see below). However for any arbitrary F, a weak form of Proposition 8.3 can be salvaged. Again, let $\Phi\colon F \times G \to H$ be a bilinear form, let $\cdot\colon \mathcal{A}^p(U;F) \times \mathcal{A}^0(U;G) \to \mathcal{A}^p(U;H)$ be as defined before Definition 8.12 and let $\wedge_\Phi$ be the wedge product associated with Φ. The following fact is proved in Cartan [30] (Section 2.4):

Proposition 8.19. *For all $\omega \in \mathcal{A}^p(U;F)$ and all $f \in \mathcal{A}^0(U;G)$, we have*

$$d(\omega \cdot f) = (d\omega) \cdot f + \omega \wedge_\Phi df.$$

Fortunately, $d \circ d$ still vanishes but this requires a completely different proof since we can't rely on Proposition 8.2 (see Cartan [30], Section 2.5). Similarly, Proposition 8.2 holds but a different proof is needed.

Proposition 8.20. *The composition $\mathcal{A}^p(U;F) \xrightarrow{d} \mathcal{A}^{p+1}(U;F) \xrightarrow{d} \mathcal{A}^{p+2}(U;F)$ is identically zero for every $p \geq 0$, that is,*

$$d \circ d = 0,$$

or using superscripts, $d^{p+1} \circ d^p = 0$.

To generalize Proposition 8.2, we use Proposition 8.19 with the product, $\cdot$, and the wedge product, $\wedge_\Phi$, induced by the bilinear form, Φ, given by scalar multiplication in F, that, is $\Phi(\lambda, f) = \lambda f$, for all $\lambda \in \mathbb{R}$ and all $f \in F$.

Proposition 8.21. *For every p form, $\omega \in \mathcal{A}^p(U;F)$, with $\omega = dx_{i_1} \wedge \cdots \wedge dx_{i_p} \cdot f$, we have*

$$d\omega = dx_{i_1} \wedge \cdots \wedge dx_{i_p} \wedge_F df,$$

where $\wedge$ is the usual wedge product on real-valued forms and $\wedge_F$ is the wedge product associated with scalar multiplication in F.

More explicitly, for every $x \in U$, for all $u_1, \ldots, u_{p+1} \in \mathbb{R}^n$, we have

$$(d\omega_x)(u_1, \ldots, u_{p+1}) = \sum_{i=1}^{p+1} (-1)^{i-1} (dx_{i_1} \wedge \cdots \wedge dx_{i_p})_x (u_1, \ldots, \widehat{u_i}, \ldots, u_{p+1}) df_x(u_i).$$

If we use the fact that

$$df = \sum_{i=1}^{n} dx_i \cdot \frac{\partial f}{\partial x_i},$$

we see easily that

$$d\omega = \sum_{j=1}^{n} dx_{i_1} \wedge \cdots \wedge dx_{i_p} \wedge dx_j \cdot \frac{\partial f}{\partial x_j},$$

the direct generalization of the real-valued case, except that the "coefficients" are functions with values in F.

The pull-back of forms in $\mathcal{A}^*(V,F)$ is defined as before. Luckily, Proposition 8.6 holds (see Cartan [30], Section 2.8).

Proposition 8.22. *Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^m$ be two open sets and let $\varphi \colon U \to V$ be a smooth map. Then*

(i) $\varphi^*(\omega \wedge \eta) = \varphi^*\omega \wedge \varphi^*\eta$, *for all* $\omega \in \mathcal{A}^p(V;F)$ *and all* $\eta \in \mathcal{A}^q(V;F)$.

(ii) $\varphi^*(f) = f \circ \varphi$, *for all* $f \in \mathcal{A}^0(V;F)$.

(iii) $d\varphi^*(\omega) = \varphi^*(d\omega)$, *for all* $\omega \in \mathcal{A}^p(V;F)$, *that is, the following diagram commutes for all* $p \geq 0$*:*

$$\begin{array}{ccc} \mathcal{A}^p(V;F) & \xrightarrow{\varphi^*} & \mathcal{A}^p(U;F) \\ \Big\downarrow{\scriptstyle d} & & \Big\downarrow{\scriptstyle d} \\ \mathcal{A}^{p+1}(V;F) & \xrightarrow{\varphi^*} & \mathcal{A}^{p+1}(U;F). \end{array}$$

Let us now consider the special case where F has finite dimension m. Pick any basis, $(f_1, \ldots, f_m)$, of F. Then, as every differential p-form, $\omega \in \mathcal{A}^p(U;F)$, can be written uniquely as

$$\omega(x) = \sum_I e^*_{i_1} \wedge \cdots \wedge e^*_{i_p} \cdot f_I(x), \qquad x \in U.$$

where each $f_I \colon U \to F$ is a smooth function on U, by expressing the f_I over the basis, $(f_1, \ldots, f_m)$, we see that ω can be written uniquely as

$$\omega = \sum_{i=1}^m \omega_i \cdot f_i,$$

where $\omega_1, \ldots, \omega_m$ are smooth real-valued differential forms in $\mathcal{A}^p(U;\mathbb{R})$ and we view f_i as the constant map with value f_i from U to F. Then, as

$$\omega'_x(u) = \sum_{i=1}^m (\omega'_i)_x(u) f_i,$$

for all $u \in \mathbb{R}^n$, we see that

$$d\omega = \sum_{i=1}^m d\omega_i \cdot f_i.$$

Actually, because $d\omega$ is defined independently of bases, the f_i do not need to be linearly independent; any choice of vectors and forms such that

$$\omega = \sum_{i=1}^k \omega_i \cdot f_i,$$

will do.

Given a bilinear map, $\Phi \colon F \times G \to H$, a simple calculation shows that for all $\omega \in \mathcal{A}^p(U;F)$ and all $\eta \in \mathcal{A}^p(U;G)$, we have

$$\omega \wedge_\Phi \eta = \sum_{i=1}^m \sum_{j=1}^{m'} \omega_i \wedge \eta_j \cdot \Phi(f_i, g_j),$$

with $\omega = \sum_{i=1}^{m} \omega_i \cdot f_i$ and $\eta = \sum_{j=1}^{m'} \eta_j \cdot g_j$, where $(f_1, \ldots, f_m)$ is a basis of F and $(g_1, \ldots, g_{m'})$ is a basis of G. From this and Proposition 8.3, it follows that Proposition 8.3 holds for finite-dimensional spaces.

Proposition 8.23. *If F, G, H are finite dimensional and $\Phi\colon F \times G \to H$ is a bilinear map, then For all $\omega \in \mathcal{A}^p(U; F)$ and all $\eta \in \mathcal{A}^q(U; G)$,*

$$d(\omega \wedge_\Phi \eta) = d\omega \wedge_\Phi \eta + (-1)^p \omega \wedge_\Phi d\eta.$$

On the negative side, in general, Proposition 8.1 still fails.

A special case of interest is the case where $F = G = H = \mathfrak{g}$ is a Lie algebra and $\Phi(a, b) = [a, b]$, is the Lie bracket of $\mathfrak{g}$. In this case, using a basis, $(f_1, \ldots, f_r)$, of $\mathfrak{g}$ if we write $\omega = \sum_i \alpha_i f_i$ and $\eta = \sum_j \beta_j f_j$, we have

$$[\omega, \eta] = \sum_{i,j} \alpha_i \wedge \beta_j [f_i, f_j],$$

where, for simplicity of notation, we dropped the subscript, Φ, on $[\omega, \eta]$ and the multiplication sign, $\cdot$. Let us figure out what $[\omega, \omega]$ is for a one-form, $\omega \in \mathcal{A}^1(U, \mathfrak{g})$. By definition,

$$[\omega, \omega] = \sum_{i,j} \omega_i \wedge \omega_j [f_i, f_j],$$

so

$$\begin{aligned}
[\omega, \omega](u, v) &= \sum_{i,j} (\omega_i \wedge \omega_j)(u, v)[f_i, f_j] \\
&= \sum_{i,j} (\omega_i(u)\omega_j(v) - \omega_i(v)\omega_j(u))[f_i, f_j] \\
&= \sum_{i,j} \omega_i(u)\omega_j(v)[f_i, f_j] - \sum_{i,j} \omega_i(v)\omega_j(u)[f_i, f_j] \\
&= [\sum_i \omega_i(u) f_i - \sum_j \omega_j(v) f_j] - [\sum_i \omega_i(v) f_i - \sum_j \omega_j(u) f_j] \\
&= [\omega(u), \omega(v)] - [\omega(v), \omega(u)] \\
&= 2[\omega(u), \omega(v)].
\end{aligned}$$

Therefore,

$$[\omega, \omega](u, v) = 2[\omega(u), \omega(v)].$$

Note that in general, $[\omega, \omega] \neq 0$, because ω is vector valued. Of course, for real-valued forms, $[\omega, \omega] = 0$. Using the Jacobi identity of the Lie algebra, we easily find that

$$[[\omega, \omega], \omega] = 0.$$

The generalization of vector-valued differential forms to manifolds is no problem, except that some results involving the wedge product fail for the same reason that they fail in the case of forms on open subsets of $\mathbb{R}^n$.

Given a smooth manifold, M, of dimension n and a vector space, F, the set, $\mathcal{A}^k(M;F)$, of *differential k-forms on M with values in F* is the set of maps, $p \mapsto \omega_p$, with $\omega_p \in \left(\bigwedge^k T_p^*M\right) \otimes F \cong \mathrm{Alt}^k(T_pM;F)$, which vary smoothly in $p \in M$. This means that the map

$$p \mapsto \omega_p(X_1(p), \ldots, X_k(p))$$

is smooth for all vector fields, $X_1, \ldots, X_k \in \mathfrak{X}(M)$. Using the operations on vector bundles described in Section 7.3, we can define $\mathcal{A}^k(M;F)$ as the set of smooth sections of the vector bundle, $\left(\bigwedge^k T^*M\right) \otimes \epsilon_F$, that is, as

$$\mathcal{A}^k(M;F) = \Gamma\Big(\Big(\bigwedge^k T^*M\Big) \otimes \epsilon_F\Big),$$

where ϵ_F is the trivial vector bundle, $\epsilon_F = M \times F$. In view of Proposition 7.12 and since $\Gamma(\epsilon_F) \cong C^\infty(M;F)$ and $\mathcal{A}^k(M) = \Gamma\left(\bigwedge^k T^*M\right)$, we have

$$\begin{aligned}
\mathcal{A}^k(M;F) &= \Gamma\Big(\Big(\bigwedge^k T^*M\Big) \otimes \epsilon_F\Big) \\
&\cong \Gamma\Big(\bigwedge^k T^*M\Big) \otimes_{C^\infty(M)} \Gamma(\epsilon_F) \\
&= \mathcal{A}^k(M) \otimes_{C^\infty(M)} C^\infty(M;F) \\
&\cong \bigwedge_{C^\infty(M)}^k (\Gamma(TM))^* \otimes_{C^\infty(M)} C^\infty(M;F) \\
&\cong \mathrm{Alt}^k_{C^\infty(M)}(\mathfrak{X}(M); C^\infty(M;F)).
\end{aligned}$$

with all of the spaces viewed as $C^\infty(M)$-modules. Therefore,

$$\mathcal{A}^k(M;F) \cong \mathcal{A}^k(M) \otimes_{C^\infty(M)} C^\infty(M;F) \cong \mathrm{Alt}^k_{C^\infty(M)}(\mathfrak{X}(M); C^\infty(M;F)),$$

which reduces to Proposition 8.12 when $F = \mathbb{R}$. The reader may want to carry out the verification that the theory generalizes to manifolds on her/his own. In Section 11.1, we will consider a generalization of the above situation where the trivial vector bundle, ϵ_F, is replaced by any vector bundle, $\xi = (E, \pi, B, V)$, and where $M = B$.

In the next section, we consider some properties of differential forms on Lie groups.

8.5 Differential Forms on Lie Groups and Maurer-Cartan Forms

Given a Lie group, G, we saw in Section 5.2 that the set of left-invariant vector fields on G is isomorphic to the Lie algebra, $\mathfrak{g} = T_1G$, of G (where 1 denotes the identity element of G). Recall that a vector field, X, on G is left-invariant iff

$$d(L_a)_b(X_b) = X_{L_a b} = X_{ab},$$

for all $a, b \in G$. In particular, for $b = 1$, we get

$$X_a = d(L_a)_1(X_1).$$

which shows that X is completely determined by its value at 1. The map, $X \mapsto X(1)$, is an isomorphism between left-invariant vector fields on G and $\mathfrak{g}$.

The above suggests looking at left-invariant differential forms on G. We will see that the set of left-invariant one-forms on G is isomorphic to $\mathfrak{g}^*$, the dual of $\mathfrak{g}$, as a vector space.

Definition 8.13. Given a Lie group, G, a differential form, $\omega \in \mathcal{A}^k(G)$, is *left-invariant* iff

$$L_a^*\omega = \omega, \qquad \text{for all } a \in G,$$

where $L_a^*\omega$ is the pull-back of ω by L_a (left multiplication by a). The left-invariant one-forms, $\omega \in \mathcal{A}^1(G)$, are also called *Maurer-Cartan forms*.

For a one-form, $\omega \in \mathcal{A}^1(G)$, left-invariance means that

$$(L_a^*\omega)_g(u) = \omega_{L_a g}(d(L_a)_g u) = \omega_{ag}(d(L_a)_g u) = \omega_g(u),$$

for all $a, g \in G$ and all $u \in T_gG$. For $a = g^{-1}$, we get

$$\omega_g(u) = \omega_1(d(L_{g^{-1}})_g u) = \omega_1(d(L_g^{-1})_g u),$$

which shows that ω_g is completely determined by its value at $g = 1$.

We claim that the map, $\omega \mapsto \omega_1$, is an isomorphism between the set of left-invariant one-forms on G and $\mathfrak{g}^*$.

First, for any linear form, $\alpha \in \mathfrak{g}^*$, the one-form, α^L, given by

$$\alpha_g^L(u) = \alpha(d(L_g^{-1})_g u)$$

is left-invariant, because

$$\begin{aligned}
(L_h^*\alpha^L)_g(u) &= \alpha^L_{hg}(d(L_h)_g(u)) \\
&= \alpha(d(L_{hg}^{-1})_{hg}(d(L_h)_g(u))) \\
&= \alpha(d(L_{hg}^{-1} \circ L_h)_g(u)) \\
&= \alpha(d(L_g^{-1})_g(u)) = \alpha_g^L(u).
\end{aligned}$$

Second, we saw that for every one-form, $\omega \in \mathcal{A}^1(G)$,

$$\omega_g(u) = \omega_1(d(L_g^{-1})_g u),$$

so $\omega_1 \in \mathfrak{g}^*$ is the unique element such that $\omega = \omega_1^L$, which shows that the map $\alpha \mapsto \alpha^L$ is an isomorphism whose inverse is the map, $\omega \mapsto \omega_1$.

Now, since every left-invariant vector field is of the form $X = u^L$, for some unique, $u \in \mathfrak{g}$, where u^L is the vector field given by $u^L(a) = d(L_a)_1 u$, and since

$$\omega_{ag}(d(L_a)_g u) = \omega_g(u),$$

for $g = 1$, we get $\omega_a(d(L_a)_1 u) = \omega_1(u)$, that is

$$\omega(X)_a = \omega_1(u), \qquad a \in G,$$

which shows that $\omega(X)$ is a constant function on G. It follows that for every vector field, Y, (not necessarily left-invariant),

$$Y[\omega(X)] = 0.$$

Recall that as a special case of Proposition 8.13, we have

$$d\omega(X, Y) = X[\omega(Y)] - Y[\omega(X)] - \omega([X, Y]).$$

Consequently, for all left-invariant vector fields, X, Y, on G, for every left-invariant one-form, ω, we have

$$d\omega(X, Y) = -\omega([X, Y]).$$

If we identify the set of left-invariant vector fields on G with $\mathfrak{g}$ and the set of left-invariant one-forms on G with $\mathfrak{g}^*$, we have

$$d\omega(X, Y) = -\omega([X, Y]), \qquad \omega \in \mathfrak{g}^*,\ X, Y \in \mathfrak{g}.$$

We summarize these facts in the following proposition:

Proposition 8.24. *Let G be any Lie group.*

(1) The set of left-invariant one-forms on G is isomorphic to $\mathfrak{g}^*$, the dual of the Lie algebra, $\mathfrak{g}$, of G, via the isomorphism, $\omega \mapsto \omega_1$.

(2) For every left-invariant one form, ω, and every left-invariant vector field, X, the value of the function $\omega(X)$ is constant and equal to $\omega_1(X_1)$.

(3) If we identify the set of left-invariant vector fields on G with $\mathfrak{g}$ and the set of left-invariant one-forms on G with $\mathfrak{g}^*$, then

$$d\omega(X, Y) = -\omega([X, Y]), \qquad \omega \in \mathfrak{g}^*,\ X, Y \in \mathfrak{g}.$$

Pick any basis, $X_1, \ldots, X_r$, of the Lie algebra, $\mathfrak{g}$, and let $\omega_1, \ldots, \omega_r$ be the dual basis of $\mathfrak{g}^*$. Then, there are some constants, c_{ij}^k, such that

$$[X_i, X_j] = \sum_{k=1}^{r} c_{ij}^k X_k.$$

The constants, c_{ij}^k are called the *structure constants* of the Lie algebra, $\mathfrak{g}$. Observe that $c_{ji}^k = -c_{ij}^k$.

As $\omega_i([X_p, X_q]) = c_{pq}^i$ and $d\omega_i(X, Y) = -\omega_i([X, Y])$, we have

$$\begin{aligned}
\sum_{j,k} c_{jk}^i \omega_j \wedge \omega_k(X_p, X_q) &= \sum_{j,k} c_{jk}^i (\omega_j(X_p)\omega_k(X_q) - \omega_j(X_q)\omega_k(X_p)) \\
&= \sum_{j,k} c_{jk}^i \omega_j(X_p)\omega_k(X_q) - \sum_{j,k} c_{jk}^i \omega_j(X_q)\omega_k(X_p) \\
&= \sum_{j,k} c_{jk}^i \omega_j(X_p)\omega_k(X_q) + \sum_{j,k} c_{kj}^i \omega_j(X_q)\omega_k(X_p) \\
&= c_{p,q}^i + c_{p,q}^i = 2c_{p,q}^i,
\end{aligned}$$

so we get the equations

$$d\omega_i = -\frac{1}{2} \sum_{j,k} c_{jk}^i \omega_j \wedge \omega_k,$$

known as the *Maurer-Cartan equations*.

These equations can be neatly described if we use differential forms valued in $\mathfrak{g}$. Let ω_{MC} be the one-form given by

$$(\omega_{\mathrm{MC}})_g(u) = d(L_g^{-1})_g u, \qquad g \in G,\ u \in T_g G.$$

The same computation that showed that α^L is left-invariant if $\alpha \in \mathfrak{g}$ shows that ω_{MC} is left-invariant and, obviously, $(\omega_{\mathrm{MC}})_1 = \mathrm{id}$.

Definition 8.14. Given any Lie group, G, the *Maurer-Cartan form on G* is the $\mathfrak{g}$-valued differential 1-form, $\omega_{\mathrm{MC}} \in \mathcal{A}^1(G, \mathfrak{g})$, given by

$$(\omega_{\mathrm{MC}})_g = d(L_g^{-1})_g, \qquad g \in G.$$

Recall that for every $g \in G$, conjugation by g is the map given by $a \mapsto gag^{-1}$, that is, $a \mapsto (L_g \circ R_{g^{-1}})a$, and the adjoint map, $\mathrm{Ad}(g)\colon \mathfrak{g} \to \mathfrak{g}$, associated with g is the derivative of $L_g \circ R_{g^{-1}}$ at 1, that is, we have

$$\mathrm{Ad}(g)(u) = d(L_g \circ R_{g^{-1}})_1(u), \qquad u \in \mathfrak{g}.$$

Furthermore, it is obvious that L_g and R_h commute.

Proposition 8.25. *Given any Lie group, G, for all $g \in G$, the Maurer-Cartan form, ω_{MC}, has the following properties:*

(1) $(\omega_{\mathrm{MC}})_1 = \mathrm{id}_{\mathfrak{g}}$.

(2) For all $g \in G$,

$$R_g^* \omega_{\mathrm{MC}} = \mathrm{Ad}(g^{-1}) \circ \omega_{\mathrm{MC}}.$$

(3) The 2-form, $d\omega \in \mathcal{A}^2(G, \mathfrak{g})$, satisfies the Maurer-Cartan equation,

$$d\omega_{\mathrm{MC}} = -\frac{1}{2}[\omega_{\mathrm{MC}}, \omega_{\mathrm{MC}}].$$

Proof. Property (1) is obvious.

(2) For simplicity of notation, if we write $\omega = \omega_{\mathrm{MC}}$, then

$$\begin{aligned}
(R_g^*\omega)_h &= \omega_{hg} \circ d(R_g)_h \\
&= d(L_{hg}^{-1})_{hg} \circ d(R_g)_h \\
&= d(L_{hg}^{-1} \circ R_g)_h \\
&= d((L_h \circ L_g)^{-1} \circ R_g)_h \\
&= d(L_g^{-1} \circ L_h^{-1} \circ R_g)_h \\
&= d(L_g^{-1} \circ R_g \circ L_h^{-1})_h \\
&= d(L_{g^{-1}} \circ R_g)_1 \circ d(L_h^{-1})_h \\
&= \mathrm{Ad}(g^{-1}) \circ \omega_h,
\end{aligned}$$

as claimed.

(3) We can easily express ω_{MC} in terms of a basis of $\mathfrak{g}$. If $X_1, \ldots, X_r$ is a basis of $\mathfrak{g}$ and $\omega_1, \ldots, \omega_r$ is the dual basis, then by Proposition 8.24 (2) and part (1) of Proposition 8.25, we have $\omega_{\mathrm{MC}}(X_i) = (\omega_{\mathrm{MC}})_1(X_i) = X_i$, for $i = 1, \ldots, r$, so ω_{MC} is given by

$$\omega_{\mathrm{MC}} = \omega_1 X_1 + \cdots + \omega_r X_r,$$

under the usual identification of left-invariant vector fields (resp. left-invariant one forms) with elements of $\mathfrak{g}$ (resp. elements of $\mathfrak{g}^*$) and, for simplicity of notation, with the sign $\cdot$ omitted between ω_i and X_i. Using this expression for ω_{MC}, a simple computation shows that the Maurer-Cartan equation is equivalent to

$$d\omega_{\mathrm{MC}} = -\frac{1}{2}[\omega_{\mathrm{MC}}, \omega_{\mathrm{MC}}],$$

as claimed. □

In the case of a matrix group, $G \subseteq \mathrm{GL}(n, \mathbb{R})$, it is easy to see that the Maurer-Cartan form is given explicitly by

$$\omega_{\mathrm{MC}} = g^{-1} dg, \qquad g \in G.$$

Thus, it is a kind of logarithmic derivative of the identity. For $n = 2$, if we let

$$g = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix},$$

we get

$$\omega_{\mathrm{MC}} = \frac{1}{\alpha\delta - \beta\gamma} \begin{pmatrix} \delta d\alpha - \beta d\gamma & \delta d\beta - \beta d\delta \\ -\gamma d\alpha + \alpha d\gamma & -\gamma d\beta + \alpha d\delta \end{pmatrix}.$$

Remarks:

(1) The quantity, $d\omega_{\mathrm{MC}} + \frac{1}{2}[\omega_{\mathrm{MC}}, \omega_{\mathrm{MC}}]$ is the *curvature* of the *connection* ω_{MC} on G. The Maurer-Cartan equation says that the curvature of the Maurer-Cartan connection is zero. We also say that ω_{MC} is a *flat* connection.

(2) As $d\omega_{\mathrm{MC}} = -\frac{1}{2}[\omega_{\mathrm{MC}}, \omega_{\mathrm{MC}}]$, we get

$$d[\omega_{\mathrm{MC}}, \omega_{\mathrm{MC}}] = 0,$$

which yields

$$[[\omega_{\mathrm{MC}}, \omega_{\mathrm{MC}}], \omega_{\mathrm{MC}}] = 0.$$

It is easy to show that the above expresses the Jacobi identity in $\mathfrak{g}$.

(3) As in the case of real-valued one-forms, for every left-invariant one-form, $\omega \in \mathcal{A}^1(G, \mathfrak{g})$, we have

$$\omega_g(u) = \omega_1(d(L_g^{-1})_g u) = \omega_1((\omega_{\mathrm{MC}})_g u),$$

for all $g \in G$ and all $u \in T_g G$ and where $\omega_1 \colon \mathfrak{g} \to \mathfrak{g}$ is a linear map. Consequently, there is a bijection between the set of left-invariant one-forms in $\mathcal{A}^1(G, \mathfrak{g})$ and $\mathrm{Hom}(\mathfrak{g}, \mathfrak{g})$.

(4) The Maurer-Cartan form can be used to define the *Darboux derivative* of a map, $f \colon M \to G$, where M is a manifold and G is a Lie group. The Darboux derivative of f is the $\mathfrak{g}$-valued one-form, $\omega_f \in \mathcal{A}^1(M, \mathfrak{g})$, on M given by

$$\omega_f = f^* \omega_{\mathrm{MC}}.$$

Then, it can be shown that when M is connected, if f_1 and f_2 have the same Darboux derivative, $\omega_{f_1} = \omega_{f_2}$, then $f_2 = L_g \circ f_1$, for some $g \in G$. Elie Cartan also characterized which $\mathfrak{g}$-valued one-forms on M are Darboux derivatives ($d\omega + \frac{1}{2}[\omega, \omega] = 0$ must hold). For more on Darboux derivatives, see Sharpe [140] (Chapter 3) and Malliavin [102] (Chapter III).

8.6 Volume Forms on Riemannian Manifolds and Lie Groups

Recall from Section 7.4 that a smooth manifold, M, is a *Riemannian manifold* iff the vector bundle, TM, has a Euclidean metric. This means that there is a family, $(\langle -, - \rangle_p)_{p \in M}$, of inner products on each tangent space, T_pM, such that $\langle -, - \rangle_p$ depends smoothly on p, which can be expessed by saying that that the maps

$$x \mapsto \langle d\varphi_x^{-1}(e_i), d\varphi_x^{-1}(e_j) \rangle_{\varphi^{-1}(x)}, \qquad x \in \varphi(U),\ 1 \leq i, j \leq n$$

are smooth, for every chart, (U, φ), of M, where $(e_1, \ldots, e_n)$ is the canonical basis of $\mathbb{R}^n$. We let

$$g_{ij}(x) = \langle d\varphi_x^{-1}(e_i), d\varphi_x^{-1}(e_j) \rangle_{\varphi^{-1}(x)}$$

and we say that the $n \times n$ matrix, $(g_{ij}(x))$, is the *local expression of the Riemannian metric on M at x* in the coordinate patch, (U, φ).

For orientability of manifolds, volume forms and related notions, please refer back to Section 3.8. If a Riemannian manifold, M, is orientable, then there is a volume form on M with some special properties.

Proposition 8.26. *Let M be a Riemannian manifold with $\dim(M) = n$. If M is orientable, then there is a uniquely determined volume form, Vol_M, on M with the following properties:*

(1) For every $p \in M$, for every positively oriented orthonormal basis $(b_1, \ldots, b_n)$ of T_pM, we have

$$\mathrm{Vol}_M(b_1, \ldots, b_n) = 1.$$

(2) In every orientation preserving local chart, (U, φ), we have

$$((\varphi^{-1})^* \mathrm{Vol}_M)_q = \sqrt{\det(g_{ij}(q))}\, dx_1 \wedge \cdots \wedge dx_n, \qquad q \in \varphi(U).$$

Proof. (1) Say the orientation of M is given by $\omega \in \mathcal{A}^n(M)$. For any two positively oriented orthonormal bases, $(b_1, \ldots, b_n)$ and $(b'_1, \ldots, b'_n)$, in T_pM, by expressing the second basis over the first, there is an orthogonal matrix, $C = (c_{ij})$, so that

$$b'_i = \sum_{j=1}^{n} c_{ij} b_j.$$

We have

$$\omega_p(b'_1, \ldots, b'_n) = \det(C)\omega_p(b_1, \ldots, b_n),$$

and as these bases are positively oriented, we conclude that $\det(C) = 1$ (as C is orthogonal, $\det(C) = \pm 1$). As a consequence, we have a well-defined function, $\rho \colon M \to \mathbb{R}$, with $\rho(p) > 0$ for all $p \in M$, such that

$$\rho(p) = \omega_p(b_1, \ldots, b_n),$$

for every positively oriented orthonormal basis, $(b_1, \ldots, b_n)$, of T_pM. If we can show that ρ is smooth, then $\mathrm{Vol}_M = \rho^{-1}\omega$ is the required volume form.

Let (U, φ) be a positively oriented chart and consider the vector fields, X_j, on $\varphi(U)$ given by

$$X_j(q) = d\varphi_q^{-1}(e_j), \qquad q \in \varphi(U),\ 1 \leq j \leq n.$$

Then, $(X_1(q), \ldots, X_n(q))$ is a positively oriented basis of $T_{\varphi^{-1}(q)}$. If we apply Gram-Schmidt orthogonalization we get an upper triangular matrix, $A(q) = (a_{ij}(q))$, of smooth functions on $\varphi(U)$ with $a_{ii}(q) > 0$ such that

$$b_i(q) = \sum_{j=1}^{n} a_{ij}(q) X_j(q), \qquad 1 \leq i \leq n,$$

and $(b_1(q), \ldots, b_n(q))$ is a positively oriented orthonormal basis of $T_{\varphi^{-1}(q)}$. We have

$$\begin{aligned} \rho(\varphi^{-1}(q)) &= \omega_{\varphi^{-1}(q)}(b_1(q), \ldots, b_n(q)) \\ &= \det(A(q))\omega_{\varphi^{-1}(q)}(X_1(q), \ldots, X_n(q)) \\ &= \det(A(q))(\varphi^{-1})^*\omega_q(e_1, \ldots, e_n), \end{aligned}$$

which shows that ρ is smooth.

(2) If we repeat the end of the proof with $\omega = \mathrm{Vol}_M$, then $\rho \equiv 1$ on M and the above formula yield

$$((\varphi^{-1})^*\mathrm{Vol}_M)_q = (\det(A(q)))^{-1} dx_1 \wedge \cdots \wedge dx_n.$$

If we compute $\langle b_i(q), b_k(q)\rangle_{\varphi^{-1}(q)}$, we get

$$\delta_{ik} = \langle b_i(q), b_k(q)\rangle_{\varphi^{-1}(q)} = \sum_{j=1}^{n}\sum_{l=1}^{n} a_{ij}(q) g_{jl}(q) a_{kl}(q),$$

and so, $I = A(q)G(q)A(q)^\top$, where $G(q) = (g_{jl}(q))$. Thus, $(\det(A(q)))^2 \det(G(q)) = 1$ and since $\det(A(q)) = \prod_i a_{ii}(q) > 0$, we conclude that

$$(\det(A(q)))^{-1} = \sqrt{\det(g_{ij}(q))},$$

which proves the formula in (2). □

We saw in Section 3.8 that a volume form, ω_0, on the sphere $S^n \subseteq \mathbb{R}^{n+1}$ is given by

$$(\omega_0)_p(u_1, \ldots u_n) = \det(p, u_1, \ldots u_n),$$

where $p \in S^n$ and $u_1, \ldots u_n \in T_pS^n$. To be more precise, we consider the n-form, $\omega_0 \in \mathcal{A}^n(\mathbb{R}^{n+1})$ given by the above formula. As

$$(\omega_0)_p(e_1, \ldots, \widehat{e_i}, \ldots, e_{n+1}) = (-1)^{i-1} p_i,$$

where $p = (p_1, \ldots, p_{n+1})$, we have

$$(\omega_0)_p = \sum_{i=1}^{n+1} (-1)^{i-1} p_i \, dx_1 \wedge \cdots \wedge \widehat{dx_i} \wedge \cdots \wedge dx_{n+1}.$$

Let $i\colon S^n \to \mathbb{R}^{n+1}$ be the inclusion map. For every $p \in S^n$, and every basis, $(u_1, \ldots, u_n)$, of $T_p S^n$, the $(n+1)$-tuple $(p, u_1, \ldots, u_n)$ is a basis of $\mathbb{R}^{n+1}$ and so, $(\omega_0)_p \neq 0$. Hence, $\omega_0 \restriction S^n = i^*\omega_0$ is a volume form on S^n. If we give S^n the Riemannian structure induced by $\mathbb{R}^{n+1}$, then the discussion above shows that

$$\mathrm{Vol}_{S^n} = \omega_0 \restriction S^n.$$

Let $r\colon \mathbb{R}^{n+1} - \{0\} \to S^n$ be the map given by

$$r(x) = \frac{x}{\|x\|}$$

and set

$$\omega = r^* \mathrm{Vol}_{S^n},$$

a closed n-form on $\mathbb{R}^{n+1} - \{0\}$. Clearly,

$$\omega \restriction S^n = \mathrm{Vol}_{S^n}.$$

Furthermore

$$\begin{aligned} \omega_x(u_1, \ldots, u_n) &= (\omega_0)_{r(x)}(dr_x(u_1), \ldots, dr_x(u_n)) \\ &= \|x\|^{-1} \det(x, dr_x(u_1), \ldots, dr_x(u_n)). \end{aligned}$$

We leave it as an exercise to prove that ω is given by

$$\omega_x = \frac{1}{\|x\|^n} \sum_{i=1}^{n+1} (-1)^{i-1} x_i \, dx_1 \wedge \cdots \wedge \widehat{dx_i} \wedge \cdots \wedge dx_{n+1}.$$

We know that there is a map, $\pi\colon S^n \to \mathbb{RP}^n$, such that $\pi^{-1}([p])$ consist of two antipodal points, for every $[p] \in \mathbb{RP}^n$. It can be shown that there is a volume form on $\mathbb{RP}^n$ iff n is even, in which case,

$$\pi^*(\mathrm{Vol}_{\mathbb{RP}^n}) = \mathrm{Vol}_{S^n}.$$

Thus, $\mathbb{RP}^n$ is orientable iff n is even.

Let G be a Lie group of dimension n. For any basis, $(\omega_1, \ldots, \omega_n)$, of the Lie algebra, $\mathfrak{g}$, of G, we have the left-invariant one-forms defined by the ω_i, also denoted ω_i, and obviously, $(\omega_1, \ldots, \omega_n)$ is a frame for TG. Therefore, $\omega = \omega_1 \wedge \cdots \wedge \omega_n$ is an n-form on G that is never zero, that is, a volume form. Since pull-back commutes with $\wedge$, the n-form ω is left-invariant. We summarize this as

Proposition 8.27. *Every Lie group, G, possesses a left-invariant volume form. Therefore, every Lie group is orientable.*

Chapter 9

Integration on Manifolds

9.1 Integration in $\mathbb{R}^n$

As we said in Section 8.1, one of the *raison d'être* for differential forms is that they are the objects that can be integrated on manifolds. We will be integrating differential forms that are at least continuous (in most cases, smooth) and with compact support. In the case of forms, ω, on $\mathbb{R}^n$, this means that the closure of the set, $\{x \in \mathbb{R}^n \mid \omega_x \neq 0\}$, is compact. Similarly, for a form, $\omega \in \mathcal{A}^*(M)$, where M is a manifod, the support, $\mathrm{supp}_M(\omega)$, of ω is the closure of the set, $\{p \in M \mid \omega_p \neq 0\}$. We let $\mathcal{A}^*_c(M)$ denote the set of differential forms with compact support on M. If M is a smooth manifold of dimension n, our ultimate goal is to define a linear operator,

$$\int_M : \mathcal{A}^n_c(M) \longrightarrow \mathbb{R},$$

which generalizes in a natural way the usual integral on $\mathbb{R}^n$.

In this section, we assume that $M = \mathbb{R}^n$, or some open subset of $\mathbb{R}^n$. Now, every n-form (with compact support) on $\mathbb{R}^n$ is given by

$$\omega_x = f(x)\, dx_1 \wedge \cdots \wedge dx_n,$$

where f is a smooth function with compact support. Thus, we set

$$\int_{\mathbb{R}^n} \omega = \int_{\mathbb{R}^n} f(x) dx_1 \cdots dx_n,$$

where the expression on the right-hand side is the usual Riemann integral of f on $\mathbb{R}^n$. Actually, we will need to integrate smooth forms, $\omega \in \mathcal{A}^n_c(U)$, with compact support defined on some open subset, $U \subseteq \mathbb{R}^n$ (with $\mathrm{supp}(\omega) \subseteq U$). However, this is no problem since we still have

$$\omega_x = f(x)\, dx_1 \wedge \cdots \wedge dx_n,$$

where $f\colon U \to \mathbb{R}$ is a smooth function with compact support contained in U and f can be smoothly extended to $\mathbb{R}^n$ by setting it to 0 on $\mathbb{R}^n - \mathrm{supp}(f)$. We write $\int_U \omega$ for this integral.

It is crucial for the generalization of the integral to manifolds to see what the change of variable formula looks like in terms of differential forms.

Proposition 9.1. *Let $\varphi\colon U \to V$ be a diffeomorphism between two open subsets of $\mathbb{R}^n$. If the Jacobian determinant, $J(\varphi)(x)$, has a constant sign, $\delta = \pm 1$ on U, then for every $\omega \in \mathcal{A}^n_c(V)$, we have*

$$\int_U \varphi^*\omega = \delta \int_V \omega.$$

Proof. We know that ω can be written as

$$\omega_x = f(x)\, dx_1 \wedge \cdots \wedge dx_n, \qquad x \in V,$$

where $f\colon V \to \mathbb{R}$ has compact support. From the example before Proposition 8.6, we have

$$\begin{aligned}(\varphi^*\omega)_y &= f(\varphi(y))J(\varphi)_y\, dy_1 \wedge \cdots \wedge dy_n \\ &= \delta f(\varphi(y))|J(\varphi)_y|\, dy_1 \wedge \cdots \wedge dy_n.\end{aligned}$$

On the other hand, the change of variable formula (using φ) is

$$\int_{\varphi(U)} f(x)\, dx_1 \cdots dx_n = \int_U f(\varphi(y))\, |J(\varphi)_y|\, dy_1 \cdots dy_n,$$

so the formula follows. □

We will promote the integral on open subsets of $\mathbb{R}^n$ to manifolds using partitions of unity.

9.2 Integration on Manifolds

Intuitively, for any n-form, $\omega \in \mathcal{A}^n_c(M)$, on a smooth n-dimensional oriented manifold, M, the integral, $\int_M \omega$, is computed by patching together the integrals on small-enough open subsets covering M using a partition of unity. If (U, φ) is a chart such that $\mathrm{supp}(\omega) \subseteq U$, then the form $(\varphi^{-1})^*\omega$ is an n-form on $\mathbb{R}^n$ and the integral, $\int_{\varphi(U)} (\varphi^{-1})^*\omega$, makes sense. The orientability of M is needed to ensure that the above integrals have a consistent value on overlapping charts.

Remark: It is also possible to define integration on non-orientable manifolds using densities but we have no need for this extra generality.

Proposition 9.2. *Let M be a smooth oriented manifold of dimension n. Then, there exists a unique linear operator,*

$$\int_M : \mathcal{A}^n_c(M) \longrightarrow \mathbb{R},$$

with the following property: For any $\omega \in \mathcal{A}^n_c(M)$, if $\mathrm{supp}(\omega) \subseteq U$, where (U, φ) is a positively oriented chart, then

$$\int_M \omega = \int_{\varphi(U)} (\varphi^{-1})^*\omega. \tag{$\dagger$}$$

Proof. First, assume that $\mathrm{supp}(\omega) \subseteq U$, where (U, φ) is a positively oriented chart. Then, we wish to set

$$\int_M \omega = \int_{\varphi(U)} (\varphi^{-1})^* \omega.$$

However, we need to prove that the above expression does not depend on the choice of the chart. Let (V, ψ) be another chart such that $\mathrm{supp}(\omega) \subseteq V$, so that $\mathrm{supp}(\omega) \subseteq U \cap V$. The map, $\theta = \psi \circ \varphi^{-1}$, is a diffeomorphism from $W = \varphi(U \cap V)$ to $W' = \psi(U \cap V)$ and, by hypothesis, its Jacobian determinant is positive on W. Since

$$\mathrm{supp}_{\varphi(U)}((\varphi^{-1})^*\omega) \subseteq W, \qquad \mathrm{supp}_{\psi(V)}((\psi^{-1})^*\omega) \subseteq W',$$

and $\theta^* \circ (\psi^{-1})^*\omega = (\varphi^{-1})^* \circ \psi^* \circ (\psi^{-1})^*\omega = (\varphi^{-1})^*\omega$, Proposition 9.1 yields

$$\int_{W'} (\psi^{-1})^*\omega = \int_W \theta^*((\psi^{-1})^*\omega) = \int_W (\varphi^{-1})^*\omega,$$

as claimed.

In the general case, using Theorem 3.32, for every open cover of M by positively oriented charts, (U_i, φ_i), we have a partition of unity, $(\rho_i)_{i \in I}$, subordinate to this cover. Recall that

$$\mathrm{supp}(\rho_i) \subseteq U_i, \qquad i \in I.$$

Thus, $\rho_i\omega$ is an n-form whose support is a subset of U_i. Furthermore, as $\sum_i \rho_i = 1$,

$$\omega = \sum_i \rho_i \omega.$$

Define

$$I(\omega) = \sum_i \int_{U_i} \rho_i \omega,$$

where each term in the sum is defined by

$$\int_{U_i} \rho_i \omega = \int_{\varphi_i(U_i)} (\varphi_i^{-1})^* \rho_i \omega,$$

where (U_i, φ_i) is the chart associated with $i \in I$. It remains to show that $I(\omega)$ does not depend on the choice of open cover and on the choice of partition of unity. Let (V_j, ψ_j) be another open cover by positively oriented charts and let $(\theta_j)_{j \in J}$ be a partition of unity subordinate to the open cover, (V_j). Note that

$$\int_{U_i} \rho_i \theta_j \omega = \int_{V_j} \rho_i \theta_j \omega,$$

since $\mathrm{supp}(\rho_i\theta_j\omega) \subseteq U_i \cap V_j$, and as $\sum_i \rho_i = 1$ and $\sum_j \theta_j = 1$, we have

$$\sum_i \int_{U_i} \rho_i \omega = \sum_{i,j} \int_{U_i} \rho_i \theta_j \omega = \sum_{i,j} \int_{V_j} \rho_i \theta_j \omega = \sum_j \int_{V_j} \theta_j \omega,$$

proving that $I(\omega)$ is indeed independent of the open cover and of the partition of unity. The uniqueness assertion is easily proved using a partition of unity. □

The integral of Definition 9.2 has the following properties:

Proposition 9.3. *Let M be an oriented manifold of dimension n. The following properties hold:*

(1) If M is connected, then for every n-form, $\omega \in \mathcal{A}_c^n(M)$, the sign of $\int_M \omega$ changes when the orientation of M is reversed.

(2) For every n-form, $\omega \in \mathcal{A}_c^n(M)$, if $\operatorname{supp}(\omega) \subseteq W$, for some open subset, W, of M, then

$$\int_M \omega = \int_W \omega,$$

where W is given the orientation induced by M.

(3) If $\varphi\colon M \to N$ is an orientation-preserving diffeomorphism, then for every $\omega \in \mathcal{A}_c^n(N)$, we have

$$\int_N \omega = \int_M \varphi^*\omega.$$

Proof. Use a partition of unity to reduce to the case where $\operatorname{supp}(\omega)$ is contained in the domain of a chart and then use Proposition 9.1 and (†) from Proposition 9.2. □

The theory or integration developed so far deals with domains that are not general enough. Indeed, for many applications, we need to integrate over domains with boundaries.

9.3 Integration on Regular Domains and Stokes' Theorem

Given a manifold, M, we define a class of subsets with boundaries that can be integrated on, and for which Stokes' Theorem holds. In Warner [148] (Chapter 4), such subsets are called *regular domains* and in Madsen and Tornehave [101] (Chapter 10) they are called *domains with smooth boundary*.

Definition 9.1. Let M be a smooth manifold of dimension n. A subset, $N \subseteq M$, is called a *domain with smooth boundary* (or *codimension zero submanifold with boundary*) iff for every $p \in M$, there is a chart, (U, φ), with $p \in U$, such that

$$\varphi(U \cap N) = \varphi(U) \cap \mathbb{H}^n, \tag{$*$}$$

where $\mathbb{H}^n$ is the closed upper-half space,

$$\mathbb{H}^n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_n \geq 0\}.$$

Note that $(*)$ is automatically satisfied when p is an interior or an exterior point of N, since we can pick a chart such that $\varphi(U)$ is contained in an open half space of $\mathbb{R}^n$ defined by either $x_n > 0$ or $x_n < 0$. If p is a boundary point of N, then $\varphi(p)$ has its last coordinate equal to 0. If M is orientable, then any orientation of M induces an orientation of ∂N, the boundary of N. This follows from the following proposition:

Proposition 9.4. *Let $\varphi\colon \mathbb{H}^n \to \mathbb{H}^n$ be a diffeomorphism with everywhere positive Jacobian determinant. Then, φ induces a diffeomorphism, $\Phi\colon \partial\mathbb{H}^n \to \partial\mathbb{H}^n$, which, viewed as a diffeomorphism of $\mathbb{R}^{n-1}$ also has everywhere positive Jacobian determinant.*

Proof. By the inverse function theorem, every interior point of $\mathbb{H}^n$ is the image of an interior point, so φ maps the boundary to itself. If $\varphi = (\varphi_1, \ldots, \varphi_n)$, then

$$\Phi = (\varphi_1(x_1, \ldots, x_{n-1}, 0), \ldots, \varphi_{n-1}(x_1, \ldots, x_{n-1}, 0)),$$

since $\varphi_n(x_1, \ldots, x_{n-1}, 0) = 0$. It follows that $\frac{\partial \varphi_n}{\partial x_i}(x_1, \ldots, x_{n-1}, 0) = 0$, for $i = 1, \ldots, n-1$, and as φ maps $\mathbb{H}^n$ to itself,

$$\frac{\partial \varphi_n}{\partial x_n}(x_1, \ldots, x_{n-1}, 0) > 0.$$

Now, the Jacobian matrix of φ at $q = \varphi(p) \in \partial\mathbb{H}^n$ is of the form

$$d\varphi_q = \begin{pmatrix} & & & * \\ & d\Phi_q & & \vdots \\ & & & * \\ 0 & \cdots & 0 & \frac{\partial \varphi_n}{\partial x_n}(q) \end{pmatrix}$$

and since $\frac{\partial \varphi_n}{\partial x_n}(q) > 0$ and by hypothesis $\det(d\varphi_q) > 0$, we have $\det(d\Phi_q) > 0$, as claimed. □

In order to make Stokes' formula sign free, if $\mathbb{R}^n$ has the orientation given by $dx_1 \wedge \cdots \wedge dx_n$, then $\mathbb{H}^n$ is given the orientation given by $(-1)^n dx_1 \wedge \cdots \wedge dx_{n-1}$ if $n \geq 2$ and -1 for $n = 1$. This choice of orientation can be explained in terms of the notion of outward directed tangent vector.

Definition 9.2. Given any domain with smooth boundary, $N \subseteq M$, a tangent vector, $w \in T_pM$, at a boundary point, $p \in \partial N$, is *outward directed* iff there is a chart, (U, φ), with $p \in U$ and $\varphi(U \cap N) = \varphi(U) \cap \mathbb{H}^n$ and such that $d\varphi_p(w)$ has a negative n^{th} coordinate $pr_n(d\varphi_p(w))$.

Let (V, ψ) be another chart with $p \in V$. Then, the transition map,

$$\theta = \psi \circ \varphi^{-1}\colon \varphi(U \cap V) \to \psi(U \cap V)$$

induces a map

$$\varphi(U \cap V) \cap \mathbb{H}^n \longrightarrow \psi(U \cap V) \cap \mathbb{H}^n$$

which restricts to a diffeomorphism

$$\Theta\colon \varphi(U\cap V)\cap \partial\mathbb{H}^n \to \psi(U\cap V)\cap \partial\mathbb{H}^n.$$

The proof of Proposition 9.4 shows that the Jacobian matrix of $d\theta_q$ at $q=\varphi(p)\in\partial\mathbb{H}^n$ is of the form

$$d\theta_q = \begin{pmatrix} & & & * \\ & d\Theta_q & & \vdots \\ & & & * \\ 0 & \cdots & 0 & \frac{\partial\theta_n}{\partial x_n}(q) \end{pmatrix}$$

with $\theta=(\theta_1,\ldots,\theta_n)$ and that $\frac{\partial\theta_n}{\partial x_n}(q)>0$. As $d\psi_p = d(\psi\circ\varphi^{-1})_q\circ d\varphi_p$, we see that for any $w\in T_pM$ with $pr_n(d\varphi_p(w))<0$, we also have $pr_n(d\psi_p(w))<0$. Therefore, the negativity condition of Definition 9.2 does not depend on the chart at p. The following proposition is then easy to show:

Proposition 9.5. *Let $N\subseteq M$ be a domain with smooth boundary where M is a smooth manifold of dimension n.*

(1) The boundary, ∂N, of N is a smooth manifold of dimension $n-1$.

(2) Assume M is oriented. If $n\geq 2$, there is an induced orientation on ∂N determined as follows: For every $p\in\partial N$, if $v_1\in T_pM$ is an outward directed tangent vector then a basis, $(v_2,\ldots,v_n)$ for $T_p\partial N$ is positively oriented iff the basis $(v_1,v_2,\ldots,v_n)$ for T_pM is positively oriented. When $n=1$, every $p\in\partial N$ has the orientation $+1$ iff for every outward directed tangent vector, $v_1\in T_pM$, the vector v_1 is a positively oriented basis of T_pM.

If M is oriented, then for every n-form, $\omega\in\mathcal{A}^n_c(M)$, the integral $\int_N\omega$ is well-defined. More precisely, Proposition 9.2 can be generalized to domains with a smooth boundary. This can be shown in various ways. In Warner, this is shown by covering N with special kinds of open subsets arising from regular simplices (see Warner [148], Chapter 4). In Madsen and Tornehave [101], it is argued that integration theory goes through for continuous n-forms with compact support. If σ is a volume form on M, then for every continuous function with compact support, f, the map

$$f\mapsto I_\sigma(f)=\int_M f\sigma$$

is a linear positive operator (which means that $I(f)\geq 0$ for $f\geq 0$). By Riesz' representation theorem, I_σ determines a positive Borel measure, μ_σ, which satisfies

$$\int_M f d\mu_\sigma = \int_M f\sigma.$$

Then, we can set

$$\int_N\omega=\int_M 1_N\omega,$$

where 1_N is the function with value 1 on N and 0 outside N.

Another way to proceed is to prove an extension of Proposition 9.1 using a slight generalization of the change of variable formula:

Proposition 9.6. *Let $\varphi\colon U \to V$ be a diffeomorphism between two open subsets of $\mathbb{R}^n$ and assume that φ maps $U \cap \mathbb{H}^n$ to $V \cap \mathbb{H}^n$. Then, for every smooth function, $f\colon V \to \mathbb{R}$, with compact support,*

$$\int_{V\cap\mathbb{H}^n} f(x)dx_1\cdots dx_n = \int_{U\cap\mathbb{H}^n} f(\varphi(y))\,|J(\varphi)_y|\,dy_1\cdots dy_n.$$

We now have all the ingredient to state Stokes's formula. We omit the proof as it can be found in many places (for example, Warner [148], Chapter 4, Bott and Tu [19], Chapter 1, and Madsen and Tornehave [101], Chapter 10). The proof is fairly easy and it is not particularly illuminating, although one has to be very careful about matters of orientation.

Theorem 9.7. *(Stokes' Theorem) Let $N \subseteq M$ be a domain with smooth boundary where M is a smooth oriented manifold of dimension n, give ∂N the orientation induced by M and let $i\colon \partial N \to M$ be the inclusion map. For every differential form with compact support, $\omega \in \mathcal{A}_c^{n-1}(M)$, we have*

$$\int_{\partial N} i^*\omega = \int_N d\omega.$$

In particular, if $N = M$ is a smooth oriented manifold with boundary, then

$$\int_{\partial M} i^*\omega = \int_M d\omega$$

and if M is a smooth oriented manifold without boundary, then

$$\int_M d\omega = 0.$$

Of course, $i^*\omega$ is the restriction of ω to ∂N and for simplicity of notation, $i^*\omega$ is usually written ω and Stokes' formula is written

$$\int_{\partial N} \omega = \int_N d\omega.$$

9.4 Integration on Riemannian Manifolds and Lie Groups

We saw in Section 8.6 that every orientable Riemannian manifold has a uniquely defined volume form, Vol_M (see Proposition 8.26). Given any smooth function, f, with compact support on M, we define the *integral of f over M* by

$$\int_M f = \int_M f\,\mathrm{Vol}_M.$$

Actually, it is possible to define the integral, $\int_M f$, even if M is not orientable but we do not need this extra generality. If M is compact, then $\int_M 1_M = \int_M \mathrm{Vol}_M$ is the *volume* of M (where 1_M is the constant function with value 1).

If M and N are Riemannian manifolds, then we have the following version of Proposition 9.3 (3):

Proposition 9.8. *If M and N are oriented Riemannian manifolds and if $\varphi\colon M \to N$ is an orientation preserving diffeomorphism, then for every function, $f \in C^\infty(N)$, with compact support, we have*

$$\int_N f \,\mathrm{Vol}_N = \int_M f \circ \varphi \,|\det(d\varphi)|\,\mathrm{Vol}_M,$$

where $f\circ\varphi\,|\det(d\varphi)|$ denotes the function, $p \mapsto f(\varphi(p))|\det(d\varphi_p)|$, with $d\varphi_p\colon T_pM \to T_{\varphi(p)}N$. In particular, if φ is an orientation preserving isometry (see Definition 7.11), then

$$\int_N f \,\mathrm{Vol}_N = \int_M f \circ \varphi \,\mathrm{Vol}_M.$$

We often denote $\int_M f\,\mathrm{Vol}_M$ by $\int_M f(t)dt$.

If G is a Lie group, we know from Section 8.6 that G is always orientable and that G possesses left-invariant volume forms. Since $\dim(\bigwedge^n \mathfrak{g}^*) = 1$ if $\dim(G) = n$ and since every left-invariant volume form is determined by its value at the identity, the space of left-invariant volume forms on G has dimension 1. If we pick some left-invariant volume form, ω, defining the orientation of G, then every other left-invariant volume form is proportional to ω. Given any smooth function, f, with compact support on G, we define the *integral of f over G (w.r.t. ω)* by

$$\int_G f = \int_G f\omega.$$

This integral depends on ω but since ω is defined up to some positive constant, so is the integral. When G is compact, we usually pick ω so that

$$\int_G \omega = 1.$$

For every $g \in G$, as ω is left-invariant, $L_g^*\omega = \omega$, so L_g^* is an orientation-preserving diffeomorphism and by Proposition 9.3 (3),

$$\int_G f\omega = \int_G L_g^*(f\omega),$$

so, using Proposition 8.9, we get

$$\int_G f = \int_G f\omega = \int_G L_g^*(f\omega) = \int_G L_g^* f\, L_g^*\omega = \int_G L_g^* f\,\omega = \int_G (f \circ L_g)\omega = \int_G f \circ L_g.$$

The property

$$\int_G f = \int_G f \circ L_g$$

is called *left-invariance*.

It is then natural to ask when our integral is right-invariant, that is, when

$$\int_G f = \int_G f \circ R_g.$$

Observe that $R_g^*\omega$ is left-invariant, since

$$L_h^* R_g^* \omega = R_g^* L_h^* \omega = R_g^* \omega.$$

It follows that $R_g^*\omega$ is some constant multiple of ω, and so, there is a function, $\overline{\Delta}\colon G \to \mathbb{R}$ such that

$$R_g^*\omega = \overline{\Delta}(g)\omega.$$

One can check that $\overline{\Delta}$ is smooth and we let

$$\Delta(g) = |\overline{\Delta}(g)|.$$

Clearly,

$$\Delta(gh) = \Delta(g)\Delta(h),$$

so Δ is a homomorphism of G into $\mathbb{R}_+$. The function Δ is called the *modular function of* G. Now, by Proposition 9.3 (3), as R_g^* is an orientation-preserving diffeomorphism,

$$\int_G f\omega = \int_G R_g^*(f\omega) = \int_G R_g^* f \, R_g^*\omega = \int_G (f \circ R_g)\Delta(g)\omega$$

or, equivalently,

$$\int_G f\omega = \Delta(g) \int_G (f \circ R_g)\omega.$$

It follows that if ω_l is any left-invariant volume form on G and if ω_r is any right-invariant volume form on G, then

$$\omega_r(g) = c\Delta(g^{-1})\omega_l(g),$$

for some constant $c \neq 0$. Indeed, if let $\omega(g) = \overline{\Delta}(g^{-1})\omega_l(g)$, then by Proposition 8.9 we have

$$\begin{aligned} R_h^*\omega &= \overline{\Delta}((hg)^{-1})R_h^*\omega_l \\ &= \overline{\Delta}(h^{-1})\overline{\Delta}(g^{-1})\overline{\Delta}(h)\omega_l \\ &= \overline{\Delta}(g^{-1})\omega_l, \end{aligned}$$

which shows that ω is right-invariant and thus, $\omega_r(g) = c\Delta(g^{-1})\omega_l(g)$, as claimed (since $\overline{\Delta}(g^{-1}) = \pm\Delta(g^{-1})$). Actually, it is not difficult to prove that

$$\Delta(g) = |\det(\mathrm{Ad}(g^{-1}))|.$$

For this, recall that $\mathrm{Ad}(g) = d(L_g \circ R_{g^{-1}})_1$. For any left-invariant n-form, $\omega \in \bigwedge^n \mathfrak{g}^*$, we claim that

$$(R_g^*\omega)_h = \det(\mathrm{Ad}(g^{-1}))\omega_h,$$

which shows that $\Delta(g) = |\det(\mathrm{Ad}(g^{-1}))|$. Indeed, for all $v_1, \ldots, v_n \in T_hG$, we have

$$\begin{aligned}
&(R_g^*\omega)_h(v_1, \ldots, v_n)\\
&\quad = \omega_{hg}(d(R_g)_h(v_1), \ldots, d(R_g)_h(v_n))\\
&\quad = \omega_{hg}(d(L_g \circ L_{g^{-1}} \circ R_g \circ L_h \circ L_{h^{-1}})_h(v_1), \ldots, d(L_g \circ L_{g^{-1}} \circ R_g \circ L_h \circ L_{h^{-1}})_h(v_n))\\
&\quad = \omega_{hg}(d(L_h \circ L_g \circ L_{g^{-1}} \circ R_g \circ L_{h^{-1}})_h(v_1), \ldots, d(L_h \circ L_g \circ L_{g^{-1}} \circ R_g \circ L_{h^{-1}})_h(v_n))\\
&\quad = \omega_{hg}(d(L_{hg} \circ L_{g^{-1}} \circ R_g \circ L_{h^{-1}})_h(v_1), \ldots, d(L_{hg} \circ L_{g^{-1}} \circ R_g \circ L_{h^{-1}})_h(v_n))\\
&\quad = \omega_{hg}\big(d(L_{hg})_1(\mathrm{Ad}(g^{-1})(d(L_{h^{-1}})_h(v_1))), \ldots, d(L_{hg})_1(\mathrm{Ad}(g^{-1})(d(L_{h^{-1}})_h(v_n)))\big)\\
&\quad = (L_{hg}^*\omega)_1\big(\mathrm{Ad}(g^{-1})(d(L_{h^{-1}})_h(v_1)), \ldots, \mathrm{Ad}(g^{-1})(d(L_{h^{-1}})_h(v_n))\big)\\
&\quad = \omega_1\big(\mathrm{Ad}(g^{-1})(d(L_{h^{-1}})_h(v_1)), \ldots, \mathrm{Ad}(g^{-1})(d(L_{h^{-1}})_h(v_n))\big)\\
&\quad = \det(\mathrm{Ad}(g^{-1}))\,\omega_1\big(d(L_{h^{-1}})_h(v_1), \ldots, d(L_{h^{-1}})_h(v_n)\big)\\
&\quad = \det(\mathrm{Ad}(g^{-1}))\,(L_{h^{-1}}^*\omega)_h(v_1, \ldots, v_n)\\
&\quad = \det(\mathrm{Ad}(g^{-1}))\,\omega_h(v_1, \ldots, v_n),
\end{aligned}$$

where we used the left-invariance of ω twice.

Consequently, our integral is right-invariant iff $\Delta \equiv 1$ on G. Thus, our integral is not always right-invariant. When it is, *i.e.* when $\Delta \equiv 1$ on G, we say that G *is unimodular*. This happens in particular when G is compact, since in this case,

$$1 = \int_G \omega = \int_G 1_G\omega = \int_G \Delta(g)\omega = \Delta(g)\int_G \omega = \Delta(g),$$

for all $g \in G$. Therefore, for a *compact Lie group,* G, our integral is both left and right invariant. We say that our integral is *bi-invariant*.

As a matter of notation, the integral $\int_G f = \int_G f\omega$ is often written $\int_G f(g)dg$. Then, left-invariance can be expressed as

$$\int_G f(g)dg = \int_G f(hg)dg$$

and right-invariance as

$$\int_G f(g)dg = \int_G f(gh)dg,$$

for all $h \in G$. If ω is left-invariant, then it can be shown (see Dieudonné [48], Chapter XIV, Section 3) that

$$\int_G f(g^{-1})\Delta(g^{-1})dg = \int_G f(g)dg.$$

Consequently, if G is unimodular, then

$$\int_G f(g^{-1})dg = \int_G f(g)dg.$$

In general, if G is not unimodular, then $\omega_l \neq \omega_r$. A simple example is the group, G, of affine transformations of the real line, which can be viewed as the group of matrices of the form

$$A = \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}, \qquad a, b, \in \mathbb{R},\ a \neq 0.$$

Then, it it is easy to see that the left-invariant volume form and the right-invariant volume form on G are given by

$$\omega_l = \frac{dadb}{a^2}, \qquad \omega_r = \frac{dadb}{a},$$

and so, $\Delta(A) = |a^{-1}|$.

Remark: By the Riesz' representation theorem, ω defines a positive measure, μ_ω, which satisfies

$$\int_G f d\mu_\omega = \int_G f\omega.$$

Using what we have shown, this measure is left-invariant. Such measures are called *left Haar measures* and similarly, we have *right Haar measures*. It can be shown that every two left Haar measures on a Lie group are proportional (see Knapp, [90], Chapter VIII). Given a left Haar measure, μ, the function, Δ, such that

$$\mu(R_g h) = \Delta(g)\mu(h)$$

for all $g, h \in G$ is the *modular function* of G. However, beware that some authors, including Knapp, use $\Delta(g^{-1})$ instead of $\Delta(g)$. As above, we have

$$\Delta(g) = |\det(\mathrm{Ad}(g^{-1}))|.$$

Beware that authors who use $\Delta(g^{-1})$ instead of $\Delta(g)$, give a formula where $\mathrm{Ad}(g)$ appears instead of $\mathrm{Ad}(g^{-1})$. Again, G is *unimodular* iff $\Delta \equiv 1$. It can be shown that compact, semisimple, reductive and nilpotent Lie groups are unimodular (for instance, see Knapp, [90], Chapter VIII). On such groups, left Haar measures are also right Haar measures (and vice versa). In this case, we can speak of *Haar measures* on G. For more details on Haar measures on locally compact groups and Lie groups, we refer the reader to Folland [55] (Chapter 2), Helgason [73] (Chapter 1) and Dieudonné [48] (Chapter XIV).

Chapter 10

Distributions and the Frobenius Theorem

10.1 Tangential Distributions, Involutive Distributions

Given any smooth manifold, M, (of dimension n) for any smooth vector field, X, on M, we know from Section 3.5 that for every point, $p \in M$, there is a unique maximal integral curve through p. Furthermore, any two distinct integral curves do not intersect each other and the union of all the integral curves is M itself. A nonvanishing vector field, X, can be viewed as the smooth assignment of a one-dimensional vector space to every point of M, namely, $p \mapsto \mathbb{R}X_p \subseteq T_pM$, where $\mathbb{R}X_p$ denotes the line spanned by X_p. Thus, it is natural to consider the more general situation where we fix some integer, r, with $1 \leq r \leq n$ and we have an assignment, $p \mapsto D(p) \subseteq T_pM$, where $D(p)$ is some r-dimensional subspace of T_pM such that $D(p)$ "varies smoothly" with $p \in M$. Is there a notion of integral manifold for such assignments? Do they always exist?

It is indeed possible to generalize the notion of integral curve and to define integral manifolds but, unlike the situation for vector fields ($r = 1$), not every assignment, D, as above, possess an integral manifold. However, there is a necessary and sufficient condition for the existence of integral manifolds given by the Frobenius Theorem. This theorem has several equivalent formulations. First, we will present a formulation in terms of vector fields. Then, we will show that there are advantages in reformulating the notion of involutivity in terms of differential ideals and we will state a differential form version of the Frobenius Theorem. The above versions of the Frobenius Theorem are "local". We will briefly discuss the notion of foliation and state a global version of the Frobenius Theorem.

Since Frobenius' Theorem is a standard result of differential geometry, we will omit most proofs and instead refer the reader to the literature. A complete treatment of Frobenius' Theorem can be found in Warner [148], Morita [115] and Lee [99].

Our first task is to define precisely what we mean by a smooth assignment, $p \mapsto D(p) \subseteq T_pM$, where $D(p)$ is an r-dimensional subspace.

Definition 10.1. Let M be a smooth manifold of dimension n. For any integer r, with $1 \leq r \leq n$, an r-dimensional *tangential distribution* (for short, a *distribution*) is a map, $D\colon M \to TM$, such that

(a) $D(p) \subseteq T_pM$ is an r-dimensional subspace for all $p \in M$.

(b) For every $p \in M$, there is some open subset, U, with $p \in U$, and r smooth vector fields, $X_1, \ldots, X_r$, defined on U, such that $(X_1(q), \ldots, X_r(q))$ is a basis of $D(q)$ for all $q \in U$. We say that D is *locally spanned* by $X_1, \ldots, X_r$.

An immersed submanifold, N, of M is an *integral manifold* of D iff $D(p) = T_pN$, for all $p \in N$. We say that D is *completely integrable* iff there exists an integral manifold of D through every point of M.

We also write D_p for $D(p)$.

Remarks:

(1) An r-dimensional distribution, D, is just a smooth subbundle of TM.

(2) An integral manifold is only an immersed submanifold, not necessarily an embedded submanifold.

(3) Some authors (such as Lee) reserve the locution "completely integrable" to a seemingly strongly condition (See Lee [99], Chapter 19, page 500). This condition is in fact equivalent to "our" definition (which seems the most commonly adopted).

(4) Morita [115] uses a stronger notion of integral manifold, namely, an integral manifold is actually an embedded manifold. Most of the results, including Frobenius Theorem still hold but maximal integral manifolds are immersed but not embedded manifolds and this is why most authors prefer to use the weaker definition (immersed manifolds).

Here is an example of a distribution which does not have any integral manifolds: This is the two-dimensional distribution in $\mathbb{R}^3$ spanned by the vector fields

$$X = \frac{\partial}{\partial x} + y\frac{\partial}{\partial z}, \qquad X = \frac{\partial}{\partial y}.$$

We leave it as an exercise to the reader to show that the above distribution is not integrable.

The key to integrability is an involutivity condition. Here is the definition.

Definition 10.2. Let M be a smooth manifold of dimension n and let D be an r-dimensional distribution on M. For any smooth vector field, X, we say that X *belongs to* D (or *lies in* D) iff $X_p \in D_p$, for all $p \in M$. We say that D is *involutive* iff for any two smooth vector fields, X, Y, on M, if X and Y belong to D, then $[X, Y]$ also belongs to D.

Proposition 10.1. *Let M be a smooth manifold of dimension n. If an r-dimensional distribution, D, is completely integrable, then D is involutive.*

Proof. A proof can be found in in Warner [148] (Chapter 1), and Lee [99] (Proposition 19.3). These proofs use Proposition 3.20. Another proof is given in Morita [115] (Section 2.3) but beware that Morita defines an integral manifold to be an embedded manifold. □

In the example before Definition 10.1, we have

$$[X, Y] = -\frac{\partial}{\partial z},$$

so this distribution is not involutive. Therefore, by Proposition 10.1, this distribution is not completely integrable.

10.2 Frobenius Theorem

Frobenius' Theorem asserts that the converse of Proposition 10.1 holds. Although we do not intend to prove it in full, we would like to explain the main idea of the proof of Frobenius' Theorem. It turns out that the involutivity condition of two vector fields is equivalent to the commutativity of their corresponding flows and this is the crucial fact used in the proof.

Given a manifold, M, we sa that two vector fields, X and Y are *mutually commutative* iff $[X, Y] = 0$. For example, on $\mathbb{R}^2$, the vector fields $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ are commutative but $\frac{\partial}{\partial x}$ and $x\frac{\partial}{\partial y}$ are not.

Recall from Definition 3.27 that we denote by Φ^X the (global) flow of the vector field, X. For every $p \in M$, the map, $t \mapsto \Phi^X(t, p) = \gamma_p(t)$ is the maximal integral curve through p. We also write $\Phi_t(p)$ for $\Phi^X(t, p)$ (dropping X). Recall that the map, $p \mapsto \Phi_t(p)$, is a diffeomorphism on its domain (an open subset of M). For the next proposition, given two vector fields, X and Y, we will write Φ for the flow associated with X and Ψ for the flow associated with Y.

Proposition 10.2. *Given a manifold, M, for any two smooth vector fields, X and Y, the following conditions are equivalent:*

(1) X and Y are mutually commutative (i.e. $[X, Y] = 0$).

(2) Y is invariant under Φ_t, that is, $(\Phi_t)_* Y = Y$, whenever the left-hand side is defined.

(3) X is invariant under Ψ_s, that is, $(\Psi_s)_* X = X$, whenever the left-hand side is defined.

(4) The maps Φ_t and Ψ_t are mutually commutative. This means that

$$\Phi_t \circ \Psi_s = \Psi_s \circ \Phi_t,$$

for all s, t such that both sides are defined.

(5) $\mathcal{L}_X Y = 0$.

(6) $\mathcal{L}_Y X = 0$.

(In (5) $\mathcal{L}_X Y$ *is the Lie derivative and similarly in (6).)*

Proof. A proof can be found in Lee [99] (Chapter 18, Proposition 18.5) and in Morita [115] (Chapter 2, Proposition 2.18). For example, to prove the implication (2) $\Longrightarrow$ (4), we observe that if φ is a diffeomorphism on some open subset, U, of M, then the integral curves of $\varphi_* Y$ through a point $p \in M$ are of the form $\varphi \circ \gamma$, where γ is the integral curve of Y through $\varphi^{-1}(p)$. Consequently, the local one-parameter group generated by $\varphi_* Y$ is $\{\varphi \circ \Psi_s \circ \varphi^{-1}\}$. If we apply this to $\varphi = \Phi_t$, as $(\Phi_t)_* Y = Y$, we get $\Phi_t \circ \Psi_s \circ \Phi_t^{-1} = \Psi_s$ and hence, $\Phi_t \circ \Psi_s = \Psi_s \circ \Phi_t$. □

In order to state our first version of the Frobenius Theorem we make the following definition:

Definition 10.3. Let M be a smooth manifold of dimension n. Given any smooth r-dimensional distribution, D, on M, a chart, (U, φ), is *flat for* D iff

$$\varphi(U) \cong U' \times U'' \subseteq \mathbb{R}^r \times \mathbb{R}^{n-r},$$

where U' and U'' are connected open subsets such that for every $p \in U$, the distribution D is spanned by the vector fields

$$\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_r}.$$

If (U, φ) is flat for D, then it is clear that each slice of (U, φ),

$$S_c = \{q \in U \mid x_{r+1} = c_{r+1}, \ldots, x_n = c_n\},$$

is an integral manifold of D, where $x_i = pr_i \circ \varphi$ is the i^{th}-coordinate function on U and $c = (c_{r+1}, \ldots, c_n) \in \mathbb{R}^{n-r}$ is a fixed vector.

Theorem 10.3. *(Frobenius) Let M be a smooth manifold of dimension n. A smooth r-dimensional distribution, D, on M is completely integrable iff it is involutive. Furthermore, for every $p \in U$, there is flat chart, (U, φ), for D with $p \in U$, so that every slice of (U, φ) is an integral manifold of D.*

Proof. A proof of Theorem 10.3 can be found in Warner [148] (Chapter 1, Theorem 1.60), Lee [99] (Chapter 19, Theorem 19.10) and Morita [115] (Chapter 2, Theorem 2.17). Since we already have Proposition 10.1, it is only necessary to prove that if a distribution is involutive then it is completely integrable. Here is a sketch of the proof, following Morita.

Pick any $p \in M$. As D is a smooth distribution, we can find some chart, (U, φ), with $p \in U$, and some vector fields, $Y_1, \ldots, Y_r$, so that $Y_1(q), \ldots, Y_r(q)$ are linearly independent and span D_q for all $q \in U$. Locally, we can write

$$Y_i = \sum_{j=1}^{n} a_{ij} \frac{\partial}{\partial x_j}, \qquad i = 1, \ldots, r.$$

Since the Y_i are linearly independent, by renumbering the coordinates if necessary, we may assume that the $r \times r$ matrices

$$A(q) = (a_{ij}(q)) \qquad q \in U$$

are invertible. Then, the inverse matrices, $B(q) = A^{-1}(q)$ define $r \times r$ functions, $b_{ij}(q)$ and let

$$X_i = \sum_{j=1}^{r} b_{ij} Y_j, \qquad j = 1, \ldots, r.$$

Now, in matrix form,

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_r \end{pmatrix} = \begin{pmatrix} A & R \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial x_1}, \\ \vdots \\ \frac{\partial}{\partial x_n}, \end{pmatrix}$$

for some $r \times (n-r)$ matrix, R and

$$\begin{pmatrix} X_1 \\ \vdots \\ X_r \end{pmatrix} = B \begin{pmatrix} Y_1 \\ \vdots \\ Y_r \end{pmatrix},$$

so we get

$$\begin{pmatrix} X_1 \\ \vdots \\ X_r \end{pmatrix} = \begin{pmatrix} I & BR \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix},$$

that is,

$$X_i = \frac{\partial}{\partial x_i} + \sum_{j=r+1}^{n} c_{ij} \frac{\partial}{\partial x_j}, \qquad i = 1, \ldots, r, \tag{$*$}$$

where the c_{ij} are functions defined on U. Obviously, $X_1, \ldots, X_r$ are linearly independent and they span D_q for all $q \in U$. Since D is involutive, there are some functions, f_k, defined on U, so that

$$[X_i, X_j] = \sum_{k=1}^{r} f_k X_k.$$

On the other hand, by $(*)$, each $[X_i, X_j]$ is a linear combination of $\frac{\partial}{\partial x_{r+1}}, \ldots, \frac{\partial}{\partial x_n}$. Therefore, $f_k = 0$, for $k = 1, \ldots, r$, which shows that

$$[X_i, X_j] = 0, \qquad 1 \leq i, j \leq r,$$

that is, the vector fields $X_1, \ldots, X_r$ are mutually commutative.

Let Φ_t^i be the local one-parameter group associated with X_i. By Proposition 10.2 (4), the Φ_t^i commute, that is,

$$\Phi_t^i \circ \Phi_s^j = \Phi_s^j \circ \Phi_t^i \qquad 1 \leq i, j \leq r,$$

whenever both sides are defined. We can pick a sufficiently open subset, V, in $\mathbb{R}^r$ containing the origin and define the map, $\Phi\colon V \to U$ by

$$\Phi(t_1, \ldots, t_r) = \Phi^1_{t_1} \circ \cdots \circ \Phi^r_{t_r}(p).$$

Clearly, Φ is smooth and using the fact that each X_i is invariant under each Φ^j_s, for $j \neq i$, and

$$d\Phi^i_p\left(\frac{\partial}{\partial t_i}\right) = X_i(p),$$

we get

$$d\Phi_p\left(\frac{\partial}{\partial t_i}\right) = X_i(p).$$

As $X_1, \ldots, X_r$ are linearly independent, we deduce that $d\Phi_p\colon T_0\mathbb{R}^r \to T_pM$ is an injection and thus, we may assume by shrinking V if necessary that our map, $\Phi\colon V \to M$, is an embedding. But then, $N = \Phi(V)$ is a submanifold of M and it only remains to prove that N is an integral manifold of D through p.

Obviously, $T_pN = D_p$, so we just have to prove that $T_qN = D_qN$ for all $q \in N$. Now, for every $q \in N$, we can write

$$q = \Phi(t_1, \ldots, t_r) = \Phi^1_{t_1} \circ \cdots \circ \Phi^r_{t_r}(p),$$

for some $(t_1, \ldots, t_r) \in V$. Since the Φ^i_t commute, for any i, with $1 \leq i \leq r$, we can write

$$q = \Phi^i_{t_i} \circ \Phi^1_{t_1} \circ \cdots \circ \Phi^{i-1}_{t_{i-1}} \circ \Phi^{i+1}_{t_{i+1}} \circ \cdots \circ \Phi^r_{t_r}(p).$$

If we fix all the t_j but t_i and vary t_i by a small amount, we obtain a curve in N through q and this is an orbit of Φ^i_t. Therefore, this curve is an integral curve of X_i through q whose velocity vector at q is equal to $X_i(q)$ and so, $X_i(q) \in T_qN$. Since the above reasoning holds for all i, we get $T_qN = D_q$, as claimed. Therefore, N is an integral manifold of D through p. □

In preparation for a global version of Frobenius Theorem in terms of foliations, we state the following Proposition proved in Lee [99] (Chapter 19, Proposition 19.12):

Proposition 10.4. *Let M be a smooth manifold of dimension n and let D be an involutive r-dimensional distribution on M. For every flat chart, (U, φ), for D, for every integral manifold, N, of D, the set $N \cap U$ is a countable disjoint union of open parallel k-dimensional slices of U, each of which is open in N and embedded in M.*

We now describe an alternative method for describing involutivity in terms of differential forms.

10.3 Differential Ideals and Frobenius Theorem

First, we give a smoothness criterion for distributions in terms of one-forms.

Proposition 10.5. *Let M be a smooth manifold of dimension n and let D be an assignment, $p \mapsto D_p \subseteq T_pM$, of some r-dimensional subspace of T_pM, for all $p \in M$. Then, D is a smooth distribution iff for every $p \in U$, there is some open subset, U, with $p \in U$, and some linearly independent one-forms, $\omega_1, \ldots, \omega_{n-r}$, defined on U, so that*

$$D_q = \{u \in T_qM \mid (\omega_1)_q(u) = \cdots = (\omega_{n-r})_q(u) = 0\}, \qquad \textit{for all } q \in U.$$

Proof. Proposition 10.5 is proved in Lee [99] (Chapter 19, Lemma 19.5). The idea is to either extend a set of linearly independent differential one-forms to a coframe and then consider the dual frame or to extend some linearly independent vector fields to a frame and then take the dual basis. □

Proposition 10.5 suggests the following definition:

Definition 10.4. Let M be a smooth manifold of dimension n and let D be an r-dimensional distibution on M. Some linearly independent one-forms, $\omega_1, \ldots, \omega_{n-r}$, defined some open subset, $U \subseteq M$, are called *local defining one-forms* for D if

$$D_q = \{u \in T_qM \mid (\omega_1)_q(u) = \cdots = (\omega_{n-r})_q(u) = 0\}, \qquad \text{for all } q \in U.$$

We say that a k-form, $\omega \in \mathcal{A}^k(M)$, *annihilates* D iff

$$\omega_q(X_1(q), \ldots, X_r(q)) = 0,$$

for all $q \in M$ and for all vector fields, $X_1, \ldots, X_r$, belonging to D. We write

$$\mathfrak{I}^k(D) = \{\omega \in \mathcal{A}^k(M) \mid \omega_q(X_1(q), \ldots, X_r(q)) = 0\},$$

for all $q \in M$ and for all vector fields, $X_1, \ldots, X_r$, belonging to D and we let

$$\mathfrak{I}(D) = \bigoplus_{k=1}^{n} \mathfrak{I}^k(D).$$

Thus, $\mathfrak{I}(D)$ is the collection of differential forms that "vanish on D." In the classical terminology, a system of local defining one-forms as above is called a *system of Pfaffian equations*.

It turns out that $\mathfrak{I}(D)$ is not only a vector space but also an ideal of $\mathcal{A}^\bullet(M)$.

A subspace, $\mathfrak{I}$, of $\mathcal{A}^\bullet(M)$ is an *ideal* iff for every $\omega \in \mathfrak{I}$, we have $\theta \wedge \omega \in \mathfrak{I}$ for every $\theta \in \mathcal{A}^\bullet(M)$.

Proposition 10.6. *Let M be a smooth n-dimensional manifold and D be an r-dimensional distribution. If $\mathfrak{I}(D)$ is the space of forms annihilating D then the following hold:*

(a) $\mathfrak{I}(D)$ is an ideal in $\mathcal{A}^\bullet(M)$.

(b) $\mathfrak{I}(D)$ is locally generated by $n-r$ linearly independent one-forms, which means: For every $p \in U$, there is some open subset, $U \subseteq M$, with $p \in U$ and a set of linearly independent one-forms, $\omega_1, \ldots, \omega_{n-r}$, defined on U, so that

(i) If $\omega \in \mathfrak{I}^k(D)$, then $\omega \restriction U$ belongs to the ideal in $\mathcal{A}^\bullet(U)$ generated by $\omega_1, \ldots, \omega_{n-r}$, that is,

$$\omega = \sum_{i=1}^{n-r} \theta_i \wedge \omega_i, \qquad \text{on } U,$$

for some $(k-1)$-forms, $\theta_i \in \mathcal{A}^{k-1}(U)$.

(ii) If $\omega \in \mathcal{A}^k(M)$ and if there is an open cover by subsets U (as above) such that for every U in the cover, $\omega \restriction U$ belongs to the ideal generated by $\omega_1, \ldots, \omega_{n-r}$, then $\omega \in \mathfrak{I}(D)$.

(c) If $\mathfrak{I} \subseteq \mathcal{A}^\bullet(M)$ is an ideal locally generated by $n-r$ linearly independent one-forms, then there exists a unique smooth r-dimensional distribution, D, for which $\mathfrak{I} = \mathfrak{I}(D)$.

Proof. Proposition 10.6 is proved in Warner (Chapter 2, Proposition 2.28). See also Morita [115] (Chapter 2, Lemma 2.19) and Lee [99] (Chapter 19, page 498-500). □

In order to characterize involutive distributions, we need the notion of differential ideal.

Definition 10.5. Let M be a smooth manifold of dimension n. An ideal, $\mathfrak{I} \subseteq \mathcal{A}^\bullet(M)$, is a *differential ideal* iff it is closed under exterior differentiation, that is

$$d\omega \in \mathfrak{I} \quad \text{whenever} \quad \omega \in \mathfrak{I},$$

which we also express by $d\mathfrak{I} \subseteq \mathfrak{I}$.

Here is the differential ideal criterion for involutivity.

Proposition 10.7. *Let M be a smooth manifold of dimension n. A smooth r-dimensional distribution, D, is involutive iff the ideal, $\mathfrak{I}(D)$, is a differential ideal.*

Proof. Proposition 10.7 is proved in Warner [148] (Chapter 2, Proposition 2.30), Morita [115] (Chapter 2, Proposition 2.20) and Lee [99] (Chapter 19, Proposition 19.19). Here is one direction of the proof. Assume $\mathfrak{I}(D)$ is a differential ideal. We know that for any one-form, ω,

$$d\omega(X, Y) = X(\omega(Y)) - Y(\omega(X)) - \omega([X, Y]),$$

for any vector fields, X, Y. Now, if $\omega_1, \ldots, \omega_{n-r}$ are linearly independent one-forms that define D locally on U, using a bump function, we can extend $\omega_1, \ldots, \omega_{n-r}$ to M and then using the above equation, for any vector fields X, Y belonging to D, we get

$$\omega_i([X,Y]) = X(\omega_i(Y)) - Y(\omega_i(X)) - d\omega_i(X,Y) = 0,$$

and since $\omega_i(X) = \omega_i(Y) = d\omega_i(X,Y) = 0$, we get $\omega_i([X,Y]) = 0$ for $i = 1, \ldots, n-r$, which means that $[X,Y]$ belongs to D. □

Using Proposition 10.6, we can give a more concrete criterion: D is involutive iff for every local defining one-forms, $\omega_1, \ldots, \omega_{n-r}$, for D (on some open subset, U), there are some one-forms, $\omega_{ij} \in \mathcal{A}^1(U)$, so that

$$d\omega_i = \sum_{j=1}^{n-r} \omega_{ij} \wedge \omega_j \qquad (i = 1, \ldots, n-r).$$

The above conditions are often called the *integrability conditions*.

Definition 10.6. Let M be a smooth manifold of dimension n. Given any ideal $\mathfrak{I} \subseteq \mathcal{A}^\bullet(M)$, an immersed manifold, (M, ψ), of M is an *integral manifold of* $\mathfrak{I}$ iff

$$\psi^*\omega = 0, \qquad \text{for all } \omega \in \mathfrak{I}.$$

A connected integral manifold of the ideal $\mathfrak{I}$ is *maximal* iff its image is not a proper subset of the image of any other connected integral manifold of $\mathfrak{I}$.

Finally, here is the differential form version of the Frobenius Theorem.

Theorem 10.8. *(Frobenius Theorem, Differential Ideal Version) Let M be a smooth manifold of dimension n. If $\mathfrak{I} \subseteq \mathcal{A}^\bullet(M)$ is a differential ideal locally generated by $n-r$ linearly independent one-forms, then for every $p \in M$, there exists a unique maximal, connected, integral manifold of $\mathfrak{I}$ through p and this integral manifold has dimension r.*

Proof. Theorem 10.8 is proved in Warner [148]. This theorem follows immediately from Theorem 1.64 in Warner [148]. □

Another version of the Frobenius Theorem goes as follows:

Theorem 10.9. *(Frobenius Theorem, Integrability Conditions Version) Let M be a smooth manifold of dimension n. An r-dimensional distribution, D, on M is completely integrable iff for every local defining one-forms, $\omega_1, \ldots, \omega_{n-r}$, for D (on some open subset, U), there are some one-forms, $\omega_{ij} \in \mathcal{A}^1(U)$, so that we have the integrability conditions*

$$d\omega_i = \sum_{j=1}^{n-r} \omega_{ij} \wedge \omega_j \qquad (i = 1, \ldots, n-r).$$

There are applications of Frobenius Theorem (in its various forms) to systems of partial differential equations but we will not deal with this subject. The reader is advised to consult Lee [99], Chapter 19, and the references there.

10.4 A Glimpse at Foliations and a Global Version of Frobenius Theorem

All the maximal integral manifolds of an r-dimensional involutive distribution on a manifold, M, yield a decomposition of M with some nice properties, those of a foliation.

Definition 10.7. Let M be a smooth manifold of dimension n. A family, $\mathcal{F} = \{\mathcal{F}_\alpha\}_\alpha$, of subsets of M is a *k-dimensional foliation* iff it is a family of pairwise disjoint, connected, immersed k-dimensional submanifolds of M, called the *leaves* of the foliation, whose union is M and such that, for every $p \in M$, there is a chart, (U, φ), with $p \in U$, called a *flat chart* for the foliation and the following property holds:

$$\varphi(U) \cong U' \times U'' \subseteq \mathbb{R}^r \times \mathbb{R}^{n-r},$$

where U' and U'' are some connected open subsets and for every leaf, $\mathcal{F}_\alpha$, of the foliation, if $\mathcal{F}_\alpha \cap U \neq \emptyset$, then $\mathcal{F}_\alpha \cap U$ is a countable union of k-dimensional slices given by

$$x_{r+1} = c_{r+1}, \ldots, x_n = c_n,$$

for some constants, $c_{r+1}, \ldots, c_n \in \mathbb{R}$.

The structure of a foliation can be very complicated. For instance, the leaves can be dense in M. For example, there are spirals on a torus that form the leaves of a foliation (see Lee [99], Example 19.9). Foliations are in one-to-one correspondence with involutive distributions.

Proposition 10.10. *Let M be a smooth manifold of dimension n. For any foliation, $\mathcal{F}$, on M, the family of tangent spaces to the leaves of $\mathcal{F}$ forms an involutive distribution on M.*

The converse to the above proposition may be viewed as a global version of Frobenius Theorem.

Theorem 10.11. *Let M be a smooth manifold of dimension n. For every r-dimensional smooth, involutive distribution, D, on M, the family of all maximal, connected, integral manifolds of D forms a foliation of M.*

Proof. The proof of Theorem 10.11 can be found in Lee [99] (Theorem 19.21). □

Chapter 11

Connections and Curvature in Vector Bundles

11.1 Connections and Connection Forms in Vector Bundles and Riemannian Manifolds

Given a manifold, M, in general, for any two points, $p, q \in M$, there is no "natural" isomorphism between the tangent spaces T_pM and T_qM. More generally, given any vector bundle, $\xi = (E, \pi, B, V)$, for any two points, $p, q \in B$, there is no "natural" isomorphism between the fibres, $E_p = \pi^{-1}(p)$ and $E_q = \pi^{-1}(q)$. Given a curve, $c\colon [0,1] \to M$, on M (resp. a curve, $c\colon [0,1] \to E$, on B), as $c(t)$ moves on M (resp. on B), how does the tangent space, $T_{c(t)}M$ (resp. the fibre $E_{c(t)} = \pi^{-1}(c(t))$) change as $c(t)$ moves?

If $M = \mathbb{R}^n$, then the spaces $T_{c(t)}\mathbb{R}^n$ are canonically isomorphic to $\mathbb{R}^n$ and any vector, $v \in T_{c(0)}\mathbb{R}^n \cong \mathbb{R}^n$, is simply moved along c by *parallel transport*, that it, at $c(t)$, the tangent vector, v, also belongs to $T_{c(t)}\mathbb{R}^n$. However, if M is curved, for example, a sphere, then it is not obvious how to "parallel transport" a tangent vector at $c(0)$ along a curve c. A way to achieve this is to define the notion of *parallel vector field* along a curve and this, in turn, can be defined in terms of the notion of *covariant derivative* of a vector field (or covariant derivative of a section, in the case of vector bundles).

Assume for simplicity that M is a surface in $\mathbb{R}^3$. Given any two vector fields, X and Y defined on some open subset, $U \subseteq \mathbb{R}^3$, for every $p \in U$, the *directional derivative, $D_XY(p)$, of Y with respect to X* is defined by

$$D_XY(p) = \lim_{t \to 0} \frac{Y(p + tX(p)) - Y(p)}{t}.$$

If $f\colon U \to \mathbb{R}$ is a differentiable function on U, for every $p \in U$, the *directional derivative, $X[f](p)$ (or $X(f)(p)$), of f with respect to X* is defined by

$$X[f](p) = \lim_{t \to 0} \frac{f(p + tX(p)) - f(p)}{t}.$$

We know that $X[f](p) = df_p(X(p))$.

It is easily shown that $D_X Y(p)$ is $\mathbb{R}$-bilinear in X and Y, is $C^\infty(U)$-linear in X and satisfies the Leibnitz derivation rule with respect to Y, that is:

Proposition 11.1. *The directional derivative of vector fields satisfies the following properties:*

$$\begin{aligned} D_{X_1+X_2}Y(p) &= D_{X_1}Y(p) + D_{X_2}Y(p) \\ D_{fX}Y(p) &= fD_X Y(p) \\ D_X(Y_1+Y_2)(p) &= D_X Y_1(p) + D_X Y_2(p) \\ D_X(fY)(p) &= X[f](p)Y(p) + f(p)D_X Y(p), \end{aligned}$$

for all $X, X_1, X_2, Y, Y_1, Y_2 \in \mathfrak{X}(U)$ *and all* $f \in C^\infty(U)$.

Now, if $p \in U$ where $U \subseteq M$ is an open subset of M, for any vector field, Y, defined on U ($Y(p) \in T_pM$, for all $p \in U$), for every $X \in T_pM$, the directional derivative, $D_X Y(p)$, makes sense and it has an orthogonal decomposition,

$$D_X Y(p) = \nabla_X Y(p) + (D_n)_X Y(p),$$

where its *horizontal (or tangential) component* is $\nabla_X Y(p) \in T_pM$ and its normal component is $(D_n)_X Y(p)$. The component, $\nabla_X Y(p)$, is the *covariant derivative* of Y with respect to $X \in T_pM$ and it allows us to define the covariant derivative of a vector field, $Y \in \mathfrak{X}(U)$, with respect to a vector field, $X \in \mathfrak{X}(M)$, on M. We easily check that $\nabla_X Y$ satisfies the four equations of Proposition 11.1.

In particular, Y, may be a vector field associated with a curve, $c\colon [0,1] \to M$. A *vector field along a curve, c,* is a vector field, Y, such that $Y(c(t)) \in T_{c(t)}M$, for all $t \in [0,1]$. We also write $Y(t)$ for $Y(c(t))$. Then, we say that Y *is parallel along* c iff $\nabla_{\partial/\partial t} Y = 0$ along c.

The notion of *parallel transport* on a surface can be defined using parallel vector fields along curves. Let p, q be any two points on the surface M and assume there is a curve, $c\colon [0,1] \to M$, joining $p = c(0)$ to $q = c(1)$. Then, using the uniqueness and existence theorem for ordinary differential equations, it can be shown that for any initial tangent vector, $Y_0 \in T_pM$, there is a unique parallel vector field, Y, along c, with $Y(0) = Y_0$. If we set $Y_1 = Y(1)$, we obtain a linear map, $Y_0 \mapsto Y_1$, from T_pM to T_qM which is also an isometry.

As a summary, given a surface, M, if we can define a notion of covariant derivative, $\nabla\colon \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$, satisfying the properties of Proposition 11.1, then we can define the notion of parallel vector field along a curve and the notion of parallel transport, which yields a natural way of relating two tangent spaces, T_pM and T_qM, using curves joining p and q. This can be generalized to manifolds and even to vector bundles using the notion of connection. We will see that the notion of connection induces the notion of curvature.

Moreover, if M has a Riemannian metric, we will see that this metric induces a unique connection with two extra properties (the *Levi-Civita* connection).

Given a manifold, M, as $\mathfrak{X}(M) = \Gamma(M, TM) = \Gamma(TM)$, the set of smooth sections of the tangent bundle, TM, it is natural that for a vector bundle, $\xi = (E, \pi, B, V)$, a connection on ξ should be some kind of bilinear map,

$$\mathfrak{X}(B) \times \Gamma(\xi) \longrightarrow \Gamma(\xi),$$

that tells us how to take the covariant derivative of sections.

Technically, it turns out that it is cleaner to define a connection on a vector bundle, ξ, as an $\mathbb{R}$-linear map,

$$\nabla \colon \Gamma(\xi) \to \mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi), \tag{$*$}$$

that satisfies the "Leibnitz rule"

$$\nabla(fs) = df \otimes s + f\nabla s,$$

with $s \in \Gamma(\xi)$ and $f \in C^\infty(B)$, where $\Gamma(\xi)$ and $\mathcal{A}^1(B)$ are treated as $C^\infty(B)$-modules. Since $\mathcal{A}^1(B) = \Gamma(B, T^*B) = \Gamma(T^*B)$ and, by Proposition 7.12,

$$\begin{aligned}
\mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi) &= \Gamma(T^*B) \otimes_{C^\infty(B)} \Gamma(\xi) \\
&\cong \Gamma(T^*B \otimes \xi) \\
&\cong \Gamma(\mathcal{H}om(TB, \xi)) \\
&\cong \mathrm{Hom}_{C^\infty(B)}(\Gamma(TB), \Gamma(\xi)) \\
&= \mathrm{Hom}_{C^\infty(B)}(\mathfrak{X}(B), \Gamma(\xi)),
\end{aligned}$$

the range of ∇ can be viewed as a space of $\Gamma(\xi)$-valued differential forms on B. Milnor and Stasheff [111] (Appendix C) use the version where

$$\nabla \colon \Gamma(\xi) \to \Gamma(T^*B \otimes \xi)$$

and Madsen and Tornehave [101] (Chapter 17) use the equivalent version stated in $(*)$. A thorough presentation of connections on vector bundles and the various ways to define them can be found in Postnikov [126] which also constitutes one of the most extensive references on differential geometry. Set

$$\mathcal{A}^1(\xi) = \mathcal{A}^1(B; \xi) = \mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi)$$

and, more generally, for any $i \geq 0$, set

$$\mathcal{A}^i(\xi) = \mathcal{A}^i(B; \xi) = \mathcal{A}^i(B) \otimes_{C^\infty(B)} \Gamma(\xi) \cong \Gamma\Big(\Big(\bigwedge^i T^*B\Big) \otimes \xi\Big).$$

Obviously, $\mathcal{A}^0(\xi) = \Gamma(\xi)$ (and recall that $\mathcal{A}^0(B) = C^\infty(B)$). The space of differential forms, $\mathcal{A}^i(B; \xi)$, with values in $\Gamma(\xi)$ is a generalization of the space, $\mathcal{A}^i(M, F)$, of differential forms with values in F encountered in Section 8.4.

If we use the isomorphism

$$\mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi) \cong \mathrm{Hom}_{C^\infty(B)}(\mathfrak{X}(B), \Gamma(\xi)),$$

then a connection is an $\mathbb{R}$-linear map,

$$\nabla \colon \Gamma(\xi) \longrightarrow \mathrm{Hom}_{C^\infty(B)}(\mathfrak{X}(B), \Gamma(\xi)),$$

satisfying a Leibnitz-type rule or equivalently, an $\mathbb{R}$-bilinear map,

$$\nabla \colon \mathfrak{X}(B) \times \Gamma(\xi) \longrightarrow \Gamma(\xi),$$

such that, for any $X \in \mathfrak{X}(B)$ and $s \in \Gamma(\xi)$, if we write $\nabla_X s$ instead of $\nabla(X, s)$, then the following properties hold for all $f \in C^\infty(B)$:

$$\begin{aligned} \nabla_{fX} s &= f \nabla_X s \\ \nabla_X(fs) &= X[f]s + f\nabla_X s. \end{aligned}$$

This second version may be considered simpler than the first since it does not involve a tensor product. Since

$$\mathcal{A}^1(B) = \Gamma(T^*B) \cong \mathrm{Hom}_{C^\infty(B)}(\mathfrak{X}(B), C^\infty(B)) = \mathfrak{X}(B)^*,$$

using Proposition 22.36, the isomorphism

$$\alpha \colon \mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi) \cong \mathrm{Hom}_{C^\infty(B)}(\mathfrak{X}(B), \Gamma(\xi))$$

can be described in terms of the evaluation map,

$$\mathrm{Ev}_X \colon \mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi) \to \Gamma(\xi),$$

given by

$$\mathrm{Ev}_X(\omega \otimes s) = \omega(X)s, \qquad X \in \mathfrak{X}(B),\ \omega \in \mathcal{A}^1(B),\ s \in \Gamma(\xi).$$

Namely, for any $\theta \in \mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi)$,

$$\alpha(\theta)(X) = \mathrm{Ev}_X(\theta).$$

In particular, the reader should check that

$$\mathrm{Ev}_X(df \otimes s) = X[f]s.$$

Then, it is easy to see that we pass from the first version of ∇, where

$$\nabla \colon \Gamma(\xi) \to \mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi) \tag{$*$}$$

with the Leibnitz rule

$$\nabla(fs) = df \otimes s + f\nabla s,$$

to the second version of ∇, denoted ∇', where

$$\nabla' \colon \mathfrak{X}(B) \times \Gamma(\xi) \to \Gamma(\xi), \tag{$**$}$$

is $\mathbb{R}$-bilinear and where the two conditions

$$\begin{aligned} \nabla'_{fX} s &= f \nabla'_X s \\ \nabla'_X (fs) &= X[f]s + f \nabla'_X s \end{aligned}$$

hold, *via* the equation

$$\nabla'_X = \mathrm{Ev}_X \circ \nabla.$$

From now on, we will simply write $\nabla_X s$ instead of $\nabla'_X s$, unless confusion arise. As summary of the above discussion, we make the following definition:

Definition 11.1. Let $\xi = (E, \pi, B, V)$ be a smooth real vector bundle. A *connection* on ξ is an $\mathbb{R}$-linear map,

$$\nabla \colon \Gamma(\xi) \to \mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi) \tag{$*$}$$

such that the *Leibnitz rule*

$$\nabla(fs) = df \otimes s + f \nabla s$$

holds, for all $s \in \Gamma(\xi)$ and all $f \in C^\infty(B)$. For every $X \in \mathfrak{X}(B)$, we let

$$\nabla_X = \mathrm{Ev}_X \circ \nabla$$

and for every $s \in \Gamma(\xi)$, we call $\nabla_X s$ the *covariant derivative of s relative to X*. Then, the family, (∇_X), induces a $\mathbb{R}$-bilinear map also denoted ∇,

$$\nabla \colon \mathfrak{X}(B) \times \Gamma(\xi) \to \Gamma(\xi), \tag{$**$}$$

such that the following two conditions hold:

$$\begin{aligned} \nabla_{fX} s &= f \nabla_X s \\ \nabla_X (fs) &= X[f]s + f \nabla_X s, \end{aligned}$$

for all $s \in \Gamma(\xi)$, all $X \in \mathfrak{X}(B)$ and all $f \in C^\infty(B)$. We refer to $(*)$ as the *first version* of a connection and to $(**)$ as the *second version* of a connection.

Observe that in terms of the $\mathcal{A}^i(\xi)$'s, a connection is a linear map,

$$\nabla \colon \mathcal{A}^0(\xi) \to \mathcal{A}^1(\xi),$$

satisfying the Leibnitz rule. When $\xi = TB$, a connection (second version) is what is known as an *affine connection* on a manifold, B.

Remark: Given two connections, ∇^1 and ∇^2, we have

$$\nabla^1(fs) - \nabla^2(fs) = df \otimes s + f\nabla^1 s - df \otimes s - f\nabla^2 s = f(\nabla^1 s - \nabla^2 s),$$

which shows that $\nabla^1 - \nabla^2$ is a $C^\infty(B)$-linear map from $\Gamma(\xi)$ to $\mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi)$. However

$$\begin{aligned}
\mathrm{Hom}_{C^\infty(B)}(\mathcal{A}^0(\xi), \mathcal{A}^i(\xi)) &= \mathrm{Hom}_{C^\infty(B)}(\Gamma(\xi), \mathcal{A}^i(B) \otimes_{C^\infty(B)} \Gamma(\xi)) \\
&\cong \Gamma(\xi)^* \otimes_{C^\infty(B)} (\mathcal{A}^i(B) \otimes_{C^\infty(B)} \Gamma(\xi)) \\
&\cong \mathcal{A}^i(B) \otimes_{C^\infty(B)} (\Gamma(\xi)^* \otimes_{C^\infty(B)} \Gamma(\xi)) \\
&\cong \mathcal{A}^i(B) \otimes_{C^\infty(B)} \mathrm{Hom}_{C^\infty(B)}(\Gamma(\xi), \Gamma(\xi)) \\
&\cong \mathcal{A}^i(B) \otimes_{C^\infty(B)} \Gamma(\mathcal{H}om(\xi, \xi)) \\
&= \mathcal{A}^i(\mathcal{H}om(\xi, \xi)).
\end{aligned}$$

Therefore, $\nabla^1 - \nabla^2 \in \mathcal{A}^1(\mathcal{H}om(\xi, \xi))$, that is, it is a one-form with values in $\Gamma(\mathcal{H}om(\xi, \xi))$. But then, the vector space, $\Gamma(\mathcal{H}om(\xi, \xi))$, acts on the space of connections (by addition) and makes the space of connections into an affine space. Given any connection, ∇ and any one-form, $\omega \in \Gamma(\mathcal{H}om(\xi, \xi))$, the expression $\nabla + \omega$ is also a connection. Equivalently, any affine combination of connections is also a connection.

A basic property of ∇ is that it is a local operator.

Proposition 11.2. *Let $\xi = (E, \pi, B, V)$ be a smooth real vector bundle and let ∇ be a connection on ξ. For every open subset, $U \subseteq B$, for every section, $s \in \Gamma(\xi)$, if $s \equiv 0$ on U, then $\nabla s \equiv 0$ on U, that is, ∇ is a local operator.*

Proof. By Proposition 3.30 applied to the constant function with value 1, for every $p \in U$, there is some open subset, $V \subseteq U$, containing p and a smooth function, $f \colon B \to \mathbb{R}$, such that $\mathrm{supp}\, f \subseteq U$ and $f \equiv 1$ on V. Consequently, fs is a smooth section which is identically zero. By applying the Leibnitz rule, we get

$$0 = \nabla(fs) = df \otimes s + f\nabla s,$$

which, evaluated at p yields $(\nabla s)(p) = 0$, since $f(p) = 1$ and $df \equiv 0$ on V. □

As an immediate consequence of Proposition 11.2, if s_1 and s_2 are two sections in $\Gamma(\xi)$ that agree on U, then $s_1 - s_2$ is zero on U, so $\nabla(s_1 - s_2) = \nabla s_1 - \nabla s_2$ is zero on U, that is, ∇s_1 and ∇s_2 agree on U.

Proposition 11.2 also implies that a connection, ∇, on ξ, restricts to a connection, $\nabla \restriction U$ on the vector bundle, $\xi \restriction U$, for every open subset, $U \subseteq B$. Indeed, let s be a section of ξ over U. Pick any $b \in U$ and define $(\nabla s)(b)$ as follows: Using Proposition 3.30, there is some open subset, $V_1 \subseteq U$, containing b and a smooth function, $f_1 \colon B \to \mathbb{R}$, such that $\mathrm{supp}\, f_1 \subseteq U$ and $f_1 \equiv 1$ on V_1 so, let $s_1 = f_1 s$, a global section of ξ. Clearly, $s_1 = s$ on V_1, and set

$$(\nabla s)(b) = (\nabla s_1)(b).$$

This definition does not depend on (V_1, f_1), because if we had used another pair, (V_2, f_2), as above, since $b \in V_1 \cap V_2$, we have

$$s_1 = f_1 s = s = f_2 s = s_2 \qquad \text{on} \quad V_1 \cap V_2$$

so, by Proposition 11.2,

$$(\nabla s_1)(b) = (\nabla s_2)(b).$$

It should also be noted that $(\nabla_X s)(b)$ only depends on $X(b)$, that is, for any two vector fields, $X, Y \in \mathfrak{X}(B)$, if $X(b) = Y(b)$ for some $b \in B$, then

$$(\nabla_X s)(b) = (\nabla_Y s)(b), \qquad \text{for every } s \in \Gamma(\xi).$$

As above, by linearity, it it enough to prove that if $X(b) = 0$, then $(\nabla_X s)(b) = 0$. To prove this, pick any local trivialization, (U, φ), with $b \in U$. Then, we can write

$$X \restriction U = \sum_{i=1}^{d} X_i \frac{\partial}{\partial x_i}.$$

However, as before, we can find a pair, (V, f), with $b \in V \subseteq U$, $\operatorname{supp} f \subseteq U$ and $f = 1$ on V, so that $f\frac{\partial}{\partial x_i}$ is a smooth vector field on B and $f\frac{\partial}{\partial x_i}$ agrees with $\frac{\partial}{\partial x_i}$ on V, for $i = 1, \ldots, n$. Clearly, $fX_i \in C^\infty(B)$ and fX_i agrees with X_i on V so if we write $\widetilde{X} = f^2 X$, then

$$\widetilde{X} = f^2 X = \sum_{i=1}^{d} fX_i \, f \frac{\partial}{\partial x_i}$$

and we have

$$f^2 \nabla_X s = \nabla_{\widetilde{X}} s = \sum_{i=1}^{d} fX_i \, \nabla_{f\frac{\partial}{\partial x_i}} s.$$

Since $X_i(b) = 0$ and $f(b) = 1$, we get $(\nabla_X s)(b) = 0$, as claimed.

Using the above property, for any point, $p \in B$, we can define the covariant derivative, $(\nabla_u s)(p)$, of a section, $s \in \Gamma(\xi)$, with respect to a tangent vector, $u \in T_p B$. Indeed, pick any vector field, $X \in \mathfrak{X}(B)$, such that $X(p) = u$ (such a vector field exists locally over the domain of a chart and then extend it using a bump function) and set $(\nabla_u s)(p) = (\nabla_X s)(p)$. By the above property, if $X(p) = Y(p)$, then $(\nabla_X s)(p) = (\nabla_Y s)(p)$ so $(\nabla_u s)(p)$ is well-defined. Since ∇ is a local operator, $(\nabla_u s)(p)$ is also well defined for any tangent vector, $u \in T_p B$, and any local section, $s \in \Gamma(U, \xi)$, defined in some open subset, U, with $p \in U$. From now on, we will use this property without any further justification.

Since ξ is locally trivial, it is interesting to see what $\nabla \restriction U$ looks like when (U, φ) is a local trivialization of ξ.

Fix once and for all some basis, $(v_1, \ldots, v_n)$, of the typical fibre, V ($n = \dim(V)$). To every local trivialization, $\varphi\colon \pi^{-1}(U) \to U \times V$, of ξ (for some open subset, $U \subseteq B$), we associate the frame, $(s_1, \ldots, s_n)$, over U given by

$$s_i(b) = \varphi^{-1}(b, v_i), \qquad b \in U.$$

Then, every section, s, over U, can be written uniquely as $s = \sum_{i=1}^n f_i s_i$, for some functions $f_i \in C^\infty(U)$ and we have

$$\nabla s = \sum_{i=1}^n \nabla(f_i s_i) = \sum_{i=1}^n (df_i \otimes s_i + f_i \nabla s_i).$$

On the other hand, each ∇s_i can be written as

$$\nabla s_i = \sum_{j=1}^n \omega_{ij} \otimes s_j,$$

for some $n \times n$ matrix, $\omega = (\omega_{ij})$, of one-forms, $\omega_{ij} \in \mathcal{A}^1(U)$, so we get

$$\nabla s = \sum_{i=1}^n df_i \otimes s_i + \sum_{i=1}^n f_i \nabla s_i = \sum_{i=1}^n df_i \otimes s_i + \sum_{i,j=1}^n f_i \omega_{ij} \otimes s_j = \sum_{j=1}^n (df_j + \sum_{i=1}^n f_i \omega_{ij}) \otimes s_j.$$

With respect to the frame, $(s_1, \ldots, s_n)$, the connection ∇ has the matrix form

$$\nabla(f_1, \ldots, f_n) = (df_1, \ldots, df_n) + (f_1, \ldots, f_n)\omega$$

and the matrix, $\omega = (\omega_{ij})$, of one-forms, $\omega_{ij} \in \mathcal{A}^1(U)$, is called the *connection form* or *connection matrix* of ∇ with respect to $\varphi\colon \pi^{-1}(U) \to U \times V$. The above computation also shows that on U, any connection is uniquely determined by a matrix of one-forms, $\omega_{ij} \in \mathcal{A}^1(U)$. In particular, the connection on U for which

$$\nabla s_1 = 0, \ldots, \nabla s_n = 0,$$

corresponding to the zero matrix is called the *flat connection on* U (w.r.t. $(s_1, \ldots, s_n)$).

Some authors (such as Morita [115]) use a notation involving subscripts and superscripts, namely

$$\nabla s_i = \sum_{j=1}^n \omega_i^j \otimes s_j.$$

But, beware, the expression $\omega = (\omega_i^j)$ denotes the $n \times n$-matrix whose rows are indexed by j and whose columns are indexed by i! Accordingly, if $\theta = \omega\eta$, then

$$\theta_j^i = \sum_k \omega_k^i \eta_j^k.$$

The matrix, (ω_j^i) is thus *the transpose* of our matrix (ω_{ij}). This has the effects that some of the results differ either by a sign (as in $\omega \wedge \omega$) or by a permutation of matrices (as in the formula for a change of frame).

Remark: If $(\theta_1, \ldots, \theta_n)$ is the dual frame of $(s_1, \ldots, s_n)$, that is, $\theta_i \in \mathcal{A}^1(U)$, is the one-form defined so that

$$\theta_i(b)(s_j(b)) = \delta_{ij}, \qquad \text{for all} \quad b \in U,\ 1 \leq i, j \leq n,$$

then we can write $\omega_{ik} = \sum_{j=1}^n \Gamma_{ji}^k \theta_j$ and so,

$$\nabla s_i = \sum_{j,k=1}^n \Gamma_{ji}^k (\theta_j \otimes s_k),$$

where the $\Gamma_{ji}^k \in C^\infty(U)$ are the *Christoffel symbols*.

Proposition 11.3. *Every vector bundle, ξ, possesses a connection.*

Proof. Since ξ is locally trivial, we can find a locally finite open cover, $(U_\alpha)_\alpha$, of B such that $\pi^{-1}(U_\alpha)$ is trivial. If (f_α) is a partition of unity subordinate to the cover $(U_\alpha)_\alpha$ and if ∇^α is any flat connection on $\xi \restriction U_\alpha$, then it is immediately verified that

$$\nabla = \sum_\alpha f_\alpha \nabla^\alpha$$

is a connection on ξ. □

If $\varphi_\alpha \colon \pi^{-1}(U_\alpha) \to U_\alpha \times V$ and $\varphi_\beta \colon \pi^{-1}(U_\beta) \to U_\beta \times V$ are two overlapping trivializations, we know that for every $b \in U_\alpha \cap U_\beta$, we have

$$\varphi_\alpha \circ \varphi_\beta^{-1}(b, u) = (b, g_{\alpha\beta}(b)u),$$

where $g_{\alpha\beta} \colon U_\alpha \cap U_\beta \to \mathrm{GL}(V)$ is the transition function. As

$$\varphi_\beta^{-1}(b, u) = \varphi_\alpha^{-1}(b, g_{\alpha\beta}(b)u),$$

if $(s_1, \ldots, s_n)$ is the frame over U_α associated with φ_α and $(t_1, \ldots, t_n)$ is the frame over U_β associated with φ_β, we see that

$$t_i = \sum_{j=1}^n g_{ij} s_j,$$

where $g_{\alpha\beta} = (g_{ij})$.

Proposition 11.4. *With the notations as above, the connection matrices, ω_α and ω_β respectively over U_α and U_β obey the tranformation rule*

$$\omega_\beta = g_{\alpha\beta} \omega_\alpha g_{\alpha\beta}^{-1} + (dg_{\alpha\beta}) g_{\alpha\beta}^{-1},$$

where $dg_{\alpha\beta} = (dg_{ij})$.

To prove the above proposition, apply ∇ to both side of the equations

$$t_i = \sum_{j=1}^n g_{ij} s_j$$

and use ω_α and ω_β to express ∇t_i and ∇s_j. The details are left as an exercise.

In Morita [115] (Proposition 5.22), the order of the matrices in the equation of Proposition 11.4 must be reversed.

If $\xi = TM$, the tangent bundle of some smooth manifold, M, then a connection on TM, also called a *connection on* M is a linear map,

$$\nabla\colon \mathfrak{X}(M) \longrightarrow \mathcal{A}^1(M) \otimes_{C^\infty(M)} \mathfrak{X}(M) \cong \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), (\mathfrak{X}(M)),$$

since $\Gamma(TM) = \mathfrak{X}(M)$. Then, for fixed $Y \in \mathcal{X}(M)$, the map ∇Y is $C^\infty(M)$-linear, which implies that ∇Y is a $(1,1)$ tensor. In a local chart, (U, φ), we have

$$\nabla_{\frac{\partial}{\partial x_i}} \left(\frac{\partial}{\partial x_j} \right) = \sum_{k=1}^{n} \Gamma_{ij}^k \frac{\partial}{\partial x_k},$$

where the Γ_{ij}^k are Christoffel symbols.

The covariant derivative, ∇_X, given by a connection, ∇, on TM, can be extended to a covariant derivative, $\nabla_X^{r,s}$, defined on tensor fields in $\Gamma(M, T^{r,s}(M))$, for all $r, s \geq 0$, where

$$T^{r,s}(M) = T^{\otimes r} M \otimes (T^* M)^{\otimes s}.$$

We already have $\nabla_X^{1,0} = \nabla_X$ and it is natural to set $\nabla_X^{0,0} f = X[f] = df(X)$. Recall that there is an isomorphism between the set of tensor fields, $\Gamma(M, T^{r,s}(M))$, and the set of $C^\infty(M)$-multilinear maps,

$$\Phi\colon \underbrace{\mathcal{A}^1(M) \times \cdots \times \mathcal{A}^1(M)}_{r} \times \underbrace{\mathfrak{X}(M) \times \cdots \times \mathfrak{X}(M)}_{s} \longrightarrow C^\infty(M),$$

where $\mathcal{A}^1(M)$ and $\mathfrak{X}(M)$ are $C^\infty(M)$-modules.

The next proposition is left as an exercise. For help, see O'Neill [120], Chapter 2, Proposition 13 and Theorem 15.

Proposition 11.5. *for every vector field, $X \in \mathfrak{X}(M)$, there is a unique family of $\mathbb{R}$-linear map, $\nabla^{r,s}\colon \Gamma(M, T^{r,s}(M)) \to \Gamma(M, T^{r,s}(M))$, with $r, s \geq 0$, such that*

(a) $\nabla_X^{0,0} f = df(X)$, for all $f \in C^\infty(M)$ and $\nabla_X^{1,0} = \nabla_X$, for all $X \in \mathfrak{X}(M)$.

(b) $\nabla_X^{r_1+r_2, s_1+s_2}(S \otimes T) = \nabla_X^{r_1,s_1}(S) \otimes T + S \otimes \nabla_X^{r_2,s_2}(T)$, for all $S \in \Gamma(M, T^{r_1,s_1}(M))$ and all $T \in \Gamma(M, T^{r_2,s_2}(M))$.

(c) $\nabla_X^{r-1,s-1}(c_{ij}(S)) = c_{ij}(\nabla_X^{r,s}(S))$, for all $S \in \Gamma(M, T^{r,s}(M))$ and all contractions, c_{ij}, of $\Gamma(M, T^{r,s}(M))$.

Furthermore,

$$(\nabla_X^{0,1}\theta)(Y) = X[\theta(Y)] - \theta(\nabla_X Y),$$

for all $X, Y \in \mathfrak{X}(M)$ and all one-forms, $\theta \in \mathcal{A}^1(M)$ and for every $S \in \Gamma(M, T^{r,s}(M))$, with $r + s \geq 2$, the covariant derivative, $\nabla^{r,s}_X(S)$, is given by

$$\begin{aligned}(\nabla^{r,s}_X S)(\theta_1, \ldots, \theta_r, X_1, \ldots, X_s) &= X[S(\theta_1, \ldots, \theta_r, X_1, \ldots, X_s)] \\ &\quad - \sum_{i=1}^{r} S(\theta_1, \ldots, \nabla^{0,1}_X \theta_i, \ldots, \theta_r, X_1, \ldots, X_s) \\ &\quad - \sum_{j=1}^{s} S(\theta_1, \ldots, \ldots, \theta_r, X_1, \ldots, \nabla_X X_j, \ldots, X_s),\end{aligned}$$

for all $X_1, \ldots, X_s \in \mathfrak{X}(M)$ and all one-forms, $\theta_1, \ldots, \theta_r \in \mathcal{A}^1(M)$.

We define the *covariant differential*, $\nabla^{r,s} S$, of a tensor, $S \in \Gamma(M, T^{r,s}(M))$, as the $(r, s+1)$-tensor given by

$$(\nabla^{r,s} S)(\theta_1, \ldots, \theta_r, X, X_1, \ldots, X_s) = (\nabla^{r,s}_X S)(\theta_1, \ldots, \theta_r, X_1, \ldots, X_s),$$

for all $X, X_j \in \mathfrak{X}(M)$ and all $\theta_i \in \mathcal{A}^1(M)$. For simplicity of notation we usually omit the superscripts r and s. In particular, for $S = g$, the Riemannian metric on M (a $(0,2)$ tensor), we get

$$\nabla_X(g)(Y, Z) = d(g(Y, Z))(X) - g(\nabla_X Y, Z) - g(Y, \nabla_X Z),$$

for all $X, Y, Z \in \mathfrak{X}(M)$. We will see later on that a connection on M is compatible with a metric, g, iff $\nabla_X(g) = 0$.

Everything we did in this section applies to complex vector bundles by considering complex vector spaces instead of real vector spaces, $\mathbb{C}$-linear maps instead of $\mathbb{R}$-linear map, and the space of smooth complex-valued functions, $C^\infty(B; \mathbb{C}) \cong C^\infty(B) \otimes_{\mathbb{R}} \mathbb{C}$. We also use spaces of complex-valued differentials forms,

$$\mathcal{A}^i(B; \mathbb{C}) = \mathcal{A}^i(B) \otimes_{C^\infty(B)} C^\infty(B; \mathbb{C}) \cong \Gamma\Big(\Big(\bigwedge^i T^*B\Big) \otimes \epsilon^1_{\mathbb{C}}\Big),$$

where $\epsilon^1_{\mathbb{C}}$ is the trivial complex line bundle, $B \times \mathbb{C}$, and we define $\mathcal{A}^i(\xi)$ as

$$\mathcal{A}^i(\xi) = \mathcal{A}^i(B; \mathbb{C}) \otimes_{C^\infty(B;\mathbb{C})} \Gamma(\xi).$$

A connection is a $\mathbb{C}$-linear map, $\nabla \colon \Gamma(\xi) \to \mathcal{A}^1(\xi)$, that satisfies the same Leibnitz-type rule as before. Obviously, every differential form in $\mathcal{A}^i(B; \mathbb{C})$ can be written uniquely as $\omega + i\eta$, with $\omega, \eta \in \mathcal{A}^i(B)$. The exterior differential,

$$d \colon \mathcal{A}^i(B; \mathbb{C}) \to \mathcal{A}^{i+1}(B; \mathbb{C})$$

is defined by $d(\omega + i\eta) = d\omega + i d\eta$. We obtain complex-valued de Rham cohomology groups,

$$H^i_{\mathrm{DR}}(M; \mathbb{C}) = H^i_{\mathrm{DR}}(M) \otimes_{\mathbb{R}} \mathbb{C}.$$

11.2 Curvature, Curvature Form and Curvature Matrix

If $\xi = B \times V$ is the trivial bundle and ∇ is a flat connection on ξ, we obviously have

$$\nabla_X \nabla_Y - \nabla_Y \nabla_X = \nabla_{[X,Y]},$$

where $[X, Y]$ is the Lie bracket of the vector fields X and Y. However, for general bundles and arbitrary connections, the above fails. The error term,

$$R(X, Y) = \nabla_X \nabla_Y - \nabla_Y \nabla_X - \nabla_{[X,Y]},$$

measures what's called the *curvature* of the connection. The curvature of a connection also turns up as the failure of a certain sequence involving the spaces $\mathcal{A}^i(\xi)$ to be a cochain complex. Recall that a connection on ξ is a linear map

$$\nabla \colon \mathcal{A}^0(\xi) \to \mathcal{A}^1(\xi)$$

satisfying a Leibnitz-type rule. It is natural to ask whether ∇ can be extended to a family of operators, $d^\nabla \colon \mathcal{A}^i(\xi) \to \mathcal{A}^{i+1}(\xi)$, with properties analogous to d on $\mathcal{A}^*(B)$.

This is indeed the case and we get a sequence of map,

$$0 \longrightarrow \mathcal{A}^0(\xi) \xrightarrow{\nabla} \mathcal{A}^1(\xi) \xrightarrow{d^\nabla} \mathcal{A}^2(\xi) \longrightarrow \cdots \longrightarrow \mathcal{A}^i(\xi) \xrightarrow{d^\nabla} \mathcal{A}^{i+1}(\xi) \longrightarrow \cdots,$$

but in general, $d^\nabla \circ d^\nabla = 0$ fails. In particular, $d^\nabla \circ \nabla = 0$ generally fails. The term $K^\nabla = d^\nabla \circ \nabla$ is the *curvature form (or tensor)* of the connection ∇. As we will see it yields our previous curvature, R, back.

Our next goal is to define d^∇. For this, we first define an $C^\infty(B)$-bilinear map

$$\wedge \colon \mathcal{A}^i(\xi) \times \mathcal{A}^j(\eta) \longrightarrow \mathcal{A}^{i+j}(\xi \otimes \eta)$$

as follows:

$$(\omega \otimes s) \wedge (\tau \otimes t) = (\omega \wedge \tau) \otimes (s \otimes t),$$

where $\omega \in \mathcal{A}^i(B)$, $\tau \in \mathcal{A}^j(B)$, $s \in \Gamma(\xi)$, and $t \in \Gamma(\eta)$, where we used the fact that

$$\Gamma(\xi \otimes \eta) = \Gamma(\xi) \otimes_{C^\infty(B)} \Gamma(\eta).$$

First, consider the case where $\xi = \epsilon^1 = B \times \mathbb{R}$, the trivial line bundle over B. In this case, $\mathcal{A}^i(\xi) = \mathcal{A}^i(B)$ and we have a bilinear map

$$\wedge \colon \mathcal{A}^i(B) \times \mathcal{A}^j(\eta) \longrightarrow \mathcal{A}^{i+j}(\eta)$$

given by

$$\omega \wedge (\tau \otimes t) = (\omega \wedge \tau) \otimes t.$$

For $j = 0$, we have the bilinear map

$$\wedge\colon \mathcal{A}^i(B) \times \Gamma(\eta) \longrightarrow \mathcal{A}^i(\eta)$$

given by

$$\omega \wedge t = \omega \otimes t.$$

It is clear that the bilinear map

$$\wedge\colon \mathcal{A}^r(B) \times \mathcal{A}^s(\eta) \longrightarrow \mathcal{A}^{r+s}(\eta)$$

has the following properties:

$$\begin{aligned} (\omega \wedge \tau) \wedge \theta &= \omega \wedge (\tau \wedge \theta) \\ 1 \wedge \theta &= \theta, \end{aligned}$$

for all $\omega \in \mathcal{A}^i(B), \tau \in \mathcal{A}^j(B)$, $\theta \in \mathcal{A}^k(\xi)$ and where 1 denotes the constant function in $C^\infty(B)$ with value 1.

Proposition 11.6. *For every vector bundle, ξ, for all $j \geq 0$, there is a unique $\mathbb{R}$-linear map (resp. $\mathbb{C}$-linear if ξ is a complex VB), $d^\nabla\colon \mathcal{A}^j(\xi) \to \mathcal{A}^{j+1}(\xi)$, such that*

(i) $d^\nabla = \nabla$ for $j = 0$.

(ii) $d^\nabla(\omega \wedge t) = d\omega \wedge t + (-1)^i \omega \wedge d^\nabla t$, for all $\omega \in \mathcal{A}^i(B)$ and all $t \in \mathcal{A}^j(\xi)$.

Proof. Recall that $\mathcal{A}^j(\xi) = \mathcal{A}^j(B) \otimes_{C^\infty(B)} \Gamma(\xi)$ and define $d^\nabla\colon \mathcal{A}^j(B) \times \Gamma(\xi) \to \mathcal{A}^{j+1}(\xi)$ by

$$d^\nabla(\omega, s) = d\omega \otimes s + (-1)^j \omega \wedge \nabla s,$$

for all $\omega \in \mathcal{A}^j(B)$ and all $s \in \Gamma(\xi)$. We claim that d^∇ induces an $\mathbb{R}$-linear map on $\mathcal{A}^j(\xi)$ but there is a complication as d^∇ is not $C^\infty(B)$-bilinear. The way around this problem is to use Proposition 22.37. For this, we need to check that d^∇ satisfies the condition of Proposition 22.37, where the right action of $C^\infty(B)$ on $\mathcal{A}^j(B)$ is equal to the left action, namely wedging:

$$f \wedge \omega = \omega \wedge f \qquad f \in C^\infty(B) = \mathcal{A}^0(B),\ \omega \in \mathcal{A}^j(B).$$

As $\wedge$ is $C^\infty(B)$-bilinear and $\tau \otimes s = \tau \wedge s$ for all $\tau \in \mathcal{A}^i(B)$ and all $s \in \Gamma(\xi)$, we have

$$\begin{aligned} d^\nabla(\omega f, s) &= d(\omega f) \otimes s + (-1)^j (\omega f) \wedge \nabla s \\ &= d(\omega f) \wedge s + (-1)^j f\omega \wedge \nabla s \\ &= ((d\omega) f + (-1)^j \omega \wedge df) \wedge s + (-1)^j f\omega \wedge \nabla s \\ &= f d\omega \wedge s + (-1)^j \omega \wedge df \wedge s + (-1)^j f\omega \wedge \nabla s \end{aligned}$$

and

$$\begin{aligned} d^\nabla(\omega, fs) &= d\omega \otimes (fs) + (-1)^j \omega \wedge \nabla(fs) \\ &= d\omega \wedge (fs) + (-1)^j \omega \wedge \nabla(fs) \\ &= f d\omega \wedge s + (-1)^j \omega \wedge (df \otimes s + f\nabla s) \\ &= f d\omega \wedge s + (-1)^j \omega \wedge (df \wedge s + f\nabla s) \\ &= f d\omega \wedge s + (-1)^j \omega \wedge df \wedge s + (-1)^j f\omega \wedge \nabla s. \end{aligned}$$

Thus, $d^\nabla(\omega f, s) = d^\nabla(\omega, fs)$, and Proposition 22.37 shows that $d^\nabla \colon \mathcal{A}^j(\xi) \to \mathcal{A}^{j+1}(\xi)$ is a well-defined $\mathbb{R}$-linear map for all $j \geq 0$. Furthermore, it is clear that $d^\nabla = \nabla$ for $j = 0$. Now, for $\omega \in \mathcal{A}^i(B)$ and $t = \tau \otimes s \in \mathcal{A}^j(\xi)$ we have

$$\begin{aligned} d^\nabla(\omega \wedge (\tau \otimes s)) &= d^\nabla((\omega \wedge \tau) \otimes s)) \\ &= d(\omega \wedge \tau) \otimes s + (-1)^{i+j}(\omega \wedge \tau) \wedge \nabla s \\ &= (d\omega \wedge \tau) \otimes s + (-1)^i(\omega \wedge d\tau) \otimes s + (-1)^{i+j}(\omega \wedge \tau) \wedge \nabla s \\ &= d\omega \wedge (\tau \otimes s + (-1)^i \omega \wedge (d\tau \otimes s) + (-1)^{i+j}\omega \wedge (\tau \wedge \nabla s) \\ &= d\omega \wedge (\tau \otimes s) + (-1)^i \omega \wedge d^\nabla(\tau \wedge s), \\ &= d\omega \wedge (\tau \otimes s) + (-1)^i \omega \wedge d^\nabla(\tau \otimes s), \end{aligned}$$

which proves (ii). □

As a consequence, we have the following sequence of linear maps:

$$0 \longrightarrow \mathcal{A}^0(\xi) \xrightarrow{\nabla} \mathcal{A}^1(\xi) \xrightarrow{d^\nabla} \mathcal{A}^2(\xi) \longrightarrow \cdots \longrightarrow \mathcal{A}^i(\xi) \xrightarrow{d^\nabla} \mathcal{A}^{i+1}(\xi) \longrightarrow \cdots .$$

but in general, $d^\nabla \circ d^\nabla = 0$ fails. Although generally $d^\nabla \circ \nabla = 0$ fails, the map $d^\nabla \circ \nabla$ is $C^\infty(B)$-linear. Indeed,

$$\begin{aligned} (d^\nabla \circ \nabla)(fs) &= d^\nabla(df \otimes s + f\nabla s) \\ &= d^\nabla(df \wedge s + f \wedge \nabla s) \\ &= ddf \wedge s - df \wedge \nabla s + df \wedge \nabla s + f \wedge d^\nabla(\nabla s) \\ &= f(d^\nabla \circ \nabla)(s)). \end{aligned}$$

Therefore, $d^\nabla \circ \nabla \colon \mathcal{A}^0(\xi) \to \mathcal{A}^2(\xi)$ is a $C^\infty(B)$-linear map. However, recall that just before Proposition 11.2 we showed that

$$\mathrm{Hom}_{C^\infty(B)}(\mathcal{A}^0(\xi), \mathcal{A}^i(\xi)) \cong \mathcal{A}^i(\mathcal{H}om(\xi, \xi)),$$

therefore, $d^\nabla \circ \nabla \in \mathcal{A}^2(\mathcal{H}om(\xi, \xi))$, that is, $d^\nabla \circ \nabla$ is a two-form with values in $\Gamma(\mathcal{H}om(\xi, \xi))$.

Definition 11.2. For any vector bundle, ξ, and any connection, ∇, on ξ, the vector-valued two-form, $R^\nabla = d^\nabla \circ \nabla \in \mathcal{A}^2(\mathcal{H}om(\xi, \xi))$ is the *curvature form* (or *curvature tensor*) of the connection ∇. We say that ∇ is a *flat connection* iff $R^\nabla = 0$.

For simplicity of notation, we also write R for R^∇. The expression R^∇ is also denoted F^∇ or K^∇. As in the case of a connection, we can express R^∇ locally in any local trivialization, $\varphi\colon \pi^{-1}(U) \to U \times V$, of ξ. Since $R^\nabla = d^\nabla \circ \nabla \in \mathcal{A}^2(\xi) = \mathcal{A}^j(B) \otimes_{C^\infty(B)} \Gamma(\xi)$, if $(s_1, \ldots, s_n)$ is the frame associated with (φ, U), then

$$R^\nabla(s_i) = \sum_{j=1}^n \Omega_{ij} \otimes s_j,$$

for some matrix, $\Omega = (\Omega_{ij})$, of two forms, $\Omega_{ij} \in \mathcal{A}^2(B)$. We call $\Omega = (\Omega_{ij})$ the *curvature matrix* (or *curvature form*) associated with the local trivialization. The relationship between the connection form, ω, and the curvature form, Ω, is simple:

Proposition 11.7. *(Structure Equations) Let ξ be any vector bundle and let ∇ be any connection on ξ. For every local trivialization, $\varphi\colon \pi^{-1}(U) \to U \times V$, the connection matrix, $\omega = (\omega_{ij})$, and the curvature matrix, $\Omega = (\Omega_{ij})$, associated with the local trivialization, (φ, U), are related by the* **structure equation***:*

$$\Omega = d\omega - \omega \wedge \omega.$$

Proof. By definition,

$$\nabla(s_i) = \sum_{j=1}^n \omega_{ij} \otimes s_j,$$

so if we apply d^∇ and use property (ii) of Proposition 11.6 we get

$$\begin{aligned}
d^\nabla(\nabla(s_i)) &= \sum_{k=1}^n \Omega_{ik} \otimes s_k \\
&= \sum_{j=1}^n d^\nabla(\omega_{ij} \otimes s_j) \\
&= \sum_{j=1}^n d\omega_{ij} \otimes s_j - \sum_{j=1}^n \omega_{ij} \wedge \nabla s_j \\
&= \sum_{j=1}^n d\omega_{ij} \otimes s_j - \sum_{j=1}^n \omega_{ij} \wedge \sum_{k=1}^n \omega_{jk} \otimes s_k \\
&= \sum_{k=1}^n d\omega_{ik} \otimes s_k - \sum_{k=1}^n \Big(\sum_{j=1}^n \omega_{ij} \wedge \omega_{jk}\Big) \otimes s_k,
\end{aligned}$$

and so,

$$\Omega_{ik} = d\omega_{ik} - \sum_{j=1}^n \omega_{ij} \wedge \omega_{jk},$$

which, means that

$$\Omega = d\omega - \omega \wedge \omega,$$

as claimed. □

Some other texts, including Morita [115] (Theorem 5.21) state the structure equations as

$$\Omega = d\omega + \omega \wedge \omega.$$

Although this is far from obvious from Definition 11.2, the curvature form, R^∇, is related to the curvature, $R(X,Y)$, defined at the beginning of Section 11.2. For this, we define the evaluation map

$$\mathrm{Ev}_{X,Y} \colon \mathcal{A}^2(\mathcal{H}om(\xi,\xi)) \to \mathcal{A}^0(\mathcal{H}om(\xi,\xi)) = \Gamma(\mathcal{H}om(\xi,\xi)),$$

as follows: For all $X, Y \in \mathfrak{X}(B)$, all $\omega \otimes h \in \mathcal{A}^2(\mathcal{H}om(\xi,\xi)) = \mathcal{A}^2(B) \otimes_{C^\infty(B)} \Gamma(\mathcal{H}om(\xi,\xi))$, set

$$\mathrm{Ev}_{X,Y}(\omega \otimes h) = \omega(X,Y)h.$$

It is clear that this map is $C^\infty(B)$-linear and thus well-defined on $\mathcal{A}^2(\mathcal{H}om(\xi,\xi))$. (Recall that $\mathcal{A}^0(\mathcal{H}om(\xi,\xi)) = \Gamma(\mathcal{H}om(\xi,\xi)) = \mathrm{Hom}_{C^\infty(B)}(\Gamma(\xi),\Gamma(\xi))$.) We write

$$R^\nabla_{X,Y} = \mathrm{Ev}_{X,Y}(R^\nabla) \in \mathrm{Hom}_{C^\infty(B)}(\Gamma(\xi),\Gamma(\xi)).$$

Proposition 11.8. *For any vector bundle, ξ, and any connection, ∇, on ξ, for all $X, Y \in \mathfrak{X}(B)$, if we let*

$$R(X,Y) = \nabla_X \circ \nabla_Y - \nabla_Y \circ \nabla_X - \nabla_{[X,Y]},$$

then

$$R(X,Y) = R^\nabla_{X,Y}.$$

Proof sketch. First, check that $R(X,Y)$ is $C^\infty(B)$-linear and then work locally using the frame associated with a local trivialization using Proposition 11.7. □

Remark: Proposition 11.8 implies that $R(Y,X) = -R(X,Y)$ and that $R(X,Y)(s)$ is $C^\infty(B)$-linear in X, Y and s.

If $\varphi_\alpha \colon \pi^{-1}(U_\alpha) \to U_\alpha \times V$ and $\varphi_\beta \colon \pi^{-1}(U_\beta) \to U_\beta \times V$ are two overlapping trivializations, the relationship between the curvature matrices Ω_α and Ω_β, is given by the following proposition which is the counterpart of Proposition 11.4 for the curvature matrix:

Proposition 11.9. *If $\varphi_\alpha \colon \pi^{-1}(U_\alpha) \to U_\alpha \times V$ and $\varphi_\beta \colon \pi^{-1}(U_\beta) \to U_\beta \times V$ are two overlapping trivializations of a vector bundle, ξ, then we have the following transformation rule for the curvature matrices Ω_α and Ω_β:*

$$\Omega_\beta = g_{\alpha\beta} \Omega_\alpha g_{\alpha\beta}^{-1},$$

where $g_{\alpha\beta} \colon U_\alpha \cap U_\beta \to \mathrm{GL}(V)$ is the transition function.

Proof sketch. Use the structure equations (Proposition 11.7) and apply d to the equations of Proposition 11.4. □

Proposition 11.7 also yields a formula for $d\Omega$, know as *Bianchi's identity* (in local form).

Proposition 11.10. *(Bianchi's Identity) For any vector bundle, ξ, any connection, ∇, on ξ, if ω and Ω are respectively the connection matrix and the curvature matrix, in some local trivialization, then*

$$d\Omega = \omega \wedge \Omega - \Omega \wedge \omega.$$

Proof. If we apply d to the structure equation, $\Omega = d\omega - \omega \wedge \omega$, we get

$$\begin{aligned} d\Omega &= dd\omega - d\omega \wedge \omega + \omega \wedge d\omega \\ &= -(\Omega + \omega \wedge \omega) \wedge \omega + \omega \wedge (\Omega + \omega \wedge \omega) \\ &= -\Omega \wedge \omega - \omega \wedge \omega \wedge \omega + \omega \wedge \Omega + \omega \wedge \omega \wedge \omega \\ &= \omega \wedge \Omega - \Omega \wedge \omega, \end{aligned}$$

as claimed. □

We conclude this section by giving a formula for $d^\nabla \circ d^\nabla(t)$, for any $t \in \mathcal{A}^i(\xi)$. Consider the special case of the bilinear map

$$\wedge\colon \mathcal{A}^i(\xi) \times \mathcal{A}^j(\eta) \longrightarrow \mathcal{A}^{i+j}(\xi \otimes \eta)$$

defined just before Proposition 11.6 with $j = 2$ and $\eta = \mathcal{H}om(\xi, \xi)$. This is the C^∞-bilinear map

$$\wedge\colon \mathcal{A}^i(\xi) \times \mathcal{A}^2(\mathcal{H}om(\xi, \xi)) \longrightarrow \mathcal{A}^{i+2}(\xi \otimes \mathcal{H}om(\xi, \xi)).$$

We also have the evaluation map,

$$\begin{aligned} \mathrm{ev}\colon \mathcal{A}^j(\xi \otimes \mathcal{H}om(\xi, \xi)) \cong \mathcal{A}^j(B) \otimes_{C^\infty(B)} \Gamma(\xi) \otimes_{C^\infty(B)} \mathrm{Hom}_{C^\infty(B)}(\Gamma(\xi), \Gamma(\xi)) \\ \longrightarrow \mathcal{A}^j(B) \otimes_{C^\infty(B)} \Gamma(\xi) = \mathcal{A}^j(\xi), \end{aligned}$$

given by

$$\mathrm{ev}(\omega \otimes s \otimes h) = \omega \otimes h(s),$$

with $\omega \in \mathcal{A}^j(B)$, $s \in \Gamma(\xi)$ and $h \in \mathrm{Hom}_{C^\infty(B)}(\Gamma(\xi), \Gamma(\xi))$. Let

$$\wedge\colon \mathcal{A}^i(\xi) \times \mathcal{A}^2(\mathcal{H}om(\xi, \xi)) \longrightarrow \mathcal{A}^{i+2}(\xi)$$

be the composition

$$\mathcal{A}^i(\xi) \times \mathcal{A}^2(\mathcal{H}om(\xi, \xi)) \xrightarrow{\wedge} \mathcal{A}^{i+2}(\xi \otimes \mathcal{H}om(\xi, \xi)) \xrightarrow{\mathrm{ev}} \mathcal{A}^{i+2}(\xi).$$

More explicitly, the above map is given (on generators) by

$$(\omega \otimes s) \wedge H = \omega \wedge H(s),$$

where $\omega \in \mathcal{A}^i(B)$, $s \in \Gamma(\xi)$ and $H \in \mathrm{Hom}_{C^\infty(B)}(\Gamma(\xi), \mathcal{A}^2(\xi)) \cong \mathcal{A}^2(\mathcal{H}om(\xi, \xi))$.

Proposition 11.11. *For any vector bundle, ξ, and any connection, ∇, on ξ the composition $d^\nabla \circ d^\nabla \colon \mathcal{A}^i(\xi) \to \mathcal{A}^{i+2}(\xi)$ maps t to $t \wedge R^\nabla$, for any $t \in \mathcal{A}^i(\xi)$.*

Proof. Any $t \in \mathcal{A}^i(\xi)$ is some linear combination of elements $\omega \otimes s \in \mathcal{A}^i(B) \otimes_{C^\infty(B)} \Gamma(\xi)$ and by Proposition 11.6, we have

$$\begin{aligned} d^\nabla \circ d^\nabla(\omega \otimes s) &= d^\nabla(d\omega \otimes s + (-1)^i \omega \wedge \nabla s) \\ &= dd\omega \otimes s + (-1)^{i+1} d\omega \wedge \nabla s + (-1)^i d\omega \wedge \nabla s + (-1)^i(-1)^i \omega \wedge d^\nabla \circ \nabla s \\ &= \omega \wedge d^\nabla \circ \nabla s \\ &= (\omega \otimes s) \wedge R^\nabla, \end{aligned}$$

as claimed. □

Proposition 11.11 shows that $d^\nabla \circ d^\nabla = 0$ iff $R^\nabla = d^\nabla \circ \nabla = 0$, that is, iff the connection ∇ is flat. Thus, the sequence

$$0 \longrightarrow \mathcal{A}^0(\xi) \xrightarrow{\nabla} \mathcal{A}^1(\xi) \xrightarrow{d^\nabla} \mathcal{A}^2(\xi) \longrightarrow \cdots \longrightarrow \mathcal{A}^i(\xi) \xrightarrow{d^\nabla} \mathcal{A}^{i+1}(\xi) \longrightarrow \cdots,$$

is a cochain complex iff ∇ is flat.

Again, everything we did in this section applies to complex vector bundles.

11.3 Parallel Transport

The notion of connection yields the notion of parallel transport in a vector bundle. First, we need to define the covariant derivative of a section along a curve.

Definition 11.3. Let $\xi = (E, \pi, B, V)$ be a vector bundle and let $\gamma \colon [a, b] \to B$ be a smooth curve in B. A *smooth section along the curve* γ is a smooth map, $X \colon [a, b] \to E$, such that $\pi(X(t)) = \gamma(t)$, for all $t \in [a, b]$. When $\xi = TB$, the tangent bundle of the manifold, B, we use the terminology *smooth vector field along* γ.

Recall that the curve $\gamma \colon [a, b] \to B$ is smooth iff γ is the restriction to $[a, b]$ of a smooth curve on some open interval containing $[a, b]$.

Proposition 11.12. *Let ξ be a vector bundle, ∇ be a connection on ξ and $\gamma \colon [a, b] \to B$ be a smooth curve in B. There is a $\mathbb{R}$-linear map, D/dt, defined on the vector space of smooth sections, X, along γ, which satisfies the following conditions:*

(1) For any smooth function, $f \colon [a, b] \to \mathbb{R}$,

$$\frac{D(fX)}{dt} = \frac{df}{dt} X + f \frac{DX}{dt}$$

(2) If X is induced by a global section, $s \in \Gamma(\xi)$, that is, if $X(t_0) = s(\gamma(t_0))$ for all $t_0 \in [a,b]$, then

$$\frac{DX}{dt}(t_0) = (\nabla_{\gamma'(t_0)}\, s)_{\gamma(t_0)}.$$

Proof. Since $\gamma([a,b])$ is compact, it can be covered by a finite number of open subsets, U_α, such that $(U_\alpha, \varphi_\alpha)$ is a local trivialization. Thus, we may assume that $\gamma\colon [a,b] \to U$ for some local trivialization, (U,φ). As $\varphi \circ \gamma\colon [a,b] \to \mathbb{R}^n$, we can write

$$\varphi \circ \gamma(t) = (u_1(t), \ldots, u_n(t)),$$

where each $u_i = pr_i \circ \varphi \circ \gamma$ is smooth. Now (see Definition 3.17), for every $g \in C^\infty(B)$, as

$$d\gamma_{t_0}\left(\left.\frac{d}{dt}\right|_{t_0}\right)(g) = \left.\frac{d}{dt}(g\circ\gamma)\right|_{t_0} = \left.\frac{d}{dt}((g\circ\varphi^{-1})\circ(\varphi\circ\gamma))\right|_{t_0} = \sum_{i=1}^n \frac{du_i}{dt}\left(\frac{\partial}{\partial x_i}\right)_{\gamma(t_0)} g,$$

since by definition of $\gamma'(t_0)$,

$$\gamma'(t_0) = d\gamma_{t_0}\left(\left.\frac{d}{dt}\right|_{t_0}\right)$$

(see the end of Section 3.2), we have

$$\gamma'(t_0) = \sum_{i=1}^n \frac{du_i}{dt}\left(\frac{\partial}{\partial x_i}\right)_{\gamma(t_0)}.$$

If $(s_1, \ldots, s_n)$ is a frame over U, we can write

$$X(t) = \sum_{i=1}^n X_i(t) s_i(\gamma(t)),$$

for some smooth functions, X_i. Then, conditions (1) and (2) imply that

$$\frac{DX}{dt} = \sum_{j=1}^n \left(\frac{dX_j}{dt}\, s_j(\gamma(t)) + X_j(t)\nabla_{\gamma'(t)}(s_j(\gamma(t)))\right)$$

and since

$$\gamma'(t) = \sum_{i=1}^n \frac{du_i}{dt}\left(\frac{\partial}{\partial x_i}\right)_{\gamma(t)},$$

there exist some smooth functions, Γ_{ij}^k, so that

$$\nabla_{\gamma'(t)}(s_j(\gamma(t))) = \sum_{i=1}^n \frac{du_i}{dt}\,\nabla_{\frac{\partial}{\partial x_i}}(s_j(\gamma(t))) = \sum_{i,k} \frac{du_i}{dt}\,\Gamma_{ij}^k s_k(\gamma(t)).$$

It follows that

$$\frac{DX}{dt} = \sum_{k=1}^{n} \left(\frac{dX_k}{dt} + \sum_{ij} \Gamma_{ij}^k \frac{du_i}{dt} X_j \right) s_k(\gamma(t)).$$

Conversely, the above expression defines a linear operator, D/dt, and it is easy to check that it satisfies (1) and (2). □

The operator, D/dt is often called *covariant derivative along* γ and it is also denoted by $\nabla_{\gamma'(t)}$ or simply $\nabla_{\gamma'}$.

Definition 11.4. Let ξ be a vector bundle and let ∇ be a connection on ξ. For every curve, $\gamma\colon [a,b] \to B$, in B, a section, X, along γ is *parallel (along γ)* iff

$$\frac{DX}{dt} = 0.$$

If M was embedded in $\mathbb{R}^d$ (for some d), then to say that X is parallel along γ would mean that the directional derivative, $(D_{\gamma'}X)(\gamma(t))$, is normal to $T_{\gamma(t)}M$.

The following proposition can be shown using the existence and uniqueness of solutions of ODE's (in our case, linear ODE's) and its proof is omitted:

Proposition 11.13. *Let ξ be a vector bundle and let ∇ be a connection on ξ. For every C^1 curve, $\gamma\colon [a,b] \to B$, in B, for every $t \in [a,b]$ and every $v \in \pi^{-1}(\gamma(t))$, there is a unique parallel section, X, along γ such that $X(t) = v$.*

For the proof of Proposition 11.13 it is sufficient to consider the portions of the curve γ contained in some local trivialization. In such a trivialization, (U, φ), as in the proof of Proposition 11.12, using a local frame, $(s_1, \ldots, s_n)$, over U, we have

$$\frac{DX}{dt} = \sum_{k=1}^{n} \left(\frac{dX_k}{dt} + \sum_{ij} \Gamma_{ij}^k \frac{du_i}{dt} X_j \right) s_k(\gamma(t)),$$

with $u_i = pr_i \circ \varphi \circ \gamma$. Consequently, X is parallel along our portion of γ iff the system of linear ODE's in the unknowns, X_k,

$$\frac{dX_k}{dt} + \sum_{ij} \Gamma_{ij}^k \frac{du_i}{dt} X_j = 0, \qquad k = 1, \ldots, n,$$

is satisfied.

Remark: Proposition 11.13 can be extended to piecewise C^1 curves.

Definition 11.5. Let ξ be a vector bundle and let ∇ be a connection on ξ. For every curve, $\gamma\colon [a,b] \to B$, in B, for every $t \in [a,b]$, the *parallel transport from $\gamma(a)$ to $\gamma(t)$ along γ* is the linear map from the fibre, $\pi^{-1}(\gamma(a))$, to the fibre, $\pi^{-1}(\gamma(t))$, which associates to any $v \in \pi^{-1}(\gamma(a))$ the vector $X_v(t) \in \pi^{-1}(\gamma(t))$, where X_v is the unique parallel section along γ with $X_v(a) = v$.

The following proposition is an immediate consequence of properties of linear ODE's:

Proposition 11.14. *Let $\xi = (E, \pi, B, V)$ be a vector bundle and let ∇ be a connection on ξ. For every C^1 curve, $\gamma\colon [a,b] \to B$, in B, the parallel transport along γ defines for every $t \in [a,b]$ a linear isomorphism, $P_\gamma\colon \pi^{-1}(\gamma(a)) \to \pi^{-1}(\gamma(t))$, between the fibres $\pi^{-1}(\gamma(a))$ and $\pi^{-1}(\gamma(t))$.*

In particular, if γ is a closed curve, that is, if $\gamma(a) = \gamma(b) = p$, we obtain a linear isomorphism, P_γ, of the fibre $E_p = \pi^{-1}(p)$, called the *holonomy of γ*. The *holonomy group of ∇ based at p*, denoted $\mathrm{Hol}_p(\nabla)$, is the subgroup of $\mathrm{GL}(V, \mathbb{R})$ given by

$$\mathrm{Hol}_p(\nabla) = \{P_\gamma \in \mathrm{GL}(V, \mathbb{R}) \mid \gamma \text{ is a closed curve based at } p\}.$$

If B is connected, then $\mathrm{Hol}_p(\nabla)$ depends on the basepoint $p \in B$ up to conjugation and so $\mathrm{Hol}_p(\nabla)$ and $\mathrm{Hol}_q(\nabla)$ are isomorphic for all $p, q \in B$. In this case, it makes sense to talk about the *holonomy group of ∇*. If $\xi = TB$, the tangent bundle of a manifold, B, by abuse of language, we call $\mathrm{Hol}_p(\nabla)$ the *holonomy group of B*.

11.4 Connections Compatible with a Metric; Levi-Civita Connections

If a vector bundle (or a Riemannian manifold), ξ, has a metric, then it is natural to define when a connection, ∇, on ξ is compatible with the metric. So, assume the vector bundle, ξ, has a metric, $\langle -, - \rangle$. We can use this metric to define pairings

$$\mathcal{A}^1(\xi) \times \mathcal{A}^0(\xi) \longrightarrow \mathcal{A}^1(B) \quad \text{and} \quad \mathcal{A}^0(\xi) \times \mathcal{A}^1(\xi) \longrightarrow \mathcal{A}^1(B)$$

as follows: Set (on generators)

$$\langle \omega \otimes s_1, s_2 \rangle = \langle s_1, \omega \otimes s_2 \rangle = \omega \langle s_1, s_2 \rangle,$$

for all $\omega \in \mathcal{A}^1(B)$, $s_1, s_2 \in \Gamma(\xi)$ and where $\langle s_1, s_2 \rangle$ is the function in $C^\infty(B)$ given by $b \mapsto \langle s_1(b), s_2(b) \rangle$, for all $b \in B$. More generally, we define a pairing

$$\mathcal{A}^i(\xi) \times \mathcal{A}^j(\xi) \longrightarrow \mathcal{A}^{i+j}(B),$$

by

$$\langle \omega \otimes s_1, \eta \otimes s_2 \rangle = \langle s_1, s_2 \rangle \omega \wedge \eta,$$

for all $\omega \in \mathcal{A}^i(B), \eta \in \mathcal{A}^j(B)$, $s_1, s_2 \in \Gamma(\xi)$.

Definition 11.6. Given any metric, $\langle -, - \rangle$, on a vector bundle, ξ, a connection, ∇, on ξ is *compatible with the metric*, for short, a *metric connection* iff

$$d\langle s_1, s_2 \rangle = \langle \nabla s_1, s_2 \rangle + \langle s_1, \nabla s_2 \rangle,$$

for all $s_1, s_2 \in \Gamma(\xi)$.

In terms of version-two of a connection, ∇_X is a metric connection iff

$$X(\langle s_1, s_2 \rangle) = \langle \nabla_X s_1, s_2 \rangle + \langle s_1, \nabla_X s_2 \rangle,$$

for every vector field, $X \in \mathfrak{X}(B)$.

Definition 11.6 remains unchanged if ξ is a complex vector bundle. The condition of compatibility with a metric is nicely expressed in a local trivialization. Indeed, let (U, φ) be a local trivialization of the vector bundle, ξ (of rank n). Then, using the Gram-Schmidt procedure, we obtain an orthonormal frame, $(s_1, \ldots, s_n)$, over U.

Proposition 11.15. *Using the above notations, if $\omega = (\omega_{ij})$ is the connection matrix of ∇ w.r.t. $(s_1, \ldots, s_n)$, then ω is skew-symmetric.*

Proof. Since

$$\nabla e_i = \sum_{j=1}^{n} \omega_{ij} \otimes s_j$$

and since $\langle s_i, s_j \rangle = \delta_{ij}$ (as $(s_1, \ldots, s_n)$ is orthonormal), we have $d\langle s_i, s_j \rangle = 0$ on U. Consequently

$$\begin{aligned} 0 &= d\langle s_i, s_j \rangle \\ &= \langle \nabla s_i, s_j \rangle + \langle s_i, \nabla s_j \rangle \\ &= \langle \sum_{k=1}^{n} \omega_{ik} \otimes s_k, s_j \rangle + \langle s_i, \sum_{l=1}^{n} \omega_{jl} \otimes s_l \rangle \\ &= \sum_{k=1}^{n} \omega_{ik} \langle s_k, s_j \rangle + \sum_{l=1}^{n} \omega_{jl} \langle s_i, s_l \rangle \\ &= \omega_{ij} + \omega_{ji}, \end{aligned}$$

as claimed. □

In Proposition 11.15, if ξ is a complex vector bundle, then ω is skew-Hermitian. This means that

$$\overline{\omega}^{\top} = -\omega,$$

where $\overline{\omega}$ is the conjugate matrix of ω, that is, $(\overline{\omega})_{ij} = \overline{\omega_{ij}}$. It is also easy to prove that metric connections exist.

Proposition 11.16. *Let ξ be a rank n vector with a metric, $\langle -, - \rangle$. Then, ξ, possesses metric connections.*

Proof. We can pick a locally finite cover, $(U_\alpha)_\alpha$, of B such that $(U_\alpha, \varphi_\alpha)$ is a local trivialization of ξ. Then, for each $(U_\alpha, \varphi_\alpha)$, we use the Gram-Schmidt procedure to obtain an orthonormal frame, $(s_1^\alpha, \ldots, s_n^\alpha)$, over U_α and we let ∇^α be the trivial connection on $\pi^{-1}(U_\alpha)$. By construction, ∇^α is compatible with the metric. We finish the argumemt by using a partition of unity, leaving the details to the reader. □

If ξ is a complex vector bundle, then we use a Hermitian metric and we call a connection compatible with this metric a *Hermitian connection*. In any local trivialization, the connection matrix, ω, is skew-Hermitian. The existence of Hermitian connections is clear.

If ∇ is a metric connection, then the curvature matrices are also skew-symmetric.

Proposition 11.17. *Let ξ be a rank n vector bundle with a metric, $\langle -, - \rangle$. In any local trivialization of ξ, the curvature matrix, $\Omega = (\Omega_{ij})$ is skew-symmetric. If ξ is a complex vector bundle, then $\Omega = (\Omega_{ij})$ is skew-Hermitian.*

Proof. By the structure equation (Proposition 11.7),

$$\Omega = d\omega - \omega \wedge \omega,$$

that is, $\Omega_{ij} = d\omega_{ij} - \sum_{k=1}^n \omega_{ik} \wedge \omega_{kj}$, so, using the skew symetry of ω_{ij} and wedge,

$$\begin{aligned} \Omega_{ji} &= d\omega_{ji} - \sum_{k=1}^n \omega_{jk} \wedge \omega_{ki} \\ &= -d\omega_{ij} - \sum_{k=1}^n \omega_{kj} \wedge \omega_{ik} \\ &= -d\omega_{ij} + \sum_{k=1}^n \omega_{ik} \wedge \omega_{kj} \\ &= -\Omega_{ij}, \end{aligned}$$

as claimed. □

We now restrict our attention to a Riemannian manifold, that is, to the case where our bundle, ξ, is the tangent bundle, $\xi = TM$, of some Riemannian manifold, M. We know from Proposition 11.16 that metric connections on TM exist. However, there are many metric connections on TM and none of them seems more relevant than the others. If M is a Riemannian manifold, the metric, $\langle -, - \rangle$, on M is often denoted g. In this case, for every chart, (U, φ), we let $g_{ij} \in C^\infty(M)$ be the function defined by

$$g_{ij}(p) = \left\langle \left(\frac{\partial}{\partial x_i}\right)_p, \left(\frac{\partial}{\partial x_j}\right)_p \right\rangle_p.$$

(Note the unfortunate clash of notation with the transitions functions!)

The notations $g = \sum_{ij} g_{ij} dx_i \otimes dx_j$ or simply $g = \sum_{ij} g_{ij} dx_i dx_j$ are often used to denote the metric in local coordinates. We observed immediately after stating Proposition 11.5 that the covariant differential, ∇g, of the Riemannian metric, g, on M is given by

$$\nabla_X(g)(Y,Z) = d(g(Y,Z))(X) - g(\nabla_X Y, Z) - g(Y, \nabla_X Z),$$

for all $X, Y, Z \in \mathfrak{X}(M)$. Therefore, a connection, ∇, on a Riemannian manifold, (M, g), is compatible with the metric iff

$$\nabla g = 0.$$

It is remarkable that if we require a certain kind of symmetry on a metric connection, then it is uniquely determined. Such a connection is known as the *Levi-Civita connection*. The Levi-Civita connection can be characterized in several equivalent ways, a rather simple way involving the notion of torsion of a connection.

Recall that one way to introduce the curvature is to view it as the "error term"

$$R(X,Y) = \nabla_X \nabla_Y - \nabla_Y \nabla_X - \nabla_{[X,Y]}.$$

Another natural error term is the *torsion*, $T(X,Y)$, of the connection, ∇, given by

$$T(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y],$$

which measures the failure of the connection to behave like the Lie bracket.

Another way to characterize the Levi-Civita connection uses the cotangent bundle, T^*M. It turns out that a connection, ∇, on a vector bundle (metric or not), ξ, naturally induces a connection, ∇^*, on the dual bundle, ξ^*. Now, if ∇ is a connection on TM, then ∇^* is is a connection on T^*M, namely, a linear map, $\nabla^* \colon \Gamma(T^*M) \to \mathcal{A}^1(M) \otimes_{C^\infty(B)} \Gamma(T^*M)$, that is

$$\nabla^* \colon \mathcal{A}^1(M) \to \mathcal{A}^1(M) \otimes_{C^\infty(B)} \mathcal{A}^1(M) \cong \Gamma(T^*M \otimes T^*M),$$

since $\Gamma(T^*M) = \mathcal{A}^1(M)$. If we compose this map with $\wedge$, we get the map

$$\mathcal{A}^1(M) \xrightarrow{\nabla^*} \mathcal{A}^1(M) \otimes_{C^\infty(B)} \mathcal{A}^1(M) \xrightarrow{\wedge} \mathcal{A}^2(M).$$

Then, miracle, a metric connection is the Levi-Civita connection iff

$$d = \wedge \circ \nabla^*,$$

where $d \colon \mathcal{A}^1(M) \to \mathcal{A}^2(M)$ is exterior differentiation. There is also a nice local expression of the above equation.

First, we consider the definition involving the torsion.

Proposition 11.18. *(Levi-Civita, Version 1) Let M be any Riemannian manifold. There is a unique, metric, torsion-free connection, ∇, on M, that is, a connection satisfying the conditions*

$$\begin{aligned} X(\langle Y, Z\rangle) &= \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle \\ \nabla_X Y - \nabla_Y X &= [X, Y], \end{aligned}$$

for all vector fields, $X, Y, Z \in \mathfrak{X}(M)$. This connection is called the **Levi-Civita connection** *(or* **canonical connection***) on M. Furthermore, this connection is determined by the* **Koszul formula**

$$\begin{aligned} 2\langle \nabla_X Y, Z\rangle &= X(\langle Y, Z\rangle) + Y(\langle X, Z\rangle) - Z(\langle X, Y\rangle) \\ &\quad - \langle Y, [X, Z]\rangle - \langle X, [Y, Z]\rangle - \langle Z, [Y, X]\rangle. \end{aligned}$$

Proof. First, we prove uniqueness. Since our metric is a non-degenerate bilinear form, it suffices to prove the Koszul formula. As our connection is compatible with the metric, we have

$$\begin{aligned} X(\langle Y, Z\rangle) &= \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle \\ Y(\langle X, Z\rangle) &= \langle \nabla_Y X, Z\rangle + \langle X, \nabla_Y Z\rangle \\ -Z(\langle X, Y\rangle) &= -\langle \nabla_Z X, Y\rangle - \langle X, \nabla_Z Y\rangle \end{aligned}$$

and by adding up the above equations, we get

$$\begin{aligned} X(\langle Y, Z\rangle) + Y(\langle X, Z\rangle) - Z(\langle X, Y\rangle) &= \langle Y, \nabla_X Z - \nabla_Z X\rangle \\ &\quad + \langle X, \nabla_Y Z - \nabla_Z Y\rangle \\ &\quad + \langle Z, \nabla_X Y + \nabla_Y X\rangle. \end{aligned}$$

Then, using the fact that the torsion is zero, we get

$$\begin{aligned} X(\langle Y, Z\rangle) + Y(\langle X, Z\rangle) - Z(\langle X, Y\rangle) &= \langle Y, [X, Z]\rangle + \langle X, [Y, Z]\rangle \\ &\quad + \langle Z, [Y, X]\rangle + 2\langle Z, \nabla_X Y\rangle \end{aligned}$$

which yields the Koszul formula.

Next, we prove existence. We begin by checking that the right-hand side of the Koszul formula is $C^\infty(M)$-linear in Z, for X and Y fixed. But then, the linear map $Z \mapsto \langle \nabla_X Y, Z\rangle$ induces a one-form and $\nabla_X Y$ is the vector field corresponding to it *via* the non-degenerate pairing. It remains to check that ∇ satisfies the properties of a connection, which it a bit tedious (for example, see Kuhnel [92], Chapter 5, Section D). □

Remark: In a chart, (U, φ), if we set

$$\partial_k g_{ij} = \frac{\partial}{\partial x_k}(g_{ij})$$

then it can be shown that the Christoffel symbols are given by

$$\Gamma_{ij}^k = \frac{1}{2}\sum_{l=1}^n g^{kl}(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}),$$

where (g^{kl}) is the inverse of the matrix (g_{kl}).

Let us now consider the second approach to torsion-freeness. For this, we have to explain how a connection, ∇, on a vector bundle, $\xi = (E, \pi, B, V)$, induces a connection, ∇^*, on the dual bundle, ξ^*. First, there is an evaluation map $\Gamma(\xi \otimes \xi^*) \longrightarrow \Gamma(\epsilon^1)$ or equivalently,

$$\langle -, - \rangle \colon \Gamma(\xi) \otimes_{C^\infty(B)} \mathrm{Hom}_{C^\infty(B)}(\Gamma(\xi), C^\infty(B)) \longrightarrow C^\infty(B),$$

given by

$$\langle s_1, s_2^* \rangle = s_2^*(s_1), \qquad s_1 \in \Gamma(\xi),\ s_2^* \in \mathrm{Hom}_{C^\infty(B)}(\Gamma(\xi), C^\infty(B))$$

and thus a map

$$\mathcal{A}^k(\xi \otimes \xi^*) = \mathcal{A}^k(B) \otimes_{C^\infty(B)} \Gamma(\xi \otimes \xi^*) \stackrel{\mathrm{id}\otimes\langle -,-\rangle}{\longrightarrow} \mathcal{A}^k(B) \otimes_{C^\infty(B)} C^\infty(B) \cong \mathcal{A}^k(B).$$

Using this map we obtain a pairing

$$(-,-) \colon \mathcal{A}^i(\xi) \otimes \mathcal{A}^j(\xi^*) \stackrel{\wedge}{\longrightarrow} \mathcal{A}^{i+j}(\xi \otimes \xi^*) \longrightarrow \mathcal{A}^{i+j}(B),$$

given by

$$(\omega \otimes s_1, \eta \otimes s_2^*) = (\omega \wedge \eta) \otimes \langle s_1, s_2^* \rangle,$$

where $\omega \in \mathcal{A}^i(B)$, $\eta \in \mathcal{A}^j(B)$, $s_1 \in \Gamma(\xi)$, $s_2^* \in \Gamma(\xi^*)$. It is easy to check that this pairing is non-degenerate. Then, given a connection, ∇, on a rank n vector bundle, ξ, we define ∇^* on ξ^* by

$$d\langle s_1, s_2^* \rangle = \big(\nabla(s_1), s_2^*\big) + \big(s_1, \nabla^*(s_2^*)\big),$$

where $s_1 \in \Gamma(\xi)$ and $s_2^* \in \Gamma(\xi^*)$. Because the pairing $(-,-)$ is non-degenerate, ∇^* is well-defined and it is immediately that it is a connection on ξ^*. Let us see how it is expressed locally. If (U, φ) is a local trivialization and $(s_1, \ldots, s_n)$ is the frame over U associated with (U, φ), then let $(\theta_1, \ldots, \theta_n)$ be the dual frame (called a *coframe*). We have

$$\langle s_j, \theta_i \rangle = \theta_i(s_j) = \delta_{ij}, \qquad 1 \leq i, j \leq n.$$

Recall that

$$\nabla s_j = \sum_{k=1}^n \omega_{jk} \otimes s_k$$

and write

$$\nabla^* \theta_i = \sum_{k=1}^n \omega_{ik}^* \otimes \theta_k.$$

Applying d to the equation $\langle s_j, \theta_i\rangle = \delta_{ij}$ and using the equation defining ∇^*, we get

$$\begin{aligned} 0 &= d\langle s_j, \theta_i\rangle \\ &= \big(\nabla(s_j), \theta_i\big) + \big(s_j, \nabla^*(\theta_i)\big) \\ &= \Big(\sum_{k=1}^n \omega_{jk} \otimes s_k, \theta_i\Big) + \Big(s_j, \sum_{l=1}^n \omega^*_{il} \otimes \theta_l\Big) \\ &= \sum_{k=1}^n \omega_{jk}\big(s_k, \theta_i\big) + \sum_{l=1}^n \omega^*_{il}\big(s_j, \theta_l\big) \\ &= \omega_{ji} + \omega^*_{ij}. \end{aligned}$$

Therefore, if we write $\omega^* = (\omega^*_{ij})$, we have

$$\omega^* = -\omega^\top.$$

If ∇ is a metric connection, then ω is skew-symmetric, that is, $\omega^\top = -\omega$. In this case, $\omega^* = -\omega^\top = \omega$.

If ξ is a complex vector bundle, then there is a problem because if $(s_1, \ldots, s_n)$ is a frame over U, then the $\theta_j(b)$'s defined by

$$\langle s_i(b), \theta_j(b)\rangle = \delta_{ij}$$

are *not* linear, but instead conjugate-linear. (Recall that a linear form, θ, is *conjugate linear* (or *semi-linear*) iff $\theta(\lambda u) = \overline{\lambda}\theta(u)$, for all $\lambda \in \mathbb{C}$.) Instead of ξ^*, we need to consider the bundle $\overline{\xi}^*$, which is the bundle whose fibre over $b \in B$ consist of all conjugate-linear forms over $\pi^{-1}(b)$. In this case, the evaluation pairing, $\langle s, \theta\rangle$ is conjugate-linear in s and we find that $\omega^* = -\overline{\omega}^\top$, where ω^* is the connection matrix of $\overline{\xi}^*$ over U. If ξ is a Hermitian bundle, as ω is skew-Hermitian, we find that $\omega^* = \omega$, which makes sense since ξ and $\overline{\xi}^*$ are canonically isomorphic. However, this does not give any information on ξ^*. For this, we consider the *conjugate bundle*, $\overline{\xi}$. This is the bundle obtained from ξ by redefining the vector space structure on each fibre, $\pi^{-1}(b)$, $b \in B$, so that

$$(x + iy)v = (x - iy)v,$$

for every $v \in \pi^{-1}(b)$. If ω is the connection matrix of ξ over U, then $\overline{\omega}$ is the connection matrix of $\overline{\xi}$ over U. If ξ has a Hermitian metric, it is easy to prove that ξ^* and $\overline{\xi}$ are canonically isomorphic (see Proposition 11.32). In fact, the Hermitian product, $\langle -, -\rangle$, establishes a pairing between $\overline{\xi}$ and ξ^* and, basically as above, we can show that if $\overline{\omega}$ is the connection matrix of $\overline{\xi}$ over U, then $\omega^* = -\omega^\top$ is the the connection matrix of ξ^* over U. As ω is skew-Hermitian, $\omega^* = \overline{\omega}$.

Going back to a connection, ∇, on a manifold, M, the connection, ∇^*, is a linear map,

$$\nabla^* \colon \mathcal{A}^1(M) \longrightarrow \mathcal{A}^1(M) \otimes \mathcal{A}^1(M) \cong (\mathfrak{X}(M))^* \otimes_{C^\infty(M)} (\mathfrak{X}(M))^* \cong (\mathfrak{X}(M) \otimes_{C^\infty(M)} \mathfrak{X}(M))^*.$$

Let us figure out what $\wedge \circ \nabla^*$ is using the above interpretation. By the definition of ∇^*,

$$\nabla^*_\theta(X, Y) = X(\theta(Y)) - \theta(\nabla_X Y),$$

for every one-form, $\theta \in \mathcal{A}^1(M)$ and all vector fields, $X, Y \in \mathfrak{X}(M)$. Applying $\wedge$, we get

$$\begin{aligned} \nabla^*_\theta(X, Y) - \nabla^*_\theta(Y, X) &= X(\theta(Y)) - \theta(\nabla_X Y) - Y(\theta(X)) + \theta(\nabla_Y X) \\ &= X(\theta(Y)) - Y(\theta(X)) - \theta(\nabla_X Y - \nabla_Y X). \end{aligned}$$

However, recall that

$$d\theta(X, Y) = X(\theta(Y)) - Y(\theta(X)) - \theta([X, Y]),$$

so we get

$$\begin{aligned} (\wedge \circ \nabla^*)(\theta)(X, Y) &= \nabla^*_\theta(X, Y) - \nabla^*_\theta(Y, X) \\ &= d\theta(X, Y) - \theta(\nabla_X Y - \nabla_Y X - [X, Y]) \\ &= d\theta(X, Y) - \theta(T(X, Y)). \end{aligned}$$

It follows that for every $\theta \in \mathcal{A}^1(M)$, we have $(\wedge \circ \nabla^*)\theta = d\theta$ iff $\theta(T(X, Y)) = 0$ for all $X, Y \in \mathfrak{X}(M)$, that is iff $T(X, Y) = 0$, for all $X, Y \in \mathfrak{X}(M)$. We record this as

Proposition 11.19. *Let ξ be a manifold with connection ∇. Then, ∇ is torsion-free (i.e., $T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y] = 0$, for all $X, Y \in \mathfrak{X}(M)$) iff*

$$\wedge \circ \nabla^* = d,$$

where $d\colon \mathcal{A}^1(M) \to \mathcal{A}^2(M)$ is exterior differentiation.

Proposition 11.19 together with Proposition 11.18 yield a second version of the Levi-Civita Theorem:

Proposition 11.20. *(Levi-Civita, Version 2) Let M be any Riemannian manifold. There is a unique, metric connection, ∇, on M, such that*

$$\wedge \circ \nabla^* = d,$$

where $d\colon \mathcal{A}^1(M) \to \mathcal{A}^2(M)$ is exterior differentiation. This connection is equal to the Levi-Civita connection in Proposition 11.18.

Remark: If ∇ is the Levi-Civita connection of some Riemannian manifold, M, for every chart, (U, φ), we have $\omega^* = \omega$, where ω is the connection matrix of ∇ over U and ω^* is the connection matrix of the dual connection ∇^*. This implies that the Christoffel symbols of ∇ and ∇^* over U are identical. Furthermore, ∇^* is a linear map

$$\nabla^*\colon \mathcal{A}^1(M) \longrightarrow \Gamma(T^*M \otimes T^*M).$$

Thus, locally in a chart, (U, φ), if (as usual) we let $x_i = pr_i \circ \varphi$, then we can write

$$\nabla^*(dx_k) = \sum_{ij} \Gamma^k_{ij} dx_i \otimes dx_j.$$

Now, if we want $\wedge \circ \nabla^* = d$, we must have $\wedge\nabla^*(dx_k) = ddx_k = 0$, that is

$$\Gamma^k_{ij} = \Gamma^k_{ji},$$

for all i, j. Therefore, torsion-freeness can indeed be viewed as a symmetry condition on the Christoffel symbols of the connection ∇.

Our third version is a local version due to Élie Cartan. Recall that locally in a chart, (U, φ), the connection, ∇^*, is given by the matrix, ω^*, such that $\omega^* = -\omega^\top$ where ω is the connection matrix of TM over U. That is, we have

$$\nabla^*\theta_i = \sum_{j=1}^n -\omega_{ji} \otimes \theta_j,$$

for some one-forms, $\omega_{ij} \in \mathcal{A}^1(M)$. Then,

$$\wedge \circ \nabla^*\theta_i = -\sum_{j=1}^n \omega_{ji} \wedge \theta_j$$

so the requirement that $d = \wedge \circ \nabla^*$ is expressed locally by

$$d\theta_i = -\sum_{j=1}^n \omega_{ji} \wedge \theta_j.$$

In addition, since our connection is metric, ω is skew-symmetric and so, $\omega^* = \omega$. Then, it is not too surprising that the following proposition holds:

Proposition 11.21. *Let M be a Riemannian manifold with metric, $\langle -, - \rangle$. For every chart, (U, φ), if $(s_1, \ldots, s_n)$ is the frame over U associated with (U, φ) and $(\theta_1, \ldots, \theta_n)$ is the corresponding coframe (dual frame), then there is a unique matrix, $\omega = (\omega_{ij})$, of one-forms, $\omega_{ij} \in \mathcal{A}^1(M)$, so that the following conditions hold:*

(i) $\omega_{ji} = -\omega_{ij}$.

(ii) $d\theta_i = \sum_{j=1}^n \omega_{ij} \wedge \theta_j$ *or, in matrix form,* $d\theta = \omega \wedge \theta$.

Proof. There is a direct proof using a combinatorial trick, for instance, see Morita [115], Chapter 5, Proposition 5.32 or Milnor and Stasheff [111], Appendix C, Lemma 8. On the other hand, if we view $\omega = (\omega_{ij})$ as a connection matrix, then we observed that (i) asserts that the connection is metric and (ii) that it is torsion-free. We conclude by applying Proposition 11.20. □

As an example, consider an orientable (compact) surface, M, with a Riemannian metric. Pick any chart, (U, φ), and choose an orthonormal coframe of one-forms, (θ_1, θ_2), such that $\mathrm{Vol} = \theta_1 \wedge \theta_2$ on U. Then, we have

$$\begin{aligned} d\theta_1 &= a_1\theta_1 \wedge \theta_2 \\ d\theta_2 &= a_2\theta_1 \wedge \theta_2 \end{aligned}$$

for some functions, a_1, a_2, and we let

$$\omega_{12} = a_1\theta_1 + a_2\theta_2.$$

Clearly,

$$\begin{pmatrix} 0 & \omega_{12} \\ -\omega_{12} & 0 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} 0 & a_1\theta_1 + a_2\theta_2 \\ -(a_1\theta_1 + a_2\theta_2) & 0 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} d\theta_1 \\ d\theta_2 \end{pmatrix}$$

which shows that

$$\omega = \omega^* = \begin{pmatrix} 0 & \omega_{12} \\ -\omega_{12} & 0 \end{pmatrix}$$

corresponds to the Levi-Civita connection on M. Let $\Omega = d\omega - \omega \wedge \omega$, we see that

$$\Omega = \begin{pmatrix} 0 & d\omega_{12} \\ -d\omega_{12} & 0 \end{pmatrix}.$$

As M is oriented and as M has a metric, the transition functions are in $\mathrm{SO}(2)$. We easily check that

$$\begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \begin{pmatrix} 0 & d\omega_{12} \\ -d\omega_{12} & 0 \end{pmatrix} \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} = \begin{pmatrix} 0 & d\omega_{12} \\ -d\omega_{12} & 0 \end{pmatrix},$$

which shows that Ω is a global two-form called the *Gauss-Bonnet* 2-form of M. Then, there is a function, κ, the *Gaussian curvature of* M such that

$$d\omega_{12} = -\kappa \mathrm{Vol},$$

where Vol is the oriented volume form on M. The *Gauss-Bonnet Theorem* for orientable surfaces asserts that

$$\int_M d\omega_{12} = 2\pi\chi(M),$$

where $\chi(M)$ is the *Euler characteristic* of M.

Remark: The Levi-Civita connection induced by a Riemannian metric, g, can also be defined in terms of the Lie derivative of the metric, g. This is the approach followed in Petersen [122] (Chapter 2). If θ_X is the one-form given by

$$\theta_X = i_X g,$$

that is, $(i_X g)(Y) = g(X, Y)$ for all $X, Y \in \mathfrak{X}(M)$ and if $L_X g$ is the Lie derivative of the symmetric $(0, 2)$ tensor, g, defined so that

$$(L_X g)(Y, Z) = X(g(Y, Z)) - g(L_X Y, Z) - g(Y, L_X Z)$$

(see Proposition 8.18), then, it is proved in Petersen [122] (Chapter 2, Theorem 1) that the Levi-Civita connection is defined implicitly by the formula

$$2g(\nabla_X Y, Z) = (L_Y g)(X, Z) + (d\theta_Y)(X, Z).$$

We conclude this section with various useful facts about torsion-free or metric connections. First, there is a nice characterization for the Levi-Civita connection induced by a Riemannian manifold over a submanifold. The proof of the next proposition is left as an exercise.

Proposition 11.22. *Let M be any Riemannian manifold and let N be any submanifold of M equipped with the induced metric. If ∇^M and ∇^N are the Levi-Civita connections on M and N, respectively, induced by the metric on M, then for any two vector fields, X and Y in $\mathfrak{X}(M)$ with $X(p), Y(p) \in T_p N$, for all $p \in N$, we have*

$$\nabla^N_X Y = (\nabla^M_X Y)^{\|},$$

where $(\nabla^M_X Y)^{\|}(p)$ is the orthogonal projection of $\nabla^M_X Y(p)$ onto $T_p N$, for every $p \in N$.

In particular, if γ is a curve on a surface, $M \subseteq \mathbb{R}^3$, then a vector field, $X(t)$, along γ is parallel iff $X'(t)$ is normal to the tangent plane, $T_{\gamma(t)} M$.

For any manifold, M, and any connection, ∇, on M, if ∇ is torsion-free, then the Lie derivative of any $(p, 0)$-tensor can be expressed in terms of ∇ (see Proposition 8.18).

Proposition 11.23. *For every $(0, q)$-tensor, $S \in \Gamma(M, (T^*M)^{\otimes q})$, we have*

$$(L_X S)(X_1, \ldots, X_q) = X[S(X_1, \ldots, X_q)] + \sum_{i=1}^{q} S(X_1, \ldots, \nabla_X X_i, \ldots, X_q),$$

for all $X_1, \ldots, X_q, X \in \mathfrak{X}(M)$.

Proposition 11.23 is proved in Gallot, Hullin and Lafontaine [61] (Chapter 2, Proposition 2.61). Using Proposition 8.13 it is also possible to give a formula for $d\omega(X_0 \ldots, X_k)$ in terms of the ∇_{X_i}, where ω is any k-form, namely

$$d\omega(X_0 \ldots, X_k) = \sum_{i=0}^{k} (-1)^i \nabla_{X_i} \omega(X_1, \ldots, X_{i-1}, X_0, X_{i+1}, \ldots, X_k).$$

Again, the above formula in proved in Gallot, Hullin and Lafontaine [61] (Chapter 2, Proposition 2.61).

If ∇ is a metric connection, then we can say more about the parallel transport along a curve. Recall from Section 11.3, Definition 11.4, that a vector field, X, along a curve, γ, is parallel iff

$$\frac{DX}{dt} = 0.$$

The following proposition will be needed:

Proposition 11.24. *Given any Riemannian manifold, M, and any metric connection, ∇, on M, for every curve, $\gamma\colon [a,b] \to M$, on M, if X and Y are two vector fields along γ, then*

$$\frac{d}{dt}\langle X(t), Y(t)\rangle = \left\langle \frac{DX}{dt}, Y \right\rangle + \left\langle X, \frac{DY}{dt} \right\rangle.$$

Proof. (After John Milnor.) Using Proposition 11.13, we can pick some parallel vector fields, $Z_1, \ldots, Z_n$, along γ, such that $Z_1(a), \ldots, Z_n(a)$ form an orthogonal frame. Then, as in the proof of Proposition 11.12, in any chart, (U, φ), the vector fields X and Y along the portion of γ in U can be expressed as

$$X = \sum_{i=1}^{n} X_i(t)\frac{\partial}{\partial x_i}, \qquad Y = \sum_{i=1}^{n} Y_i(t)\frac{\partial}{\partial x_i},$$

and

$$\gamma'(t_0) = \sum_{i=1}^{n} \frac{du_i}{dt}\left(\frac{\partial}{\partial x_i}\right)_{\gamma(t_0)},$$

with $u_i = pr_i \circ \varphi \circ \gamma$. Let $\widetilde{X}$ and $\widetilde{Y}$ be two parallel vector fields along γ. As the vector fields, $\frac{\partial}{\partial x_i}$, can be extended over the whole space, M, as ∇ is a metric connection and as $\widetilde{X}$ and $\widetilde{Y}$ are parallel along γ, we get

$$d(\langle \widetilde{X}, \widetilde{Y}\rangle)(\gamma') = \gamma'[\langle \widetilde{X}, \widetilde{Y}\rangle] = \langle \nabla_{\gamma'}\widetilde{X}, \widetilde{Y}\rangle + \langle \widetilde{X}, \nabla_{\gamma'}\widetilde{Y}\rangle = 0.$$

So, $\langle \widetilde{X}, \widetilde{Y}\rangle$ is constant along the portion of γ in U. But then, $\langle \widetilde{X}, \widetilde{Y}\rangle$ is constant along γ. Applying this to the $Z_i(t)$, we see that $Z_1(t), \ldots, Z_n(t)$ is an orthogonal frame, for every $t \in [a,b]$. Then, we can write

$$X = \sum_{i=1} x_i Z_i, \qquad Y = \sum_{j=1} y_j Z_j,$$

where $x_i(t)$ and $y_i(t)$ are smooth real-valued functions. It follows that

$$\langle X(t), Y(t)\rangle = \sum_{i=1}^{n} x_i y_i$$

and that

$$\frac{DX}{dt} = \frac{dx_i}{dt}Z_i + x_i\frac{DZ_i}{dt} = \frac{dx_i}{dt}Z_i, \qquad \frac{DY}{dt} = \frac{dy_i}{dt}Z_i + y_i\frac{DZ_i}{dt} = \frac{dy_i}{dt}Z_i.$$

Therefore,

$$\left\langle \frac{DX}{dt}, Y\right\rangle + \left\langle X, \frac{DY}{dt}\right\rangle = \sum_{i=1}^{n}\left(\frac{dx_i}{dt}\,y_i + x_i\,\frac{dy_i}{dt}\right) = \frac{d}{dt}\,\langle X(t), Y(t)\rangle,$$

as claimed. □

Using Proposition 11.24 we get

Proposition 11.25. *Given any Riemannian manifold, M, and any metric connection, ∇, on M, for every curve, $\gamma\colon [a,b] \to M$, on M, if X and Y are two vector fields along γ that are parallel, then*

$$\langle X, Y\rangle = C,$$

for some constant, C. In particular, $\|X(t)\|$ is constant. Furthermore, the linear isomorphism, $P_\gamma\colon T_{\gamma(a)} \to T_{\gamma(b)}$, is an isometry.

Proof. From Proposition 11.24, we have

$$\frac{d}{dt}\,\langle X(t), Y(t)\rangle = \left\langle \frac{DX}{dt}, Y\right\rangle + \left\langle X, \frac{DY}{dt}\right\rangle.$$

As X and Y are parallel along γ, we have $DX/dt = 0$ and $DY/dt = 0$, so

$$\frac{d}{dt}\,\langle X(t), Y(t)\rangle = 0,$$

which shows that $\langle X(t), Y(t)\rangle$ is constant. Therefore, for all $v, w \in T_{\gamma(a)}$, if X and Y are the unique vector fields parallel along γ such that $X(a) = v$ and $Y(a) = w$ given by Proposition 11.13, we have

$$\langle P_\gamma(v), P_\gamma(w)\rangle = \langle X(b), Y(b)\rangle = \langle X(a), Y(a)\rangle = \langle u, v\rangle,$$

which proves that P_γ is an isometry. □

In particular, Proposition 11.25 shows that the holonomy group, $\mathrm{Hol}_p(\nabla)$, based at p, is a subgroup of $\mathbf{O}(n)$.

11.5 Duality between Vector Fields and Differential Forms and their Covariant Derivatives

If $(M, \langle -, - \rangle)$ is a Riemannian manifold, then the inner product, $\langle -, - \rangle_p$, on T_pM, establishes a canonical duality between T_pM and T_p^*M, as explained in Section 22.1. Namely, we have the isomorphism, $\flat \colon T_pM \to T_p^*M$, defined such that for every $u \in T_pM$, the linear form, $u^\flat \in T_p^*M$, is given by

$$u^\flat(v) = \langle u, v \rangle_p \qquad v \in T_pM.$$

The inverse isomorphism, $\sharp \colon T_p^*M \to T_pM$, is defined such that for every $\omega \in T_p^*M$, the vector, $\omega^\sharp$, is the unique vector in T_pM so that

$$\langle \omega^\sharp, v \rangle_p = \omega(v), \qquad v \in T_pM.$$

The isomorphisms $\flat$ and $\sharp$ induce isomorphisms between vector fields, $X \in \mathfrak{X}(M)$, and one-forms, $\omega \in \mathcal{A}^1(M)$: A vector field, $X \in \mathfrak{X}(M)$, yields the one-form, $X^\flat \in \mathcal{A}^1(M)$, given by

$$(X^\flat)_p = (X_p)^\flat$$

and a one-form, $\omega \in \mathcal{A}^1(M)$, yields the vector field, $\omega^\sharp \in \mathfrak{X}(M)$, given by

$$(\omega^\sharp)_p = (\omega_p)^\sharp,$$

so that

$$\omega_p(v) = \langle (\omega_p)^\sharp, v \rangle_p, \qquad v \in T_pM,\ p \in M.$$

In particular, for every smooth function, $f \in C^\infty(M)$, the vector field corresponding to the one-form, df, is the *gradient*, $\operatorname{grad} f$, of f. The gradient of f is uniquely determined by the condition

$$\langle (\operatorname{grad} f)_p, v \rangle_p = df_p(v), \qquad v \in T_pM,\ p \in M.$$

Recall from Proposition 11.5 that the covariant derivative, $\nabla_X \omega$, of any one-form, $\omega \in \mathcal{A}^1(M)$, is the one-form given by

$$(\nabla_X \omega)(Y) = X(\omega(Y)) - \omega(\nabla_X Y).$$

If ∇ is a metric connection, then the vector field, $(\nabla_X \omega)^\sharp$, corresponding to $\nabla_X \omega$ is nicely expressed in terms of $\omega^\sharp$: Indeed, we have

$$(\nabla_X \omega)^\sharp = \nabla_X \omega^\sharp.$$

The proof goes as follows:

$$\begin{aligned}
(\nabla_X \omega)(Y) &= X(\omega(Y)) - \omega(\nabla_X Y) \\
&= X(\langle \omega^\sharp, Y \rangle) - \langle \omega^\sharp, \nabla_X Y \rangle \\
&= \langle \nabla_X \omega^\sharp, Y \rangle + \langle \omega^\sharp, \nabla_X Y \rangle - \langle \omega^\sharp, \nabla_X Y \rangle \\
&= \langle \nabla_X \omega^\sharp, Y \rangle,
\end{aligned}$$

where we used the fact that the connection is compatible with the metric in the third line and so,

$$(\nabla_X \omega)^\sharp = \nabla_X \omega^\sharp,$$

as claimed.

11.6 Pontrjagin Classes and Chern Classes, a Glimpse

This section can be omitted at first reading. Its purpose is to introduce the reader to Pontrjagin Classes and Chern Classes which are fundamental invariants of real (resp. complex) vector bundles. We focus on motivations and intuitions and omit most proofs but we give precise pointers to the literature for proofs.

Given a real (resp. complex) rank n vector bundle, $\xi = (E, \pi, B, V)$, we know that locally, ξ "looks like" a trivial bundle, $U \times V$, for some open subset, U, of the base space, B, but globally, ξ can be very twisted and one of the main issues is to understand and quantify "how twisted" ξ really is. Now, we know that every vector bundle admit a connection, say ∇, and the curvature, R^∇, of this connection is some measure of the twisting of ξ. However, R^∇ depends on ∇, so curvature is not intrinsic to ξ, which is unsatisfactory as we seek invariants that depend only on ξ.

Pontrjagin, Stiefel and Chern (starting from the late 1930's) discovered that invariants with "good" properties could be defined if we took these invariants to belong to various cohomology groups associated with B. Such invariants are usually called *characteristic classes*. Roughly, there are two main methods for defining characteristic classes, one using topology and the other, due to Chern and Weil, using differential forms. A masterly exposition of these methods is given in the classic book by Milnor and Stasheff [111]. Amazingly, the method of Chern and Weil using differential forms is quite accessible for someone who has reasonably good knowledge of differential forms and de Rham cohomology as long as one is willing to gloss over various technical details.

As we said earlier, one of the problems with curvature is that is depends on a connection. The way to circumvent this difficuty rests on the simple, yet subtle observation that locally, given any two overlapping local trivializations $(U_\alpha, \varphi_\alpha)$ and (U_β, φ_β), the transformation rule for the curvature matrices Ω_α and Ω_β is

$$\Omega_\beta = g_{\alpha\beta} \Omega_\alpha g_{\alpha\beta}^{-1},$$

where $g_{\alpha\beta} \colon U_\alpha \cap U_\beta \to \mathrm{GL}(V)$ is the transition function. The matrices of two-forms, Ω_α, are local, but the stroke of genius is to glue them together to form a global form using *invariant polynomials*.

Indeed, the Ω_α are $n \times n$ matrices so, consider the algebra of polynomials, $\mathbb{R}[X_1, \ldots, X_{n^2}]$ (or $\mathbb{C}[X_1, \ldots, X_{n^2}]$ in the complex case) in n^2 variables, considered as the entries of an $n \times n$ matrix. It is more convenient to use the set of variables $\{X_{ij} \mid 1 \leq i, j \leq n\}$, and to let X be the $n \times n$ matrix $X = (X_{ij})$.

Definition 11.7. A polynomial, $P \in \mathbb{R}[\{X_{ij} \mid 1 \leq i, j \leq n\}]$ (or $P \in \mathbb{C}[\{X_{ij} \mid 1 \leq i, j \leq n\}]$) is *invariant* iff

$$P(AXA^{-1}) = P(X),$$

for all $A \in \mathrm{GL}(n, \mathbb{R})$ (resp. $A \in \mathrm{GL}(n, \mathbb{C})$). The *algebra of invariant polynomials* over $n \times n$ matrices is denoted by I_n.

Examples of invariant polynomials are, the trace, $\mathrm{tr}(X)$, and the determinant, $\det(X)$, of the matrix X. We will characterize shortly the algebra I_n.

Now comes the punch line: For any homogeneous invariant polynomial, $P \in I_n$, of degree k, we can substitute Ω_α for X, that is, substitute ω_{ij} for X_{ij}, and evaluate $P(\Omega_\alpha)$. This is because Ω is a matrix of two-forms and the wedge product is commutative for forms of even degree. Therefore, $P(\Omega_\alpha) \in \mathcal{A}^{2k}(U_\alpha)$. But, the formula for a change of trivialization yields

$$P(\Omega_\alpha) = P(g_{\alpha\beta}\Omega_\alpha g_{\alpha\beta}^{-1}) = P(\Omega_\beta),$$

so the forms $P(\Omega_\alpha)$ and $P(\Omega_\beta)$ agree on overlaps and thus, they define a *global form* denoted $P(R^\nabla) \in \mathcal{A}^{2k}(B)$.

Now, we know how to obtain global $2k$-forms, $P(R^\nabla) \in \mathcal{A}^{2k}(B)$, but they still seem to depend on the connection and how do they define a cohomology class? Both problems are settled thanks to the following Theorems:

Theorem 11.26. *For every real rank n vector bundle, ξ, for every connection, ∇, on ξ, for every invariant homogeneous polynomial, P, of degree k, the $2k$-form, $P(R^\nabla) \in \mathcal{A}^{2k}(B)$, is closed. If ξ is a complex vector bundle, then the $2k$-form, $P(R^\nabla) \in \mathcal{A}^{2k}(B; \mathbb{C})$, is closed.*

Theorem 11.26 implies that the $2k$-form, $P(R^\nabla) \in \mathcal{A}^{2k}(B)$, defines a cohomology class, $[P(R^\nabla)] \in H^{2k}_{\mathrm{DR}}(B)$. We will come back to the proof of Theorem 11.26 later.

Theorem 11.27. *For every real (resp. complex) rank n vector bundle, for every invariant homogeneous polynomial, P, of degree k, the cohomology class, $[P(R^\nabla)] \in H^{2k}_{\mathrm{DR}}(B)$ (resp. $[P(R^\nabla)] \in H^{2k}_{\mathrm{DR}}(B; \mathbb{C})$) is independent of the choice of the connection ∇.*

The cohomology class, $[P(R^\nabla)]$, which does not depend on ∇ is denoted $P(\xi)$ and is called the *characteristic class* of ξ corresponding to P.

The proof of Theorem 11.27 involves a kind of homotopy argument, see Madsen and Tornehave [101] (Lemma 18.2), Morita [115] (Proposition 5.28) or see Milnor and Stasheff [111] (Appendix C).

The upshot is that Theorems 11.26 and 11.27 give us a method for producing invariants of a vector bundle that somehow reflect how curved (or twisted) the bundle is. However, it appears that we need to consider infinitely many invariants. Fortunately, we can do better because the algebra, I_n, of invariant polynomials is finitely generated and in fact, has very nice sets of generators. For this, we recall the *elementary symmetric functions* in n variables.

Given n variables, $\lambda_1, \ldots, \lambda_n$, we can write

$$\prod_{i=1}^{n}(1 + t\lambda_i) = 1 + \sigma_1 t + \sigma_2 t^2 + \cdots + \sigma_n t^n,$$

where the σ_i are symmetric, homogeneous polynomials of degree i in $\lambda_1, \ldots, \lambda_n$ called *elementary symmetric functions* in n variables. For example,

$$\sigma_1 = \sum_{i=1}^{n} \lambda_i, \quad \sigma_1 = \sum_{1 \leq i < j \leq n} \lambda_i \lambda_j, \quad \sigma_n = \lambda_1 \cdots \lambda_n.$$

To be more precise, we write $\sigma_i(\lambda_1, \ldots, \lambda_n)$ instead of σ_i.

Given any $n \times n$ matrix, $X = (X_{ij})$, we define $\sigma_i(X)$ by the formula

$$\det(I + tX) = 1 + \sigma_1(X)t + \sigma_2(X)t^2 + \cdots + \sigma_n(X)t^n.$$

We claim that

$$\sigma_i(X) = \sigma_i(\lambda_1, \ldots, \lambda_n),$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of X. Indeed, $\lambda_1, \ldots, \lambda_n$ are the roots the the polynomial $\det(\lambda I - X) = 0$, and as

$$\det(\lambda I - X) = \prod_{i=1}^{n}(\lambda - \lambda_i) = \lambda^n + \sum_{i=1}^{n}(-1)^i \sigma_i(\lambda_1, \ldots, \lambda_n)\lambda^{n-i},$$

by factoring λ^n and replacing λ^{-1} by $-\lambda^{-1}$, we get

$$\det(I + (-\lambda^{-1})X) = 1 + \sum_{i=1}^{n} \sigma_i(\lambda_1, \ldots, \lambda_n)(-\lambda^{-1})^n,$$

which proves our claim.

Observe that

$$\sigma_1(X) = \mathrm{tr}(X), \qquad \sigma_n(X) = \det(X).$$

Also, $\sigma_k(X^\top) = \sigma_k(X)$, since $\det(I + tX) = \det((I + tX)^\top) = \det(I + tX^\top)$. It is not very difficult to prove the following theorem:

Theorem 11.28. *The algebra, I_n, of invariant polynomials in n^2 variables is generated by $\sigma_1(X), \ldots, \sigma_n(X)$, that is*

$$I_n \cong \mathbb{R}[\sigma_1(X), \ldots, \sigma_n(X)] \qquad (\textit{resp.} \quad I_n \cong \mathbb{C}[\sigma_1(X), \ldots, \sigma_n(X)]).$$

For a proof of Theorem 11.28, see Milnor and Stasheff [111] (Appendix C, Lemma 6), Madsen and Tornehave [101] (Appendix B) or Morita [115] (Theorem 5.26). The proof uses the fact that for every matrix, X, there is an upper-triangular matrix, T, and an invertible matrix, B, so that

$$X = BTB^{-1}.$$

Then, we can replace B by the matrix $\mathrm{diag}(\epsilon, \epsilon^2, \ldots, \epsilon^n)B$, where ϵ is very small, and make the off diagonal entries arbitrarily small. By continuity, it follows that $P(X)$ depends only on the diagonal entries of BTB^{-1}, that is, on the eigenvalues of X. So, $P(X)$ must be a symmetric function of these eigenvalues and the classical theory of symmetric functions completes the proof.

It turns out that there are situations where it is more convenient to use another set of generators instead of $\sigma_1, \ldots, \sigma_n$. Define $s_i(X)$ by

$$s_i(X) = \mathrm{tr}(X^i).$$

Of course,

$$s_i(X) = \lambda_1^i + \cdots + \lambda_n^i,$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of X. Now, the $\sigma_i(X)$ and $s_i(X)$ are related to each other by *Newton's formula*, namely:

$$s_i(X) - \sigma_1(X)s_{i-1}(X) + \sigma_2(X)s_{i-2}(X) + \cdots + (-1)^{i-1}\sigma_{i-1}(X)s_1(X) + (-1)^i i\sigma_i(X) = 0$$

with $1 \leq i \leq n$. A "cute" proof of the Newton formulae is obtained by computing the derivative of $\log(h(t))$, where

$$h(t) = \prod_{i=1}^{n}(1 + t\lambda_i) = 1 + \sigma_1 t + \sigma_2 t^2 + \cdots + \sigma_n t^n,$$

see Madsen and Tornehave [101] (Appendix B) or Morita [115] (Exercise 5.7).

Consequently, we can inductively compute s_i in terms of $\sigma_1, \ldots, \sigma_i$ and conversely, σ_i in terms of $s_1, \ldots, s_i$. For example,

$$s_1 = \sigma_1, \quad s_2 = \sigma_1^2 - 2\sigma_2, \quad s_3 = \sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3.$$

It follows that

$$I_n \cong \mathbb{R}[s_1(X), \ldots, s_n(X)] \qquad (\text{resp.} \quad I_n \cong \mathbb{C}[s_1(X), \ldots, s_n(X)]).$$

Using the above, we can give a simple proof of Theorem 11.26, using Theorem 11.28.

Proof. (Proof of Theorem 11.26). Since $s_1, \ldots, s_n$ generate I_n, it is enough to prove that $s_i(R^\nabla)$ is closed. We need to prove that $ds_i(R^\nabla) = 0$ and for this, it is enough to prove

it in every local trivialization, $(U_\alpha, \varphi_\alpha)$. To simplify notation, we write Ω for Ω_α. Now, $s_i(\Omega) = \mathrm{tr}(\Omega^i)$, so

$$ds_i(\Omega) = d\mathrm{tr}(\Omega^i) = \mathrm{tr}(d\Omega^i),$$

and we use Bianchi's identity (Proposition 11.10),

$$d\Omega = \omega \wedge \Omega - \Omega \wedge \omega.$$

We have

$$\begin{aligned}
d\Omega^i &= d\Omega \wedge \Omega^{i-1} + \Omega \wedge d\Omega \wedge \Omega^{i-2} + \cdots + \Omega^k \wedge d\Omega \wedge \Omega^{i-k-1} + \cdots + \Omega^{i-1} \wedge d\Omega \\
&= (\omega \wedge \Omega - \Omega \wedge \omega) \wedge \Omega^{i-1} + \Omega \wedge (\omega \wedge \Omega - \Omega \wedge \omega) \wedge \Omega^{i-2} \\
&\quad + \cdots + \Omega^k \wedge (\omega \wedge \Omega - \Omega \wedge \omega) \wedge \Omega^{i-k-1} + \Omega^{k+1} \wedge (\omega \wedge \Omega - \Omega \wedge \omega) \wedge \Omega^{i-k-2} \\
&\quad + \cdots + \Omega^{i-1} \wedge (\omega \wedge \Omega - \Omega \wedge \omega) \\
&= \omega \wedge \Omega^i - \Omega \wedge \omega \wedge \Omega^{i-1} + \Omega \wedge \omega \wedge \Omega^{i-1} - \Omega^2 \wedge \omega \wedge \Omega^{i-2} + \cdots + \\
&\quad \Omega^k \wedge \omega \wedge \Omega^{i-k} - \Omega^{k+1} \wedge \omega \wedge \Omega^{i-k-1} + \Omega^{k+1} \wedge \omega \wedge \Omega^{i-k-1} - \Omega^{k+2} \wedge \omega \wedge \Omega^{i-k-2} \\
&\quad + \cdots + \Omega^{i-1} \wedge \omega \wedge \Omega - \Omega^i \wedge \omega \\
&= \omega \wedge \Omega^i - \Omega^i \wedge \omega.
\end{aligned}$$

However, the entries in ω are one-forms, the entries in Ω are two-forms and since

$$\eta \wedge \theta = \theta \wedge \eta$$

for all $\eta \in \mathcal{A}^1(B)$ and all $\theta \in \mathcal{A}^2(B)$ and $\mathrm{tr}(XY) = \mathrm{tr}(YX)$ for all matrices X and Y with commuting entries, we get

$$\mathrm{tr}(d\Omega^i) = \mathrm{tr}(\omega \wedge \Omega^i - \Omega^i \wedge \omega) = \mathrm{tr}(\omega \wedge \Omega^i) - \mathrm{tr}(\Omega^i \wedge \omega) = 0,$$

as required. □

A more elegant proof (also using Bianchi's identity) can be found in Milnor and Stasheff [111] (Appendix C, page 296-298).

For real vector bundles, only invariant polynomials of even degrees matter.

Proposition 11.29. *If ξ is a real vector bundle, then for every homogeneous invariant polynomial, P, of odd degree, k, we have $P(\xi) = 0 \in H^{2k}_{\mathrm{DR}}(B)$.*

Proof. As $I_n \cong \mathbb{R}[s_1(X), \ldots, s_n(X)]$ and $s_i(X)$ is homogeneous of degree i, it is enough to prove Proposition 11.29 for $s_i(X)$ with i odd. By Theorem 11.27, we may assume that we pick a metric connection on ξ, so that Ω_α is skew-symmetric in every local trivialization. Then, Ω^i_α is also skew symmetric and

$$\mathrm{tr}(\Omega^i_\alpha) = 0,$$

since the diagonal entries of a real skew-symmetric matrix are all zero. It follows that $s_i(\Omega_\alpha) = \mathrm{tr}(\Omega^i_\alpha) = 0$. □

Proposition 11.29 implies that for a real vector bundle, ξ, non-zero characteristic classes can only live in the cohomology groups $H^{4k}_{\mathrm{DR}}(B)$ of dimension $4k$. This property is specific to real vector bundles and generally fails for complex vector bundles.

Before defining Pontrjagin and Chern classes, we state another important properties of the homology classes, $P(\xi)$:

Proposition 11.30. *If $\xi = (E, \pi, B, V)$ and $\xi' = (E', \pi', B', V)$ are real (resp. complex) vector bundles, for every bundle map*

$$\begin{array}{ccc} E & \xrightarrow{f_E} & E' \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\pi'} \\ B & \xrightarrow[f]{} & B', \end{array}$$

for every homogeneous invariant polynomial, P, of degree k, we have

$$P(\xi) = f^*(P(\xi')) \in H^{2k}_{\mathrm{DR}}(B) \qquad (\textit{resp.} \quad P(\xi) = f^*(P(\xi')) \in H^{2k}_{\mathrm{DR}}(B; \mathbb{C})).$$

In particular, for every smooth map, $f\colon N \to B$, we have

$$P(f^*\xi) = f^*(P(\xi)) \in H^{2k}_{\mathrm{DR}}(N) \qquad (\textit{resp.} \quad P(f^*\xi) = f^*(P(\xi)) \in H^{2k}_{\mathrm{DR}}(N; \mathbb{C})).$$

The above proposition implies that isomorphic vector bundles have identical characteristic classes. We finally define Pontrjagin classes and Chern classes.

Definition 11.8. If ξ be a real rank n vector bundle, then the k^{th} *Pontrjagin class* of ξ, denoted $p_k(\xi)$, where $1 \leq 2k \leq n$, is the cohomology class

$$p_k(\xi) = \left[\frac{1}{(2\pi)^{2k}}\,\sigma_{2k}(R^\nabla)\right] \in H^{4k}_{\mathrm{DR}}(B),$$

for any connection, ∇, on ξ.

If ξ be a complex rank n vector bundle, then the k^{th} *Chern class* of ξ, denoted $c_k(\xi)$, where $1 \leq k \leq n$, is the cohomology class

$$c_k(\xi) = \left[\left(\frac{-1}{2\pi i}\right)^k \sigma_k(R^\nabla)\right] \in H^{2k}_{\mathrm{DR}}(B),$$

for any connection, ∇, on ξ. We also set $p_0(\xi) = 1$ and $c_0(\xi) = 1$ in the complex case.

The strange coefficient in $p_k(\xi)$ is present so that our expression matches the topological definition of Pontrjagin classes. The equally strange coefficient in $c_k(\xi)$ is there to insure that $c_k(\xi)$ actually belongs to the *real* cohomology group $H^{2k}_{\mathrm{DR}}(B)$, as stated (from the definition

we can only claim that $c_k(\xi) \in H^{2k}_{\mathrm{DR}}(B;\mathbb{C})$). This requires a proof which can be found in Morita [115] (Proposition 5.30) or in Madsen and Tornehave [101] (Chapter 18). One can use the fact that every complex vector bundle admits a Hermitian connection. Locally, the curvature matrices are skew-Hermitian and this easily implies that the Chern classes are real since if Ω is skew-Hermitian, then $i\Omega$ is Hermitian. (Actually, the topological version of Chern classes shows that $c_k(\xi) \in H^{2k}(B;\mathbb{Z})$.)

If ξ is a real rank n vector bundle and n is odd, say $n = 2m+1$, then the "top" Pontrjagin class, $p_m(\xi)$, corresponds to $\sigma_{2m}(R^\nabla)$, which is not $\det(R^\nabla)$. However, if n is even, say $n = 2m$, then the "top" Pontrjagin class $p_m(\xi)$ corresponds to $\det(R^\nabla)$.

It is also useful to introduce the *Pontrjagin polynomial*, $p(\xi)(t) \in H^\bullet_{\mathrm{DR}}(B)[t]$, given by

$$p(\xi)(t) = \left[\det\left(I + \frac{t}{2\pi}\, R^\nabla\right)\right] = 1 + p_1(\xi)t + p_2(\xi)t^2 + \cdots + p_{\lfloor \frac{n}{2}\rfloor}(\xi)t^{\lfloor \frac{n}{2}\rfloor}$$

and the *Chern polynomial*, $c(\xi)(t) \in H^\bullet_{\mathrm{DR}}(B)[t]$, given by

$$c(\xi)(t) = \left[\det\left(I - \frac{t}{2\pi i}\, R^\nabla\right)\right] = 1 + c_1(\xi)t + c_2(\xi)t^2 + \cdots + c_n(\xi)t^n.$$

If a vector bundle is trivial, then all its Pontrjagin classes (or Chern classes) are zero for all $k \geq 1$. If ξ is the real tangent bundle, $\xi = TB$, of some manifold of dimension n, then the $\lfloor \frac{n}{4}\rfloor$ Pontrjagin classes of TB are denoted $p_1(B), \ldots, p_{\lfloor \frac{n}{4}\rfloor}(B)$.

For complex vector bundles, the manifold, B, is often the real manifold corresponding to a complex manifold. If B has complex dimension, n, then B has real dimension $2n$. In this case, the tangent bundle, TB, is a rank n complex vector bundle over the real manifold of dimension, $2n$, and thus, it has n Chern classes, denoted $c_1(B), \ldots, c_n(B)$. The determination of the Pontrjagin classes (or Chern classes) of a manifold is an important step for the study of the geometric/topological structure of the manifold. For example, it is possible to compute the Chern classes of complex projective space, $\mathbb{CP}^n$ (as a complex manifold).

The Pontrjagin classes of a real vector bundle, ξ, are related to the Chern classes of its complexification, $\xi_{\mathbb{C}} = \xi \otimes \epsilon^1_{\mathbb{C}}$ (where $\epsilon^1_{\mathbb{C}}$ is the trivial complex line bundle $B \times \mathbb{C}$).

Proposition 11.31. *For every real rank n vector bundle, $\xi = (E, \pi, B, V)$, if $\xi_{\mathbb{C}} = \xi \otimes \epsilon^1_{\mathbb{C}}$ is the complexification of ξ, then*

$$p_k(\xi) = (-1)^k c_{2k}(\xi_{\mathbb{C}}) \in H^{4k}_{\mathrm{DR}}(B) \qquad k \geq 0.$$

Basically, the reason why Proposition 11.31 holds is that

$$\frac{1}{(2\pi)^{2k}} = (-1)^k \left(\frac{-1}{2\pi i}\right)^{2k}$$

We conclude this section by stating a few more properties of Chern classes.

Proposition 11.32. *For every complex rank n vector bundle, ξ, the following properties hold:*

(1) If ξ has a Hermitian metric, then we have a canonical isomorphism, $\xi^* \cong \overline{\xi}$.

(2) The Chern classes of ξ, ξ^* and $\overline{\xi}$ satisfy:

$$c_k(\xi^*) = c_k(\overline{\xi}) = (-1)^k c_k(\xi).$$

(3) For any complex vector bundles, ξ and η,

$$c_k(\xi \oplus \eta) = \sum_{i=0}^{k} c_i(\xi) c_{k-i}(\eta)$$

or equivalently

$$c(\xi \oplus \eta)(t) = c(\xi)(t)c(\eta)(t)$$

and similarly for Pontrjagin classes when ξ and η are real vector bundles.

To prove (2) we can use the fact that ξ can be given a Hermitian metric. Then, we saw earlier that if ω is the connection matrix of ξ over U then $\overline{\omega} = -\omega^\top$ is the connection matrix of $\overline{\xi}$ over U. However, it is clear that $\sigma_k(-\Omega_\alpha^\top) = (-1)^k \sigma_k(\Omega_\alpha)$ and so, $c_k(\overline{\xi}) = (-1)^k c_k(\xi)$.

Remark: For a real vector bundle, ξ, it is easy to see that $(\xi_{\mathbb{C}})^* = (\xi^*)_{\mathbb{C}}$, which implies that $c_k((\xi_{\mathbb{C}})^*) = c_k(\xi_{\mathbb{C}})$ (as $\xi \cong \xi^*$) and (2) implies that $c_k(\xi_{\mathbb{C}}) = 0$ for k odd. This proves again that the Pontrjagin classes exit only in dimension $4k$.

A complex rank n vector bundle, ξ, can also be viewed as a rank $2n$ vector bundle, $\xi_{\mathbb{R}}$ and we have:

Proposition 11.33. *For every rank n complex vector bundle, ξ, if $p_k = p_k(\xi_{\mathbb{R}})$ and $c_k = c_k(\xi)$, then we have*

$$1 - p_1 + p_2 + \cdots + (-1)^n p_n = (1 + c_1 + c_2 + \cdots + c_n)(1 - c_1 + c_2 + \cdots + (-1)^n c_n).$$

11.7 Euler Classes and The Generalized Gauss-Bonnet Theorem

Let $\xi = (E, \pi, B, V)$ be a real vector bundle of rank $n = 2m$ and let ∇ be any metric connection on ξ. Then, if ξ is orientable (as defined in Section 7.4, see Definition 7.12 and the paragraph following it), it is possible to define a global form, $\mathrm{eu}(R^\nabla) \in \mathcal{A}^{2m}(B)$, which turns out to be closed. Furthermore, the cohomology class, $[\mathrm{eu}(R^\nabla)] \in H^{2m}_{\mathrm{DR}}(B)$, is independent of the choice of ∇. This cohomology class, denoted $e(\xi)$, is called the *Euler class* of ξ and has some very interesting properties. For example, $p_m(\xi) = e(\xi)^2$.

As ∇ is a metric connection, in a trivialization, $(U_\alpha, \varphi_\alpha)$, the curvature matrix, Ω_α, is a skew symmetric $2m \times 2m$ matrix of 2-forms. Therefore, we can substitute the 2-forms in Ω_α for the variables of the Pfaffian of degree m (see Section 22.20) and we obtain the $2m$-form, $\mathrm{Pf}(\Omega_\alpha) \in \mathcal{A}^{2m}(B)$. Now, as ξ is orientable, the transition functions take values in $\mathbf{SO}(2m)$, so by Proposition 11.9, since

$$\Omega_\beta = g_{\alpha\beta}\Omega_\alpha g_{\alpha\beta}^{-1},$$

we conclude from Proposition 22.38 (ii) that

$$\mathrm{Pf}(\Omega_\alpha) = \mathrm{Pf}(\Omega_\beta).$$

Therefore, the local $2m$-forms, $\mathrm{Pf}(\Omega_\alpha)$, patch and define a global form, $\mathrm{Pf}(R^\nabla) \in \mathcal{A}^{2m}(B)$.

The following propositions can be shown:

Proposition 11.34. *For every real, orientable, rank $2m$ vector bundle, ξ, for every metric connection, ∇, on ξ the $2m$-form, $\mathrm{Pf}(R^\nabla) \in \mathcal{A}^{2m}(B)$, is closed.*

Proposition 11.35. *For every real, orientable, rank $2m$ vector bundle, ξ, the cohomology class, $[\mathrm{Pf}(R^\nabla)] \in H^{2m}_{\mathrm{DR}}(B)$, is independent of the metric connection, ∇, on ξ.*

Proofs of Propositions 11.34 and 11.35 can be found in Madsen and Tornehave [101] (Chapter 19) or Milnor and Stasheff [111] (Appendix C) (also see Morita [115], Chapters 5 and 6).

Definition 11.9. Let $\xi = (E, \pi, B, V)$ be any real, orientable, rank $2m$ vector bundle. For any metric connection, ∇, on ξ the *Euler form* associated with ∇ is the closed form

$$\mathrm{eu}(R^\nabla) = \frac{1}{(2\pi)^n}\,\mathrm{Pf}(R^\nabla) \in \mathcal{A}^{2m}(B)$$

and the *Euler class* of ξ is the cohomology class,

$$e(\xi) = \big[\mathrm{eu}(R^\nabla)\big] \in H^{2m}_{\mathrm{DR}}(B),$$

which does not depend on ∇.

Some authors, including Madsen and Tornehave [101], have a negative sign in front of R^∇ in their definition of the Euler form, that is, they define $\mathrm{eu}(R^\nabla)$ by

$$\mathrm{eu}(R^\nabla) = \frac{1}{(2\pi)^n}\,\mathrm{Pf}(-R^\nabla).$$

However these authors use a Pfaffian with the opposite sign convention from ours and this Pfaffian differs from ours by the factor $(-1)^n$ (see the warning in Section 22.20). Madsen and Tornehave [101] seem to have overlooked this point and with their definition of the Pfaffian (which is the one we have adopted) Proposition 11.37 is incorrect.

Here is the relationship between the Euler class, $e(\xi)$, and the top Pontrjagin class, $p_m(\xi)$:

Proposition 11.36. *For every real, orientable, rank $2m$ vector bundle, $\xi = (E, \pi, B, V)$, we have*

$$p_m(\xi) = e(\xi)^2 \in H^{4m}_{\mathrm{DR}}(B).$$

Proof. The top Pontrjagin class, $p_m(\xi)$, is given by

$$p_m(\xi) = \left[\frac{1}{(2\pi)^{2m}}\,\det(R^\nabla)\right],$$

for any (metric) connection, ∇ and

$$e(\xi) = \left[\mathrm{eu}(R^\nabla)\right]$$

with

$$\mathrm{eu}(R^\nabla) = \frac{1}{(2\pi)^n}\,\mathrm{Pf}(R^\nabla).$$

From Proposition 22.38 (i), we have

$$\det(R^\nabla) = \mathrm{Pf}(R^\nabla)^2,$$

which yields the desired result. □

A rank m complex vector bundle, $\xi = (E, \pi, B, V)$, can be viewed as a real rank $2m$ vector bundle, $\xi_{\mathbb{R}}$, by viewing V as a $2m$ dimensional real vector space. Then, it turns out that $\xi_{\mathbb{R}}$ is naturally orientable. Here is the reason.

For any basis, $(e_1, \ldots, e_m)$, of V over $\mathbb{C}$, observe that $(e_1, ie_1, \ldots, e_m, ie_m)$ is a basis of V over $\mathbb{R}$ (since $v = \sum_{i=1}^m (\lambda_i + i\mu_i)e_i = \sum_{i=1}^m \lambda_i e_i + \sum_{i=1}^m \mu_i ie_i$). But, any $m \times m$ invertible matrix, A, over $\mathbb{C}$ becomes a real $2m \times 2m$ invertible matrix, $A_{\mathbb{R}}$, obtained by replacing the entry $a_{jk} + ib_{jk}$ in A by the real 2×2 matrix

$$\begin{pmatrix} a_{jk} & -b_{jk} \\ b_{jk} & a_{jk}. \end{pmatrix}$$

Indeed, if $v_k = \sum_{j=1}^m a_{jk}e_j + \sum_{j=1}^m b_{jk}ie_j$, then $iv_k = \sum_{j=1}^m -b_{jk}e_j + \sum_{j=1}^m a_{jk}ie_j$ and when we express v_k and iv_k over the basis $(e_1, ie_1, \ldots, e_m, ie_m)$, we get a matrix $A_{\mathbb{R}}$ consisting of 2×2 blocks as above. Clearly, the map $r\colon A \mapsto A_{\mathbb{R}}$ is a continuous injective homomorphism from $\mathrm{GL}(m, \mathbb{C})$ to $\mathrm{GL}(2m, \mathbb{R})$. Now, it is known $\mathrm{GL}(m, \mathbb{C})$ is connected, thus $\mathrm{Im}(r) = r(\mathrm{GL}(m, \mathbb{C}))$ is connected and as $\det(I_{2m}) = 1$, we conclude that all matrices in $\mathrm{Im}(r)$ have positive determinant.[1] Therefore, the transition functions of $\xi_{\mathbb{R}}$ which take values in $\mathrm{Im}(r)$ have positive determinant and $\xi_{\mathbb{R}}$ is orientable. We can give $\xi_{\mathbb{R}}$ an orientation by fixing some basis of V over $\mathbb{R}$. Then, we have the following relationship between $e(\xi_{\mathbb{R}})$ and the top Chern class, $c_m(\xi)$:

[1] One can also prove directly that every matrix in $\mathrm{Im}(r)$ has positive determinant by expressing $r(A)$ as a product of simple matrices whose determinants are easily computed.

Proposition 11.37. *For every complex, rank m vector bundle, $\xi = (E, \pi, B, V)$, we have*

$$c_m(\xi) = e(\xi) \in H^{2m}_{\mathrm{DR}}(B).$$

Proof. Pick some metric connection, ∇. Recall that

$$c_m(\xi) = \left[\left(\frac{-1}{2\pi i}\right)^m \det(R^\nabla)\right] = i^m \left[\left(\frac{1}{2\pi}\right)^m \det(R^\nabla)\right].$$

On the other hand,

$$e(\xi) = \left[\frac{1}{(2\pi)^m} \operatorname{Pf}(R^\nabla_{\mathbb{R}})\right].$$

Here, $R^\nabla_{\mathbb{R}}$ denotes the global $2m$-form wich, locally, is equal to $\Omega_{\mathbb{R}}$, where Ω is the $m \times m$ curvature matrix of ξ over some trivialization. By Proposition 22.39,

$$\operatorname{Pf}(\Omega_{\mathbb{R}}) = i^n \det(\Omega),$$

so $c_m(\xi) = e(\xi)$, as claimed. □

The Euler class enjoys many other nice properties. For example, if $f\colon \xi_1 \to \xi_2$ is an orientation preserving bundle map, then

$$e(f^*\xi_2) = f^*(e(\xi_2)),$$

where $f^*\xi_2$ is given the orientation induced by ξ_2. Also, the Euler class can be defined by topological means and it belongs to the integral cohomology group $H^{2m}(B; \mathbb{Z})$.

Although this result lies beyond the scope of these notes we cannot resist stating one of the most important and most beautiful theorems of differential geometry usually called the *Generalized Gauss-Bonnet Theorem* or *Gauss-Bonnet-Chern Theorem*.

For this, we need the notion of Euler characteristic. Since we haven't discussed triangulations of manifolds, we will use a defintion in terms of cohomology. Although concise, this definition is hard to motivate and we appologize for this. Given a smooth n-dimensional manifold, M, we define its *Euler characteristic*, $\chi(M)$, as

$$\chi(M) = \sum_{i=0}^{n} (-1)^i \dim(H^i_{\mathrm{DR}}).$$

The integers, $b_i = \dim(H^i_{\mathrm{DR}})$, are known as the *Betti numbers* of M. For example, $\chi(S^2) = 2$.

It turns out that if M is an odd dimensional manifold, then $\chi(M) = 0$. This explains partially why the Euler class is only defined for even dimensional bundles.

The Generalized Gauss-Bonnet Theorem (or Gauss-Bonnet-Chern Theorem) is a generalization of the Gauss-Bonnet Theorem for surfaces. In the general form stated below it was first proved by Allendoerfer and Weil (1943), and Chern (1944).

Theorem 11.38. *(Generalized Gauss-Bonnet Formula) Let M be an orientable, smooth, compact manifold of dimension $2m$. For every metric connection, ∇, on TM, (in particular, the Levi-Civita connection for a Riemannian manifold) we have*

$$\int_M \mathrm{eu}(R^\nabla) = \chi(M).$$

A proof of Theorem 11.38 can be found in Madsen and Tornehave [101] (Chapter 21), but beware of some sign problems. The proof uses another famous theorem of differential topology, the *Poincaré-Hopf Theorem*. A sketch of the proof is also given in Morita [115], Chapter 5.

Theorem 11.38 is remarkable because it establishes a relationship between the geometry of the manifold (its curvature) and the topology of the manifold (the number of "holes"), somehow encoded in its Euler characteristic.

Characteristic classes are a rich and important topic and we've only scratched the surface. We refer the reader to the texts mentioned earlier in this section as well as to Bott and Tu [19] for comprehensive expositions.

Chapter 12

Geodesics on Riemannian Manifolds

12.1 Geodesics, Local Existence and Uniqueness

If (M, g) is a Riemannian manifold, then the concept of *length* makes sense for any piecewise smooth (in fact, C^1) curve on M. Then, it possible to define the structure of a metric space on M, where $d(p, q)$ is the greatest lower bound of the length of all curves joining p and q. Curves on M which locally yield the shortest distance between two points are of great interest. These curves called *geodesics* play an important role and the goal of this chapter is to study some of their properties. Since geodesics are a standard chapter of every differential geometry text, we will omit most proofs and instead give precise pointers to the literature. Among the many presentations of this subject, in our opinion, Milnor's account [107] (Part II, Section 11) is still one of the best, certainly by its clarity and elegance. We acknowledge that our presentation was heavily inspired by this beautiful work. We also relied heavily on Gallot, Hulin and Lafontaine [61] (Chapter 2), Do Carmo [51], O'Neill [120], Kuhnel [92] and class notes by Pierre Pansu (see http://www.math.u-psud.fr/%7Epansu/web_dea/resume_dea_04.html in http://www.math.u-psud.fr~pansu/). Another reference that is remarkable by its clarity and the completeness of its coverage is Postnikov [126].

Given any $p \in M$, for every $v \in T_pM$, the *(Riemannian) norm* of v, denoted $\|v\|$, is defined by

$$\|v\| = \sqrt{g_p(v, v)}.$$

The Riemannian inner product, $g_p(u, v)$, of two tangent vectors, $u, v \in T_pM$, will also be denoted by $\langle u, v\rangle_p$, or simply $\langle u, v\rangle$. Recall the following definitions regarding curves:

Definition 12.1. Given any Riemannian manifold, M, a *smooth parametric curve* (for short, *curve*) on M is a map, $\gamma\colon I \to M$, where I is some open interval of $\mathbb{R}$. For a closed interval, $[a, b] \subseteq \mathbb{R}$, a map $\gamma\colon [a, b] \to M$ is a *smooth curve from $p = \gamma(a)$ to $q = \gamma(b)$* iff γ can be extended to a smooth curve $\widetilde{\gamma}\colon (a - \epsilon, b + \epsilon) \to M$, for some $\epsilon > 0$. Given any two points, $p, q \in M$, a continuous map, $\gamma\colon [a, b] \to M$, is a *piecewise smooth curve from p to q* iff

(1) There is a sequence $a = t_0 < t_1 < \cdots < t_{k-1} < t_k = b$ of numbers, $t_i \in \mathbb{R}$, so that each map, $\gamma_i = \gamma \restriction [t_i, t_{i+1}]$, called a *curve segment* is a smooth curve for $i = 0, \ldots, k-1$.

(2) $\gamma(a) = p$ and $\gamma(b) = q$.

The set of all piecewise smooth curves from p to q is denoted by $\Omega(M; p, q)$ or briefly by $\Omega(p, q)$ (or even by Ω, when p and q are understood).

The set $\Omega(M; p, q)$ is an important object sometimes called the *path space* of M (from p to q). Unfortunately it is an infinite-dimensional manifold, which makes it hard to investigate its properties.

Observe that at any junction point, $\gamma_{i-1}(t_i) = \gamma_i(t_i)$, there may be a jump in the velocity vector of γ. We let $\gamma'((t_i)_+) = \gamma_i'(t_i)$ and $\gamma'((t_i)_-) = \gamma_{i-1}'(t_i)$.

Given any curve, $\gamma \in \Omega(M; p, q)$, the *length*, $L(\gamma)$, of γ is defined by

$$L(\gamma) = \sum_{i=0}^{k-1} \int_{t_i}^{t_{i+1}} \|\gamma'(t)\| \, dt = \sum_{i=0}^{k-1} \int_{t_i}^{t_{i+1}} \sqrt{g(\gamma'(t), \gamma'(t))}\, dt.$$

It is easy to see that $L(\gamma)$ is unchanged by a monotone reparametrization (that is, a map $h\colon [a, b] \to [c, d]$, whose derivative, h', has a constant sign).

Let us now assume that our Riemannian manifold, (M, g), is equipped with the Levi-Civita connection and thus, for every curve, γ, on M, let $\frac{D}{dt}$ be the associated covariant derivative along γ, also denoted $\nabla_{\gamma'}$

Definition 12.2. Let (M, g) be a Riemannian manifold. A curve, $\gamma\colon I \to M$, (where $I \subseteq \mathbb{R}$ is any interval) is a *geodesic* iff $\gamma'(t)$ is parallel along γ, that is, iff

$$\frac{D\gamma'}{dt} = \nabla_{\gamma'}\gamma' = 0.$$

If M was embedded in $\mathbb{R}^d$, a geodesic would be a curve, γ, such that the acceleration vector, $\gamma'' = \frac{D\gamma'}{dt}$, is normal to $T_{\gamma(t)}M$.

By Proposition 11.25, $\|\gamma'(t)\| = \sqrt{g(\gamma'(t), \gamma'(t))}$ is constant, say $\|\gamma'(t)\| = c$. If we define the *arc-length* function, $s(t)$, relative to a, where a is any chosen point in I, by

$$s(t) = \int_a^t \sqrt{g(\gamma'(t), \gamma'(t))}\, dt = c(t - a), \qquad t \in I,$$

we conclude that for a geodesic, $\gamma(t)$, the parameter, t, is an affine function of the arc-length. When $c = 1$, which can be achieved by an affine reparametrization, we say that the geodesic is *normalized*.

The geodesics in $\mathbb{R}^n$ are the straight lines parametrized by constant velocity. The geodesics of the 2-sphere are the great circles, parametrized by arc-length. The geodesics

of the Poincaré half-plane are the lines $x = a$ and the half-circles centered on the x-axis. The geodesics of an ellipsoid are quite fascinating. They can be completely characterized and they are parametrized by elliptic functions (see Hilbert and Cohn-Vossen [76], Chapter 4, Section and Berger and Gostiaux [17], Section 10.4.9.5). If M is a submanifold of $\mathbb{R}^n$, geodesics are curves whose acceleration vector, $\gamma'' = (D\gamma')/dt$ is normal to M (that is, for every $p \in M$, γ'' is normal to T_pM).

In a local chart, (U, φ), since a geodesic is characterized by the fact that its velocity vector field, $\gamma'(t)$, along γ is parallel, by Proposition 11.13, it is the solution of the following system of second-order ODE's in the unknowns, u_k:

$$\frac{d^2u_k}{dt^2} + \sum_{ij} \Gamma^k_{ij} \frac{du_i}{dt} \frac{du_j}{dt} = 0, \qquad k = 1, \ldots, n,$$

with $u_i = pr_i \circ \varphi \circ \gamma$ ($n = \dim(M)$).

The standard existence and uniqueness results for ODE's can be used to prove the following proposition (see O'Neill [120], Chapter 3):

Proposition 12.1. *Let (M, g) be a Riemannian manifold. For every point, $p \in M$, and every tangent vector, $v \in T_pM$, there is some interval, $(-\eta, \eta)$, and a unique geodesic,*

$$\gamma_v \colon (-\eta, \eta) \to M,$$

satisfying the conditions

$$\gamma_v(0) = p, \qquad \gamma_v'(0) = v.$$

The following proposition is used to prove that every geodesic is contained in a unique maximal geodesic (*i.e*, with largest possible domain). For a proof, see O'Neill [120], Chapter 3 or Petersen [122] (Chapter 5, Section 2, Lemma 7).

Proposition 12.2. *For any two geodesics, $\gamma_1 \colon I_1 \to M$ and $\gamma_2 \colon I_2 \to M$, if $\gamma_1(a) = \gamma_2(a)$ and $\gamma_1'(a) = \gamma_2'(a)$, for some $a \in I_1 \cap I_2$, then $\gamma_1 = \gamma_2$ on $I_1 \cap I_2$.*

Propositions 12.1 and 12.2 imply that for every $p \in M$ and every $v \in T_pM$, there is a unique geodesic, denoted γ_v, such that $\gamma(0) = p$, $\gamma'(0) = v$, and the domain of γ is the largest possible, that is, cannot be extended. We call γ_v a *maximal geodesic* (with initial conditions $\gamma_v(0) = p$ and $\gamma_v'(0) = v$).

Observe that the system of differential equations satisfied by geodesics has the following homogeneity property: If $t \mapsto \gamma(t)$ is a solution of the above system, then for every constant, c, the curve $t \mapsto \gamma(ct)$ is also a solution of the system. We can use this fact together with standard existence and uniqueness results for ODE's to prove the proposition below. For proofs, see Milnor [107] (Part II, Section 10), or Gallot, Hulin and Lafontaine [61] (Chapter 2).

Proposition 12.3. *Let (M, g) be a Riemannian manifold. For every point, $p_0 \in M$, there is an open subset, $U \subseteq M$, with $p_0 \in U$, and some $\epsilon > 0$, so that: For every $p \in U$ and every tangent vector, $v \in T_pM$, with $\|v\| < \epsilon$, there is a unique geodesic,*

$$\gamma_v \colon (-2, 2) \to M,$$

satisfying the conditions

$$\gamma_v(0) = p, \qquad \gamma_v'(0) = v.$$

If $\gamma_v \colon (-\eta, \eta) \to M$ is a geodesic with initial conditions $\gamma_v(0) = p$ and $\gamma_v'(0) = v \neq 0$, for any constant, $c \neq 0$, the curve, $t \mapsto \gamma_v(ct)$, is a geodesic defined on $(-\eta/c, \eta/c)$ (or $(\eta/c, -\eta/c)$ if $c < 0$) such that $\gamma'(0) = cv$. Thus,

$$\gamma_v(ct) = \gamma_{cv}(t), \qquad ct \in (-\eta, \eta).$$

This fact will be used in the next section.

Given any function, $f \in C^\infty(M)$, for any $p \in M$ and for any $u \in T_pM$, the value of the Hessian, $\mathrm{Hess}_p(f)(u, u)$, can be computed using geodesics. Indeed, for any geodesic, $\gamma \colon [0, \epsilon] \to M$, such that $\gamma(0) = p$ and $\gamma'(0) = u$, we have

$$\mathrm{Hess}_p(u, u) = \gamma'(\gamma'(f)) - (\nabla_{\gamma'}\gamma')(f) = \gamma'(\gamma'(f))$$

since $\nabla_{\gamma'}\gamma' = 0$ because γ is a geodesic and

$$\gamma'(\gamma'(f)) = \gamma'(df(\gamma')) = \gamma'\left(\frac{d}{dt} f(\gamma(t))\Big|_{t=0}\right) = \frac{d^2}{dt^2} f(\gamma(t))\Big|_{t=0},$$

and thus,

$$\mathrm{Hess}_p(u, u) = \frac{d^2}{dt^2} f(\gamma(t))\Big|_{t=0}.$$

12.2 The Exponential Map

The idea behind the exponential map is to parametrize a Riemannian manifold, M, locally near any $p \in M$ in terms of a map from the tangent space T_pM to the manifold, this map being defined in terms of geodesics.

Definition 12.3. Let (M, g) be a Riemannian manifold. For every $p \in M$, let $\mathcal{D}(p)$ (or simply, $\mathcal{D}$) be the open subset of T_pM given by

$$\mathcal{D}(p) = \{v \in T_pM \mid \gamma_v(1) \quad \text{is defined}\},$$

where γ_v is the unique maximal geodesic with initial conditions $\gamma_v(0) = p$ and $\gamma_v'(0) = v$. The *exponential map* is the map, $\exp_p \colon \mathcal{D}(p) \to M$, given by

$$\exp_p(v) = \gamma_v(1).$$

It is easy to see that $\mathcal{D}(p)$ is *star-shaped*, which means that if $w \in \mathcal{D}(p)$, then the line segment $\{tw \mid 0 \leq t \leq 1\}$ is contained in $\mathcal{D}(p)$. In view of the remark made at the end of the previous section, the curve

$$t \mapsto \exp_p(tv), \qquad tv \in \mathcal{D}(p)$$

is the geodesic, γ_v, through p such that $\gamma_v'(0) = v$. Such geodesics are called *radial geodesics*. The point, $\exp_p(tv)$, is obtained by running along the geodesic, γ_v, an arc length equal to $t\,\|v\|$, starting from p.

In general, $\mathcal{D}(p)$ is a proper subset of T_pM. For example, if U is a bounded open subset of $\mathbb{R}^n$, since we can identify T_pU with $\mathbb{R}^n$ for all $p \in U$, then $\mathcal{D}(p) \subseteq U$, for all $p \in U$.

Definition 12.4. A Riemannian manifold, (M, g), is *geodesically complete* iff $\mathcal{D}(p) = T_pM$, for all $p \in M$, that is, iff the exponential, $\exp_p(v)$, is defined for all $p \in M$ and for all $v \in T_pM$.

Equivalently, (M, g) is geodesically complete iff every geodesic can be extended indefinitely. Geodesically complete manifolds have nice properties, some of which will be investigated later.

Observe that $d(\exp_p)_0 = \mathrm{id}_{T_pM}$. This is because, for every $v \in \mathcal{D}(p)$, the map $t \mapsto \exp_p(tv)$ is the geodesic, γ_v, and

$$\frac{d}{dt}(\gamma_v(t))|_{t=0} = v = \frac{d}{dt}(\exp_p(tv))|_{t=0} = d(\exp_p)_0(v).$$

It follows from the inverse function theorem that $\exp_p$ is a diffeomorphism from some open ball in T_pM centered at 0 to M. The following slightly stronger proposition can be shown (Milnor [107], Chapter 10, Lemma 10.3):

Proposition 12.4. *Let (M, g) be a Riemannian manifold. For every point, $p \in M$, there is an open subset, $W \subseteq M$, with $p \in W$ and a number $\epsilon > 0$, so that*

(1) Any two points q_1, q_2 of W are joined by a unique geodesic of length $< \epsilon$.

(2) This geodesic depends smoothly upon q_1 and q_2, that is, if $t \mapsto \exp_{q_1}(tv)$ is the geodesic joining q_1 and q_2 $(0 \leq t \leq 1)$, then $v \in T_{q_1}M$ depends smoothly on (q_1, q_2).

(3) For every $q \in W$, the map $\exp_q$ is a diffeomorphism from the open ball, $B(0, \epsilon) \subseteq T_qM$, to its image, $U_q = \exp_q(B(0, \epsilon)) \subseteq M$, with $W \subseteq U_q$ and U_q open.

For any $q \in M$, an open neighborhood of q of the form, $U_q = \exp_q(B(0, \epsilon))$, where $\exp_q$ is a diffeomorphism from the open ball $B(0, \epsilon)$ onto U_q, is called a *normal neighborhood*.

Definition 12.5. Let (M, g) be a Riemannian manifold. For every point, $p \in M$, the *injectivity radius of M at p*, denoted $i(p)$, is the least upper bound of the numbers, $r > 0$, such that $\exp_p$ is a diffeomorphism on the open ball $B(0, r) \subseteq T_pM$. The *injectivity radius, $i(M)$, of M* is the greatest lower bound of the numbers, $i(p)$, where $p \in M$.

For every $p \in M$, we get a chart, (U_p, φ), where $U_p = \exp_p(B(0, i(p)))$ and $\varphi = \exp^{-1}$, called a *normal chart*. If we pick any orthonormal basis, $(e_1, \ldots, e_n)$, of T_pM, then the x_i's, with $x_i = pr_i \circ \exp^{-1}$ and pr_i the projection onto $\mathbb{R}e_i$, are called *normal coordinates* at p (here, $n = \dim(M)$). These are defined up to an isometry of T_pM. The following proposition shows that Riemannian metrics do not admit any local invariants of order one. The proof is left as an exercise.

Proposition 12.5. *Let (M, g) be a Riemannian manifold. For every point, $p \in M$, in normal coordinates at p,*

$$g\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right)_p = \delta_{ij} \qquad \text{and} \qquad \Gamma^k_{ij}(p) = 0.$$

For the next proposition, known as *Gauss Lemma*, we need to define *polar coordinates* on T_pM. If $n = \dim(M)$, observe that the map, $(0, \infty) \times S^{n-1} \longrightarrow T_pM - \{0\}$, given by

$$(r, v) \mapsto rv, \qquad r > 0,\ v \in S^{n-1}$$

is a diffeomorphism, where S^{n-1} is the sphere of radius $r = 1$ in T_pM. Then, the map, $f\colon (0, i(p)) \times S^{n-1} \to U_p - \{p\}$, given by

$$(r, v) \mapsto \exp_p(rv), \qquad 0 < r < i(p),\ v \in S^{n-1}$$

is also a diffeomorphism.

Proposition 12.6. *(Gauss Lemma) Let (M, g) be a Riemannian manifold. For every point, $p \in M$, the images, $\exp_p(S(0, r))$, of the spheres, $S(0, r) \subseteq T_pM$, centered at 0 by the exponential map, $\exp_p$, are orthogonal to the radial geodesics, $r \mapsto \exp_p(rv)$, through p, for all $r < i(p)$. Furthermore, in polar coordinates, the pull-back metric, $\exp^* g$, induced on T_pM is of the form*

$$exp^* g = dr^2 + g_r,$$

where g_r is a metric on the unit sphere, S^{n-1}, with the property that g_r/r^2 converges to the standard metric on S^{n-1} (induced by $\mathbb{R}^n$) when r goes to zero (here, $n = \dim(M)$).

Proof sketch. after Milnor, see [107], Chapter II, Section 10. Pick any curve, $t \mapsto v(t)$ on the unit sphere, S^{n-1}. We must show that the corresponding curve on M,

$$t \mapsto \exp_p(rv(t)),$$

with r fixed, is orthogonal to the radial geodesic,

$$r \mapsto \exp_p(rv(t)),$$

with t fixed, $0 \leq r < i(p)$. In terms of the parametrized surface,

$$f(r, t) = \exp_p(rv(t)),$$

we must prove that

$$\left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle = 0,$$

for all (r, t). However, as we are using the Levi-Civita connection which is compatible with the metric, we have

$$\frac{\partial}{\partial r} \left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle = \left\langle \frac{D}{\partial r} \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle + \left\langle \frac{\partial f}{\partial r}, \frac{D}{\partial r} \frac{\partial f}{\partial t} \right\rangle.$$

The first expression on the right is zero since the curves

$$t \mapsto f(r, t)$$

are geodesics. For the second expression, we have

$$\left\langle \frac{\partial f}{\partial r}, \frac{D}{\partial r} \frac{\partial f}{\partial t} \right\rangle = \frac{1}{2} \frac{\partial}{\partial t} \left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial r} \right\rangle = 0,$$

since $1 = \|v(t)\| = \|\partial f / \partial r\|$. Therefore,

$$\left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle$$

is independent of r. But, for $r = 0$, we have

$$f(0, t) = \exp_p(0) = p,$$

hence

$$\partial f / \partial t(0, t) = 0$$

and thus,

$$\left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle = 0$$

for all r, t, which concludes the proof of the first statement. For the proof of the second statement, see Pansu's class notes, Chapter 3, Section 3.5. □

Consider any piecewise smooth curve

$$\omega \colon [a, b] \to U_p - \{p\}.$$

We can write each point $\omega(t)$ uniquely as

$$\omega(t) = \exp_p(r(t)v(t)),$$

with $0 < r(t) < i(p)$, $v(t) \in T_pM$ and $\|v(t)\| = 1$.

Proposition 12.7. *Let (M, g) be a Riemannian manifold. We have*

$$\int_a^b \|\omega'(t)\|\, dt \geq |r(b) - r(a)|,$$

where equality holds only if the function r is monotone and the function v is constant. Thus, the shortest path joining two concentric spherical shells, $\exp_p(S(0, r_1))$ and $\exp_p(S(0, r_2))$, is a radial geodesic.

Proof. (After Milnor, see [107], Chapter II, Section 10.) Again, let $f(r, t) = \exp_p(rv(t))$, so that $\omega(t) = f(r(t), t)$. Then,

$$\frac{d\omega}{dt} = \frac{\partial f}{\partial r}\, r'(t) + \frac{\partial f}{\partial t}.$$

The proof of the previous proposition showed that the two vectors on the right-hand side are orthogonal and since $\|\partial f/\partial r\| = 1$, this gives

$$\left\|\frac{d\omega}{dt}\right\|^2 = |r'(t)|^2 + \left\|\frac{\partial f}{\partial t}\right\|^2 \geq |r'(t)|^2$$

where equality holds only if $\partial f/\partial t = 0$; hence only if $v'(t) = 0$. Thus,

$$\int_a^b \left\|\frac{d\omega}{dt}\right\| dt \geq \int_a^b |r'(t)| dt \geq |r(b) - r(a)|$$

where equality holds only if $r(t)$ is monotone and $v(t)$ is constant. □

We now get the following important result from Proposition 12.6 and Proposition 12.7:

Theorem 12.8. *Let (M, g) be a Riemannian manifold. Let W and ϵ be as in Proposition 12.4 and let $\gamma\colon [0, 1] \to M$ be the geodesic of length $< \epsilon$ joining two points q_1, q_2 of W. For any other piecewise smooth path, ω, joining q_1 and q_2, we have*

$$\int_0^1 \|\gamma'(t)\|\, dt \leq \int_0^1 \|\omega'(t)\|\, dt$$

where equality can holds only if the images $\omega([0, 1])$ and $\gamma([0, 1])$ coincide. Thus, γ is the shortest path from q_1 to q_2.

Proof. (After Milnor, see [107], Chapter II, Section 10.) Consider any piecewise smooth path, ω, from $q_1 = \gamma(0)$ to some point

$$q_2 = \exp_{q_1}(rv) \in U_{q_1},$$

where $0 < r < \epsilon$ and $\|v\| = 1$. Then, for any δ with $0 < \delta < r$, the path ω must contain a segment joining the spherical shell of radius δ to the spherical shell of radius r, and lying between these two shells. The length of this segment will be at least $r - \delta$; hence if we let δ go to zero, the length of ω will be at least r. If $\omega([0, 1]) \neq \gamma([0, 1])$, we easily obtain a strict inequality. □

Here is an important consequence of Theorem 12.8.

Corollary 12.9. *Let (M, g) be a Riemannian manifold. If $\omega\colon [0, b] \to M$ is any curve parametrized by arc-length and ω has length less than or equal to the length of any other curve from $\omega(0)$ to $\omega(b)$, then ω is a geodesic.*

Proof. Consider any segment of ω lying within an open set, W, as above, and having length $< \epsilon$. By Theorem 12.8, this segment must be a geodesic. Hence, the entire curve is a geodesic. □

Definition 12.6. Let (M, g) be a Riemannian manifold. A geodesic, $\gamma\colon [a, b] \to M$, is *minimal* iff its length is less than or equal to the length of any other piecewise smooth curve joining its endpoints.

Theorem 12.8 asserts that any sufficiently small segment of a geodesic is minimal. On the other hand, a long geodesic may not be minimal. For example, a great circle arc on the unit sphere is a geodesic. If such an arc has length greater than π, then it is not minimal. Minimal geodesics are generally not unique. For example, any two antipodal points on a sphere are joined by an infinite number of minimal geodesics.

A *broken geodesic* is a piecewise smooth curve as in Definition 12.1, where each curve segment is a geodesic.

Proposition 12.10. *A Riemannian manifold, (M, g), is connected iff any two points of M can be joined by a broken geodesic.*

Proof. Assume M is connected, pick any $p \in M$, and let $S_p \subseteq M$ be the set of all points that can be connected to p by a broken geodesic. For any $q \in M$, choose a normal neighborhood, U, of q. If $q \in S_p$, then it is clear that $U \subseteq S_p$. On the other hand, if $q \notin S_p$, then $U \subseteq M - S_p$. Therefore, $S_p \neq \emptyset$ is open and closed, so $S_p = M$. The converse is obvious. □

In general, if M is connected, then it is not true that any two points are joined by a geodesic. However, this will be the case if M is geodesically complete, as we will see in the next section.

Next, we will see that a Riemannian metric induces a distance on the manifold whose induced topology agrees with the original metric.

12.3 Complete Riemannian Manifolds, the Hopf-Rinow Theorem and the Cut Locus

Every connected Riemannian manifold, (M, g), is a metric space in a natural way. Furthermore, M is a complete metric space iff M is geodesically complete. In this section, we explore briefly some properties of complete Riemannian manifolds.

Proposition 12.11. *Let (M, g) be a connected Riemannian manifold. For any two points, $p, q \in M$, let $d(p, q)$ be the greatest lower bound of the lengths of all piecewise smooth curves joining p to q. Then, d is a metric on M and the topology of the metric space, (M, d), coincides with the original topology of M.*

A proof of the above proposition can be found in Gallot, Hulin and Lafontaine [61] (Chapter 2, Proposition 2.91) or O'Neill [120] (Chapter 5, Proposition 18).

The distance, d, is often called the *Riemannian distance* on M. For any $p \in M$ and any $\epsilon > 0$, the *metric ball of center p and radius ϵ* is the subset, $B_\epsilon(p) \subseteq M$, given by

$$B_\epsilon(p) = \{q \in M \mid d(p, q) < \epsilon\}.$$

The next proposition follows easily from Proposition 12.4 (Milnor [107], Section 10, Corollary 10.8).

Proposition 12.12. *Let (M, g) be a connected Riemannian manifold. For any compact subset, $K \subseteq M$, there is a number $\delta > 0$ so that any two points, $p, q \in K$, with distance $d(p, q) < \delta$ are joined by a unique geodesic of length less than δ. Furthermore, this geodesic is minimal and depends smoothly on its endpoints.*

Recall from Definition 12.4 that (M, g) is geodesically complete iff the exponential map, $v \mapsto \exp_p(v)$, is defined for all $p \in M$ and for all $v \in T_pM$. We now prove the following important theorem due to Hopf and Rinow (1931):

Theorem 12.13. *(Hopf-Rinow) Let (M, g) be a connected Riemannian manifold. If there is a point, $p \in M$, such that $\exp_p$ is defined on the entire tangent space, T_pM, then any point, $q \in M$, can be joined to p by a minimal geodesic. As a consequence, if M is geodesically complete, then any two points of M can be joined by a minimal geodesic.*

Proof. We follow Milnor's proof in [107], Chapter 10, Theorem 10.9. Pick any two points, $p, q \in M$ and let $r = d(p, q)$. By Proposition 12.4, there is some open subset, W, with $p \in W$ and some $\epsilon > 0$ so that any two points of W are joined by a unique geodesic and the exponential map is a diffeomorphism between the open ball, $B(0, \epsilon)$, and its image, $U_p = \exp_p(B(0, \epsilon))$. For $\delta < \epsilon$, let $S = \exp_p(S(0, \delta))$, where $S(0, \delta)$ is the sphere of radius δ. Since $S \subseteq U_p$ is compact, there is some point,

$$p_0 = \exp_p(\delta v), \qquad \text{with } \|v\| = 1,$$

on S for which the distance to q is minimized. We will prove that

$$\exp_p(rv) = q,$$

which will imply that the geodesic, γ, given by $\gamma(t) = \exp_p(tv)$ is actually a minimal geodesic from p to q (with $t \in [0, r]$). Here, we use the fact that the exponential $\exp_p$ is defined everywhere on T_pM.

The proof amounts to showing that a point which moves along the geodesic γ must get closer and closer to q. In fact, for each $t \in [\delta, r]$, we prove

$$d(\gamma(t), q) = r - t. \tag{$*_t$}$$

We get the proof by setting $t = r$.

First, we prove $(*_\delta)$. Since every path from p to q must pass through S, by the choice of p_0, we have

$$r = d(p, q) = \min_{s \in S}\{d(p, s) + d(s, q)\} = \delta + d(p_0, q).$$

Therefore, $d(p_0, q) = r - \delta$ and since $p_0 = \gamma(\delta)$, this proves $(*_\delta)$.

Define $t_0 \in [\delta, r]$ by

$$t_0 = \sup\{t \in [\delta, r] \mid d(\gamma(t), q) = r - t\}.$$

As the set, $\{t \in [\delta, r] \mid d(\gamma(t), q) = r - t\}$, is closed, it contains its upper bound, t_0, so the equation $(*_{t_0})$ also holds. We claim that if $t_0 < r$, then we obtain a contradiction.

As we did with p, there is some small $\delta' > 0$ so that if $S' = \exp_{\gamma(t_0)}(B(0, \delta'))$, then there is some point, p_0', on S' with minimum distance from q and p_0' is joined to $\gamma(t_0)$ by a mimimal geodesic. We have

$$r - t_0 = d(\gamma(t_0), q) = \min_{s \in S'}\{d(\gamma(t_0), s) + d(s, q)\} = \delta' + d(p_0', q),$$

hence

$$d(p_0', q) = r - t_0 - \delta'. \tag{$\dagger$}$$

We claim that $p_0' = \gamma(t_0 + \delta')$.

By the triangle inequality and using $(\dagger)$ (recall that $d(p, q) = r$), we have

$$d(p, p_0') \geq d(p, q) - d(p_0', q) = t_0 + \delta'.$$

But, a path of length precisely $t_0 + \delta'$ from p to p_0' is obtained by following γ from p to $\gamma(t_0)$, and then following a minimal geodesic from $\gamma(t_0)$ to p_0'. Since this broken geodesic has minimal length, by Corollary 12.9, it is a genuine (unbroken) geodesic, and so, it coincides with γ. But then, as $p_0' = \gamma(t_0 + \delta')$, equality $(\dagger)$ becomes $(*_{t_0+\delta'})$, namely

$$d(\gamma(t_0 + \delta'), q) = r - (t_0 + \delta'),$$

contradicting the maximality of t_0. Therefore, we must have $t_0 = r$ and $q = \exp_p(rv)$, as desired. □

Remark: Theorem 12.13 is proved is every decent book on Riemannian geometry. Among those, we mention Gallot, Hulin and Lafontaine [61], Chapter 2, Theorem 2.103 and O'Neill [120], Chapter 5, Lemma 24.

Theorem 12.13 implies the following result (often known as the *Hopf-Rinow Theorem*):

Theorem 12.14. *Let (M, g) be a connected, Riemannian manifold. The following statements are equivalent:*

(1) The manifold (M, g) is geodesically complete, that is, for every $p \in M$, every geodesic through p can be extended to a geodesic defined on all of $\mathbb{R}$.

(2) For every point, $p \in M$, the map $\exp_p$ is defined on the entire tangent space, T_pM.

(3) There is a point, $p \in M$, such that $\exp_p$ is defined on the entire tangent space, T_pM.

(4) Any closed and bounded subset of the metric space, (M, d), is compact.

(5) The metric space, (M, d), is complete (that is, every Cauchy sequence converges).

Proofs of Theorem 12.14 can be found in Gallot, Hulin and Lafontaine [61], Chapter 2, Corollary 2.105 and O'Neill [120], Chapter 5, Theorem 21.

In view of Theorem 12.14, a connected Riemannian manifold, (M, g), is geodesically complete iff the metric space, (M, d), is complete. We will refer simply to M as a *complete Riemannian manifold* (it is understood that M is connected). Also, by (4), every compact, Riemannian manifold is complete. If we remove any point, p, from a Riemannian manifold, M, then $M - \{p\}$ is not complete since every geodesic that formerly went through p yields a geodesic that can't be extended.

Assume (M, g) is a complete Riemannian manifold. Given any point, $p \in M$, it is interesting to consider the subset, $\mathcal{U}_p \subseteq T_pM$, consisting of all $v \in T_pM$ such that the geodesic

$$t \mapsto \exp_p(tv)$$

is a minimal geodesic up to $t = 1 + \epsilon$, for some $\epsilon > 0$. The subset $\mathcal{U}_p$ is open and star-shaped and it turns out that $\exp_p$ is a diffeomorphism from $\mathcal{U}_p$ onto its image, $\exp_p(\mathcal{U}_p)$, in M. The left-over part, $M - \exp_p(\mathcal{U}_p)$ (if nonempty), is actually equal to $\exp_p(\partial\mathcal{U}_p)$ and it is an important subset of M called the *cut locus of p*. The following proposition is needed to establish properties of the cut locus:

Proposition 12.15. *Let (M, g) be a complete Riemannian manifold. For any geodesic, $\gamma\colon [0, a] \to M$, from $p = \gamma(0)$ to $q = \gamma(a)$, the following properties hold:*

(i) If there is no geodesic shorter than γ between p and q, then γ is minimal on $[0, a]$.

(ii) If there is another geodesic of the same length as γ between p and q, then γ is no longer minimal on any larger interval, $[0, a + \epsilon]$.

(iii) If γ is minimal on any interval, I, then γ is also minimal on any subinterval of I.

Proof. Part (iii) is an immediate consequence of the triangle inequality. As M is complete, by the Hopf-Rinow Theorem, there is a minimal geodesic from p to q, so γ must be minimal too. This proves part (i). Part (ii) is proved in Gallot, Hulin and Lafontaine [61], Chapter 2, Corollary 2.111. □

Again, assume (M, g) is a complete Riemannian manifold and let $p \in M$ be any point. For every $v \in T_pM$, let

$$I_v = \{s \in \mathbb{R} \cup \{\infty\} \mid \text{the geodesic} \quad t \mapsto \exp_p(tv) \quad \text{is minimal on } [0, s]\}.$$

It is easy to see that I_v is a closed interval, so $I_v = [0, \rho(v)]$ (with $\rho(v)$ possibly infinite). It can be shown that if $w = \lambda v$, then $\rho(v) = \lambda\rho(w)$, so we can restrict our attention to unit vectors, v. It can also be shown that the map, $\rho\colon S^{n-1} \to \mathbb{R}$, is continuous, where S^{n-1} is the unit sphere of center 0 in T_pM, and that $\rho(v)$ is bounded below by a strictly positive number.

Definition 12.7. Let (M, g) be a complete Riemannian manifold and let $p \in M$ be any point. Define $\mathcal{U}_p$ by

$$\mathcal{U}_p = \left\{ v \in T_pM \;\middle|\; \rho\left(\frac{v}{\|v\|}\right) > \|v\| \right\} = \{v \in T_pM \mid \rho(v) > 1\}$$

and the *cut locus* of p by

$$\mathrm{Cut}(p) = \exp_p(\partial\mathcal{U}_p) = \{\exp_p(\rho(v)v) \mid v \in S^{n-1}\}.$$

The set $\mathcal{U}_p$ is open and star-shaped. The boundary, $\partial\mathcal{U}_p$, of $\mathcal{U}_p$ in T_pM is sometimes called the *tangential cut locus* of p and is denoted $\widetilde{\mathrm{Cut}}(p)$.

Remark: The cut locus was first introduced for convex surfaces by Poincaré (1905) under the name *ligne de partage*. According to Do Carmo [51] (Chapter 13, Section 2), for Riemannian manifolds, the cut locus was introduced by J.H.C. Whitehead (1935). But it was Klingenberg (1959) who revived the interest in the cut locus and showed its usefuleness.

Proposition 12.16. *Let (M, g) be a complete Riemannian manifold. For any point, $p \in M$, the sets $\exp_p(\mathcal{U}_p)$ and $\mathrm{Cut}(p)$ are disjoint and*

$$M = \exp_p(\mathcal{U}_p) \cup \mathrm{Cut}(p).$$

Proof. From the Hopf-Rinow Theorem, for every $q \in M$, there is a minimal geodesic, $t \mapsto \exp_p(vt)$ such that $\exp_p(v) = q$. This shows that $\rho(v) \geq 1$, so $v \in \overline{\mathcal{U}_p}$ and

$$M = \exp_p(\mathcal{U}_p) \cup \mathrm{Cut}(p).$$

It remains to show that this is a disjoint union. Assume $q \in \exp_p(\mathcal{U}_p) \cap \mathrm{Cut}(p)$. Since $q \in \exp_p(\mathcal{U}_p)$, there is a geodesic, γ, such that $\gamma(0) = p$, $\gamma(a) = q$ and γ is minimal on $[0, a + \epsilon]$, for some $\epsilon > 0$. On the other hand, as $q \in \mathrm{Cut}(p)$, there is some geodesic, $\widetilde{\gamma}$, with $\widetilde{\gamma}(0) = p$, $\widetilde{\gamma}(b) = q$, $\widetilde{\gamma}$ minimal on $[0, b]$, but $\widetilde{\gamma}$ not minimal after b. As γ and $\widetilde{\gamma}$ are both minimal from p to q, they have the same length from p to q. But then, as γ and $\widetilde{\gamma}$ are distinct, by Proposition 12.15 (ii), the geodesic γ can't be minimal after q, a contradiction. □

Observe that the injectivity radius, $i(p)$, of M at p is equal to the distance from p to the cut locus of p:

$$i(p) = d(p, \mathrm{Cut}(p)) = \inf_{q \in \mathrm{Cut}(p)} d(p, q).$$

Consequently, the injectivity radius, $i(M)$, of M is given by

$$i(M) = \inf_{p \in M} d(p, \mathrm{Cut}(p)).$$

If M is compact, it can be shown that $i(M) > 0$. It can also be shown using Jacobi fields that $\exp_p$ is a diffeomorphism from $\mathcal{U}_p$ onto its image, $\exp_p(\mathcal{U}_p)$. Thus, $\exp_p(\mathcal{U}_p)$ is diffeomorphic to an open ball in $\mathbb{R}^n$ (where $n = \dim(M)$) and the cut locus is closed. Hence, the manifold, M, is obtained by gluing together an open n-ball onto the cut locus of a point. In some sense the topology of M is "contained" in its cut locus.

Given any sphere, S^{n-1}, the cut locus of any point, p, is its antipodal point, $\{-p\}$. For more examples, consult Gallot, Hulin and Lafontaine [61] (Chapter 2, Section 2C7), Do Carmo [51] (Chapter 13, Section 2) or Berger [16] (Chapter 6). In general, the cut locus is very hard to compute. In fact, according to Berger [16], even for an ellipsoid, the determination of the cut locus of an arbitrary point is still a matter of conjecture!

12.4 The Calculus of Variations Applied to Geodesics; The First Variation Formula

Given a Riemannian manifold, (M, g), the path space, $\Omega(p, q)$, was introduced in Definition 12.1. It is an "infinite dimensional" manifold. By analogy with finite dimensional manifolds, we define a kind of tangent space to $\Omega(p, q)$ at a "point" ω. In this section, it is convenient to assume that paths in $\Omega(p, q)$ are parametrized over the interval $[0, 1]$.

Definition 12.8. For every "point", $\omega \in \Omega(p, q)$, we define the "*tangent space*", $T_\omega\Omega(p, q)$, of $\Omega(p, q)$ at ω, to be the space of all piecewise smooth vector fields, W, along ω, for which $W(0) = W(1) = 0$.

Now, if $F \colon \Omega(p, q) \to \mathbb{R}$ is a real-valued function on $\Omega(p, q)$, it is natural to ask what the induced "tangent map",

$$dF_\omega \colon T_\omega\Omega(p, q) \to \mathbb{R},$$

should mean (here, we are identifying $T_{F(\omega)}\mathbb{R}$ with $\mathbb{R}$). Observe that $\Omega(p, q)$ is not even a topological space so the answer is far from obvious! In the case where $f \colon M \to \mathbb{R}$ is a function on a manifold, there are various equivalent ways to define df, one of which involves curves. For every $v \in T_pM$, if $\alpha \colon (-\epsilon, \epsilon) \to M$ is a curve such that $\alpha(0) = p$ and $\alpha'(0) = v$, then we know that

$$df_p(v) = \left.\frac{d(f(\alpha(t)))}{dt}\right|_{t=0}.$$

We may think of α as a small *variation* of p. Recall that p is a *critical point* of f iff $df_p(v) = 0$, for all $v \in T_pM$.

Rather than attempting to define dF_ω (which requires some conditions on F), we will mimic what we did with functions on manifolds and define what is a *critical path* of a function, $F\colon \Omega(p,q) \to \mathbb{R}$, using the notion of *variation*. Now, geodesics from p to q are special paths in $\Omega(p,q)$ and they turn out to be the critical paths of the *energy function*,

$$E_a^b(\omega) = \int_a^b \|\omega'(t)\|^2\, dt,$$

where $\omega \in \Omega(p,q)$, and $0 \leq a < b \leq 1$.

Definition 12.9. Given any path, $\omega \in \Omega(p,q)$, a *variation of* ω *(keeping endpoints fixed)* is a function, $\widetilde{\alpha}\colon (-\epsilon,\epsilon) \to \Omega(p,q)$, for some $\epsilon > 0$, such that

(1) $\widetilde{\alpha}(0) = \omega$

(2) There is a subdivision, $0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = 1$ of $[0,1]$ so that the map

$$\alpha\colon (-\epsilon,\epsilon) \times [0,1] \to M$$

defined by $\alpha(u,t) = \widetilde{\alpha}(u)(t)$ is smooth on each strip $(-\epsilon,\epsilon)\times[t_i,t_{i+1}]$, for $i = 0,\ldots,k-1$.

If U is an open subset of $\mathbb{R}^n$ containing the origin and if we replace $(-\epsilon,\epsilon)$ by U in the above, then $\widetilde{\alpha}\colon U \to \Omega(p,q)$ is called an *n-parameter variation* of ω.

The function α is also called a *variation* of ω. Since each $\widetilde{\alpha}(u)$ belongs to $\Omega(p,q)$, note that

$$\alpha(u,0) = p, \quad \alpha(u,1) = q, \quad \text{for all } u \in (-\epsilon,\epsilon).$$

The function, $\widetilde{\alpha}$, may be considered as a "smooth path" in $\Omega(p,q)$, since for every $u \in (-\epsilon,\epsilon)$, the map $\widetilde{\alpha}(u)$ is a curve in $\Omega(p,q)$ called a *curve in the variation (or longitudinal curve of the variation)*. The "velocity vector", $\frac{d\widetilde{\alpha}}{du}(0) \in T_\omega\Omega(p,q)$, is defined to be the vector field, W, along ω, given by

$$W_t = \frac{d\widetilde{\alpha}}{du}(0)_t = \frac{\partial \alpha}{\partial u}\,(0,t),$$

Clearly, $W \in T_\omega\Omega(p,q)$. In particular, $W(0) = W(1) = 0$. The vector field, W, is also called the *variation vector field* associated with the variation α.

Besides the curves in the variation, $\widetilde{\alpha}(u)$ (with $u \in (-\epsilon,\epsilon)$), for every $t \in [0,1]$, we have a curve, $\alpha_t\colon (-\epsilon,\epsilon) \to M$, called a *transversal curve of the variation*, defined by

$$\alpha_t(u) = \widetilde{\alpha}(u)(t),$$

and W_t is equal to the velocity vector, $\alpha_t'(0)$, at the point $\omega(t) = \alpha_t(0)$. For ϵ sufficiently small, the vector field, W_t, is an infinitesimal model of the variation $\widetilde{\alpha}$.

We can show that for any $W \in T_\omega\Omega(p,q)$ there is a variation, $\widetilde{\alpha}\colon (-\epsilon,\epsilon) \to \Omega(p,q)$, which satisfies the conditions

$$\widetilde{\alpha}(0) = \omega, \qquad \frac{d\widetilde{\alpha}}{du}(0) = W.$$

Sketch of the proof. By the compactness of $\omega([0,1])$, it is possible to find a $\delta > 0$ so that $\exp_{\omega(t)}$ is defined for all $t \in [0,1]$ and all $v \in T_{\omega(t)}M$, with $\|v\| < \delta$. Then, if

$$N = \max_{t\in[0,1]} \|W_t\|,$$

for any ϵ such that $0 < \epsilon < \frac{\delta}{N}$, it can be shown that

$$\widetilde{\alpha}(u)(t) = \exp_{\omega(t)}(uW_t)$$

works (for details, see Do Carmo [51], Chapter 9, Proposition 2.2). □

As we said earlier, given a function, $F\colon \Omega(p,q) \to \mathbb{R}$, we do not attempt to define the differential, dF_ω, but instead, the notion of critical path.

Definition 12.10. Given a function, $F\colon \Omega(p,q) \to \mathbb{R}$, we say that a path, $\omega \in \Omega(p,q)$, is a *critical path* for F iff

$$\left.\frac{dF(\widetilde{\alpha}(u))}{du}\right|_{u=0} = 0,$$

for every variation, $\widetilde{\alpha}$, of ω (which implies that the derivative $\left.\frac{dF(\widetilde{\alpha}(u))}{du}\right|_{u=0}$ is defined for every variation, $\widetilde{\alpha}$, of ω).

For example, if F takes on its minimum on a path ω_0 and if the derivatives $\frac{dF(\widetilde{\alpha}(u))}{du}$ are all defined, then ω_0 is a critical path of F.

We will apply the above to two functions defined on $\Omega(p,q)$:

(1) The *energy function* (also called *action integral*):

$$E_a^b(\omega) = \int_a^b \|\omega'(t)\|^2\, dt.$$

(We write $E = E_0^1$.)

(2) The *arc-length function*,

$$L_a^b(\omega) = \int_a^b \|\omega'(t)\|\, dt.$$

The quantities $E_a^b(\omega)$ and $L_a^b(\omega)$ can be compared as follows: if we apply the Cauchy-Schwarz's inequality,

$$\left(\int_a^b f(t)g(t)dt\right)^2 \leq \left(\int_a^b f^2(t)dt\right)\left(\int_a^b g^2(t)dt\right)$$

with $f(t) \equiv 1$ and $g(t) = \|\omega'(t)\|$, we get

$$(L_a^b(\omega))^2 \leq (b-a)E_a^b,$$

where equality holds iff g is constant; that is, iff the parameter t is proportional to arc-length.

Now, suppose that there exists a minimal geodesic, γ, from p to q. Then,

$$E(\gamma) = L(\gamma)^2 \leq L(\omega)^2 \leq E(\omega),$$

where the equality $L(\gamma)^2 = L(\omega)^2$ holds only if ω is also a minimal geodesic, possibly reparametrized. On the other hand, the equality $L(\omega) = E(\omega)^2$ can hold only if the parameter is proportional to arc-length along ω. This proves that $E(\gamma) < E(\omega)$ unless ω is also a minimal geodesic. We just proved:

Proposition 12.17. *Let (M, g) be a complete Riemannian manifold. For any two points, $p, q \in M$, if $d(p, q) = \delta$, then the energy function, $E\colon \Omega(p, q) \to \mathbb{R}$, takes on its minimum, δ^2, precisely on the set of minimal geodesics from p to q.*

Next, we are going to show that the critical paths of the energy function are exactly the geodesics. For this, we need the *first variation formula*.

Let $\widetilde{\alpha}\colon (-\epsilon, \epsilon) \to \Omega(p, q)$ be a variation of ω and let

$$W_t = \frac{\partial \alpha}{\partial u}(0, t)$$

be its associated variation vector field. Furthermore, let

$$V_t = \frac{d\omega}{dt} = \omega'(t),$$

the velocity vector of ω and

$$\Delta_t V = V_{t_+} - V_{t_-},$$

the discontinuity in the velocity vector at t, which is nonzero only for $t = t_i$, with $0 < t_i < 1$ (see the definition of $\gamma'((t_i)_+)$ and $\gamma'((t_i)_-)$ just after Definition 12.1).

Theorem 12.18. *(First Variation Formula) For any path, $\omega \in \Omega(p, q)$, we have*

$$\frac{1}{2}\frac{dE(\widetilde{\alpha}(u))}{du}\bigg|_{u=0} = -\sum_i \langle W_t, \Delta_t V\rangle - \int_0^1 \left\langle W_t, \frac{D}{dt}V_t\right\rangle dt,$$

where $\widetilde{\alpha}\colon (-\epsilon, \epsilon) \to \Omega(p, q)$ is any variation of ω.

Proof. (After Milnor, see [107], Chapter II, Section 12, Theorem 12.2.) By Proposition 11.24, we have

$$\frac{\partial}{\partial u}\left\langle \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t}\right\rangle = 2\left\langle \frac{D}{\partial u}\frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t}\right\rangle.$$

Therefore,

$$\frac{dE(\widetilde{\alpha}(u))}{du} = \frac{d}{du}\int_0^1 \left\langle \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t}\right\rangle dt = 2\int_0^1 \left\langle \frac{D}{\partial u}\frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t}\right\rangle dt.$$

Now, because we are using the Levi-Civita connection, which is torsion-free, it is not hard to prove that

$$\frac{D}{\partial t}\frac{\partial \alpha}{\partial u} = \frac{D}{\partial u}\frac{\partial \alpha}{\partial t},$$

so

$$\frac{dE(\widetilde{\alpha}(u))}{du} = 2\int_0^1 \left\langle \frac{D}{\partial t}\frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t}\right\rangle dt.$$

We can choose $0 = t_0 < t_1 < \cdots < t_k = 1$ so that α is smooth on each strip $(-\epsilon, \epsilon) \times [t_{i-1}, t_i]$. Then, we can "integrate by parts" on $[t_{i-1}, t_i]$ as follows: The equation

$$\frac{\partial}{\partial t}\left\langle \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t}\right\rangle = \left\langle \frac{D}{\partial t}\frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t}\right\rangle + \left\langle \frac{\partial \alpha}{\partial u}, \frac{D}{\partial t}\frac{\partial \alpha}{\partial t}\right\rangle$$

implies that

$$\int_{t_{i-1}}^{t_i} \left\langle \frac{D}{\partial t}\frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t}\right\rangle dt = \left\langle \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t}\right\rangle\Big|_{t=(t_{i-1})_+}^{t=(t_i)_-} - \int_{t_{i-1}}^{t_i} \left\langle \frac{\partial \alpha}{\partial u}, \frac{D}{\partial t}\frac{\partial \alpha}{\partial t}\right\rangle dt.$$

Adding up these formulae for $i = 1, \ldots k-1$ and using the fact that $\frac{\partial \alpha}{\partial u} = 0$ for $t = 0$ and $t = 1$, we get

$$\frac{1}{2}\frac{dE(\widetilde{\alpha}(u))}{du} = -\sum_{i=1}^{k-1}\left\langle \frac{\partial \alpha}{\partial u}, \Delta_{t_i}\frac{\partial \alpha}{\partial t}\right\rangle - \int_0^1 \left\langle \frac{\partial \alpha}{\partial u}, \frac{D}{\partial t}\frac{\partial \alpha}{\partial t}\right\rangle dt.$$

Setting $u = 0$, we obtain the formula

$$\frac{1}{2}\frac{dE(\widetilde{\alpha}(u))}{du}\Big|_{u=0} = -\sum_i \langle W_t, \Delta_t V\rangle - \int_0^1 \left\langle W_t, \frac{D}{dt}V_t\right\rangle dt,$$

as claimed. □

Intuitively, the first term on the right-hand side shows that varying the path ω in the direction of decreasing "kink" tends to decrease E. The second term shows that varying the curve in the direction of its acceleration vector, $\frac{D}{dt}\omega'(t)$, also tends to reduce E.

A geodesic, γ, (parametrized over $[0, 1]$) is smooth on the entire interval $[0, 1]$ and its acceleration vector, $\frac{D}{dt}\gamma'(t)$, is identically zero along γ. This gives us half of

Theorem 12.19. *Let (M, g) be a Riemanian manifold. For any two points, $p, q \in M$, a path, $\omega \in \Omega(p, q)$ (parametrized over $[0, 1]$), is critical for the energy function, E, iff ω is a geodesic.*

Proof. From the first variation formula, it is clear that a geodesic is a critical path of E.

Conversely, assume ω is a critical path of E. There is a variation, $\widetilde{\alpha}$, of ω such that its associated variation vector field is of the form

$$W(t) = f(t)\frac{D}{dt}\,\omega'(t),$$

with $f(t)$ smooth and positive except that it vanishes at the t_i's. For this variation, we get

$$\frac{1}{2}\,\frac{dE(\widetilde{\alpha}(u))}{du}\bigg|_{u=0} = -\int_0^1 f(t)\left\langle \frac{D}{dt}\,\omega'(t), \frac{D}{dt}\,\gamma'(t)\right\rangle dt.$$

This expression is zero iff

$$\frac{D}{dt}\,\omega'(t) = 0 \qquad \text{on } [0, 1].$$

Hence, the restriction of ω to each $[t_i, t_{i+1}]$ is a geodesic.

It remains to prove that ω is smooth on the entire interval $[0, 1]$. For this, pick a variation $\widetilde{\alpha}$ such that

$$W(t_i) = \Delta_{t_i} V.$$

Then, we have

$$\frac{1}{2}\,\frac{dE(\widetilde{\alpha}(u))}{du}\bigg|_{u=0} = -\sum_{i=1}^{k}\langle \Delta_{t_i} V, \Delta_{t_i} V\rangle.$$

If the above expression is zero, then $\Delta_{t_i} V = 0$ for $i = 1, \ldots, k-1$, which means that ω is C^1 everywhere on $[0, 1]$. By the uniqueness theorem for ODE's, ω must be smooth everywhere on $[0, 1]$, and thus, it is an unbroken geodesic. □

Remark: If $\omega \in \Omega(p, q)$ is parametrized by arc-length, it is easy to prove that

$$\frac{dL(\widetilde{\alpha}(u))}{du}\bigg|_{u=0} = \frac{1}{2}\,\frac{dE(\widetilde{\alpha}(u))}{du}\bigg|_{u=0}.$$

As a consequence, a path, $\omega \in \Omega(p, q)$ is critical for the arc-length function, L, iff it can be reparametrized so that it is a geodesic (see Gallot, Hulin and Lafontaine [61], Chapter 3, Theorem 3.31).

In order to go deeper into the study of geodesics we need Jacobi fields and the "second variation formula", both involving a curvature term. Therefore, we now proceed with a more thorough study of curvature on Riemannian manifolds.

Chapter 13

Curvature in Riemannian Manifolds

13.1 The Curvature Tensor

If $(M, \langle -, - \rangle)$ is a Riemannian manifold and ∇ is a connection on M (that is, a connection on TM), we saw in Section 11.2 (Proposition 11.8) that the curvature induced by ∇ is given by

$$R(X, Y) = \nabla_X \circ \nabla_Y - \nabla_Y \circ \nabla_X - \nabla_{[X,Y]},$$

for all $X, Y \in \mathfrak{X}(M)$, with $R(X, Y) \in \Gamma(\mathcal{H}om(TM, TM)) \cong \mathrm{Hom}_{C^\infty(M)}(\Gamma(TM), \Gamma(TM))$. Since sections of the tangent bundle are vector fields ($\Gamma(TM) = \mathfrak{X}(M)$), R defines a map

$$R\colon \mathfrak{X}(M) \times \mathfrak{X}(M) \times \mathfrak{X}(M) \longrightarrow \mathfrak{X}(M),$$

and, as we observed just after stating Proposition 11.8, $R(X, Y)Z$ is $C^\infty(M)$-linear in X, Y, Z and skew-symmetric in X and Y. It follows that R defines a $(1, 3)$-tensor, also denoted R, with

$$R_p\colon T_pM \times T_pM \times T_pM \longrightarrow T_pM.$$

Experience shows that it is useful to consider the $(0, 4)$-tensor, also denoted R, given by

$$R_p(x, y, z, w) = \langle R_p(x, y)z, w \rangle_p$$

as well as the expression $R(x, y, y, x)$, which, for an orthonormal pair, of vectors (x, y), is known as the *sectional curvature*, $K(x, y)$.

This last expression brings up a dilemma regarding the choice for the sign of R. With our present choice, the sectional curvature, $K(x, y)$, is given by $K(x, y) = R(x, y, y, x)$ but many authors define K as $K(x, y) = R(x, y, x, y)$. Since $R(x, y)$ is skew-symmetric in x, y, the latter choice corresponds to using $-R(x, y)$ instead of $R(x, y)$, that is, to define $R(X, Y)$ by

$$R(X, Y) = \nabla_{[X,Y]} + \nabla_Y \circ \nabla_X - \nabla_X \circ \nabla_Y.$$

As pointed out by Milnor [107] (Chapter II, Section 9), the latter choice for the sign of R has the advantage that, in coordinates, the quantity, $\langle R(\partial/\partial x_h, \partial/\partial x_i)\partial/\partial x_j, \partial/\partial x_k \rangle$ coincides

with the classical Ricci notation, R_{hijk}. Gallot, Hulin and Lafontaine [61] (Chapter 3, Section A.1) give other reasons supporting this choice of sign. Clearly, the choice for the sign of R is mostly a matter of taste and we apologize to those readers who prefer the first choice but we will adopt the second choice advocated by Milnor and others. Therefore, we make the following formal definition:

Definition 13.1. Let $(M, \langle -, - \rangle)$ be a Riemannian manifold equipped with the Levi-Civita connection. The *curvature tensor* is the $(1,3)$-tensor, R, defined by

$$R_p(x,y)z = \nabla_{[X,Y]}Z + \nabla_Y \nabla_X Z - \nabla_X \nabla_Y Z,$$

for every $p \in M$ and for any vector fields, $X, Y, Z \in \mathfrak{X}(M)$, such that $x = X(p)$, $y = Y(p)$ and $z = Z(p)$. The $(0,4)$-tensor associated with R, also denoted R, is given by

$$R_p(x,y,z,w) = \langle (R_p(x,y)z, w \rangle,$$

for all $p \in M$ and all $x, y, z, w \in T_pM$.

Locally in a chart, we write

$$R\left(\frac{\partial}{\partial x_h}, \frac{\partial}{\partial x_i}\right)\frac{\partial}{\partial x_j} = \sum_l R^l_{jhi}\frac{\partial}{\partial x_l}$$

and

$$R_{hijk} = \left\langle R\left(\frac{\partial}{\partial x_h}, \frac{\partial}{\partial x_i}\right)\frac{\partial}{\partial x_j}, \frac{\partial}{\partial x_k}\right\rangle = \sum_l g_{lk} R^l_{jhi}.$$

The coefficients, R^l_{jhi}, can be expressed in terms of the Christoffel symbols, Γ^k_{ij}, in terms of a rather unfriendly formula (see Gallot, Hulin and Lafontaine [61] (Chapter 3, Section 3.A.3) or O'Neill [120] (Chapter III, Lemma 38). Since we have adopted O'Neill's conventions for the order of the subscripts in R^l_{jhi}, here is the formula from O'Neill:

$$R^l_{jhi} = \partial_i \Gamma^l_{hj} - \partial_h \Gamma^l_{ij} + \sum_m \Gamma^l_{im}\Gamma^m_{hj} - \sum_m \Gamma^l_{hm}\Gamma^m_{ij}.$$

There is another way of defining the curvature tensor which is useful for comparing second covariant derivatives of one-forms. Recall that for any fixed vector field, Z, the map, $Y \mapsto \nabla_Y Z$, is a $(1,1)$ tensor that we will denote $\nabla_- Z$. Thus, using Proposition 11.5, the covariant derivative $\nabla_X \nabla_- Z$ of $\nabla_- Z$ makes sense and is given by

$$(\nabla_X(\nabla_- Z))(Y) = \nabla_X(\nabla_Y Z) - (\nabla_{\nabla_X Y})Z.$$

Usually, $(\nabla_X(\nabla_- Z))(Y)$ is denoted by $\nabla^2_{X,Y} Z$ and

$$\nabla^2_{X,Y} Z = \nabla_X(\nabla_Y Z) - \nabla_{\nabla_X Y} Z$$

is called the *second covariant derivative* of Z with respect to X and Y. Then, we have

$$\begin{aligned}\nabla^2_{Y,X}Z - \nabla^2_{X,Y}Z &= \nabla_Y(\nabla_X Z) - \nabla_{\nabla_Y X}Z - \nabla_X(\nabla_Y Z) + \nabla_{\nabla_X Y}Z\\ &= \nabla_Y(\nabla_X Z) - \nabla_X(\nabla_Y Z) + \nabla_{\nabla_X Y - \nabla_Y X}Z\\ &= \nabla_Y(\nabla_X Z) - \nabla_X(\nabla_Y Z) + \nabla_{[X,Y]}Z\\ &= R(X,Y)Z,\end{aligned}$$

since $\nabla_X Y - \nabla_Y X = [X,Y]$, as the Levi-Civita connection is torsion-free. Therefore, the curvature tensor can also be defined by

$$R(X,Y)Z = \nabla^2_{Y,X}Z - \nabla^2_{X,Y}Z.$$

We already know that the curvature tensor has some symmetry properties, for example, $R(y,x)z = -R(x,y)z$ but when it is induced by the Levi-Civita connection, it has more remarkable properties stated in the next proposition.

Proposition 13.1. *For a Riemannian manifold, $(M, \langle -,-\rangle)$, equipped with the Levi-Civita connection, the curvature tensor satisfies the following properties:*

(1) $R(x,y)z = -R(y,x)z$

(2) (First Bianchi Identity) $R(x,y)z + R(y,z)x + R(z,x)y = 0$

(3) $R(x,y,z,w) = -R(x,y,w,z)$

(4) $R(x,y,z,w) = R(z,w,x,y)$.

The proof of Proposition 13.1 uses the fact that $R_p(x,y)z = R(X,Y)Z$, for any vector fields X, Y, Z such that $x = X(p)$, $y = Y(p)$ and $Z = Z(p)$. In particular, X, Y, Z can be chosen so that their pairwise Lie brackets are zero (choose a coordinate system and give X, Y, Z constant components). Part (1) is already known. Part (2) follows from the fact that the Levi-Civita connection is torsion-free. Parts (3) and (4) are a little more tricky. Complete proofs can be found in Milnor [107] (Chapter II, Section 9), O'Neill [120] (Chapter III) and Kuhnel [92] (Chapter 6, Lemma 6.3).

If $\omega \in \mathcal{A}^1(M)$ is a one-form, then the covariant derivative of ω defines a $(0,2)$-tensor, T, given by $T(Y,Z) = (\nabla_Y \omega)(Z)$. Thus, we can define the second covariant derivative, $\nabla^2_{X,Y}\omega$, of ω as the covariant derivative of T (see Proposition 11.5), that is,

$$(\nabla_X T)(Y,Z) = X(T(Y,Z)) - T(\nabla_X Y, Z) - T(Y, \nabla_X Z),$$

and so

$$\begin{aligned}(\nabla^2_{X,Y}\omega)(Z) &= X((\nabla_Y\omega)(Z)) - (\nabla_{\nabla_X Y}\omega)(Z) - (\nabla_Y\omega)(\nabla_X Z)\\ &= X((\nabla_Y\omega)(Z)) - (\nabla_Y\omega)(\nabla_X Z) - (\nabla_{\nabla_X Y}\omega)(Z)\\ &= (\nabla_X(\nabla_Y\omega))(Z) - (\nabla_{\nabla_X Y}\omega)(Z).\end{aligned}$$

Therefore,

$$\nabla^2_{X,Y}\omega = \nabla_X(\nabla_Y\omega) - \nabla_{\nabla_X Y}\omega,$$

that is, $\nabla^2_{X,Y}\omega$ is formally the same as $\nabla^2_{X,Y}Z$. Then, it is natural to ask what is

$$\nabla^2_{X,Y}\omega - \nabla^2_{Y,X}\omega.$$

The answer is given by the following proposition which plays a crucial role in the proof of a version of Bochner's formula:

Proposition 13.2. *For any vector fields, $X, Y, Z \in \mathfrak{X}(M)$, and any one-form, $\omega \in \mathcal{A}^1(M)$, on a Riemannian manifold, M, we have*

$$((\nabla^2_{X,Y} - \nabla^2_{Y,X})\omega)(Z) = \omega(R(X,Y)Z).$$

Proof. Recall that we proved in Section 11.5 that

$$(\nabla_X\omega)^\sharp = \nabla\omega^\sharp.$$

We claim that we also have

$$(\nabla^2_{X,Y}\omega)^\sharp = \nabla^2_{X,Y}\omega^\sharp.$$

This is because

$$\begin{aligned}
(\nabla^2_{X,Y}\omega)^\sharp &= (\nabla_X(\nabla_Y\omega))^\sharp - (\nabla_{\nabla_X Y}\omega)^\sharp \\
&= \nabla_X(\nabla_Y\omega)^\sharp - \nabla_{\nabla_X Y}\omega^\sharp \\
&= \nabla_X(\nabla_Y\omega^\sharp) - \nabla_{\nabla_X Y}\omega^\sharp \\
&= \nabla^2_{X,Y}\omega^\sharp.
\end{aligned}$$

Thus, we deduce that

$$((\nabla^2_{X,Y} - \nabla^2_{Y,X})\omega)^\sharp = (\nabla^2_{X,Y} - \nabla^2_{Y,X})\omega^\sharp = R(Y,X)\omega^\sharp.$$

Consequently,

$$\begin{aligned}
((\nabla^2_{X,Y} - \nabla^2_{Y,X})\omega)(Z) &= \langle((\nabla^2_{X,Y} - \nabla^2_{Y,X})\omega)^\sharp, Z\rangle \\
&= \langle R(Y,X)\omega^\sharp, Z\rangle \\
&= R(Y,X,\omega^\sharp,Z) \\
&= R(X,Y,Z,\omega^\sharp) \\
&= \langle R(X,Y)Z, \omega^\sharp\rangle \\
&= \omega(R(X,Y)Z),
\end{aligned}$$

where we used properties (3) and (4) of Proposition 13.1. □

The next proposition will be needed in the proof of the second variation formula. If $\alpha\colon U \to M$ is a parametrized surface, where U is some open subset of $\mathbb{R}^2$, we say that a vector field, $V \in \mathfrak{X}(M)$, is a *vector field along* α iff $V(x,y) \in T_{\alpha(x,y)}M$, for all $(x,y) \in U$. For any smooth vector field, V, along α, we also define the covariant derivatives, $DV/\partial x$ and $DV/\partial y$ as follows: For each fixed y_0, if we restrict V to the curve

$$x \mapsto \alpha(x, y_0)$$

we obtain a vector field, V_{y_0}, along this curve and we set

$$\frac{DX}{\partial x}(x, y_0) = \frac{DV_{y_0}}{dx}.$$

Then, we let y_0 vary so that $(x, y_0) \in U$ and this yields $DV/\partial x$. We define $DV/\partial y$ is a similar manner, using a fixed x_0.

Proposition 13.3. *For a Riemannian manifold, $(M, \langle -, - \rangle)$, equipped with the Levi-Civita connection, for every parametrized surface, $\alpha\colon \mathbb{R}^2 \to M$, for every vector field, $V \in \mathfrak{X}(M)$ along α, we have*

$$\frac{D}{\partial y}\frac{D}{\partial x}V - \frac{D}{\partial x}\frac{D}{\partial y}V = R\left(\frac{\partial \alpha}{\partial x}, \frac{\partial \alpha}{\partial y}\right)V.$$

Proof. Express both sides in local coordinates in a chart and make use of the identity

$$\nabla_{\frac{\partial}{\partial x_j}} \nabla_{\frac{\partial}{\partial x_i}} \frac{\partial}{\partial x_k} - \nabla_{\frac{\partial}{\partial x_i}} \nabla_{\frac{\partial}{\partial x_j}} \frac{\partial}{\partial x_k} = R\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right)\frac{\partial}{\partial x_k}.$$ □

Remark: Since the Levi-Civita connection is torsion-free, it is easy to check that

$$\frac{D}{\partial x}\frac{\partial \alpha}{\partial y} = \frac{D}{\partial y}\frac{\partial \alpha}{\partial x}.$$

We used this identity in the proof of Theorem 12.18.

The curvature tensor is a rather complicated object. Thus, it is quite natural to seek simpler notions of curvature. The sectional curvature is indeed a simpler object and it turns out that the curvature tensor can be recovered from it.

13.2 Sectional Curvature

Basically, the sectional curvature is the curvature of two-dimensional sections of our manifold. Given any two vectors, $u, v \in T_pM$, recall by Cauchy-Schwarz that

$$\langle u, v\rangle_p^2 \leq \langle u, u\rangle_p \langle v, v\rangle_p,$$

with equality iff u and v are linearly dependent. Consequently, if u and v are linearly independent, we have

$$\langle u, u\rangle_p \langle v, v\rangle_p - \langle u, v\rangle_p^2 \neq 0.$$

In this case, we claim that the ratio

$$K(u, v) = \frac{R_p(u, v, u, v)}{\langle u, u\rangle_p \langle v, v\rangle_p - \langle u, v\rangle_p^2}$$

is independent of the plane, Π, spanned by u and v. If (x, y) is another basis of Π, then

$$\begin{aligned} x &= au + bv \\ y &= cu + dv. \end{aligned}$$

We get

$$\langle x, x\rangle_p \langle y, y\rangle_p - \langle x, y\rangle_p^2 = (ad - bc)^2(\langle u, u\rangle_p \langle v, v\rangle_p - \langle u, v\rangle_p^2)$$

and similarly,

$$R_p(x, y, x, y) = \langle R_p(x, y)x, y\rangle_p = (ad - bc)^2 R_p(u, v, u, v),$$

which proves our assertion.

Definition 13.2. Let $(M, \langle -, -\rangle)$ be any Riemannian manifold equipped with the Levi-Civita connection. For every $p \in T_pM$, for every 2-plane, $\Pi \subseteq T_pM$, the *sectional curvature*, $K(\Pi)$, of Π, is given by

$$K(\Pi) = K(x, y) = \frac{R_p(x, y, x, y)}{\langle x, x\rangle_p \langle y, y\rangle_p - \langle x, y\rangle_p^2},$$

for any basis, (x, y), of Π.

Observe that if (x, y) is an orthonormal basis, then the denominator is equal to 1. The expression $R_p(x, y, x, y)$ is often denoted $\kappa_p(x, y)$. Remarkably, κ_p determines R_p. We denote the function $p \mapsto \kappa_p$ by κ. We state the following proposition without proof:

Proposition 13.4. *Let $(M, \langle -, -\rangle)$ be any Riemannian manifold equipped with the Levi-Civita connection. The function κ determines the curvature tensor, R. Thus, the knowledge of all the sectional curvatures determines the curvature tensor. Moreover, we have*

$$\begin{aligned} 6\langle R(x, y)z, w\rangle &= \kappa(x + w, y + z) - \kappa(x, y + z) - \kappa(w, y + z) \\ &\quad - \kappa(y + w, x + z) + \kappa(y, x + z) + \kappa(w, x + z) \\ &\quad - \kappa(x + w, y) + \kappa(x, y) + \kappa(w, y) \\ &\quad - \kappa(x + w, z) + \kappa(x, z) + \kappa(w, z) \\ &\quad + \kappa(y + w, x) - \kappa(y, x) - \kappa(w, x) \\ &\quad + \kappa(y + w, z) - \kappa(y, z) - \kappa(w, z). \end{aligned}$$

For a proof of this formidable equation, see Kuhnel [92] (Chapter 6, Theorem 6.5). A different proof of the above proposition (without an explicit formula) is also given in O'Neill [120] (Chapter III, Corollary 42).

Let

$$R_1(x,y)z = \langle x,z\rangle y - \langle y,z\rangle x.$$

Observe that

$$\langle R_1(x,y)x,y\rangle = \langle x,x\rangle\langle y,y\rangle - \langle x,y\rangle^2.$$

As a corollary of Proposition 13.4, we get:

Proposition 13.5. *Let $(M,\langle -,-\rangle)$ be any Riemannian manifold equipped with the Levi-Civita connection. If the sectional curvature, $K(\Pi)$ does not depend on the plane, Π, but only on $p \in M$, in the sense that K is a scalar function, $K\colon M \to \mathbb{R}$, then*

$$R = KR_1.$$

Proof. By hypothesis,

$$\kappa_p(x,y) = K(p)(\langle x,x\rangle_p\langle y,y\rangle_p - \langle x,y\rangle_p^2),$$

for all x, y. As the right-hand side of the formula in Proposition 13.4 consists of a sum of terms, we see that the right-hand side is equal to K times a similar sum with κ replaced by

$$\langle R_1(x,y)x,y\rangle = \langle x,x\rangle\langle y,y\rangle - \langle x,y\rangle^2,$$

so it is clear that $R = KR_1$. □

In particular, in dimension $n = 2$, the assumption of Proposition 13.5 holds and K is the well-known *Gaussian curvature* for surfaces.

Definition 13.3. A Riemannian manifold, $(M,\langle -,-\rangle)$ is said to have *constant (resp. negative, resp. positive) curvature* iff its sectional curvature is constant (resp. negative, resp. positive).

In dimension $n \geq 3$, we have the following somewhat surprising theorem due to F. Schur:

Proposition 13.6. *(F. Schur, 1886) Let $(M,\langle -,-\rangle)$ be a connected Riemannian manifold. If $\dim(M) \geq 3$ and if the sectional curvature, $K(\Pi)$, does not depend on the plane, $\Pi \subseteq T_pM$, but only on the point, $p \in M$, then K is constant (i.e., does not depend on p).*

The proof, which is quite beautiful, can be found in Kuhnel [92] (Chapter 6, Theorem 6.7).

If we replace the metric, $g = \langle -, - \rangle$ by the metric $\widetilde{g} = \lambda \langle -, - \rangle$ where $\lambda > 0$ is a constant, some simple calculations show that the Christoffel symbols and the Levi-Civita connection are unchanged, as well as the curvature tensor, but the sectional curvature is changed, with

$$\widetilde{K} = \lambda^{-1} K.$$

As a consequence, if M is a Riemannian manifold of constant curvature, by rescaling the metric, we may assume that either $K = -1$, or $K = 0$, or $K = +1$. Here are standard examples of spaces with constant curvature.

(1) The sphere, $S^n \subseteq \mathbb{R}^{n+1}$, with the metric induced by $\mathbb{R}^{n+1}$, where

$$S^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1^2 + \cdots + x_{n+1}^2 = 1\}.$$

The sphere, S^n, has constant sectional curvature, $K = +1$. This can be shown by using the fact that the stabilizer of the action of $\mathbf{SO}(n+1)$ on S^n is isomorphic to $\mathbf{SO}(n)$. Then, it is easy to see that the action of $\mathbf{SO}(n)$ on $T_p S^n$ is transitive on 2-planes and from this, it follows that $K = 1$ (for details, see Gallot, Hulin and Lafontaine [61] (Chapter 3, Proposition 3.14).

(2) Euclidean space, $\mathbb{R}^{n+1}$, with its natural Euclidean metric. Of course, $K = 0$.

(3) The *hyperbolic space*, $\mathcal{H}_n^+(1)$, from Definition 2.10. Recall that this space is defined in terms of the *Lorentz innner product*, $\langle -, - \rangle_1$, on $\mathbb{R}^{n+1}$, given by

$$\langle (x_1, \ldots, x_{n+1}), (y_1, \ldots, y_{n+1}) \rangle_1 = -x_1 y_1 + \sum_{i=2}^{n+1} x_i y_i.$$

By definition, $\mathcal{H}_n^+(1)$, written simply H^n, is given by

$$H^n = \{x = (x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid \langle x, x \rangle_1 = -1,\ x_1 > 0\}.$$

Given any points, $p = (x_1, \ldots, x_{n+1}) \in H^n$, it is easy to see that the set of tangent vectors, $u \in T_p H^n$, are given by the equation

$$\langle p, u \rangle_1 = 0,$$

that is, $T_p H^n$ is orthogonal to p with respect to the Lorentz inner-product. Since $p \in H^n$, we have $\langle p, p \rangle_1 = -1$, that is, u is lightlike, so by Proposition 2.10, all vectors in $T_p H^n$ are spacelike, that is,

$$\langle u, u \rangle_1 > 0, \qquad \text{for all } u \in T_p H^n,\ u \neq 0.$$

Therefore, the restriction of $\langle -, - \rangle_1$ to H^n is positive, definite, which means that it is a metric on $T_p H^n$. The space H^n equipped with this metric, g_H, is called *hyperbolic space* and it has constant curvature, $K = -1$. This can be shown by using the fact that the stabilizer of the action of $\mathbf{SO}_0(n, 1)$ on H^n is isomorphic to $\mathbf{SO}(n)$ (see Proposition 2.11). Then, it is easy to see that the action of $\mathbf{SO}(n)$ on $T_p H^n$ is transitive on 2-planes and from this, it follows that $K = -1$ (for details, see Gallot, Hulin and Lafontaine [61] (Chapter 3, Proposition 3.14).

There are other isometric models of H^n that are perhaps intuitively easier to grasp but for which the metric is more complicated. For example, there is a map, $\mathrm{PD}\colon B^n \to H^n$, where $B^n = \{x \in \mathbb{R}^n \mid \|x\| < 1\}$ is the open unit ball in $\mathbb{R}^n$, given by

$$\mathrm{PD}(x) = \left(\frac{1 + \|x\|^2}{1 - \|x\|^2}, \frac{2x}{1 - \|x\|^2}\right).$$

It is easy to check that $\langle \mathrm{PD}(x), \mathrm{PD}(x)\rangle_1 = -1$ and that PD is bijective and an isometry. One also checks that the pull-back metric, $g_{\mathrm{PD}} = \mathrm{PD}^* g_H$, on B^n, is given by

$$g_{\mathrm{PD}} = \frac{4}{(1 - \|x\|^2)^2}\,(dx_1^2 + \cdots + dx_n^2).$$

The metric, g_{PD} is called the *conformal disc metric* and the Riemannian manifold, (B^n, g_{PD}) is called the *Poincaré disc model* or *conformal disc model*. The metric g_{PD} is proportional to the Euclidean metric and thus, angles are preserved under the map PD. Another model is the *Poincaré half-plane model*, $\{x \in \mathbb{R}^n \mid x_1 > 0\}$, with the metric

$$g_{\mathrm{PH}} = \frac{1}{x_1^2}\,(dx_1^2 + \cdots + dx_n^2).$$

We already encountered this space for $n = 2$.

The metrics for S^n, $\mathbb{R}^{n+1}$ and H^n have a nice expression in polar coordinates but we prefer to discuss the Ricci curvature next.

13.3 Ricci Curvature

The Ricci tensor is another important notion of curvature. It is mathematically simpler than the sectional curvature (since it is symmetric) but it plays an important role in the theory of gravitation as it occurs in the Einstein field equations. The Ricci tensor is an example of contraction, in this case, the trace of a linear map. Recall that if $f\colon E \to E$ is a linear map from a finite-dimensional Euclidean vector space to itself, given any orthonormal basis, $(e_1, \ldots, e_n)$, we have

$$\mathrm{tr}(f) = \sum_{i=1}^{n} \langle f(e_i), e_i\rangle.$$

Definition 13.4. Let $(M, \langle -, -\rangle)$ be a Riemannian manifold (equipped with the Levi-Civita connection). The *Ricci curvature*, Ric, of M is the $(0,2)$-tensor defined as follows: For every $p \in M$, for all $x, y \in T_pM$, set $\mathrm{Ric}_p(x, y)$ to be the trace of the endomorphism, $v \mapsto R_p(x, v)y$. With respect to any orthonormal basis, $(e_1, \ldots, e_n)$, of T_pM, we have

$$\mathrm{Ric}_p(x, y) = \sum_{j=1}^{n} \langle R_p(x, e_j)y, e_j\rangle_p = \sum_{j=1}^{n} R_p(x, e_j, y, e_j).$$

The *scalar curvature*, S, of M, is the trace of the Ricci curvature, that is, for every $p \in M$,

$$S(p) = \sum_{i \neq j} R(e_i, e_j, e_i, e_j) = \sum_{i \neq j} K(e_i, e_j),$$

where $K(e_i, e_j)$ denotes the sectional curvature of the plane spanned by e_i, e_j.

In view of Proposition 13.1 (4), the Ricci curvature is symmetric. The tensor Ric is a $(0,2)$-tensor but it can be interpreted as a $(1,1)$-tensor as follows: We let $\mathrm{Ric}_p^{\#}$ be the $(1,1)$-tensor given by

$$\langle \mathrm{Ric}_p^{\#} u, v\rangle_p = \mathrm{Ric}(u, v),$$

for all $u, v \in T_pM$. Then, it is easy to see that

$$S(p) = \mathrm{tr}(\mathrm{Ric}_p^{\#}).$$

This is why we said (by abuse of language) that S is the trace of Ric. Observe that if $(e_1, \ldots, e_n)$ is any orthonormal basis of T_pM, as

$$\begin{aligned} \mathrm{Ric}_p(u, v) &= \sum_{j=1}^{n} R_p(u, e_j, v, e_j) \\ &= \sum_{j=1}^{n} R_p(e_j, u, e_j, v) \\ &= \sum_{j=1}^{n} \langle R_p(e_j, u)e_j, v\rangle_p, \end{aligned}$$

we have

$$\mathrm{Ric}_p^{\#}(u) = \sum_{j=1}^{n} R_p(e_j, u)e_j.$$

Observe that in dimension $n = 2$, we get $S(p) = 2K(p)$. Therefore, in dimension 2, the scalar curvature determines the curvature tensor. In dimension $n = 3$, it turns out that the Ricci tensor completely determines the curvature tensor, although this is not obvious. We will come back to this point later.

Since $\mathrm{Ric}(x, y)$ is symmetric, $\mathrm{Ric}(x, x)$ determines $\mathrm{Ric}(x, y)$ completely (Use the polarization identity for a symmetric bilinear form, φ:

$$2\varphi(x, y) = \varphi(x + y) - \varphi(x) - \varphi(y).)$$

Observe that for any orthonormal frame, $(e_1, \ldots, e_n)$, of T_pM, using the definition of the sectional curvature, K, we have

$$\mathrm{Ric}(e_1, e_1) = \sum_{i=1}^{n} \langle (R(e_1, e_i)e_1, e_i\rangle = \sum_{i=2}^{n} K(e_1, e_i).$$

Thus, $\mathrm{Ric}(e_1, e_1)$ is the sum of the sectional curvatures of any $n-1$ orthogonal planes orthogonal to e_1 (a unit vector).

For a Riemannian manifold with constant sectional curvature, we see that

$$\mathrm{Ric}(x, x) = (n-1)Kg(x, x), \qquad S = n(n-1)K,$$

where $g = \langle -, - \rangle$ is the metric on M. Indeed, if K is constant, then we know that $R = KR_1$ and so,

$$\begin{aligned}
\mathrm{Ric}(x, x) &= K \sum_{i=1}^{n} g(R_1(x, e_i)x, e_i) \\
&= K \sum_{i=1}^{n} (g(e_i, e_i)g(x, x) - g(e_i, x)^2) \\
&= K(ng(x, x) - \sum_{i=1}^{n} g(e_i, x)^2) \\
&= (n-1)Kg(x, x).
\end{aligned}$$

Spaces for which the Ricci tensor is proportional to the metric are called Einstein spaces.

Definition 13.5. A Riemannian manifold, (M, g), is called an *Einstein space* iff the Ricci curvature is proportional to the metric, g, that is:

$$\mathrm{Ric}(x, y) = \lambda g(x, y),$$

for some function, $\lambda \colon M \to \mathbb{R}$.

If M is an Einstein space, observe that $S = n\lambda$.

Remark: For any Riemanian manifold, (M, g), the quantity

$$G = \mathrm{Ric} - \frac{S}{2} g$$

is called the *Einstein tensor* (or *Einstein gravitation tensor* for space-times spaces). The Einstein tensor plays an important role in the theory of general relativity. For more on this topic, see Kuhnel [92] (Chapters 6 and 8) O'Neill [120] (Chapter 12).

13.4 Isometries and Local Isometries

Recall that a *local isometry* between two Riemannian manifolds, M and N, is a smooth map, $\varphi \colon M \to N$, so that

$$\langle (d\varphi)_p(u), (d\varphi_p)(v) \rangle_{\varphi(p)} = \langle u, v \rangle_p,$$

for all $p \in M$ and all $u, v \in T_pM$. An *isometry* is a local isometry and a diffeomorphism.

By the inverse function theorem, if $\varphi\colon M \to N$ is a local isometry, then for every $p \in M$, there is some open subset, $U \subseteq M$, with $p \in U$, so that $\varphi \restriction U$ is an isometry between U and $\varphi(U)$.

Also recall that if $\varphi\colon M \to N$ is a diffeomorphism, then for any vector field, X, on M, the vector field, $\varphi_* X$, on N (called the *push-forward* of X) is given by

$$(\varphi_* X)_q = d\varphi_{\varphi^{-1}(q)} X(\varphi^{-1}(q)), \qquad \text{for all } q \in N,$$

or equivalently, by

$$(\varphi_* X)_{\varphi(p)} = d\varphi_p X(p), \qquad \text{for all } p \in M.$$

For any smooth function, $h\colon N \to \mathbb{R}$, for any $q \in N$, we have

$$\begin{aligned} X_*(h)_q &= dh_q(X_*(q)) \\ &= dh_q(d\varphi_{\varphi^{-1}(q)} X(\varphi^{-1}(q))) \\ &= d(h \circ \varphi)_{\varphi^{-1}(q)} X(\varphi^{-1}(q)) \\ &= X(h \circ \varphi)_{\varphi^{-1}(q)}, \end{aligned}$$

that is

$$X_*(h)_q = X(h \circ \varphi)_{\varphi^{-1}(q)},$$

or

$$X_*(h)_{\varphi(p)} = X(h \circ \varphi)_p.$$

It is natural to expect that isometries preserve all "natural" Riemannian concepts and this is indeed the case. We begin with the Levi-Civita connection.

Proposition 13.7. *If $\varphi\colon M \to N$ is an isometry, then*

$$\varphi_*(\nabla_X Y) = \nabla_{\varphi_* X}(\varphi_* Y), \qquad \textit{for all } X, Y \in \mathfrak{X}(M),$$

where $\nabla_X Y$ is the Levi-Civita connection induced by the metric on M and similarly on N.

Proof. We use the Koszul formula (Proposition 11.18),

$$\begin{aligned} 2\langle \nabla_X Y, Z\rangle &= X(\langle Y, Z\rangle) + Y(\langle X, Z\rangle) - Z(\langle X, Y\rangle) \\ &\quad - \langle Y, [X, Z]\rangle - \langle X, [Y, Z]\rangle - \langle Z, [Y, X]\rangle. \end{aligned}$$

We have

$$(\varphi_*(\nabla_X Y))_{\varphi(p)} = d\varphi_p(\nabla_X Y)_p,$$

and as φ is an isometry,

$$\langle d\varphi_p(\nabla_X Y)_p, d\varphi_p Z_p\rangle_{\varphi(p)} = \langle (\nabla_X Y)_p, Z_p\rangle_p,$$

so, Koszul yields

$$\begin{aligned} 2\langle\varphi_*(\nabla_X Y),\varphi_* Z\rangle_{\varphi(p)} &= X(\langle Y,Z\rangle_p)+Y(\langle X,Z\rangle_p)-Z(\langle X,Y\rangle_p)\\ &\quad -\langle Y,[X,Z]\rangle_p-\langle X,[Y,Z]\rangle_p-\langle Z,[Y,X]\rangle_p. \end{aligned}$$

Next, we need to compute

$$\langle\nabla_{\varphi_* X}(\varphi_* Y),\varphi_* Z\rangle_{\varphi(p)}.$$

When we plug $\varphi_* X$, $\varphi_* Y$ and $\varphi_* Z$ into the Koszul formula, as φ is an isometry, for the fourth term on the right-hand side, we get

$$\begin{aligned} \langle\varphi_* Y,[\varphi_* X,\varphi_* Z]\rangle_{\varphi(p)} &= \langle d\varphi_p Y_p,[d\varphi_p X_p,d\varphi_p Z_p]\rangle_{\varphi(p)}\\ &= \langle Y_p,[X_p,Z_p]\rangle_p \end{aligned}$$

and similarly for the fifth and sixth term on the right-hand side. For the first term on the right-hand side, we get

$$\begin{aligned} (\varphi_* X)(\langle\varphi_* Y,\varphi_* Z\rangle)_{\varphi(p)} &= (\varphi_* X)(\langle d\varphi_p Y_p,d\varphi_p Z_p\rangle)_{\varphi(p)}\\ &= (\varphi_* X)(\langle Y_p,Z_p\rangle_{\varphi^{-1}(\varphi(p))})_{\varphi(p)}\\ &= (\varphi_* X)(\langle Y,Z\rangle\circ\varphi^{-1})_{\varphi(p)}\\ &= X(\langle Y,Z\rangle\circ\varphi^{-1}\circ\varphi)_p\\ &= X(\langle Y,Z\rangle)_p \end{aligned}$$

and similarly for the second and third term. Consequently, we get

$$\begin{aligned} 2\langle\nabla_{\varphi_* X}(\varphi_* Y),\varphi_* Z\rangle_{\varphi(p)} &= X(\langle Y,Z\rangle_p)+Y(\langle X,Z\rangle_p)-Z(\langle X,Y\rangle_p)\\ &\quad -\langle Y,[X,Z]\rangle_p-\langle X,[Y,Z]\rangle_p-\langle Z,[Y,X]\rangle_p. \end{aligned}$$

By comparing right-hand sides, we get

$$2\langle\varphi_*(\nabla_X Y),\varphi_* Z\rangle_{\varphi(p)}=2\langle\nabla_{\varphi_* X}(\varphi_* Y),\varphi_* Z\rangle_{\varphi(p)}$$

for all X,Y,Z, and as φ is a diffeomorphism, this implies

$$\varphi_*(\nabla_X Y)=\nabla_{\varphi_* X}(\varphi_* Y),$$

as claimed. □

As a corollary of Proposition 13.7, the curvature induced by the connection is preserved, that is

$$\varphi_* R(X,Y)Z=R(\varphi_* X,\varphi_* Y)\varphi_* Z,$$

as well as the parallel transport, the covariant derivative of a vector field along a curve, the exponential map, sectional curvature, Ricci curvature and geodesics. Actually, all concepts that are local in nature are preserved by local diffeomorphisms! So, except for the Levi-Civita connection and if we consider the Riemann tensor on vectors, all the above concepts are preserved under local diffeomorphisms. For the record, we state:

Proposition 13.8. *If $\varphi\colon M \to N$ is a local isometry, then the following concepts are preserved:*

(1) The covariant derivative of vector fields along a curve, γ, that is

$$d\varphi_{\gamma(t)}\frac{DX}{dt} = \frac{D\varphi_* X}{dt},$$

for any vector field, X, along γ, with $(\varphi_* X)(t) = d\varphi_{\gamma(t)}Y(t)$, for all t.

(2) Parallel translation along a curve. If P_γ denotes parallel transport along the curve γ and if $P_{\varphi\circ\gamma}$ denotes parallel transport along the curve $\varphi\circ\gamma$, then

$$d\varphi_{\gamma(1)} \circ P_\gamma = P_{\varphi\circ\gamma} \circ d\varphi_{\gamma(0)}.$$

(3) Geodesics. If γ is a geodesic in M, then $\varphi\circ\gamma$ is a geodesic in N. Thus, if γ_v is the unique geodesic with $\gamma(0) = p$ and $\gamma_v'(0) = v$, then

$$\varphi \circ \gamma_v = \gamma_{d\varphi_p v},$$

wherever both sides are defined. Note that the domain of $\gamma_{d\varphi_p v}$ may be strictly larger than the domain of γ_v. For example, consider the inclusion of an open disc into $\mathbb{R}^2$.

(4) Exponential maps. We have

$$\varphi \circ \exp_p = \exp_{\varphi(p)} \circ d\varphi_p,$$

wherever both sides are defined.

(5) Riemannian curvature tensor. We have

$$d\varphi_p R(x,y)z = R(d\varphi_p x, d\varphi_p y)d\varphi_p z, \qquad \text{for all } x, y, z \in T_pM.$$

(6) Sectional, Ricci and Scalar curvature. We have

$$K(d\varphi_p x, d\varphi_p y) = K(x,y)_p,$$

for all linearly independent vectors, $x, y \in T_pM$;

$$\mathrm{Ric}(d\varphi_p x, d\varphi_p y) = \mathrm{Ric}(x,y)_p$$

for all $x, y \in T_pM$;

$$S_M = S_N \circ \varphi.$$

where S_M is the scalar curvature on M and S_N is the scalar curvature on N.

A useful property of local diffeomorphisms is stated below. For a proof, see O'Neill [120] (Chapter 3, Proposition 62):

Proposition 13.9. *Let $\varphi, \psi \colon M \to N$ be two local isometries. If M is connected and if $\varphi(p) = \psi(p)$ and $d\varphi_p = d\psi_p$ for some $p \in M$, then $\varphi = \psi$.*

The idea is to prove that

$$\{p \in M \mid d\varphi_p = d\psi_p\}$$

is both open and closed and for this, to use the preservation of the exponential under local diffeomorphisms.

13.5 Riemannian Covering Maps

The notion of covering map discussed in Section 3.9 can be extended to Riemannian manifolds.

Definition 13.6. If M and N are two Riemannian manifold, then a map, $\pi \colon M \to N$, is a *Riemannian covering* iff the following conditions hold:

(1) The map π is a smooth covering map.

(2) The map π is a local isometry.

Recall from Section 3.9 that a covering map is a local diffeomorphism. A way to obtain a metric on a manifold, M, is to pull-back the metric, g, on a manifold, N, along a local diffeomorphism, $\varphi \colon M \to N$ (see Section 7.4). If φ is a covering map, then it becomes a Riemannian covering map.

Proposition 13.10. *Let $\pi \colon M \to N$ be a smooth covering map. For any Riemannian metric, g, on N, there is a unique metric, $\pi^* g$, on M, so that π is a Riemannian covering.*

Proof. We define the *pull-back metric*, $\pi^* g$, on M induced by g as follows: For all $p \in M$, for all $u, v \in T_p M$,

$$(\pi^* g)_p(u, v) = g(d\pi_p(u), d\pi_p(v)).$$

We need to check that $(\pi^* g)_p$ is an inner product, which is very easy since $d\pi_p$ is a linear isomorphism. Our map, π, between the two Riemannian manifolds $(M, \pi^* g)$ and (N, g) becomes a local isometry. Now, every metric on M making π a local isometry has to satisfy the equation defining, $\pi^* g$, so this metric is unique. □

As a corollary of Proposition 13.10 and Theorem 3.41, every connected Riemmanian manifold, M, has a simply connected covering map, $\pi \colon \widetilde{M} \to M$, where π is a Riemannian covering. Furthermore, if $\pi \colon M \to N$ is a Riemannian covering and $\varphi \colon P \to N$ is a local isometry, it is easy to see that its lift, $\widetilde{\varphi} \colon P \to M$, is also a local isometry. In particular, the deck-transformations of a Riemannian covering are isometries.

In general, a local isometry is not a Riemannian covering. However, this is the case when the source space is complete.

Proposition 13.11. *Let $\pi\colon M \to N$ be a local isometry with N connected. If M is a complete manifold, then π is a Riemannian covering map.*

Proof. We follow the proof in Sakai [131] (Chapter III, Theorem 5.4). Because π is a local isometry, geodesics in M can be projected onto geodesics in N and geodesics in N can be lifted back to M. The proof makes heavy use of these facts.

First, we prove that N is complete. Pick any $p \in M$ and let $q = \pi(p)$. For any geodesic, γ_u, of N with initial point, $q \in N$, and initial direction the unit vector, $u \in T_qN$, consider the geodesic, $\widetilde{\gamma}_u$, of M, with initial point p and with $u = d\pi_q^{-1}(v) \in T_pM$. As π is a local isometry, it preserves geodesic, so

$$\gamma_v = \pi \circ \widetilde{\gamma}_u,$$

and since $\widetilde{\gamma}_u$ is defined on $\mathbb{R}$ because M is complete, so if γ_v. As $\exp_q$ is defined on the whole of T_qN, by Hopf-Rinow, N is complete.

Next, we prove that π is surjective. As N is complete, for any $q_1 \in N$, there is a minimal geodesic, $\gamma\colon [0, b] \to N$, joining q to q_1 and for the geodesic, $\widetilde{\gamma}$, in M, emanating from p and with initial direction $d\pi_q^{-1}(\gamma'(0))$, we have $\pi(\widetilde{\gamma}(b)) = \gamma(b) = q_1$, establishing surjectivity.

For any $q \in N$, pick $r > 0$ wih $r < i(q)$, where $i(q)$ denotes the injectivity radius of N at q and consider the open metric ball, $B_r(q) = \exp_q(B(0_q, r))$ (where $B(0_q, r)$ is the open ball of radius r in T_qN). Let

$$\pi^{-1}(q) = \{p_i\}_{i\in I} \subseteq M.$$

We claim that the following properties hold:

(1) Each map, $\pi \upharpoonright B_r(p_i)\colon B_r(p_i) \longrightarrow B_r(q)$, is a diffeomorphism, in fact, an isometry.

(2) $\pi^{-1}(B_r(q)) = \bigcup_{i\in I} B_r(p_i)$.

(3) $B_r(p_i) \cap B_r(p_j) = \emptyset$ whenever $i \neq j$.

It follows from (1), (2) and (3) that $B_r(q)$ is evenly covered by the family of open sets, $\{B_r(p_i)\}_{i\in I}$, so π is a covering map.

(1) Since π is a local isometry, it maps geodesics emanating from p_i to geodesics emanating from q so the following diagram commutes:

$$\begin{array}{ccc} B(0_{p_i}, r) & \xrightarrow{d\pi_{p_i}} & B(0_q, r) \\ {\scriptstyle \exp_{p_i}}\downarrow & & \downarrow{\scriptstyle \exp_q} \\ B_r(p_i) & \xrightarrow[\pi]{} & B_r(q). \end{array}$$

Since $\exp_q \circ d\pi_{p_i}$ is a diffeomorphism, $\pi \upharpoonright B_r(p_i)$ must be injective and since $\exp_{p_i}$ is surjective, so is $\pi \upharpoonright B_r(p_i)$. Then, $\pi \upharpoonright B_r(p_i)$ is a bijection and as π is a local diffeomorphism, $\pi \upharpoonright B_r(p_i)$ is a diffeomorphism.

(2) Obviously, $\bigcup_{i\in I} B_r(p_i) \subseteq \pi^{-1}(B_r(q))$, by (1). Conversely, pick $p_1 \in \pi^{-1}(B_r(q))$. For $q_1 = \pi(p_1)$, we can write $q_1 = \exp_q v$, for some $v \in B(0_q, r)$ and the map $\gamma(t) = \exp_q(1-t)v$, for $t \in [0,1]$, is a geodesic in N joining q_1 to q. Then, we have the geodesic, $\widetilde{\gamma}$, emanating from p_1 with initial direction $d\pi_{q_1}^{-1}(\gamma'(0))$ and as $\pi \circ \widetilde{\gamma}(1) = \gamma(1) = q$, we have $\widetilde{\gamma}(1) = p_i$ for some α. Since γ has length less than r, we get $p_1 \in B_r(p_i)$.

(3) Suppose $p_1 \in B_r(p_i) \cap B_r(p_j)$. We can pick a minimal geodesic, $\widetilde{\gamma}$, in $B_r(p_i)$, (resp $\widetilde{\omega}$ in $B_r(p_j)$) joining p_i to p (resp. joining p_j to p). Then, the geodesics $\pi \circ \widetilde{\gamma}$ and $\pi \circ \widetilde{\omega}$ are geodesics in $B_r(q)$ from q to $\pi(p_1)$ and their length is less than r. Since $r < i(q)$, these geodesics are minimal so they must coincide. Therefore, $\gamma = \omega$, which implies $i = j$. □

13.6 The Second Variation Formula and the Index Form

In Section 12.4, we discovered that the geodesics are exactly the critical paths of the energy functional (Theorem 12.19). For this, we derived the First Variation Formula (Theorem 12.18). It is not too surprising that a deeper understanding is achieved by investigating the second derivative of the energy functional at a critical path (a geodesic). By analogy with the Hessian of a real-valued function on $\mathbb{R}^n$, it is possible to define a bilinear functional,

$$I_\gamma \colon T_\gamma\Omega(p,q) \times T_\gamma\Omega(p,q) \to \mathbb{R},$$

when γ is a critical point of the energy function, E (that is, γ is a geodesic). This bilinear form is usually called the *index form*. Note that Milnor denotes I_γ by E_{**} and refers to it as the *Hessian* of E but this is a bit confusing since I_γ is only defined for critical points, whereas the Hessian is defined for all points, critical or not.

Now, if $f \colon M \to \mathbb{R}$ is a real-valued function on a finite-dimensional manifold, M, and if p is a critical point of f, which means that $df_p = 0$, it is not hard to prove that there is a symmetric bilinear map, $I \colon T_pM \times T_pM \to \mathbb{R}$, such that

$$I(X(p), Y(p)) = X_p(Yf) = Y_p(Xf),$$

for all vector fields, $X, Y \in \mathfrak{X}(M)$. Furthermore, $I(u,v)$ can be computed as follows, for any $u, v \in T_pM$: for any smooth map, $\alpha \colon \mathbb{R}^2 \to \mathbb{R}$, such that

$$\alpha(0,0) = p, \quad \frac{\partial \alpha}{\partial x}(0,0) = u, \quad \frac{\partial \alpha}{\partial y}(0,0) = v,$$

we have

$$I(u,v) = \left.\frac{\partial^2 (f \circ \alpha)(x,y)}{\partial x \partial y}\right|_{(0,0)}.$$

The above suggests that in order to define

$$I_\gamma \colon T_\gamma\Omega(p,q) \times T_\gamma\Omega(p,q) \to \mathbb{R},$$

that is, to define $I_\gamma(W_1, W_2)$, where $W_1, W_2 \in T_\gamma\Omega(p,q)$ are vector fields along γ (with $W_1(0) = W_2(0) = 0$ and $W_1(1) = W_2(1) = 0$), we consider 2-parameter variations,

$$\alpha\colon U \times [0,1] \to M,$$

where U is an open subset of $\mathbb{R}^2$ with $(0,0) \in U$, such that

$$\alpha(0,0,t) = \gamma(t), \quad \frac{\partial\alpha}{\partial u_1}(0,0,t) = W_1(t), \quad \frac{\partial\alpha}{\partial u_2}(0,0,t) = W_2(t).$$

Then, we set

$$I_\gamma(W_1, W_2) = \left.\frac{\partial^2(E \circ \widetilde{\alpha})(u_1,u_2)}{\partial u_1 \partial u_2}\right|_{(0,0)},$$

where $\widetilde{\alpha} \in \Omega(p,q)$ is the path given by

$$\widetilde{\alpha}(u_1,u_2)(t) = \alpha(u_1,u_2,t).$$

For simplicity of notation, the above derivative if often written as $\frac{\partial^2 E}{\partial u_1 \partial u_2}(0,0)$.

To prove that $I_\gamma(W_1, W_2)$ is actually well-defined, we need the following result:

Theorem 13.12. *(Second Variation Formula) Let $\alpha\colon U \times [0,1] \to M$ be a 2-parameter variation of a geodesic, $\gamma \in \Omega(p,q)$, with variation vector fields $W_1, W_2 \in T_\gamma\Omega(p,q)$ given by*

$$W_1(t) = \frac{\partial\alpha}{\partial u_1}(0,0,t), \quad W_2(t) = \frac{\partial\alpha}{\partial u_2}(0,0,t).$$

Then, we have the formula

$$\frac{1}{2}\left.\frac{\partial^2(E \circ \widetilde{\alpha})(u_1,u_2)}{\partial u_1 \partial u_2}\right|_{(0,0)} = -\sum_t \left\langle W_2(t), \Delta_t \frac{dW_1}{dt}\right\rangle - \int_0^1 \left\langle W_2, \frac{D^2W_1}{dt^2} + R(V, W_1)V\right\rangle dt,$$

where $V(t) = \gamma'(t)$ is the velocity field,

$$\Delta_t \frac{dW_1}{dt} = \frac{dW_1}{dt}(t_+) - \frac{dW_1}{dt}(t_-)$$

is the jump in $\frac{dW_1}{dt}$ at one of its finitely many points of discontinuity in $(0,1)$ and E is the energy function on $\Omega(p,q)$.

Proof. (After Milnor, see [107], Chapter II, Section 13, Theorem 13.1.) By the First Variation Formula (Theorem 12.18), we have

$$\frac{1}{2}\frac{\partial E(\widetilde{\alpha}(u_1,u_2))}{\partial u_2} = -\sum_i \left\langle \frac{\partial\alpha}{\partial u_2}, \Delta_t \frac{\partial\alpha}{\partial t}\right\rangle - \int_0^1 \left\langle \frac{\partial\alpha}{\partial u_2}, \frac{D}{\partial t}\frac{\partial\alpha}{\partial t}\right\rangle dt.$$

Thus, we get

$$\begin{aligned}\frac{1}{2}\frac{\partial^2(E\circ\widetilde{\alpha})(u_1,u_2)}{\partial u_1\partial u_2} &= -\sum_i\left\langle\frac{D}{\partial u_1}\frac{\partial\alpha}{\partial u_2},\Delta_t\frac{\partial\alpha}{\partial t}\right\rangle-\sum_i\left\langle\frac{\partial\alpha}{\partial u_2},\frac{D}{\partial u_1}\Delta_t\frac{\partial\alpha}{\partial t}\right\rangle\\ &\quad-\int_0^1\left\langle\frac{D}{\partial u_1}\frac{\partial\alpha}{\partial u_2},\frac{D}{\partial t}\frac{\partial\alpha}{\partial t}\right\rangle dt-\int_0^1\left\langle\frac{\partial\alpha}{\partial u_2},\frac{D}{\partial u_1}\frac{D}{\partial t}\frac{\partial\alpha}{\partial t}\right\rangle dt.\end{aligned}$$

Let us evaluate this expression for $(u_1,u_2)=(0,0)$. Since $\gamma=\widetilde{\alpha}(0,0)$ is an unbroken geodesic, we have

$$\Delta_t\frac{\partial\alpha}{\partial t}=0,\qquad\frac{D}{\partial t}\frac{\partial\alpha}{\partial t}=0,$$

so that the first and third term are zero. As

$$\frac{D}{\partial u_1}\frac{\partial\alpha}{\partial t}=\frac{D}{\partial t}\frac{\partial\alpha}{\partial u_1},$$

(see the remark just after Proposition 13.3), we can rewrite the second term and we get

$$\frac{1}{2}\frac{\partial^2(E\circ\widetilde{\alpha})(u_1,u_2)}{\partial u_1\partial u_2}(0,0)=-\sum_i\left\langle W_2,\Delta_t\frac{D}{\partial t}W_1\right\rangle-\int_0^1\left\langle W_2,\frac{D}{\partial u_1}\frac{D}{\partial t}V\right\rangle dt.\qquad(*)$$

In order to interchange the operators $\frac{D}{\partial u_1}$ and $\frac{D}{\partial t}$, we need to bring in the curvature tensor. Indeed, by Proposition 13.3, we have

$$\frac{D}{\partial u_1}\frac{D}{\partial t}V-\frac{D}{\partial t}\frac{D}{\partial u_1}V=R\left(\frac{\partial\alpha}{\partial t},\frac{\partial\alpha}{\partial u_1}\right)V=R(V,W_1)V.$$

Together with the equation

$$\frac{D}{\partial u_1}V=\frac{D}{\partial u_1}\frac{\partial\alpha}{\partial t}=\frac{D}{\partial t}\frac{\partial\alpha}{\partial u_1}=\frac{D}{\partial t}W_1,$$

this yields

$$\frac{D}{\partial u_1}\frac{D}{\partial t}V=\frac{D^2W_1}{dt^2}+R(V,W_1)V.$$

Substituting this last expression in $(*)$, we get the Second Variation Formula. □

Theorem 13.12 shows that the expression

$$\left.\frac{\partial^2(E\circ\widetilde{\alpha})(u_1,u_2)}{\partial u_1\partial u_2}\right|_{(0,0)}$$

only depends on the variation fields W_1 and W_2 and thus, $I_\gamma(W_1,W_2)$ is actually well-defined. If no confusion arises, we write $I(W_1,W_2)$ for $I_\gamma(W_1,W_2)$.

Proposition 13.13. *Given any geodesic, $\gamma \in \Omega(p,q)$, the map, $I\colon T_\gamma\Omega(p,q)\times T_\gamma\Omega(p,q) \to \mathbb{R}$, defined so that, for all $W_1, W_2 \in T_\gamma\Omega(p,q)$,*

$$I(W_1,W_2) = \left.\frac{\partial^2(E\circ\widetilde{\alpha})(u_1,u_2)}{\partial u_1\partial u_2}\right|_{(0,0)},$$

only depends on W_1 and W_2 and is bilinear and symmetric, where $\alpha\colon U\times[0,1]\to M$ is any 2-parameter variation, with

$$\alpha(0,0,t) = \gamma(t), \quad \frac{\partial\alpha}{\partial u_1}(0,0,t) = W_1(t), \quad \frac{\partial\alpha}{\partial u_2}(0,0,t) = W_2(t).$$

Proof. We already observed that the Second Variation Formula implies that $I(W_1,W_2)$ is well defined. This formula also shows that I is bilinear. As

$$\frac{\partial^2(E\circ\widetilde{\alpha})(u_1,u_2)}{\partial u_1\partial u_2} = \frac{\partial^2(E\circ\widetilde{\alpha})(u_1,u_2)}{\partial u_2\partial u_1},$$

I is symmetric (but this is not obvious from the right-handed side of the Second Variation Formula). □

On the diagonal, $I(W,W)$ can be described in terms of a 1-parameter variation of γ. In fact,

$$I(W,W) = \frac{d^2E(\widetilde{\alpha})}{du^2}(0),$$

where $\widetilde{\alpha}\colon(-\epsilon,\epsilon)\to\Omega(p,q)$ denotes any variation of γ with variation vector field, $\frac{d\widetilde{\alpha}}{du}(0)$ equal to W. To prove this equation it is only necessary to introduce the 2-parameter variation

$$\widetilde{\beta}(u_1,u_2) = \widetilde{\alpha}(u_1+u_2)$$

and to observe that

$$\frac{\partial\widetilde{\beta}}{\partial u_i} = \frac{d\widetilde{\alpha}}{du}, \qquad \frac{\partial^2(E\circ\widetilde{\beta})}{\partial u_1\partial u_2} = \frac{d^2(E\circ\widetilde{\alpha})}{du^2}.$$

As an application of the above remark we have the following result:

Proposition 13.14. *If $\gamma\in\Omega(p,q)$ is a minimal geodesic, then the bilinear index form, I, is positive semi-definite, which means that $I(W,W)\geq 0$, for all $W\in T_\gamma\Omega(p,q)$.*

Proof. The inequality

$$E(\widetilde{\alpha}(u)) \geq E(\gamma) = E(\widetilde{\alpha}(0))$$

implies that

$$\frac{d^2E(\widetilde{\alpha})}{du^2}(0) \geq 0,$$

which is exactly what needs to be proved. □

If we define the *index* of

$$I\colon T_\gamma\Omega(p,q)\times T_\gamma\Omega(p,q)\to\mathbb{R}$$

as the maximum dimension of a subspace of $T_\gamma\Omega(p,q)$ on which I is negative definite, then Proposition 13.14 says that the index of I is zero (for the minimal geodesic γ). It turns out that the index of I is finite for any geodesic, γ (this is a consequence of the *Morse Index Theorem*).

13.7 Jacobi Fields and Conjugate Points

Jacobi fields arise naturally when considering the expression involved under the integral sign in the Second Variation Formula and also when considering the derivative of the exponential.

If $B\colon E\times E\to\mathbb{R}$ is a symmetric bilinear form defined on some vector space, E (possibly infinite dimentional), recall that the *nullspace* of B is the subset, $\mathrm{null}(B)$, of E given by

$$\mathrm{null}(B)=\{u\in E\mid B(u,v)=0,\quad \text{for all } v\in E\}.$$

The *nullity*, ν, of B is the dimension of its nullspace. The bilinear form, B, is *nondegenerate* iff $\mathrm{null}(B)=(0)$ iff $\nu=0$. If U is a subset of E, we say that B is *positive definite* (resp. *negative definite*) *on* U iff $B(u,u)>0$ (resp. $B(u,u)<0$) for all $u\in U$, with $u\neq 0$. The *index* of B is the maximum dimension of a subspace of E on which B is negative definite. We will determine the nullspace of the symmetric bilinear form,

$$I\colon T_\gamma\Omega(p,q)\times T_\gamma\Omega(p,q)\to\mathbb{R},$$

where γ is a geodesic from p to q in some Riemannian manifold, M. Now, if W is a vector field in $T_\gamma\Omega(p,q)$ and W satisfies the equation

$$\frac{D^2W}{dt^2}+R(V,W)V=0, \tag{$*$}$$

where $V(t)=\gamma'(t)$ is the velocity field of the geodesic, γ, since W is smooth along γ, it is obvious from the Second Variation Formula that

$$I(W,W_2)=0,\quad \text{for all } W_2\in T_\gamma\Omega(p,q).$$

Therefore, any vector field in the nullspace of I must satisfy equation $(*)$. Such vector fields are called *Jacobi fields*.

Definition 13.7. Given a geodesic, $\gamma\in\Omega(p,q)$, a vector field, J, along γ is a *Jacobi field* iff it satisfies the *Jacobi differential equation*

$$\frac{D^2J}{dt^2}+R(\gamma',J)\gamma'=0.$$

The equation of Definition 13.7 is a linear second-order differential equation that can be transformed into a more familiar form by picking some orthonormal parallel vector fields, $X_1, \ldots, X_n$, along γ. To do this, pick any orthonormal basis, $(e_1, \ldots, e_n)$ in T_pM, with $e_1 = \gamma'(0)/\|\gamma'(0)\|$, and use parallel transport along γ to get $X_1, \ldots, X_n$. Then, we can write $J = \sum_{i=1}^n y_i X_i$, for some smooth functions, y_i, and the Jacobi equation becomes the system of second-order linear ODE's,

$$\frac{d^2 y_i}{dt^2} + \sum_{j=1}^{n} R(\gamma', E_j, \gamma', E_i) y_j = 0, \qquad 1 \leq i \leq n.$$

By the existence and uniqueness theorem for ODE's, for every pair of vectors, $u, v \in T_pM$, there is a unique Jacobi fields, J, so that $J(0) = u$ and $\frac{DJ}{dt}(0) = v$. Since T_pM has dimension n, it follows that the dimension of the space of Jacobi fields along γ is $2n$. If $J(0)$ and $\frac{DJ}{dt}(0)$ are orthogonal to $\gamma'(0)$, then $J(t)$ is orthogonal to $\gamma'(t)$ for all $t \in [0, 1]$. Indeed, the ODE for $\frac{d^2 y_1}{dt^2}$ yields

$$\frac{d^2 y_1}{dt^2} = 0,$$

and as $y_1(0) = 0$ and $\frac{dy_1}{dt}(0) = 0$, we get $y_1(t) = 0$ for all $t \in [0, 1]$. Furthermore, if J is orthogonal to γ, which means that $J(t)$ is orthogonal to $\gamma'(t)$, for all $t \in [0, 1]$, then $\frac{DJ}{dt}$ is also orthogonal to γ. Indeed, as γ is a geodesic,

$$0 = \frac{d}{dt}\langle J, \gamma' \rangle = \langle \frac{DJ}{dt}, \gamma' \rangle.$$

Therefore, the dimension of the space of Jacobi fields normal to γ is $2n - 2$. These facts prove part of the following

Proposition 13.15. *If $\gamma \in \Omega(p, q)$ is a geodesic in a Riemannian manifold of dimension n, then the following properties hold:*

(1) For all $u, v \in T_pM$, there is a unique Jacobi fields, J, so that $J(0) = u$ and $\frac{DJ}{dt}(0) = v$. Consequently, the vector space of Jacobi fields has dimension n.

(2) The subspace of Jacobi fields orthogonal to γ has dimension $2n - 2$. The vector fields γ' and $t \mapsto t\gamma'(t)$ are Jacobi fields that form a basis of the subspace of Jacobi fields parallel to γ (that is, such that $J(t)$ is collinear with $\gamma'(t)$, for all $t \in [0, 1]$.)

(3) If J is a Jacobi field, then J is orthogonal to γ iff there exist $a, b \in [0, 1]$, with $a \neq b$, so that $J(a)$ and $J(b)$ are both orthogonal to γ iff there is some $a \in [0, 1]$ so that $J(a)$ and $\frac{DJ}{dt}(a)$ are both orthogonal to γ.

(4) For any two Jacobi fields, X, Y, along γ, the expression $\langle \nabla_{\gamma'} X, Y \rangle - \langle \nabla_{\gamma'} Y, X \rangle$ is a constant and if X and Y vanish at some point on γ, then $\langle \nabla_{\gamma'} X, Y \rangle - \langle \nabla_{\gamma'} Y, X \rangle = 0$.

Proof. We already proved (1) and part of (2). If J is parallel to γ, then $J(t) = f(t)\gamma(t)$ and the Jacobi equation becomes

$$\frac{d^2 f}{dt} = 0.$$

Therefore,

$$J(t) = (\alpha + \beta t)\gamma'(t).$$

It is easily shown that γ' and $t \mapsto t\gamma'(t)$ are linearly independent (as vector fields).

To prove (3), using the Jacobi equation, observe that

$$\frac{d^2}{dt^2}\langle J, \gamma'\rangle = \langle \frac{D^2 J}{dt^2}, \gamma'\rangle = -R(J, \gamma', \gamma', \gamma') = 0.$$

Therefore,

$$\langle J, \gamma'\rangle = \alpha + \beta t$$

and the result follows. We leave (4) as an exercise. □

Following Milnor, we will show that the Jacobi fields in $T_\gamma\Omega(p, q)$ are exactly the vector fields in the nullspace of the index form, I. First, we define the important notion of conjugate points.

Definition 13.8. Let $\gamma \in \Omega(p, q)$ be a geodesic. Two distinct parameter values, $a, b \in [0, 1]$, with $a < b$, are *conjugate along* γ iff there is some Jacobi field, J, not identically zero, such that $J(a) = J(b) = 0$. The dimension, k, of the space, $\mathfrak{J}_{a,b}$, consisting of all such Jacobi fields is called the *multiplicity* (or *order of conjugacy)* of a and b as conjugate parameters. We also say that the points $p_1 = \gamma(a)$ and $p_2 = \gamma(b)$ are *conjugate along* γ.

Remark: As remarked by Milnor and others, as γ may have self-intersections, the above definition is ambiguous if we replace a and b by $p_1 = \gamma(a)$ and $p_2 = \gamma(b)$, even though many authors make this slight abuse. Although it makes sense to say that the points p_1 and p_2 are conjugate, the space of Jacobi fields vanishing at p_1 and p_2 is not well defined. Indeed, if $p_1 = \gamma(a)$ for distinct values of a (or $p_2 = \gamma(b)$ for distinct values of b), then we don't know which of the spaces, $\mathfrak{J}_{a,b}$, to pick. We will say that some points p_1 and p_2 on γ are *conjugate* iff there are parameter values, $a < b$, such that $p_1 = \gamma(a)$, $p_2 = \gamma(b)$, and a and b are conjugate along γ.

However, for the endpoints p and q of the geodesic segment γ, we may assume that $p = \gamma(0)$ and $q = \gamma(1)$, so that when we say that p and q are conjugate we consider the space of Jacobi fields vanishing for $t = 0$ and $t = 1$. This is the definition adopted Gallot, Hulin and Lafontaine [61] (Chapter 3, Section 3E).

In view of Proposition 13.15 (3), the Jacobi fields involved in the definition of conjugate points are orthogonal to γ. The dimension of the space of Jacobi fields such that $J(a) = 0$ is obviously n, since the only remaining parameter determining J is $\frac{dJ}{dt}(a)$. Furthermore, the

Jacobi field, $t \mapsto (t-a)\gamma'(t)$, vanishes at a but not at b, so the multiplicity of conjugate parameters (points) is at most $n-1$.

For example, if M is a flat manifold, that is, iff its curvature tensor is identically zero, then the Jacobi equation becomes

$$\frac{D^2J}{dt^2} = 0.$$

It follows that $J \equiv 0$, and thus, there are no conjugate points. More generally, the Jacobi equation can be solved explicitly for spaces of constant curvature.

Theorem 13.16. *Let $\gamma \in \Omega(p,q)$ be a geodesic. A vector field, $W \in T_\gamma\Omega(p,q)$, belongs to the nullspace of the index form, I, iff W is a Jacobi field. Hence, I is degenerate if p and q are conjugate. The nullity of I is equal to the multiplicity of p and q.*

Proof. (After Milnor [107], Theorem 14.1). We already observed that a Jacobi field vanishing at 0 and 1 belong to the nullspace of I.

Conversely, assume that $W_1 \in T_\gamma\Omega(p,q)$ belongs to the nullspace of I. Pick a subdivision, $0 = t_0 < t_1 < \cdots < t_k = 1$ of $[0,1]$ so that $W_1 \upharpoonright [t_i, t_{i+1}]$ is smooth for all $i = 0, \ldots, k-1$ and let $f\colon [0,1] \to [0,1]$ be a smooth function which vanishes for the parameter values $t_0, \ldots, t_k$ and is strictly positive otherwise. Then, if we let

$$W_2(t) = f(t)\left(\frac{D^2W_1}{dt^2} + R(\gamma', W_1)\gamma'\right)_t,$$

by the Second Variation Formula, we get

$$0 = -\frac{1}{2}I(W_1, W_2) = \sum 0 + \int_0^1 f(t)\left\|\frac{D^2W_1}{dt^2} + R(\gamma', W_1)\gamma'\right\|^2 dt.$$

Consequently, $W_1 \upharpoonright [t_i, t_{i+1}]$ is a Jacobi field for all $i = 0, \ldots, k-1$.

Now, let $W_2' \in T_\gamma\Omega(p,q)$ be a field such that

$$W_2'(t_i) = \Delta_{t_i}\frac{DW_1}{dt}, \qquad i = 1, \ldots, k-1.$$

We get

$$0 = -\frac{1}{2}I(W_1, W_2') = \sum_{i=1}^{k-1}\left\|\Delta_{t_i}\frac{DW_1}{dt}\right\|^2 + \int_0^1 0\, dt.$$

Hence, $\frac{DW_1}{dt}$ has no jumps. Now, a solution, W_1, of the Jacobi equation is completely determined by the vectors $W_1(t_i)$ and $\frac{DW_1}{dt}(t_i)$, so the k Jacobi fields, $W_1 \upharpoonright [t_i, t_{i+1}]$, fit together to give a Jacobi field, W_1, which is smooth throughout $[0,1]$. □

Theorem 13.16 implies that the nullity of I is finite, since the vector space of Jacobi fields vanishing at 0 and 1 has dimension at most n. In fact, we observed that the dimension of this space is at most $n-1$.

Corollary 13.17. *The nullity, ν, of I satisfies $0 \leq \nu \leq n-1$, where $n = \dim(M)$.*

Jacobi fields turn out to be induced by certain kinds of variations called *geodesic variations*.

Definition 13.9. Given a geodesic, $\gamma \in \Omega(p,q)$, a *geodesic variation of* γ is a smooth map,

$$\alpha\colon (-\epsilon,\epsilon)\times[0,1] \to M,$$

such that

(1) $\alpha(0,t) = \gamma(t)$, for all $t \in [0,1]$.

(2) For every $u \in (-\epsilon,\epsilon)$, the curve $\widetilde{\alpha}(u)$ is a geodesic, where

$$\widetilde{\alpha}(u)(t) = \alpha(u,t), \qquad t \in [0,1].$$

Note that the geodesics, $\widetilde{\alpha}(u)$, do not necessarily begin at p and end at q and so, a geodesic variation is not a "fixed endpoints" variation.

Proposition 13.18. *If $\alpha\colon (-\epsilon,\epsilon)\times[0,1] \to M$ is a geodesic variation of $\gamma \in \Omega(p,q)$, then the vector field, $W(t) = \frac{\partial \alpha}{\partial u}(0,t)$, is a Jacobi field along γ.*

Proof. As α is a geodesic variation, we have

$$\frac{D}{dt}\frac{\partial \alpha}{\partial t} = 0.$$

Hence, using Proposition 13.3, we have

$$\begin{aligned} 0 &= \frac{D}{\partial u}\frac{D}{\partial t}\frac{\partial \alpha}{\partial t} \\ &= \frac{D}{\partial t}\frac{D}{\partial u}\frac{\partial \alpha}{\partial t} + R\left(\frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial u}\right)\frac{\partial \alpha}{\partial t} \\ &= \frac{D^2}{\partial t^2}\frac{\partial \alpha}{\partial u} + R\left(\frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial u}\right)\frac{\partial \alpha}{\partial t}, \end{aligned}$$

where we used the fact (already used before) that

$$\frac{D}{\partial t}\frac{\partial \alpha}{\partial u} = \frac{D}{\partial u}\frac{\partial \alpha}{\partial t},$$

as the Levi-Civita connection is torsion-free. □

For example, on the sphere, S^n, for any two antipodal points, p and q, rotating the sphere keeping p and q fixed, the variation field along a geodesic, γ, through p and q (a great circle) is a Jacobi field vanishing at p and q. Rotating in $n-1$ different directions one obtains $n-1$ linearly independent Jacobi fields and thus, p and q are conjugate along γ with multiplicity $n-1$.

Interestingly, the converse of Proposition 13.18 holds.

Proposition 13.19. *For every Jacobi field, $W(t)$, along a geodesic, $\gamma \in \Omega(p,q)$, there is some geodesic variation, $\alpha\colon (-\epsilon,\epsilon)\times[0,1]\to M$ of γ, such that $W(t) = \frac{\partial \alpha}{\partial u}(0,t)$. Furthermore, for every point, $\gamma(a)$, there is an open subset, U, containing $\gamma(a)$, such that the Jacobi fields along a geodesic segment in U are uniquely determined by their values at the endpoints of the geodesic.*

Proof. (After Milnor, see [107], Chapter III, Lemma 14.4.) We begin by proving the second assertion. By Proposition 12.4 (1), there is an open subset, U, with $\gamma(0)\in U$, so that any two points of U are joined by a unique minimal geodesic which depends differentially on the endpoints. Suppose that $\gamma(t)\in U$ for $t\in[0,\delta]$. We will construct a Jacobi field, W, along $\gamma\restriction[0,\delta]$ with arbitrarily prescribed values, u, at $t=0$ and v at $t=\delta$. Choose some curve, $c_0\colon(-\epsilon,\epsilon)\to U$, so that $c_0(0)=\gamma(0)$ and $c_0'(0)=u$ and some curve, $c_\delta\colon(-\epsilon,\epsilon)\to U$, so that $c_\delta(0)=\gamma(\delta)$ and $c_\delta'(0)=v$. Now, define the map,

$$\alpha\colon(-\epsilon,\epsilon)\times[0,\delta]\to M,$$

by letting $\widetilde{\alpha}(u)\colon[0,\delta]\to M$ be the unique minimal geodesic from $c_0(u)$ to $c_\delta(u)$. It is easily checked that α is a geodesic variation of $\gamma\restriction[0,\delta]$ and that

$$J(t) = \frac{\partial\alpha}{\partial u}(0,t)$$

is a Jacobi field such that $J(0)=u$ and $J(\delta)=v$.

We claim that every Jacobi field along $\gamma\restriction[0,\delta]$ can be obtained uniquely in this way. If $\mathfrak{J}_\delta$ denotes the vector space of all Jacobi fields along $\gamma\restriction[0,\delta]$, the map $J\mapsto(J(0),J(\delta))$ defines a linear map

$$\ell\colon\mathfrak{J}_\delta\to T_{\gamma(0)}M\times T_{\gamma(\delta)}M.$$

The above argument shows that ℓ is onto. However, both vector spaces have the same dimension, $2n$, so ℓ is an isomorphism. Therefore, every Jacobi field in $\mathfrak{J}_\delta$ is determined by its values at $\gamma(0)$ and $\gamma(\delta)$.

Now, the above argument can be repeated for every point, $\gamma(a)$, on γ, so we get an open cover, $\{(l_a,r_a)\}$, of $[0,1]$, such that every Jacobi field along $\gamma\restriction[l_a,r_a]$ is uniquely determined by its endpoints. By compactness of $[0,1]$, the above cover possesses some finite subcover and we get a geodesic variation, α, defined on the entire interval $[0,1]$ whose variation field is equal to the original Jacobi field, W. □

Remark: The proof of Proposition 13.19 also shows that there is some open interval $(-\delta, \delta)$, such that if $t \in (-\delta, \delta)$, then $\gamma(t)$ is not conjugate to $\gamma(0)$ along γ. In fact, the Morse Index Theorem implies that for any geodesic segment, $\gamma\colon [0,1] \to M$, there are only finitely many points which are conjugate to $\gamma(0)$ along γ (see Milnor [107], Part III, Corollary 15.2).

There is also an intimate connection between Jacobi fields and the differential of the exponential map and between conjugate points and critical points of the exponential map.

Recall that if $f\colon M \to N$ is a smooth map between manifolds, a point, $p \in M$, is a *critical* point of f iff the tangent map at p,

$$df_p\colon T_pM \to T_{f(p)}N,$$

is not surjective. If M and N have the same dimension, which will be the case in the sequel, df_p is not surjective iff it is not injective, so p is a critical point of f iff there is some nonzero vector, $u \in T_pM$, such that $df_p(u) = 0$.

If $\exp_p\colon T_pM \to M$ is the exponential map, for any $v \in T_pM$ where $\exp_p(v)$ is defined, we have the derivative of $\exp_p$ at v;

$$(d\exp_p)_v\colon T_v(T_pM) \to T_pM.$$

Since T_pM is a finite-dimensional vector space, $T_v(T_pM)$ is isomorphic to T_pM, so we identify $T_v(T_pM)$ with T_pM.

Proposition 13.20. *Let $\gamma \in \Omega(p,q)$ be a geodesic. The point, $r = \gamma(t)$, with $t \in (0,1]$, is conjugate to p along γ iff $v = t\gamma'(0)$ is a critical point of $\exp_p$. Furthermore, the multiplicity of p and r as conjugate points is equal to the dimension of the kernel of $(d\exp_p)_v$.*

A proof of Proposition 13.20 can be found in various places, including Do Carmo [51] (Chapter 5, Proposition 3.5), O'Neill [120] (Chapter 10, Proposition 10), or Milnor [107] (Part III, Theorem 18.1).

Using Proposition 13.19 it is easy to characterize conjugate points in terms of geodesic variations.

Proposition 13.21. *If $\gamma \in \Omega(p,q)$ is a geodesic, then q is conjugate to p iff there is a geodesic variation, α, of γ, such that every geodesic, $\widetilde{\alpha}(u)$, starts from p, the Jacobi field, $J(t) = \frac{\partial \alpha}{\partial u}(0,t)$ does not vanish identically, and $J(1) = 0$.*

Jacobi fields can also be used to compute the derivative of the exponential (see Gallot, Hulin and Lafontaine [61], Chapter 3, Corollary 3.46).

Proposition 13.22. *Given any point, $p \in M$, for any vectors $u, v \in T_pM$, if $\exp_p v$ is defined, then*

$$J(t) = (d\exp_p)_{tv}(tu), \qquad 0 \leq t \leq 1,$$

is a Jacobi field such that $\frac{DJ}{dt}(0) = u$.

Remark: If $u, v \in T_pM$ are orthogonal unit vectors, then $R(u, v, u, v) = K(u, v)$, the sectional curvature of the plane spanned by u and v in T_pM, and for t small enough, we have

$$\|J(t)\| = t - \frac{1}{6}\, K(u, v)t^3 + o(t^3).$$

(Here, $o(t^3)$ stands for an expression of the form $t^4 R(t)$, such that $\lim_{t \mapsto 0} R(t) = 0$.) Intuitively, this formula tells us how fast the geodesics that start from p and are tangent to the plane spanned by u and v spread apart. Locally, for $K(u, v) > 0$, the radial geodesics spread apart less than the rays in T_pM and for $K(u, v) < 0$, they spread apart more than the rays in T_pM. More details, see Do Carmo [51] (Chapter 5, Section 2).

There is also another version of "Gauss lemma" (see Gallot, Hulin and Lafontaine [61], Chapter 3, Lemma 3.70):

Proposition 13.23. *(Gauss Lemma) Given any point, $p \in M$, for any vectors $u, v \in T_pM$, if $\exp_p v$ is defined, then*

$$\langle d(\exp_p)_{tv}(u), d(\exp_p)_{tv}(v)\rangle = \langle u, v\rangle, \qquad 0 \leq t \leq 1.$$

As our (connected) Riemannian manifold, M, is a metric space, the path space, $\Omega(p, q)$, is also a metric space if we use the metric, d^*, given by

$$d^*(\omega_1, \omega_2) = \max_t(d(\omega_1(t), \omega_2(t))),$$

where d is the metric on M induced by the Riemannian metric.

Remark: The topology induced by d^* turns out to be the compact open topology on $\Omega(p, q)$.

Theorem 13.24. *Let $\gamma \in \Omega(p, q)$ be a geodesic. Then, the following properties hold:*

(1) If there are no conjugate points to p along γ, then there is some open subset, $\mathcal{V}$, of $\Omega(p, q)$, with $\gamma \in \mathcal{V}$, such that

$$L(\omega) \geq L(\gamma) \quad \text{and} \quad E(\omega) \geq E(\gamma), \qquad \text{for all } \omega \in \mathcal{V},$$

with strict inequality when $\omega([0, 1]) \neq \gamma([0, 1])$. We say that γ is a local minimum.

(2) If there is some $t \in (0, 1)$ such that p and $\gamma(t)$ are conjugate along γ, then there is a fixed endpoints variation, α, such that

$$L(\widetilde{\alpha}(u)) < L(\gamma) \quad \text{and} \quad E(\widetilde{\alpha}(u)) < E(\gamma), \qquad \text{for } u \text{ small enough.}$$

A proof of Theorem 13.24 can be found in Gallot, Hulin and Lafontaine [61] (Chapter 3, Theorem 3.73) or in O'Neill [120] (Chapter 10, Theorem 17 and Remark 18).

13.8 Convexity, Convexity Radius

Proposition 12.4 shows that if (M, g) is a Riemannian manifold, then for every point, $p \in M$, there is an open subset, $W \subseteq M$, with $p \in W$ and a number $\epsilon > 0$, so that any two points q_1, q_2 of W are joined by a unique geodesic of length $< \epsilon$. However, there is no garantee that this unique geodesic between q_1 and q_2 stays inside W. Intuitively this says that W may not be convex.

The notion of convexity can be generalized to Riemannian manifolds but there are some subtleties. In this short section, we review various definition or convexity found in the literature and state one basic result. Following Sakai [131] (Chapter IV, Section 5), we make the following definition:

Definition 13.10. Let $C \subseteq M$ be a nonempty subset of some Riemannian manifold, M.

(1) The set C is called *strongly convex* iff for any two points, $p, q \in C$, there exists a unique minimal geodesic, γ, from p to q in M and γ is contained in C.

(2) If for every point, $p \in \overline{C}$, there is some $\epsilon(p) > 0$, so that $C \cap B_{\epsilon(p)}(p)$ is strongly convex, then we say that C is *locally convex* (where $B_{\epsilon(p)}(p)$ is the metric ball of center 0 and radius $\epsilon(p)$).

(3) The set C is called *totally convex* iff for any two points, $p, q \in C$, all geodesics from p to q in M are contained in C.

It is clear that if C is strongly convex or totally convex, then C is locally convex. If M is complete and any two points are joined by a unique geodesic, then the three conditions of Definition 13.10 are equivalent. The next Proposition will show that a metric ball with sufficiently small radius is strongly convex.

Definition 13.11. For any $p \in M$, the *convexity radius at* p, denoted, $r(p)$, is the least upper bound of the numbers, $r > 0$, such that for any metric ball, $B_\epsilon(q)$, if $B_\epsilon(q) \subseteq B_r(p)$, then $B_\epsilon(q)$ is strongly convex and every geodesic contained in $B_r(p)$ is a minimal geodesic joining its endpoints. The *convexity radius of* M, $r(M)$, as the greatest lower bound of the set $\{r(p) \mid p \in M\}$.

Note that it is possible that $r(p) = 0$ if M is not compact.

The following proposition is proved in Sakai [131] (Chapter IV, Section 5, Theorem 5.3).

Proposition 13.25. *If M is a Riemannian manifold, then $r(p) > 0$ for every $p \in M$ and the map, $p \mapsto r(p) \in \mathbb{R}_+ \cup \{\infty\}$ is continuous. Furthermore, if $r(p) = \infty$ for some $p \in M$, then $r(q) = \infty$ for all $q \in M$.*

That $r(p) > 0$ is also proved in Do Carmo [51] (Chapter 3, Section 4, Proposition 4.2). More can be said about the structure of connected locally convex subsets of M, see Sakai [131] (Chapter IV, Section 5).

Remark: The following facts are stated in Berger [16] (Chapter 6):

(1) If M is compact, then the convexity radius, $r(M)$, is strictly positive.

(2) $r(M) \leq \frac{1}{2} i(M)$, where $i(M)$ is the injectivity radius of M.

Berger also points out that if M is compact, then the existence of a finite cover by convex balls can used to triangulate M. This method was proposed by Hermann Karcher (see Berger [16], Chapter 3, Note 3.4.5.3).

13.9 Applications of Jacobi Fields and Conjugate Points

Jacobi fields and conjugate points are basic tools that can be used to prove many global results of Riemannian geometry. The flavor of these results is that certain constraints on curvature (sectional, Ricci, sectional) have a significant impact on the topology. One may want consider the effect of non-positive curvature, constant curvature, curvature bounded from below by a positive constant, *etc*. This is a vast subject and we highly recommend Berger's Panorama of Riemannian Geometry [16] for a masterly survey. We will content ourselves with three results:

(1) Hadamard and Cartan's Theorem about complete manifolds of non-positive sectional curvature.

(2) Myers' Theorem about complete manifolds of Ricci curvature bounded from below by a positive number.

(3) The Morse Index Theorem.

First, on the way to Hadamard and Cartan we begin with a proposition.

Proposition 13.26. *Let M be a complete Riemannian manifold with non-positive curvature, $K \leq 0$. Then, for every geodesic, $\gamma \in \Omega(p, q)$, there are no conjugate points to p along γ. Consequently, the exponential map, $\exp_p \colon T_pM \to M$, is a local diffeomorphism for all $p \in M$.*

Proof. Let J be a Jacobi field along γ. Then,

$$\frac{D^2 J}{dt^2} + R(\gamma', J)\gamma' = 0$$

so that, by the definition of the sectional curvature,

$$\left\langle \frac{D^2J}{dt^2}, J \right\rangle = -\langle R(\gamma', J)\gamma', J) = -R(\gamma', J, \gamma', J) \geq 0.$$

It follows that

$$\frac{d}{dt}\left\langle \frac{DJ}{dt}, J \right\rangle = \left\langle \frac{D^2J}{dt^2}, J \right\rangle + \left\| \frac{DJ}{dt} \right\|^2 \geq 0.$$

Thus, the function, $t \mapsto \left\langle \frac{DJ}{dt}, J \right\rangle$ is monotonic increasing and, strictly so if $\frac{DJ}{dt} \neq 0$. If J vanishes at both 0 and t, for any given $t \in (0, 1]$, then so does $\left\langle \frac{DJ}{dt}, J \right\rangle$, and hence $\left\langle \frac{DJ}{dt}, J \right\rangle$ must vanish throughout the interval $[0, t]$. This implies

$$J(0) = \frac{DJ}{dt}(0) = 0,$$

so that J is identically zero. Therefore, t is not conjugate to 0 along γ. □

Theorem 13.27. *(Hadamard–Cartan) Let M be a complete Riemannian manifold. If M has non-positive sectional curvature, $K \leq 0$, then the following hold:*

(1) For every $p \in M$, the map, $\exp_p \colon T_pM \to M$, is a Riemannian covering.

(2) If M is simply connected then M is diffeomorphic to $\mathbb{R}^n$, where $n = \dim(M)$; more precisely, $\exp_p \colon T_pM \to M$ is a diffeomorphism for all $p \in M$. Furthermore, any two points on M are joined by a unique minimal geodesic.

Proof. We follow the proof in Sakai [131] (Chapter V, Theorem 4.1).

(1) By Proposition 13.26, the exponential map, $\exp_p \colon T_pM \to M$, is a local diffeomorphism for all $p \in M$. Let $\widetilde{g}$ be the pullback metric, $\widetilde{g} = (\exp_p)^* g$, on T_pM (where g denotes the metric on M). We claim that $(T_pM, \widetilde{g})$ is complete.

This is because, for every nonzero $u \in T_pM$, the line, $t \mapsto tu$, is mapped to the geodesic, $t \mapsto \exp_p(tu)$, in M, which is defined for all $t \in \mathbb{R}$ since M is complete, and thus, this line is a geodedic in $(T_pM, \widetilde{g})$. Since this holds for all $u \in T_pM$, $(T_pM, \widetilde{g})$ is geodesically complete at 0, so by Hopf-Rinow, it is complete. But now, $\exp_p \colon T_pM \to M$ is a local isometry and by Proposition 13.11, it is a Riemannian covering map.

(2) If M is simply connected, then by Proposition 3.44, the covering map $\exp_p \colon T_pM \to M$ is a diffeomorphism (T_pM is connected). Therefore, $\exp_p \colon T_pM \to M$ is a diffeomorphism for all $p \in M$. □

Other proofs of Theorem 13.27 can be found in Do Carmo [51] (Chapter 7, Theorem 3.1), Gallot, Hulin and Lafontaine [61] (Chapter 3, Theorem 3.87), Kobayashi and Nomizu [91] (Chapter VIII, Theorem 8.1) and Milnor [107] (Part III, Theorem 19.2).

Remark: A version of Theorem 13.27 was first proved by Hadamard and then extended by Cartan.

Theorem 13.27 was generalized by Kobayashi, see Kobayashi and Nomizu [91] (Chapter VIII, Remark 2 after Corollary 8.2). Also, it is shown in Milnor [107] that if M is complete, assuming non-positive sectional curvature, then all homotopy groups, $\pi_i(M)$, vanish, for $i > 1$, and that $\pi_1(M)$ has no element of finite order except the identity. Finally, non-positive sectional curvature implies that the exponential map does not decrease distance (Kobayashi and Nomizu [91], Chapter VIII, Section 8, Lemma 3).

We now turn to manifolds with strictly positive curvature bounded away from zero and to Myers' Theorem. The first version of such a theorem was first proved by Bonnet for surfaces with positive sectional curvature bounded away from zero. It was then generalized by Myers in 1941. For these reasons, this theorem is sometimes called the *Bonnet-Myers' Theorem*. The proof of Myers Theorem involves a beautiful "trick".

Given any metric space, X, recall that the *diameter* of X is defined by

$$\mathrm{diam}(X) = \sup\{d(p, q) \mid p, q \in X\}.$$

The diameter of X may be infinite.

Theorem 13.28. *(Myers) Let M be a complete Riemannian manifold of dimension n and assume that*

$$\mathrm{Ric}(u, u) \geq (n-1)/r^2, \qquad \textit{for all unit vectors, } u \in T_pM, \textit{ and for all } p \in M,$$

with $r > 0$. Then,

(1) The diameter of M is bounded by πr and M is compact.

(2) The fundamental group of M is finite.

Proof. (1) Pick any two points $p, q \in M$ and let $d(p, q) = L$. As M is complete, by Hopf and Rinow's Theorem, there is a minimal geodesic, γ, joining p and q and by Proposition 13.14, the bilinear index form, I, associated with γ is positive semi-definite, which means that $I(W, W) \geq 0$, for all vector fields, $W \in T_\gamma\Omega(p, q)$. Pick an orthonormal basis, $(e_1, \ldots, e_n)$, of T_pM, with $e_1 = \gamma'(0)/L$. Using parallel transport, we get a field of orthonormal frames, $(X_1, \ldots, X_n)$, along γ, with $X_1(t) = \gamma'(t)/L$. Now comes Myers' beautiful trick. Define new vector fields, Y_i, along γ, by

$$W_i(t) = \sin(\pi t)X_i(t), \qquad 2 \leq i \leq n.$$

We have

$$\gamma'(t) = LX_1 \quad \text{and} \quad \frac{DX_i}{dt} = 0.$$

Then, by the second variation formula,

$$\begin{aligned}\frac{1}{2}\, I(W_i, W_i) &= -\int_0^1 \left\langle W_i, \frac{D^2 W_i}{dt^2} + R(\gamma', W_i)\gamma' \right\rangle dt \\ &= \int_0^1 (\sin(\pi t))^2(\pi^2 - L^2 \langle R(X_1, X_i)X_1, X_i\rangle)dt,\end{aligned}$$

for $i = 2, \ldots, n$. Adding up these equations and using the fact that

$$\mathrm{Ric}(X_1(t), X_1(t)) = \sum_{i=2}^{n} \langle R(X_1(t), X_i(t))X_1(t), X_i(t)\rangle,$$

we get

$$\frac{1}{2}\sum_{i=2}^{n} I(W_i, W_i) = \int_0^1 (\sin(\pi t))^2[(n-1)\pi^2 - L^2\,\mathrm{Ric}(X_1(t), X_1(t))]dt.$$

Now, by hypothesis,

$$\mathrm{Ric}(X_1(t), X_1(t)) \geq (n-1)/r^2,$$

so

$$0 \leq \frac{1}{2}\sum_{i=2}^{n} I(W_i, W_i) \leq \int_0^1 (\sin(\pi t))^2 \left[(n-1)\pi^2 - (n-1)\frac{L^2}{r^2}\right] dt,$$

which implies $\frac{L^2}{r^2} \leq \pi^2$, that is

$$d(p, q) = L \leq \pi r.$$

As the above holds for every pair of points, $p, q \in M$, we conclude that

$$\mathrm{diam}(M) \leq \pi r.$$

Since closed and bounded subsets in a complete manifold are compact, M itself must be compact.

(2) Since the universal covering space, $\widetilde{M}$, of M has the pullback of the metric on M, this metric satisfies the same assumption on its Ricci curvature as that of M. Therefore, $\widetilde{M}$ is also compact, which implies that the fundamental group, $\pi_1(M)$, is finite (see the discussion at the end of Section 3.9). □

Remarks:

(1) The condition on the Ricci curvature cannot be weakened to $\mathrm{Ric}(u, u) > 0$ for all unit vectors. Indeed, the paraboloid of revolution, $z = x^2 + y^2$, satisfies the above condition, yet it is not compact.

(2) Theorem 13.28 also holds under the stronger condition that the sectional curvature $K(u, v)$ satisfies

$$K(u, v) \geq (n-1)/r^2,$$

for all orthonormal vectors, u, v. In this form, it is due to Bonnet (for surfaces).

It would be a pity not to include in this section a beautiful theorem due to Morse.

Theorem 13.29. *(Morse Index Theorem) Given a geodesic, $\gamma \in \Omega(p, q)$, the index, λ, of the index form, $I \colon T_\gamma\Omega(p, q) \times T_\gamma\Omega(p, q) \to \mathbb{R}$, is equal to the number of points, $\gamma(t)$, with $0 \leq t \leq 1$, such that $\gamma(t)$ is conjugate to $p = \gamma(0)$ along γ, each such conjugate point counted with its multiplicity. The index λ is always finite.*

As a corollary of Theorem 13.29, we see that there are only finitely many points which are conjugate to $p = \gamma(0)$ along γ.

A proof of Theorem 13.29 can be found in Milnor [107] (Part III, Section 15) and also in Do Carmo [51] (Chapter 11) or Kobayashi and Nomizu [91] (Chapter VIII, Section 6).

A key ingredient of the proof is that the vector space, $T_\gamma\Omega(p, q)$, can be split into a direct sum of subspaces mutually orthogonal with respect to I, on one of which (denoted T') I is positive definite. Furthermore, the subspace orthogonal to T' is finite-dimensional. This space is obtained as follows: Since for every point, $\gamma(t)$, on γ, there is some open subset, U_t, containing $\gamma(t)$ such that any two points in U_t are joined by a unique minimal geodesic, by compactness of $[0, 1]$, there is a subdivision, $0 = t_0 < t_1 < \cdots < t_k = 1$ of $[0, 1]$ so that $\gamma \restriction [t_i, t_{i+1}]$ lies within an open where it is a minimal geodesic.

Let $T_\gamma\Omega(t_0, \ldots, t_k) \subseteq T_\gamma\Omega(p, q)$ be the vector space consisting of all vector fields, W, along γ such that

(1) $W \restriction [t_i, t_{i+1}]$ is a Jacobi field along $\gamma \restriction [t_i, t_{i+1}]$, for $i = 0, \ldots, k-1$.

(2) $W(0) = W(1) = 0$.

The space $T_\gamma\Omega(t_0, \ldots, t_k) \subseteq T_\gamma\Omega(p, q)$ is a finite-dimensional vector space consisting of broken Jacobi fields. Let $T' \subseteq T_\gamma\Omega(p, q)$ be the vector space consisting of all vector fields, $W \in T_\gamma\Omega(p, q)$, for which

$$W(t_i) = 0, \qquad 0 \leq i \leq k.$$

It is not hard to prove that

$$T_\gamma\Omega(p, q) = T_\gamma\Omega(t_0, \ldots, t_k) \oplus T',$$

that $T_\gamma\Omega(t_0, \ldots, t_k)$ and T' are orthogonal *w.r.t* I and that $I \restriction T'$ is positive definite. The reason why $I(W, W) \geq 0$ for $W \in T'$ is that each segment, $\gamma \restriction [t_i, t_{i+1}]$, is a minimal geodesic, which has smaller energy than any other path between its endpoints.

As a consequence, the index (or nullity) of I is equal to the index (or nullity) of I restricted to the finite dimensional vector space, $T_\gamma\Omega(t_0, \ldots, t_k)$. This shows that the index is always finite.

In the next section, we will use conjugate points to give a more precise characterization of the cut locus.

13.10 Cut Locus and Injectivity Radius: Some Properties

We begin by reviewing the definition of the cut locus from a slightly different point of view. Let M be a complete Riemannian manifold of dimension n. There is a bundle, UM, called the *unit tangent bundle*, such that the fibre at any $p \in M$ is the unit sphere, $S^{n-1} \subseteq T_pM$ (check the details). As usual, we let $\pi\colon UM \to M$ denote the projection map which sends every point in the fibre over p to p. Then, we have the function,

$$\rho\colon UM \to \mathbb{R},$$

defined so that for all $p \in M$, for all $v \in S^{n-1} \subseteq T_pM$,

$$\begin{aligned}\rho(v) &= \sup_{t\in\mathbb{R}\cup\{\infty\}} d(\pi(v), \exp_p(tv)) = t\\ &= \sup\{t \in \mathbb{R} \cup \{\infty\} \mid \text{the geodesic} \quad t \mapsto \exp_p(tv) \quad \text{is minimal on } [0,t]\}.\end{aligned}$$

The number $\rho(v)$ is called the *cut value* of v. It can be shown that ρ is continuous and for every $p \in M$, we let

$$\widetilde{\mathrm{Cut}}(p) = \{\rho(v)v \in T_pM \mid v \in UM \cap T_pM,\ \rho(v) \text{ is finite}\}$$

be the *tangential cut locus of* p and

$$\mathrm{Cut}(p) = \exp_p(\widetilde{\mathrm{Cut}}(p))$$

be the *cut locus of* p. The point, $\exp_p(\rho(v)v)$, in M is called the *cut point* of the geodesic, $t \mapsto \exp_p(vt)$, and so, the cut locus of p is the set of cut points of all the geodesics emanating from p. Also recall from Definition 12.7 that

$$\mathcal{U}_p = \{v \in T_pM \mid \rho(v) > 1\}$$

and that $\mathcal{U}_p$ is open and star-shaped. It can be shown that

$$\widetilde{\mathrm{Cut}}(p) = \partial\mathcal{U}_p$$

and the following property holds:

Theorem 13.30. *If M is a complete Riemannian manifold, then for every $p \in M$, the exponential map, $\exp_p$, is a diffeomorphism between $\mathcal{U}_p$ and its image, $\exp_p(\mathcal{U}_p) = M - \mathrm{Cut}(p)$, in M.*

Proof. The fact that $\exp_p$ is injective on $\mathcal{U}_p$ was shown in Proposition 12.16. Now, for any $v \in \mathcal{U}$, as $t \mapsto \exp_p(tv)$ is a minimal geodesic for $t \in [0,1]$, by Theorem 13.24 (2), the point $\exp_p v$ is not conjugate to p, so $d(\exp_p)_v$ is bijective, which implies that $\exp_p$ is a local diffeomorphism. As $\exp_p$ is also injective, it is a diffeomorphism. □

Theorem 13.30 implies that the cut locus is closed.

Remark: In fact, $M - \mathrm{Cut}(p)$ can be retracted homeomorphically onto a ball around p and $\mathrm{Cut}(p)$ is a deformation retract of $M - \{p\}$.

The following Proposition gives a rather nice characterization of the cut locus in terms of minimizing geodesics and conjugate points:

Proposition 13.31. *Let M be a complete Riemannian manifold. For every pair of points, $p, q \in M$, the point q belongs to the cut locus of p iff one of the two (not mutually exclusive from each other) properties hold:*

(a) There exist two distinct minimizing geodesics from p to q.

(b) There is a minimizing geodesic, γ, from p to q and q is the first conjugate point to p along γ.

A proof of Proposition 13.31 can be found in Do Carmo [51] (Chapter 13, Proposition 2.2) Kobayashi and Nomizu [91] (Chapter VIII, Theorem 7.1) or Klingenberg [89] (Chapter 2, Lemma 2.1.11).

Observe that Proposition 13.31 implies the following symmetry property of the cut locus: $q \in \mathrm{Cut}(p)$ iff $p \in \mathrm{Cut}(q)$. Furthermore, if M is compact, we have

$$p = \bigcap_{q \in \mathrm{Cut}(p)} \mathrm{Cut}(q).$$

Proposition 13.31 admits the following sharpening:

Proposition 13.32. *Let M be a complete Riemannian manifold. For all $p, q \in M$, if $q \in \mathrm{Cut}(p)$, then:*

(a) If among the minimizing geodesics from p to q, there is one, say γ, such that q is not conjugate to p along γ, then there is another minimizing geodesic $\omega \neq \gamma$ from p to q.

(b) Suppose $q \in \mathrm{Cut}(p)$ realizes the distance from p to $\mathrm{Cut}(p)$ (i.e., $d(p,q) = d(p,\mathrm{Cut}(p))$). If there are no minimal geodesics from p to q such that q is conjugate to p along this geodesic, then there are exactly two minimizing geodesics, γ_1 and γ_2, from p to q, with $\gamma_2'(1) = -\gamma_1'(1)$. Moreover, if $d(p,q) = i(M)$ (the injectivity radius), then γ_1 and γ_2 together form a closed geodesic.

Except for the last statement, Proposition 13.32 is proved in Do Carmo [51] (Chapter 13, Proposition 2.12). The last statement is from Klingenberg [89] (Chapter 2, Lemma 2.1.11).

We also have the following characterization of $\widetilde{\mathrm{Cut}}(p)$:

Proposition 13.33. *Let M be a complete Riemannian manifold. For any $p \in M$, the set of vectors, $u \in \widetilde{\mathrm{Cut}}(p)$, such that is some $v \in \widetilde{\mathrm{Cut}}(p)$ with $v \neq u$ and $\exp_p(u) = \exp_p(v)$, is dense in $\widetilde{\mathrm{Cut}}(p)$.*

Proposition 13.33 is proved in Klingenberg [89] (Chapter 2, Theorem 2.1.14).

We conclude this section by stating a classical theorem of Klingenberg about the injectivity radius of a manifold of bounded positive sectional curvature.

Theorem 13.34. *(Klingenberg) Let M be a complete Riemannian manifold and assume that there are some positive constants, $K_{\min}$, $K_{\max}$, such that the sectional curvature of K satisfies*

$$0 < K_{\min} \leq K \leq K_{\max}.$$

Then, M is compact and either

(a) $i(M) \geq \pi/\sqrt{K_{\max}}$, or

(b) There is a closed geodesic, γ, of minimal length among all closed geodesics in M and such that

$$i(M) = \frac{1}{2}\, L(\gamma).$$

The proof of Theorem 13.34 is quite hard. A proof using Rauch's comparison Theorem can be found in Do Carmo [51] (Chapter 13, Proposition 2.13).

Chapter 14

Discrete Curvatures and Geodesics on Polyhedral Surfaces

Chapter 15

The Laplace-Beltrami Operator, Harmonic Forms, The Connection Laplacian and Weitzenböck Formulae

15.1 The Gradient, Hessian and Hodge $*$ Operators on Riemannian Manifolds

The Laplacian is a very important operator because it shows up in many of the equations used in physics to describe natural phenomena such as heat diffusion or wave propagation. Therefore, it is highly desirable to generalize the Laplacian to functions defined on a manifold. Furthermore, in the late 1930's George de Rham (inspired by Élie Cartan) realized that it was fruitful to define a version of the Laplacian operating on differential forms, because of a fundamental and almost miraculous relationship between harmonics forms (those in the kernel of the Laplacian) and the de Rham cohomology groups on a (compact, orientable) smooth manifold. Indeed, as we will see in Section 15.3, for every cohomology group, $H^k_{\mathrm{DR}}(M)$, every cohomology class, $[\omega] \in H^k_{\mathrm{DR}}(M)$, is represented by a *unique harmonic k-form*, ω. This connection between analysis and topology lies deep and has many important consequences. For example, *Poincaré duality* follows as an "easy" consequence of the Hodge Theorem.

Technically, the Laplacian can be defined on differential forms using the Hodge $*$ operator (Section 22.16). On functions, there are alternate definitions of the Laplacian using only the covariant derivative and obtained by generalizing the notions of gradient and divergence to functions on manifolds.

Another version of the Laplacian can be defined in terms of the adjoint of the connection, ∇, on differential forms, viewed as a linear map from $\mathcal{A}^*(M)$ to $\mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^*(M))$. We obtain the *connection Laplacian* (also called *Bochner Laplacian*), $\nabla^*\nabla$. Then, it is natural to wonder how the Hodge Laplacian, Δ, differs from the connection Laplacian, $\nabla^*\nabla$? Remarkably, there is a formula known as *Weitzenböck's formula* (or *Bochner's formula*) of

the form

$$\Delta = \nabla^*\nabla + C(R_\nabla),$$

where $C(R_\nabla)$ is a contraction of a version of the curvature tensor on differential forms (a fairly complicated term). In the case of one-forms,

$$\Delta = \nabla^*\nabla + \mathrm{Ric},$$

where Ric is a suitable version of the Ricci curvature operating on one-forms.

Weitzenböck-type formulae are at the root of the so-called "Bochner Technique", which consists in exploiting curvature information to deduce topological information. For example, if the Ricci curvature on a compact, orientable Riemannian manifold is strictly positive, then $H^1_{\mathrm{DR}}(M) = (0)$, a theorem due to Bochner.

If $(M, \langle -, - \rangle)$ is a Riemannian manifold of dimension n, then for every $p \in M$, the inner product, $\langle -, - \rangle_p$, on T_pM yields a canonical isomorphism, $\flat \colon T_pM \to T_p^*M$, as explained in Sections 22.1 and 11.5. Namely, for any $u \in T_pM$, $u^\flat = \flat(u)$ is the linear form in T_p^*M defined by

$$u^\flat(v) = \langle u, v \rangle_p, \qquad v \in T_pM.$$

Recall that the inverse of the map $\flat$ is the map $\sharp \colon T_p^*M \to T_pM$. As a consequence, for every smooth function, $f \in C^\infty(M)$, we get smooth vector field, $\mathrm{grad}\, f = (df)^\sharp$, defined so that

$$(\mathrm{grad}\, f)_p = (df_p)^\sharp,$$

that is, we have

$$\langle (\mathrm{grad}\, f)_p, u \rangle_p = df_p(u), \quad \text{for all} \quad u \in T_pM.$$

The vector field, $\mathrm{grad}\, f$, is the *gradient* of the function f.

Conversely, a vector field, $X \in \mathfrak{X}(M)$, yields the one-form, $X^\flat \in \mathcal{A}^1(M)$, given by

$$(X^\flat)_p = (X_p)^\flat.$$

The *Hessian*, $\mathrm{Hess}(f)$, (or $\nabla^2(f)$) of a function, $f \in C^\infty(M)$, is the $(0,2)$-tensor defined by

$$\mathrm{Hess}(f)(X, Y) = X(Y(f)) - (\nabla_X Y)(f) = X(df(Y)) - df(\nabla_X Y),$$

for all vector fields, $X, Y \in \mathfrak{X}(M)$.

Recall from Proposition 11.5 that the covariant derivative, $\nabla_X \theta$, of any one-form, $\theta \in \mathcal{A}^1(M)$, is the one-form given by

$$(\nabla_X \theta)(Y) = X(\theta(Y)) - \theta(\nabla_X Y)$$

Recall from Proposition 11.5 that the covariant derivative, $\nabla_X \theta$, of any one-form, $\theta \in \mathcal{A}^1(M)$, is the one-form given by

$$(\nabla_X \theta)(Y) = X(\theta(Y)) - \theta(\nabla_X Y)$$

so, the Hessian, $\mathrm{Hess}(f)$, is also defined by

$$\mathrm{Hess}(f)(X,Y) = (\nabla_X df)(Y).$$

Since ∇ is torsion-free, we get

$$\mathrm{Hess}(f)(X,Y) = X(Y(f)) - (\nabla_X Y)(f) = Y(X(f)) - (\nabla_Y X)(f) = \mathrm{Hess}(f)(Y,X),$$

which means that the Hessian is a *symmetric* $(0,2)$-tensor. We also have the equation

$$\mathrm{Hess}(f)(X,Y) = \langle \nabla_X \operatorname{grad} f, Y\rangle.$$

Indeed,

$$\begin{aligned} X(Y(f)) &= X(df(Y)) \\ &= X(\langle \operatorname{grad} f, Y\rangle) \\ &= \langle \nabla_X \operatorname{grad} f, Y\rangle + \langle \operatorname{grad} f, \nabla_X Y\rangle \\ &= \langle \nabla_X \operatorname{grad} f, Y\rangle + (\nabla_X Y)(f) \end{aligned}$$

which yields

$$\langle \nabla_X \operatorname{grad} f, Y\rangle = X(Y(f)) - (\nabla_X Y)(f) = \mathrm{Hess}(f)(X,Y).$$

A function, $f \in C^\infty(M)$, is *convex* (resp. *strictly convex*) iff its Hessian, $\mathrm{Hess}(f)$, is positive semi-definite (resp. positive definite).

By the results of Section 22.16, the inner product, $\langle -, -\rangle_p$, on T_pM induces an inner product on $\bigwedge^k T_p^*M$. Therefore, for any two k-forms, $\omega, \eta \in \mathcal{A}^k(M)$, we get the smooth function, $\langle \omega, \eta\rangle$, given by

$$\langle \omega, \eta\rangle(p) = \langle \omega_p, \eta_p\rangle_p.$$

Furthermore, if M is oriented, then we can apply the results of Section 22.16 so the vector bundle, T^*M, is oriented (by giving T_p^*M the orientation induced by the orientation of T_pM, for every $p \in M$) and for every $p \in M$, we get a Hodge $*$-operator,

$$*\colon \bigwedge^k T_p^*M \to \bigwedge^{n-k} T_p^*M.$$

Then, given any k-form, $\omega \in \mathcal{A}^k(M)$, we can define $*\omega$ by

$$(*\omega)_p = *(\omega_p), \qquad p \in M.$$

We have to check that $*\omega$ is indeed a smooth form in $\mathcal{A}^{n-k}(M)$, but this is not hard to do in local coordinates (for help, see Morita [115], Chapter 4, Section 1). Therefore, if M is a Riemannian oriented manifold of dimension n, we have Hodge $*$-operators,

$$*\colon \mathcal{A}^k(M) \to \mathcal{A}^{n-k}(M).$$

Observe that $*1$ is just the volume form, Vol_M, induced by the metric. Indeed, we know from Section 22.1 that in local coordinates, $x_1, \ldots, x_n$, near p, the metric on T_p^*M is given by the inverse, (g^{ij}), of the metric, (g_{ij}), on T_pM and by the results of Section 22.16,

$$\begin{aligned} *(1) &= \frac{1}{\sqrt{\det(g^{ij})}}\, dx_1 \wedge \cdots \wedge dx_n \\ &= \sqrt{\det(g_{ij})}\, dx_1 \wedge \cdots \wedge dx_n = \mathrm{Vol}_M. \end{aligned}$$

Proposition 22.25 yields the following:

Proposition 15.1. *If M is a Riemannian oriented manifold of dimension n, then we have the following properties:*

(i) $*(f\omega + g\eta) = f\, *\omega + g\, *\eta$, *for all* $\omega, \eta \in \mathcal{A}^k(M)$ *and all* $f, g \in C^\infty(M)$.

(ii) $** = (-\mathrm{id})^{k(n-k)}$.

(iii) $\omega \wedge *\eta = \eta \wedge *\omega = \langle \omega, \eta \rangle \mathrm{Vol}_M$, *for all* $\omega, \eta \in \mathcal{A}^k(M)$.

(iv) $*(\omega \wedge *\eta) = *(\eta \wedge *\omega) = \langle \omega, \eta \rangle$, *for all* $\omega, \eta \in \mathcal{A}^k(M)$.

(v) $\langle *\omega, *\eta \rangle = \langle \omega, \eta \rangle$, *for all* $\omega, \eta \in \mathcal{A}^k(M)$.

Recall that exterior differentiation, d, is a map, $d\colon \mathcal{A}^k(M) \to \mathcal{A}^{k+1}(M)$. Using the Hodge $*$-operator, we can define an operator, $\delta\colon \mathcal{A}^k(M) \to \mathcal{A}^{k-1}(M)$, that will turn out to be adjoint to d with respect to an inner product on $\mathcal{A}^\bullet(M)$.

Definition 15.1. Let M be an oriented Riemannian manifold of dimension n. For any k, with $1 \leq k \leq n$, let

$$\delta = (-1)^{n(k+1)+1} * d * .$$

Clearly, δ is a map, $\delta\colon \mathcal{A}^k(M) \to \mathcal{A}^{k-1}(M)$, and $\delta = 0$ on $\mathcal{A}^0(M) = C^\infty(M)$. It is easy to see that

$$*\delta = (-1)^k d*, \quad \delta* = (-1)^{k+1} * d, \quad \delta \circ \delta = 0.$$

15.2 The Laplace-Beltrami and Divergence Operators on Riemannian Manifolds

Using d and δ, we can generalize the Laplacian to an operator on differential forms.

Definition 15.2. Let M be an oriented Riemannian manifold of dimension n. The *Laplace-Beltrami operator*, for short, *Laplacian*, is the operator, $\Delta\colon \mathcal{A}^k(M) \to \mathcal{A}^k(M)$, defined by

$$\Delta = d\delta + \delta d.$$

A form, $\omega \in \mathcal{A}^k(M)$, such that $\Delta\omega = 0$ is a *harmonic form*. In particular, a function, $f \in \mathcal{A}^0(M) = C^\infty(M)$, such that $\Delta f = 0$ is called a *harmonic function*.

The Laplacian in Definition 15.2 is also called the *Hodge Laplacian*.

If $M = \mathbb{R}^n$ with the Euclidean metric and f is a smooth function, a laborious computation yields

$$\Delta f = -\sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2},$$

that is, the usual Laplacian *with a negative sign* in front (the computation can be found in Morita [115], Example 4.12 or Jost [84], Chapter 2, Section 2.1). It is also easy to see that Δ commutes with $*$, that is,

$$\Delta * = * \Delta.$$

Given any vector field, $X \in \mathfrak{X}(M)$, its *divergence*, $\operatorname{div} X$, is defined by

$$\operatorname{div} X = \delta X^\flat.$$

Now, for a function, $f \in C^\infty(M)$, we have $\delta f = 0$, so $\Delta f = \delta df$. However,

$$\operatorname{div}(\operatorname{grad} f) = \delta(\operatorname{grad} f)^\flat = \delta((df)^\sharp)^\flat = \delta df,$$

so

$$\Delta f = \operatorname{div} \operatorname{grad} f,$$

as in the case of $\mathbb{R}^n$.

Remark: Since the definition of δ involves two occurrences of the Hodge $*$-operator, δ also makes sense on non-orientable manifolds by using a local definition. Therefore, the Laplacian, Δ, also makes sense on non-orientable manifolds.

In the rest of this section, we assume that M is orientable.

The relationship between δ and d can be made clearer by introducing an inner product on forms with compact support. Recall that $\mathcal{A}_c^k(M)$ denotes the space of k-forms with compact support (an infinite dimensional vector space). For any two k-forms with compact support, $\omega, \eta \in \mathcal{A}_c^k(M)$, set

$$(\omega, \eta) = \int_M \langle \omega, \eta \rangle \operatorname{Vol}_M = \int_M \langle \omega, \eta \rangle \ * (1).$$

Using Proposition 15.1, we have

$$(\omega, \eta) = \int_M \langle \omega, \eta \rangle \operatorname{Vol}_M = \int_M \omega \wedge * \eta = \int_M \eta \wedge * \omega,$$

so it is easy to check that $(-,-)$ is indeed an inner product on k-forms with compact support. We can extend this inner product to forms with compact support in $\mathcal{A}_c^\bullet(M) = \bigoplus_{k=0}^n \mathcal{A}_c^k(M)$ by making $\mathcal{A}_c^h(M)$ and $\mathcal{A}_c^k(M)$ orthogonal if $h \neq k$.

Proposition 15.2. *If M is an orientable Riemannian manifold, then δ is (formally) adjoint to d, that is,*

$$(d\omega, \eta) = (\omega, \delta\eta),$$

for all k-forms, ω, η, with compact support.

Proof. By linearity and orthogonality of the $\mathcal{A}_c^k(M)$ the proof reduces to the case where $\omega \in \mathcal{A}_c^{k-1}(M)$ and $\eta \in \mathcal{A}_c^k(M)$ (both with compact support). By definition of δ and the fact that

$$** = (-\mathrm{id})^{(k-1)(n-k+1)}$$

for $*\colon \mathcal{A}^{k-1}(M) \to \mathcal{A}^{n-k+1}(M)$, we have

$$*\delta = (-1)^k d*,$$

and since

$$\begin{aligned} d(\omega \wedge *\eta) &= d\omega \wedge *\eta + (-1)^{k-1}\omega \wedge d * \eta \\ &= d\omega \wedge *\eta - \omega \wedge *\delta\eta \end{aligned}$$

we get

$$\begin{aligned} \int_M d(\omega \wedge *\eta) &= \int_M d\omega \wedge *\eta - \int_M \omega \wedge *\delta\eta \\ &= (d\omega, \eta) - (\omega, \delta\eta). \end{aligned}$$

However, by Stokes Theorem (Theorem 9.7),

$$\int_M d(\omega \wedge *\eta) = 0,$$

so $(d\omega, \eta) - (\omega, \delta\eta) = 0$, that is, $(d\omega, \eta) = (\omega, \delta\eta)$, as claimed. □

Corollary 15.3. *If M is an orientable Riemannian manifold, then the Laplacian, Δ is self-adjoint that is,*

$$(\Delta\omega, \eta) = (\omega, \Delta\eta),$$

for all k-forms, ω, η, with compact support.

We also obtain the following useful fact:

Proposition 15.4. *If M is an orientable Riemannian manifold, then for every k-form, ω, with compact support, $\Delta\omega = 0$ iff $d\omega = 0$ and $\delta\omega = 0$.*

Proof. Since $\Delta = d\delta + \delta d$, is is obvious that if $d\omega = 0$ and $\delta\omega = 0$, then $\Delta\omega = 0$. Conversely,

$$(\Delta\omega, \omega) = ((d\delta + \delta d)\omega, \omega) = (d\delta\omega, \omega) + (\delta d\omega, \omega) = (\delta\omega, \delta\omega) + (d\omega, d\omega).$$

Thus, if $\Delta\omega = 0$, then $(\delta\omega, \delta\omega) = (d\omega, d\omega) = 0$, which implies $d\omega = 0$ and $\delta\omega = 0$. □

As a consequence of Proposition 15.4, if M is a connected, orientable, compact Riemannian manifold, then every harmonic function on M is a constant.

For practical reasons, we need a formula for the Laplacian of a function, $f \in C^\infty(M)$, in local coordinates. If (U, φ) is a chart near p, as usual, let

$$\frac{\partial f}{\partial x_j}(p) = \frac{\partial (f \circ \varphi^{-1})}{\partial u_j}(\varphi(p)),$$

where $(u_1, \ldots, u_n)$ are the coordinate functions in $\mathbb{R}^n$. Write $|g| = \det(g_{ij})$, where (g_{ij}) is the symmetric, positive definite matrix giving the metric in the chart (U, φ).

Proposition 15.5. *If M is an orientable Riemannian manifold, then for every local chart, (U, φ), for every function, $f \in C^\infty(M)$, we have*

$$\Delta f = -\frac{1}{\sqrt{|g|}} \sum_{i,j} \frac{\partial}{\partial x_i} \left(\sqrt{|g|}\, g^{ij}\, \frac{\partial f}{\partial x_j} \right).$$

Proof. We follow Jost [84], Chapter 2, Section 1. Pick any function, $h \in C^\infty(M)$, with compact support. We have

$$\begin{aligned}
\int_M (\Delta f) h * (1) &= (\Delta f, h) \\
&= (\delta df, h) \\
&= (df, dh) \\
&= \int_M \langle df, dh \rangle * (1) \\
&= \int_M \sum_{ij} g^{ij} \frac{\partial f}{\partial x_i} \frac{\partial h}{\partial x_j} * (1) \\
&= -\int_M \sum_{ij} \frac{1}{\sqrt{|g|}} \frac{\partial}{\partial x_j} \left(\sqrt{|g|}\, g^{ij} \frac{\partial f}{\partial x_i} \right) h * (1),
\end{aligned}$$

where we have used integration by parts in the last line. Since the above equation holds for all h, we get our result. □

It turns out that in a Riemannian manifold, the divergence of a vector field and the Laplacian of a function can be given a definition that uses the covariant derivative (see Chapter 11, Section 11.1) instead of the Hodge $*$-operator. For the sake of completeness, we present this alternate definition which is the one used in Gallot, Hulin and Lafontaine [61] (Chapter 4) and O'Neill [120] (Chapter 3). If ∇ is the Levi-Civita connection induced by the Riemannian metric, then the *divergence* of a vector field, $X \in \mathfrak{X}(M)$, is the function, $\operatorname{div} X \colon M \to \mathbb{R}$, defined so that

$$(\operatorname{div} X)(p) = \operatorname{tr}(Y(p) \mapsto (-\nabla_Y X)_p),$$

namely, for every p, $(\operatorname{div} X)(p)$ is the trace of the linear map, $Y(p) \mapsto (-\nabla_Y X)_p$. Of course, for any function, $f \in C^\infty(M)$, we define Δf by

$$\Delta f = \operatorname{div} \operatorname{grad} f.$$

Observe that the above definition of the divergence (and of the Laplacian) makes sense even if M is non-orientable. For orientable manifolds, the equivalence of this new definition of the divergence with our definition is proved in Petersen [122], see Chapter 3, Proposition 31. The main reason is that

$$L_X \operatorname{Vol}_M = -(\operatorname{div} X)\operatorname{Vol}_M$$

and by Cartan's Formula (Proposition 8.15), $L_X = i(X) \circ d + d \circ i(X)$, as $d\operatorname{Vol}_M = 0$, we get

$$(\operatorname{div} X)\operatorname{Vol}_M = -d(i(X)\operatorname{Vol}_M).$$

The above formulae also holds for a local volume form (*i.e.* for a volume form on a local chart).

The operator, $\delta\colon \mathcal{A}^1(M) \to \mathcal{A}^0(M)$, can also be defined in terms of the covariant derivative (see Gallot, Hulin and Lafontaine [61], Chapter 4). For any one-form, $\omega \in \mathcal{A}^1(M)$, recall that

$$(\nabla_X \omega)(Y) = X(\omega(Y)) - \omega(\nabla_X Y).$$

Then, it turns out that

$$\delta\omega = -\operatorname{tr} \nabla\omega,$$

where the trace should be interpreted as the trace of the $\mathbb{R}$-bilinear map, $X, Y \mapsto (\nabla_X \omega)(Y)$, as in Chapter 22, see Proposition 22.2. This means that in any chart, (U, φ),

$$\delta\omega = -\sum_{i=1}^{n} (\nabla_{E_i}\omega)(E_i),$$

for any orthonormal frame field, $(E_1, \ldots, E_n)$ over U. It can be shown that

$$\delta(f df) = f\Delta f - \langle \operatorname{grad} f, \operatorname{grad} f\rangle,$$

and, as a consequence,

$$(\Delta f, f) = \int_M \langle \operatorname{grad} f, \operatorname{grad} f\rangle \operatorname{Vol}_M,$$

for any orientable, compact manifold, M.

Since the proof of the next proposition is quite technical, we omit the proof.

Proposition 15.6. *If M is an orientable and compact Riemannian manifold, then for every vector field, $X \in \mathfrak{X}(M)$, we have*

$$\operatorname{div} X = \delta X^\flat.$$

Consequently, for the Laplacian, we have

$$\Delta f = \delta df = \operatorname{div} \operatorname{grad} f.$$

Remark: Some authors omit the negative sign in the definition of the divergence, that is, they define

$$(\operatorname{div} X)(p) = \operatorname{tr}(Y(p) \mapsto (\nabla_Y X)_p).$$

Here is a frequently used corollary of Proposition 15.6:

Proposition 15.7. *(Green's Formula) If M is an orientable and compact Riemannian manifold without boundary, then for every vector field, $X \in \mathfrak{X}(M)$, we have*

$$\int_M (\operatorname{div} X)\operatorname{Vol}_M = 0.$$

Proofs of proposition 15.7 can be found in Gallot, Hulin and Lafontaine [61] (Chapter 4, Proposition 4.9) and Helgason [73] (Chapter 2, Section 2.4).

There is a generalization of the formula expressing $\delta\omega$ over an orthonormal frame, $E_1, \ldots,$ E_n, for a one-form, ω, that applies to any differential form. In fact, there are formulae expressing both d and δ over an orthornormal frame and its coframe and these are often handy in proofs. Recall that for every vector field, $X \in \mathfrak{X}(M)$, the interior product, $i(X)\colon \mathcal{A}^{k+1}(M) \to \mathcal{A}^k(M)$, is defined by

$$(i(X)\omega)(Y_1, \ldots, Y_k) = \omega(X, Y_1, \ldots, Y_k),$$

for all $Y_1, \ldots, Y_k \in \mathfrak{X}(M)$.

Proposition 15.8. *Let M be a compact, orientable, Riemannian manifold. For every $p \in M$, for every local chart, (U, φ), with $p \in M$, if $(E_1, \ldots, E_n)$ is an orthonormal frame over U and $(\theta_1, \ldots, \theta_n)$ is its dual coframe, then for every k-form, $\omega \in \mathcal{A}^k(M)$, we have:*

$$\begin{aligned} d\omega &= \sum_{i=1}^n \theta_i \wedge \nabla_{E_i}\omega \\ \delta\omega &= -\sum_{i=1}^n i(E_i)\nabla_{E_i}\omega. \end{aligned}$$

A proof of Proposition 15.8 can be found in Petersen [122] (Chapter 7, proposition 37) or Jost [84] (Chapter 3, Lemma 3.3.4). When ω is a one-form, $\delta\omega_p$ is just a number and indeed,

$$\delta\omega = -\sum_{i=1}^n i(E_i)\nabla_{E_i}\omega = -\sum_{i=1}^n (\nabla_{E_i}\omega)(E_i),$$

as stated earlier.

15.3 Harmonic Forms, the Hodge Theorem, Poincaré Duality

Let us now assume that M is orientable and compact.

Definition 15.3. Let M be an orientable and compact Riemannian manifold of dimension n. For every k, with $0 \leq k \leq n$, let

$$\mathbb{H}^k(M) = \{\omega \in \mathcal{A}^k(M) \mid \Delta\omega = 0\},$$

the space of *harmonic k-forms*.

The following proposition is left as an easy exercise:

Proposition 15.9. *Let M be an orientable and compact Riemannian manifold of dimension n. The Laplacian commutes with the Hodge $*$-operator and we have a linear map,*

$$*\colon \mathbb{H}^k(M) \to \mathbb{H}^{n-k}(M).$$

One of the deepest and most important theorems about manifolds is the *Hodge decomposition theorem* which we now state.

Theorem 15.10. *(Hodge Decomposition Theorem) Let M be an orientable and compact Riemannian manifold of dimension n. For every k, with $0 \leq k \leq n$, the space, $\mathbb{H}^k(M)$, is finite dimensional and we have the following orthogonal direct sum decomposition of the space of k-forms:*

$$\mathcal{A}^k(M) = \mathbb{H}^k(M) \oplus d(\mathcal{A}^{k-1}(M)) \oplus \delta(\mathcal{A}^{k+1}(M)).$$

The proof of Theorem 15.10 involves a lot of analysis and it is long and complicated. A complete proof can be found in Warner [148], Chapter 6. Other treatments of Hodge theory can be found in Morita [115] (Chapter 4) and Jost [84] (Chapter 2).

The Hodge Decomposition Theorem has a number of important corollaries, one of which is *Hodge Theorem*:

Theorem 15.11. *(Hodge Theorem) Let M be an orientable and compact Riemannian manifold of dimension n. For every k, with $0 \leq k \leq n$, there is an isomorphism between $\mathbb{H}^k(M)$ and the de Rham cohomology vector space, $H^k_{\mathrm{DR}}(M)$:*

$$H^k_{\mathrm{DR}}(M) \cong \mathbb{H}^k(M).$$

Proof. Since by Proposition 15.4, every harmonic form, $\omega \in \mathbb{H}^k(M)$, is closed, we get a linear map from $\mathbb{H}^k(M)$ to $H^k_{\mathrm{DR}}(M)$ by assigning its cohomology class, $[\omega]$, to ω. This map is injective. Indeed if $[\omega] = 0$ for some $\omega \in \mathbb{H}^k(M)$, then $\omega = d\eta$, for some $\eta \in \mathcal{A}^{k-1}(M)$ so

$$(\omega, \omega) = (d\eta, \omega) = (\eta, \delta\omega).$$

But, as $\omega \in \mathbb{H}^k(M)$ we have $\delta\omega = 0$ by Proposition 15.4, so $(\omega, \omega) = 0$, that is, $\omega = 0$.

Our map is also surjective, this is the hard part of Hodge Theorem. By the Hodge Decomposition Theorem, for every closed form, $\omega \in \mathcal{A}^k(M)$, we can write

$$\omega = \omega_H + d\eta + \delta\theta,$$

with $\omega_H \in \mathbb{H}^k(M)$, $\eta \in \mathcal{A}^{k-1}(M)$ and $\theta \in \mathcal{A}^{k+1}(M)$. Since ω is closed and $\omega_H \in \mathbb{H}^k(M)$, we have $d\omega = 0$ and $d\omega_H = 0$, thus

$$d\delta\theta = 0$$

and so

$$0 = (d\delta\theta, \theta) = (\delta\theta, \delta\theta),$$

that is, $\delta\theta = 0$. Therefore, $\omega = \omega_H + d\eta$, which implies $[\omega] = [\omega_H]$, with $\omega_H \in \mathbb{H}^k(M)$, proving the surjectivity of our map. □

The Hodge Theorem also implies the *Poincaré Duality Theorem*. If M is a compact, orientable, n-dimensional smooth manifold, for each k, with $0 \leq k \leq n$, we define a bilinear map,

$$((-,-))\colon H^k_{\mathrm{DR}}(M) \times H^{n-k}_{\mathrm{DR}}(M) \longrightarrow \mathbb{R},$$

by setting

$$(([\omega], [\eta])) = \int_M \omega \wedge \eta.$$

We need to check that this definition does not depend on the choice of closed forms in the cohomology classes $[\omega]$ and $[\eta]$. However, as $d\omega = d\eta = 0$, we have

$$d(\alpha \wedge \eta + (-1)^k \omega \wedge \beta + \alpha \wedge d\beta) = d\alpha \wedge \eta + \omega \wedge d\beta + d\alpha \wedge d\beta,$$

so by Stokes' Theorem,

$$\begin{aligned} \int_M (\omega + d\alpha) \wedge (\eta + d\beta) &= \int_M \omega \wedge \eta + \int_M d(\alpha \wedge \eta + (-1)^k \omega \wedge \beta + \alpha \wedge d\beta) \\ &= \int_M \omega \wedge \eta. \end{aligned}$$

Theorem 15.12. *(Poincaré Duality) If M is a compact, orientable, smooth manifold of dimension n, then the bilinear map*

$$((-,-))\colon H^k_{\mathrm{DR}}(M) \times H^{n-k}_{\mathrm{DR}}(M) \longrightarrow \mathbb{R}$$

defined above is a nondegenerate pairing and hence, yields an isomorphism

$$H^k_{\mathrm{DR}}(M) \cong (H^{n-k}_{\mathrm{DR}}(M))^*.$$

Proof. Pick any Riemannian metric on M. It is enough to show that for every nonzero cohomology class, $[\omega] \in H^k_{\mathrm{DR}}(M)$, there is some $[\eta] \in H^{n-k}_{\mathrm{DR}}(M)$ such that

$$(([\omega],[\eta])) = \int_M \omega \wedge \eta \neq 0.$$

By Hodge Theorem, we may assume that ω is a nonzero harmonic form. By Proposition 15.9, $\eta = *\omega$ is also harmonic and $\eta \in \mathbb{H}^{n-k}(M)$. Then, we get

$$(\omega,\omega) = \int_M \omega \wedge *\omega = (([\omega],[\eta]))$$

and indeed, $(([\omega],[\eta])) \neq 0$, since $\omega \neq 0$. □

15.4 The Connection Laplacian, Weitzenböck Formula and the Bochner Technique

If M is compact, orientable, Riemannian manifold, then the inner product, $\langle -,-\rangle_p$, on T_pM (with $p \in M$) induces an inner product on differential forms, as we explained in Section 15.2. We also get an inner product on vector fields if, for any two vector field, $X, Y \in \mathfrak{X}(M)$, we define (X,Y) by

$$(X,Y) = \int_M \langle X,Y\rangle \,\mathrm{Vol}_M,$$

where $\langle X,Y\rangle$ is the function defined pointwise by

$$\langle X,Y\rangle(p) = \langle X(p),Y(p)\rangle_p.$$

Using Proposition 11.5, we can define the covariant derivative, $\nabla_X\omega$, of any k-form, $\omega \in \mathcal{A}^k(M)$, as the k-form given by

$$(\nabla_X\omega)(Y_1,\ldots,Y_k) = X(\omega(Y_1,\ldots,Y_k)) - \sum_{j=1}^{k} \omega(Y_1,\ldots,\nabla_X Y_j,\ldots,Y_k).$$

We can view ∇ as linear map,

$$\nabla\colon \mathcal{A}^k(M) \to \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M),\mathcal{A}^k(M)),$$

where $\nabla\omega$ is the $C^\infty(M)$-linear map, $X \mapsto \nabla_X\omega$. The inner product on $\mathcal{A}^k(M)$ allows us to define the (formal) adjoint, ∇^*, of ∇, as a linear map

$$\nabla^*\colon \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M),\mathcal{A}^k(M)) \to \mathcal{A}^k(M).$$

For any linear map, $A \in \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M),\mathcal{A}^k(M))$, let A^* be the adjoint of A defined by

$$(AX,\theta) = (X,A^*\theta),$$

for all vector fields $X \in \mathfrak{X}(M)$ and all k-forms, $\theta \in \mathcal{A}^k(M)$. It can be verified that $A^* \in \mathrm{Hom}_{C^\infty(M)}(\mathcal{A}^k(M), \mathfrak{X}(M))$. Then, given $A, B \in \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$, the expression $\mathrm{tr}(A^*B)$ is a smooth function on M and it can be verified that

$$\langle A, B\rangle = \mathrm{tr}(A^*B)$$

defines a non-degenerate pairing on $\mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$. Using this pairing we obtain the ($\mathbb{R}$-valued) inner product on $\mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$ given by

$$(A, B) = \int_M \mathrm{tr}(A^*B)\,\mathrm{Vol}_M.$$

Using all this, the (formal) adjoint, ∇^*, of $\nabla\colon \mathcal{A}^k(M) \to \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$ is the linear map, $\nabla^*\colon \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M)) \to \mathcal{A}^k(M)$, defined implicitly by

$$(\nabla^* A, \omega) = (A, \nabla\omega),$$

that is,

$$\int_M \langle \nabla^* A, \omega\rangle\,\mathrm{Vol}_M = \int_M \langle A, \nabla\omega\rangle\,\mathrm{Vol}_M,$$

for all $A \in \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$ and all $\omega \in \mathcal{A}^k(M)$.

The notation ∇^* for the adjoint of ∇ should not be confused with the dual connection on T^*M of a connection, ∇, on TM! Here, ∇ denotes the connection on $\mathcal{A}^*(M)$ induced by the orginal connection, ∇, on TM. The argument type (differential form or vector field) should make it clear which ∇ is intended but it might have been better to use a notation such as $\nabla^\top$ instead of ∇^*.

What we just did also applies to $\mathcal{A}^*(M) = \bigoplus_{k=0}^n \mathcal{A}^k(M)$ (where $\dim(M) = n$) and so we can view the connection, ∇, as a linear map, $\nabla\colon \mathcal{A}^*(M) \to \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^*(M))$ and its adjoint as a linear map, $\nabla^*\colon \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^*(M)) \to \mathcal{A}^*(M)$.

Definition 15.4. Given a compact, orientable, Riemannian manifold, M, the *connection Laplacian* (or *Bochner Laplacian*), $\nabla^*\nabla$, is defined as the composition of the connection, $\nabla\colon \mathcal{A}^*(M) \to \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^*(M))$, with its adjoint, $\nabla^*\colon \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^*(M)) \to \mathcal{A}^*(M)$, as defined above.

Observe that

$$(\nabla^*\nabla\omega, \omega) = (\nabla\omega, \nabla\omega) = \int_M \langle \nabla\omega, \nabla\omega\rangle\,\mathrm{Vol}_M,$$

for all $\omega \in \mathcal{A}^k(M)$. Consequently, the "harmonic forms", ω, with respect to $\nabla^*\nabla$ must satisfy

$$\nabla\omega = 0,$$

but this condition is not equivalent to the harmonicity of ω with respect to the Hodge Laplacian. Thus, in general, $\nabla^*\nabla$ and Δ are different operators. The relationship between the two is given by formulae involving contractions of the curvature tensor and known as *Weitzenböck formulae*. We will state such a formula in case of one-forms later on. But first, we can give another definition of the connection Laplacian using *second covariant derivatives* of forms. Given any k-form, $\omega \in \mathcal{A}^k(M)$, for any two vector fields, $X, Y \in \mathfrak{X}(M)$, we define $\nabla^2_{X,Y}\omega$ by

$$\nabla^2_{X,Y}\omega = \nabla_X(\nabla_Y\omega) - \nabla_{\nabla_X Y}\omega.$$

Given any local chart, (U, φ), and given any orthormal frame, $(E_1, \ldots, E_n)$, over U, we can take the trace, $\mathrm{tr}(\nabla^2\omega)$, of $\nabla^2_{X,Y}\omega$, defined by

$$\mathrm{tr}(\nabla^2\omega) = \sum_{i=1}^{n} \nabla^2_{E_i,E_i}\omega.$$

It is easily seen that $\mathrm{tr}(\nabla^2\omega)$ does not depend on the choice of local chart and orthonormal frame.

Proposition 15.13. *If is M a compact, orientable, Riemannian manifold, then the connection Laplacian, $\nabla^*\nabla$, is given by*

$$\nabla^*\nabla\omega = -\mathrm{tr}(\nabla^2\omega),$$

for all differential forms, $\omega \in \mathcal{A}^*(M)$.

The proof of Proposition 15.13, which is quite technical, can be found in Petersen [122] (Chapter 7, Proposition 34).

We are now ready to prove the Weitzenböck formulae for one-forms.

Theorem 15.14. *(Weitzenböck–Bochner Formula) If is M a compact, orientable, Riemannian manifold, then for every one-form, $\omega \in \mathcal{A}^1(M)$, we have*

$$\Delta\omega = \nabla^*\nabla\omega + \mathrm{Ric}(\omega),$$

where $\mathrm{Ric}(\omega)$ *is the one-form given by*

$$\mathrm{Ric}(\omega)(X) = \omega(\mathrm{Ric}^\sharp(X)),$$

where $\mathrm{Ric}^\sharp$ *is the Ricci curvature viewed as a $(1,1)$-tensor (that is, $\langle \mathrm{Ric}^\sharp(u), v\rangle_p = \mathrm{Ric}(u, v)$, for all $u, v \in T_pM$ and all $p \in M$).*

Proof. For any $p \in M$, pick any normal local chart, (U, φ), with $p \in U$, and pick any orthonormal frame, $(E_1, \ldots, E_n)$, over U. Because (U, φ) is a normal chart, at p, we have

$(\nabla_{E_j} E_j)_p = 0$ for all i, j. Recall from the discussion at the end of Section 15.2 that for every one-form, ω, we have

$$\delta\omega = -\sum_i \nabla_{E_i}\omega(E_i),$$

and so

$$d\delta\omega = -\sum_i \nabla_X \nabla_{E_i}\omega(E_i).$$

Also recall that

$$d\omega(X, Y) = \nabla_X \omega(Y) - \nabla_Y \omega(X),$$

and using Proposition 15.8 we can show that

$$\delta d\omega(X) = -\sum_i \nabla_{E_i}\nabla_{E_i}\omega(X) + \sum_i \nabla_{E_i}\nabla_X \omega(E_i).$$

Thus, we get

$$\begin{aligned}
\Delta\omega(X) &= -\sum_i \nabla_{E_i}\nabla_{E_i}\omega(X) + \sum_i (\nabla_{E_i}\nabla_X - \nabla_X\nabla_{E_i})\omega(E_i) \\
&= -\sum_i \nabla^2_{E_i,E_i}\omega(X) + \sum_i (\nabla^2_{E_i,X} - \nabla^2_{X,E_i})\omega(E_i) \\
&= \nabla^*\nabla\omega(X) + \sum_i \omega(R(E_i, X)E_i) \\
&= \nabla^*\nabla\omega(X) + \omega(\mathrm{Ric}^\sharp(X)),
\end{aligned}$$

using the fact that $(\nabla_{E_j} E_j)_p = 0$ for all i, j and using Proposition 13.2 and Proposition 15.13. □

For simplicity of notation, we will write $\mathrm{Ric}(u)$ for $\mathrm{Ric}^\sharp(u)$. There should be no confusion since $\mathrm{Ric}(u, v)$ denotes the Ricci curvature, a $(0, 2)$-tensor. There is another way to express $\mathrm{Ric}(\omega)$ which will be useful in the proof of the next theorem. Observe that

$$\begin{aligned}
\mathrm{Ric}(\omega)(Z) &= \omega(\mathrm{Ric}(Z)) \\
&= \langle \omega^\sharp, \mathrm{Ric}(Z)\rangle \\
&= \langle \mathrm{Ric}(Z), \omega^\sharp\rangle \\
&= \mathrm{Ric}(Z, \omega^\sharp) \\
&= \mathrm{Ric}(\omega^\sharp, Z) \\
&= \langle \mathrm{Ric}(\omega^\sharp), Z\rangle \\
&= (\mathrm{Ric}(\omega^\sharp))^\flat(Z),
\end{aligned}$$

and thus,

$$\mathrm{Ric}(\omega)(Z) = (\mathrm{Ric}(\omega^\sharp))^\flat(Z).$$

Consequently the Weitzenböck formula can be written as

$$\Delta\omega = \nabla^*\nabla\omega + (\mathrm{Ric}(\omega^\sharp))^\flat.$$

The Weitzenböck–Bochner Formula implies the following theorem due to Bochner:

Theorem 15.15. *(Bochner) If M is a compact, orientable, connected Riemannian manifold, then the following properties hold:*

(i) If the Ricci curvature is non-negative, that is $\mathrm{Ric}(u,u) \geq 0$ *for all* $p \in M$ *and all* $u \in T_pM$ *and if* $\mathrm{Ric}(u,u) > 0$ *for some* $p \in M$ *and all* $u \in T_pM$*, then* $H^1_{\mathrm{DR}}M = (0)$.

(ii) If the Ricci curvature is non-negative, then $\nabla\omega = 0$ *for all* $\omega \in \mathcal{A}^1(M)$ *and* $\dim H^1_{\mathrm{DR}}M \leq \dim M$.

Proof. (i) Assume $H^1_{\mathrm{DR}}M \neq (0)$. Then, by the Hodge Theorem, there is some nonzero harmonic one-form, ω. The Weitzenböck–Bochner Formula implies that

$$(\Delta\omega,\omega) = (\nabla^*\nabla\omega,\omega) + ((\mathrm{Ric}(\omega^\sharp))^\flat,\omega).$$

Since $\Delta\omega = 0$, we get

$$\begin{aligned} 0 &= (\nabla^*\nabla\omega,\omega) + ((\mathrm{Ric}(\omega^\sharp))^\flat,\omega) \\ &= (\nabla\omega,\nabla\omega) + \int_M \langle(\mathrm{Ric}(\omega^\sharp))^\flat,\omega\rangle \,\mathrm{Vol}_M \\ &= (\nabla\omega,\nabla\omega) + \int_M \langle\mathrm{Ric}(\omega^\sharp),\omega^\sharp\rangle \,\mathrm{Vol}_M \\ &= (\nabla\omega,\nabla\omega) + \int_M \mathrm{Ric}(\omega^\sharp,\omega^\sharp) \,\mathrm{Vol}_M. \end{aligned}$$

However, $(\nabla\omega,\nabla\omega) \geq 0$ and by the assumption on the Ricci curvature, the integrand is nonnegative and strictly positive at some point, so the integral is strictly positive, a contradiction.

(ii) Again, for any one-form, ω, we have

$$(\Delta\omega,\omega) = (\nabla\omega,\nabla\omega) + \int_M \mathrm{Ric}(\omega^\sharp,\omega^\sharp) \,\mathrm{Vol}_M,$$

and so, if the Ricci curvature is non-negative, $\Delta\omega = 0$ iff $\nabla\omega = 0$. This means that ω is invariant by parallel transport (see Section 11.3) and thus, ω is completely determined by its value, ω_p, at some point, $p \in M$, so there is an injection, $\mathbb{H}^1(M) \longrightarrow T^*_pM$, which implies that $\dim H^1_{\mathrm{DR}}M = \dim \mathbb{H}^1(M) \leq \dim M$. □

There is a version of the Weitzenböck formula for p-forms but it involves a more complicated curvature term and its proof is also more complicated. The Bochner technique can

also be generalized in various ways, in particular, to *spin manifolds*, but these considerations are beyond the scope of these notes. Let me just say that Weitzenböck formulae involving the Dirac operator play an important role in physics and 4-manifold geometry. We refer the interested reader to Gallot, Hulin and Lafontaine [61] (Chapter 4) Petersen [122] (Chapter 7), Jost [84] (Chaper 3) and Berger [16] (Section 15.6) for more details on Weitzenböck formulae and the Bochner technique.

Chapter 16

Spherical Harmonics and Linear Representations of Lie Groups

16.1 Introduction, Spherical Harmonics on the Circle

In this chapter, we discuss spherical harmonics and take a glimpse at the linear representation of Lie groups. Spherical harmonics on the sphere, S^2, have interesting applications in computer graphics and computer vision so this material is not only important for theoretical reasons but also for practical reasons.

Joseph Fourier (1768-1830) invented Fourier series in order to solve the heat equation [56]. Using Fourier series, every square-integrable periodic function, f, (of period 2π) can be expressed uniquely as the sum of a power series of the form

$$f(\theta) = a_0 + \sum_{k=1}^{\infty}(a_k \cos k\theta + b_k \cos k\theta),$$

where the *Fourier coefficients*, a_k, b_k, of f are given by the formulae

$$a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\theta)\, d\theta, \quad a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)\cos k\theta\, d\theta, \quad b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\theta)\sin k\theta\, d\theta,$$

for $k \geq 1$. The reader will find the above formulae in Fourier's famous book [56] in Chapter III, Section 233, page 256, essentially using the notation that we use nowdays.

This remarkable discovery has many theoretical and practical applications in physics, signal processing, engineering, *etc.* We can describe Fourier series in a more conceptual manner if we introduce the following inner product on square-integrable functions:

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(\theta)g(\theta)\, d\theta,$$

which we will also denote by

$$\langle f, g\rangle = \int_{S^1} f(\theta)g(\theta)\, d\theta,$$

where S^1 denotes the unit circle. After all, periodic functions of (period 2π) can be viewed as functions on the circle. With this inner product, the space $L^2(S^1)$ is a complete normed vector space, that is, a Hilbert space. Furthermore, if we define the subspaces, $\mathcal{H}_k(S^1)$, of $L^2(S^1)$, so that $\mathcal{H}_0(S^1)\,(=\mathbb{R})$ is the set of constant functions and $\mathcal{H}_k(S^1)$ is the two-dimensional space spanned by the functions $\cos k\theta$ and $\sin k\theta$, then it turns out that we have a Hilbert sum decomposition

$$L^2(S^1) = \bigoplus_{k=0}^{\infty} \mathcal{H}_k(S^1)$$

into pairwise orthogonal subspaces, where $\bigcup_{k=0}^{\infty} \mathcal{H}_k(S^1)$ is dense in $L^2(S^1)$. The functions $\cos k\theta$ and $\sin k\theta$ are also orthogonal in $\mathcal{H}_k(S^1)$.

Now, it turns out that the spaces, $\mathcal{H}_k(S^1)$, arise naturally when we look for homogeneous solutions of the Laplace equation, $\Delta f = 0$, in $\mathbb{R}^2$ (Pierre-Simon Laplace, 1749-1827). Roughly speaking, a homogeneous function in $\mathbb{R}^2$ is a function that can be expressed in polar coordinates, (r, θ), as

$$f(r, \theta) = r^k g(\theta).$$

Recall that the Laplacian on $\mathbb{R}^2$ expressed in cartesian coordinates, (x, y), is given by

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2},$$

where $f\colon \mathbb{R}^2 \to \mathbb{R}$ is a function which is at least of class C^2. In polar coordinates, (r, θ), where $(x, y) = (r\cos\theta, r\sin\theta)$ and $r > 0$, the Laplacian is given by

$$\Delta f = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial f}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2}.$$

If we restrict f to the unit circle, S^1, then the Laplacian on S^1 is given by

$$\Delta_{s^1} f = \frac{\partial^2 f}{\partial \theta^2}.$$

It turns out that *the space $\mathcal{H}_k(S^1)$ is the eigenspace of Δ_{S^1} for the eigenvalue $-k^2$.*

To show this, we consider another question, namely, *what are the harmonic functions on* $\mathbb{R}^2$, that is, the functions, f, that are solutions of the *Laplace equation*,

$$\Delta f = 0.$$

Our ancestors had the idea that the above equation can be solved by *separation of variables*. This means that we write $f(r, \theta) = F(r)g(\theta)$, where $F(r)$ and $g(\theta)$ are independent functions.

To make things easier, let us assume that $F(r) = r^k$, for some integer $k \geq 0$, which means that we assume that f is a *homogeneous function of degree* k. Recall that a function, $f\colon \mathbb{R}^2 \to \mathbb{R}$, is *homogeneous of degree* k iff

$$f(tx, ty) = t^k f(x, y) \qquad \text{for all } t > 0.$$

Now, using the Laplacian in polar coordinates, we get

$$\begin{aligned}
\Delta f &= \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial(r^k g(\theta))}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2(r^k g(\theta))}{\partial \theta^2} \\
&= \frac{1}{r}\frac{\partial}{\partial r}\left(k r^k g\right) + r^{k-2}\frac{\partial^2 g}{\partial \theta^2} \\
&= r^{k-2}k^2 g + r^{k-2}\frac{\partial^2 g}{\partial \theta^2} \\
&= r^{k-2}(k^2 g + \Delta_{S^1} g).
\end{aligned}$$

Thus, we deduce that

$$\Delta f = 0 \quad \text{iff} \quad \Delta_{S^1} g = -k^2 g,$$

that is, g is an eigenfunction of Δ_{S^1} for the eigenvalue $-k^2$. But, the above equation is equivalent to the second-order differential equation

$$\frac{d^2 g}{d\theta^2} + k^2 g = 0,$$

whose general solution is given by

$$g(\theta) = a_n \cos k\theta + b_n \sin k\theta.$$

In summary, we found that the integers, $0, -1, -4, -9, \ldots, -k^2, \ldots$ are eigenvalues of Δ_{S^1} and that the functions $\cos k\theta$ and $\sin k\theta$ are eigenfunctions for the eigenvalue $-k^2$, with $k \geq 0$. So, it looks like the dimension of the eigenspace corresponding to the eigenvalue $-k^2$ is 1 when $k = 0$ and 2 when $k \geq 1$.

It can indeed be shown that Δ_{S^1} has no other eigenvalues and that the dimensions claimed for the eigenspaces are correct. Observe that if we go back to our homogeneous harmonic functions, $f(r, \theta) = r^k g(\theta)$, we see that this space is spanned by the functions

$$u_k = r^k \cos k\theta, \quad v_k = r^k \sin k\theta.$$

Now, $(x + iy)^k = r^k(\cos k\theta + i \sin k\theta)$, and since $\Re(x + iy)^k$ and $\Im(x + iy)^k$ are homogeneous polynomials, we see that u_k and v_k are homogeneous polynomials called *harmonic polynomials*. For example, here is a list of a basis for the harmonic polynomials (in two variables) of degree $k = 0, 1, 2, 3, 4$:

$k = 0$	1
$k = 1$	$x,\ y$
$k = 2$	$x^2 - y^2,\ xy$
$k = 3$	$x^3 - 3xy^2,\ 3x^2y - y^3$
$k = 4$	$x^4 - 6x^2y^2 + y^4,\ x^3y - xy^3.$

Therefore, the eigenfunctions of the Laplacian on S^1 are the restrictions of the harmonic polynomials on $\mathbb{R}^2$ to S^1 and we have a Hilbert sum decomposition, $L^2(S^1) = \bigoplus_{k=0}^{\infty} \mathcal{H}_k(S^1)$. It turns out that this phenomenon generalizes to the sphere $S^n \subseteq \mathbb{R}^{n+1}$ for all $n \geq 1$.

Let us take a look at next case, $n = 2$.

16.2 Spherical Harmonics on the 2-Sphere

The material of section is very classical and can be found in many places, for example Andrews, Askey and Roy [2] (Chapter 9), Sansone [133] (Chapter III), Hochstadt [79] (Chapter 6) and Lebedev [98] (Chapter). We recommend the exposition in Lebedev [98] because we find it particularly clear and uncluttered. We have also borrowed heavily from some lecture notes by Hermann Gluck for a course he offered in 1997-1998.

Our goal is to find the homogeneous solutions of the Laplace equation, $\Delta f = 0$, in $\mathbb{R}^3$, and to show that they correspond to spaces, $\mathcal{H}_k(S^2)$, of eigenfunctions of the Laplacian, Δ_{S^2}, on the 2-sphere,

$$S^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}.$$

Then, the spaces $\mathcal{H}_k(S^2)$ will give us a Hilbert sum decomposition of the Hilbert space, $L^2(S^2)$, of square-integrable functions on S^2. This is the generalization of Fourier series to the 2-sphere and the functions in the spaces $\mathcal{H}_k(S^2)$ are called *spherical harmonics*.

The Laplacian in $\mathbb{R}^3$ is of course given by

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}.$$

If we use spherical coordinates

$$\begin{aligned} x &= r \sin\theta \cos\varphi \\ y &= r \sin\theta \sin\varphi \\ z &= r \cos\theta, \end{aligned}$$

in $\mathbb{R}^3$, where $0 \leq \theta < \pi$, $0 \leq \varphi < 2\pi$ and $r > 0$ (recall that φ is the so-called *azimuthal angle* in the xy-plane originating at the x-axis and θ is the so-called *polar angle* from the z-axis, angle defined in the plane obtained by rotating the xz-plane around the z-axis by the angle φ), then the Laplacian in spherical coordinates is given by

$$\Delta f = \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial f}{\partial r} \right) + \frac{1}{r^2} \Delta_{S^2} f,$$

where

$$\Delta_{S^2} f = \frac{1}{\sin\theta} \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{\sin^2\theta} \frac{\partial^2 f}{\partial \varphi^2},$$

is the Laplacian on the sphere, S^2 (for example, see Lebedev [98], Chapter 8 or Section 16.3, where we derive this formula). Let us look for homogeneous harmonic functions, $f(r,\theta,\varphi) = r^k g(\theta,\varphi)$, on $\mathbb{R}^3$, that is, solutions of the Laplace equation

$$\Delta f = 0.$$

We get

$$\begin{aligned}
\Delta f &= \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial (r^k g)}{\partial r}\right) + \frac{1}{r^2}\Delta_{S^2}(r^k g)\\
&= \frac{1}{r^2}\frac{\partial}{\partial r}\left(k r^{k+1} g\right) + r^{k-2}\Delta_{S^2} g\\
&= r^{k-2}k(k+1)g + r^{k-2}\Delta_{S^2} g\\
&= r^{k-2}(k(k+1)g + \Delta_{S^2} g).
\end{aligned}$$

Therefore,

$$\Delta f = 0 \quad \text{iff} \quad \Delta_{S^2} g = -k(k+1)g,$$

that is, g is an eigenfunction of Δ_{S^2} for the eigenvalue $-k(k+1)$.

We can look for solutions of the above equation using the separation of variables method. If we let $g(\theta,\varphi) = \Theta(\theta)\Phi(\varphi)$, then we get the equation

$$\frac{\Phi}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Theta}{\partial\theta}\right) + \frac{\Theta}{\sin^2\theta}\frac{\partial^2\Phi}{\partial\varphi^2} = -k(k+1)\Theta\Phi,$$

that is, dividing by $\Theta\Phi$ and multiplying by $\sin^2\theta$,

$$\frac{\sin\theta}{\Theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Theta}{\partial\theta}\right) + k(k+1)\sin^2\theta = -\frac{1}{\Phi}\frac{\partial^2\Phi}{\partial\varphi^2}.$$

Since Θ and Φ are independent functions, the above is possible only if both sides are equal to a constant, say μ. This leads to two equations

$$\begin{aligned}
&\frac{\partial^2\Phi}{\partial\varphi^2} + \mu\Phi = 0\\
&\frac{\sin\theta}{\Theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Theta}{\partial\theta}\right) + k(k+1)\sin^2\theta - \mu = 0.
\end{aligned}$$

However, we want Φ to be a periodic in φ since we are considering functions on the sphere, so μ be must of the form $\mu = m^2$, for some non-negative integer, m. Then, we know that the space of solutions of the equation

$$\frac{\partial^2\Phi}{\partial\varphi^2} + m^2\Phi = 0$$

is two-dimensional and is spanned by the two functions

$$\Phi(\varphi) = \cos m\varphi, \qquad \Phi(\varphi) = \sin m\varphi.$$

We still have to solve the equation

$$\sin\theta \frac{\partial}{\partial\theta}\left(\sin\theta \frac{\partial\Theta}{\partial\theta}\right) + (k(k+1)\sin^2\theta - m^2)\Theta = 0,$$

which is equivalent to

$$\sin^2\theta\,\Theta'' + \sin\theta\cos\theta\,\Theta' + (k(k+1)\sin^2\theta - m^2)\Theta = 0.$$

a variant of Legendre's equation. For this, we use the change of variable, $t = \cos\theta$, and we consider the function, u, given by $u(\cos\theta) = \Theta(\theta)$ (recall that $0 \leq \theta < \pi$), so we get the second-order differential equation

$$(1-t^2)u'' - 2tu' + \left(k(k+1) - \frac{m^2}{1-t^2}\right)u = 0$$

sometimes called the *general Legendre equation* (Adrien-Marie Legendre, 1752-1833). The trick to solve this equation is to make the substitution

$$u(t) = (1-t^2)^{\frac{m}{2}} v(t),$$

see Lebedev [98], Chapter 7, Section 7.12. Then, we get

$$(1-t^2)v'' - 2(m+1)tv' + (k(k+1) - m(m+1))v = 0.$$

When $m = 0$, we get the *Legendre equation*:

$$(1-t^2)v'' - 2tv' + k(k+1)v = 0,$$

see Lebedev [98], Chapter 7, Section 7.3. This equation has two fundamental solution, $P_k(t)$ and $Q_k(t)$, called the *Legendre functions of the first and second kinds*. The $P_k(t)$ are actually polynomials and the $Q_k(t)$ are given by power series that diverge for $t = 1$, so we only keep the *Legendre polynomials*, $P_k(t)$. The Legendre polynomials can be defined in various ways. One definition is in terms of *Rodrigues' formula*:

$$P_n(t) = \frac{1}{2^n n!}\frac{d^n}{dt^n}(t^2-1)^n,$$

see Lebedev [98], Chapter 4, Section 4.2. In this version of the Legendre polynomials they are normalized so that $P_n(1) = 1$. There is also the following recurrence relation:

$$\begin{aligned} P_0 &= 1 \\ P_1 &= t \\ (n+1)P_{n+1} &= (2n+1)tP_n - nP_{n-1} \qquad n \geq 1, \end{aligned}$$

see Lebedev [98], Chapter 4, Section 4.3. For example, the first six Legendre polynomials are:

$$\begin{aligned}
&1\\
&t\\
&\frac{1}{2}(3t^2-1)\\
&\frac{1}{2}(5t^3-3t)\\
&\frac{1}{8}(35t^4-30t^2+3)\\
&\frac{1}{8}(63t^5-70t^3+15t).
\end{aligned}$$

Let us now return to our differential equation

$$(1-t^2)v''-2(m+1)tv'+(k(k+1)-m(m+1))v=0. \qquad (*)$$

Observe that if we differentiate with respect to t, we get the equation

$$(1-t^2)v'''-2(m+2)tv''+(k(k+1)-(m+1)(m+2))v'=0.$$

This shows that if v is a solution of our equation $(*)$ for given k and m, then v' is a solution of the same equation for k and $m+1$. Thus, if $P_k(t)$ solves $(*)$ for given k and $m=0$, then $P_k'(t)$ solves $(*)$ for the same k and $m=1$, $P_k''(t)$ solves $(*)$ for the same k and $m=2$, and in general, $d^m/dt^m(P_k(t))$ solves $(*)$ for k and m. Therefore, our original equation,

$$(1-t^2)u''-2tu'+\left(k(k+1)-\frac{m^2}{1-t^2}\right)u=0 \qquad (\dagger)$$

has the solution

$$u(t)=(1-t^2)^{\frac{m}{2}}\frac{d^m}{dt^m}(P_k(t)).$$

The function $u(t)$ is traditionally denoted $P_k^m(t)$ and called an *associated Legendre function*, see Lebedev [98], Chapter 7, Section 7.12. The index k is often called the *band index*. Obviously, $P_k^m(t)\equiv 0$ if $m>k$ and $P_k^0(t)=P_k(t)$, the Legendre polynomial of degree k. An associated Legendre function is not a polynomial in general and because of the factor $(1-t^2)^{\frac{m}{2}}$ it is only defined on the closed interval $[-1,1]$.

Certain authors add the factor $(-1)^m$ in front of the expression for the associated Legendre function $P_k^m(t)$, as in Lebedev [98], Chapter 7, Section 7.12, see also footnote 29 on page 193. This seems to be common practice in the quantum mechanics literature where it is called the *Condon Shortley phase factor*.

The associated Legendre functions satisfy various recurrence relations that allows us to compute them. For example, for fixed $m \geq 0$, we have (see Lebedev [98], Chapter 7, Section 7.12) the recurrence

$$(k-m+1)P_{k+1}^m(t) = (2k+1)tP_k^m(t) - (k+m)P_{k-1}^m(t), \qquad k \geq 1$$

and for fixed $k \geq 2$ we have

$$P_k^{m+2}(t) = \frac{2(m+1)t}{(t^2-1)^{\frac{1}{2}}}P_k^{m+1}(t) + (k-m)(k+m+1)P_k^m(t), \qquad 0 \leq m \leq k-2$$

which can also be used to compute P_k^m starting from

$$\begin{aligned} P_k^0(t) &= P_k(t) \\ P_k^1(t) &= \frac{kt}{(t^2-1)^{\frac{1}{2}}}P_k(t) - \frac{k}{(t^2-1)^{\frac{1}{2}}}P_{k-1}(t). \end{aligned}$$

Observe that the recurrence relation for m fixed yields the following equation for $k = m$ (as $P_{m-1}^m = 0$):

$$P_{m+1}^m(t) = (2m+1)tP_m^m(t).$$

It it also easy to see that

$$P_m^m(t) = \frac{(2m)!}{2^m m!}(1-t^2)^{\frac{m}{2}}.$$

Observe that

$$\frac{(2m)!}{2^m m!} = (2m-1)(2m-3)\cdots 5 \cdot 3 \cdot 1,$$

an expression that is sometimes denoted $(2m-1)!!$ and called the *double factorial*.

Beware that some papers in computer graphics adopt the definition of associated Legendre functions with the scale factor $(-1)^m$ added so this factor is present in these papers, for example, Green [65].

The equation above allows us to "lift" P_m^m to the higher band $m+1$. The computer graphics community (see Green [65]) uses the following three rules to compute $P_k^m(t)$ where $0 \leq m \leq k$:

(1) Compute

$$P_m^m(t) = \frac{(2m)!}{2^m m!}(1-t^2)^{\frac{m}{2}}.$$

If $m = k$, stop. Otherwise do step 2 once:

(2) Compute $P_{m+1}^m(t) = (2m+1)tP_m^m(t)$. If $k = m+1$, stop. Otherwise, iterate step 3:

(3) Starting from $i = m + 1$, compute

$$(i - m + 1)P_{i+1}^m(t) = (2i + 1)tP_i^m(t) - (i + m)P_{i-1}^m(t)$$

until $i + 1 = k$.

If we recall that equation (†) was obtained from the equation

$$\sin^2\theta\,\Theta'' + \sin\theta\cos\theta\,\Theta' + (k(k+1)\sin^2\theta - m^2)\Theta = 0$$

using the substitution $u(\cos\theta) = \Theta(\theta)$, we see that

$$\Theta(\theta) = P_k^m(\cos\theta)$$

is a solution of the above equation. Putting everything together, as $f(r,\theta,\varphi) = r^k\Theta(\theta)\Phi(\varphi)$, we proved that the homogeneous functions,

$$f(r,\theta,\varphi) = r^k\cos m\varphi\, P_k^m(\cos\theta), \qquad f(r,\theta,\varphi) = r^k\sin m\varphi\, P_k^m(\cos\theta),$$

are solutions of the Laplacian, Δ, in $\mathbb{R}^3$, and that the functions

$$\cos m\varphi\, P_k^m(\cos\theta), \qquad \sin m\varphi\, P_k^m(\cos\theta),$$

are eigenfunctions of the Laplacian, Δ_{S^2}, on the sphere for the eigenvalue $-k(k+1)$. For k fixed, as $0 \leq m \leq k$, we get $2k + 1$ linearly independent functions.

The notation for the above functions varies quite a bit essentially because of the choice of normalization factors used in various fields (such as physics, seismology, geodesy, spectral analysis, magnetics, quantum mechanics *etc.*). We will adopt the notation y_l^m, where l is a nonnegative integer but m is allowed to be negative, with $-l \leq m \leq l$. Thus, we set

$$y_l^m(\theta,\varphi) = \begin{cases} N_l^0 P_l(\cos\theta) & \text{if } m = 0 \\ \sqrt{2}N_l^m\cos m\varphi\, P_l^m(\cos\theta) & \text{if } m > 0 \\ \sqrt{2}N_l^m\sin(-m\varphi)\, P_l^{-m}(\cos\theta) & \text{if } m < 0 \end{cases}$$

for $l = 0, 1, 2, \ldots$, and where the N_l^m are scaling factors. In physics and computer graphics, N_l^m is chosen to be

$$N_l^m = \sqrt{\frac{(2l+1)(l-|m|)!}{4\pi(l+|m|)!}}.$$

The functions y_l^m are called the *real spherical harmonics of degree l and order m*. The index l is called the *band index*.

The functions, y_l^m, have some very nice properties but to explain these we need to recall the Hilbert space structure of the space, $L^2(S^2)$, of square-integrable functions on the sphere. Recall that we have an inner product on $L^2(S^2)$ given by

$$\langle f, g\rangle = \int_{S^2} fg\,\Omega_2 = \int_0^{2\pi}\int_0^{\pi} f(\theta,\varphi)g(\theta,\varphi)\sin\theta d\theta d\varphi,$$

where $f, g \in L^2(S^2)$ and where Ω_2 is the volume form on S^2 (induced by the metric on $\mathbb{R}^3$). With this inner product, $L^2(S^2)$ is a complete normed vector space using the norm, $\|f\| = \sqrt{\langle f, f\rangle}$, associated with this inner product, that is, $L^2(S^2)$ is a *Hilbert space*. Now, it can be shown that the Laplacian, Δ_{S^2}, on the sphere is a self-adjoint linear operator with respect to this inner product. As the functions, $y_{l_1}^{m_1}$ and $y_{l_2}^{m_2}$ with $l_1 \neq l_2$ are eigenfunctions corresponding to distinct eigenvalues ($-l_1(l_1+1)$ and $-l_2(l_2+1)$), they are orthogonal, that is,

$$\langle y_{l_1}^{m_1}, y_{l_2}^{m_2}\rangle = 0, \qquad \text{if} \quad l_1 \neq l_2.$$

It is also not hard to show that for a fixed l,

$$\langle y_l^{m_1}, y_l^{m_2}\rangle = \delta_{m_1,m_2},$$

that is, the functions y_l^m with $-l \leq m \leq l$ form an orthonormal system and we denote by $\mathcal{H}_l(S^2)$ the $(2l+1)$-dimensional space spanned by these functions. It turns out that the functions y_l^m form a basis of the eigenspace, E_l, of Δ_{S^2} associated with the eigenvalue $-l(l+1)$ so that $E_l = \mathcal{H}_l(S^2)$ and that Δ_{S^2} has no other eigenvalues. More is true. It turns out that $L^2(S^2)$ is the orthogonal Hilbert sum of the eigenspaces, $\mathcal{H}_l(S^2)$. This means that the $\mathcal{H}_l(S^2)$ are

(1) mutually orthogonal

(2) closed, and

(3) The space $L^2(S^2)$ is the Hilbert sum, $\bigoplus_{l=0}^{\infty} \mathcal{H}_l(S^2)$, which means that for every function, $f \in L^2(S^2)$, there is a unique sequence of spherical harmonics, $f_j \in \mathcal{H}_l(S^2)$, so that

$$f = \sum_{l=0}^{\infty} f_l,$$

that is, the sequence $\sum_{j=0}^{l} f_j$, converges to f (in the norm on $L^2(S^2)$). Observe that each f_l is a unique linear combination, $f_l = \sum_{m_l} a_{m_l\, l}\, y_l^{m_l}$.

Therefore, (3) gives us a *Fourier decomposition on the sphere* generalizing the familiar Fourier decomposition on the circle. Furthermore, the *Fourier coefficients*, $a_{m_l l}$, can be computed using the fact that the y_l^m form an orthonormal Hilbert basis:

$$a_{m_l\, l} = \langle f, y_l^{m_l}\rangle.$$

We also have the corresponding homogeneous harmonic functions, $H_l^m(r, \theta, \varphi)$, on $\mathbb{R}^3$ given by

$$H_l^m(r, \theta, \varphi) = r^l y_l^m(\theta, \varphi).$$

If one starts computing explicity the H_l^m for small values of l and m, one finds that it is always possible to express these functions in terms of the cartesian coordinates x, y, z as

homogeneous polynomials! This remarkable fact holds in general: The eigenfunctions of the Laplacian, Δ_{S^2}, and thus, the spherical harmonics, are the restrictions of homogeneous harmonic polynomials in $\mathbb{R}^3$. Here is a list of bases of the homogeneous harmonic polynomials of degree k in three variables up to $k = 4$ (thanks to Herman Gluck):

$$
\begin{array}{ll}
k = 0 & 1 \\
k = 1 & x,\ y,\ z \\
k = 2 & x^2 - y^2,\ x^2 - z^2,\ xy,\ xz,\ yz \\
k = 3 & x^3 - 3xy^2,\ 3x^2y - y^3,\ x^3 - 3xz^2,\ 3x^2z - z^3, \\
 & y^3 - 3yz^2,\ 3y^2z - z^3,\ xyz \\
k = 4 & x^4 - 6x^2y^2 + y^4,\ x^4 - 6x^2z^2 + z^4,\ y^4 - 6y^2z^2 + z^4, \\
 & x^3y - xy^3,\ x^3z - xz^3,\ y^3z - yz^3, \\
 & 3x^2yz - yz^3,\ 3xy^2z - xz^3,\ 3xyz^2 - x^3y.
\end{array}
$$

Subsequent sections will be devoted to a proof of the important facts stated earlier.

16.3 The Laplace-Beltrami Operator

In order to define rigorously the Laplacian on the sphere, $S^n \subseteq \mathbb{R}^{n+1}$, and establish its relationship with the Laplacian on $\mathbb{R}^{n+1}$, we need the definition of the Laplacian on a Riemannian manifold, (M, g), the *Laplace-Beltrami operator*, as defined in Section 15.2 (Eugenio Beltrami, 1835-1900). In that section, the Laplace-Beltrami operator is defined as an operator on differential forms but a more direct definition can be given for the Laplacian-Beltrami operator on functions (using the covariant derivative, see the paragraph preceding Proposition 15.6). For the benefit of the reader who may not have read Section 15.2, we present this definition of the divergence again.

Recall that a Riemannian metric, g, on a manifold, M, is a smooth family of inner products, $g = (g_p)$, where g_p is an inner product on the tangent space, T_pM, for every $p \in M$. The inner product, g_p, on T_pM, establishes a canonical duality between T_pM and T_p^*M, namely, we have the isomorphism, $\flat\colon T_pM \to T_p^*M$, defined such that for every $u \in T_pM$, the linear form, $u^\flat \in T_p^*M$, is given by

$$u^\flat(v) = g_p(u, v), \qquad v \in T_pM.$$

The inverse isomorphism, $\sharp\colon T_p^*M \to T_pM$, is defined such that for every $\omega \in T_p^*M$, the vector, $\omega^\sharp$, is the unique vector in T_pM so that

$$g_p(\omega^\sharp, v) = \omega(v), \qquad v \in T_pM.$$

The isomorphisms $\flat$ and $\sharp$ induce isomorphisms between vector fields, $X \in \mathfrak{X}(M)$, and one-forms, $\omega \in \mathcal{A}^1(M)$. In particular, for every smooth function, $f \in C^\infty(M)$, the vector field

corresponding to the one-form, df, is the *gradient*, $\operatorname{grad} f$, of f. The gradient of f is uniquely determined by the condition

$$g_p((\operatorname{grad} f)_p, v) = df_p(v), \qquad v \in T_pM,\ p \in M.$$

If ∇_X is the covariant derivative associated with the Levi-Civita connection induced by the metric, g, then the *divergence* of a vector field, $X \in \mathfrak{X}(M)$, is the function, $\operatorname{div} X \colon M \to \mathbb{R}$, defined so that

$$(\operatorname{div} X)(p) = \operatorname{tr}(Y(p) \mapsto (\nabla_Y X)_p),$$

namely, for every p, $(\operatorname{div} X)(p)$ is the trace of the linear map, $Y(p) \mapsto (\nabla_Y X)_p$. Then, the *Laplace-Beltrami operator*, for short, *Laplacian*, is the linear operator,
$\Delta \colon C^\infty(M) \to C^\infty(M)$, given by

$$\Delta f = \operatorname{div} \operatorname{grad} f.$$

Observe that the definition just given differs from the definition given in Section 15.2 by a negative sign. We adopted this sign convention to conform with most of the literature on spherical harmonics (where the negative sign is omitted). A consequence of this choice is that the eigenvalues of the Laplacian are negative.

For more details on the Laplace-Beltrami operator, we refer the reader to Chapter 15 or to Gallot, Hulin and Lafontaine [61] (Chapter 4) or O'Neill [120] (Chapter 3), Postnikov [126] (Chapter 13), Helgason [73] (Chapter 2) or Warner [148] (Chapters 4 and 6).

All this being rather abstact, it is useful to know how $\operatorname{grad} f$, $\operatorname{div} X$ and Δf are expressed in a chart. If (U, φ) is a chart of M, with $p \in M$ and if, as usual,

$$\left(\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p\right)$$

denotes the basis of T_pM induced by φ, the local expression of the metric g at p is given by the $n \times n$ matrix, $(g_{ij})_p$, with

$$(g_{ij})_p = g_p\left(\left(\frac{\partial}{\partial x_i}\right)_p, \left(\frac{\partial}{\partial x_j}\right)_p\right).$$

The matrix $(g_{ij})_p$ is symmetric, positive definite and its inverse is denoted $(g^{ij})_p$. We also let $|g|_p = \det(g_{ij})_p$. For simplicity of notation we often omit the subscript p. Then, it can be shown that for every function, $f \in C^\infty(M)$, in local coordinates given by the chart (U, φ), we have

$$\operatorname{grad} f = \sum_{ij} g^{ij} \frac{\partial f}{\partial x_j} \frac{\partial}{\partial x_i},$$

where, as usual

$$\frac{\partial f}{\partial x_j}(p) = \left(\frac{\partial}{\partial x_j}\right)_p f = \frac{\partial (f \circ \varphi^{-1})}{\partial u_j}(\varphi(p))$$

and $(u_1, \ldots, u_n)$ are the coordinate functions in $\mathbb{R}^n$. There are formulae for $\operatorname{div} X$ and Δf involving the Christoffel symbols but the following formulae will be more convenient for our purposes: For every vector field, $X \in \mathfrak{X}(M)$, expressed in local coordinates as

$$X = \sum_{i=1}^{n} X_i \frac{\partial}{\partial x_i}$$

we have

$$\operatorname{div} X = \frac{1}{\sqrt{|g|}} \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left(\sqrt{|g|}\, X_i \right)$$

and for every function, $f \in C^\infty(M)$, the Laplacian, Δf, is given by

$$\Delta f = \frac{1}{\sqrt{|g|}} \sum_{i,j} \frac{\partial}{\partial x_i} \left(\sqrt{|g|}\, g^{ij} \frac{\partial f}{\partial x_j} \right).$$

The above formula is proved in Proposition 15.5, assuming M is orientable. A different derivation is given in Postnikov [126] (Chapter 13, Section 5).

One should check that for $M = \mathbb{R}^n$ with its standard coordinates, the Laplacian is given by the familiar formula

$$\Delta f = \frac{\partial^2 f}{\partial x_1^2} + \cdots + \frac{\partial^2 f}{\partial x_n^2}.$$

Remark: A different sign convention is also used in defining the divergence, namely,

$$\operatorname{div} X = -\frac{1}{\sqrt{|g|}} \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left(\sqrt{|g|}\, X_i \right).$$

With this convention, which is the one used in Section 15.2, the Laplacian also has a negative sign. This has the advantage that the eigenvalues of the Laplacian are nonnegative.

As an application, let us derive the formula for the Laplacian in spherical coordinates,

$$\begin{aligned} x &= r \sin\theta \cos\varphi \\ y &= r \sin\theta \sin\varphi \\ z &= r \cos\theta. \end{aligned}$$

We have

$$\begin{aligned} \frac{\partial}{\partial r} &= \sin\theta\cos\varphi \frac{\partial}{\partial x} + \sin\theta\sin\varphi \frac{\partial}{\partial y} + \cos\theta \frac{\partial}{\partial z} = \widehat{r} \\ \frac{\partial}{\partial \theta} &= r\left(\cos\theta\cos\varphi \frac{\partial}{\partial x} + \cos\theta\sin\varphi \frac{\partial}{\partial y} - \sin\theta \frac{\partial}{\partial z} \right) = r\widehat{\theta} \\ \frac{\partial}{\partial \varphi} &= r\left(-\sin\theta\sin\varphi \frac{\partial}{\partial x} + \sin\theta\cos\varphi \frac{\partial}{\partial y} \right) = r\widehat{\varphi}. \end{aligned}$$

Observe that $\widehat{r}$, $\widehat{\theta}$ and $\widehat{\varphi}$ are pairwise orthogonal. Therefore, the matrix (g_{ij}) is given by

$$(g_{ij}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2\theta \end{pmatrix}$$

and $|g| = r^4 \sin^2\theta$. The inverse of (g_{ij}) is

$$(g^{ij}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^{-2} & 0 \\ 0 & 0 & r^{-2} \sin^{-2}\theta \end{pmatrix}.$$

We will let the reader finish the computation to verify that we get

$$\Delta f = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial f}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial f}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2 f}{\partial \varphi^2}.$$

Since (θ, φ) are coordinates on the sphere S^2 *via*

$$\begin{aligned} x &= \sin\theta\cos\varphi \\ y &= \sin\theta\sin\varphi \\ z &= \cos\theta, \end{aligned}$$

we see that in these coordinates, the metric, $(\widetilde{g}_{ij})$, on S^2 is given by the matrix

$$(\widetilde{g}_{ij}) = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2\theta \end{pmatrix},$$

that $|\widetilde{g}| = \sin^2\theta$, and that the inverse of $(\widetilde{g}_{ij})$ is

$$(\widetilde{g}^{ij}) = \begin{pmatrix} 1 & 0 \\ 0 & \sin^{-2}\theta \end{pmatrix}.$$

It follows immediately that

$$\Delta_{S^2} f = \frac{1}{\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial f}{\partial \theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2 f}{\partial \varphi^2},$$

so we have verified that

$$\Delta f = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial f}{\partial r}\right) + \frac{1}{r^2}\Delta_{S^2} f.$$

Let us now generalize the above formula to the Laplacian, Δ, on $\mathbb{R}^{n+1}$ and the Laplacian, Δ_{S^n}, on S^n, where

$$S^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1^2 + \cdots + x_{n+1}^2 = 1\}.$$

Following Morimoto [114] (Chapter 2, Section 2), let us use "polar coordinates". The map from $\mathbb{R}_+ \times S^n$ to $\mathbb{R}^{n+1} - \{0\}$ given by

$$(r, \sigma) \mapsto r\sigma$$

is clearly a diffeomorphism. Thus, for any system of coordinates, $(u_1, \ldots, u_n)$, on S^n, the tuple $(u_1, \ldots, u_n, r)$ is a system of coordinates on $\mathbb{R}^{n+1} - \{0\}$ called *polar coordinates*. Let us establish the relationship between the Laplacian, Δ, on $\mathbb{R}^{n+1} - \{0\}$ in polar coordinates and the Laplacian, Δ_{S^n}, on S^n in local coordinates $(u_1, \ldots, u_n)$.

Proposition 16.1. *If Δ is the Laplacian on $\mathbb{R}^{n+1} - \{0\}$ in polar coordinates $(u_1, \ldots, u_n, r)$ and Δ_{S^n} is the Laplacian on the sphere, S^n, in local coordinates $(u_1, \ldots, u_n)$, then*

$$\Delta f = \frac{1}{r^n}\frac{\partial}{\partial r}\left(r^n \frac{\partial f}{\partial r}\right) + \frac{1}{r^2}\Delta_{S^n} f.$$

Proof. Let us compute the $(n+1) \times (n+1)$ matrix, $G = (g_{ij})$, expressing the metric on $\mathbb{R}^{n+1}$ is polar coordinates and the $n \times n$ matrix, $\widetilde{G} = (\widetilde{g}_{ij})$, expressing the metric on S^n. Recall that if $\sigma \in S^n$, then $\sigma \cdot \sigma = 1$ and so,

$$\frac{\partial \sigma}{\partial u_i} \cdot \sigma = 0,$$

as

$$\frac{\partial \sigma}{\partial u_i} \cdot \sigma = \frac{1}{2}\frac{\partial(\sigma \cdot \sigma)}{\partial u_i} = 0.$$

If $x = r\sigma$ with $\sigma \in S^n$, we have

$$\frac{\partial x}{\partial u_i} = r\frac{\partial \sigma}{\partial u_i}, \qquad 1 \leq i \leq n,$$

and

$$\frac{\partial x}{\partial r} = \sigma.$$

It follows that

$$\begin{aligned} g_{ij} &= \frac{\partial x}{\partial u_i} \cdot \frac{\partial x}{\partial u_j} = r^2 \frac{\partial \sigma}{\partial u_i} \cdot \frac{\partial \sigma}{\partial u_j} = r^2 \widetilde{g}_{ij} \\ g_{in+1} &= \frac{\partial x}{\partial u_i} \cdot \frac{\partial x}{\partial r} = r\frac{\partial \sigma}{\partial u_i} \cdot \sigma = 0 \\ g_{n+1n+1} &= \frac{\partial x}{\partial r} \cdot \frac{\partial x}{\partial r} = \sigma \cdot \sigma = 1. \end{aligned}$$

Consequently, we get

$$G = \begin{pmatrix} r^2\widetilde{G} & 0 \\ 0 & 1 \end{pmatrix},$$

$|g| = r^{2n}|\widetilde{g}|$ and

$$G^{-1} = \begin{pmatrix} r^{-2}\widetilde{G}^{-1} & 0 \\ 0 & 1 \end{pmatrix}.$$

Using the above equations and

$$\Delta f = \frac{1}{\sqrt{|g|}} \sum_{i,j} \frac{\partial}{\partial x_i} \left(\sqrt{|g|}\, g^{ij} \frac{\partial f}{\partial x_j} \right),$$

we get

$$\begin{aligned} \Delta f &= \frac{1}{r^n \sqrt{|\widetilde{g}|}} \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} \left(r^n \sqrt{|\widetilde{g}|}\, \frac{1}{r^2} \widetilde{g}^{ij} \frac{\partial f}{\partial x_j} \right) + \frac{1}{r^n \sqrt{|\widetilde{g}|}} \frac{\partial}{\partial r} \left(r^n \sqrt{|\widetilde{g}|}\, \frac{\partial f}{\partial r} \right) \\ &= \frac{1}{r^2 \sqrt{|\widetilde{g}|}} \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} \left(\sqrt{|\widetilde{g}|}\, \widetilde{g}^{ij} \frac{\partial f}{\partial x_j} \right) + \frac{1}{r^n} \frac{\partial}{\partial r} \left(r^n \frac{\partial f}{\partial r} \right) \\ &= \frac{1}{r^2} \Delta_{S^n} f + \frac{1}{r^n} \frac{\partial}{\partial r} \left(r^n \frac{\partial f}{\partial r} \right), \end{aligned}$$

as claimed. □

It is also possible to express Δ_{S^n} in terms of $\Delta_{S^{n-1}}$. If $e_{n+1} = (0, \ldots, 0, 1) \in \mathbb{R}^{n+1}$, then we can view S^{n-1} as the intersection of S^n with the hyperplane, $x_{n+1} = 0$, that is, as the set

$$S^{n-1} = \{\sigma \in S^n \mid \sigma \cdot e_{n+1} = 0\}.$$

If $(u_1, \ldots, u_{n-1})$ are local coordinates on S^{n-1}, then $(u_1, \ldots, u_{n-1}, \theta)$ are local coordinates on S^n, by setting

$$\sigma = \sin\theta\, \widetilde{\sigma} + \cos\theta\, e_{n+1},$$

with $\widetilde{\sigma} \in S^{n-1}$ and $0 \leq \theta < \pi$. Using these local coordinate systems, it is a good exercise to find the relationship between Δ_{S^n} and $\Delta_{S^{n-1}}$, namely

$$\Delta_{S^n} f = \frac{1}{\sin^{n-1}\theta} \frac{\partial}{\partial\theta} \left(\sin^{n-1}\theta \frac{\partial f}{\partial\theta} \right) + \frac{1}{\sin^2\theta} \Delta_{S_{n-1}} f.$$

A fundamental property of the divergence is known as *Green's Formula*. There are actually two Greens' Formulae but we will only need the version for an orientable manifold without boundary given in Proposition 15.7. Recall that Green's Formula states that if M is a compact, orientable, Riemannian manifold without boundary, then, for every smooth vector field, $X \in \mathfrak{X}(M)$, we have

$$\int_M (\operatorname{div} X)\, \Omega_M = 0,$$

where Ω_M is the volume form on M induced by the metric.

If M is a compact, orientable Riemannian manifold, then for any two smooth functions, $f, h \in C^\infty(M)$, we define $\langle f, h\rangle$ by

$$\langle f, h\rangle = \int_M fh\,\Omega_M.$$

Then, it is not hard to show that $\langle -, -\rangle$ is an inner product on $\mathcal{C}^\infty(M)$.

An important property of the Laplacian on a compact, orientable Riemannian manifold is that it is a self-adjoint operator. This fact has already been proved in the more general case of an inner product on differential forms in Proposition 15.3 but it might be instructive to give another proof in the special case of functions using Green's Formula.

For this, we prove the following properties: For any two functions, $f, h \in C^\infty(M)$, and any vector field, $X \in \mathcal{C}^\infty(M)$, we have:

$$\begin{aligned} \operatorname{div}(fX) &= f\operatorname{div} X + X(f) = f\operatorname{div} X + g(\operatorname{grad} f, X) \\ \operatorname{grad} f\,(h) &= g(\operatorname{grad} f, \operatorname{grad} h) = \operatorname{grad} h\,(f). \end{aligned}$$

Using these identities, we obtain the following important special case of Proposition 15.3:

Proposition 16.2. *Let M be a compact, orientable, Riemannian manifold without boundary. The Laplacian on M is self-adjoint, that is, for any two functions, $f, h \in C^\infty(M)$, we have*

$$\langle \Delta f, h\rangle = \langle f, \Delta h\rangle$$

or equivalently

$$\int_M f\Delta h\,\Omega_M = \int_M h\Delta f\,\Omega_M.$$

Proof. By the two identities before Proposition 16.2,

$$f\Delta h = f\operatorname{div}\operatorname{grad} h = \operatorname{div}(f\operatorname{grad} h) - g(\operatorname{grad} f, \operatorname{grad} h)$$

and

$$h\Delta f = h\operatorname{div}\operatorname{grad} f = \operatorname{div}(h\operatorname{grad} f) - g(\operatorname{grad} h, \operatorname{grad} f),$$

so we get

$$f\Delta h - h\Delta f = \operatorname{div}(f\operatorname{grad} h - h\operatorname{grad} f).$$

By Green's Formula,

$$\int_M (f\Delta h - h\Delta f)\Omega_M = \int_M \operatorname{div}(f\operatorname{grad} h - h\operatorname{grad} f)\Omega_M = 0,$$

which proves that Δ is self-adjoint. □

The importance of Proposition 16.2 lies in the fact that as $\langle -, -\rangle$ is an inner product on $\mathcal{C}^\infty(M)$, the eigenspaces of Δ for distinct eigenvalues are pairwise orthogonal. We will make heavy use of this property in the next section on harmonic polynomials.

16.4 Harmonic Polynomials, Spherical Harmonics and $L^2(S^n)$

Harmonic homogeneous polynomials and their restrictions to S^n, where

$$S^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1^2 + \cdots + x_{n+1}^2 = 1\},$$

turn out to play a crucial role in understanding the structure of the eigenspaces of the Laplacian on S^n (with $n \geq 1$). The results in this section appear in one form or another in Stein and Weiss [143] (Chapter 4), Morimoto [114] (Chapter 2), Helgason [73] (Introduction, Section 3), Dieudonné [44] (Chapter 7), Axler, Bourdon and Ramey [12] (Chapter 5) and Vilenkin [147] (Chapter IX). Some of these sources assume a fair amount of mathematical background and consequently, uninitiated readers will probably find the exposition rather condensed, especially Helgason. We tried hard to make our presentation more "user-friendly".

Definition 16.1. Let $\mathcal{P}_k(n+1)$ (resp. $\mathcal{P}_k^{\mathbb{C}}(n+1)$) denote the space of homogeneous polynomials of degree k in $n+1$ variables with real coefficients (resp. complex coefficients) and let $\mathcal{P}_k(S^n)$ (resp. $\mathcal{P}_k^{\mathbb{C}}(S^n)$) denote the restrictions of homogeneous polynomials in $\mathcal{P}_k(n+1)$ to S^n (resp. the restrictions of homogeneous polynomials in $\mathcal{P}_k^{\mathbb{C}}(n+1)$ to S^n). Let $\mathcal{H}_k(n+1)$ (resp. $\mathcal{H}_k^{\mathbb{C}}(n+1)$) denote the space of *(real) harmonic polynomials* (resp. *complex harmonic polynomials*), with

$$\mathcal{H}_k(n+1) = \{P \in \mathcal{P}_k(n+1) \mid \Delta P = 0\}$$

and

$$\mathcal{H}_k^{\mathbb{C}}(n+1) = \{P \in \mathcal{P}_k^{\mathbb{C}}(n+1) \mid \Delta P = 0\}.$$

Harmonic polynomials are sometimes called *solid harmonics*. Finally, Let $\mathcal{H}_k(S^n)$ (resp. $\mathcal{H}_k^{\mathbb{C}}(S^n)$) denote the space of *(real) spherical harmonics* (resp. *complex spherical harmonics*) be the set of restrictions of harmonic polynomials in $\mathcal{H}_k(n+1)$ to S^n (resp. restrictions of harmonic polynomials in $\mathcal{H}_k^{\mathbb{C}}(n+1)$ to S^n).

A function, $f\colon \mathbb{R}^n \to \mathbb{R}$ (resp. $f\colon \mathbb{R}^n \to \mathbb{C}$), is *homogeneous of degree k* iff

$$f(tx) = t^k f(x), \qquad \text{for all } x \in \mathbb{R}^n \text{ and } t > 0.$$

The restriction map, $\rho\colon \mathcal{H}_k(n+1) \to \mathcal{H}_k(S^n)$, is a surjective linear map. In fact, it is a bijection. Indeed, if $P \in \mathcal{H}_k(n+1)$, observe that

$$P(x) = \|x\|^k P\left(\frac{x}{\|x\|}\right), \qquad \text{with} \quad \frac{x}{\|x\|} \in S^n,$$

for all $x \neq 0$. Consequently, if $P \restriction S^n = Q \restriction S^n$, that is, $P(\sigma) = Q(\sigma)$ for all $\sigma \in S^n$, then

$$P(x) = \|x\|^k P\left(\frac{x}{\|x\|}\right) = \|x\|^k Q\left(\frac{x}{\|x\|}\right) = Q(x)$$

for all $x \neq 0$, which implies $P = Q$ (as P and Q are polynomials). Therefore, we have a linear isomorphism between $\mathcal{H}_k(n+1)$ and $\mathcal{H}_k(S^n)$ (and between $\mathcal{H}_k^{\mathbb{C}}(n+1)$ and $\mathcal{H}_k^{\mathbb{C}}(S^n)$).

It will be convenient to introduce some notation to deal with homogeneous polynomials. Given $n \geq 1$ variables, $x_1, \ldots, x_n$, and any n-tuple of nonnegative integers, $\alpha = (\alpha_1, \ldots, \alpha_n)$, let $|\alpha| = \alpha_1 + \cdots + \alpha_n$, let $x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ and let $\alpha! = \alpha_1! \cdots \alpha_n!$. Then, every homogeneous polynomial, P, of degree k in the variables $x_1, \ldots, x_n$ can be written uniquely as

$$P = \sum_{|\alpha|=k} c_\alpha x^\alpha,$$

with $c_\alpha \in \mathbb{R}$ or $c_\alpha \in \mathbb{C}$. It is well known that $\mathcal{P}_k(n)$ is a (real) vector space of dimension

$$d_k = \binom{n+k-1}{k}$$

and $\mathcal{P}_k^{\mathbb{C}}(n)$ is a complex vector space of the same dimension, d_k.

We can define an Hermitian inner product on $\mathcal{P}_k^{\mathbb{C}}(n)$ whose restriction to $\mathcal{P}_k(n)$ is an inner product by viewing a homogeneous polynomial as a differential operator as follows: For every $P = \sum_{|\alpha|=k} c_\alpha x^\alpha \in \mathcal{P}_k^{\mathbb{C}}(n)$, let

$$\partial(P) = \sum_{|\alpha|=k} c_\alpha \frac{\partial^k}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}.$$

Then, for any two polynomials, $P, Q \in \mathcal{P}_k^{\mathbb{C}}(n)$, let

$$\langle P, Q \rangle = \partial(P)\overline{Q}.$$

A simple computation shows that

$$\left\langle \sum_{|\alpha|=k} a_\alpha x^\alpha, \sum_{|\alpha|=k} b_\alpha x^\alpha \right\rangle = \sum_{|\alpha|=k} \alpha!\, a_\alpha \overline{b}_\alpha.$$

Therefore, $\langle P, Q \rangle$ is indeed an inner product. Also observe that

$$\partial(x_1^2 + \cdots + x_n^2) = \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2} = \Delta.$$

Another useful property of our inner product is this:

$$\langle P, QR \rangle = \langle \partial(Q)P, R \rangle.$$

Indeed.

$$\begin{aligned}\langle P, QR\rangle &= \langle QR, P\rangle \\ &= \partial(QR)\overline{P} \\ &= \partial(Q)(\partial(R)\overline{P}) \\ &= \partial(R)(\partial(Q)\overline{P}) \\ &= \langle R, \partial(Q)P\rangle \\ &= \langle \partial(Q)P, R\rangle.\end{aligned}$$

In particular,

$$\langle (x_1^2 + \cdots + x_n^2)P, Q\rangle = \langle P, \partial(x_1^2 + \cdots + x_n^2)Q\rangle = \langle P, \Delta Q\rangle.$$

Let us write $\|x\|^2$ for $x_1^2 + \cdots + x_n^2$. Using our inner product, we can prove the following important theorem:

Theorem 16.3. *The map, $\Delta\colon \mathcal{P}_k(n) \to \mathcal{P}_{k-2}(n)$, is surjective for all $n, k \geq 2$ (and similarly for $\Delta\colon \mathcal{P}_k^{\mathbb{C}}(n) \to \mathcal{P}_{k-2}^{\mathbb{C}}(n)$). Furthermore, we have the following orthogonal direct sum decompositions:*

$$\mathcal{P}_k(n) = \mathcal{H}_k(n) \oplus \|x\|^2 \mathcal{H}_{k-2}(n) \oplus \cdots \oplus \|x\|^{2j} \mathcal{H}_{k-2j}(n) \oplus \cdots \oplus \|x\|^{2[k/2]} \mathcal{H}_{[k/2]}(n)$$

and

$$\mathcal{P}_k^{\mathbb{C}}(n) = \mathcal{H}_k^{\mathbb{C}}(n) \oplus \|x\|^2 \mathcal{H}_{k-2}^{\mathbb{C}}(n) \oplus \cdots \oplus \|x\|^{2j} \mathcal{H}_{k-2j}^{\mathbb{C}}(n) \oplus \cdots \oplus \|x\|^{2[k/2]} \mathcal{H}_{[k/2]}^{\mathbb{C}}(n),$$

with the understanding that only the first term occurs on the right-hand side when $k < 2$.

Proof. If the map $\Delta\colon \mathcal{P}_k^{\mathbb{C}}(n) \to \mathcal{P}_{k-2}^{\mathbb{C}}(n)$ is not surjective, then some nonzero polynomial, $Q \in \mathcal{P}_{k-2}^{\mathbb{C}}(n)$, is orthogonal to the image of Δ. In particular, Q must be orthogonal to ΔP with $P = \|x\|^2 Q \in \mathcal{P}_k^{\mathbb{C}}(n)$. So, using a fact established earlier,

$$0 = \langle Q, \Delta P\rangle = \langle \|x\|^2 Q, P\rangle = \langle P, P\rangle,$$

which implies that $P = \|x\|^2 Q = 0$ and thus, $Q = 0$, a contradiction. The same proof is valid in the real case.

We claim that we have an orthogonal direct sum decomposition,

$$\mathcal{P}_k^{\mathbb{C}}(n) = \mathcal{H}_k^{\mathbb{C}}(n) \oplus \|x\|^2 \mathcal{P}_{k-2}^{\mathbb{C}}(n),$$

and similarly in the real case, with the understanding that the second term is missing if $k < 2$. If $k = 0, 1$, then $\mathcal{P}_k^{\mathbb{C}}(n) = \mathcal{H}_k^{\mathbb{C}}(n)$ so this case is trivial. Assume $k \geq 2$. Since $\operatorname{Ker}\Delta = \mathcal{H}_k^{\mathbb{C}}(n)$ and Δ is surjective, $\dim(\mathcal{P}_k^{\mathbb{C}}(n)) = \dim(\mathcal{H}_k^{\mathbb{C}}(n)) + \dim(\mathcal{P}_{k-2}^{\mathbb{C}}(n))$, so it is

sufficient to prove that $\mathcal{H}_k^{\mathbb{C}}(n)$ is orthogonal to $\|x\|^2 \mathcal{P}_{k-2}^{\mathbb{C}}(n)$. Now, if $H \in \mathcal{H}_k^{\mathbb{C}}(n)$ and $P = \|x\|^2 Q \in \|x\|^2 \mathcal{P}_{k-2}^{\mathbb{C}}(n)$, we have

$$\langle \|x\|^2 Q, H\rangle = \langle Q, \Delta H\rangle = 0,$$

so $\mathcal{H}_k^{\mathbb{C}}(n)$ and $\|x\|^2 \mathcal{P}_{k-2}^{\mathbb{C}}(n)$ are indeed orthogonal. Using induction, we immediately get the orthogonal direct sum decomposition

$$\mathcal{P}_k^{\mathbb{C}}(n) = \mathcal{H}_k^{\mathbb{C}}(n) \oplus \|x\|^2 \mathcal{H}_{k-2}^{\mathbb{C}}(n) \oplus \cdots \oplus \|x\|^{2j} \mathcal{H}_{k-2j}^{\mathbb{C}}(n) \oplus \cdots \oplus \|x\|^{2[k/2]} \mathcal{H}_{[k/2]}^{\mathbb{C}}(n)$$

and the corresponding real version. □

Remark: Theorem 16.3 also holds for $n = 1$.

Theorem 16.3 has some important corollaries. Since every polynomial in $n+1$ variables is the sum of homogeneous polynomials, we get:

Corollary 16.4. *The restriction to S^n of every polynomial (resp. complex polynomial) in $n+1 \geq 2$ variables is a sum of restrictions to S^n of harmonic polynomials (resp. complex harmonic polynomials).*

We can also derive a formula for the dimension of $\mathcal{H}_k(n)$ (and $\mathcal{H}_k^{\mathbb{C}}(n)$).

Corollary 16.5. *The dimension, $a_{k,n}$, of the space of harmonic polynomials, $\mathcal{H}_k(n)$, is given by the formula*

$$a_{k,n} = \binom{n+k-1}{k} - \binom{n+k-3}{k-2}$$

if $n, k \geq 2$, with $a_{0,n} = 1$ and $a_{1,n} = n$, and similarly for $\mathcal{H}_k^{\mathbb{C}}(n)$. As $\mathcal{H}_k(n+1)$ is isomorphic to $\mathcal{H}_k(S^n)$ (and $\mathcal{H}_k^{\mathbb{C}}(n+1)$ is isomorphic to $\mathcal{H}_k^{\mathbb{C}}(S^n)$) we have

$$\dim(\mathcal{H}_k^{\mathbb{C}}(S^n)) = \dim(\mathcal{H}_k(S^n)) = a_{k,n+1} = \binom{n+k}{k} - \binom{n+k-2}{k-2}.$$

Proof. The cases $k = 0$ and $k = 1$ are trivial since in this case $\mathcal{H}_k(n) = \mathcal{P}_k(n)$. For $k \geq 2$, the result follows from the direct sum decomposition

$$\mathcal{P}_k(n) = \mathcal{H}_k(n) \oplus \|x\|^2 \mathcal{P}_{k-2}(n)$$

proved earlier. The proof is identical in the complex case. □

Observe that when $n = 2$, we get $a_{k,2} = 2$ for $k \geq 1$ and when $n = 3$, we get $a_{k,3} = 2k+1$ for all $k \geq 0$, which we already knew from Section 16.2. The formula even applies for $n = 1$ and yields $a_{k,1} = 0$ for $k \geq 2$.

Remark: It is easy to show that

$$a_{k,n+1} = \binom{n+k-1}{n-1} + \binom{n+k-2}{n-1}$$

for $k \geq 2$, see Morimoto [114] (Chapter 2, Theorem 2.4) or Dieudonné [44] (Chapter 7, formula 99), where a different proof technique is used.

Let $L^2(S^n)$ be the space of (real) square-integrable functions on the sphere, S^n. We have an inner product on $L^2(S^n)$ given by

$$\langle f, g\rangle = \int_{S^n} f g \, \Omega_n,$$

where $f, g \in L^2(S^n)$ and where Ω_n is the volume form on S^n (induced by the metric on $\mathbb{R}^{n+1}$). With this inner product, $L^2(S^n)$ is a complete normed vector space using the norm, $\|f\| = \|f\|_2 = \sqrt{\langle f, f\rangle}$, associated with this inner product, that is, $L^2(S^n)$ is a *Hilbert space*. In the case of complex-valued functions, we use the Hermitian inner product

$$\langle f, g\rangle = \int_{S^n} f \, \overline{g} \, \Omega_n$$

and we get the complex Hilbert space, $L^2_{\mathbb{C}}(S^n)$. We also denote by $C(S^n)$ the space of continuous (real) functions on S^n with the L^∞ norm, that is,

$$\|f\|_\infty = \sup\{|f(x)|\}_{x \in S^n}$$

and by $C_{\mathbb{C}}(S^n)$ the space of continuous complex-valued functions on S^n also with the L^∞ norm. Recall that $C(S^n)$ is dense in $L^2(S^n)$ (and $C_{\mathbb{C}}(S^n)$ is dense in $L^2_{\mathbb{C}}(S^n)$). The following proposition shows why the spherical harmonics play an important role:

Proposition 16.6. *The set of all finite linear combinations of elements in $\bigcup_{k=0}^\infty \mathcal{H}_k(S^n)$ (resp. $\bigcup_{k=0}^\infty \mathcal{H}^{\mathbb{C}}_k(S^n)$) is*

(i) dense in $C(S^n)$ (resp. in $C_{\mathbb{C}}(S^n)$) with respect to the L^∞-norm;

(ii) dense in $L^2(S^n)$ (resp. dense in $L^2_{\mathbb{C}}(S^n)$).

Proof. (i) As S^n is compact, by the Stone-Weierstrass approximation theorem (Lang [94], Chapter III, Corollary 1.3), if g is continuous on S^n, then it can be approximated uniformly by polynomials, P_j, restricted to S^n. By Corollary 16.4, the restriction of each P_j to S^n is a linear combination of elements in $\bigcup_{k=0}^\infty \mathcal{H}_k(S^n)$.

(ii) We use the fact that $C(S^n)$ is dense in $L^2(S^n)$. Given $f \in L^2(S^n)$, for every $\epsilon > 0$, we can choose a continuous function, g, so that $\|f - g\|_2 < \epsilon/2$. By (i), we can find a linear combination, h, of elements in $\bigcup_{k=0}^\infty \mathcal{H}_k(S^n)$ so that $\|g - h\|_\infty < \epsilon/(2\sqrt{\mathrm{vol}(S^n)})$, where $\mathrm{vol}(S^n)$ is the volume of S^n (really, area). Thus, we get

$$\|f - h\|_2 \leq \|f - g\|_2 + \|g - h\|_2 < \epsilon/2 + \sqrt{\mathrm{vol}(S^n)}\,\|g - h\|_\infty < \epsilon/2 + \epsilon/2 = \epsilon,$$

which proves (ii). The proof in the complex case is identical. □

We need one more proposition before showing that the spaces $\mathcal{H}_k(S^n)$ constitute an orthogonal Hilbert space decomposition of $L^2(S^n)$.

Proposition 16.7. *For every harmonic polynomial, $P \in \mathcal{H}_k(n+1)$ (resp. $P \in \mathcal{H}_k^{\mathbb{C}}(n+1)$), the restriction, $H \in \mathcal{H}_k(S^n)$ (resp. $H \in \mathcal{H}_k^{\mathbb{C}}(S^n)$), of P to S^n is an eigenfunction of Δ_{S^n} for the eigenvalue $-k(n+k-1)$.*

Proof. We have

$$P(r\sigma) = r^k H(\sigma), \qquad r > 0,\ \sigma \in S^n,$$

and by Proposition 16.1, for any $f \in C^\infty(\mathbb{R}^{n+1})$, we have

$$\Delta f = \frac{1}{r^n}\frac{\partial}{\partial r}\left(r^n \frac{\partial f}{\partial r}\right) + \frac{1}{r^2}\Delta_{S^n} f.$$

Consequently,

$$\begin{aligned}
\Delta P = \Delta(r^k H) &= \frac{1}{r^n}\frac{\partial}{\partial r}\left(r^n \frac{\partial (r^k H)}{\partial r}\right) + \frac{1}{r^2}\Delta_{S^n}(r^k H) \\
&= \frac{1}{r^n}\frac{\partial}{\partial r}\left(k r^{n+k-1} H\right) + r^{k-2}\Delta_{S^n} H \\
&= \frac{1}{r^n} k(n+k-1) r^{n+k-2} H + r^{k-2}\Delta_{S^n} H \\
&= r^{k-2}(k(n+k-1)H + \Delta_{S^n} H).
\end{aligned}$$

Thus,

$$\Delta P = 0 \quad \text{iff} \quad \Delta_{S^n} H = -k(n+k-1)H,$$

as claimed. □

From Proposition 16.7, we deduce that the space $\mathcal{H}_k(S^n)$ is a subspace of the eigenspace, E_k, of Δ_{S^n}, associated with the eigenvalue $-k(n+k-1)$ (and similarly for $\mathcal{H}_k^{\mathbb{C}}(S^n)$). Remarkably, $E_k = \mathcal{H}_k(S^n)$ but it will take more work to prove this.

What we can deduce immediately is that $\mathcal{H}_k(S^n)$ *and* $\mathcal{H}_l(S^n)$ *are pairwise orthogonal whenever* $k \neq l$. This is because, by Proposition 16.2, the Laplacian is self-adjoint and thus, any two eigenspaces, E_k and E_l are pairwise orthogonal whenever $k \neq l$ and as $\mathcal{H}_k(S^n) \subseteq E_k$ and $\mathcal{H}_l(S^n) \subseteq E_l$, our claim is indeed true. Furthermore, by Proposition 16.5, each $\mathcal{H}_k(S^n)$ is finite-dimensional and thus, closed. Finally, we know from Proposition 16.6 that $\bigcup_{k=0}^{\infty} \mathcal{H}_k(S^n)$ is dense in $L^2(S^n)$. But then, we can apply a standard result from Hilbert space theory (for example, see Lang [94], Chapter V, Proposition 1.9) to deduce the following important result:

Theorem 16.8. *The family of spaces, $\mathcal{H}_k(S^n)$ (resp. $\mathcal{H}_k^{\mathbb{C}}(S^n)$) yields a Hilbert space direct sum decomposition*

$$L^2(S^n) = \bigoplus_{k=0}^{\infty} \mathcal{H}_k(S^n) \qquad (\textit{resp.} \quad L^2_{\mathbb{C}}(S^n) = \bigoplus_{k=0}^{\infty} \mathcal{H}_k^{\mathbb{C}}(S^n)),$$

which means that the summands are closed, pairwise orthogonal, and that every $f \in L^2(S^n)$ (resp. $f \in L^2_{\mathbb{C}}(S^n)$) is the sum of a converging series

$$f = \sum_{k=0}^{\infty} f_k,$$

in the L^2-norm, where the $f_k \in \mathcal{H}_k(S^n)$ (resp. $f_k \in \mathcal{H}_k^{\mathbb{C}}(S^n)$) are uniquely determined functions. Furthermore, given any orthonormal basis, $(Y_k^1, \ldots, Y_k^{a_{k,n+1}})$, of $\mathcal{H}_k(S^n)$, we have

$$f_k = \sum_{m_k=1}^{a_{k,n+1}} c_{k,m_k} Y_k^{m_k}, \qquad \textit{with} \quad c_{k,m_k} = \langle f, Y_k^{m_k} \rangle.$$

The coefficients c_{k,m_k} are "generalized" *Fourier coefficients* with respect to the Hilbert basis $\{Y_k^{m_k} \mid 1 \leq m_k \leq a_{k,n+1}, k \geq 0\}$. We can finally prove the main theorem of this section.

Theorem 16.9.

(1) The eigenspaces (resp. complex eigenspaces) of the Laplacian, Δ_{S^n}, on S^n are the spaces of spherical harmonics,

$$E_k = \mathcal{H}_k(S^n) \qquad (\textit{resp.} \quad E_k = \mathcal{H}_k^{\mathbb{C}}(S^n))$$

and E_k corresponds to the eigenvalue $-k(n+k-1)$.

(2) We have the Hilbert space direct sum decompositions

$$L^2(S^n) = \bigoplus_{k=0}^{\infty} E_k \qquad (\textit{resp.} \quad L^2_{\mathbb{C}}(S^n) = \bigoplus_{k=0}^{\infty} E_k).$$

(3) The complex polynomials of the form $(c_1 x_1 + \cdots + c_{n+1} x_{n+1})^k$, with $c_1^2 + \cdots + c_{n+1}^2 = 0$, span the space $\mathcal{H}_k^{\mathbb{C}}(n+1)$, for $k \geq 1$.

Proof. We follow essentially the proof in Helgason [73] (Introduction, Theorem 3.1). In (1) and (2) we only deal with the real case, the proof in the complex case being identical.

(1) We already know that the integers $-k(n+k-1)$ are eigenvalues of Δ_{S^n} and that $\mathcal{H}_k(S^n) \subseteq E_k$. We will prove that Δ_{S^n} has no other eigenvalues and no other eigenvectors using the Hilbert basis, $\{Y_k^{m_k} \mid 1 \leq m_k \leq a_{k,n+1},\ k \geq 0\}$, given by Theorem 16.8. Let λ be any eigenvalue of Δ_{S^n} and let $f \in L^2(S^n)$ be any eigenfunction associated with λ so that

$$\Delta f = \lambda f.$$

We have a unique series expansion

$$f = \sum_{k=0}^{\infty} \sum_{m_k=1}^{a_{k,n+1}} c_{k,m_k} Y_k^{m_k},$$

with $c_{k,m_k} = \langle f, Y_k^{m_k}\rangle$. Now, as Δ_{S^n} is self-adjoint and $\Delta Y_k^{m_k} = -k(n+k-1)Y_k^{m_k}$, the Fourier coefficients, d_{k,m_k}, of Δf are given by

$$d_{k,m_k} = \langle \Delta f, Y_k^{m_k}\rangle = \langle f, \Delta Y_k^{m_k}\rangle = -k(n+k-1)\langle f, Y_k^{m_k}\rangle = -k(n+k-1)c_{k,m_k}.$$

On the other hand, as $\Delta f = \lambda f$, the Fourier coefficients of Δf are given by

$$d_{k,m_k} = \lambda c_{k,m_k}.$$

By uniqueness of the Fourier expansion, we must have

$$\lambda c_{k,m_k} = -k(n+k-1)c_{k,m_k} \qquad \text{for all } k \geq 0.$$

Since $f \neq 0$, there some k such that $c_{k,m_k} \neq 0$ and we must have

$$\lambda = -k(n+k-1)$$

for any such k. However, the function $k \mapsto -k(n+k-1)$ reaches its maximum for $k = -\frac{n-1}{2}$ and as $n \geq 1$, it is strictly decreasing for $k \geq 0$, which implies that k is unique and that

$$c_{j,m_j} = 0 \qquad \text{for all } j \neq k.$$

Therefore, $f \in \mathcal{H}_k(S^n)$ and the eigenvalues of Δ_{S^n} are exactly the integers $-k(n+k-1)$ so $E_k = \mathcal{H}_k(S^n)$, as claimed.

Since we just proved that $E_k = \mathcal{H}_k(S^n)$, (2) follows immediately from the Hilbert decomposition given by Theorem 16.8.

(3) If $H = (c_1x_1 + \cdots + c_{n+1}x_{n+1})^k$, with $c_1^2 + \cdots + c_{n+1}^2 = 0$, then for $k \leq 1$ is is obvious that $\Delta H = 0$ and for $k \geq 2$ we have

$$\Delta H = k(k-1)(c_1^2 + \cdots + c_{n+1}^2)(c_1x_1 + \cdots + c_{n+1}x_{n+1})^{k-2} = 0,$$

so $H \in \mathcal{H}_k^{\mathbb{C}}(n+1)$. A simple computation shows that for every $Q \in \mathcal{P}_k^{\mathbb{C}}(n+1)$, if $c = (c_1, \ldots, c_{n+1})$, then we have

$$\partial(Q)(c_1x_1 + \cdots + c_{n+1}x_{n+1})^m = m(m-1)\cdots(m-k+1)Q(c)(c_1x_1 + \cdots + c_{n+1}x_{n+1})^{m-k},$$

for all $m \geq k \geq 1$.

Assume that $\mathcal{H}_k^{\mathbb{C}}(n+1)$ is not spanned by the complex polynomials of the form $(c_1x_1 + \cdots + c_{n+1}x_{n+1})^k$, with $c_1^2 + \cdots + c_{n+1}^2 = 0$, for $k \geq 1$. Then, some $Q \in \mathcal{H}_k^{\mathbb{C}}(n+1)$ is orthogonal to all polynomials of the form $H = (c_1x_1 + \cdots + c_{n+1}x_{n+1})^k$, with $c_1^2 + \cdots + c_{n+1}^2 = 0$. Recall that

$$\langle P, \partial(Q)H\rangle = \langle QP, H\rangle$$

and apply this equation to $P = Q(c)$, H and Q. Since

$$\partial(Q)H = \partial(Q)(c_1x_1 + \cdots + c_{n+1}x_{n+1})^k = k!Q(c),$$

as Q is orthogonal to H, we get

$$k!\langle Q(c), Q(c)\rangle = \langle Q(c), k!Q(c)\rangle = \langle Q(c), \partial(Q)H\rangle = \langle Q\,Q(c), H\rangle = Q(c)\langle Q, H\rangle = 0,$$

which implies $Q(c) = 0$. Consequently, $Q(x_1, \ldots, x_{n+1})$ vanishes on the complex algebraic variety,

$$\{(x_1, \ldots, x_{n+1}) \in \mathbb{C}^{n+1} \mid x_1^2 + \cdots + x_{n+1}^2 = 0\}.$$

By the Hilbert *Nullstellensatz*, some power, Q^m, belongs to the ideal, $(x_1^2 + \cdots + x_{n+1}^2)$, generated by $x_1^2+\cdots+x_{n+1}^2$. Now, if $n \geq 2$, it is well-known that the polynomial $x_1^2+\cdots+x_{n+1}^2$ is irreducible so the ideal $(x_1^2 + \cdots + x_{n+1}^2)$ is a prime ideal and thus, Q is divisible by $x_1^2+\cdots+x_{n+1}^2$. However, we know from the proof of Theorem 16.3 that we have an orthogonal direct sum

$$\mathcal{P}_k^{\mathbb{C}}(n+1) = \mathcal{H}_k^{\mathbb{C}}(n+1) \oplus \|x\|^2\, \mathcal{P}_{k-2}^{\mathbb{C}}(n+1).$$

Since $Q \in \mathcal{H}_k^{\mathbb{C}}(n+1)$ and Q is divisible by $x_1^2 + \cdots + x_{n+1}^2$, we must have $Q = 0$. Therefore, if $n \geq 2$, we proved (3). However, when $n = 1$, we know from Section 16.1 that the complex harmonic homogeneous polynomials in two variables, $P(x, y)$, are spanned by the real and imaginary parts, U_k, V_k of the polynomial $(x + iy)^k = U_k + iV_k$. Since $(x - iy)^k = U_k - iV_k$ we see that

$$U_k = \frac{1}{2}\left((x+iy)^k + (x-iy)^k\right), \qquad V_k = \frac{1}{2i}\left((x+iy)^k - (x-iy)^k\right),$$

and as $1 + i^2 = 1 + (-i)^2 = 0$, the space $\mathcal{H}_k^{\mathbb{C}}(\mathbb{R}^2)$ is spanned by $(x + iy)^k$ and $(x - iy)^k$ (for $k \geq 1$), so (3) holds for $n = 1$ as well. □

As an illustration of part (3) of Theorem 16.9, the polynomials $(x_1 + i\cos\theta x_2 + i\sin\theta x_3)^k$ are harmonic. Of course, the real and imaginary part of a complex harmonic polynomial $(c_1x_1 + \cdots + c_{n+1}x_{n+1})^k$ are real harmonic polynomials.

In the next section, we try to show how spherical harmonics fit into the broader framework of linear respresentations of (Lie) groups.

16.5 Spherical Functions and Linear Representations of Lie Groups; A Glimpse

In this section, we indicate briefly how Theorem 16.9 (except part (3)) can be viewed as a special case of a famous theorem known as the *Peter-Weyl Theorem* about unitary representations of compact Lie groups (Herman, Klauss, Hugo Weyl, 1885-1955). First, we review the notion of a linear representation of a group. A good and easy-going introduction to representations of Lie groups can be found in Hall [71]. We begin with finite-dimensional representations.

Definition 16.2. Given a Lie group, G, and a vector space, V, of dimension n, a *linear representation* of G of *dimension* (or *degree* n) is a group homomorphism, $U : G \to \mathbf{GL}(V)$, such that the map, $g \mapsto U(g)(u)$, is continuous for every $u \in V$ and where $\mathbf{GL}(V)$ denotes the group of invertible linear maps from V to itself. The space, V, called the *representation space* may be a real or a complex vector space. If V has a Hermitian (resp Euclidean) inner product, $\langle -, - \rangle$, we say that $U : G \to \mathbf{GL}(V)$ is a *unitary representation* iff

$$\langle U(g)(u), U(g)(v) \rangle = \langle u, v \rangle, \qquad \text{for all } g \in G \text{ and all } u, v \in V.$$

Thus, a linear representation of G is a map, $U : G \to \mathbf{GL}(V)$, satisfying the properties:

$$\begin{aligned} U(gh) &= U(g)U(h) \\ U(g^{-1}) &= U(g)^{-1} \\ U(1) &= I. \end{aligned}$$

For simplicity of language, we usually abbreviate *linear representation* as *representation*. The representation space, V, is also called a *G-module* since the representation, $U : G \to \mathbf{GL}(V)$, is equivalent to the left action, $\cdot : G \times V \to V$, with $g \cdot v = U(g)(v)$. The representation such that $U(g) = I$ for all $g \in G$ is called the *trivial representation*.

As an example, we describe a class of representations of $\mathbf{SL}(2, \mathbb{C})$, the group of complex matrices with determinant $+1$,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad ad - bc = 1.$$

Recall that $\mathcal{P}_k^{\mathbb{C}}(2)$ denotes the vector space of complex homogeneous polynomials of degree k in two variables, (z_1, z_2). For every matrix, $A \in \mathbf{SL}(2, \mathbb{C})$, with

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

for every homogeneous polynomial, $Q \in \mathcal{P}_k^{\mathbb{C}}(2)$, we define $U_k(A)(Q(z_1, z_2))$ by

$$U_k(A)(Q(z_1, z_2)) = Q(dz_1 - bz_2, -cz_1 + az_2).$$

If we think of the homogeneous polynomial, $Q(z_1, z_2)$, as a function, $Q\binom{z_1}{z_2}$, of the vector $\binom{z_1}{z_2}$, then

$$U_k(A)\left(Q\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}\right) = QA^{-1}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = Q\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.$$

The expression above makes it clear that

$$U_k(AB) = U_k(A)U_k(B)$$

for any two matrices, $A, B \in \mathbf{SL}(2,\mathbb{C})$, so U_k is indeed a representation of $\mathbf{SL}(2,\mathbb{C})$ into $\mathcal{P}_k^{\mathbb{C}}(2)$. It can be shown that the representations, U_k, are irreducible and that every representation of $\mathbf{SL}(2,\mathbb{C})$ is equivalent to one of the U_k's (see Bröcker and tom Dieck [25], Chapter 2, Section 5). The representations, U_k, are also representations of $\mathbf{SU}(2)$. Again, they are irreducible representations of $\mathbf{SU}(2)$ and they constitute all of them (up to equivalence). The reader should consult Hall [71] for more examples of representations of Lie groups.

One might wonder why we considered $\mathbf{SL}(2,\mathbb{C})$ rather than $\mathbf{SL}(2,\mathbb{R})$. This is because it can be shown that $\mathbf{SL}(2,\mathbb{R})$ has *no* nontrivial unitary (finite-dimensional) representations! For more on representations of $\mathbf{SL}(2,\mathbb{R})$, see Dieudonné [44] (Chapter 14).

Given any basis, $(e_1,\ldots,e_n)$, of V, each $U(g)$ is represented by an $n \times n$ matrix, $U(g) = (U_{ij}(g))$. We may think of the scalar functions, $g \mapsto U_{ij}(g)$, as *special functions* on G. As explained in Dieudonné [44] (see also Vilenkin [147]), essentially all special functions (Legendre polynomials, ultraspherical polynomials, Bessel functions, *etc.*) arise in this way by choosing some suitable G and V. There is a natural and useful notion of equivalence of representations:

Definition 16.3. Given any two representations, $U_1 \colon G \to \mathbf{GL}(V_1)$ and $U_2 \colon G \to \mathbf{GL}(V_2)$, a *$G$-map* (or *morphism of representations*), $\varphi \colon U_1 \to U_2$, is a linear map, $\varphi \colon V_1 \to V_2$, so that the following diagram commutes for every $g \in G$:

$$\begin{array}{ccc} V_1 & \xrightarrow{U_1(g)} & V_1 \\ {\scriptstyle\varphi}\downarrow & & \downarrow{\scriptstyle\varphi} \\ V_2 & \xrightarrow{U_2(g)} & V_2. \end{array}$$

The space of all G-maps between two representations as above is denoted $\mathrm{Hom}_G(U_1, U_2)$. Two representations $U_1 \colon G \to \mathbf{GL}(V_1)$ and $U_2 \colon G \to \mathbf{GL}(V_2)$ are *equivalent* iff $\varphi \colon V_1 \to V_2$ is an invertible linear map (which implies that $\dim V_1 = \dim V_2$). In terms of matrices, the representations $U_1 \colon G \to \mathbf{GL}(V_1)$ and $U_2 \colon G \to \mathbf{GL}(V_2)$ are equivalent iff there is some invertible $n \times n$ matrix, P, so that

$$U_2(g) = PU_1(g)P^{-1}, \qquad g \in G.$$

If $W \subseteq V$ is a subspace of V, then in some cases, a representation $U \colon G \to \mathbf{GL}(V)$ yields a representation $U \colon G \to \mathbf{GL}(W)$. This is interesting because under certain conditions on G (*e.g.*, G compact) every representation may be decomposed into a "sum" of so-called irreducible representations and thus, the study of all representations of G boils down to the study of irreducible representations of G (for instance, see Knapp [90] (Chapter 4, Corollary 4.7) or Bröcker and tom Dieck [25] (Chapter 2, Proposition 1.9).

Definition 16.4. Let $U \colon G \to \mathbf{GL}(V)$ be a representation of G. If $W \subseteq V$ is a subspace of V, then we say that W is *invariant* (or *stable*) under U iff $U(g)(w) \in W$, for all $g \in G$ and all

$w \in W$. If W is invariant under U, then we have a homomorphism, $U\colon G \to \mathbf{GL}(W)$, called a *subrepresentation* of G. A representation, $U\colon G \to \mathbf{GL}(V)$, with $V \neq (0)$ is *irreducible* iff it only has the two subrepresentations, $U\colon G \to \mathbf{GL}(W)$, corresponding to $W = (0)$ or $W = V$.

An easy but crucial lemma about irreducible representations is "Schur's Lemma".

Lemma 16.10. *(Schur's Lemma) Let $U_1\colon G \to \mathbf{GL}(V)$ and $U_2\colon G \to \mathbf{GL}(W)$ be any two real or complex representations of a group, G. If U_1 and U_2 are irreducible, then the following properties hold:*

(i) Every G-map, $\varphi\colon U_1 \to U_2$, is either the zero map or an isomorphism.

(ii) If U_1 is a complex representation, then every G-map, $\varphi\colon U_1 \to U_1$, is of the form, $\varphi = \lambda \mathrm{id}$, for some $\lambda \in \mathbb{C}$.

Proof. (i) Observe that the kernel, $\operatorname{Ker}\varphi \subseteq V$, of φ is invariant under U_1. Indeed, for every $v \in \operatorname{Ker}\varphi$ and every $g \in G$, we have

$$\varphi(U_1(g)(v)) = U_2(g)(\varphi(v)) = U_2(g)(0) = 0,$$

so $U_1(g)(v) \in \operatorname{Ker}\varphi$. Thus, $U_1\colon G \to \mathbf{GL}(\operatorname{Ker}\varphi)$ is a subrepresentation of U_1 and as U_1 is irreducible, either $\operatorname{Ker}\varphi = (0)$ or $\operatorname{Ker}\varphi = V$. In the second case, $\varphi = 0$. If $\operatorname{Ker}\varphi = (0)$, then φ is injective. However, $\varphi(V) \subseteq W$ is invariant under U_2 since for every $v \in V$ and every $g \in G$,

$$U_2(g)(\varphi(v)) = \varphi(U_1(g)(v)) \in \varphi(V),$$

and as $\varphi(V) \neq (0)$ (as $V \neq (0)$ since U_1 is irreducible) and U_2 is irreducible, we must have $\varphi(V) = W$, that is, φ is an isomorphism.

(ii) Since V is a complex vector space, the linear map, φ, has some eigenvalue, $\lambda \in \mathbb{C}$. Let $E_\lambda \subseteq V$ be the eigenspace associated with λ. The subspace E_λ is invariant under U_1 since for every $u \in E_\lambda$ and every $g \in G$, we have

$$\varphi(U_1(g)(u)) = U_1(g)(\varphi(u)) = U_1(g)(\lambda u) = \lambda U_1(g)(u),$$

so $U_1\colon G \to \mathbf{GL}(E_\lambda)$ is a subrepresentation of U_1 and as U_1 is irreducible and $E_\lambda \neq (0)$, we must have $E_\lambda = V$. □

An interesting corollary of Schur's Lemma is that every complex irreducible representtaion of a commutative group is one-dimensional.

Let us now restrict our attention to compact Lie groups. If G is a compact Lie group, then it is known that it has a left and right-invariant volume form, ω_G, so we can define the integral of a (real or complex) continuous function, f, defined on G by

$$\int_G f = \int_G f\,\omega_G,$$

also denoted $\int_G f\, d\mu_G$ or simply $\int_G f(t)\, dt$, with ω_G normalized so that $\int_G \omega_G = 1$. (See Section 9.4, or Knapp [90], Chapter 8, or Warner [148], Chapters 4 and 6.) Because G is compact, the *Haar measure*, μ_G, induced by ω_G is both left and right-invariant (G is a *unimodular group*) and our integral has the following invariance properties:

$$\int_G f(t)\, dt = \int_G f(st)\, dt = \int_G f(tu)\, dt = \int_G f(t^{-1})\, dt,$$

for all $s, u \in G$ (see Section 9.4).

Since G is a compact Lie group, we can use an "averaging trick" to show that every (finite-dimensional) representation is equivalent to a unitary representation (see Bröcker and tom Dieck [25] (Chapter 2, Theorem 1.7) or Knapp [90] (Chapter 4, Proposition 4.6).

If we define the Hermitian inner product,

$$\langle f, g\rangle = \int_G f\, \overline{g}\, \omega_G,$$

then, with this inner product, the space of square-integrable functions, $L^2_{\mathbb{C}}(G)$, is a *Hilbert space*. We can also define the *convolution*, $f * g$, of two functions, $f, g \in L^2_{\mathbb{C}}(G)$, by

$$(f * g)(x) = \int_G f(xt^{-1})g(t)dt = \int_G f(t)g(t^{-1}x)dt$$

In general, $f * g \neq g * f$ unless G is commutative. With the convolution product, $L^2_{\mathbb{C}}(G)$ becomes an associative algebra (non-commutative in general).

This leads us to consider unitary representations of G into the infinite-dimensional vector space, $L^2_{\mathbb{C}}(G)$. The definition is the same as in Definition 16.2, except that $\mathbf{GL}(L^2_{\mathbb{C}}(G))$ is the group of automorphisms (unitary operators), $\mathrm{Aut}(L^2_{\mathbb{C}}(G))$, of the Hilbert space, $L^2_{\mathbb{C}}(G)$ and

$$\langle U(g)(u), U(g)(v)\rangle = \langle u, v\rangle$$

with respect to the inner product on $L^2_{\mathbb{C}}(G)$. Also, in the definition of an irreducible representation, $U\colon G \to V$, we require that the only *closed* subrepresentations, $U\colon G \to W$, of the representation, $U\colon G \to V$, correspond to $W = (0)$ or $W = V$.

The *Peter Weyl Theorem* gives a decomposition of $L^2_{\mathbb{C}}(G)$ as a Hilbert sum of spaces that correspond to irreducible unitary representations of G. We present a version of the Peter Weyl Theorem found in Dieudonné [44] (Chapters 3-8) and Dieudonné [45] (Chapter XXI, Sections 1-4), which contains complete proofs. Other versions can be found in Bröcker and tom Dieck [25] (Chapter 3), Knapp [90] (Chapter 4) or Duistermaat and Kolk [54] (Chapter 4). A good preparation for these fairly advanced books is Deitmar [41].

Theorem 16.11. *(Peter-Weyl (1927)) Given a compact Lie group, G, there is a decomposition of $L^2_{\mathbb{C}}(G)$ as a Hilbert sum,*

$$L^2_{\mathbb{C}}(G) = \bigoplus_{\rho} \mathfrak{a}_\rho,$$

of countably many two-sided ideals, $\mathfrak{a}_\rho$, where each $\mathfrak{a}_\rho$ is isomorphic to a finite-dimensional algebra of $n_\rho \times n_\rho$ complex matrices. More precisely, there is a basis of $\mathfrak{a}_\rho$ consisting of smooth pairwise orthogonal functions, $m_{ij}^{(\rho)}$, satisfying various properties, including

$$\langle m_{ij}^{(\rho)}, m_{ij}^{(\rho)} \rangle = n_\rho,$$

and if we form the matrix, $M_\rho(g) = (\frac{1}{n_\rho} m_{ij}^{(\rho)}(g))$, then the map, $g \mapsto M_\rho(g)$ is an **irreducible unitary representation** *of G in the vector space $\mathbb{C}^{n_\rho}$. Furthermore, every irreducible representation of G is equivalent to some M_ρ, so the set of indices, ρ, corresponds to the set of equivalence classes of irreducible unitary representations of G. The function, u_ρ, given by*

$$u_\rho(g) = \sum_{j=1}^{n_\rho} m_{jj}^{(\rho)}(g) = n_\rho \mathrm{tr}(M_\rho(g))$$

is the unit of the algebra $\mathfrak{a}_\rho$ and the orthogonal projection of $L^2_{\mathbb{C}}(G)$ onto $\mathfrak{a}_\rho$ is the map

$$f \mapsto u_\rho * f,$$

that is, convolution with u_ρ.

Remark: The function, $\chi_\rho = \frac{1}{n_\rho} u_\rho = \mathrm{tr}(M_\rho)$, is the *character* of G associated with the representation of G into M_ρ. The functions, χ_ρ, form an orthogonal system. Beware that they are *not* homomorphisms of G into $\mathbb{C}$ unless G is commutative. The characters of G are the key to the definition of the Fourier transform on a (compact) group, G.

A complete proof of Theorem 16.11 is given in Dieudonné [45], Chapter XXI, Section 2, but see also Sections 3 and 4.

There is more to the Peter Weyl Theorem: It gives a description of all unitary representations of G into a separable Hilbert space (see Dieudonné [45], Chapter XXI, Section 4). If $V\colon G \to \mathrm{Aut}(E)$ is such a representation, then for every ρ as above, the map

$$x \mapsto V(u_\rho)(x) = \int_G (V(s)(x)) u_\rho(s)\, ds$$

is an orthogonal projection of E onto a closed subspace, E_ρ. Then, E is the Hilbert sum, $E = \bigoplus_\rho E_\rho$, of those E_ρ such that $E_\rho \neq (0)$ and each such E_ρ is itself a (countable) Hilbert sum of closed spaces invariant under V. The subrepresentations of V corresponding to these subspaces of E_ρ are all equivalent to $M_{\overline{\rho}} = \overline{M_\rho}$ and hence, irreducible. This is why *every* (unitary) representation of G is equivalent to some representation of the form M_ρ.

An interesting special case is the case of the so-called regular representation of G in $L^2_{\mathbb{C}}(G)$ itself. The (left) *regular representation*, $\mathbf{R}$, of G in $L^2_{\mathbb{C}}(G)$ is defined by

$$(\mathbf{R}_s(f))(t) = \lambda_s(f)(t) = f(s^{-1}t), \qquad f \in L^2_{\mathbb{C}}(G),\ s, t \in G.$$

It turns out that we also get the same Hilbert sum,

$$L^2_{\mathbb{C}}(G) = \bigoplus_{\rho} \mathfrak{a}_\rho,$$

but this time, the $\mathfrak{a}_\rho$ generally do not correspond to irreducible subrepresentations. However, $\mathfrak{a}_\rho$ splits into n_ρ left ideals, $\mathfrak{b}^{(\rho)}_j$, where $\mathfrak{b}^{(\rho)}_j$ corresponds to the jth columm of M_ρ and all the subrepresentations of G in $\mathfrak{b}^{(\rho)}_j$ are equivalent to $M_{\overline{\rho}}$ and thus, are irreducible (see Dieudonné [44], Chapter 3).

Finally, assume that besides the compact Lie group, G, we also have a closed subgroup, K, of G. Then, we know that $M = G/K$ is a manifold called a *homogeneous space* and G acts on M on the left. For example, if $G = \mathbf{SO}(n+1)$ and $K = \mathbf{SO}(n)$, then $S^n = \mathbf{SO}(n+1)/\mathbf{SO}(n)$ (for instance, see Warner [148], Chapter 3). The subspace of $L^2_{\mathbb{C}}(G)$ consisting of the functions $f \in L^2_{\mathbb{C}}(G)$ that are right-invariant under the action of K, that is, such that

$$f(su) = f(s) \qquad \text{for all } s \in G \text{ and all } u \in K$$

form a closed subspace of $L^2_{\mathbb{C}}(G)$ denoted $L^2_{\mathbb{C}}(G/K)$. For example, if $G = \mathbf{SO}(n+1)$ and $K = \mathbf{SO}(n)$, then $L^2_{\mathbb{C}}(G/K) = L^2_{\mathbb{C}}(S^n)$.

It turns out that $L^2_{\mathbb{C}}(G/K)$ is invariant under the regular representation, $\mathbf{R}$, of G in $L^2_{\mathbb{C}}(G)$, so we get a subrepresentation (of the regular representation) of G in $L^2_{\mathbb{C}}(G/K)$. Again, the Peter-Weyl gives us a Hilbert sum decomposition of $L^2_{\mathbb{C}}(G/K)$ of the form

$$L^2_{\mathbb{C}}(G/K) = \bigoplus_{\rho} L_\rho = L^2_{\mathbb{C}}(G/K) \cap \mathfrak{a}_\rho,$$

for the same ρ's as before. However, these subrepresentations of $\mathbf{R}$ in L_ρ are not necessarily irreducible. What happens is that there is some d_ρ with $0 \leq d_\rho \leq n_\rho$ so that if $d_\rho \geq 1$, then L_σ is the direct sum of the first d_ρ columns of M_ρ (see Dieudonné [44], Chapter 6 and Dieudonné [46], Chapter XXII, Sections 4-5).

We can also consider the subspace of $L^2_{\mathbb{C}}(G)$ consisting of the functions, $f \in L^2_{\mathbb{C}}(G)$, that are left-invariant under the action of K, that is, such that

$$f(ts) = f(s) \qquad \text{for all } s \in G \text{ and all } t \in K.$$

This is a closed subspace of $L^2_{\mathbb{C}}(G)$ denoted $L^2_{\mathbb{C}}(K\backslash G)$. Then, we get a Hilbert sum decomposition of $L^2_{\mathbb{C}}(K\backslash G)$ of the form

$$L^2_{\mathbb{C}}(K\backslash G) = \bigoplus_{\rho} L'_\rho = L^2_{\mathbb{C}}(K\backslash G) \cap \mathfrak{a}_\rho,$$

and for the same d_ρ as before, L'_σ is the direct sum of the first d_ρ rows of M_ρ. We can also consider

$$\begin{aligned} L^2_{\mathbb{C}}(K\backslash G/K) &= L^2_{\mathbb{C}}(G/K) \cap L^2_{\mathbb{C}}(K\backslash G) \\ &= \{f \in L^2_{\mathbb{C}}(G) \mid f(tsu) = f(s)\} \qquad \text{for all } s \in G \text{ and all } t, u \in K. \end{aligned}$$

From our previous discussion, we see that we have a Hilbert sum decomposition

$$L^2_{\mathbb{C}}(K\backslash G/K) = \bigoplus_{\rho} L_\rho \cap L'_\rho$$

and each $L_\rho \cap L'_\rho$ for which $d_\rho \geq 1$ is a matrix algebra of dimension d_ρ^2. As a consequence, *the algebra $L^2_{\mathbb{C}}(K\backslash G/K)$ is commutative iff $d_\rho \leq 1$ for all ρ.*

If the algebra $L^2_{\mathbb{C}}(K\backslash G/K)$ is commutative (for the convolution product), we say that (G, K) is a *Gelfand pair* (see Dieudonné [44], Chapter 8 and Dieudonné [46], Chapter XXII, Sections 6-7). In this case, the L_ρ in the Hilbert sum decomposition of $L^2_{\mathbb{C}}(G/K)$ are nontrivial of dimension n_ρ iff $d_\rho = 1$ and the subrepresentation, $\mathbf{U}$, (of the regular representation) of G into L_ρ is irreducible and equivalent to $M_{\overline{\rho}}$. The space L_ρ is generated by the functions, $m^{(\rho)}_{1,1}, \ldots, m^{(\rho)}_{n_\rho,1}$, but the function

$$\omega_\rho(s) = \frac{1}{n_\rho}\, m^{(\rho)}_{1,1}(s)$$

plays a special role. This function called a *zonal spherical function* has some interesting properties. First, $\omega_\rho(e) = 1$ (where e is the identity element of the group, G) and

$$\omega_\rho(ust) = \omega_\rho(s) \qquad \text{for all } s \in G \text{ and all } u, t \in K.$$

In addition, ω_ρ is of positive type. A function, $f\colon G \to \mathbb{C}$, is of *positive type* iff

$$\sum_{j,k=1}^{n} f(s_j^{-1}s_k) z_j \overline{z}_k \geq 0,$$

for every finite set, $\{s_1, \ldots, s_n\}$, of elements of G and every finite tuple, $(z_1, \ldots, z_n) \in \mathbb{C}^n$. Because the subrepresentation of G into L_ρ is irreducible, the function ω_ρ generates L_ρ under left translation. This means the following: If we recall that for any function, f, on G,

$$\lambda_s(f)(t) = f(s^{-1}t), \qquad s, t \in G,$$

then, L_ρ is generated by the functions $\lambda_s(\omega_\rho)$, as s varies in G. The function ω_ρ also satisfies the following property:

$$\omega_\rho(s) = \langle \mathbf{U}(s)(\omega_\rho), \omega_\rho \rangle.$$

The set of zonal spherical functions on G/K is denoted $S(G/K)$. It is a countable set.

The notion of Gelfand pair also applies to locally-compact unimodular groups that are not necessary compact but we will not discuss this notion here. Curious readers may consult Dieudonné [44] (Chapters 8 and 9) and Dieudonné [46] (Chapter XXII, Sections 6-9).

It turns out that $G = \mathbf{SO}(n+1)$ and $K = \mathbf{SO}(n)$ form a Gelfand pair (see Dieudonné [44], Chapters 7-8 and Dieudonné [47], Chapter XXIII, Section 38). In this particular case,

$\rho = k$ is any nonnegative integer and $L_\rho = E_k$, the eigenspace of the Laplacian on S^n corresponding to the eigenvalue $-k(n+k-1)$. Therefore, the regular representation of $\mathbf{SO}(n)$ into $E_k = \mathcal{H}_k^{\mathbb{C}}(S^n)$ is irreducible. This can be proved more directly, for example, see Helgason [73] (Introduction, Theorem 3.1) or Bröcker and tom Dieck [25] (Chapter 2, Proposition 5.10).

The zonal spherical harmonics, ω_k, can be expressed in terms of the *ultraspherical polynomials* (also called *Gegenbauer polynomials*), $P_k^{(n-1)/2}$ (up to a constant factor), see Stein and Weiss [143] (Chapter 4), Morimoto [114] (Chapter 2) and Dieudonné [44] (Chapter 7). For $n = 2$, $P_k^{\frac{1}{2}}$ is just the ordinary Legendre polynomial (up to a constant factor). We will say more about the zonal spherical harmonics and the ultraspherical polynomials in the next two sections.

The material in this section belongs to the overlapping areas of *representation theory* and *noncommutative harmonic analysis*. These are deep and vast areas. Besides the references cited earlier, for noncommutative harmonic analysis, the reader may consult Folland [55] or Taylor [145], but they may find the pace rather rapid. Another great survey on both topics is Kirillov [88], although it is not geared for the beginner.

16.6 Reproducing Kernel, Zonal Spherical Functions and Gegenbauer Polynomials

We now return to S^n and its spherical harmonics. The previous section suggested that zonal spherical functions play a special role. In this section, we describe the zonal spherical functions on S^n and show that they essentially come from certain polynomials generalizing the Legendre polyomials known as the *Gegenbauer Polynomials*. Most proof will be omitted. We refer the reader to Stein and Weiss [143] (Chapter 4) and Morimoto [114] (Chapter 2) for a complete exposition with proofs.

Recall that the space of spherical harmonics, $\mathcal{H}_k^{\mathbb{C}}(S^n)$, is the image of the space of homogeneous harmonic poynomials, $\mathcal{P}_k^{\mathbb{C}}(n+1)$, under the restriction map. It is a finite-dimensional space of dimension

$$a_{k,n+1} = \binom{n+k}{k} - \binom{n+k-2}{k-2},$$

if $n \geq 1$ and $k \geq 2$, with $a_{0,n+1} = 1$ and $a_{1,n+1} = n+1$. Let $(Y_k^1, \ldots, Y_k^{a_{k,n+1}})$ be any orthonormal basis of $\mathcal{H}_k^{\mathbb{C}}(S^n)$ and define $F_k(\sigma, \tau)$ by

$$F_k(\sigma, \tau) = \sum_{i=1}^{a_{k,n+1}} Y_k^i(\sigma)\overline{Y_k^i(\tau)}, \qquad \sigma, \tau \in S^n.$$

The following proposition is easy to prove (see Morimoto [114], Chapter 2, Lemma 1.19 and Lemma 2.20):

Proposition 16.12. *The function F_k is independent of the choice of orthonormal basis. Furthermore, for every orthogonal transformation, $R \in \mathbf{O}(n+1)$, we have*

$$F_k(R\sigma, R\tau) = F_k(\sigma, \tau), \qquad \sigma, \tau \in S^n.$$

Clearly, F_k is a symmetric function. Since we can pick an orthonormal basis of real orthogonal functions for $\mathcal{H}_k^{\mathbb{C}}(S^n)$ (pick a basis of $\mathcal{H}_k(S^n)$), Proposition 16.12 shows that F_k is a real-valued function.

The function F_k satisfies the following property which justifies its name, the *reproducing kernel* for $\mathcal{H}_k^{\mathbb{C}}(S^n)$:

Proposition 16.13. *For every spherical harmonic, $H \in \mathcal{H}_j^{\mathbb{C}}(S^n)$, we have*

$$\int_{S^n} H(\tau) F_k(\sigma, \tau)\, d\tau = \delta_{j\,k} H(\sigma), \qquad \sigma, \tau \in S^n,$$

for all $j, k \geq 0$.

Proof. When $j \neq k$, since $\mathcal{H}_k^{\mathbb{C}}(S^n)$ and $\mathcal{H}_j^{\mathbb{C}}(S^n)$ are orthogonal and since $F_k(\sigma, \tau) = \sum_{i=1}^{a_{k,n+1}} Y_k^i(\sigma)\overline{Y_k^i(\tau)}$, it is clear that the integral in Proposition 16.13 vanishes. When $j = k$, we have

$$\begin{aligned}
\int_{S^n} H(\tau) F_k(\sigma, \tau)\, d\tau &= \int_{S^n} H(\tau) \sum_{i=1}^{a_{k,n+1}} Y_k^i(\sigma)\overline{Y_k^i(\tau)}\, d\tau \\
&= \sum_{i=1}^{a_{k,n+1}} Y_k^i(\sigma) \int_{S^n} H(\tau)\overline{Y_k^i(\tau)}\, d\tau \\
&= \sum_{i=1}^{a_{k,n+1}} Y_k^i(\sigma)\, \langle H, Y_k^i \rangle \\
&= H(\sigma),
\end{aligned}$$

since $(Y_k^1, \ldots, Y_k^{a_{k,n+1}})$ is an orthonormal basis. □

In Stein and Weiss [143] (Chapter 4), the function $F_k(\sigma, \tau)$ is denoted by $Z_\sigma^{(k)}(\tau)$ and it is called the *zonal harmonic of degree k with pole σ*.

The value, $F_k(\sigma, \tau)$, of the function F_k depends only on $\sigma \cdot \tau$, as stated in Proposition 16.15 which is proved in Morimoto [114] (Chapter 2, Lemma 2.23). The following proposition also proved in Morimoto [114] (Chapter 2, Lemma 2.21) is needed to prove Proposition 16.15:

Proposition 16.14. *For all $\sigma, \tau, \sigma', \tau' \in S^n$, with $n \geq 1$, the following two conditions are equivalent:*

(i) There is some orthogonal transformation, $R \in \mathbf{O}(n+1)$, such that $R(\sigma) = \sigma'$ and $R(\tau) = \tau'$.

(ii) $\sigma \cdot \tau = \sigma' \cdot \tau'$.

Propositions 16.13 and 16.14 immediately yield

Proposition 16.15. *For all $\sigma, \tau, \sigma', \tau' \in S^n$, if $\sigma \cdot \tau = \sigma' \cdot \tau'$, then $F_k(\sigma, \tau) = F_k(\sigma', \tau')$. Consequently, there is some function, $\varphi\colon \mathbb{R} \to \mathbb{R}$, such that $F_k(\omega, \tau) = \varphi(\omega \cdot \tau)$.*

We are now ready to define zonal functions. Remarkably, the function φ in Proposition 16.15 comes from a real polynomial. We need the following proposition which is of independent interest:

Proposition 16.16. *If P is any (complex) polynomial in n variables such that*

$$P(R(x)) = P(x) \qquad \text{for all rotations, } R \in \mathbf{SO}(n), \text{ and all } x \in \mathbb{R}^n,$$

then P is of the form

$$P(x) = \sum_{j=0}^{m} c_j (x_1^2 + \cdots + x_n^2)^j,$$

for some $c_0, \ldots, c_m \in \mathbb{C}$.

Proof. Write P as the sum of its homogeneous pieces, $P = \sum_{l=0}^{k} Q_l$, where Q_l is homogeneous of degree l. Then, for every $\epsilon > 0$ and every rotation, R, we have

$$\sum_{l=0}^{k} \epsilon^l Q_l(x) = P(\epsilon x) = P(R(\epsilon x)) = P(\epsilon R(x)) = \sum_{l=0}^{k} \epsilon^l Q_l(R(x)),$$

which implies that

$$Q_l(R(x)) = Q_l(x), \qquad l = 0, \ldots, k.$$

If we let $F_l(x) = \|x\|^{-l} Q_l(x)$, then F_l is a homogeneous function of degree 0 and F_l is invariant under all rotations. This is only possible if F_l is a constant function, thus $F_l(x) = a_l$ for all $x \in \mathbb{R}^n$. But then, $Q_l(x) = a_l \|x\|^l$. Since Q_l is a polynomial, l must be even whenever $a_l \neq 0$. It follows that

$$P(x) = \sum_{j=0}^{m} c_j \|x\|^{2j}$$

with $c_j = a_{2j}$ for $j = 0, \ldots, m$ and where m is the largest integer $\leq k/2$. □

Proposition 16.16 implies that if a polynomial function on the sphere, S^n, in particular, a spherical harmonic, is invariant under all rotations, then it is a constant. If we relax this condition to invariance under all rotations leaving some given point, $\tau \in S^n$, invariant, then we obtain zonal harmonics.

The following theorem from Morimoto [114] (Chapter 2, Theorem 2.24) gives the relationship between zonal harmonics and the Gegenbauer polynomials:

Theorem 16.17. *Fix any $\tau \in S^n$. For every constant, $c \in \mathbb{C}$, there is a unique homogeneous harmonic polynomial, $Z_k^\tau \in \mathcal{H}_k^{\mathbb{C}}(n+1)$, satisfying the following conditions:*

(1) $Z_k^\tau(\tau) = c$;

(2) For every rotation, $R \in \mathbf{SO}(n+1)$, if $R\tau = \tau$, then $Z_k^\tau(R(x)) = Z_k^\tau(x)$, for all $x \in \mathbb{R}^{n+1}$.

Furthermore, we have

$$Z_k^\tau(x) = c\,\|x\|^k\, P_{k,n}\left(\frac{x}{\|x\|}\cdot\tau\right),$$

for some polynomial, $P_{k,n}(t)$, of degree k.

Remark: The proof given in Morimoto [114] is essentially the same as the proof of Theorem 2.12 in Stein and Weiss [143] (Chapter 4) but Morimoto makes an implicit use of Proposition 16.16 above. Also, Morimoto states Theorem 16.17 only for $c = 1$ but the proof goes through for any $c \in \mathbb{C}$, including $c = 0$, and we will need this extra generality in the proof of the Funk-Hecke formula.

Proof. Let $e_{n+1} = (0, \ldots, 0, 1) \in \mathbb{R}^{n+1}$ and for any $\tau \in S^n$, let R_τ be some rotation such that $R_\tau(e_{n+1}) = \tau$. Assume $Z \in \mathcal{H}_k^{\mathbb{C}}(n+1)$ satisfies conditions (1) and (2) and let Z' be given by $Z'(x) = Z(R_\tau(x))$. As $R_\tau(e_{n+1}) = \tau$, we have $Z'(e_{n+1}) = Z(\tau) = c$. Furthermore, for any rotation, S, such that $S(e_{n+1}) = e_{n+1}$, observe that

$$R_\tau \circ S \circ R_\tau^{-1}(\tau) = R_\tau \circ S(e_{n+1}) = R_\tau(e_{n+1}) = \tau,$$

and so, as Z satisfies property (2) for the rotation $R_\tau \circ S \circ R_\tau^{-1}$, we get

$$Z'(S(x)) = Z(R_\tau \circ S(x)) = Z(R_\tau \circ S \circ R_\tau^{-1} \circ R_\tau(x)) = Z(R_\tau(x)) = Z'(x),$$

which proves that Z' is a harmonic polynomial satisfying properties (1) and (2) with respect to e_{n+1}. Therefore, we may assume that $\tau = e_{n+1}$.

Write

$$Z(x) = \sum_{j=0}^{k} x_{n+1}^{k-j} P_j(x_1, \ldots, x_n),$$

where $P_j(x_1, \ldots, x_n)$ is a homogeneous polynomial of degree j. Since Z is invariant under every rotation, R, fixing e_{n+1} and since the monomials x_{n+1}^{k-j} are clearly invariant under such a rotation, we deduce that every $P_j(x_1, \ldots, x_n)$ is invariant under all rotations of $\mathbb{R}^n$ (clearly, there is a one-two-one correspondence between the rotations of $\mathbb{R}^{n+1}$ fixing e_{n+1} and the rotations of $\mathbb{R}^n$). By Proposition 16.16, we conclude that

$$P_j(x_1, \ldots, x_n) = c_j(x_1^2 + \cdots + x_n^2)^{\frac{j}{2}},$$

which implies that $P_j = 0$ is j is odd. Thus, we can write

$$Z(x) = \sum_{i=0}^{[k/2]} c_i x_{n+1}^{k-2i}(x_1^2 + \cdots + x_n^2)^i$$

where $[k/2]$ is the greatest integer, m, such that $2m \leq k$. If $k < 2$, then $Z = c_0$, so $c_0 = c$ and Z is uniquely determined. If $k \geq 2$, we know that Z is a harmonic polynomial so we assert that $\Delta Z = 0$. A simple computation shows that

$$\Delta(x_1^2 + \cdots + x_n^2)^i = 2i(n + 2i - 2)(x_1^2 + \cdots + x_n^2)^{i-1}$$

and

$$\begin{aligned}\Delta x_{n+1}^{k-2i}(x_1^2 + \cdots + x_n^2)^i &= (k-2i)(k-2i-1)x_{n+1}^{k-2i-2}(x_1^2 + \cdots + x_n^2)^i \\ &\quad + x_{n+1}^{k-2i}\Delta(x_1^2 + \cdots + x_n^2)^i \\ &= (k-2i)(k-2i-1)x_{n+1}^{k-2i-2}(x_1^2 + \cdots + x_n^2)^i \\ &\quad + 2i(n+2i-2)x_{n+1}^{k-2i}(x_1^2 + \cdots + x_n^2)^{i-1},\end{aligned}$$

so we get

$$\Delta Z = \sum_{i=0}^{[k/2]-1} \left((k-2i)(k-2i-1)c_i + 2(i+1)(n+2i)c_{i+1}\right) x_{n+1}^{k-2i-2}(x_1^2 + \cdots + x_n^2)^i.$$

Then, $\Delta Z = 0$ yields the relations

$$2(i+1)(n+2i)c_{i+1} = -(k-2i)(k-2i-1)c_i, \qquad i = 0, \ldots, [k/2] - 1,$$

which shows that Z is uniquely determined up to the constant c_0. Since we are requiring $Z(e_{n+1}) = c$, we get $c_0 = c$ and Z is uniquely determined. Now, on S^n, we have $x_1^2 + \cdots + x_{n+1}^2 = 1$, so if we let $t = x_{n+1}$, for $c_0 = 1$, we get a polynomial in one variable,

$$P_{k,n}(t) = \sum_{i=0}^{[k/2]} c_i t^{k-2i}(1-t^2)^i.$$

Thus, we proved that when $Z(e_{n+1}) = c$, we have

$$Z(x) = c\,\|x\|^k\; P_{k,n}\left(\frac{x_{n+1}}{\|x\|}\right) = c\,\|x\|^k\; P_{k,n}\left(\frac{x}{\|x\|}\cdot e_{n+1}\right).$$

When $Z(\tau) = c$, we write $Z = Z' \circ R_\tau^{-1}$ with $Z' = Z \circ R_\tau$ and where R_τ is a rotation such that $R_\tau(e_{n+1}) = \tau$. Then, as $Z'(e_{n+1}) = c$, using the formula above for Z', we have

$$\begin{aligned}Z(x) = Z'(R_\tau^{-1}(x)) &= c\left\|R_\tau^{-1}(x)\right\|^k\; P_{k,n}\left(\frac{R_\tau^{-1}(x)}{\|R_\tau^{-1}(x)\|}\cdot e_{n+1}\right) \\ &= c\,\|x\|^k\; P_{k,n}\left(\frac{x}{\|x\|}\cdot R_\tau(e_{n+1})\right) \\ &= c\,\|x\|^k\; P_{k,n}\left(\frac{x}{\|x\|}\cdot \tau\right),\end{aligned}$$

since R_τ is an isometry. □

The function, Z_k^τ, is called a *zonal function* and its restriction to S^n is a *zonal spherical function*. The polynomial, $P_{k,n}$, is called the *Gegenbauer polynomial* of degree k and dimension $n+1$ or *ultraspherical polynomial*. By definition, $P_{k,n}(1) = 1$.

The proof of Theorem 16.17 shows that for k even, say $k = 2m$, the polynomial $P_{2m,n}$ is of the form

$$P_{2m,n} = \sum_{j=0}^{m} c_{m-j} t^{2j} (1-t^2)^{m-j}$$

and for k odd, say $k = 2m+1$, the polynomial $P_{2m+1,n}$ is of the form

$$P_{2m+1,n} = \sum_{j=0}^{m} c_{m-j} t^{2j+1} (1-t^2)^{m-j}.$$

Consequently, $P_{k,n}(-t) = (-1)^k P_{k,n}(t)$, for all $k \geq 0$. The proof also shows that the "natural basis" for these polynomials consists of the polynomials, $t^i(1-t^2)^{\frac{k-i}{2}}$, with $k-i$ even. Indeed, with this basis, there are simple recurrence equations for computing the coefficients of $P_{k,n}$.

Remark: Morimoto [114] calls the polynomials, $P_{k,n}$, "Legendre polynomials". For $n = 2$, they are indeed the Legendre polynomials. Stein and Weiss denotes our (and Morimoto's) $P_{k,n}$ by $P_k^{\frac{n-1}{2}}$ (up to a constant factor) and Dieudonné [44] (Chapter 7) by $G_{k,n+1}$.

When $n = 2$, using the notation of Section 16.2, the zonal functions on S^2 are the spherical harmonics, y_l^0, for which $m = 0$, that is (up to a constant factor),

$$y_l^0(\theta, \varphi) = \sqrt{\frac{(2l+1)}{4\pi}}\, P_l(\cos\theta),$$

where P_l is the Legendre polynomial of degree l. For example, for $l = 2$, $P_l(t) = \frac{1}{2}(3t^2 - 1)$.

If we put $Z(r^k\sigma) = r^k F_k(\sigma, \tau)$ for a fixed τ, then by the definition of $F_k(\sigma, \tau)$ it is clear that Z is a homogeneous harmonic polynomial. The value $F_k(\tau, \tau)$ does not depend of τ because by transitivity of the action of $\mathbf{SO}(n+1)$ on S^n, for any other $\sigma \in S^n$, there is some rotation, R, so that $R\tau = \sigma$ and by Proposition 16.12, we have $F_k(\sigma, \sigma) = F_k(R\tau, R\tau) = F_k(\tau, \tau)$. To compute $F_k(\tau, \tau)$, since

$$F_k(\tau, \tau) = \sum_{i=1}^{a_{k,n+1}} \left\| Y_k^i(\tau) \right\|^2,$$

and since $(Y_k^1, \ldots, Y_k^{a_{k,n+1}})$ is an orthonormal basis of $\mathcal{H}_k^{\mathbb{C}}(S^n)$, observe that

$$\begin{aligned} a_{k,n+1} &= \sum_{i=1}^{a_{k,n+1}} \int_{S^n} \left\| Y_k^i(\tau) \right\|^2 d\tau && (16.1) \\ &= \int_{S^n} \left(\sum_{i=1}^{a_{k,n+1}} \left\| Y_k^i(\tau) \right\|^2 \right) d\tau && (16.2) \\ &= \int_{S^n} F_k(\tau, \tau)\, d\tau = F_k(\tau, \tau)\, \mathrm{vol}(S^n). && (16.3) \end{aligned}$$

Therefore,

$$F_k(\tau,\tau) = \frac{a_{k,n+1}}{\mathrm{vol}(S^n)}.$$

Beware that Morimoto [114] uses the normalized measure on S^n, so the factor involving $\mathrm{vol}(S^n)$ does not appear.

Remark: Recall that

$$\mathrm{vol}(S^{2d}) = \frac{2^{d+1}\pi^d}{1\cdot 3\cdots(2d-1)} \quad \text{if} \quad d \geq 1 \quad \text{and} \quad \mathrm{vol}(S^{2d+1}) = \frac{2\pi^{d+1}}{d!} \quad \text{if} \quad d \geq 0.$$

Now, if $R\tau = \tau$, then Proposition 16.12 shows that

$$Z(R(r^k\sigma)) = Z(r^k R(\sigma)) = r^k F_k(R\sigma,\tau) = r^k F_k(R\sigma,R\tau) = r^k F_k(\sigma,\tau) = Z(r^k\sigma).$$

Therefore, the function Z_k^τ satisfies conditions (1) and (2) of Theorem 16.17 with $c = \frac{a_{k,n+1}}{\mathrm{vol}(S^n)}$ and by uniqueness, we get

$$F_k(\sigma,\tau) = \frac{a_{k,n+1}}{\mathrm{vol}(S^n)}\, P_{k,n}(\sigma\cdot\tau).$$

Consequently, we have obtained the so-called *addition formula*:

Proposition 16.18. *(Addition Formula) If $(Y_k^1,\ldots,Y_k^{a_{k,n+1}})$ is any orthonormal basis of $\mathcal{H}_k^{\mathbb{C}}(S^n)$, then*

$$P_{k,n}(\sigma\cdot\tau) = \frac{\mathrm{vol}(S^n)}{a_{k,n+1}} \sum_{i=1}^{a_{k,n+1}} Y_k^i(\sigma)\overline{Y_k^i(\tau)}.$$

Again, beware that Morimoto [114] does not have the factor $\mathrm{vol}(S^n)$.

For $n = 1$, we can write $\sigma = (\cos\theta, \sin\theta)$ and $\tau = (\cos\varphi, \sin\varphi)$ and it is easy to see that the addition formula reduces to

$$P_{k,1}(\cos(\theta-\varphi)) = \cos k\theta\cos k\varphi + \sin k\theta \sin k\varphi = \cos k(\theta-\varphi),$$

the standard addition formula for trigonometric functions.

Proposition 16.18 implies that $P_{k,n}$ has real coefficients. Furthermore Proposition 16.13 is reformulated as

$$\frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \int_{S^n} P_{k,n}(\sigma\cdot\tau)H(\tau)\, d\tau = \delta_{j\,k} H(\sigma), \tag{rk}$$

showing that the Gengenbauer polynomials are reproducing kernels. A neat application of this formula is a formula for obtaining the kth spherical harmonic component of a function, $f \in L^2_{\mathbb{C}}(S^n)$.

Proposition 16.19. *For every function, $f \in L^2_{\mathbb{C}}(S^n)$, if $f = \sum_{k=0}^{\infty} f_k$ is the unique decomposition of f over the Hilbert sum $\bigoplus_{k=0}^{\infty} \mathcal{H}_k^{\mathbb{C}}(S^k)$, then f_k is given by*

$$f_k(\sigma) = \frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \int_{S^n} f(\tau) P_{k,n}(\sigma \cdot \tau)\, d\tau,$$

for all $\sigma \in S^n$.

Proof. If we recall that $\mathcal{H}_k^{\mathbb{C}}(S^k)$ and $\mathcal{H}_j^{\mathbb{C}}(S^k)$ are orthogonal for all $j \neq k$, using the formula (rk), we have

$$\begin{aligned}
\frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \int_{S^n} f(\tau) P_{k,n}(\sigma \cdot \tau)\, d\tau &= \frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \int_{S^n} \sum_{j=0}^{\infty} f_j(\tau) P_{k,n}(\sigma \cdot \tau)\, d\tau \\
&= \frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \sum_{j=0}^{\infty} \int_{S^n} f_j(\tau) P_{k,n}(\sigma \cdot \tau)\, d\tau \\
&= \frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \int_{S^n} f_k(\tau) P_{k,n}(\sigma \cdot \tau)\, d\tau \\
&= f_k(\sigma),
\end{aligned}$$

as claimed. □

We know from the previous section that the kth zonal function generates $\mathcal{H}_k^{\mathbb{C}}(S^n)$. Here is an explicit way to prove this fact.

Proposition 16.20. *If $H_1, \ldots, H_m \in \mathcal{H}_k^{\mathbb{C}}(S^n)$ are linearly independent, then there are m points, $\sigma_1, \ldots, \sigma_m$, on S^n, so that the $m \times m$ matrix, $(H_j(\sigma_i))$, is invertible.*

Proof. We proceed by induction on m. The case $m = 1$ is trivial. For the induction step, we may assume that we found m points, $\sigma_1, \ldots, \sigma_m$, on S^n, so that the $m \times m$ matrix, $(H_j(\sigma_i))$, is invertible. Consider the function

$$\sigma \mapsto \begin{vmatrix} H_1(\sigma) & \ldots & H_m(\sigma) & H_{m+1}(\sigma) \\ H_1(\sigma_1) & \ldots & H_m(\sigma_1) & H_{m+1}(\sigma_1) \\ \vdots & \ddots & \vdots & \vdots \\ H_1(\sigma_m) & \ldots & H_m(\sigma_m) & H_{m+1}(\sigma_m). \end{vmatrix}$$

Since $H_1, \ldots, H_{m+1}$ are linearly independent, the above function does not vanish for all σ since otherwise, by expanding this determinant with respect to the first row, we get a linear dependence among the H_j's where the coefficient of H_{m+1} is nonzero. Therefore, we can find σ_{m+1} so that the $(m+1) \times (m+1)$ matrix, $(H_j(\sigma_i))$, is invertible. □

We say that $a_{k,n+1}$ points, $\sigma_1, \ldots, \sigma_{a_{k,n+1}}$ on S^n form a *fundamental system* iff the $a_{k,n+1} \times a_{k,n+1}$ matrix, $(P_{n,k}(\sigma_i \cdot \sigma_j))$, is invertible.

Theorem 16.21. *The following properties hold:*

(i) There is a fundamental system, $\sigma_1, \ldots, \sigma_{a_{k,n+1}}$, for every $k \geq 1$.

(ii) Every spherical harmonic, $H \in \mathcal{H}_k^{\mathbb{C}}(S^n)$, can be written as

$$H(\sigma) = \sum_{j=1}^{a_{k,n+1}} c_j \, P_{k,n}(\sigma_j \cdot \sigma),$$

for some unique $c_j \in \mathbb{C}$.

Proof. (i) By the addition formula,

$$P_{k,n}(\sigma_i \cdot \sigma_j) = \frac{\mathrm{vol}(S^n)}{a_{k,n+1}} \sum_{l=1}^{a_{k,n+1}} Y_k^l(\sigma_i)\overline{Y_k^l(\sigma_j)}$$

for any orthonormal basis, $(Y_k^1, \ldots, Y_k^{a_{k,n+1}})$. It follows that the matrix $(P_{k,n}(\sigma_i \cdot \sigma_j))$ can be written as

$$(P_{k,n}(\sigma_i \cdot \sigma_j)) = \frac{\mathrm{vol}(S^n)}{a_{k,n+1}} \, YY^*,$$

where $Y = (Y_k^l(\sigma_i))$, and by Proposition 16.20, we can find $\sigma_1, \ldots, \sigma_{a_{k,n+1}} \in S^n$ so that Y and thus also Y^* are invertible and so, $(P_{n,k}(\sigma_i \cdot \sigma_j))$ is invertible.

(ii) Again, by the addition formula,

$$P_{k,n}(\sigma \cdot \sigma_j) = \frac{\mathrm{vol}(S^n)}{a_{k,n+1}} \sum_{i=1}^{a_{k,n+1}} Y_k^i(\sigma)\overline{Y_k^i(\sigma_j)}.$$

However, as $(Y_k^1, \ldots, Y_k^{a_{k,n+1}})$ is an orthonormal basis, (i) proved that the matrix Y^* is invertible so the $Y_k^i(\sigma)$ can be expressed uniquely in terms of the $P_{k,n}(\sigma \cdot \sigma_j)$, as claimed. □

A neat geometric characterization of the zonal spherical functions is given in Stein and Weiss [143]. For this, we need to define the notion of a parallel on S^n. A *parallel of S^n orthogonal to a point $\tau \in S^n$* is the intersection of S^n with any (affine) hyperplane orthogonal to the line through the center of S^n and τ. Clearly, any rotation, R, leaving τ fixed leaves every parallel orthogonal to τ globally invariant and for any two points, σ_1 and σ_2, on such a parallel there is a rotation leaving τ fixed that maps σ_1 to σ_2. Consequently, the zonal function, Z_k^τ, defined by τ is constant on the parallels orthogonal to τ. In fact, this property characterizes zonal harmonics, up to a constant.

The theorem below is proved in Stein and Weiss [143] (Chapter 4, Theorem 2.12). The proof uses Proposition 16.16 and it is very similar to the proof of Theorem 16.17 so, to save space, it is omitted.

Theorem 16.22. *Fix any point, $\tau \in S^n$. A spherical harmonic, $Y \in \mathcal{H}_k^{\mathbb{C}}(S^n)$, is constant on parallels orthogonal to τ iff $Y = cZ_k^\tau$, for some constant, $c \in \mathbb{C}$.*

In the next section, we show how the Gegenbauer polynomials can actually be computed.

16.7 More on the Gegenbauer Polynomials

The Gegenbauer polynomials are characterized by a formula generalizing the Rodrigues formula defining the Legendre polynomials (see Section 16.2). The expression

$$\left(k+\frac{n-2}{2}\right)\left(k-1+\frac{n-2}{2}\right)\cdots\left(1+\frac{n-2}{2}\right)$$

can be expressed in terms of the Γ function as

$$\frac{\Gamma\left(k+\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}.$$

Recall that the Γ function is a generalization of factorial that satisfies the equation

$$\Gamma(z+1)=z\Gamma(z).$$

For $z=x+iy$ with $x>0$, $\Gamma(z)$ is given by

$$\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}\,dt,$$

where the integral converges absolutely. If n is an integer $n\geq 0$, then $\Gamma(n+1)=n!$.

It is proved in Morimoto [114] (Chapter 2, Theorem 2.35) that

Proposition 16.23. *The Gegenbauer polynomial, $P_{k,n}$, is given by Rodrigues' formula:*

$$P_{k,n}(t)=\frac{(-1)^k}{2^k}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(k+\frac{n}{2}\right)}\frac{1}{(1-t^2)^{\frac{n-2}{2}}}\frac{d^k}{dt^k}(1-t^2)^{k+\frac{n-2}{2}},$$

with $n\geq 2$.

The Gegenbauer polynomials satisfy the following orthogonality properties with respect to the kernel $(1-t^2)^{\frac{n-2}{2}}$ (see Morimoto [114] (Chapter 2, Theorem 2.34):

Proposition 16.24. *The Gegenbauer polynomial, $P_{k,n}$, have the following properties:*

$$\begin{aligned}\int_{-1}^{1}(P_{k,n}(t))^2(1-t^2)^{\frac{n-2}{2}}\,dt &= \frac{\mathrm{vol}(S^n)}{a_{k,n+1}\mathrm{vol}(S^{n-1})}\\ \int_{-1}^{1}P_{k,n}(t)P_{l,n}(t)(1-t^2)^{\frac{n-2}{2}}\,dt &= 0,\qquad k\neq l.\end{aligned}$$

The Gegenbauer polynomials satisfy a second-order differential equation generalizing the Legendre equation from Section 16.2.

Proposition 16.25. *The Gegenbauer polynomial, $P_{k,n}$, are solutions of the differential equation*

$$(1-t^2)P''_{k,n}(t) - ntP'_{k,n}(t) + k(k+n-1)P_{k,n}(t) = 0.$$

Proof. For a fixed τ, the function H given by $H(\sigma) = P_{k,n}(\sigma \cdot \tau) = P_{k,n}(\cos\theta)$, belongs to $\mathcal{H}_k^{\mathbb{C}}(S^n)$, so

$$\Delta_{S^n} H = -k(k+n-1)H.$$

Recall from Section 16.3 that

$$\Delta_{S^n} f = \frac{1}{\sin^{n-1}\theta}\frac{\partial}{\partial\theta}\left(\sin^{n-1}\theta\,\frac{\partial f}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\Delta_{S_{n-1}} f,$$

in the local coordinates where

$$\sigma = \sin\theta\,\widetilde{\sigma} + \cos\theta\,e_{n+1},$$

with $\widetilde{\sigma} \in S^{n-1}$ and $0 \leq \theta < \pi$. If we make the change of variable $t = \cos\theta$, then it is easy to see that the above formula becomes

$$\Delta_{S^n} f = (1-t^2)\frac{\partial^2 f}{\partial t^2} - nt\frac{\partial f}{\partial t} + \frac{1}{1-t^2}\Delta_{S^{n-1}}$$

(see Morimoto [114], Chapter 2, Theorem 2.9.) But, H being zonal, it only depends on θ, that is, on t, so $\Delta_{S^{n-1}} H = 0$ and thus,

$$-k(k+n-1)P_{k,n}(t) = \Delta_{S^n} P_{k,n}(t) = (1-t^2)\frac{\partial^2 P_{k,n}}{\partial t^2} - nt\frac{\partial P_{k,n}}{\partial t},$$

which yields our equation. □

Note that for $n = 2$, the differential equation of Proposition 16.25 is the Legendre equation from Section 16.2.

The Gegenbauer poynomials also appear as coefficients in some simple generating functions. The following proposition is proved in Morimoto [114] (Chapter 2, Theorem 2.53 and Theorem 2.55):

Proposition 16.26. *For all r and t such that $-1 < r < 1$ and $-1 \leq t \leq 1$, for all $n \geq 1$, we have the following generating formula:*

$$\sum_{k=0}^{\infty} a_{k,n+1}\, r^k P_{k,n}(t) = \frac{1-r^2}{(1-2rt+r^2)^{\frac{n+1}{2}}}.$$

Furthermore, for all r and t such that $0 \leq r < 1$ and $-1 \leq t \leq 1$, if $n = 1$, then

$$\sum_{k=1}^{\infty} \frac{r^k}{k}\, P_{k,1}(t) = -\frac{1}{2}\log(1-2rt+r^2)$$

and if $n \geq 2$*, then*

$$\sum_{k=0}^{\infty} \frac{n-1}{2k+n-1}\, a_{k,n+1}\, r^k P_{k,n}(t) = \frac{1}{(1-2rt+r^2)^{\frac{n-1}{2}}}.$$

In Stein and Weiss [143] (Chapter 4, Section 2), the polynomials, $P_k^\lambda(t)$, where $\lambda > 0$ are defined using the following generating formula:

$$\sum_{k=0}^{\infty} r^k P_k^\lambda(t) = \frac{1}{(1-2rt+r^2)^\lambda}.$$

Each polynomial, $P_k^\lambda(t)$, has degree k and is called an *ultraspherical polynomial of degree* k *associated with* λ. In view of Proposition 16.26, we see that

$$P_k^{\frac{n-1}{2}}(t) = \frac{n-1}{2k+n-1}\, a_{k,n+1}\, P_{k,n}(t),$$

as we mentionned ealier. There is also an integral formula for the Gegenbauer polynomials known as *Laplace representation*, see Morimoto [114] (Chapter 2, Theorem 2.52).

16.8 The Funk-Hecke Formula

The Funk-Hecke Formula (also known as Hecke-Funk Formula) basically allows one to perform a sort of convolution of a "kernel function" with a spherical function in a convenient way. Given a measurable function, K, on $[-1, 1]$ such that the integral

$$\int_{-1}^{1} |K(t)|(1-t^2)^{\frac{n-2}{2}}\, dt$$

makes sense, given a function $f \in L^2_{\mathbb{C}}(S^n)$, we can view the expression

$$K \star f(\sigma) = \int_{S^n} K(\sigma \cdot \tau) f(\tau)\, d\tau$$

as a sort of *convolution* of K and f. Actually, the use of the term convolution is really unfortunate because in a "true" convolution, $f * g$, either the argument of f or the argument of g should be multiplied by the inverse of the variable of integration, which means that the integration should really be taking place over the group $\mathbf{SO}(n+1)$. We will come back to this point later. For the time being, let us call the expression $K \star f$ defined above a *pseudo-convolution*. Now, if f is expressed in terms of spherical harmonics as

$$f = \sum_{k=0}^{\infty} \sum_{m_k=1}^{a_{k,n+1}} c_{k,m_k} Y_k^{m_k},$$

then the Funk-Hecke Formula states that

$$K \star Y_k^{m_k}(\sigma) = \alpha_k Y_k^{m_k}(\sigma),$$

for some fixed constant, α_k, and so

$$K \star f = \sum_{k=0}^{\infty} \sum_{m_k=1}^{a_{k,n+1}} \alpha_k c_{k,m_k} Y_k^{m_k}.$$

Thus, if the constants, α_k are known, then it is "cheap" to compute the pseudo-convolution $K \star f$.

This method was used in a ground-breaking paper by Basri and Jacobs [14] to compute the reflectance function, r, from the lighting function, ℓ, as a pseudo-convolution $K \star \ell$ (over S^2) with the *Lambertian kernel*, K, given by

$$K(\sigma \cdot \tau) = \max(\sigma \cdot \tau, 0).$$

Below, we give a proof of the Funk-Hecke formula due to Morimoto [114] (Chapter 2, Theorem 2.39) but see also Andrews, Askey and Roy [2] (Chapter 9). This formula was first published by Funk in 1916 and then by Hecke in 1918.

Theorem 16.27. *(Funk-Hecke Formula) Given any measurable function, K, on $[-1, 1]$ such that the integral*

$$\int_{-1}^{1} |K(t)|(1-t^2)^{\frac{n-2}{2}}\, dt$$

makes sense, for every function, $H \in \mathcal{H}_k^{\mathbb{C}}(S^n)$, we have

$$\int_{S^n} K(\sigma \cdot \xi) H(\xi)\, d\xi = \left(\mathrm{vol}(S_{n-1}) \int_{-1}^{1} K(t) P_{k,n}(t)(1-t^2)^{\frac{n-2}{2}}\, dt\right) H(\sigma).$$

Observe that when $n = 2$, the term $(1-t^2)^{\frac{n-2}{2}}$ is missing and we are simply requiring that $\int_{-1}^{1} |K(t)|\, dt$ makes sense.

Proof. We first prove the formula in the case where H is a zonal harmonic and then use the fact that the $P_{k,n}$'s are reproducing kernels (formula (rk)).

For all $\sigma, \tau \in S^n$ define H by

$$H(\sigma) = P_{k,n}(\sigma \cdot \tau)$$

and F by

$$F(\sigma, \tau) = \int_{S^n} K(\sigma \cdot \xi) P_{k,n}(\xi \cdot \tau)\, d\xi.$$

Since the volume form on the sphere is invariant under orientation-preserving isometries, for every $R \in \mathbf{SO}(n+1)$, we have

$$F(R\sigma, R\tau) = F(\sigma, \tau).$$

On the other hand, for σ fixed, it is not hard to see that as a function in τ, the function $F(\sigma, -)$ is a spherical harmonic, because $P_{k,n}$ satisfies a differential equation that implies that $\Delta_{S^2} F(\sigma, -) = -k(k+n-1)F(\sigma, -)$. Now, for every rotation, R, that fixes σ,

$$F(\sigma, \tau) = F(R\sigma, R\tau) = F(\sigma, R\tau),$$

which means that $F(\sigma, -)$ satisfies condition (2) of Theorem 16.17. By Theorem 16.17, we get

$$F(\sigma, \tau) = F(\sigma, \sigma) P_{k,n}(\sigma \cdot \tau).$$

If we use local coordinates on S^n where

$$\sigma = \sqrt{1-t^2}\, \widetilde{\sigma} + t\, e_{n+1},$$

with $\widetilde{\sigma} \in S^{n-1}$ and $-1 \leq t \leq 1$, it is not hard to show that the volume form on S^n is given by

$$d\sigma_{S^n} = (1-t^2)^{\frac{n-2}{2}}\, dt d\sigma_{S^{n-1}}.$$

Using this, we have

$$F(\sigma, \sigma) = \int_{S^n} K(\sigma \cdot \xi) P_{k,n}(\xi \cdot \sigma)\, d\xi = \mathrm{vol}(S^{n-1}) \int_{-1}^{1} K(t) P_{k,n}(t)(1-t^2)^{\frac{n-2}{2}}\, dt,$$

and thus,

$$F(\sigma, \tau) = \left(\mathrm{vol}(S^{n-1}) \int_{-1}^{1} K(t) P_{k,n}(t)(1-t^2)^{\frac{n-2}{2}}\, dt \right) P_{k,n}(\sigma \cdot \tau),$$

which is the Funk-Hecke formula when $H(\sigma) = P_{k,n}(\sigma \cdot \tau)$.

Let us now consider any function, $H \in \mathcal{H}_k^{\mathbb{C}}(S^n)$. Recall that by the reproducing kernel property (rk), we have

$$\frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \int_{S^n} P_{k,n}(\xi \cdot \tau) H(\tau)\, d\tau = H(\xi).$$

Then, we can compute $\int_{S^n} K(\sigma \cdot \xi) H(\xi)\, d\xi$ using Fubini's Theorem and the Funk-Hecke

formula in the special case where $H(\sigma) = P_{k,n}(\sigma \cdot \tau)$, as follows:

$$\begin{aligned}
&\int_{S^n} K(\sigma \cdot \xi) H(\xi)\, d\xi \\
&= \int_{S^n} K(\sigma \cdot \xi) \left(\frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \int_{S^n} P_{k,n}(\xi \cdot \tau) H(\tau)\, d\tau \right) d\xi \\
&= \frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \int_{S^n} H(\tau) \left(\int_{S^n} K(\sigma \cdot \xi) P_{k,n}(\xi \cdot \tau)\, d\xi \right) d\tau \\
&= \frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \int_{S^n} H(\tau) \left(\left(\mathrm{vol}(S^{n-1}) \int_{-1}^{1} K(t) P_{k,n}(t) (1-t^2)^{\frac{n-2}{2}}\, dt \right) P_{k,n}(\sigma \cdot \tau) \right) d\tau \\
&= \left(\mathrm{vol}(S^{n-1}) \int_{-1}^{1} K(t) P_{k,n}(t) (1-t^2)^{\frac{n-2}{2}}\, dt \right) \left(\frac{a_{k,n+1}}{\mathrm{vol}(S^n)} \int_{S^n} P_{k,n}(\sigma \cdot \tau) H(\tau)\, d\tau \right) \\
&= \left(\mathrm{vol}(S^{n-1}) \int_{-1}^{1} K(t) P_{k,n}(t) (1-t^2)^{\frac{n-2}{2}}\, dt \right) H(\sigma),
\end{aligned}$$

which proves the Funk-Hecke formula in general. □

The Funk-Hecke formula can be used to derive an "addition theorem" for the ultraspherical polynomials (Gegenbauer polynomials). We omit this topic and we refer the interested reader to Andrews, Askey and Roy [2] (Chapter 9, Section 9.8).

Remark: Oddly, in their computation of $K \star \ell$, Basri and Jacobs [14] first expand K in terms of spherical harmonics as

$$K = \sum_{n=0}^{\infty} k_n Y_n^0,$$

and then use the Funk-Hecke formula to compute $K \star Y_n^m$ and they get (see page 222)

$$K \star Y_n^m = \alpha_n Y_n^m, \quad \text{with} \quad \alpha_n = \sqrt{\frac{4\pi}{2n+1}}\, k_n,$$

for some constant, k_n, given on page 230 of their paper (see below). However, there is no need to expand K as the Funk-Hecke formula yields directly

$$K \star Y_n^m(\sigma) = \int_{S^2} K(\sigma \cdot \xi) Y_n^m(\xi)\, d\xi = \left(\int_{-1}^{1} K(t) P_n(t)\, dt \right) Y_n^m(\sigma),$$

where $P_n(t)$ is the standard Legendre polynomial of degree n since we are in the case of S^2. By the definition of K ($K(t) = \max(t, 0)$) and since $\mathrm{vol}(S^1) = 2\pi$, we get

$$K \star Y_n^m = \left(2\pi \int_{0}^{1} t P_n(t)\, dt \right) Y_n^m,$$

which is equivalent to Basri and Jacobs' formula (14) since their α_n on page 222 is given by

$$\alpha_n = \sqrt{\frac{4\pi}{2n+1}}\, k_n,$$

but from page 230,

$$k_n = \sqrt{(2n+1)\pi} \int_0^1 t P_n(t)\, dt.$$

What remains to be done is to compute $\int_0^1 t P_n(t)\, dt$, which is done by using the Rodrigues Formula and integrating by parts (see Appendix A of Basri and Jacobs [14]).

16.9 Convolution on G/K, for a Gelfand Pair (G, K)

Chapter 17

Discrete Laplacians on Polyhedral Surfaces

Chapter 18

Metrics and Curvature on Lie Groups

18.1 Left (resp. Right) Invariant Metrics

Since a Lie group, G, is a smooth manifold, we can endow G with a Riemannian metric. Among all the Riemannian metrics on a Lie groups, those for which the left translations (or the right translations) are isometries are of particular interest because they take the group structure of G into account. As a consequence, it is possible to find explicit formulae for the Levi-Civita connection and the various curvatures, especially in the case of metrics which are both left and right-invariant. This chapter makes extensive use of results from a beautiful paper of Milnor [110].

Definition 18.1. A metric, $\langle -, - \rangle$, on a Lie group, G, is called *left-invariant* (resp. *right-invariant*) iff

$$\langle u, v\rangle_b = \langle (dL_a)_b u, (dL_a)_b v\rangle_{ab},$$

(resp.

$$\langle u, v\rangle_b = \langle (dR_a)_b u, (dR_a)_b v\rangle_{ba}),$$

or all $a, b \in G$ and all $u, v \in T_bG$. A Riemannian metric that is both left and right-invariant is called a *bi-invariant metric*.

As shown in the next proposition, left-invariant (resp. right-invariant) metrics on G are induced by inner products on the Lie algebra, $\mathfrak{g}$, of G. In the sequel, the identity element of the Lie group, G, will be denoted by e or 1.

Proposition 18.1. *There is a bijective correspondence between left-invariant (resp. right invariant) metrics on a Lie group, G, and inner products on the Lie algebra, $\mathfrak{g}$, of G.*

Proof. If the metric on G is left-invariant, then for all $a \in G$ and all $u, v \in T_aG$, we have

$$\begin{aligned}
\langle u, v\rangle_a &= \langle d(L_a \circ L_{a^{-1}})_a u, d(L_a \circ L_{a^{-1}})_a v\rangle_a \\
&= \langle (dL_a)_e((dL_{a^{-1}})_a u), (dL_a)_e((dL_{a^{-1}})_a v)\rangle_a \\
&= \langle (dL_{a^{-1}})_a u, (dL_{a^{-1}})_a v\rangle_e,
\end{aligned}$$

which shows that our metric is completely determined by its restriction to $\mathfrak{g} = T_eG$. Conversely, let $\langle -, - \rangle$ be an inner product on $\mathfrak{g}$ and set

$$\langle u, v \rangle_g = \langle (dL_{g^{-1}})_g u, (dL_{g^{-1}})_g v \rangle,$$

for all $u, v \in T_gG$ and all $g \in G$. Obviously, the family of inner products, $\langle -, - \rangle_g$, yields a Riemannian metric on G. To prove that it is left-invariant, we use the chain rule and the fact that left translations are group isomorphisms. For all $a, b \in G$ and all $u, v \in T_bG$, we have

$$\begin{aligned}
\langle (dL_a)_b u, (dL_a)_b v \rangle_{ab} &= \langle (dL_{(ab)^{-1}})_{ab}((dL_a)_b u), (dL_{(ab)^{-1}})_{ab}((dL_a)_b v) \rangle \\
&= \langle d(L_{(ab)^{-1}} \circ L_a)_b u, d(L_{(ab)^{-1}} \circ L_a)_b v \rangle \\
&= \langle d(L_{b^{-1}a^{-1}} \circ L_a)_b u, d(L_{b^{-1}a^{-1}} \circ L_a)_b v \rangle \\
&= \langle (dL_{b^{-1}})_b u, (dL_{b^{-1}})_b v \rangle \\
&= \langle u, v \rangle_b,
\end{aligned}$$

as desired.

To get a right-invariant metric on G, set

$$\langle u, v \rangle_g = \langle (dR_{g^{-1}})_g u, (dR_{g^{-1}})_g v \rangle,$$

for all $u, v \in T_gG$ and all $g \in G$. The verification that this metric is right-invariant is analogous. □

If G has dimension n, then since inner products on $\mathfrak{g}$ are in one-to-one correspondence with $n \times n$ positive definite matrices, we see that G possesses a family of left-invariant metrics of dimension $\frac{1}{2}\, n(n+1)$.

If G has a left-invariant (resp. right-invariant) metric, since left-invariant (resp. right-invariant) translations are isometries and act transitively on G, the space G is called a *homogeneous Riemannian manifold*.

Proposition 18.2. *Every Lie group, G, equipped with a left-invariant (resp. right-invariant) metric is complete.*

Proof. As G is locally compact, we can pick some $\epsilon > 0$ small enough so that the closed ϵ-ball about the identity is compact. By translation, every ϵ-ball is compact, hence every Cauchy sequence eventually lies within a compact set and thus, converges. □

We now give several characterizations of bi-invariant metrics.

18.2 Bi-Invariant Metrics

Recall that the adjoint representation, $\mathrm{Ad}\colon G \to \mathrm{GL}(\mathfrak{g})$, is the map defined such that $\mathrm{Ad}_a\colon \mathfrak{g} \to \mathfrak{g}$ is the linear isomorphism given by

$$\mathrm{Ad}_a = d(R_{a^{-1}} \circ L_a)_e,$$

for every $a \in G$. Clearly,

$$\mathrm{Ad}_a = (dR_{a^{-1}})_a \circ (dL_a)_e.$$

Here is the first of four criteria for the existence of a bi-invariant metric on a Lie group.

Proposition 18.3. *There is a bijective correspondence between bi-invariant metrics on a Lie group, G, and* Ad*-invariant inner products on the Lie algebra, $\mathfrak{g}$, of G, that is, inner products, $\langle -, - \rangle$, on $\mathfrak{g}$ such that Ad_a is an isometry of $\mathfrak{g}$ for all $a \in G$; more explicitly, inner products such that*

$$\langle \mathrm{Ad}_a u, \mathrm{Ad}_a v \rangle = \langle u, v \rangle,$$

for all $a \in G$ and all $u, v \in \mathfrak{g}$.

Proof. If $\langle -, - \rangle$ is a bi-invariant metric on G, as

$$\mathrm{Ad}_a = (dR_{a^{-1}})_a \circ (dL_a)_e.$$

it is clear that Ad_a is an isometry on $\mathfrak{g}$.

Conversely, if $\langle -, - \rangle$ is any inner product on $\mathfrak{g}$ such that Ad_a is an isometry of $\mathfrak{g}$ for all $a \in G$, we need to prove that the metric on G given by

$$\langle u, v \rangle_g = \langle (dL_{g^{-1}})_g u, (dL_{g^{-1}})_g v \rangle$$

is also right-invariant. We have

$$\begin{aligned}
\langle (dR_a)_b u, (dR_a)_b v \rangle_{ba} &= \langle (dL_{(ba)^{-1}})_{ba}((dR_a)_b u), (dL_{(ba)^{-1}})_{ba}((dR_a)_b v) \rangle \\
&= \langle d(L_{a^{-1}} \circ L_{b^{-1}} \circ R_a)_b u, d(L_{a^{-1}} \circ L_{b^{-1}} \circ R_a)_b v \rangle \\
&= \langle d(R_a \circ L_{a^{-1}} \circ L_{b^{-1}})_b u, d(R_a \circ L_{a^{-1}} \circ L_{b^{-1}})_b v \rangle \\
&= \langle d(R_a \circ L_{a^{-1}})_e \circ d(L_{b^{-1}})_b u, d(R_a \circ L_{a^{-1}})_e \circ d(L_{b^{-1}})_b v \rangle \\
&= \langle \mathrm{Ad}_{a^{-1}} \circ d(L_{b^{-1}})_b u, \mathrm{Ad}_{a^{-1}} \circ d(L_{b^{-1}})_b v \rangle \\
&= \langle u, v \rangle,
\end{aligned}$$

as $\langle -, - \rangle$ is left-invariant and Ag_g-invariant for all $g \in G$. □

Proposition 18.3 shows that if a Lie group, G, possesses a bi-invariant metric, then every linear map, Ad_a, is an orthogonal transformation of $\mathfrak{g}$. It follows that $\mathrm{Ad}(G)$ is a subgroup of the orthogonal group of $\mathfrak{g}$ and so, its closure, $\overline{\mathrm{Ad}(G)}$, is compact. It turns out that this condition is also sufficient!

To prove the above fact, we make use of an "averaging trick" used in representation theory. Recall that a *representation* of a Lie group, G, is a (smooth) homomorphism, $\rho\colon G \to \mathrm{GL}(V)$, where V is some finite-dimensional vector space. For any $g \in G$ and any $u \in V$, we often write $g \cdot u$ for $\rho(g)(u)$. We say that an inner-product, $\langle -, - \rangle$, on V is *G-invariant* iff

$$\langle g \cdot u, g \cdot v \rangle = \langle u, v \rangle,$$

for all $g \in G$ and all $u, v \in V$. If G is compact, then the "averaging trick", also called "Weyl's unitarian trick", yields the following important result:

Theorem 18.4. *If G is a compact Lie group, then for every representation, $\rho\colon G \to \mathrm{GL}(V)$, there is a G-invariant inner product on V.*

Proof. Recall from Section 9.4 that as a Lie group is orientable, it has a left-invariant volume form, ω, and for every continuous function, f, with compact support, we can define the integral, $\int_M f = \int_G f\omega$. Furthermore, when G is compact, we may assume that our integral is normalized so that $\int_G \omega = 1$ and in this case, our integral is both left and right invariant. Now, given any inner product, $\langle -, - \rangle$ on V, set

$$\langle\langle u, v \rangle\rangle = \int_G \langle g \cdot u, g \cdot v \rangle,$$

for all $u, v \in V$, where $\langle g \cdot u, g \cdot v \rangle$ denotes the function $g \mapsto \langle g \cdot u, g \cdot v \rangle$. It is easily checked that $\langle\langle -, - \rangle\rangle$ is an inner product on V. Furthermore, using the right-invariance of our integral (that is, $\int_G f = \int_G (R_h \circ f)$, for all $h \in G$), we have

$$\begin{aligned}
\langle\langle h \cdot u, h \cdot v \rangle\rangle &= \int_G \langle g \cdot (h \cdot u), g \cdot (h \cdot v) \rangle \\
&= \int_G \langle (gh) \cdot u, (gh) \cdot v \rangle \\
&= \int_G \langle g \cdot u, g \cdot v \rangle \\
&= \langle\langle u, v \rangle\rangle,
\end{aligned}$$

which shows that $\langle\langle -, - \rangle\rangle$ is G-invariant. □

Using Theorem 18.4, we can prove the following result giving a criterion for the existence of a G-invariant inner product for any representation of a Lie group, G (see Sternberg [144], Chapter 5, Theorem 5.2).

Theorem 18.5. *Let $\rho\colon G \to \mathrm{GL}(V)$ be a (finite-dimensional) representation of a Lie group, G. There is a G-invariant inner product on V iff $\overline{\rho(G)}$ is compact. In particular, if G is compact, then there is a G-invariant inner product on V.*

Proof. If V has a G-invariant inner product on V, then each linear map, $\rho(g)$, is an isometry, so $\rho(G)$ is a subgroup of the orthogonal group, $\mathbf{O}(V)$, of V. As $\mathbf{O}(V)$ is compact, $\overline{\rho(G)}$ is also compact.

Conversely, assume that $\overline{\rho(G)}$ is compact. In this case, $H = \overline{\rho(G)}$ is a closed subgroup of the lie group, $\mathrm{GL}(V)$, so by Theorem 5.12, H is a compact Lie subgroup of $\mathrm{GL}(V)$. Now, the inclusion homomorphism, $H \hookrightarrow \mathrm{GL}(V)$, is a representation of H ($f \cdot u = f(u)$, for all $f \in H$ and all $u \in V$), so by Theorem 18.4, there is an inner product on V which is H-invariant. However, for any $g \in G$, if we write $f = \rho(g) \in H$, then we have

$$\langle g \cdot u, g \cdot v\rangle = \langle f(u), f(v)\rangle = \langle u, v\rangle,$$

proving that $\langle -, -\rangle$ is G-invariant as well. □

Applying Theorem 18.5 to the adjoint representation, $\mathrm{Ad}\colon G \to \mathrm{GL}(\mathfrak{g})$, we get our second criterion for the existence of a bi-invariant metric on a Lie group.

Proposition 18.6. *Given any Lie group, G, an inner product, $\langle -, -\rangle$, on $\mathfrak{g}$ induces a bi-invariant metric on G iff $\overline{\mathrm{Ad}(G)}$ is compact. In particular, every compact Lie group has a bi-invariant metric.*

Proof. Proposition 18.3 is equivalent to the fact that G possesses a bi-invariant metric iff there is some Ad-invariant inner product on $\mathfrak{g}$. By Theorem 18.5, there is some Ad-invariant inner product on $\mathfrak{g}$ iff $\overline{\mathrm{Ad}(G)}$ is compact, which is the statement of our theorem. □

Proposition 18.6 can be used to prove that certain Lie groups do not have a bi-invariant metric. For example, Arsigny, Pennec and Ayache use Proposition 18.6 to give a short and elegant proof of the fact that $\mathbf{SE}(n)$ does not have any bi-invariant metric for all $n \geq 1$. As noted by these authors, other proofs found in the literature are a lot more complicated and only cover the case $n = 3$.

Recall the adjoint representation of $\mathfrak{g}$,

$$\mathrm{ad}\colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g}),$$

given by $\mathrm{ad} = d\mathrm{Ad}_1$. Here is our third criterion for the existence of a bi-invariant metric on a connected Lie group.

Proposition 18.7. *If G is a connected Lie group, an inner product, $\langle -, -\rangle$, on $\mathfrak{g}$ induces a bi-invariant metric on G iff the linear map, $\mathrm{ad}(u)\colon \mathfrak{g} \to \mathfrak{g}$, is skew-adjoint for all $u \in \mathfrak{g}$, which means that*

$$\langle \mathrm{ad}(u)(v), w\rangle = -\langle v, \mathrm{ad}(u)(w)\rangle$$

for all $u, v, w \in \mathfrak{g}$ iff

$$\langle [x, y], z\rangle = \langle x, [y, z]\rangle$$

for all $x, y, z \in \mathfrak{g}$.

Proof. We follow Milnor [110], Lemma 7.2. By Proposition 18.3, an inner product on $\mathfrak{g}$ induces a bi-invariant metric on G iff Ad_g is an isometry for all $g \in G$. We know that we can choose a small enough open subset, U, of $\mathfrak{g}$ containing 0 so that $\exp\colon \mathfrak{g} \to G$ is a diffeomorphism from U to $\exp(U)$. For any $g \in \exp(U)$, there is a unique, $u \in \mathfrak{g}$, so that $g = \exp(u)$. By Proposition 5.6,

$$\mathrm{Ad}(g) = \mathrm{Ad}(\exp(u)) = e^{\mathrm{ad}(u)}.$$

Now, $\mathrm{Ad}(g)$ is an isometry iff $\mathrm{Ad}(g)^{-1} = \mathrm{Ad}(g)^*$, where $\mathrm{Ad}(g)^*$ denotes the adjoint of $\mathrm{Ad}(g)$ and we know that

$$\mathrm{Ad}(g)^{-1} = e^{-\mathrm{ad}(u)} \quad \text{and} \quad \mathrm{Ad}(g)^* = e^{\mathrm{ad}(u)^*},$$

so we deduce that $\mathrm{Ad}(g)^{-1} = \mathrm{Ad}(g)^*$ iff

$$\mathrm{ad}(u)^* = -\mathrm{ad}(u),$$

which means that $\mathrm{ad}(u)$ is skew-adjoint. Since a connected Lie group is generated by any open subset containing the identity and since products of isometries are isometries, our results holds for all $g \in G$.

The skew-adjointness of $\mathrm{ad}(u)$ means that

$$\langle \mathrm{ad}(u)(v), w\rangle = -\langle v, \mathrm{ad}(u)(w)\rangle$$

for all $u, v, w \in \mathfrak{g}$ and since $\mathrm{ad}(u)(v) = [u, v]$ and $[u, v] = -[v, u]$, we get

$$\langle [v, u], w\rangle = \langle v, [u, w]\rangle$$

which is the last claim of the proposition after renaming u, v, w as y, x, z. □

It will be convenient to say that an inner product on $\mathfrak{g}$ is *bi-invariant* iff every $\mathrm{ad}(u)$ is skew-adjoint.

If G is a connected Lie group, then the existence of a bi-invariant metric on G places a heavy restriction on its group structure as shown by the following result from Milnor's paper [110] (Lemma 7.5):

Theorem 18.8. *A connected Lie group, G, admits a bi-invariant metric iff it is isomorphic to the cartesian product of a compact group and a vector space ($\mathbb{R}^m$, for some $m \geq 0$).*

A proof of Theorem 18.8 can be found in Milnor [110] (Lemma 7.4 and Lemma 7.5). The proof uses the universal covering group and it is a bit involved. We will outline the structure of the proof because it is really quite beautiful.

In a first step, it is shown that if G has a bi-invariant metric, then its Lie algebra, $\mathfrak{g}$, can be written as an orthogonal coproduct

$$\mathfrak{g} = \mathfrak{g}_1 \oplus \cdots \oplus \mathfrak{g}_k,$$

where each $\mathfrak{g}_i$ is either a simple ideal or a one-dimensional abelian ideal, that is, $\mathfrak{g}_i \cong \mathbb{R}$.

First, a few definitions.

Definition 18.2. A subset, $\mathfrak{h}$, of a Lie algebra, $\mathfrak{g}$, is a *Lie subalgebra* iff it is a subspace of $\mathfrak{g}$ (as a vector space) and if it is closed under the bracket operation on $\mathfrak{g}$. A subalgebra, $\mathfrak{h}$, is *abelian* iff $[x, y] = 0$ for all $x, y \in \mathfrak{h}$. An *ideal* in $\mathfrak{g}$ is a Lie subalgebra, $\mathfrak{h}$, such that

$$[h, g] \in \mathfrak{h}, \qquad \text{for all } h \in \mathfrak{h} \text{ and all } g \in \mathfrak{g}.$$

The *center*, $Z(\mathfrak{g})$, of a Lie algebra, $\mathfrak{g}$, is the set of all elements, $u \in \mathfrak{g}$, so that $[u, v] = 0$ for all $v \in \mathfrak{g}$, or equivalently, so that $\mathrm{ad}(u) = 0$. A Lie algebra, $\mathfrak{g}$, is *simple* iff it is non-abelian and if it has no ideal other than (0) and $\mathfrak{g}$. A Lie algebra, $\mathfrak{g}$, is *semisimple* iff it has no abelian ideal other than (0). A Lie group is *simple* (resp. *semisimple*) iff its Lie algebra is simple (resp. *semisimple*)

Clearly, the trivial subalgebras (0) and $\mathfrak{g}$ itself are ideals and the center is an abelian ideal.

Note that, by definition, simple and semisimple Lie algebras are non-abelian and a simple algebra is a semisimple algebra. It turns out that a Lie algebra, $\mathfrak{g}$, is semisimple iff it can be expressed as a direct sum of ideals, $\mathfrak{g}_i$, with each $\mathfrak{g}_i$ a simple algebra (see Knapp [90], Chapter I, Theorem 1.54). If we drop the requirement that a simple Lie algebra be non-abelian, thereby allowing one dimensional Lie algebras to be simple, we run into the trouble that a simple Lie algebra is no longer semisimple and the above theorem fails for this stupid reason. Thus, it seems technically advantageous to require that simple Lie algebras be non-abelian.

Nevertheless, in certain situations, it is desirable to drop the requirement that a simple Lie algebra be non-abelian and this is what Milnor does in his paper because it is more convenient for one of his proofs. This is a minor point but it could be confusing for uninitiated readers.

The next step is to lift the ideals, $\mathfrak{g}_i$, to the simply connected normal subgroups, G_i, of the universal covering group, $\widetilde{G}$, of $\mathfrak{g}$. For every simple ideal, $\mathfrak{g}_i$, in the decomposition it is proved that there is some constant, $c_i > 0$, so that all Ricci curvatures are strictly positive and bounded from below by c_i. Therefore, by Myers' Theorem (Theorem 13.28), G_i is compact. It follows that $\widetilde{G}$ is isomorphic to a product of compact simple Lie groups and some vector space, $\mathbb{R}^m$. Finally, we know that G is isomorphic to the quotient of $\widetilde{G}$ by a discrete normal subgroup of $\widetilde{G}$, which yields our theorem.

Because it is a fun proof, we prove the statement about the structure of a Lie algebra for which each $\mathrm{ad}(u)$ is skew-adjoint.

Proposition 18.9. *Let $\mathfrak{g}$ be a Lie algebra with an inner product such that the linear map, $\mathrm{ad}(u)$, is skew-adjoint for every $u \in \mathfrak{g}$. The orthogonal complement, $\mathfrak{a}^{\perp}$, of any ideal, $\mathfrak{a}$, is itself an ideal. Consequently, $\mathfrak{g}$ can be expressed as an orthogonal direct sum*

$$\mathfrak{g} = \mathfrak{g}_1 \oplus \cdots \oplus \mathfrak{g}_k,$$

where each $\mathfrak{g}_i$ is either a simple ideal or a one-dimensional abelian ideal, that is, $\mathfrak{g}_i \cong \mathbb{R}$.

Proof. Assume $u \in \mathfrak{g}$ is orthogonal to $\mathfrak{a}$. We need to prove that $[u, v]$ is orthogonal to $\mathfrak{a}$ for all $v \in \mathfrak{g}$. But, as $\mathrm{ad}(u)$ is skew-adjoint, $\mathrm{ad}(u)(v) = [u, v]$, and $\mathfrak{a}$ is an ideal, we have

$$\langle [u, v], a\rangle = -\langle u, [v, a]\rangle = 0, \qquad \text{for all } a \in \mathfrak{a},$$

which shows that $\mathfrak{a}^\perp$ is an ideal.

For the second statement, we use induction on the dimension of $\mathfrak{g}$ *but for this proof, we redefine a simple Lie algebra to be an algebra with no nontrivial proper ideals*. The case where $\dim \mathfrak{g} = 1$ is clear.

For the induction step, if $\mathfrak{g}$ is simple, we are done. Else, $\mathfrak{g}$ has some nontrivial proper ideal, $\mathfrak{h}$, and if we pick $\mathfrak{h}$ of minimal dimension, p, with $1 \leq p < n = \dim \mathfrak{g}$, then $\mathfrak{h}$ is simple. Now, $\mathfrak{h}^\perp$ is also an ideal and $\dim \mathfrak{h}^\perp < n$, so the induction hypothesis applies. Therefore, we have an orthogonal direct sum

$$\mathfrak{g} = \mathfrak{g}_1 \oplus \cdots \oplus \mathfrak{g}_k,$$

where each $\mathfrak{g}_i$ is simple *in our relaxed sense*. However, if $\mathfrak{g}_i$ is not abelian, then it is simple in the usual sense and if $\mathfrak{g}_i$ is abelian, having no proper nontrivial ideal, it must be one-dimensional and we get our decomposition. □

We now investigate connections and curvature on Lie groups with a left-invariant metric.

18.3 Connections and Curvature of Left-Invariant Metrics on Lie Groups

If G is a Lie group equipped with a left-invariant metric, then it is possible to express the Levi-Civita connection and the sectional curvature in terms of quantities defined over the Lie algebra of G, at least for left-invariant vector fields. When the metric is bi-invariant, much nicer formulae can be obtained.

If $\langle -, -\rangle$ is a left-invariant metric on G, then for any two left-invariant vector fields, X, Y, we have

$$\langle X, Y\rangle_g = \langle X(g), Y(g)\rangle_g = \langle (dL_g)_e X(e), (dL_g)_e Y(e)\rangle_e = \langle X_e, Y_e\rangle_e = \langle X, Y\rangle_e,$$

which shows that the function, $g \mapsto \langle X, Y\rangle_g$, is constant. Therefore, for any vector field, Z,

$$Z(\langle X, Y\rangle) = 0.$$

If we go back to the Koszul formula (Proposition 11.18)

$$\begin{aligned} 2\langle \nabla_X Y, Z\rangle &= X(\langle Y, Z\rangle) + Y(\langle X, Z\rangle) - Z(\langle X, Y\rangle) \\ &\quad - \langle Y, [X, Z]\rangle - \langle X, [Y, Z]\rangle - \langle Z, [Y, X]\rangle, \end{aligned}$$

we deduce that for all left-invariant vector fields, X, Y, Z, we have

$$2\langle \nabla_X Y, Z\rangle = -\langle Y, [X, Z]\rangle - \langle X, [Y, Z]\rangle - \langle Z, [Y, X]\rangle,$$

which can be rewritten as

$$2\langle \nabla_X Y, Z\rangle = \langle [X, Y], Z\rangle - \langle [Y, Z], X\rangle + \langle [Z, X], Y\rangle. \tag{$\dagger$}$$

The above yields the formula

$$\nabla_u v = \frac{1}{2}\left([u, v] - \mathrm{ad}(u)^* v - \mathrm{ad}(v)^* u\right), \qquad u, v \in \mathfrak{g},$$

where $\mathrm{ad}(x)^*$ denotes the adjoint of $\mathrm{ad}(x)$.

Following Milnor, if we pick an orthonormal basis, $(e_1, \ldots, e_n)$, *w.r.t.* our inner product on $\mathfrak{g}$ and if we define the constants, α_{ijk}, by

$$\alpha_{ijk} = \langle [e_i, e_j], e_k\rangle,$$

we see that

$$\nabla_{e_i} e_j = \frac{1}{2}\sum_k (\alpha_{ijk} - \alpha_{jki} + \alpha_{kij}) e_k. \tag{$*$}$$

Now, for orthonormal vectors, u, v, the sectional curvature is given by

$$K(u, v) = \langle R(u, v)u, v\rangle,$$

with

$$R(u, v) = \nabla_{[u,v]} - \nabla_u \nabla_v + \nabla_v \nabla_u.$$

If we plug the expressions from equation $(*)$ into the defintions we obtain the following proposition from Milnor [110] (Lemma 1.1):

Proposition 18.10. *Given a Lie group, G, equipped with a left-invariant metric, for any orthonormal basis, $(e_1, \ldots, e_n)$, of $\mathfrak{g}$ and with the structure constants, $\alpha_{ijk} = \langle [e_i, e_j], e_k\rangle$, the sectional curvature, $K(e_1, e_2)$, is given by*

$$\begin{aligned} K(e_1, e_2) \;=\; & \sum_k \frac{1}{2}(\alpha_{12k}(-\alpha_{12k} + \alpha_{2k1} + \alpha_{k12}) \\ & -\frac{1}{4}(\alpha_{12k} - \alpha_{2k1} + \alpha_{k12})(\alpha_{12k} + \alpha_{2k1} - \alpha_{k12}) - \alpha_{k11}\alpha_{k22}). \end{aligned}$$

Although the above formula is not too useful in general, in some cases of interest, a great deal of cancellation takes place so that a more useful formula can be obtained. An example of this situation is provided by the next proposition (Milnor [110], Lemma 1.2).

Proposition 18.11. *Given a Lie group, G, equipped with a left-invariant metric, for any $u \in \mathfrak{g}$, if the linear map,* $\mathrm{ad}(u)$*, is self-adjoint then*

$$K(u, v) \geq 0$$

for all $v \in \mathfrak{g}$, where equality holds iff u is orthogonal to $[v, \mathfrak{g}] = \{[v, x] \mid x \in \mathfrak{g}\}$.

Proof. We may asssume that u and v are orthonormal. If we pick an orthonormal basis such that $e_1 = u$ and $e_2 = v$, the fact that $\mathrm{ad}(e_1)$ is skew-adjoint means that the array (α_{1jk}) is skew-symmetric (in the indices j and k). It follows that the formula of Proposition 18.10 reduces to

$$K(e_1, e_2) = \frac{1}{4} \sum_k \alpha_{2k1}^2,$$

so $K(e_1, e_2) \geq 0$, as claimed. Furthermore, $K(e_1, e_2) = 0$ iff $\alpha_{2k1} = 0$ for $k = 1, \ldots, n$, that is $\langle [e_2, e_k], e_1 \rangle = 0$ for $k = 1, \ldots, n$, which means that e_1 is orthogonal to $[e_2, \mathfrak{g}]$. □

Proposition 18.12. *Given a Lie group, G, equipped with a left-invariant metric, for any u in the center, $Z(\mathfrak{g})$, of $\mathfrak{g}$,*

$$K(u, v) \geq 0$$

for all $v \in \mathfrak{g}$.

Proof. For any element, u, in the center of $\mathfrak{g}$, we have $\mathrm{ad}(u) = 0$, and the zero map is obviously skew-adjoint. □

Recall that the Ricci curvature, $\mathrm{Ric}(u, v)$, is the trace of the linear map, $y \mapsto R(u, y)v$. With respect to any orthonormal basis, $(e_1, \ldots, e_n)$, of $\mathfrak{g}$, we have

$$\mathrm{Ric}(u, v) = \sum_{j=1}^{n} \langle R(u, e_j)v, e_j \rangle = \sum_{j=1}^{n} R(u, e_j, v, e_j).$$

The Ricci curvature is a symmetric form, so it is completely determined by the quadratic form

$$r(u) = \mathrm{Ric}(u, u) = \sum_{j=1}^{n} R(u, e_j, u, e_j).$$

When u is a unit vector, $r(u)$ is called the *Ricci curvature in the direction u*. If we pick an orthonormal basis such that $e_1 = u$, then

$$r(e_1) = \sum_{i=2}^{n} K(e_1, e_i).$$

For computational purposes it may be more convenient to introduce the *Ricci transformation*, $\widehat{r}$, defined by

$$\widehat{r}(x) = \sum_{i=1}^{n} R(e_i, x)e_i.$$

The Ricci transformation is self-adjoint and it is also the unique map so that

$$r(x) = \langle \widehat{r}(x), x\rangle, \qquad \text{for all } x \in \mathfrak{g}.$$

The eigenvalues of $\widehat{r}$ are called the *principal Ricci curvatures*.

Proposition 18.13. *Given a Lie group, G, equipped with a left-invariant metric, if the linear map, $\mathrm{ad}(u)$, is skew-adjoint, then $r(u) \geq 0$, where equality holds iff u is orthogonal to the commutator ideal, $[\mathfrak{g}, \mathfrak{g}]$.*

Proof. This follows from Proposition 18.11. □

In particular, if u is in the center of $\mathfrak{g}$, then $r(u) \geq 0$.

As a corollary of Proposition 18.13, we have the following result which is used in the proof of Theorem 18.8:

Proposition 18.14. *If G is a connected Lie group equipped with a bi-invariant metric and if the Lie algebra of G is simple, then there is a constant, $c > 0$, so that $r(u) \geq c$ for all unit vector, $u \in T_gG$, for all $g \in G$.*

Proof. First of all, the linear maps, $\mathrm{ad}(u)$, are skew-adjoint for all $u \in \mathfrak{g}$, which implies that $r(u) \geq 0$. As $\mathfrak{g}$ is simple, the commutator ideal, $[\mathfrak{g}, \mathfrak{g}]$ is either (0) or $\mathfrak{g}$. But, if $[\mathfrak{g}, \mathfrak{g}] = (0)$, then then $\mathfrak{g}$ is abelian, which is impossible since $\mathfrak{g}$ is simple. Therefore $[\mathfrak{g}, \mathfrak{g}] = \mathfrak{g}$, which implies $r(u) > 0$ for all $u \neq 0$ (otherwise, u would be orthogonal to $[\mathfrak{g}, \mathfrak{g}] = \mathfrak{g}$, which is impossible). As the set of unit vectors in $\mathfrak{g}$ is compact, the function, $u \mapsto r(u)$, achieves it minimum, c, and $c > 0$ as $r(u) > 0$ for all $u \neq 0$. But, $dL_g \colon \mathfrak{g} \to T_gG$ is an isometry for all $g \in G$, so $r(u) \geq c$ for all unit vectors, $u \in T_gG$, for all $g \in G$. □

By Myers' Theorem (Theorem 13.28), the Lie group G is compact and has a finite fundamental group.

The following interesting theorem is proved in Milnor (Milnor [110], Theorem 2.2):

Theorem 18.15. *A connected Lie group, G, admits a left-invariant metric with $r(u) > 0$ for all unit vectors $u \in \mathfrak{g}$ (all Ricci curvatures are strictly positive) iff G is compact and has finite fundamental group.*

The following criterion for obtaining a direction of negative curvature is also proved in Milnor (Milnor [110], Lemma 2.3):

Proposition 18.16. *Given a Lie group, G, equipped with a left-invariant metric, if u is orthogonal to the commutator ideal, $[\mathfrak{g}, \mathfrak{g}]$, then $r(u) \leq 0$, where equality holds iff $\mathrm{ad}(u)$ is self-adjoint.*

When G possesses a bi-invariant metric, much nicer formulae are obtained. First of all, as

$$\langle [u,v], w\rangle = \langle u, [v,w]\rangle,$$

the last two terms in equation (†) cancel out and we get

$$\nabla_u v = \frac{1}{2}\,[u,v],$$

for all $u, v \in \mathfrak{g}$. Then, we get

$$R(u,v) = \frac{1}{2}\mathrm{ad}([u,v]) - \frac{1}{4}\mathrm{ad}(u)\mathrm{ad}(v) + \frac{1}{4}\mathrm{ad}(v)\mathrm{ad}(u).$$

Using the Jacobi identity,

$$\mathrm{ad}([u,v]) = \mathrm{ad}(u)\mathrm{ad}(v) - \mathrm{ad}(v)\mathrm{ad}(u),$$

we get

$$R(u,v) = \frac{1}{4}\,\mathrm{ad}[u,v],$$

so

$$R(u,v)w = \frac{1}{4}\,[[u,v],w].$$

Hence, for unit orthogonal vectors, u, v, the sectional curvature, $K(u,v) = \langle R(u,v)u, v\rangle$, is given by

$$K(u,v) = \frac{1}{4}\,\langle [[u,v],u], v\rangle,$$

which (as $\langle [x,y],z\rangle = \langle x,[y,z]\rangle$) is rewritten as

$$K(u,v) = \frac{1}{4}\,\langle [u,v],[u,v]\rangle.$$

To compute the Ricci curvature, $\mathrm{Ric}(u,v)$, we observe that $\mathrm{Ric}(u,v)$ is the trace of the linear map,

$$y \mapsto R(u,y)v = \frac{1}{4}\,[[u,y],v] = -\frac{1}{4}\,[v,[u,y]] = -\frac{1}{4}\,\mathrm{ad}(v)\circ\mathrm{ad}(u)(y).$$

However, the bilinear form, B, on $\mathfrak{g}$, given by

$$B(u,v) = \mathrm{tr}(\mathrm{ad}(u)\circ\mathrm{ad}(v))$$

is a famous object known as the *Killing form* of the Lie algebra $\mathfrak{g}$. We will take a closer look at the Killing form shortly. For the time being, we observe that as $\mathrm{tr}(\mathrm{ad}(u)\circ\mathrm{ad}(v)) = \mathrm{tr}(\mathrm{ad}(v)\circ\mathrm{ad}(u))$, we get

$$\mathrm{Ric}(u,v) = -\frac{1}{4}\,B(u,v),$$

for all $u, v \in \mathfrak{g}$.

We summarize all this in

Proposition 18.17. *For any Lie group, G, equipped with a left-invariant metric, the following properties hold:*

(a) The connection, $\nabla_u v$, is given by

$$\nabla_u v = \frac{1}{2}\,[u, v], \qquad \text{for all } u, v \in \mathfrak{g}$$

(b) The curvature tensor, $R(u, v)$, is given by

$$R(u, v) = \frac{1}{4}\,\mathrm{ad}[u, v], \qquad \text{for all } u, v \in \mathfrak{g},$$

or equivalently,

$$R(u, v)w = \frac{1}{4}\,[[u, v], w], \qquad \text{for all } u, v, w \in \mathfrak{g}.$$

(c) The sectional curvature, $K(u, v)$, is given by

$$K(u, v) = \frac{1}{4}\,\langle [u, v], [u, v]\rangle,$$

for all pairs of orthonormal vectors, $u, v \in \mathfrak{g}$.

(d) The Ricci curvature, $\mathrm{Ric}(u, v)$, is given by

$$\mathrm{Ric}(u, v) = -\frac{1}{4}\,B(u, v), \qquad \text{for all } u, v \in \mathfrak{g},$$

where B is the Killing form, with

$$B(u, v) = \mathrm{tr}(\mathrm{ad}(u) \circ \mathrm{ad}(v)), \qquad \text{for all } u, v \in \mathfrak{g}.$$

Consequently, $K(u, v) \geq 0$, with equality iff $[u, v] = 0$ and $r(u) \geq 0$, with equality iff u belongs to the center of $\mathfrak{g}$.

Remark: Proposition 18.17 shows that if a Lie group admits a bi-invariant metric, then its Killing form is negative semi-definite.

What are the geodesics in a Lie group equipped with a bi-invariant metric? The answer is simple: they are the integral curves of left-invariant vector fields.

Proposition 18.18. *For any Lie group, G, equipped with a bi-invariant metric, we have:*

(1) The inversion map, $\iota\colon g \mapsto g^{-1}$, is an isometry.

(2) For every $a \in G$, if I_a denotes the map given by

$$I_a(b) = ab^{-1}a, \qquad \text{for all } a, b \in G,$$

then I_a is an isometry fixing a which reverses geodesics, that is, for every geodesic, γ, through a

$$I_a(\gamma)(t) = \gamma(-t).$$

(3) The geodesics through e are the integral curves, $t \mapsto \exp(tu)$, where $u \in \mathfrak{g}$, that is, the one-parameter groups. Consequently, the Lie group exponential map, $\exp\colon \mathfrak{g} \to G$, coincides with the Riemannian exponential map (at e) from T_eG to G, where G is viewed as a Riemannian manifold.

Proof. (1) Since

$$\iota(g) = g^{-1} = g^{-1}h^{-1}h = (hg)^{-1}h = (R_h \circ \iota \circ L_h)(g),$$

we have

$$\iota = R_h \circ \iota \circ L_h, \qquad \text{for all } h \in G.$$

In particular, for $h = g^{-1}$, we get

$$d\iota_g = (dR_{g^{-1}})_e \circ d\iota_e \circ (dL_{g^{-1}})_g.$$

As $(dR_{g^{-1}})_e$ and $d(L_{g^{-1}})_g$ are isometries (since G has a bi-invariant metric), $d\iota_g$ is an isometry iff $d\iota_e$ is. Thus, it remains to show that $d\iota_e$ is an isometry. However, $d\iota_e = -\mathrm{id}$, so $d\iota_g$ is an isometry for all $g \in G$.

It remains to prove that $d\iota_e = -\mathrm{id}$. This can be done in several ways. If we denote the multiplication of the group by $\mu\colon G \times G \to G$, then $T_e(G \times G) = T_eG \oplus T_eG = \mathfrak{g} \oplus \mathfrak{g}$ and it is easy to see that

$$d\mu_{(e,e)}(u, v) = u + v, \qquad \text{for all } u, v \in \mathfrak{g}.$$

This is because $d\mu_{(e,e)}$ is a homomorphism and because $g \mapsto \mu(e, g)$ and $g \mapsto \mu(g, e)$ are the identity map. As the map, $g \mapsto \mu(g\iota(g))$, is the constant map with value e, by differentiating and using the chain rule, we get

$$d\iota_e(u) = -u,$$

as desired. (Another proof makes use of the fact that for every, $u \in \mathfrak{g}$, the integral curve, γ, through e with $\gamma'(0) = u$ is a group homomorphism. Therefore,

$$\iota(\gamma(t)) = \gamma(t)^{-1} = \gamma(-t)$$

and by differentiating, we get $d\iota_e(u) = -u$.)

(2) We follow Milnor [107] (Lemma 21.1). From (1), the map ι is an isometry so, by Proposition 13.8 (3), it preserves geodesics through e. Since $d\iota_e$ reverses $T_eG = \mathfrak{g}$, it reverses geodesics through e. Observe that

$$I_a = R_a \circ \iota \circ R_{a^{-1}},$$

so by (1), I_a is an isometry and obviously, $I_a(a) = a$. Again, by Proposition 13.8 (3), the isometry I_a preserve geodesics, and since R_a and $R_{a^{-1}}$ translate geodesics but ι reverses geodesics, it follows that I_a reverses geodesics.

(3) We follow Milnor [107] (Lemma 21.2). Assume γ is the unique geodesic through e such that $\gamma'(0) = u$, and let X be the left invariant vector field such that $X(e) = u$. The first step is to prove that γ has domain $\mathbb{R}$ and that it is a group homomorphism, that is,

$$\gamma(s + t) = \gamma(s)\gamma(t).$$

Details of this argument are given in Milnor [107] (Lemma 20.1 and Lemma 21.2) and in Gallot, Hulin and Lafontaine [61] (Appendix B, Solution of Exercise 2.90). We present Milnor's proof.

Claim. The isometries, I_a, have the following property: For every geodesic, γ, through a, if we let $p = \gamma(0)$ and $q = \gamma(r)$, then

$$I_q \circ I_p(\gamma(t)) = \gamma(t + 2r),$$

whenever $\gamma(t)$ and $\gamma(t + 2r)$ are defined.

Let $\alpha(t) = \gamma(t+r)$. Then, α is a geodesic with $\alpha(0) = q$. As I_p reverses geodesics through p (and similarly for I_q), we get

$$\begin{aligned} I_q \circ I_p(\gamma(t)) &= I_q(\gamma(-t)) \\ &= I_q(\alpha(-t - r)) \\ &= \alpha(t + r) = \gamma(t + 2r). \end{aligned}$$

It follows from the claim that γ can be indefinitely extended, that is, the domain of γ is $\mathbb{R}$.

Next, we prove that γ is a homomorphism. By the Claim, $I_{\gamma(t)} \circ I_e$ takes $\gamma(u)$ into $\gamma(u + 2t)$. Now, by definition of I_a and I_e,

$$I_{\gamma(t)} \circ I_e(a) = \gamma(t)a\gamma(t),$$

so, with $a = \gamma(u)$, we get

$$\gamma(t)\gamma(u)\gamma(t) = \gamma(u + 2t).$$

By induction, it follows that

$$\gamma(nt) = \gamma(t)^n, \qquad \text{for all } n \in \mathbb{Z}.$$

We now use the (usual) trick of approximating every real by a rational number. For all $r, s \in \mathbb{R}$ with $s \neq 0$, if r/s is rational, say $r/s = m/n$ where m, n are integers, then $r = mt$ and $s = nt$ with $t = r/m = s/n$ and we get

$$\gamma(r + s) = \gamma(t)^{m+n} = \gamma(t)^m\gamma(t)^n = \gamma(r)\gamma(s).$$

Given any $t_1, t_2 \in \mathbb{R}$ with $t_2 \neq 0$, since t_1 and t_2 can be approximated by rationals r and s, as r/s is rational, $\gamma(r+s) = \gamma(r)\gamma(s)$, and by continuity, we get

$$\gamma(t_1 + t_2) = \gamma(t_1)\gamma(t_2),$$

as desired (the case $t_2 = 0$ is trivial as $\gamma(0) = e$).

As γ is a homomorphism, by differentiating the equation $\gamma(s+t) = \gamma(s)\gamma(t)$, we get

$$\frac{d}{dt}(\gamma(s+t))|_{t=0} = (dL_{\gamma(s)})_e \left(\frac{d}{dt}(\gamma(t))|_{t=0} \right),$$

that is

$$\gamma'(s) = (dL_{\gamma(s)})_e(\gamma'(0)) = X(\gamma(s)),$$

which means that γ is the integral curve of the left-invariant vector field, X, a one-parameter group.

Conversely, let c be the one-parameter group determined by a left-invariant vector field, X, with $X(e) = u$ and let γ be the unique geodesic through e such that $\gamma'(0) = u$. Since we have just shown that γ is a homomorphism with $\gamma'(0) = u$, by uniqueness of one-parameter groups, $c = \gamma$, that is, c is a geodesic. □

Remarks:

(1) As $R_g = \iota \circ L_{g^{-1}} \circ \iota$, we deduce that if G has a left-invariant metric, then this metric is also right-invariant iff ι is an isometry.

(2) Property (2) of Proposition 18.18 says that a Lie group with a bi-invariant metric is a *symmetric space*, an important class of Riemannian spaces invented and studied extensively by Elie Cartan.

(3) The proof of 18.18 (3) given in O'Neill [120] (Chapter 11, equivalence of (5) and (6) in Proposition 9) appears to be missing the "hard direction", namely, that a geodesic is a one-parameter group. Also, since left and right translations are isometries and since isometries map geodesics to geodesics, the geodesics through any point, $a \in G$, are the left (or right) translates of the geodesics through e and thus, are expressed in terms of the group exponential. Therefore, the geodesics through $a \in G$ are of the form

$$\gamma(t) = L_a(\exp(tu)),$$

where $u \in \mathfrak{g}$. Observe that $\gamma'(0) = (dL_a)_e(u)$.

(4) Some of the other facts stated in Proposition 18.17 and Proposition 18.18 are equivalent to the fact that a left-invariant metric is also bi-invariant, see O'Neill [120] (Chapter 11, Proposition 9).

Many more interesting results about left-invariant metrics on Lie groups can be found in Milnor's paper [110]. For example, flat left-invariant metrics on Lie a group are characterized (Theorem 1.5). We conclude this section by stating the following proposition (Milnor [110], Lemma 7.6):

Proposition 18.19. *If G is any compact, simple, Lie group, G, then the bi-invariant metric is unique up to a constant. Such a metric necessarily has constant Ricci curvature.*

18.4 The Killing Form

The Killing form showed the tip of its nose in Proposition 18.17. It is an important concept and, in this section, we establish some of its main properties. First, we recall its definition.

Definition 18.3. For any Lie algebra, $\mathfrak{g}$, the *Killing form, B, of* $\mathfrak{g}$ is the symmetric bilinear form, $B\colon \mathfrak{g} \times \mathfrak{g} \to \mathbb{R}$, given by

$$B(u, v) = \mathrm{tr}(\mathrm{ad}(u) \circ \mathrm{ad}(v)), \qquad \text{for all } u, v \in \mathfrak{g}.$$

If $\mathfrak{g}$ is the Lie algebra of a Lie group, G, we also refer to B as the *Killing form of* G.

Remark: According to the experts (see Knapp [90], page 754) the *Killing form* as above was not defined by Killing and is closer to a variant due to Elie Cartan. On the other hand, the notion of "Cartan matrix" is due to Wilhelm Killing!

For example, consider the group $\mathbf{SU}(2)$. Its Lie algebra, $\mathfrak{su}(2)$, consists of all skew-Hermitian 2×2 matrices with zero trace, that is matrices of the form

$$\begin{pmatrix} ai & b + ic \\ -b + ic & -ai \end{pmatrix}, \qquad a, b, c \in \mathbb{R},$$

a three-dimensional algebra. By picking a suitable basis of $\mathfrak{su}(2)$, it can be shown that

$$B(X, Y) = 4\mathrm{tr}(XY).$$

Now, if we consider the group $\mathbf{U}(2)$, its Lie algebra, $\mathfrak{u}(2)$, consists of all skew-Hermitian 2×2 matrices, that is matrices of the form

$$\begin{pmatrix} ai & b + ic \\ -b + ic & id \end{pmatrix}, \qquad a, b, c, d \in \mathbb{R},$$

a four-dimensional algebra. This time, it can be shown that

$$B(X, Y) = 4\mathrm{tr}(XY) - 2\mathrm{tr}(X)\mathrm{tr}(Y).$$

For $\mathbf{SO}(3)$, we know that $\mathfrak{so}(3) = \mathfrak{su}(2)$ and we get

$$B(X, Y) = \mathrm{tr}(XY).$$

Actually, it can be shown that

$$\begin{aligned}
\mathbf{U}(n)&: \quad B(X,Y) = 2n\mathrm{tr}(XY) - 2\mathrm{tr}(X)\mathrm{tr}(Y)\\
\mathbf{SU}(n)&: \quad B(X,Y) = 2n\mathrm{tr}(XY)\\
\mathbf{SO}(n)&: \quad B(X,Y) = (n-2)\mathrm{tr}(XY).
\end{aligned}$$

Recall that a homomorphism of Lie algebras, $\varphi\colon \mathfrak{g} \to \mathfrak{h}$, is a linear map that preserves brackets, that is,

$$\varphi([u,v]) = [\varphi(u), \varphi(v)].$$

Proposition 18.20. *The Killing form, B, of a Lie algebra, $\mathfrak{g}$, has the following properties:*

(1) It is a symmetric bilinear form invariant under all automorphisms of $\mathfrak{g}$. In particular, if $\mathfrak{g}$ is the Lie algebra of a Lie group, G, then B is Ad_g-invariant, for all $g \in G$.

(2) The linear map, $\mathrm{ad}(u)$, is skew-adjoint w.r.t B for all $u \in \mathfrak{g}$, that is

$$B(\mathrm{ad}(u)(v), w) = -B(v, \mathrm{ad}(u)(w)), \qquad \textit{for all } u, v, w \in \mathfrak{g}$$

or, equivalently

$$B([u,v], w) = B(u, [v,w]), \qquad \textit{for all } u, v, w \in \mathfrak{g}.$$

Proof. (1) The form B is clearly bilinear and as $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, it is symmetric. If φ is an automorphism of $\mathfrak{g}$, the preservation of the bracket implies that

$$\mathrm{ad}(\varphi(u)) \circ \varphi = \varphi \circ \mathrm{ad}(u),$$

so

$$\mathrm{ad}(\varphi(u)) = \varphi \circ \mathrm{ad}(u) \circ \varphi^{-1}.$$

From $\mathrm{tr}(XY) = \mathrm{tr}(YX)$, we get $\mathrm{tr}(A) = \mathrm{tr}(BAB^{-1})$, so we get

$$\begin{aligned}
B(\varphi(u), \varphi(v)) &= \mathrm{tr}(\mathrm{ad}(\varphi(u)) \circ \mathrm{ad}(\varphi(v))\\
&= \mathrm{tr}(\varphi \circ \mathrm{ad}(u) \circ \varphi^{-1} \circ \varphi \circ \mathrm{ad}(v) \circ \varphi^{-1})\\
&= \mathrm{tr}(\mathrm{ad}(u) \circ \mathrm{ad}(v)) = B(u,v).
\end{aligned}$$

Since Ad_g is an automorphism of $\mathfrak{g}$ for all $g \in G$, B is Ad_g-invariant.

(2) We have

$$B(\mathrm{ad}(u)(v), w) = B([u,v], w) = \mathrm{tr}(\mathrm{ad}([u,v]) \circ \mathrm{ad}(w))$$

and

$$B(v, \mathrm{ad}(u)(w)) = B(v, [u,w]) = \mathrm{tr}(\mathrm{ad}(v) \circ \mathrm{ad}([u,w])).$$

However, the Jacobi identity is equivalent to

$$\mathrm{ad}([u,v]) = \mathrm{ad}(u)\circ\mathrm{ad}(v) - \mathrm{ad}(v)\circ\mathrm{ad}(u).$$

Consequently,

$$\begin{aligned}\mathrm{tr}(\mathrm{ad}([u,v])\circ\mathrm{ad}(w)) &= \mathrm{tr}((\mathrm{ad}(u)\circ\mathrm{ad}(v) - \mathrm{ad}(v)\circ\mathrm{ad}(u))\circ\mathrm{ad}(w))\\ &= \mathrm{tr}(\mathrm{ad}(u)\circ\mathrm{ad}(v)\circ\mathrm{ad}(w)) - \mathrm{tr}(\mathrm{ad}(v)\circ\mathrm{ad}(u)\circ\mathrm{ad}(w))\end{aligned}$$

and

$$\begin{aligned}\mathrm{tr}(\mathrm{ad}(v)\circ\mathrm{ad}([u,w])) &= \mathrm{tr}(\mathrm{ad}(v)\circ(\mathrm{ad}(u)\circ\mathrm{ad}(w) - \mathrm{ad}(w)\circ\mathrm{ad}(u)))\\ &= \mathrm{tr}(\mathrm{ad}(v)\circ\mathrm{ad}(u)\circ\mathrm{ad}(w)) - \mathrm{tr}(\mathrm{ad}(v)\circ\mathrm{ad}(w)\circ\mathrm{ad}(u)).\end{aligned}$$

As

$$\mathrm{tr}(\mathrm{ad}(u)\circ\mathrm{ad}(v)\circ\mathrm{ad}(w)) = \mathrm{tr}(\mathrm{ad}(v)\circ\mathrm{ad}(w)\circ\mathrm{ad}(u)),$$

we deduce that

$$B(\mathrm{ad}(u)(v),w) = \mathrm{tr}(\mathrm{ad}([u,v])\circ\mathrm{ad}(w)) = -\mathrm{tr}(\mathrm{ad}(v)\circ\mathrm{ad}([u,w])) = -B(v,\mathrm{ad}(u)(w)),$$

as claimed. □

Remarkably, the Killing form yields a simple criterion due to Elie Cartan for testing whether a Lie algebra is semisimple.

Theorem 18.21. *(Cartan's Criterion for Semisimplicity) A lie algebra, $\mathfrak{g}$, is semisimple iff its Killing form, B, is non-degenerate.*

As far as we know, all the known proofs of Cartan's criterion are quite involved. A fairly easy going proof can be found in Knapp [90] (Chapter 1, Theorem 1.45). A more concise proof is given in Serre [137] (Chapter VI, Theorem 2.1). As a corollary of Theorem 18.21, we get:

Proposition 18.22. *If G is a semisimple Lie group, then the center of its Lie algebra is trivial, that is, $Z(\mathfrak{g}) = (0)$.*

Proof. Since $u \in \mathfrak{g}$ iff $\mathrm{ad}(u) = 0$, we have

$$B(u,u) = \mathrm{tr}(\mathrm{ad}(u)\circ\mathrm{ad}(u)) = 0.$$

As B is nondegenerate, we must have $u = 0$. □

Since a Lie group with trivial Lie algebra is discrete, this implies that the center of a simple Lie group is discrete (because the Lie algebra of the center of a Lie group is the center of its Lie algebra. Prove it!).

We can also characterize which Lie groups have a Killing form which is negative definite.

Theorem 18.23. *A connected Lie group is compact and semisimple iff its Killing form is negative definite.*

Proof. First, assume that G is compact and semisimple. Then, by Proposition 18.6, there is an inner product on $\mathfrak{g}$ inducing a bi-invariant metric on G and by Proposition 18.7, every linear map, $\mathrm{ad}(u)$, is skew-adjoint. Therefore, if we pick an orthonormal basis of $\mathfrak{g}$, the matrix, X, representing $\mathrm{ad}(u)$ is skew-symmetric and

$$B(u,u) = \mathrm{tr}(\mathrm{ad}(u)\circ \mathrm{ad}(u)) = \mathrm{tr}(XX) = \sum_{i,j=1}^{n} a_{ij}a_{ji} = -\sum_{i,j=1}^{n} a_{ij}^2 \leq 0.$$

Since G is semisimple, B is nondegenerate, and so, it is negative definite.

Now, assume that B is negative definite. If so, $-B$ is an inner product on $\mathfrak{g}$ and by Proposition 18.20, it is Ad-invariant. By Proposition 18.3, the inner product $-B$ induces a bi-invariant metric on G and by Proposition 18.17 (d), the Ricci curvature is given by

$$\mathrm{Ric}(u,v) = -\frac{1}{4}\, B(u,v),$$

which shows that $r(u) > 0$ for all units vectors, $u \in \mathfrak{g}$. As in the proof of Proposition 18.14, there is some constant, $c > 0$, which is a lower bound on all Ricci curvatures, $r(u)$, and by Myers' Theorem (Theorem 13.28), G is compact (with finite fundamental group). By Cartan's Criterion, as B is non-degenerate, G is also semisimple. □

Remark: A compact semisimple Lie group equipped with $-B$ as a metric is an Einstein manifold, since Ric is proportional to the metric (see Definition 13.5).

Using Theorem 18.23 and since the Killing forms for $\mathbf{U}(n)$, $\mathbf{SU}(n)$ and $\mathbf{S})(n)$ are given by

$$\begin{aligned}
\mathbf{U}(n)\colon &\qquad B(X,Y) = 2n\mathrm{tr}(XY) - 2\mathrm{tr}(X)\mathrm{tr}(Y)\\
\mathbf{SU}(n)\colon &\qquad B(X,Y) = 2n\mathrm{tr}(XY)\\
\mathbf{SO}(n)\colon &\qquad B(X,Y) = (n-2)\mathrm{tr}(XY),
\end{aligned}$$

we see that $\mathbf{SU}(n)$ and $\mathbf{SO}(n)$ are compact and semisimple but $\mathbf{U}(n)$, even though it is compact, is not semisimple.

Semisimple Lie algebras and semisimple Lie groups have been investigated extensively, starting with the complete classification of the complex semisimple Lie algebras by Killing (1888) and corrected by Elie Cartan in his thesis (1894). One should read the Notes, especially on Chapter II, at the end of Knapp's book [90] for a fascinating account of the history of the theory of semisimple Lie algebras.

The theories and the body of results that emerged from these investigations play a very important role not only in mathematics but also in physics and constitute one of the most

beautiful chapters of mathematics. A quick introduction to these theories can be found in Arvanitoyeogos [8] and in Carter, Segal, Macdonald [31]. A more comprehensive but yet still introductory presentation is given in Hall [71]. The most comprehensive treatment is probably Knapp [90]. An older is classic is Helgason [74], which also discusses differential geometric aspects of Lie groups. Other "advanced" presentations can be found in Bröcker and tom Dieck [25], Serre [138, 137], Samelson [132], Humphreys [82] and Kirillov [87].

Chapter 19

The Log-Euclidean Framework Applied to SPD Matrices and Polyaffine Transformations

19.1 Introduction

In this Chapter, we use what we have learned in previous chapters to describe an approach due to Arsigny, Fillard, Pennec and Ayache to define a Lie group structure and a class of metrics on symmetric, positive-definite matrices (SPD matrices) which yield a new notion of mean on SPD matrices generalizing the standard notion of geometric mean.

SPD matrices are used in diffusion tensor magnetic resonance imaging (for short, DTI) and they are also a basic tool in numerical analysis, for example, in the generation of meshes to solve partial differential equations more efficiently.

As a consequence, there is a growing need to interpolate or to perform statistics on SPD matrices, such as computing the mean of a finite number of SPD matrices.

Recall that the set of $n \times n$ SPD matrices, $\mathbf{SPD}(n)$, is not a vector space (because if $A \in \mathbf{SPD}(n)$, then $\lambda A \notin \mathbf{SPD}(n)$ if $\lambda < 0$) but it is a convex cone. Thus, the *arithmetic mean* of n SPD matrices, $S_1, \ldots, S_n$, can be defined as $(S_1 + \cdots + S_n)/n$, which is SPD. However, there are many situations, especially in DTI, where this mean is not adequate. There are essentially two problems:

(1) The arithmetic mean is not invariant under inversion, which means that if $S = (S_1 + \cdots + S_n)/n$, then in general, $S^{-1} \neq (S_1^{-1} + \cdots + S_n^{-1})/n$.

(2) The swelling effect: the determinant, $\det(S)$, of the mean, S, may be strictly larger than the original determinants, $\det(S_i)$. This effect is undesirable in DTI because it amounts to introducing more diffusion, which is physically unacceptable.

To circumvent these difficulties, various metrics on SPD matrices have been proposed. One class of metrics is the *affine-invariant metrics* (see Arsigny, Pennec and Ayache [6]). The swelling effect disappears and the new mean is invariant under inversion but computing this new mean has a high computational cost and, in general, there is no closed-form formula for this new kind of mean.

Arsigny, Fillard, Pennec and Ayache [5] have defined a new family of metrics on $\mathbf{SPD}(n)$ named *Log-Euclidean metrics* and have also defined a novel structure of Lie group on $\mathbf{SPD}(n)$ which yields a notion of mean that has the same advantages as the affine mean but is a lot cheaper to compute. Furthermore, this new mean, called *Log-Euclidean mean*, is given by a simple closed-form formula. We will refer to this approach as the *Log-Euclidean Framework*.

The key point behind the Log-Euclidean Framework is the fact that the exponential map, $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$, is a bijection, where $\mathbf{S}(n)$ is the space of $n \times n$ symmetric matrices (see Gallier [60], Chapter 14, Lemma 14.3.1). Consequently, the exponential map has a well-defined inverse, the *logarithm*, $\log\colon \mathbf{SPD}(n) \to \mathbf{S}(n)$.

But more is true. It turns out that $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$ is a diffeomorphism, a fact stated as Theorem 2.8 in Arsigny, Fillard, Pennec and Ayache [5].

Since exp is a bijection, the above result follows from the fact that exp is a local diffeomorphism on $\mathbf{S}(n)$, because $d\exp_S$ is non-singular for all $S \in \mathbf{S}(n)$. In Arsigny, Fillard, Pennec and Ayache [5], it is proved that the non-singularity of $d\exp_I$ near 0, which is well-known, "propagates" to the whole of $\mathbf{S}(n)$.

Actually, the non-singularity of $d\exp$ on $\mathbf{S}(n)$ is a consequence of a more general result of some interest whose proof can be found in in Mmeimné and Testard [112], Chapter 3, Theorem 3.8.4 (see also Bourbaki [22], Chapter III, Section 6.9, Proposition 17, and also Theorem 6).

Let $\mathcal{S}(n)$ denote the set of all real matrices whose eigenvalues, $\lambda + i\mu$, lie in the horizontal strip determined by the condition $-\pi < \mu < \pi$. Then, we have the following theorem:

Theorem 19.1. *The restriction of the exponential map to $\mathcal{S}(n)$ is a diffeomorphism of $\mathcal{S}(n)$ onto its image, $\exp(\mathcal{S}(n))$. Furthermore, $\exp(\mathcal{S}(n))$ consists of all invertible matrices that have no real negative eigenvalues; it is an open subset of $\mathbf{GL}(n,\mathbb{R})$; it contains the open ball, $B(I,1) = \{A \in \mathbf{GL}(n,\mathbb{R}) \mid \|A - I\| < 1\}$, for every norm $\|\ \|$ on $n \times n$ matrices satisfying the condition $\|AB\| \leq \|A\|\,\|B\|$.*

Part of the proof consists in showing that exp is a local diffeomorphism and for this, to prove that $d\exp_X$ is invertible for every $X \in \mathcal{S}(n)$. This requires finding an explicit formula for the derivative of the exponential, which can be done.

With this preparation we are ready to present the natural Lie group structure on $\mathbf{SPD}(n)$ introduced by Arsigny, Fillard, Pennec and Ayache [5] (see also Arsigny's thesis [3]).

19.2 A Lie-Group Structure on $\mathbf{SPD}(n)$

Using the diffeomorphism, $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$, and its inverse, $\log\colon \mathbf{SPD}(n) \to \mathbf{S}(n)$, an abelian group structure can be defined on $\mathbf{SPD}(n)$ as follows:

Definition 19.1. For any two matrices, $S_1, S_2 \in \mathbf{SPD}(n)$, define the *logarithmic product*, $S_1 \odot S_2$, by

$$S_1 \odot S_2 = \exp(\log(S_1) + \log(S_2)).$$

Obviously, the multiplication operation, $\odot$, is commutative. The following proposition is shown in Arsigny, Fillard, Pennec and Ayache [5] (Proposition 3.2):

Proposition 19.2. *The set, $\mathbf{SPD}(n)$, with the binary operation, $\odot$, is an abelian group with identity, I, and with inverse operation the usual inverse of matrices. Whenever S_1 and S_2 commute, then $S_1 \odot S_2 = S_1 S_2$ (the usual multiplication of matrices).*

For the last statement, we need to show that if $S_1, S_2 \in \mathbf{SPD}(n)$ commute, then $S_1 S_2$ is also in $\mathbf{SPD}(n)$ and that $\log(S_1)$ and $\log(S_2)$ commute, which follows from the fact that if two diagonalizable matrices commute, then they can be diagonalized over the same basis of eigenvectors.

Actually, $(\mathbf{SPD}(n), \odot, I)$ is an abelian Lie group isomorphic to the vector space (also an abelian Lie group!) $\mathbf{S}(n)$, as shown in Arsigny, Fillard, Pennec and Ayache [5] (Theorem 3.3 and Proposition 3.4):

Theorem 19.3. *The abelian group, $(\mathbf{SPD}(n), \odot, I)$ is a Lie group isomorphic to its Lie algebra, $\mathfrak{spd}(n) = \mathbf{S}(n)$. In particular, the Lie group exponential in $\mathbf{SPD}(n)$ is identical to the usual exponential on $\mathbf{S}(n)$.*

We now investigate bi-invariant metrics on the Lie group, $\mathbf{SPD}(n)$.

19.3 Log-Euclidean Metrics on $\mathbf{SPD}(n)$

If G is a lie group, recall that we have the operations of left multiplication, L_a, and right multiplication, R_a, given by

$$L_a(b) = ab, \qquad R_a(b) = ba,$$

for all $a, b \in G$. A Riemannian metric, $\langle -, - \rangle$, on G is *left-invariant* iff dL_a is an isometry for all $a \in G$, that is,

$$\langle u, v \rangle_b = \langle (dL_a)_b(u), (dL_a)_b(v) \rangle_{ab},$$

for all $b \in G$ and all $u, v \in T_b G$. Similarly, $\langle -, - \rangle$ is *right-invariant* iff dR_a is an isometry for all $a \in G$ and $\langle -, - \rangle$ is *bi-invariant* iff it is both left and right invariant. In general, a Lie

group does not admit a bi-invariant metric but an abelian Lie group always does because $\mathrm{Ad}_g = \mathrm{id} \in \mathbf{GL}(\mathfrak{g})$ for all $g \in G$ and so, the adjoint representation, $\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g})$, is trivial (that is, $\mathrm{Ad}(G) = \{\mathrm{id}\}$) and then, the existence of bi-invariant metrics is a consequence of Proposition 18.3, which we repeat here for the convenience of the reader:

Proposition 19.4. *There is a bijective correspondence between bi-invariant metrics on a Lie group, G, and* Ad*-invariant inner products on the Lie algebra, $\mathfrak{g}$, of G, that is, inner products, $\langle -, - \rangle$, on $\mathfrak{g}$ such that Ad_a is an isometry of $\mathfrak{g}$ for all $a \in G$; more explicitly, inner products such that*

$$\langle \mathrm{Ad}_a u, \mathrm{Ad}_a v \rangle = \langle u, v \rangle,$$

for all $a \in G$ and all $u, v \in \mathfrak{g}$.

Then, given any inner product, $\langle -, - \rangle$ on G, the induced bi-invariant metric on G is given by

$$\langle u, v \rangle_g = \langle (dL_{g^{-1}})_g u, (dL_{g^{-1}})_g v \rangle.$$

Now, the geodesics on a Lie group equipped with a bi-invariant metric are the left (or right) translates of the geodesics through e and the geodesics through e are given by the group exponential, as stated in Proposition 18.18 (3) which we repeat for the convenience of the reader:

Proposition 19.5. *For any Lie group, G, equipped with a bi-invariant metric, we have:*

(1) The inversion map, $\iota\colon g \mapsto g^{-1}$, is an isometry.

(2) For every $a \in G$, if I_a denotes the map given by

$$I_a(b) = ab^{-1}a, \qquad \text{for all } a, b \in G,$$

then I_a is an isometry fixing a which reverses geodesics, that is, for every geodesic, γ, through a

$$I_a(\gamma)(t) = \gamma(-t).$$

(3) The geodesics through e are the integral curves, $t \mapsto \exp(tu)$, where $u \in \mathfrak{g}$, that is, the one-parameter groups. Consequently, the Lie group exponential map, $\exp\colon \mathfrak{g} \to G$, coincides with the Riemannian exponential map (at e) from T_eG to G, where G is viewed as a Riemannian manifold.

If we apply Proposition 19.5 to the abelian Lie group, $\mathbf{SPD}(n)$, we find that the geodesics through S are of the form

$$\gamma(t) = S \odot e^{tV},$$

where $V \in \mathbf{S}(n)$. But $S = e^{\log S}$, so

$$S \odot e^{tV} = e^{\log S} \odot e^{tV} = e^{\log S + tV},$$

so every geodesic through S is of the form

$$\gamma(t) = e^{\log S + tV} = \exp(\log S + tV).$$

To avoid confusion between the exponential and the logarithm as Lie group maps and as Riemannian manifold maps, we will denote the former by exp and log and their Riemannian counterparts by Exp and Log. Note that

$$\gamma'(0) = d\exp_{\log S}(V)$$

and since the exponential map of $\mathbf{SPD}(n)$, as a Riemannian manifold, is given by

$$\mathrm{Exp}_S(U) = \gamma_U(1),$$

where γ_U is the unique geodesic such that $\gamma_U(0) = S$ and $\gamma'_U(0) = U$, we must have $d\exp_{\log S}(V) = U$, so $V = (d\exp_{\log S})^{-1}(U)$ and

$$\mathrm{Exp}_S(U) = e^{\log S + V} = e^{\log S + (d\exp_{\log S})^{-1}(U)}.$$

However, $\log \circ \exp = \mathrm{id}$ so, by differentiation, we get

$$(d\exp_{\log S})^{-1}(U) = d\log_S(U),$$

which yields

$$\mathrm{Exp}_S(U) = e^{\log S + d\log_S(U)}.$$

To get a formula for $\mathrm{Log}_S T$, we solve the equation $T = \mathrm{Exp}_S(U)$ with respect to U, that is

$$e^{\log S + (d\exp_{\log S})^{-1}(U)} = T$$

which yields

$$\log S + (d\exp_{\log S})^{-1}(U) = \log T,$$

that is, $U = d\exp_{\log S}(\log T - \log S)$. Therefore,

$$\mathrm{Log}_S T = d\exp_{\log S}(\log T - \log S).$$

Finally, we can find an explicit formula for the Riemannian metric,

$$\langle U, V\rangle_S = \langle d(L_{S^{-1}})_S(U), d(L_{S^{-1}})_S(V)\rangle,$$

because $d(L_{S^{-1}})_S = d\log_S$, which can be shown as follows: Observe that

$$(\log \circ L_{S^{-1}})(T) = \log S^{-1} + \log T,$$

so $d(\log \circ L_{S^{-1}})_T = d\log_T$, that is

$$d\log_{S^{-1}\odot T} \circ d(L_{S^{-1}})_T = d\log_T,$$

which, for $T = S$, yields $(dL_{S^{-1}})_S = d\log_S$, since $d\log_I = I$. Therefore,

$$\langle U, V\rangle_S = \langle d\log_S(U), d\log_S(V)\rangle.$$

Now, a Lie group with a bi-invariant metric is complete, so given any two matrices, $S, T \in \mathbf{SPD}(n)$, their distance is the length of the geodesic segment, γ_V, such that $\gamma_V(0) = S$ and $\gamma_V(1) = T$, namely $\|V\|$, but $V = \log_S T$ so that

$$d(S,T) = \|\log_S T\|_S\,,$$

where $\|\ \|_S$ is the norm given by the Riemannian metric. Using the equation

$$\mathrm{Log}_S T = d\exp_{\log S}(\log T - \log S),$$

and the fact that $d\log \circ d\exp = \mathrm{id}$, we get

$$d(S,T) = \|\log T - \log S\|\,,$$

where $\|\ \|$ is the norm corresponding to the inner product on $\mathfrak{spd}(n) = \mathbf{S}(n)$. Since $\langle -, -\rangle$ is a bi-invariant metric on $\mathbf{S}(n)$ and since

$$\langle U, V\rangle_S = \langle d\log_S(U), d\log_S(V)\rangle,$$

we see that the map, $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$, is an isometry (since $d\exp \circ d\log = \mathrm{id}$).

In summary, we have proved Corollary 3.9 of Arsigny, Fillard, Pennec and Ayache [5]:

Theorem 19.6. *For any inner product, $\langle -, -\rangle$, on $\mathbf{S}(n)$, if we give the Lie group, $\mathbf{SPD}(n)$, the bi-invariant metric induced by $\langle -, -\rangle$, then the following properties hold:*

(1) For any $S \in \mathbf{SPD}(n)$, the geodesics through S are of the form

$$\gamma(t) = e^{\log S + tV}, \qquad V \in \mathbf{S}(n).$$

(2) The exponential and logarithm associated with the bi-invariant metric on $\mathbf{SPD}(n)$ are given by

$$\begin{aligned}\mathrm{Exp}_S(U) &= e^{\log S + d\log_S(U)}\\ \mathrm{Log}_S(T) &= d\exp_{\log S}(\log T - \log S),\end{aligned}$$

for all $S, T \in \mathbf{SPD}(n)$ and all $U \in \mathbf{S}(n)$.

(3) The bi-invariant metric on $\mathbf{SPD}(n)$ is given by

$$\langle U, V\rangle_S = \langle d\log_S(U), d\log_S(V)\rangle,$$

for all $U, V \in \mathbf{S}(n)$ and all $S \in \mathbf{SPD}(n)$ and the distance, $d(S,T)$, between any two matrices, $S, T \in \mathbf{SPD}(n)$, is given by

$$d(S,T) = \|\log T - \log S\|\,,$$

where $\|\ \|$ is the norm corresponding to the inner product on $\mathfrak{spd}(n) = \mathbf{S}(n)$.

(4) The map, $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$*, is an isometry.*

In view of Theorem 19.6, part (3), bi-invariant metrics on the Lie group $\mathbf{SPD}(n)$ are called *Log-Euclidean metrics*. Since $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$ is an isometry and $\mathbf{S}(n)$ is a vector space, the Riemannian Lie group, $\mathbf{SPD}(n)$, is a complete, simply-connected and flat manifold (the sectional curvature is zero at every point) that is, a flat *Hadamard manifold* (see Sakai [131], Chapter V, Section 4).

Although, in general, Log-Euclidean metrics are not invariant under the action of arbitary invertible matrices, they are invariant under similarity transformations (an isometry composed with a scaling). Recall that $\mathbf{GL}(n)$ acts on $\mathbf{SPD}(n)$, *via*,

$$A \cdot S = ASA^\top,$$

for all $A \in \mathbf{GL}(n)$ and all $S \in \mathbf{SPD}(n)$. We say that a Log-Euclidean metric is *invariant under* $A \in \mathbf{GL}(n)$ iff

$$d(A \cdot S, A \cdot T) = d(S, T),$$

for all $S, T \in \mathbf{SPD}(n)$. The following result is proved in Arsigny, Fillard, Pennec and Ayache [5] (Proposition 3.11):

Proposition 19.7. *There exist metrics on* $\mathbf{S}(n)$ *that are invariant under all similarity transformations, for example, the metric* $\langle S, T\rangle = \operatorname{tr}(ST)$.

19.4 A Vector Space Structure on $\mathbf{SPD}(n)$

The vector space structure on $\mathbf{S}(n)$ can also be transfered onto $\mathbf{SPD}(n)$.

Definition 19.2. For any matrix, $S \in \mathbf{SPD}(n)$, for any scalar, $\lambda \in \mathbb{R}$, define the scalar multiplication, $\lambda \circledast S$, by

$$\lambda \circledast S = \exp(\lambda \log(S)).$$

It is easy to check that $(\mathbf{SPD}(n), \odot, \circledast)$ is a vector space with addition $\odot$ and scalar multiplication, $\circledast$. By construction, the map, $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$, is a linear isomorphism. What happens is that the vector space structure on $\mathbf{S}(n)$ is transfered onto $\mathbf{SPD}(n)$ *via* the log and exp maps.

19.5 Log-Euclidean Means

One of the major advantages of Log-Euclidean metrics is that they yield a computationally inexpensive notion of mean with many desirable properties. If $(x_1, \ldots, x_n)$ is a list of n data

points in $\mathbb{R}^m$, then it is an easy exercise to see that the mean, $\overline{x} = (x_1 + \cdots + x_n)/n$, is the unique minimum of the map

$$x \mapsto \sum_{i=1}^{n} d(x, x_i)_2^2,$$

where d_2 is the Euclidean distance on $\mathbb{R}^m$. We can think of the quantity,

$$\sum_{i=1}^{n} d(x, x_i)_2^2,$$

as the *dispersion* of the data. More generally, if (X, d) is a metric space, for any $\alpha > 0$ and any positive weights, $w_1, \ldots, w_n$, with $\sum_{i=1}^{n} w_i = 1$, we can consider the problem of minimizing the function,

$$x \mapsto \sum_{i=1}^{n} w_i d(x, x_i)^\alpha.$$

The case $\alpha = 2$ corresponds to a generalization of the notion of mean in a vector space and was investigated by Fréchet. In this case, any minimizer of the above function is known as a *Fréchet mean*. Fréchet means are not unique but if X is a complete Riemannian manifold, certain sufficient conditions on the dispersion of the data are known that ensure the existence and uniqueness of the Fréchet mean (see Pennec [121]). The case $\alpha = 1$ corresponds to a generalization of the notion of *median*. When the weights are all equal, the points that minimize the map,

$$x \mapsto \sum_{i=1}^{n} d(x, x_i),$$

are called *Steiner points*. On a Hadamard manifold, Steiner points can be characterized (see Sakai [131], Chapter V, Section 4, Proposition 4.9).

In the case where $X = \mathbf{SPD}(n)$ and d is a Log-Euclidean metric, it turns out that the Fréchet mean is unique and is given by a simple closed-form formula. This is easy to see and we have the following theorem from Arsigny, Fillard, Pennec and Ayache [5] (Theorem 3.13):

Theorem 19.8. *Given N matrices, $S_1, \ldots, S_N \in \mathbf{SPD}(n)$, their Log-Euclidean Fréchet mean exists and is uniquely determined by the formula*

$$\mathbb{E}_{\mathrm{LE}}(S_1, \ldots, S_N) = \exp\left(\frac{1}{N}\sum_{i=1}^{N} \log(S_i)\right).$$

Furthermore, the Log-Euclidean mean is similarity-invariant, invariant by group multiplication and inversion and exponential-invariant.

Similarity-invariance means that for any similarity, A,

$$\mathbb{E}_{\mathrm{LE}}(AS_1A^\top, \ldots, AS_NA^\top) = A\mathbb{E}_{\mathrm{LE}}(S_1, \ldots, S_N)A^\top$$

and similarly for the other types of invariance.

Observe that the Log-Euclidean mean is a generalization of the notion of geometric mean. Indeed, if $x_1, \ldots, x_n$ are n positive numbers, then their *geometric mean* is given by

$$\mathbb{E}_{\mathrm{geom}}(x_1, \ldots, x_n) = (x_1 \cdots x_n)^{\frac{1}{n}} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(x_i)\right).$$

The Log-Euclidean mean also has a good behavior with respect to determinants. The following theorem is proved in Arsigny, Fillard, Pennec and Ayache [5] (Theorem 4.2):

Theorem 19.9. *Given N matrices, $S_1, \ldots, S_N \in \mathbf{SPD}(n)$, we have*

$$\det(\mathbb{E}_{\mathrm{LE}}(S_1, \ldots, S_N)) = \mathbb{E}_{\mathrm{geom}}(\det(S_1), \ldots, \det(S_N)).$$

Remark: The last line of the proof in Arsigny, Fillard, Pennec and Ayache [5] seems incorrect.

Arsigny, Fillard, Pennec and Ayache [5] also compare the Log-Euclidean mean with the affine mean. We highly recommend the above paper as well as Arsigny's thesis [3] for further details.

Next, we discuss the application of the Log-Euclidean framework to the blending of locally affine transformations, known as Log-Euclidean polyaffine transformations, as presented in Arsigny, Commowick, Pennec and Ayache [4].

19.6 Log-Euclidean Polyaffine Transformations

The registration of medical images is an important and difficult problem. The work described in Arsigny, Commowick, Pennec and Ayache [4] (and Arsigny's thesis [3]) makes an orginal and valuable contribution to this problem by describing a method for parametrizing a class of non-rigid deformations with a small number of degrees of freedom. After a global affine alignment, this sort of parametrization allows a finer local registration with very smooth transformations. This type of parametrization is particularly well adpated to the registration of histological slices, see Arsigny, Pennec and Ayache [6].

The goal is to fuse some affine or rigid transformations in such a way that the resulting transformation is invertible and smooth. The direct approach which consists in blending N

global affine or rigid transformations, $T_1, \ldots, T_N$ using weights, $w_1, \ldots, w_N$, does not work because the resulting transformation,

$$T = \sum_{i=1}^{N} w_i T_i,$$

is not necessarily invertible. The purpose of the weights is to define the domain of influence in space of each T_i.

The key idea is to associate to each rigid (or affine) transformation, T, of $\mathbb{R}^n$, a vector field, V, and to view T as the diffeomorphism, Φ_1^V, corresponding to the time $t = 1$, where Φ_t^V is the global flow associated with V. In other words, T is the result of integrating an ODE

$$X' = V(X, t),$$

starting with some initial condition, X_0, and $T = X(1)$.

Now, it would be highly desirable if the vector field, V, did not depend on the time parameter, and this is indeed possible for a large class of affine transformations, which is one of the nice contributions of the work of Arsigny, Commowick, Pennec and Ayache [4]. Recall that an affine transformation, $X \mapsto LX + v$, (where L is an $n \times n$ matrix and $X, v \in \mathbb{R}^n$) can be conveniently represented as a linear transformation from $\mathbb{R}^{n+1}$ to itself if we write

$$\begin{pmatrix} X \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} L & v \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix}.$$

Then, the ODE with constant coefficients

$$X' = LX + v,$$

can be written

$$\begin{pmatrix} X' \\ 0 \end{pmatrix} = \begin{pmatrix} L & v \\ 0 & 0 \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix}$$

and, for every initial condition, $X = X_0$, its unique solution is given by

$$\begin{pmatrix} X(t) \\ 1 \end{pmatrix} = \exp\left(t \begin{pmatrix} L & v \\ 0 & 0 \end{pmatrix} \right) \begin{pmatrix} X_0 \\ 1 \end{pmatrix}.$$

Therefore, if we can find reasonable conditions on matrices, $T = \begin{pmatrix} M & t \\ 0 & 1 \end{pmatrix}$, to ensure that they have a unique real logarithm,

$$\log(T) = \begin{pmatrix} L & v \\ 0 & 0 \end{pmatrix},$$

then we will be able to associate a vector field, $V(X) = LX + v$, to T, in such a way that T is recovered by integrating the ODE, $X' = LX + v$. Furthermore, given N transformations,

$T_1, \ldots, T_N$, such that $\log(T_1), \ldots, \log(T_N)$ are uniquely defined, we can fuse $T_1, \ldots, T_N$ at the *infinitesimal level* by defining the ODE obtained by blending the vector fields, $V_1, \ldots, V_N$, associated with $T_1, \ldots, T_N$ (with $V_i(X) = L_i X + v_i$), namely

$$V(X) = \sum_{i=1}^{N} w_i(X)(L_i X + v_i).$$

Then, it is easy to see that the ODE,

$$X' = V(X),$$

has a unique solution for every $X = X_0$ defined for all t, and the fused transformation is just $T = X(1)$. Thus, the fused vector field,

$$V(X) = \sum_{i=1}^{N} w_i(X)(L_i X + v_i),$$

yields a one-parameter group of diffeomorphisms, Φ_t. Each transformation, Φ_t, is smooth and invertible and is called a *Log-Euclidean polyaffine tranformation*, for short, *LEPT*. Of course, we have the equation

$$\Phi_{s+t} = \Phi_s \circ \Phi_t,$$

for all $s, t \in \mathbb{R}$ so, in particular, the inverse of Φ_t is Φ_{-t}. We can also interpret Φ_s as $(\Phi_1)^s$, which will yield a fast method for computing Φ_s. Observe that when the weight are scalars, the one-parameter group is given by

$$\begin{pmatrix} \Phi_t(X) \\ 1 \end{pmatrix} = \exp\left(t \sum_{i=1}^{N} w_i \begin{pmatrix} L_i & v_i \\ 0 & 0 \end{pmatrix} \right) \begin{pmatrix} X \\ 1 \end{pmatrix},$$

which is the Log-Euclidean mean of the affine transformations, T_i's (w.r.t. the weights w_i).

Fortunately, there is a sufficient condition for a real matrix to have a unique real logarithm and this condition is not too restrictive in practice.

Recall that $\mathcal{S}(n)$ denotes the set of all real matrices whose eigenvalues, $\lambda + i\mu$, lie in the horizontal strip determined by the condition $-\pi < \mu < \pi$. We have the following version of Theorem 19.1:

Theorem 19.10. *The image,* $\exp(\mathcal{S}(n))$*, of* $\mathcal{S}(n)$ *by the exponential map is the set of real invertible matrices with no negative eigenvalues and* $\exp\colon \mathcal{S}(n) \to \exp(\mathcal{S}(n))$ *is a bijection.*

Theorem 19.10 is stated in Kenney and Laub [85] without proof. Instead, Kenney and Laub cite DePrima and Johnson [42] for a proof but this latter paper deals with complex matrices and does not contain a proof of our result either. The injectivity part of Theorem 19.10 can be found in Mmeimné and Testard [112], Chapter 3, Theorem 3.8.4.

In fact, $\exp\colon \mathcal{S}(n) \to \exp(\mathcal{S}(n))$ is a diffeomorphism, a result proved in Bourbaki [22], see Chapter III, Section 6.9, Proposition 17 and Theorem 6. Curious readers should read Gallier [59] for the full story.

For any matrix, $A \in \exp(\mathcal{S}(n))$, we refer to the unique matrix, $X \in \mathcal{S}(n)$, such that $e^X = A$, as the *principal logarithm* of A and we denote it as $\log A$.

Observe that if T is an affine transformation given in matrix form by

$$T = \begin{pmatrix} M & t \\ 0 & 1 \end{pmatrix},$$

since the eigenvalues of T are those of M plus the eigenvalue 1, the matrix T has no negative eigenvalues iff M has no negative eigenvalues and thus the principal logarithm of T exists iff the principal logarithm of M exists.

It is proved in Arsigny, Commowick, Pennec and Ayache that LEPT's are affine invariant, see [4], Section 2.3. This shows that LEPT's are produced by a truly geometric kind of blending, since the result does not depend at all on the choice of the coordinate system.

In the next section, we describe a fast method for computing due to Arsigny, Commowick, Pennec and Ayache [4].

19.7 Fast Polyaffine Transforms

Recall that since LEPT's are members of the one-parameter group, $(\Phi_t)_{t\in\mathbb{R}}$, we have

$$\Phi_{2t} = \Phi_{t+t} = \Phi_t^2$$

and thus,

$$\Phi_1 = (\Phi_{1/2^N})^{2^N}.$$

Observe the formal analogy of the above formula with the formula

$$\exp(M) = \exp\left(\frac{M}{2^N}\right)^{2^N},$$

for computing the exponential of a matrix, M, by the *scaling and squaring method*.

It turns out that the "scaling and squaring method" is one of the most efficient methods for computing the exponential of a matrix, see Kenney and Laub [85] and Higham [75]. The key idea is that $\exp(M)$ is easy to compute if M is close zero since, in this case, one can use a few terms of the exponential series, or better, a Padé approximant (see Higham [75]). The scaling and squaring method for computing the exponential of a matrix, M, can be sketched as follows:

1. *Scaling Step*: Divide M by a factor, 2^N, so that $\frac{M}{2^N}$ is close enough to zero.

2. *Exponentiation step*: Compute $\exp\left(\frac{M}{2^N}\right)$ with high precision, for example, using a Padé approximant.

3. *Squaring Step*: Square $\exp\left(\frac{M}{2^N}\right)$ repeatedly N times to obtain $\exp\left(\frac{M}{2^N}\right)^{2^N}$, a very accurate approximation of e^M.

There is also a so-called *inverse scaling and squaring method* to compute efficiently the principal logarithm of a real matrix, see Cheng, Higham, Kenney and Laub [32].

Arsigny, Commowick, Pennec and Ayache made the very astute observation that the scaling and squaring method can be adpated to compute LEPT's very efficiently [4]. This method, called *fast polyaffine transform*, computes the values of a Log-Euclidean polyaffine transformation, $T = \Phi_1$, at the vertices of a regular n-dimensional grid (in practice, for $n = 2$ or $n = 3$). Recall that T is obtained by integrating an ODE, $X' = V(X)$, where the vector field, V, is obtained by blending the vector fields associated with some affine transformations, $T_1, \ldots, T_n$, having a principal logarithm.

Here are the three steps of the **fast polyaffine transform**:

1. *Scaling Step*: Divide the vector field, V, by a factor, 2^N, so that $\frac{V}{2^N}$ is close enough to zero.

2. *Exponentiation step*: Compute $\Phi_{1/2^N}$, using some adequate numerical integration method.

3. *Squaring Step*: Compose $\Phi_{1/2^N}$ with itself recursively N times to obtain an accurate approximation of $T = \Phi_1$.

Of course, one has to provide practical methods to achieve step 2 and step 3. Several methods to achieve step 2 and step 3 are proposed in Arsigny, Commowick, Pennec and Ayache [4]. One also has to worry about boundary effects, but this problem can be alleviated too, using bounding boxes. At this point, the reader is urged to read the full paper [4] for complete details and beautiful pictures illustrating the use of LEPT's in medical imaging.

To conclude our survey of the Log-Euclidean polyaffine framework for locally affine registration, we briefly discuss how the Log-Euclidean framework can be generalized to rigid and affine transformations.

19.8 A Log-Euclidean Framework for Transformations in $\exp(\mathcal{S}(n))$

Arsigny, Commowick, Pennec and Ayache observed that if T_1 and T_2 are two affine transformations in $\exp(\mathcal{S}(n))$, then we can define their distance as

$$d(T_1, T_2) = \|\log(T_1) - \log(T_2)\|,$$

where $\| \|$ is any norm on $n \times n$ matrices (see [4], Appendix A.1). We can go a little further and make $\mathcal{S}(n)$ and $\exp(\mathcal{S}(n))$ into Riemannian manifolds in such a way that the exponential map, $\exp\colon \mathcal{S}(n) \to \exp(\mathcal{S}(n))$, is an isometry.

Since $\mathcal{S}(n)$ is an open subset of the vector space, $\mathrm{M}(n, \mathbb{R})$, of all $n \times n$ real matrices, $\mathcal{S}(n)$ is a manifold, and since $\exp(\mathcal{S}(n))$ is an open subset of the manifold, $\mathbf{GL}(n, \mathbb{R})$, it is also a manifold. Obviously, $T_L\mathcal{S}(n) \cong \mathrm{M}(n, \mathbb{R})$ and $T_S \exp(\mathcal{S}(n)) \cong \mathrm{M}(n, \mathbb{R})$, for all $L \in \mathcal{S}(n)$ and all $S \in \exp(\mathcal{S}(n))$ and the maps, $d\exp_L\colon T_L\mathcal{S}(n) \to T_{\exp(L)} \exp(\mathcal{S}(n))$ and $d\log_S\colon T_S \exp(\mathcal{S}(n)) \to T_{\log(S)}\mathcal{S}(n)$, are linear isomorphisms. We can make $\mathcal{S}(n)$ into a Riemannian manifold by giving it the induced metric induced by any norm, $\| \|$, on $\mathrm{M}(n, \mathbb{R})$, and make $\exp(\mathcal{S}(n))$ into a Riemannian manifold by defining the metric, $\langle -, - \rangle_S$, on $T_S \exp(\mathcal{S}(n))$, by

$$\langle A, B \rangle_S = \|d\log_S(A) - d\log_S(B)\|,$$

for all $S \in \exp(\mathcal{S}(n))$ and all $A, B \in \mathrm{M}(n, \mathbb{R})$. Then, it is easy to check that $\exp\colon \mathcal{S}(n) \to \exp(\mathcal{S}(n))$ is indeed an isometry and, as a consequence, the Riemannian distance between two matrices, $T_1, T_2 \in \exp(\mathcal{S}(n))$, is given by

$$d(T_1, T_2) = \|\log(T_1) - \log(T_2)\|,$$

again called the *Log-Euclidean distance*.

Since every affine transformation, T, can be represented in matrix form as

$$T = \begin{pmatrix} M & t \\ 0 & 1 \end{pmatrix},$$

and, as we saw in section 19.6, since the principal logarithm of T exists iff the principal logarithm of M exists, we can view the set of affine transformations that have a principal logarithm as a subset of $\exp(\mathcal{S}(n+1))$.

Unfortunately, this time, even though they are both flat, $\mathcal{S}(n)$ and $\exp(\mathcal{S}(n))$ are not complete manifolds and so, the Fréchet mean of N matrices, $T_1, \ldots, T_n \in \exp(\mathcal{S}(n))$, may not exist.

However, recall that from Theorem 19.1 that the open ball,

$$B(I, 1) = \{A \in \mathbf{GL}(n, \mathbb{R}) \mid \|A - I\|' < 1\},$$

is contained in $\exp(\mathcal{S}(n))$ for any norm, $\| \|'$, on matrices (not necessarily equal to the norm defining the Riemannian metric on $\mathcal{S}(n)$) such that $\|AB\|' \leq \|A\|' \, \|B\|'$ so, for any matrices $T_1, \ldots, T_n \in B(I, 1)$, the Fréchet mean is well defined and is uniquely determined by

$$\mathbb{E}_{\mathrm{LE}}(T_1, \ldots, T_N) = \exp\left(\frac{1}{N} \sum_{i=1}^{N} \log(T_i)\right),$$

namely, it is their *Log-Euclidean mean*.

From a practical point of view, one only needs to ckeck that the eigenvalues, ξ, of $\frac{1}{N}\sum_{i=1}^{N}\log(T_i)$ are in the horizontal strip, $-\pi < \Im(\xi) < \pi$.

Provided that $\mathbb{E}_{\mathrm{LE}}(T_1,\ldots,T_N)$ is defined, it is easy to show, as in the case of SPD matrices, that $\det(\mathbb{E}_{\mathrm{LE}}(T_1,\ldots,T_N))$ is the geometric mean of the determinants of the T_i's.

The Riemannian distance on $\exp(\mathcal{S}(n))$ is not affine invariant but it is invariant under inversion, under rescaling by a positive scalar, and under rotation for certain norms on $\mathcal{S}(n)$ (see [4], Appendix A.2). However, the Log-Euclidean mean of matrices in $\exp(\mathcal{S}(n))$ is invariant under conjugation by any matrix, $A \in \mathbf{GL}(n,\mathbb{R})$, since $ASA^{-1} \in \exp(\mathcal{S}(n))$ for any $S \in \exp(\mathcal{S}(n))$ and since $\log(ASA^{-1}) = A\log(S)A^{-1}$. In particular, the Log-Euclidean mean of affine transformations in $\exp(\mathcal{S}(n+1))$ is invariant under arbitrary invertible affine transformations (again, see [4], Appendix A.2).

For more details on the Log-Euclidean framework for locally rigid or affine deformation, for example, about regularization, the reader should read Arsigny, Commowick, Pennec and Ayache [4].

Chapter 20

Fréchet Mean and Statistics on Riemannian Manifolds; Applications to Medical Image Analysis

Chapter 21

Clifford Algebras, Clifford Groups, and the Groups Pin(n) and Spin(n)

21.1 Introduction: Rotations As Group Actions

The main goal of this chapter is to explain how rotations in $\mathbb{R}^n$ are induced by the action of a certain group, $\mathbf{Spin}(n)$, on $\mathbb{R}^n$, in a way that generalizes the action of the unit complex numbers, $\mathbf{U}(1)$, on $\mathbb{R}^2$, and the action of the unit quaternions, $\mathbf{SU}(2)$, on $\mathbb{R}^3$ (*i.e.*, the action is defined in terms of multiplication in a larger algebra containing both the group $\mathbf{Spin}(n)$ and $\mathbb{R}^n$). The group $\mathbf{Spin}(n)$, called a *spinor group*, is defined as a certain subgroup of units of an algebra, Cl_n, the *Clifford algebra* associated with $\mathbb{R}^n$. Furthermore, for $n \geq 3$, we are lucky, because the group $\mathbf{Spin}(n)$ is topologically simpler than the group $\mathbf{SO}(n)$. Indeed, for $n \geq 3$, the group $\mathbf{Spin}(n)$ is simply connected (a fact that it not so easy to prove without some machinery), whereas $\mathbf{SO}(n)$ is not simply connected. Intuitively speaking, $\mathbf{SO}(n)$ is more twisted than $\mathbf{Spin}(n)$. In fact, we will see that $\mathbf{Spin}(n)$ is a double cover of $\mathbf{SO}(n)$.

Since the spinor groups are certain well chosen subroups of units of Clifford algebras, it is necessary to investigate Clifford algebras to get a firm understanding of spinor groups. This chapter provides a tutorial on Clifford algebra and the groups **Spin** and **Pin**, including a study of the structure of the Clifford algebra $\mathrm{Cl}_{p,q}$ associated with a nondegenerate symmetric bilinear form of signature (p, q) and culminating in the beautiful "8-periodicity theorem" of Elie Cartan and Raoul Bott (with proofs). We also explain when $\mathbf{Spin}(p, q)$ is a double-cover of $\mathbf{SO}(p, q)$. The reader should be warned that a certain amount of algebraic (and topological) background is expected. This being said, perseverant readers will be rewarded by being exposed to some beautiful and nontrivial concepts and results, including Elie Cartan and Raoul Bott "8-periodicity theorem."

Going back to rotations as transformations induced by group actions, recall that if V is a vector space, a *linear action (on the left) of a group G on V* is a map, $\alpha\colon G \times V \to V$, satisfying the following conditions, where, for simplicity of notation, we denote $\alpha(g, v)$ by $g \cdot v$:

(1) $g \cdot (h \cdot v) = (gh) \cdot v$, for all $g, h \in G$ and $v \in V$;

(2) $1 \cdot v = v$, for all $v \in V$, where 1 is the identity of the group G;

(3) The map $v \mapsto g \cdot v$ is a linear isomorphism of V for every $g \in G$.

For example, the (multiplicative) group, $\mathbf{U}(1)$, of unit complex numbers acts on $\mathbb{R}^2$ (by identifying $\mathbb{R}^2$ and $\mathbb{C}$) *via* complex multiplication: For every $z = a + ib$ (with $a^2 + b^2 = 1$), for every $(x, y) \in \mathbb{R}^2$ (viewing (x, y) as the complex number $x + iy$),

$$z \cdot (x, y) = (ax - by, ay + bx).$$

Now, every unit complex number is of the form $\cos\theta + i\sin\theta$, and thus, the above action of $z = \cos\theta + i\sin\theta$ on $\mathbb{R}^2$ corresponds to the rotation of angle θ around the origin. In the case $n = 2$, the groups $\mathbf{U}(1)$ and $\mathbf{SO}(2)$ are isomorphic, but this is an exception.

To represent rotations in $\mathbb{R}^3$ and $\mathbb{R}^4$, we need the quaternions. For our purposes, it is convenient to define the quaternions as certain 2×2 complex matrices. Let $\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{k}$ be the matrices

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{i} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \qquad \mathbf{j} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad \mathbf{k} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix},$$

and let $\mathbb{H}$ be the set of all matrices of the form

$$X = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d, \quad a, b, c, d \in \mathbb{R}.$$

Thus, every matrix in $\mathbb{H}$ is of the form

$$X = \begin{pmatrix} a + ib & c + id \\ -(c - id) & a - ib \end{pmatrix}, \quad a, b, c, d \in \mathbb{R}.$$

The quaternions $\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{k}$ satisfy the famous identities discovered by Hamilton:

$$\begin{aligned} &\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -\mathbf{1}, \\ &\mathbf{ij} = -\mathbf{ji} = \mathbf{k}, \\ &\mathbf{jk} = -\mathbf{kj} = \mathbf{i}, \\ &\mathbf{ki} = -\mathbf{ik} = \mathbf{j}. \end{aligned}$$

As a consequence, it can be verified that $\mathbb{H}$ is a skew field (a noncommutative field) called the *quaternions*. It is also a real vector space of dimension 4 with basis $(\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{k})$; thus, as a vector space, $\mathbb{H}$ is isomorphic to $\mathbb{R}^4$. The *unit quaternions* are the quaternions such that

$$\det(X) = a^2 + b^2 + c^2 + d^2 = 1.$$

It is easy to check that the matrices associated with the unit quaternions are exactly the matrices in $\mathbf{SU}(2)$. Thus, we call $\mathbf{SU}(2)$ the group of unit quaternions.

Now we can define an action of the group of unit quaternions, $\mathbf{SU}(2)$, on $\mathbb{R}^3$. For this, we use the fact that $\mathbb{R}^3$ can be identified with the pure quaternions in $\mathbb{H}$, namely, the quaternions of the form $x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k}$, where $(x_1, x_2, x_3) \in \mathbb{R}^3$. Then, we define the action of $\mathbf{SU}(2)$ over $\mathbb{R}^3$ by

$$Z \cdot X = ZXZ^{-1} = ZX\overline{Z},$$

where $Z \in \mathbf{SU}(2)$ and X is any pure quaternion. Now, it turns out that the map ρ_Z (where $\rho_Z(X) = ZX\overline{Z}$) is indeed a rotation, and that the map $\rho\colon Z \mapsto \rho_Z$ is a surjective homomorphism, $\rho\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$, whose kernel is $\{-\mathbf{1}, \mathbf{1}\}$, where $\mathbf{1}$ denotes the multiplicative unit quaternion. (For details, see Gallier [60], Chapter 8).

We can also define an action of the group $\mathbf{SU}(2) \times \mathbf{SU}(2)$ over $\mathbb{R}^4$, by identifying $\mathbb{R}^4$ with the quaternions. In this case,

$$(Y, Z) \cdot X = YX\overline{Z},$$

where $(Y, Z) \in \mathbf{SU}(2)\times\mathbf{SU}(2)$ and $X \in \mathbb{H}$ is any quaternion. Then, the map $\rho_{Y,\overline{Z}}$ is a rotation (where $\rho_{Y,\overline{Z}}(X) = YX\overline{Z}$), and the map $\rho\colon (Y, Z) \mapsto \rho_{Y,\overline{Z}}$ is a surjective homomorphism, $\rho\colon \mathbf{SU}(2) \times \mathbf{SU}(2) \to \mathbf{SO}(4)$, whose kernel is $\{(\mathbf{1}, \mathbf{1}), (-\mathbf{1}, -\mathbf{1})\}$. (For details, see Gallier [60], Chapter 8).

Thus, we observe that for $n = 2, 3, 4$, the rotations in $\mathbf{SO}(n)$ can be realized *via* the linear action of some group (the case $n = 1$ is trivial, since $\mathbf{SO}(1) = \{1, -1\}$). It is also the case that the action of each group can be somehow be described in terms of multiplication in some larger algebra "containing" the original vector space $\mathbb{R}^n$ ($\mathbb{C}$ for $n = 2$, $\mathbb{H}$ for $n = 3, 4$). However, these groups appear to have been discovered in an ad hoc fashion, and there does not appear to be any universal way to define the action of these groups on $\mathbb{R}^n$. It would certainly be nice if the action was always of the form

$$Z \cdot X = ZXZ^{-1}(= ZX\overline{Z}).$$

A systematic way of constructing groups realizing rotations in terms of linear action, using a uniform notion of action, does exist. Such groups are the spinor groups, to be described in the following sections.

21.2 Clifford Algebras

We explained in Section 21.1 how the rotations in $\mathbf{SO}(3)$ can be realized by the linear action of the group of unit quaternions, $\mathbf{SU}(2)$, on $\mathbb{R}^3$, and how the rotations in $\mathbf{SO}(4)$ can be realized by the linear action of the group $\mathbf{SU}(2) \times \mathbf{SU}(2)$ on $\mathbb{R}^4$.

The main reasons why the rotations in $\mathbf{SO}(3)$ can be represented by unit quaternions are the following:

(1) For every nonzero vector $u \in \mathbb{R}^3$, the reflection s_u about the hyperplane perpendicular to u is represented by the map

$$v \mapsto -uvu^{-1},$$

where u and v are viewed as pure quaternions in $\mathbb{H}$ (i.e., if $u = (u_1, u_2, u_2)$, then view u as $u_1\mathbf{i} + u_2\mathbf{j} + u_3\mathbf{k}$, and similarly for v).

(2) The group $\mathbf{SO}(3)$ is generated by the reflections.

As one can imagine, a successful generalization of the quaternions, i.e., the discovery of a group, G inducing the rotations in $\mathbf{SO}(n)$ *via* a linear action, depends on the ability to generalize properties (1) and (2) above. Fortunately, it is true that the group $\mathbf{SO}(n)$ is generated by the hyperplane reflections. In fact, this is also true for the orthogonal group, $\mathbf{O}(n)$, and more generally, for the group of direct isometries, $\mathbf{O}(\Phi)$, of any nondegenerate quadratic form, Φ, by the *Cartan-Dieudonné theorem* (for instance, see Bourbaki [20], or Gallier [60], Chapter 7, Theorem 7.2.1). In order to generalize (2), we need to understand how the group G acts on $\mathbb{R}^n$. Now, the case $n = 3$ is special, because the underlying space, $\mathbb{R}^3$, on which the rotations act, can be embedded as the pure quaternions in $\mathbb{H}$. The case $n = 4$ is also special, because $\mathbb{R}^4$ is the underlying space of $\mathbb{H}$. The generalization to $n \geq 5$ requires more machinery, namely, the notions of Clifford groups and Clifford algebras. As we will see, for every $n \geq 2$, there is a compact, connected (and simply connected when $n \geq 3$) group, $\mathbf{Spin}(n)$, the "spinor group," and a surjective homomorphism, $\rho\colon \mathbf{Spin}(n) \to \mathbf{SO}(n)$, whose kernel is $\{-1, 1\}$. This time, $\mathbf{Spin}(n)$ acts directly on $\mathbb{R}^n$, because $\mathbf{Spin}(n)$ is a certain subgroup of the group of units of the *Clifford algebra*, Cl_n, and $\mathbb{R}^n$ is naturally a subspace of Cl_n.

The group of unit quaternions $\mathbf{SU}(2)$ turns out to be isomorphic to the spinor group $\mathbf{Spin}(3)$. Because $\mathbf{Spin}(3)$ acts directly on $\mathbb{R}^3$, the representation of rotations in $\mathbf{SO}(3)$ by elements of $\mathbf{Spin}(3)$ may be viewed as more natural than the representation by unit quaternions. The group $\mathbf{SU}(2) \times \mathbf{SU}(2)$ turns out to be isomorphic to the spinor group $\mathbf{Spin}(4)$, but this isomorphism is less obvious.

In summary, we are going to define a group $\mathbf{Spin}(n)$ representing the rotations in $\mathbf{SO}(n)$, for any $n \geq 1$, in the sense that there is a linear action of $\mathbf{Spin}(n)$ on $\mathbb{R}^n$ which induces a surjective homomorphism, $\rho\colon \mathbf{Spin}(n) \to \mathbf{SO}(n)$, whose kernel is $\{-1, 1\}$. Furthermore, the action of $\mathbf{Spin}(n)$ on $\mathbb{R}^n$ is given in terms of multiplication in an algebra, Cl_n, containing $\mathbf{Spin}(n)$, and in which $\mathbb{R}^n$ is also embedded. It turns out that as a bonus, for $n \geq 3$, the group $\mathbf{Spin}(n)$ is topologically simpler than $\mathbf{SO}(n)$, since $\mathbf{Spin}(n)$ is simply connected, but $\mathbf{SO}(n)$ is not. By being astute, we can also construct a group, $\mathbf{Pin}(n)$, and a linear action of $\mathbf{Pin}(n)$ on $\mathbb{R}^n$ that induces a surjective homomorphism, $\rho\colon \mathbf{Pin}(n) \to \mathbf{O}(n)$, whose kernel is $\{-1, 1\}$. The difficulty here is the presence of the negative sign in (2). We will see how Atiyah, Bott and Shapiro circumvent this problem by using a "twisted adjoint action," as opposed to the usual adjoint action (where $v \mapsto uvu^{-1}$).

Our presentation is heavily influenced by Bröcker and tom Dieck [25], Chapter 1, Section 6, where most details can be found. This Chapter is almost entirely taken from the first 11 pages of the beautiful and seminal paper by Atiyah, Bott and Shapiro [11], Clifford Modules, and we highly recommend it. Another excellent (but concise) exposition can be found in Kirillov [86]. A very thorough exposition can be found in two places:

1. Lawson and Michelsohn [97], where the material on $\mathbf{Pin}(p,q)$ and $\mathbf{Spin}(p,q)$ can be found in Chapter I.

2. Lounesto's excellent book [100].

One may also want to consult Baker [13], Curtis [39], Porteous [125], Fulton and Harris (Lecture 20) [58], Choquet-Bruhat [36], Bourbaki [20], or Chevalley [35], a classic. The original source is Elie Cartan's book (1937) whose translation in English appears in [28].

We begin by recalling what is an algebra over a field. Let K denote any (commutative) field, although for our purposes, we may assume that $K = \mathbb{R}$ (and occasionally, $K = \mathbb{C}$). Since we will only be dealing with associative algebras with a multiplicative unit, we only define algebras of this kind.

Definition 21.1. Given a field, K, a *K-algebra* is a K-vector space, A, together with a bilinear operation, $\cdot\colon A \times A \to A$, called *multiplication*, which makes A into a ring with unity, 1 (or 1_A, when we want to be very precise). This means that $\cdot$ is associative and that there is a multiplicative identity element, 1, so that $1 \cdot a = a \cdot 1 = a$, for all $a \in A$. Given two K-algebras A and B, a *K-algebra homomorphism,* $h\colon A \to B$, is a linear map that is also a ring homomorphism, with $h(1_A) = 1_B$.

For example, the ring, $M_n(K)$, of all $n \times n$ matrices over a field, K, is a K-algebra.

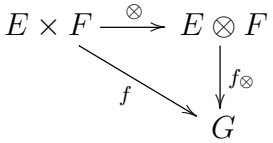
There is an obvious notion of *ideal* of a K-algebra: An ideal, $\mathfrak{A} \subseteq A$, is a linear subspace of A that is also a two-sided ideal with respect to multiplication in A. If the field K is understood, we usually simply say an algebra instead of a K-algebra.

We will also need tensor products. A rather detailed exposition of tensor products is given in Chapter 22 and the reader may want to review Section 22.1. For the reader's convenience, we recall the definition of the tensor product of vector spaces. The basic idea is that tensor products allow us to view multilinear maps as linear maps. The maps become simpler, but the spaces (product spaces) become more complicated (tensor products). For more details, see Section 22.1 or Atiyah and Macdonald [9].

Definition 21.2. Given two K-vector spaces, E and F, a *tensor product of E and F* is a pair, $(E \otimes F, \otimes)$, where $E \otimes F$ is a K-vector space and $\otimes\colon E \times F \to E \otimes F$ is a bilinear map, so that for every K-vector space, G, and every bilinear map, $f\colon E \times F \to G$, there is a unique linear map, $f_\otimes\colon E \otimes F \to G$, with

$$f(u,v) = f_\otimes(u \otimes v) \quad \text{for all } u \in E \text{ and all } v \in V,$$

as in the diagram below:



The vector space $E \otimes F$ is defined up to isomorphism. The vectors $u \otimes v$, where $u \in E$ and $v \in F$, generate $E \otimes F$.

Remark: We should really denote the tensor product of E and F by $E \otimes_K F$, since it depends on the field K. Since we usually deal with a fixed field K, we use the simpler notation $E \otimes F$.

As shown in Section 22.3, we have natural isomorphisms

$$(E \otimes F) \otimes G \approx E \otimes (F \otimes G) \quad \text{and} \quad E \otimes F \approx F \otimes E.$$

Given two linear maps $f\colon E \to F$ and $g\colon E' \to F'$, we have a unique bilinear map $f \times g\colon E \times E' \to F \times F'$ so that

$$(f \times g)(a, a') = (f(a), g(a')) \quad \text{for all } a \in E \text{ and all } a' \in E'.$$

Thus, we have the bilinear map $\otimes \circ (f \times g)\colon E \times E' \to F \otimes F'$, and so, there is a unique linear map $f \otimes g\colon E \otimes E' \to F \otimes F'$, so that

$$(f \otimes g)(a \otimes a') = f(a) \otimes g(a') \quad \text{for all } a \in E \text{ and all } a' \in E'.$$

Let us now assume that E and F are K-algebras. We want to make $E \otimes F$ into a K-algebra. Since the multiplication operations $m_E\colon E \times E \to E$ and $m_F\colon F \times F \to F$ are bilinear, we get linear maps $m'_E\colon E \otimes E \to E$ and $m'_F\colon F \otimes F \to F$, and thus, the linear map

$$m'_E \otimes m'_F\colon (E \otimes E) \otimes (F \otimes F) \to E \otimes F.$$

Using the isomorphism $\tau\colon (E \otimes E) \otimes (F \otimes F) \to (E \otimes F) \otimes (E \otimes F)$, we get a linear map

$$m_{E \otimes F}\colon (E \otimes F) \otimes (E \otimes F) \to E \otimes F,$$

which defines a multiplication m on $E \otimes F$ (namely, $m(u, v) = m_{E \otimes F}(u \otimes v)$). It is easily checked that $E \otimes F$ is indeed a K-algebra under the multiplication m. Using the simpler notation $\cdot$ for m, we have

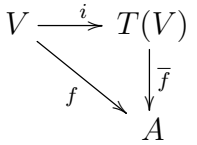
$$(a \otimes a') \cdot (b \otimes b') = (ab) \otimes (a'b')$$

for all $a, b \in E$ and all $a', b' \in F$.

Given any vector space, V, over a field, K, there is a special K-algebra, $T(V)$, together with a linear map, $i\colon V \to T(V)$, with the following universal mapping property: Given any K-algebra, A, for any linear map, $f\colon V \to A$, there is a unique K-algebra homomorphism, $\overline{f}\colon T(V) \to A$, so that

$$f = \overline{f} \circ i,$$

as in the diagram below:



The algebra, $T(V)$, is the *tensor algebra of* V, see Section 22.5. The algebra $T(V)$ may be constructed as the direct sum

$$T(V) = \bigoplus_{i \geq 0} V^{\otimes i},$$

where $V^0 = K$, and $V^{\otimes i}$ is the i-fold tensor product of V with itself. For every $i \geq 0$, there is a natural injection $\iota_n \colon V^{\otimes n} \to T(V)$, and in particular, an injection $\iota_0 \colon K \to T(V)$. The multiplicative unit, $\mathbf{1}$, of $T(V)$ is the image, $\iota_0(1)$, in $T(V)$ of the unit, 1, of the field K. Since every $v \in T(V)$ can be expressed as a finite sum

$$v = v_1 + \cdots + v_k,$$

where $v_i \in V^{\otimes n_i}$ and the n_i are natural numbers with $n_i \neq n_j$ if $i \neq j$, to define multiplication in $T(V)$, using bilinearity, it is enough to define the multiplication $V^{\otimes m} \times V^{\otimes n} \longrightarrow V^{\otimes (m+n)}$. Of course, this is defined by

$$(v_1 \otimes \cdots \otimes v_m) \cdot (w_1 \otimes \cdots \otimes w_n) = v_1 \otimes \cdots \otimes v_m \otimes w_1 \otimes \cdots \otimes w_n.$$

(This has to be made rigorous by using isomorphisms involving the associativity of tensor products, for details, see see Atiyah and Macdonald [9].) The algebra $T(V)$ is an example of a *graded algebra*, where the *homogeneous elements of rank* n are the elements in $V^{\otimes n}$.

Remark: It is important to note that multiplication in $T(V)$ is **not** commutative. Also, in all rigor, the unit, $\mathbf{1}$, of $T(V)$ is **not equal** to 1, the unit of the field K. However, in view of the injection $\iota_0 \colon K \to T(V)$, for the sake of notational simplicity, we will denote $\mathbf{1}$ by 1. More generally, in view of the injections $\iota_n \colon V^{\otimes n} \to T(V)$, we identify elements of $V^{\otimes n}$ with their images in $T(V)$.

Most algebras of interest arise as well-chosen quotients of the tensor algebra $T(V)$. This is true for the *exterior algebra,* $\bigwedge^{\bullet} V$ (also called *Grassmann algebra*), where we take the quotient of $T(V)$ modulo the ideal generated by all elements of the form $v \otimes v$, where $v \in V$, see Section 22.15.

A Clifford algebra may be viewed as a refinement of the exterior algebra, in which we take the quotient of $T(V)$ modulo the ideal generated by all elements of the form $v \otimes v - \Phi(v) \cdot 1$, where Φ is the quadratic form associated with a symmetric bilinear form, $\varphi \colon V \times V \to K$, and $\cdot \colon K \times T(V) \to T(V)$ denotes the scalar product of the algebra $T(V)$. For simplicity, let us assume that we are now dealing with real algebras.

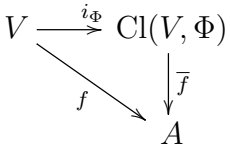
Definition 21.3. Let V be a real finite-dimensional vector space together with a symmetric bilinear form, $\varphi \colon V \times V \to \mathbb{R}$, and associated quadratic form, $\Phi(v) = \varphi(v, v)$. A *Clifford algebra associated with* V *and* Φ is a real algebra, $\mathrm{Cl}(V, \Phi)$, together with a linear map, $i_\Phi \colon V \to \mathrm{Cl}(V, \Phi)$, satisfying the condition $(i(v))^2 = \Phi(v) \cdot 1$ for all $v \in V$ and so that for every real algebra, A, and every linear map, $f \colon V \to A$, with

$$(f(v))^2 = \Phi(v) \cdot 1 \quad \text{for all } v \in V,$$

there is a unique algebra homomorphism, $\overline{f}\colon \mathrm{Cl}(V,\Phi) \to A$, so that

$$f = \overline{f} \circ i_\Phi,$$

as in the diagram below:



We use the notation, $\lambda \cdot u$, for the product of a scalar, $\lambda \in \mathbb{R}$, and of an element, u, in the algebra $\mathrm{Cl}(V,\Phi)$ and juxtaposition, uv, for the multiplication of two elements, u and v, in the algebra $\mathrm{Cl}(V,\Phi)$.

By a familiar argument, any two Clifford algebras associated with V and Φ are isomorphic. We often denote i_Φ by i.

To show the existence of $\mathrm{Cl}(V,\Phi)$, observe that $T(V)/\mathfrak{A}$ does the job, where $\mathfrak{A}$ is the ideal of $T(V)$ generated by all elements of the form $v \otimes v - \Phi(v) \cdot 1$, where $v \in V$. The map $i_\Phi\colon V \to \mathrm{Cl}(V,\Phi)$ is the composition

$$V \xrightarrow{\iota_1} T(V) \xrightarrow{\pi} T(V)/\mathfrak{A},$$

where π is the natural quotient map. We often denote the Clifford algebra $\mathrm{Cl}(V,\Phi)$ simply by $\mathrm{Cl}(\Phi)$.

Remark: Observe that Definition 21.3 does not assert that i_Φ is injective or that there is an injection of $\mathbb{R}$ into $\mathrm{Cl}(V,\Phi)$, but we will prove later that both facts are true when V is finite-dimensional. Also, as in the case of the tensor algebra, the unit of the algebra $\mathrm{Cl}(V,\Phi)$ and the unit of the field $\mathbb{R}$ are **not equal**.

Since

$$\Phi(u+v) - \Phi(u) - \Phi(v) = 2\varphi(u,v)$$

and

$$(i(u+v))^2 = (i(u))^2 + (i(v))^2 + i(u)i(v) + i(v)i(u),$$

using the fact that

$$i(u)^2 = \Phi(u) \cdot 1,$$

we get

$$i(u)i(v) + i(v)i(u) = 2\varphi(u,v) \cdot 1.$$

As a consequence, if $(u_1, \ldots, u_n)$ is an orthogonal basis w.r.t. φ (which means that $\varphi(u_j, u_k) = 0$ for all $j \neq k$), we have

$$i(u_j)i(u_k) + i(u_k)i(u_j) = 0 \quad \text{for all } j \neq k.$$

Remark: Certain authors drop the unit, 1, of the Clifford algebra $\mathrm{Cl}(V, \Phi)$ when writing the identities

$$i(u)^2 = \Phi(u) \cdot 1$$

and

$$2\varphi(u, v) \cdot 1 = i(u)i(v) + i(v)i(u),$$

where the second identity is often written as

$$\varphi(u, v) = \frac{1}{2}(i(u)i(v) + i(v)i(u)).$$

This is very confusing and technically wrong, because we only have an injection of $\mathbb{R}$ into $\mathrm{Cl}(V, \Phi)$, but $\mathbb{R}$ is **not** a subset of $\mathrm{Cl}(V, \Phi)$.

We warn the readers that Lawson and Michelsohn [97] adopt the opposite of our sign convention in defining Clifford algebras, i.e., they use the condition

$$(f(v))^2 = -\Phi(v) \cdot 1 \quad \text{for all } v \in V.$$

The most confusing consequence of this is that their $\mathrm{Cl}(p, q)$ is our $\mathrm{Cl}(q, p)$.

Observe that when $\Phi \equiv 0$ is the quadratic form identically zero everywhere, then the Clifford algebra $\mathrm{Cl}(V, 0)$ is just the exterior algebra, $\bigwedge^\bullet V$.

Example 21.1. Let $V = \mathbb{R}$, $e_1 = 1$, and assume that $\Phi(x_1 e_1) = -x_1^2$. Then, $\mathrm{Cl}(\Phi)$ is spanned by the basis $(1, e_1)$. We have

$$e_1^2 = -1.$$

Under the bijection

$$e_1 \mapsto i,$$

the Clifford algebra, $\mathrm{Cl}(\Phi)$, also denoted by Cl_1, is isomorphic to the algebra of complex numbers, $\mathbb{C}$.

Now, let $V = \mathbb{R}^2$, (e_1, e_2) be the canonical basis, and assume that $\Phi(x_1 e_1 + x_2 e_2) = -(x_1^2 + x_2^2)$. Then, $\mathrm{Cl}(\Phi)$ is spanned by the basis by $(1, e_1, e_2, e_1 e_2)$. Furthermore, we have

$$e_2 e_1 = -e_1 e_2, \quad e_1^2 = -1, \quad e_2^2 = -1, \quad (e_1 e_2)^2 = -1.$$

Under the bijection

$$e_1 \mapsto \mathbf{i}, \quad e_2 \mapsto \mathbf{j}, \quad e_1 e_2 \mapsto \mathbf{k},$$

it is easily checked that the quaternion identities

$$\begin{aligned} \mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 &= -\mathbf{1}, \\ \mathbf{ij} &= -\mathbf{ji} = \mathbf{k}, \\ \mathbf{jk} &= -\mathbf{kj} = \mathbf{i}, \\ \mathbf{ki} &= -\mathbf{ik} = \mathbf{j}, \end{aligned}$$

hold, and thus, the Clifford algebra $\mathrm{Cl}(\Phi)$, also denoted by Cl_2, is isomorphic to the algebra of quaternions, $\mathbb{H}$.

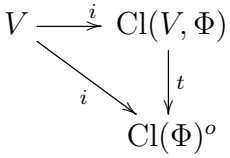
Our prime goal is to define an action of $\mathrm{Cl}(\Phi)$ on V in such a way that by restricting this action to some suitably chosen multiplicative subgroups of $\mathrm{Cl}(\Phi)$, we get surjective homomorphisms onto $\mathbf{O}(\Phi)$ and $\mathbf{SO}(\Phi)$, respectively. The key point is that a reflection in V about a hyperplane H orthogonal to a vector w can be defined by such an action, but some negative sign shows up. A correct handling of signs is a bit subtle and requires the introduction of a canonical anti-automorphism, t, and of a canonical automorphism, α, defined as follows:

Proposition 21.1. *Every Clifford algebra,* $\mathrm{Cl}(\Phi)$*, possesses a canonical anti-automorphism,* $t\colon \mathrm{Cl}(\Phi) \to \mathrm{Cl}(\Phi)$*, satisfying the properties*

$$t(xy) = t(y)t(x), \quad t \circ t = \mathrm{id}, \quad \textit{and} \quad t(i(v)) = i(v),$$

for all $x, y \in \mathrm{Cl}(\Phi)$ *and all* $v \in V$*. Furthermore, such an anti-automorphism is unique.*

Proof. Consider the opposite algebra $\mathrm{Cl}(\Phi)^o$, in which the product of x and y is given by yx. It has the universal mapping property. Thus, we get a unique isomorphism, t, as in the diagram below:



□

We also denote $t(x)$ by x^t. When V is finite-dimensional, for a more palatable description of t in terms of a basis of V, see the paragraph following Theorem 21.4.

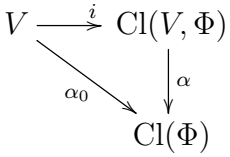
The canonical automorphism, α, is defined using the proposition

Proposition 21.2. *Every Clifford algebra,* $\mathrm{Cl}(\Phi)$*, has a unique canonical automorphism,* $\alpha\colon \mathrm{Cl}(\Phi) \to \mathrm{Cl}(\Phi)$*, satisfying the properties*

$$\alpha \circ \alpha = \mathrm{id}, \quad \textit{and} \quad \alpha(i(v)) = -i(v),$$

for all $v \in V$*.*

Proof. Consider the linear map $\alpha_0\colon V \to \mathrm{Cl}(\Phi)$ defined by $\alpha_0(v) = -i(v)$, for all $v \in V$. We get a unique homomorphism, α, as in the diagram below:



Furthermore, every $x \in \mathrm{Cl}(\Phi)$ can be written as

$$x = x_1 \cdots x_m,$$

with $x_j \in i(V)$, and since $\alpha(x_j) = -x_j$, we get $\alpha \circ \alpha = \mathrm{id}$. It is clear that α is bijective. □

Again, when V is finite-dimensional, a more palatable description of α in terms of a basis of V can be given. If $(e_1, \ldots, e_n)$ is a basis of V, then the Clifford algebra $\mathrm{Cl}(\Phi)$ consists of certain kinds of "polynomials," linear combinations of monomials of the form $\sum_J \lambda_J e_J$, where $J = \{i_1, i_2, \ldots, i_k\}$ is any subset (possibly empty) of $\{1, \ldots, n\}$, with $1 \leq i_1 < i_2 \cdots < i_k \leq n$, and the monomial e_J is the "product" $e_{i_1} e_{i_2} \cdots e_{i_k}$. The map α is the linear map defined on monomials by

$$\alpha(e_{i_1} e_{i_2} \cdots e_{i_k}) = (-1)^k e_{i_1} e_{i_2} \cdots e_{i_k}.$$

For a more rigorous explanation, see the paragraph following Theorem 21.4.

We now show that if V has dimension n, then i is injective and $\mathrm{Cl}(\Phi)$ has dimension 2^n. A clever way of doing this is to introduce a graded tensor product.

First, observe that

$$\mathrm{Cl}(\Phi) = \mathrm{Cl}^0(\Phi) \oplus \mathrm{Cl}^1(\Phi),$$

where

$$\mathrm{Cl}^i(\Phi) = \{x \in \mathrm{Cl}(\Phi) \mid \alpha(x) = (-1)^i x\}, \quad \text{where } i = 0, 1.$$

We say that we have a $\mathbb{Z}/2$-*grading*, which means that if $x \in \mathrm{Cl}^i(\Phi)$ and $y \in \mathrm{Cl}^j(\Phi)$, then $xy \in \mathrm{Cl}^{i+j \,(\mathrm{mod}\, 2)}(\Phi)$.

When V is finite-dimensional, since every element of $\mathrm{Cl}(\Phi)$ is a linear combination of the form $\sum_J \lambda_J e_J$, as explained earlier, in view of the description of α given above, we see that the elements of $\mathrm{Cl}^0(\Phi)$ are those for which the monomials e_J are products of an even number of factors, and the elements of $\mathrm{Cl}^1(\Phi)$ are those for which the monomials e_J are products of an odd number of factors.

Remark: Observe that $\mathrm{Cl}^0(\Phi)$ is a subalgebra of $\mathrm{Cl}(\Phi)$, whereas $\mathrm{Cl}^1(\Phi)$ is not.

Given two $\mathbb{Z}/2$-graded algebras $A = A^0 \oplus A^1$ and $B = B^0 \oplus B^1$, their *graded tensor product* $A \mathbin{\widehat{\otimes}} B$ is defined by

$$\begin{aligned} (A \mathbin{\widehat{\otimes}} B)^0 &= (A^0 \oplus B^0) \otimes (A^1 \oplus B^1), \\ (A \mathbin{\widehat{\otimes}} B)^1 &= (A^0 \oplus B^1) \otimes (A^1 \oplus B^0), \end{aligned}$$

with multiplication

$$(a' \otimes b)(a \otimes b') = (-1)^{ij}(a'a) \otimes (bb'),$$

for $a \in A^i$ and $b \in B^j$. The reader should check that $A \mathbin{\widehat{\otimes}} B$ is indeed $\mathbb{Z}/2$-graded.

Proposition 21.3. *Let V and W be finite dimensional vector spaces with quadratic forms Φ and Ψ. Then, there is a quadratic form, $\Phi \oplus \Psi$, on $V \oplus W$ defined by*

$$(\Phi + \Psi)(v, w) = \Phi(v) + \Psi(w).$$

If we write $i \colon V \to \mathrm{Cl}(\Phi)$ and $j \colon W \to \mathrm{Cl}(\Psi)$, we can define a linear map,

$$f \colon V \oplus W \to \mathrm{Cl}(\Phi) \mathbin{\widehat{\otimes}} \mathrm{Cl}(\Psi),$$

by

$$f(v, w) = i(v) \otimes 1 + 1 \otimes j(w).$$

Furthermore, the map f induces an isomorphism (also denoted by f)

$$f\colon \mathrm{Cl}(V \oplus W) \to \mathrm{Cl}(\Phi) \mathbin{\widehat{\otimes}} \mathrm{Cl}(\Psi).$$

Proof. See Bröcker and tom Dieck [25], Chapter 1, Section 6, page 57. □

As a corollary, we obtain the following result:

Theorem 21.4. *For every vector space, V, of finite dimension n, the map $i\colon V \to \mathrm{Cl}(\Phi)$ is injective. Given a basis $(e_1, \ldots, e_n)$ of V, the $2^n - 1$ products*

$$i(e_{i_1})i(e_{i_2}) \cdots i(e_{i_k}), \quad 1 \leq i_1 < i_2 \cdots < i_k \leq n,$$

and 1 form a basis of $\mathrm{Cl}(\Phi)$. Thus, $\mathrm{Cl}(\Phi)$ has dimension 2^n.

Proof. The proof is by induction on $n = \dim(V)$. For $n = 1$, the tensor algebra $T(V)$ is just the polynomial ring $\mathbb{R}[X]$, where $i(e_1) = X$. Thus, $\mathrm{Cl}(\Phi) = \mathbb{R}[X]/(X^2 - \Phi(e_1))$, and the result is obvious. Since

$$i(e_j)i(e_k) + i(e_k)i(e_j) = 2\varphi(e_i, e_j) \cdot 1,$$

it is clear that the products

$$i(e_{i_1})i(e_{i_2}) \cdots i(e_{i_k}), \quad 1 \leq i_1 < i_2 \cdots < i_k \leq n,$$

and 1 generate $\mathrm{Cl}(\Phi)$. Now, there is always a basis that is orthogonal with respect to φ (for example, see Artin [7], Chapter 7, or Gallier [60], Chapter 6, Problem 6.14), and thus, we have a splitting

$$(V, \Phi) = \bigoplus_{k=1}^{n} (V_k, \Phi_k),$$

where V_k has dimension 1. Choosing a basis so that $e_k \in V_k$, the theorem follows by induction from Proposition 21.3. □

Since i is injective, for simplicity of notation, from now on, we write u for $i(u)$. Theorem 21.4 implies that if $(e_1, \ldots, e_n)$ is an orthogonal basis of V, then $\mathrm{Cl}(\Phi)$ is the algebra presented by the generators $(e_1, \ldots, e_n)$ and the relations

$$\begin{aligned} e_j^2 &= \Phi(e_j) \cdot 1, \quad 1 \leq j \leq n, \quad \text{and} \\ e_j e_k &= -e_k e_j, \quad 1 \leq j, k \leq n,\ j \neq k. \end{aligned}$$

If V has finite dimension n and $(e_1, \ldots, e_n)$ is a basis of V, by Theorem 21.4, the maps t and α are completely determined by their action on the basis elements. Namely, t is defined by

$$\begin{aligned} t(e_i) &= e_i \\ t(e_{i_1} e_{i_2} \cdots e_{i_k}) &= e_{i_k} e_{i_{k-1}} \cdots e_{i_1}, \end{aligned}$$

where $1 \leq i_1 < i_2 \cdots < i_k \leq n$, and, of course, $t(1) = 1$. The map α is defined by

$$\begin{aligned} \alpha(e_i) &= -e_i \\ \alpha(e_{i_1} e_{i_2} \cdots e_{i_k}) &= (-1)^k e_{i_1} e_{i_2} \cdots e_{i_k} \end{aligned}$$

where $1 \leq i_1 < i_2 \cdots < i_k \leq n$, and, of course, $\alpha(1) = 1$. Furthermore, the even-graded elements (the elements of $\mathrm{Cl}^0(\Phi)$) are those generated by 1 and the basis elements consisting of an even number of factors, $e_{i_1} e_{i_2} \cdots e_{i_{2k}}$, and the odd-graded elements (the elements of $\mathrm{Cl}^1(\Phi)$) are those generated by the basis elements consisting of an odd number of factors, $e_{i_1} e_{i_2} \cdots e_{i_{2k+1}}$.

We are now ready to define the Clifford group and investigate some of its properties.

21.3 Clifford Groups

First, we define *conjugation* on a Clifford algebra, $\mathrm{Cl}(\Phi)$, as the map

$$x \mapsto \overline{x} = t(\alpha(x)) \quad \text{for all } x \in \mathrm{Cl}(\Phi).$$

Observe that

$$t \circ \alpha = \alpha \circ t.$$

If V has finite dimension n and $(e_1, \ldots, e_n)$ is a basis of V, in view of previous remarks, conjugation is defined by

$$\begin{aligned} \overline{e_i} &= -e_i \\ \overline{e_{i_1} e_{i_2} \cdots e_{i_k}} &= (-1)^k e_{i_k} e_{i_{k-1}} \cdots e_{i_1} \end{aligned}$$

where $1 \leq i_1 < i_2 \cdots < i_k \leq n$, and, of course, $\overline{1} = 1$. Conjugation is an anti-automorphism.

The multiplicative group of invertible elements of $\mathrm{Cl}(\Phi)$ is denoted by $\mathrm{Cl}(\Phi)^*$. Observe that for any $x \in V$, if $\Phi(x) \neq 0$, then x is invertible because $x^2 = \Phi(x)$; that is, $x \in \mathrm{Cl}(\Phi)^*$.

Definition 21.4. Given a finite dimensional vector space, V, and a quadratic form, Φ, on V, the *Clifford group of* Φ is the group

$$\Gamma(\Phi) = \{x \in \mathrm{Cl}(\Phi)^* \mid \alpha(x) v x^{-1} \in V \quad \text{for all } v \in V\}.$$

The map $N \colon \mathrm{Cl}(Q) \to \mathrm{Cl}(Q)$ given by

$$N(x) = x\overline{x}$$

is called the *norm* of $\mathrm{Cl}(\Phi)$.

We see that the group $\Gamma(\Phi)$ acts on V *via*

$$x \cdot v = \alpha(x) v x^{-1},$$

where $x \in \Gamma(\Phi)$ and $v \in V$. Actually, it is not entirely obvious why the action $\Gamma(\Phi) \times V \longrightarrow V$ is a linear action, and for that matter, why $\Gamma(\Phi)$ is a group.

This is because V is finite-dimensional and α is an automorphism. As a consequence, for any $x \in \Gamma(\Phi)$, the map ρ_x from V to V defined by

$$v \mapsto \alpha(x) v x^{-1}$$

is linear and injective, and thus bijective, since V has finite dimension. It follows that $x^{-1} \in \Gamma(\Phi)$ (the reader should fill in the details).

We also define the group $\Gamma^+(\Phi)$, called the *special Clifford group*, by

$$\Gamma^+(\Phi) = \Gamma(\Phi) \cap \mathrm{Cl}^0(\Phi).$$

Observe that $N(v) = -\Phi(v) \cdot 1$ for all $v \in V$. Also, if $(e_1, \ldots, e_n)$ is a basis of V, we leave it as an exercise to check that

$$N(e_{i_1} e_{i_2} \cdots e_{i_k}) = (-1)^k \Phi(e_{i_1}) \Phi(e_{i_2}) \cdots \Phi(e_{i_k}) \cdot 1.$$

Remark: The map $\rho \colon \Gamma(\Phi) \to \mathbf{GL}(V)$ given by $x \mapsto \rho_x$ is called the *twisted adjoint representation*. It was introduced by Atiyah, Bott and Shapiro [11]. It has the advantage of not introducing a spurious negative sign, i.e., when $v \in V$ and $\Phi(v) \neq 0$, the map ρ_v is the reflection s_v about the hyperplane orthogonal to v (see Proposition 21.6). Furthermore, when Φ is nondegenerate, the kernel $\mathrm{Ker}\,(\rho)$ of the representation ρ is given by $\mathrm{Ker}\,(\rho) = \mathbb{R}^* \cdot 1$, where $\mathbb{R}^* = \mathbb{R} - \{0\}$. The earlier *adjoint representation* (used by Chevalley [35] and others) is given by

$$v \mapsto x v x^{-1}.$$

Unfortunately, in this case, ρ_x represents $-s_v$, where s_v is the reflection about the hyperplane orthogonal to v. Furthermore, the kernel of the representation ρ is generally bigger than $\mathbb{R}^* \cdot 1$. This is the reason why the twisted adjoint representation is preferred (and must be used for a proper treatment of the **Pin** group).

Proposition 21.5. *The maps α and t induce an automorphism and an anti-automorphism of the Clifford group, $\Gamma(\Phi)$.*

Proof. It is not very instructive, see Bröcker and tom Dieck [25], Chapter 1, Section 6, page 58. □

The following proposition shows why Clifford groups generalize the quaternions.

Proposition 21.6. *Let V be a finite dimensional vector space and Φ a quadratic form on V. For every element x, of the Clifford group $\Gamma(\Phi)$, if $x \in V$ and $\Phi(x) \neq 0$, then the map $\rho_x \colon V \to V$ given by*

$$v \mapsto \alpha(x) v x^{-1} \quad \textit{for all } v \in V$$

is the reflection about the hyperplane H orthogonal to the vector x.

Proof. Recall that the reflection s about the hyperplane H orthogonal to the vector x is given by

$$s(u) = u - 2\,\frac{\varphi(u,x)}{\Phi(x)} \cdot x.$$

However, we have

$$x^2 = \Phi(x) \cdot 1 \quad \text{and} \quad ux + xu = 2\varphi(u,x) \cdot 1.$$

Thus, we have

$$\begin{aligned} s(u) &= u - 2\,\frac{\varphi(u,x)}{\Phi(x)} \cdot x \\ &= u - 2\varphi(u,x) \cdot \left(\frac{1}{\Phi(x)} \cdot x\right) \\ &= u - 2\varphi(u,x) \cdot x^{-1} \\ &= u - 2\varphi(u,x) \cdot (1x^{-1}) \\ &= u - (2\varphi(u,x) \cdot 1)x^{-1} \\ &= u - (ux + xu)x^{-1} \\ &= -xux^{-1} \\ &= \alpha(x) u x^{-1}, \end{aligned}$$

since $\alpha(x) = -x$, for $x \in V$. □

In general, we have a map

$$\rho \colon \Gamma(\Phi) \to \mathbf{GL}(V)$$

defined by

$$\rho(x)(v) = \alpha(x) v x^{-1},$$

for all $x \in \Gamma(\Phi)$ and all $v \in V$. We would like to show that ρ is a surjective homomorphism from $\Gamma(\Phi)$ onto $\mathbf{O}(\varphi)$ and a surjective homomorphism from $\Gamma^+(\Phi)$ onto $\mathbf{SO}(\varphi)$. For this, we will need to assume that φ is nondegenerate, which means that for every $v \in V$, if $\varphi(v,w) = 0$ for all $w \in V$, then $v = 0$. For simplicity of exposition, we first assume that Φ is the quadratic form on $\mathbb{R}^n$ defined by

$$\Phi(x_1, \ldots, x_n) = -(x_1^2 + \cdots + x_n^2).$$

Let Cl_n denote the Clifford algebra $\mathrm{Cl}(\Phi)$ and Γ_n denote the Clifford group $\Gamma(\Phi)$. The following lemma plays a crucial role:

Lemma 21.7. *The kernel of the map* $\rho\colon \Gamma_n \to \mathbf{GL}(n)$ *is* $\mathbb{R}^* \cdot 1$*, the multiplicative group of nonzero scalar multiples of* $1 \in \mathrm{Cl}_n$*.*

Proof. If $\rho(x) = \mathrm{id}$, then

$$\alpha(x)v = vx \quad \text{for all } v \in \mathbb{R}^n. \tag{1}$$

Since $\mathrm{Cl}_n = \mathrm{Cl}_n^0 \oplus \mathrm{Cl}_n^1$, we can write $x = x^0 + x^1$, with $x^i \in \mathrm{Cl}_n^i$ for $i = 1, 2$. Then, equation (1) becomes

$$x^0 v = vx^0 \quad \text{and} \quad -x^1 v = vx^1 \quad \text{for all } v \in \mathbb{R}^n. \tag{2}$$

Using Theorem 21.4, we can express x^0 as a linear combination of monomials in the canonical basis $(e_1, \ldots, e_n)$, so that

$$x^0 = a^0 + e_1 b^1, \quad \text{with } a^0 \in \mathrm{Cl}_n^0,\ b^1 \in \mathrm{Cl}_n^1,$$

where neither a^0 nor b^1 contains a summand with a factor e_1. Applying the first relation in (2) to $v = e_1$, we get

$$e_1 a^0 + e_1^2 b^1 = a^0 e_1 + e_1 b^1 e_1. \tag{3}$$

Now, the basis $(e_1, \ldots, e_n)$ is orthogonal w.r.t. Φ, which implies that

$$e_j e_k = -e_k e_j \quad \text{for all } j \neq k.$$

Since each monomial in a^0 is of even degre and contains no factor e_1, we get

$$a^0 e_1 = e_1 a^0.$$

Similarly, since b^1 is of odd degree and contains no factor e_1, we get

$$e_1 b^1 e_1 = -e_1^2 b^1.$$

But then, from (3), we get

$$e_1 a^0 + e_1^2 b^1 = a^0 e_1 + e_1 b^1 e_1 = e_1 a^0 - e_1^2 b^1,$$

and so, $e_1^2 b^1 = 0$. However, $e_1^2 = -1$, and so, $b_1 = 0$. Therefore, x_0 contains no monomial with a factor e_1. We can apply the same argument to the other basis elements $e_2, \ldots, e_n$, and thus, we just proved that $x^0 \in \mathbb{R} \cdot 1$.

A similar argument applying to the second equation in (2), with $x^1 = a^1 + e_1 b^0$ and $v = e_1$ shows that $b^0 = 0$. We also conclude that $x^1 \in \mathbb{R} \cdot 1$. However, $\mathbb{R} \cdot 1 \subseteq \mathrm{Cl}_n^0$, and so, $x^1 = 0$. Finally, $x = x^0 \in (\mathbb{R} \cdot 1) \cap \Gamma_n = \mathbb{R}^* \cdot 1$. □

Remark: If Φ is any nondegenerate quadratic form, we know (for instance, see Artin [7], Chapter 7, or Gallier [60], Chapter 6, Problem 6.14) that there is an orthogonal basis $(e_1, \ldots, e_n)$ with respect to φ (i.e. $\varphi(e_j, e_k) = 0$ for all $j \neq k$). Thus, the commutation relations

$$\begin{aligned} e_j^2 &= \Phi(e_j) \cdot 1, \quad \text{with } \Phi(e_j) \neq 0, \quad 1 \leq j \leq n, \quad \text{and} \\ e_j e_k &= -e_k e_j, \quad 1 \leq j, k \leq n,\ j \neq k \end{aligned}$$

hold, and since the proof only rests on these facts, Lemma 21.7 holds for any nondegenerate quadratic form.

However, Lemma 21.7 may fail for degenerate quadratic forms. For example, if $\Phi \equiv 0$, then $\mathrm{Cl}(V, 0) = \bigwedge^{\bullet} V$. Consider the element $x = 1 + e_1 e_2$. Clearly, $x^{-1} = 1 - e_1 e_2$. But now, for any $v \in V$, we have

$$\alpha(1 + e_1 e_2) v (1 + e_1 e_2)^{-1} = (1 + e_1 e_2) v (1 - e_1 e_2) = v.$$

Yet, $1 + e_1 e_2$ is not a scalar multiple of 1.

The following proposition shows that the notion of norm is well-behaved.

Proposition 21.8. *If $x \in \Gamma_n$, then $N(x) \in \mathbb{R}^* \cdot 1$.*

Proof. The trick is to show that $N(x)$ is in the kernel of ρ. To say that $x \in \Gamma_n$ means that

$$\alpha(x) v x^{-1} \in \mathbb{R}^n \quad \text{for all } v \in \mathbb{R}^n.$$

Applying t, we get

$$t(x)^{-1} v t(\alpha(x)) = \alpha(x) v x^{-1},$$

since t is the identity on $\mathbb{R}^n$. Thus, we have

$$v = t(x)\alpha(x) v (t(\alpha(x))x)^{-1} = \alpha(\overline{x}x) v (\overline{x}x)^{-1},$$

so $\overline{x}x \in \mathrm{Ker}\,(\rho)$. By Proposition 21.5, we have $\overline{x} \in \Gamma_n$, and so, $x\overline{x} = \overline{\overline{x}}\,\overline{x} \in \mathrm{Ker}\,(\rho)$. □

Remark: Again, the proof also holds for the Clifford group $\Gamma(\Phi)$ associated with any nondegenerate quadratic form Φ. When $\Phi(v) = -\|v\|^2$, where $\|v\|$ is the standard Euclidean norm of v, we have $N(v) = \|v\|^2 \cdot 1$ for all $v \in V$. However, for other quadratic forms, it is possible that $N(x) = \lambda \cdot 1$ where $\lambda < 0$, and this is a difficulty that needs to be overcome.

Proposition 21.9. *The restriction of the norm, N, to Γ_n is a homomorphism, $N \colon \Gamma_n \to \mathbb{R}^* \cdot 1$, and $N(\alpha(x)) = N(x)$ for all $x \in \Gamma_n$.*

Proof. We have

$$N(xy) = xy\overline{y}\,\overline{x} = xN(y)\overline{x} = x\overline{x}N(y) = N(x)N(y),$$

where the third equality holds because $N(x) \in \mathbb{R}^* \cdot 1$. We also have

$$N(\alpha(x)) = \alpha(x)\alpha(\overline{x}) = \alpha(x\overline{x}) = \alpha(N(x)) = N(x).$$ □

Remark: The proof also holds for the Clifford group $\Gamma(\Phi)$ associated with any nondegenerate quadratic form Φ.

Proposition 21.10. *We have $\mathbb{R}^n - \{0\} \subseteq \Gamma_n$ and $\rho(\Gamma_n) \subseteq \mathbf{O}(n)$.*

Proof. Let $x \in \Gamma_n$ and $v \in \mathbb{R}^n$, with $v \neq 0$. We have

$$N(\rho(x)(v)) = N(\alpha(x)vx^{-1}) = N(\alpha(x))N(v)N(x^{-1}) = N(x)N(v)N(x)^{-1} = N(v),$$

since $N\colon \Gamma_n \to \mathbb{R}^* \cdot 1$. However, for $v \in \mathbb{R}^n$, we know that

$$N(v) = -\Phi(v) \cdot 1.$$

Thus, $\rho(x)$ is norm-preserving, and so, $\rho(x) \in \mathbf{O}(n)$. □

Remark: The proof that $\rho(\Gamma(\Phi)) \subseteq \mathbf{O}(\Phi)$ also holds for the Clifford group $\Gamma(\Phi)$ associated with any nondegenerate quadratic form Φ. The first statement needs to be replaced by the fact that every non-isotropic vector in $\mathbb{R}^n$ (a vector is non-isotropic if $\Phi(x) \neq 0$) belongs to $\Gamma(\Phi)$. Indeed, $x^2 = \Phi(x) \cdot 1$, which implies that x is invertible.

We are finally ready for the introduction of the groups $\mathbf{Pin}(n)$ and $\mathbf{Spin}(n)$.

21.4 The Groups Pin(n) and Spin(n)

Definition 21.5. We define the *pinor group,* $\mathbf{Pin}(n)$, as the kernel $\mathrm{Ker}\,(N)$ of the homomorphism $N\colon \Gamma_n \to \mathbb{R}^* \cdot 1$, and the *spinor group,* $\mathbf{Spin}(n)$, as $\mathbf{Pin}(n) \cap \Gamma_n^+$.

Observe that if $N(x) = 1$, then x is invertible and $x^{-1} = \overline{x}$, since $x\overline{x} = N(x) = 1$. Thus, we can write

$$\mathbf{Pin}(n) = \{x \in \mathrm{Cl}_n \mid xvx^{-1} \in \mathbb{R}^n \quad \text{for all } v \in \mathbb{R}^n, \quad N(x) = 1\},$$

and

$$\mathbf{Spin}(n) = \{x \in \mathrm{Cl}_n^0 \mid xvx^{-1} \in \mathbb{R}^n \quad \text{for all } v \in \mathbb{R}^n, \quad N(x) = 1\}.$$

Remark: According to Atiyah, Bott and Shapiro, the use of the name $\mathbf{Pin}(k)$ is a joke due to Jean-Pierre Serre (Atiyah, Bott and Shapiro [11], page 1).

Theorem 21.11. *The restriction of ρ to the pinor group, $\mathbf{Pin}(n)$, is a surjective homomorphism, $\rho\colon \mathbf{Pin}(n) \to \mathbf{O}(n)$, whose kernel is $\{-1, 1\}$, and the restriction of ρ to the spinor group, $\mathbf{Spin}(n)$, is a surjective homomorphism, $\rho\colon \mathbf{Spin}(n) \to \mathbf{SO}(n)$, whose kernel is $\{-1, 1\}$.*

Proof. By Proposition 21.10, we have a map $\rho\colon \mathbf{Pin}(n) \to \mathbf{O}(n)$. The reader can easily check that ρ is a homomorphism. By the Cartan-Dieudonné theorem (see Bourbaki [20], or Gallier [60], Chapter 7, Theorem 7.2.1), every isometry $f \in \mathbf{SO}(n)$ is the composition $f = s_1 \circ \cdots \circ s_k$ of hyperplane reflections s_j. If we assume that s_j is a reflection about the

hyperplane H_j orthogonal to the nonzero vector w_j, by Proposition 21.6, $\rho(w_j) = s_j$. Since $N(w_j) = \|w_j\|^2 \cdot 1$, we can replace w_j by $w_j/\|w_j\|$, so that $N(w_1 \cdots w_k) = 1$, and then

$$f = \rho(w_1 \cdots w_k),$$

and ρ is surjective. Note that

$$\mathrm{Ker}\,(\rho \mid \mathbf{Pin}(n)) = \mathrm{Ker}\,(\rho) \cap \ker(N) = \{t \in \mathbb{R}^* \cdot 1 \mid N(t) = 1\} = \{-1, 1\}.$$

As to $\mathbf{Spin}(n)$, we just need to show that the restriction of ρ to $\mathbf{Spin}(n)$ maps Γ_n into $\mathbf{SO}(n)$. If this was not the case, there would be some improper isometry $f \in \mathbf{O}(n)$ so that $\rho(x) = f$, where $x \in \Gamma_n \cap \mathrm{Cl}_n^0$. However, we can express f as the composition of an odd number of reflections, say

$$f = \rho(w_1 \cdots w_{2k+1}).$$

Since

$$\rho(w_1 \cdots w_{2k+1}) = \rho(x),$$

we have $x^{-1} w_1 \cdots w_{2k+1} \in \mathrm{Ker}\,(\rho)$. By Lemma 21.7, we must have

$$x^{-1} w_1 \cdots w_{2k+1} = \lambda \cdot 1$$

for some $\lambda \in \mathbb{R}^*$, and thus,

$$w_1 \cdots w_{2k+1} = \lambda \cdot x,$$

where x has even degree and $w_1 \cdots w_{2k+1}$ has odd degree, which is impossible. □

Let us denote the set of elements $v \in \mathbb{R}^n$ with $N(v) = 1$ (with norm 1) by S^{n-1}. We have the following corollary of Theorem 21.11:

Corollary 21.12. *The group* $\mathbf{Pin}(n)$ *is generated by* S^{n-1} *and every element of* $\mathbf{Spin}(n)$ *can be written as the product of an even number of elements of* S^{n-1}.

Example 21.2. The reader should verify that

$$\mathbf{Pin}(1) \approx \mathbb{Z}/4\mathbb{Z}, \quad \mathbf{Spin}(1) = \{-1, 1\} \approx \mathbb{Z}/2\mathbb{Z},$$

and also that

$$\mathbf{Pin}(2) \approx \{ae_1 + be_2 \mid a^2 + b^2 = 1\} \cup \{c1 + de_1e_2 \mid c^2 + d^2 = 1\}, \quad \mathbf{Spin}(2) = \mathbf{U}(1).$$

We may also write $\mathbf{Pin}(2) = \mathbf{U}(1) + \mathbf{U}(1)$, where $\mathbf{U}(1)$ is the group of complex numbers of modulus 1 (the unit circle in $\mathbb{R}^2$). It can also be shown that $\mathbf{Spin}(3) \approx \mathbf{SU}(2)$ and $\mathbf{Spin}(4) \approx \mathbf{SU}(2) \times \mathbf{SU}(2)$. The group $\mathbf{Spin}(5)$ is isomorphic to the symplectic group $\mathbf{Sp}(2)$, and $\mathbf{Spin}(6)$ is isomorphic to $\mathbf{SU}(4)$ (see Curtis [39] or Porteous [125]).

Let us take a closer look at **Spin**(2). The Clifford algebra Cl_2 is generated by the four elements

$$1,\ e_1,\ e_2,\ , e_1e_2,$$

and they satisfy the relations

$$e_1^2 = -1, \quad e_2^2 = -1, \quad e_1e_2 = -e_2e_1.$$

The group **Spin**(2) consists of all products

$$\prod_{i=1}^{2k}(a_ie_1 + b_ie_2)$$

consisting of an even number of factors and such that $a_i^2 + b_i^2 = 1$. In view of the above relations, every such element can be written as

$$x = a1 + be_1e_2,$$

where x satisfies the conditions that $xvx^{-1} \in \mathbb{R}^2$ for all $v \in \mathbb{R}^2$, and $N(x) = 1$. Since

$$\overline{X} = a1 - be_1e_2,$$

we get

$$N(x) = a^2 + b^2,$$

and the condition $N(x) = 1$ is simply $a^2 + b^2 = 1$. We claim that $xvx^{-1} \in \mathbb{R}^2$ if $x \in \mathrm{Cl}_2^0$. Indeed, since $x \in \mathrm{Cl}_2^0$ and $v \in \mathrm{Cl}_2^1$, we have $xvx^{-1} \in \mathrm{Cl}_2^1$, which implies that $xvx^{-1} \in \mathbb{R}^2$, since the only elements of Cl_2^1 are those in $\mathbb{R}^2$. Then, **Spin**(2) consists of those elements $x = a1 + be_1e_2$ so that $a^2 + b^2 = 1$. If we let $\mathbf{i} = e_1e_2$, we observe that

$$\begin{aligned}
\mathbf{i}^2 &= -1,\\
e_1\mathbf{i} &= -\mathbf{i}e_1 = -e_2,\\
e_2\mathbf{i} &= -\mathbf{i}e_2 = e_1.
\end{aligned}$$

Thus, **Spin**(2) is isomorphic to $\mathbf{U}(1)$. Also note that

$$e_1(a1 + b\mathbf{i}) = (a1 - b\mathbf{i})e_1.$$

Let us find out explicitly what is the action of **Spin**(2) on $\mathbb{R}^2$. Given $X = a1 + b\mathbf{i}$, with $a^2 + b^2 = 1$, for any $v = v_1e_1 + v_2e_2$, we have

$$\begin{aligned}
\alpha(X)vX^{-1} &= X(v_1e_1 + v_2e_2)X^{-1}\\
&= X(v_1e_1 + v_2e_2)(-e_1e_1)\overline{X}\\
&= X(v_1e_1 + v_2e_2)(-e_1)(e_1\overline{X})\\
&= X(v_11 + v_2\mathbf{i})Xe_1\\
&= X^2(v_11 + v_2\mathbf{i})e_1\\
&= (((a^2 - b^2)v_1 - 2abv_2)1 + (a^2 - b^2)v_2 + 2abv_1)\mathbf{i})e_1\\
&= ((a^2 - b^2)v_1 - 2abv_2)e_1 + (a^2 - b^2)v_2 + 2abv_1)e_2.
\end{aligned}$$

Since $a^2+b^2=1$, we can write $X = a1+b\mathbf{i} = (\cos\theta)1+(\sin\theta)\mathbf{i}$, and the above derivation shows that

$$\alpha(X)vX^{-1} = (\cos 2\theta v_1 - \sin 2\theta v_2)e_1 + (\cos 2\theta v_2 + \sin 2\theta v_1)e_2.$$

This means that the rotation ρ_X induced by $X \in \mathbf{Spin}(2)$ is the rotation of angle 2θ around the origin. Observe that the maps

$$v \mapsto v(-e_1), \quad X \mapsto Xe_1$$

establish bijections between $\mathbb{R}^2$ and $\mathbf{Spin}(2) \simeq \mathbf{U}(1)$. Also, note that the action of $X = \cos\theta + i\sin\theta$ viewed as a complex number yields the rotation of angle θ, whereas the action of $X = (\cos\theta)1 + (\sin\theta)\mathbf{i}$ viewed as a member of $\mathbf{Spin}(2)$ yields the rotation of angle 2θ. There is nothing wrong. In general, $\mathbf{Spin}(n)$ is a two–to–one cover of $\mathbf{SO}(n)$.

Next, let us take a closer look at $\mathbf{Spin}(3)$. The Clifford algebra Cl_3 is generated by the eight elements

$$1,\ e_1,\ e_2,\ ,e_3,\ ,e_1e_2,\ e_2e_3,\ e_3e_1,\ e_1e_2e_3,$$

and they satisfy the relations

$$e_i^2 = -1, \quad e_je_j = -e_je_i, \quad 1 \le i,j \le 3,\ i \ne j.$$

The group $\mathbf{Spin}(3)$ consists of all products

$$\prod_{i=1}^{2k}(a_ie_1 + b_ie_2 + c_ie_3)$$

consisting of an even number of factors and such that $a_i^2+b_i^2+c_i^2=1$. In view of the above relations, every such element can be written as

$$x = a1 + be_2e_3 + ce_3e_1 + de_1e_2,$$

where x satisfies the conditions that $xvx^{-1} \in \mathbb{R}^3$ for all $v \in \mathbb{R}^3$, and $N(x)=1$. Since

$$\overline{X} = a1 - be_2e_3 - ce_3e_1 - de_1e_2,$$

we get

$$N(x) = a^2+b^2+c^2+d^2,$$

and the condition $N(x)=1$ is simply $a^2+b^2+c^2+d^2=1$.

It turns out that the conditions $x \in \mathrm{Cl}_3^0$ and $N(x)=1$ imply that $xvx^{-1} \in \mathbb{R}^3$ for all $v \in \mathbb{R}^3$. To prove this, first observe that $N(x)=1$ implies that $x^{-1} = \pm\overline{x}$, and that $\overline{v} = -v$ for any $v \in \mathbb{R}^3$, and so,

$$\overline{xvx^{-1}} = -xvx^{-1}.$$

Also, since $x \in \mathrm{Cl}_3^0$ and $v \in \mathrm{Cl}_3^1$, we have $xvx^{-1} \in \mathrm{Cl}_3^1$. Thus, we can write

$$xvx^{-1} = u + \lambda e_1e_2e_3, \quad \text{for some } u \in \mathbb{R}^3 \text{ and some } \lambda \in \mathbb{R}.$$

But

$$\overline{e_1e_2e_3} = -e_3e_2e_1 = e_1e_2e_3,$$

and so,

$$\overline{xvx^{-1}} = -u + \lambda e_1e_2e_3 = -xvx^{-1} = -u - \lambda e_1e_2e_3,$$

which implies that $\lambda = 0$. Thus, $xvx^{-1} \in \mathbb{R}^3$, as claimed. Then, $\mathbf{Spin}(3)$ consists of those elements $x = a1 + be_2e_3 + ce_3e_1 + de_1e_2$ so that $a^2 + b^2 + c^2 + d^2 = 1$. Under the bijection

$$\mathbf{i} \mapsto e_2e_3,\ \mathbf{j} \mapsto e_3e_1,\ \mathbf{k} \mapsto e_1e_2,$$

we can check that we have an isomorphism between the group $\mathbf{SU}(2)$ of unit quaternions and $\mathbf{Spin}(3)$. If $X = a1 + be_2e_3 + ce_3e_1 + de_1e_2 \in \mathbf{Spin}(3)$, observe that

$$X^{-1} = \overline{X} = a1 - be_2e_3 - ce_3e_1 - de_1e_2.$$

Now, using the identification

$$\mathbf{i} \mapsto e_2e_3,\ \mathbf{j} \mapsto e_3e_1,\ \mathbf{k} \mapsto e_1e_2,$$

we can easily check that

$$\begin{aligned}
(e_1e_2e_3)^2 &= 1,\\
(e_1e_2e_3)\mathbf{i} &= \mathbf{i}(e_1e_2e_3) = -e_1,\\
(e_1e_2e_3)\mathbf{j} &= \mathbf{j}(e_1e_2e_3) = -e_2,\\
(e_1e_2e_3)\mathbf{k} &= \mathbf{k}(e_1e_2e_3) = -e_3,\\
(e_1e_2e_3)e_1 &= -\mathbf{i},\\
(e_1e_2e_3)e_2 &= -\mathbf{j},\\
(e_1e_2e_3)e_3 &= -\mathbf{k}.
\end{aligned}$$

Then, if $X = a1 + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} \in \mathbf{Spin}(3)$, for every $v = v_1e_1 + v_2e_2 + v_3e_3$, we have

$$\begin{aligned}
\alpha(X)vX^{-1} &= X(v_1e_1 + v_2e_2 + v_3e_3)X^{-1}\\
&= X(e_1e_2e_3)^2(v_1e_1 + v_2e_2 + v_3e_3)X^{-1}\\
&= (e_1e_2e_3)X(e_1e_2e_3)(v_1e_1 + v_2e_2 + v_3e_3)X^{-1}\\
&= -(e_1e_2e_3)X(v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k})X^{-1}.
\end{aligned}$$

This shows that the rotation $\rho_X \in \mathbf{SO}(3)$ induced by $X \in \mathbf{Spin}(3)$ can be viewed as the rotation induced by the quaternion $a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ on the pure quaternions, using the maps

$$v \mapsto -(e_1e_2e_3)v, \quad X \mapsto -(e_1e_2e_3)X$$

to go from a vector $v = v_1e_1 + v_2e_2 + v_3e_3$ to the pure quaternion $v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$, and back.

We close this section by taking a closer look at $\mathbf{Spin}(4)$. The group $\mathbf{Spin}(4)$ consists of all products

$$\prod_{i=1}^{2k}(a_ie_1 + b_ie_2 + c_ie_3 + d_ie_4)$$

consisting of an even number of factors and such that $a_i^2 + b_i^2 + c_i^2 + d_i^2 = 1$. Using the relations

$$e_i^2 = -1, \quad e_je_j = -e_je_i, \quad 1 \leq i, j \leq 4,\ i \neq j,$$

every element of $\mathbf{Spin}(4)$ can be written as

$$x = a_1 1 + a_2e_1e_2 + a_3e_2e_3 + a_4e_3e_1 + a_5e_4e_3 + a_6e_4e_1 + a_7e_4e_2 + a_8e_1e_2e_3e_4,$$

where x satisfies the conditions that $xvx^{-1} \in \mathbb{R}^4$ for all $v \in \mathbb{R}^4$, and $N(x) = 1$. Let

$$\mathbf{i} = e_1e_2,\ \mathbf{j} = e_2e_3,\ \mathbf{k} = 3_3e_1,\ \mathbf{i}' = e_4e_3,\ \mathbf{j}' = e_4e_1,\ \mathbf{k}' = e_4e_2,$$

and $\mathbb{I} = e_1e_2e_3e_4$. The reader will easily verify that

$$\begin{aligned}
\mathbf{ij} &= \mathbf{k}\\
\mathbf{jk} &= \mathbf{i}\\
\mathbf{ki} &= \mathbf{j}\\
\mathbf{i}^2 &= -1, \quad \mathbf{j}^2 = -1, \quad \mathbf{k}^2 = -1\\
\mathbf{i}\mathbb{I} &= \mathbb{I}\mathbf{i} = \mathbf{i}'\\
\mathbf{j}\mathbb{I} &= \mathbb{I}\mathbf{j} = \mathbf{j}'\\
\mathbf{k}\mathbb{I} &= \mathbb{I}\mathbf{k} = \mathbf{k}'\\
\mathbb{I}^2 &= 1, \quad \overline{\mathbb{I}} = \mathbb{I}.
\end{aligned}$$

Then, every $x \in \mathbf{Spin}(4)$ can be written as

$$x = u + \mathbb{I}v, \quad \text{with} \quad u = a1 + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} \quad \text{and} \quad v = a'1 + b'\mathbf{i} + c'\mathbf{j} + d'\mathbf{k},$$

with the extra conditions stated above. Using the above identities, we have

$$(u + \mathbb{I}v)(u' + \mathbb{I}v') = uu' + vv' + \mathbb{I}(uv' + vu').$$

As a consequence,

$$N(u + \mathbb{I}v) = (u + \mathbb{I}v)(\overline{u} + \mathbb{I}\overline{v}) = u\overline{u} + v\overline{v} + \mathbb{I}(u\overline{v} + v\overline{u}),$$

and thus, $N(u + \mathbb{I}v) = 1$ is equivalent to

$$u\overline{u} + v\overline{v} = 1 \quad \text{and} \quad u\overline{v} + v\overline{u} = 0.$$

As in the case $n = 3$, it turns out that the conditions $x \in \mathrm{Cl}_4^0$ and $N(x) = 1$ imply that $xvx^{-1} \in \mathbb{R}^4$ for all $v \in \mathbb{R}^4$. The only change to the proof is that $xvx^{-1} \in \mathrm{Cl}_4^1$ can be written as

$$xvx^{-1} = u + \sum_{i,j,k} \lambda_{i,j,k} e_i e_j e_k, \quad \text{for some } u \in \mathbb{R}^4, \quad \text{with } \{i,j,k\} \subseteq \{1,2,3,4\}.$$

As in the previous proof, we get $\lambda_{i,j,k} = 0$. Then, $\mathbf{Spin}(4)$ consists of those elements $u + \mathbb{I}v$ so that

$$u\overline{u} + v\overline{v} = 1 \quad \text{and} \quad u\overline{v} + v\overline{u} = 0,$$

with u and v of the form $a1 + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$. Finally, we see that $\mathbf{Spin}(4)$ is isomorphic to $\mathbf{Spin}(2) \times \mathbf{Spin}(2)$ under the isomorphism

$$u + v\mathbb{I} \mapsto (u + v, u - v).$$

Indeed, we have

$$N(u+v) = (u+v)(\overline{u} + \overline{v}) = 1,$$

and

$$N(u-v) = (u-v)(\overline{u} - \overline{v}) = 1,$$

since

$$u\overline{u} + v\overline{v} = 1 \quad \text{and} \quad u\overline{v} + v\overline{u} = 0,$$

and

$$(u+v, u-v)(u'+v', u'-v') = (uu' + vv' + uv' + vu', uu' + vv' - (uv' + vu')).$$

Remark: It can be shown that the assertion if $x \in \mathrm{Cl}_n^0$ and $N(x) = 1$, then $xvx^{-1} \in \mathbb{R}^n$ for all $v \in \mathbb{R}^n$, is true up to $n = 5$ (see Porteous [125], Chapter 13, Proposition 13.58). However, this is already false for $n = 6$. For example, if $X = 1/\sqrt{2}(1 + e_1e_2e_3e_4e_5e_6)$, it is easy to see that $N(X) = 1$, and yet, $Xe_1X^{-1} \notin \mathbb{R}^6$.

21.5 The Groups $\mathbf{Pin}(p,q)$ and $\mathbf{Spin}(p,q)$

For every nondegenerate quadratic form Φ over $\mathbb{R}$, there is an orthogonal basis with respect to which Φ is given by

$$\Phi(x_1, \ldots, x_{p+q}) = x_1^2 + \cdots + x_p^2 - (x_{p+1}^2 + \cdots + x_{p+q}^2),$$

where p and q only depend on Φ. The quadratic form corresponding to (p,q) is denoted $\Phi_{p,q}$ and we call (p,q) the *signature of* $\Phi_{p,q}$. Let $n = p + q$. We define the group $\mathbf{O}(p,q)$ as the group of isometries w.r.t. $\Phi_{p,q}$, i.e., the group of linear maps f so that

$$\Phi_{p,q}(f(v)) = \Phi_{p,q}(v) \quad \text{for all } v \in \mathbb{R}^n$$

and the group $\mathbf{SO}(p,q)$ as the subgroup of $\mathbf{O}(p,q)$ consisting of the isometries, $f \in \mathbf{O}(p,q)$, with $\det(f) = 1$. We denote the Clifford algebra $\mathrm{Cl}(\Phi_{p,q})$ where $\Phi_{p,q}$ has signature (p,q) by $\mathrm{Cl}_{p,q}$, the corresponding Clifford group by $\Gamma_{p,q}$, and the special Clifford group $\Gamma_{p,q} \cap \mathrm{Cl}_{p,q}^0$ by $\Gamma_{p,q}^+$. Note that with this new notation, $\mathrm{Cl}_n = \mathrm{Cl}_{0,n}$.

As we mentioned earlier, since Lawson and Michelsohn [97] adopt the opposite of our sign convention in defining Clifford algebras, their $\mathrm{Cl}(p,q)$ is our $\mathrm{Cl}(q,p)$.

As we mentioned in Section 21.3, we have the problem that $N(v) = -\Phi(v)\cdot 1$ but $-\Phi(v)$ is not necessarily positive (where $v \in \mathbb{R}^n$). The fix is simple: Allow elements $x \in \Gamma_{p,q}$ with $N(x) = \pm 1$.

Definition 21.6. We define the *pinor group,* $\mathbf{Pin}(p,q)$, as the group

$$\mathbf{Pin}(p,q) = \{x \in \Gamma_{p,q} \mid N(x) = \pm 1\},$$

and the *spinor group,* $\mathbf{Spin}(p,q)$, as $\mathbf{Pin}(p,q) \cap \Gamma^+_{p,q}$.

Remarks:

(1) It is easily checked that the group $\mathbf{Spin}(p,q)$ is also given by

$$\mathbf{Spin}(p,q) = \{x \in \mathrm{Cl}^0_{p,q} \mid xv\overline{x} \in \mathbb{R}^n \quad \text{for all } v \in \mathbb{R}^n, \quad N(x) = 1\}.$$

This is because $\mathbf{Spin}(p,q)$ consists of elements of even degree.

(2) One can check that if $N(x) \neq 0$, then

$$\alpha(x)vx^{-1} = xvt(x)/N(x).$$

Thus, we have

$$\mathbf{Pin}(p,q) = \{x \in \mathrm{Cl}_{p,q} \mid xvt(x)N(x) \in \mathbb{R}^n \quad \text{for all } v \in \mathbb{R}^n, \quad N(x) = \pm 1\}.$$

When $\Phi(x) = -\|x\|^2$, we have $N(x) = \|x\|^2$, and

$$\mathbf{Pin}(n) = \{x \in \mathrm{Cl}_n \mid xvt(x) \in \mathbb{R}^n \quad \text{for all } v \in \mathbb{R}^n, \quad N(x) = 1\}.$$

Theorem 21.11 generalizes as follows:

Theorem 21.13. *The restriction of ρ to the pinor group, $\mathbf{Pin}(p,q)$, is a surjective homomorphism, $\rho\colon \mathbf{Pin}(p,q) \to \mathbf{O}(p,q)$, whose kernel is $\{-1,1\}$, and the restriction of ρ to the spinor group, $\mathbf{Spin}(p,q)$, is a surjective homomorphism, $\rho\colon \mathbf{Spin}(p,q) \to \mathbf{SO}(p,q)$, whose kernel is $\{-1,1\}$.*

Proof. The Cartan-Dieudonné also holds for any nondegenerate quadratic form Φ, in the sense that every isometry in $\mathbf{O}(\Phi)$ is the composition of reflections defined by hyperplanes orthogonal to non-isotropic vectors (see Dieudonné [43], Chevalley [35], Bourbaki [20], or Gallier [60], Chapter 7, Problem 7.14). Thus, Theorem 21.11 also holds for any nondegenerate quadratic form Φ. The only change to the proof is the following: Since $N(w_j) = -\Phi(w_j)\cdot 1$, we can replace w_j by $w_j/\sqrt{|\Phi(w_j)|}$, so that $N(w_1 \cdots w_k) = \pm 1$, and then

$$f = \rho(w_1 \cdots w_k),$$

and ρ is surjective. □

If we consider $\mathbb{R}^n$ equipped with the quadratic form $\Phi_{p,q}$ (with $n = p+q$), we denote the set of elements $v \in \mathbb{R}^n$ with $N(v) = 1$ by $S^{n-1}_{p,q}$. We have the following corollary of Theorem 21.13 (generalizing Corollary 21.14):

Corollary 21.14. *The group* $\mathbf{Pin}(p,q)$ *is generated by* $S^{n-1}_{p,q}$ *and every element of* $\mathbf{Spin}(p,q)$ *can be written as the product of an even number of elements of* $S^{n-1}_{p,q}$.

Example 21.3. The reader should check that

$$\mathrm{Cl}_{0,1} \approx \mathbb{C}, \quad \mathrm{Cl}_{1,0} \approx \mathbb{R} \oplus \mathbb{R}.$$

We also have

$$\mathbf{Pin}(0,1) \approx \mathbb{Z}/4\mathbb{Z}, \quad \mathbf{Pin}(1,0) \approx \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z},$$

from which we get $\mathbf{Spin}(0,1) = \mathbf{Spin}(1,0) \approx \mathbb{Z}/2\mathbb{Z}$. Also, show that

$$\mathrm{Cl}_{0,2} \approx \mathbb{H}, \quad \mathrm{Cl}_{1,1} \approx M_2(\mathbb{R}), \quad \mathrm{Cl}_{2,0} \approx M_2(\mathbb{R}),$$

where $M_n(\mathbb{R})$ denotes the algebra of $n \times n$ matrices. One can also work out what are $\mathbf{Pin}(2,0)$, $\mathbf{Pin}(1,1)$, and $\mathbf{Pin}(0,2)$; see Choquet-Bruhat [36], Chapter I, Section 7, page 26. Show that

$$\mathbf{Spin}(0,2) = \mathbf{Spin}(2,0) \approx \mathbf{U}(1),$$

and

$$\mathbf{Spin}(1,1) = \{a1 + be_1e_2 \mid a^2 - b^2 = 1\}.$$

Observe that $\mathbf{Spin}(1,1)$ is not connected.

More generally, it can be shown that $\mathrm{Cl}^0_{p,q}$ and $\mathrm{Cl}^0_{q,p}$ are isomorphic, from which it follows that $\mathbf{Spin}(p,q)$ and $\mathbf{Spin}(q,p)$ are isomorphic, but $\mathbf{Pin}(p,q)$ and $\mathbf{Pin}(q,p)$ are not isomorphic in general, and in particular, $\mathbf{Pin}(p,0)$ and $\mathbf{Pin}(0,p)$ are not isomorphic in general (see Choquet-Bruhat [36], Chapter I, Section 7). However, due to the "8-periodicity" of the Clifford algebras (to be discussed in the next section), it follows that $\mathrm{Cl}_{p,q}$ and $\mathrm{Cl}_{q,p}$ are isomorphic when $|p-q| = 0 \bmod 4$.

21.6 Periodicity of the Clifford Algebras $\mathrm{Cl}_{p,q}$

It turns out that the real algebras $\mathrm{Cl}_{p,q}$ can be build up as tensor products of the basic algebras $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{H}$. As pointed out by Lounesto (Section 23.16 [100]), the description of the real algebras $\mathrm{Cl}_{p,q}$ as matrix algebras and the 8-periodicity was first observed by Elie Cartan in 1908; see Cartan's article, *Nombres Complexes*, based on the original article in German by E. Study, in Molk [113], article I-5 (fasc. 3), pages 329-468. These algebras are defined in Section 36 under the name "'Systems of Clifford and Lipschitz," page 463-466. Of course, Cartan used a very different notation; see page 464 in the article cited above.

These facts were rediscovered independently by Raoul Bott in the 1960's (see Raoul Bott's comments in Volume 2 of his Collected papers.).

We will use the notation $\mathbb{R}(n)$ (resp. $\mathbb{C}(n)$) for the algebra $M_n(\mathbb{R})$ of all $n \times n$ real matrices (resp. the algebra $M_n(\mathbb{C})$ of all $n \times n$ complex matrices). As mentioned in Example 21.3, it is not hard to show that

$$\begin{aligned} \mathrm{Cl}_{0,1} &= \mathbb{C} \qquad \mathrm{Cl}_{1,0} = \mathbb{R} \oplus \mathbb{R} \\ \mathrm{Cl}_{0,2} &= \mathbb{H} \qquad \mathrm{Cl}_{2,0} = \mathbb{R}(2) \end{aligned}$$

and

$$\mathrm{Cl}_{1,1} = \mathbb{R}(2).$$

The key to the classification is the following lemma:

Lemma 21.15. *We have the isomorphisms*

$$\begin{aligned} \mathrm{Cl}_{0,n+2} &\approx \mathrm{Cl}_{n,0} \otimes \mathrm{Cl}_{0,2} \\ \mathrm{Cl}_{n+2,0} &\approx \mathrm{Cl}_{0,n} \otimes \mathrm{Cl}_{2,0} \\ \mathrm{Cl}_{p+1,q+1} &\approx \mathrm{Cl}_{p,q} \otimes \mathrm{Cl}_{1,1}, \end{aligned}$$

for all $n, p, q \geq 0$.

Proof. Let $\Phi_{0,n}(x) = -\|x\|^2$, where $\|x\|$ is the standard Euclidean norm on $\mathbb{R}^{n+2}$, and let $(e_1, \ldots, e_{n+2})$ be an orthonormal basis for $\mathbb{R}^{n+2}$ under the standard Euclidean inner product. We also let $(e'_1, \ldots, e'_n)$ be a set of generators for $\mathrm{Cl}_{n,0}$ and (e''_1, e''_2) be a set of generators for $\mathrm{Cl}_{0,2}$. We can define a linear map $f\colon \mathbb{R}^{n+2} \to \mathrm{Cl}_{n,0} \otimes \mathrm{Cl}_{0,2}$ by its action on the basis $(e_1, \ldots, e_{n+2})$ as follows:

$$f(e_i) = \begin{cases} e'_i \otimes e''_1 e''_2 & \text{for } 1 \leq i \leq n \\ 1 \otimes e''_{i-n} & \text{for } n+1 \leq i \leq n+2. \end{cases}$$

Observe that for $1 \leq i, j \leq n$, we have

$$f(e_i)f(e_j) + f(e_j)f(e_i) = (e'_i e'_j + e'_j e'_i) \otimes (e''_1 e''_e)^2 = -2\delta_{ij} 1 \otimes 1,$$

since $e''_1 e''_2 = -e''_2 e''_1$, $(e''_1)^2 = -1$, and $(e''_2)^2 = -1$, and $e'_i e'_j = -e'_j e'_i$, for all $i \neq j$, and $(e'_i)^2 = 1$, for all i with $1 \leq i \leq n$. Also, for $n+1 \leq i, j \leq n+2$, we have

$$f(e_i)f(e_j) + f(e_j)f(e_i) = 1 \otimes (e''_{i-n} e''_{j-n} + e''_{j-n} e''_{i-n}) = -2\delta_{ij} 1 \otimes 1,$$

and

$$f(e_i)f(e_k) + f(e_k)f(e_i) = 2e'_i \otimes (e''_1 e''_2 e''_{n-k} + e''_{n-k} e''_1 e''_2) = 0,$$

for $1 \leq i, j \leq n$ and $n+1 \leq k \leq n+2$ (since $e''_{n-k} = e''_1$ or $e''_{n-k} = e''_2$). Thus, we have

$$f(x)^2 = -\|x\|^2 \cdot 1 \otimes 1 \quad \text{for all } x \in \mathbb{R}^{n+2},$$

and by the universal mapping property of $\mathrm{Cl}_{0,n+2}$, we get an algebra map

$$\widetilde{f}\colon \mathrm{Cl}_{0,n+2} \to \mathrm{Cl}_{n,0} \otimes \mathrm{Cl}_{0,2}.$$

Since $\widetilde{f}$ maps onto a set of generators, it is surjective. However

$$\dim(\mathrm{Cl}_{0,n+2}) = 2^{n+2} = 2^n \cdot 2 = \dim(\mathrm{Cl}_{n,0})\dim(\mathrm{Cl}_{0,2}) = \dim(\mathrm{Cl}_{n,0} \otimes \mathrm{Cl}_{0,2}),$$

and $\widetilde{f}$ is an isomorphism.

The proof of the second identity is analogous. For the third identity, we have

$$\Phi_{p,q}(x_1, \ldots, x_{p+q}) = x_1^2 + \cdots + x_p^2 - (x_{p+1}^2 + \cdots + x_{p+q}^2),$$

and let $(e_1, \ldots, e_{p+1}, \epsilon_1, \ldots, \epsilon_{q+1})$ be an orthogonal basis for $\mathbb{R}^{p+q+2}$ so that $\Phi_{p+1,q+1}(e_i) = +1$ and $\Phi_{p+1,q+1}(\epsilon_j) = -1$ for $i = 1, \ldots, p+1$ and $j = 1, \ldots, q+1$. Also, let $(e'_1, \ldots, e'_p, \epsilon'_1, \ldots, \epsilon'_q)$ be a set of generators for $\mathrm{Cl}_{p,q}$ and (e''_1, ϵ''_1) be a set of generators for $\mathrm{Cl}_{1,1}$. We define a linear map $f\colon \mathbb{R}^{p+q+2} \to \mathrm{Cl}_{p,q} \otimes \mathrm{Cl}_{1,1}$ by its action on the basis as follows:

$$f(e_i) = \begin{cases} e'_i \otimes e''_1\epsilon''_1 & \text{for } 1 \leq i \leq p \\ 1 \otimes e''_1 & \text{for } i = p+1, \end{cases}$$

and

$$f(\epsilon_j) = \begin{cases} \epsilon'_j \otimes e''_1\epsilon''_1 & \text{for } 1 \leq j \leq q \\ 1 \otimes \epsilon''_1 & \text{for } j = q+1. \end{cases}$$

We can check that

$$f(x)^2 = \Phi_{p+1,q+1}(x) \cdot 1 \otimes 1 \quad \text{for all } x \in \mathbb{R}^{p+q+2},$$

and we finish the proof as in the first case. □

To apply this lemma, we need some further isomorphisms among various matrix algebras.

Proposition 21.16. *The following isomorphisms hold:*

$$\begin{aligned} \mathbb{R}(m) \otimes \mathbb{R}(n) &\approx \mathbb{R}(mn) && \text{for all } m, n \geq 0 \\ \mathbb{R}(n) \otimes_{\mathbb{R}} K &\approx K(n) && \text{for } K = \mathbb{C} \text{ or } K = \mathbb{H} \text{ and all } n \geq 0 \\ \mathbb{C} \otimes_{\mathbb{R}} \mathbb{C} &\approx \mathbb{C} \oplus \mathbb{C} \\ \mathbb{C} \otimes_{\mathbb{R}} \mathbb{H} &\approx \mathbb{C}(2) \\ \mathbb{H} \otimes_{\mathbb{R}} \mathbb{H} &\approx \mathbb{R}(4). \end{aligned}$$

Proof. Details can be found in Lawson and Michelsohn [97]. The first two isomorphisms are quite obvious. The third isomorphism $\mathbb{C} \oplus \mathbb{C} \to \mathbb{C} \otimes \mathbb{C}$ is obtained by sending

$$(1, 0) \mapsto \frac{1}{2}(1 \otimes 1 + i \otimes i), \quad (0, 1) \mapsto \frac{1}{2}(1 \otimes 1 - i \otimes i).$$

The field $\mathbb{C}$ is isomorphic to the subring of $\mathbb{H}$ generated by $\mathbf{i}$. Thus, we can view $\mathbb{H}$ as a $\mathbb{C}$-vector space under left scalar multiplication. Consider the $\mathbb{R}$-bilinear map $\pi\colon \mathbb{C}\times\mathbb{H}\to \mathrm{Hom}_{\mathbb{C}}(\mathbb{H},\mathbb{H})$ given by

$$\pi_{y,z}(x) = yx\overline{z},$$

where $y\in\mathbb{C}$ and $x,z\in\mathbb{H}$. Thus, we get an $\mathbb{R}$-linear map $\pi\colon \mathbb{C}\otimes_{\mathbb{R}}\mathbb{H}\to \mathrm{Hom}_{\mathbb{C}}(\mathbb{H},\mathbb{H})$. However, we have $\mathrm{Hom}_{\mathbb{C}}(\mathbb{H},\mathbb{H})\approx\mathbb{C}(2)$. Furthermore, since

$$\pi_{y,z}\circ\pi_{y',z'} = \pi_{yy',zz'},$$

the map π is an algebra homomorphism

$$\pi\colon \mathbb{C}\times\mathbb{H}\to\mathbb{C}(2).$$

We can check on a basis that π is injective, and since

$$\dim_{\mathbb{R}}(\mathbb{C}\times\mathbb{H}) = \dim_{\mathbb{R}}(\mathbb{C}(2)) = 8,$$

the map π is an isomorphism. The last isomorphism is proved in a similar fashion. □

We now have the main periodicity theorem.

Theorem 21.17. *(Cartan/Bott) For all $n\geq 0$, we have the following isomorphisms:*

$$\begin{aligned}\mathrm{Cl}_{0,n+8} &\approx \mathrm{Cl}_{0,n}\otimes\mathrm{Cl}_{0,8}\\ \mathrm{Cl}_{n+8,0} &\approx \mathrm{Cl}_{n,0}\otimes\mathrm{Cl}_{8,0}.\end{aligned}$$

Furthermore,

$$\mathrm{Cl}_{0,8} = \mathrm{Cl}_{8,0} = \mathbb{R}(16).$$

Proof. By Lemma 21.15 we have the isomorphisms

$$\begin{aligned}\mathrm{Cl}_{0,n+2} &\approx \mathrm{Cl}_{n,0}\otimes\mathrm{Cl}_{0,2}\\ \mathrm{Cl}_{n+2,0} &\approx \mathrm{Cl}_{0,n}\otimes\mathrm{Cl}_{2,0},\end{aligned}$$

and thus,

$$\mathrm{Cl}_{0,n+8}\approx\mathrm{Cl}_{n+6,0}\otimes\mathrm{Cl}_{0,2}\approx\mathrm{Cl}_{0,n+4}\otimes\mathrm{Cl}_{2,0}\otimes\mathrm{Cl}_{0,2}\approx\cdots\approx\mathrm{Cl}_{0,n}\otimes\mathrm{Cl}_{2,0}\otimes\mathrm{Cl}_{0,2}\otimes\mathrm{Cl}_{2,0}\otimes\mathrm{Cl}_{0,2}.$$

Since $\mathrm{Cl}_{0,2}=\mathbb{H}$ and $\mathrm{Cl}_{2,0}=\mathbb{R}(2)$, by Proposition 21.16, we get

$$\mathrm{Cl}_{2,0}\otimes\mathrm{Cl}_{0,2}\otimes\mathrm{Cl}_{2,0}\otimes\mathrm{Cl}_{0,2}\approx\mathbb{H}\otimes\mathbb{H}\otimes\mathbb{R}(2)\otimes\mathbb{R}(2)\approx\mathbb{R}(4)\otimes\mathbb{R}(4)\approx\mathbb{R}(16).$$

The second isomorphism is proved in a similar fashion. □

From all this, we can deduce the following table:

n	0	1	2	3	4	5	6	7	8
$\mathrm{Cl}_{0,n}$	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{H}$	$\mathbb{H}\oplus\mathbb{H}$	$\mathbb{H}(2)$	$\mathbb{C}(4)$	$\mathbb{R}(8)$	$\mathbb{R}(8)\oplus\mathbb{R}(8)$	$\mathbb{R}(16)$
$\mathrm{Cl}_{n,0}$	$\mathbb{R}$	$\mathbb{R}\oplus\mathbb{R}$	$\mathbb{R}(2)$	$\mathbb{C}(2)$	$\mathbb{H}(2)$	$\mathbb{H}(2)\oplus\mathbb{H}(2)$	$\mathbb{H}(4)$	$\mathbb{C}(8)$	$\mathbb{R}(16)$

A table of the Clifford groups $\mathrm{Cl}_{p,q}$ for $0 \leq p, q \leq 7$ can be found in Kirillov [86], and for $0 \leq p, q \leq 8$, in Lawson and Michelsohn [97] (but beware that their $\mathrm{Cl}_{p,q}$ is our $\mathrm{Cl}_{q,p}$). It can also be shown that

$$\begin{aligned} \mathrm{Cl}_{p+1,q} &\approx \mathrm{Cl}_{q+1,p} \\ \mathrm{Cl}_{p,q} &\approx \mathrm{Cl}_{p-4,q+4} \end{aligned}$$

with $p \geq 4$ in the second identity (see Lounesto [100], Chapter 16, Sections 16.3 and 16.4). Using the second identity, if $|p-q| = 4k$, it is easily shown by induction on k that $\mathrm{Cl}_{p,q} \approx \mathrm{Cl}_{q,p}$, as claimed in the previous section.

We also have the isomorphisms

$$\mathrm{Cl}_{p,q} \approx \mathrm{Cl}^0_{p,q+1},$$

frow which it follows that

$$\mathbf{Spin}(p,q) \approx \mathbf{Spin}(q,p)$$

(see Choquet-Bruhat [36], Chapter I, Sections 4 and 7). However, in general, $\mathbf{Pin}(p,q)$ and $\mathbf{Pin}(q,p)$ are not isomorphic. In fact, $\mathbf{Pin}(0,n)$ and $\mathbf{Pin}(n,0)$ are not isomorphic if $n \neq 4k$, with $k \in \mathbb{N}$ (see Choquet-Bruhat [36], Chapter I, Section 7, page 27).

21.7 The Complex Clifford Algebras $\mathrm{Cl}(n,\mathbb{C})$

One can also consider Clifford algebras over the complex field $\mathbb{C}$. In this case, it is well-known that every nondegenerate quadratic form can be expressed by

$$\Phi^{\mathbb{C}}_n(x_1,\ldots,x_n) = x_1^2 + \cdots + x_n^2$$

in some orthonormal basis. Also, it is easily shown that the complexification $\mathbb{C} \otimes_{\mathbb{R}} \mathrm{Cl}_{p,q}$ of the real Clifford algebra $\mathrm{Cl}_{p,q}$ is isomorphic to $\mathrm{Cl}(\Phi^{\mathbb{C}}_n)$. Thus, all these complex algebras are isomorphic for $p+q=n$, and we denote them by $\mathrm{Cl}(n,\mathbb{C})$. Theorem 21.15 yields the following periodicity theorem:

Theorem 21.18. *The following isomorphisms hold:*

$$\mathrm{Cl}(n+2,\mathbb{C}) \approx \mathrm{Cl}(n,\mathbb{C}) \otimes_{\mathbb{C}} \mathrm{Cl}(2,\mathbb{C}),$$

with $\mathrm{Cl}(2,\mathbb{C}) = \mathbb{C}(2)$.

Proof. Since $\mathrm{Cl}(n,\mathbb{C}) = \mathbb{C}\otimes_{\mathbb{R}}\mathrm{Cl}_{0,n} = \mathbb{C}\otimes_{\mathbb{R}}\mathrm{Cl}_{n,0}$, by Lemma 21.15, we have

$$\mathrm{Cl}(n+2,\mathbb{C}) = \mathbb{C}\otimes_{\mathbb{R}}\mathrm{Cl}_{0,n+2} \approx \mathbb{C}\otimes_{\mathbb{R}}(\mathrm{Cl}_{n,0}\otimes_{\mathbb{R}}\mathrm{Cl}_{0,2}) \approx (\mathbb{C}\otimes_{\mathbb{R}}\mathrm{Cl}_{n,0})\otimes_{\mathbb{C}}(\mathbb{C}\otimes_{\mathbb{R}}\mathrm{Cl}_{0,2}).$$

However, $\mathrm{Cl}_{0,2} = \mathbb{H}$, $\mathrm{Cl}(n,\mathbb{C}) = \mathbb{C}\otimes_{\mathbb{R}}\mathrm{Cl}_{n,0}$, and $\mathbb{C}\otimes_{\mathbb{R}}\mathbb{H}\approx\mathbb{C}(2)$, so we get $\mathrm{Cl}(2,\mathbb{C}) = \mathbb{C}(2)$ and

$$\mathrm{Cl}(n+2,\mathbb{C}) \approx \mathrm{Cl}(n,\mathbb{C})\otimes_{\mathbb{C}}\mathbb{C}(2),$$

and the theorem is proved. □

As a corollary of Theorem 21.18, we obtain the fact that

$$\mathrm{Cl}(2k,\mathbb{C}) \approx \mathbb{C}(2^k) \quad \text{and} \quad \mathrm{Cl}(2k+1,\mathbb{C}) \approx \mathbb{C}(2^k)\oplus\mathbb{C}(2^k).$$

The table of the previous section can also be completed as follows:

n	0	1	2	3	4	5	6	7	8
$\mathrm{Cl}_{0,n}$	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{H}$	$\mathbb{H}\oplus\mathbb{H}$	$\mathbb{H}(2)$	$\mathbb{C}(4)$	$\mathbb{R}(8)$	$\mathbb{R}(8)\oplus\mathbb{R}(8)$	$\mathbb{R}(16)$
$\mathrm{Cl}_{n,0}$	$\mathbb{R}$	$\mathbb{R}\oplus\mathbb{R}$	$\mathbb{R}(2)$	$\mathbb{C}(2)$	$\mathbb{H}(2)$	$\mathbb{H}(2)\oplus\mathbb{H}(2)$	$\mathbb{H}(4)$	$\mathbb{C}(8)$	$\mathbb{R}(16)$
$\mathrm{Cl}(n,\mathbb{C})$	$\mathbb{C}$	$2\mathbb{C}$	$\mathbb{C}(2)$	$2\mathbb{C}(2)$	$\mathbb{C}(4)$	$2\mathbb{C}(4)$	$\mathbb{C}(8)$	$2\mathbb{C}(8)$	$\mathbb{C}(16)$.

where $2\mathbb{C}(k)$ is an abbrevation for $\mathbb{C}(k)\oplus\mathbb{C}(k)$.

21.8 The Groups $\mathbf{Pin}(p,q)$ and $\mathbf{Spin}(p,q)$ as double covers of $\mathbf{O}(p,q)$ and $\mathbf{SO}(p,q)$

It turns out that the groups $\mathbf{Pin}(p,q)$ and $\mathbf{Spin}(p,q)$ have nice topological properties w.r.t. the groups $\mathbf{O}(p,q)$ and $\mathbf{SO}(p,q)$. To explain this, we review the definition of covering maps and covering spaces (for details, see Fulton [57], Chapter 11). Another interesting source is Chevalley [34], where is is proved that $\mathbf{Spin}(n)$ is a universal double cover of $\mathbf{SO}(n)$ for all $n \geq 3$.

Since $C_{p,q}$ is an algebra of dimension 2^{p+q}, it is a topological space as a vector space isomorphic to $V = \mathbb{R}^{2^{p+q}}$. Now, the group $C^*_{p,q}$ of units of $C_{p,q}$ is open in $C_{p,q}$, because $x \in C_{p,q}$ is a unit if the linear map $y \mapsto xy$ is an isomorphism, and $\mathbf{GL}(V)$ is open in $\mathrm{End}(V)$, the space of endomorphisms of V. Thus, $C^*_{p,q}$ is a Lie group, and since $\mathbf{Pin}(p,q)$ and $\mathbf{Spin}(p,q)$ are clearly closed subgroups of $C^*_{p,q}$, they are also Lie groups.

The definition below is analogous to the definition of a covering map given in Section 3.9 (Definition 3.37) except that now, we are only dealing with topological spaces and not manifolds.

Definition 21.7. Given two topological spaces X and Y, a *covering map* is a continuous surjective map, $p\colon Y \to X$, with the property that for every $x \in X$, there is some open subset, $U \subseteq X$, with $x \in U$, so that $p^{-1}(U)$ is the disjoint union of open subsets, $V_\alpha \subseteq Y$,

and the restriction of p to each V_α is a homeomorphism onto U. We say that U is *evenly covered by* p. We also say that Y is a *covering space of* X. A covering map $p\colon Y \to X$ is called *trivial* if X itself is evenly covered by p (i.e., Y is the disjoint union of open subsets, Y_α, each homeomorphic to X), and *nontrivial*, otherwise. When each fiber, $p^{-1}(x)$, has the same finite cardinaly n for all $x \in X$, we say that p is an *n-covering* (or *n-sheeted covering*).

Note that a covering map, $p\colon Y \to X$, is not always trivial, but always *locally trivial* (i.e., for every $x \in X$, it is trivial in some open neighborhood of x). A covering is trivial iff Y is isomorphic to a product space of $X \times T$, where T is any set with the discrete topology. Also, if Y is connected, then the covering map is nontrivial.

Definition 21.8. An *isomorphism* φ between covering maps $p\colon Y \to X$ and $p'\colon Y' \to X$ is a homeomorphism, $\varphi\colon Y \to Y'$, so that $p = p' \circ \varphi$.

Typically, the space X is connected, in which case it can be shown that all the fibers $p^{-1}(x)$ have the same cardinality.

One of the most important properties of covering spaces is the path–lifting property, a property that we will use to show that $\mathbf{Spin}(n)$ is path-connected. The proposition below is the analog of Proposition 3.39 for topological spaces and continuous curves.

Proposition 21.19. *(Path lifting) Let $p\colon Y \to X$ be a covering map, and let $\gamma\colon [a,b] \to X$ be any continuous curve from $x_a = \gamma(a)$ to $x_b = \gamma(b)$ in X. If $y \in Y$ is any point so that $p(y) = x_a$, then there is a unique curve, $\widetilde{\gamma}\colon [a,b] \to Y$, so that $y = \widetilde{\gamma}(a)$ and*

$$p \circ \widetilde{\gamma}(t) = \gamma(t) \quad \textit{for all } t \in [a,b].$$

Proof. See Fulton [58], Chapter 11, Lemma 11.6. □

Many important covering maps arise from the action of a group G on a space Y. If Y is a topological space, an *action (on the left) of a group G on Y* is a map $\alpha\colon G \times Y \to Y$ satisfying the following conditions, where, for simplicity of notation, we denote $\alpha(g,y)$ by $g \cdot y$:

(1) $g \cdot (h \cdot y) = (gh) \cdot y$, for all $g, h \in G$ and $y \in Y$;

(2) $1 \cdot y = y$, for all $\in Y$, where 1 is the identity of the group G;

(3) The map $y \mapsto g \cdot y$ is a homeomorphism of Y for every $g \in G$.

We define an equivalence relation on Y as follows: $x \equiv y$ iff $y = g \cdot x$ for some $g \in G$ (check that this is an equivalence relation). The equivalence class $G \cdot x = \{g \cdot x \mid g \in G\}$ of any $x \in Y$ is called the *orbit of x*. We obtain the quotient space Y/G and the projection map $p\colon Y \to Y/G$ sending every $y \in Y$ to its orbit. The space Y/G is given the quotient topology (a subset U of Y/G is open iff $p^{-1}(U)$ is open in Y).

Given a subset V of Y and any $g \in G$, we let

$$g \cdot V = \{g \cdot y \mid y \in V\}.$$

We say that G *acts evenly on* Y if for every $y \in Y$ there is an open subset V containing y so that $g \cdot V$ and $h \cdot V$ are disjoint for any two distinct elements $g, h \in G$.

The importance of the notion a group acting evenly is that such actions induce a covering map.

Proposition 21.20. *If G is a group acting evenly on a space Y, then the projection map, $p\colon Y \to Y/G$, is a covering map.*

Proof. See Fulton [58], Chapter 11, Lemma 11.17. □

The following proposition shows that $\mathbf{Pin}(p,q)$ and $\mathbf{Spin}(p,q)$ are nontrivial covering spaces unless $p = q = 1$.

Proposition 21.21. *For all $p, q \geq 0$, the groups $\mathbf{Pin}(p,q)$ and $\mathbf{Spin}(p,q)$ are double covers of $\mathbf{O}(p,q)$ and $\mathbf{SO}(p,q)$, respectively. Furthermore, they are nontrivial covers unless $p = q = 1$.*

Proof. We know that kernel of the homomorphism $\rho\colon \mathbf{Pin}(p,q) \to \mathbf{O}(p,q)$ is $\mathbb{Z}_2 = \{-1, 1\}$. If we let $\mathbb{Z}_2$ act on $\mathbf{Pin}(p,q)$ in the natural way, then $\mathbf{O}(p,q) \approx \mathbf{Pin}(p,q)/\mathbb{Z}_2$, and the reader can easily check that $\mathbb{Z}_2$ acts evenly. By Proposition 21.20, we get a double cover. The argument for $\rho\colon \mathbf{Spin}(p,q) \to \mathbf{SO}(p,q)$ is similar.

Let us now assume that $p \neq 1$ and $q \neq 1$. In order to prove that we have nontrivial covers, it is enough to show that -1 and 1 are connected by a path in $\mathbf{Pin}(p,q)$ (If we had $\mathbf{Pin}(p,q) = U_1 \cup U_2$ with U_1 and U_2 open, disjoint, and homeomorphic to $\mathbf{O}(p,q)$, then -1 and 1 would not be in the same U_i, and so, they would be in disjoint connected components. Thus, -1 and 1 can't be path–connected, and similarly with $\mathbf{Spin}(p,q)$ and $\mathbf{SO}(p,q)$.) Since $(p,q) \neq (1,1)$, we can find two orthogonal vectors e_1 and e_2 so that $\Phi_{p,q}(e_1) = \Phi_{p,q}(e_2) = \pm 1$. Then,

$$\gamma(t) = \pm\cos(2t)\,1 + \sin(2t)\,e_1e_2 = (\cos t\,e_1 + \sin t\,e_2)(\sin t\,e_2 - \cos t\,e_1),$$

for $0 \leq t \leq \pi$, defines a path in $\mathbf{Spin}(p,q)$, since

$$(\pm\cos(2t)\,1 + \sin(2t)\,e_1e_2)^{-1} = \pm\cos(2t)\,1 - \sin(2t)\,e_1e_2,$$

as desired. □

In particular, if $n \geq 2$, since the group $\mathbf{SO}(n)$ is path-connected, the group $\mathbf{Spin}(n)$ is also path-connected. Indeed, given any two points x_a and x_b in $\mathbf{Spin}(n)$, there is a path γ from $\rho(x_a)$ to $\rho(x_b)$ in $\mathbf{SO}(n)$ (where $\rho\colon \mathbf{Spin}(n) \to \mathbf{SO}(n)$ is the covering map). By Proposition 21.19, there is a path $\widetilde{\gamma}$ in $\mathbf{Spin}(n)$ with origin x_a and some origin $\widetilde{x_b}$ so that

$\rho(\widetilde{x}_b) = \rho(x_b)$. However, $\rho^{-1}(\rho(x_b)) = \{-x_b, x_b\}$, and so, $\widetilde{x}_b = \pm x_b$. The argument used in the proof of Proposition 21.21 shows that x_b and $-x_b$ are path-connected, and so, there is a path from x_a to x_b, and $\mathbf{Spin}(n)$ is path-connected. In fact, for $n \geq 3$, it turns out that $\mathbf{Spin}(n)$ is simply connected. Such a covering space is called a *universal cover* (for instance, see Chevalley [34]).

This last fact requires more algebraic topology than we are willing to explain in detail, and we only sketch the proof. The notions of fibre bundle, fibration, and homotopy sequence associated with a fibration are needed in the proof. We refer the perseverant readers to Bott and Tu [19] (Chapter 1 and Chapter 3, Sections 16–17) or Rotman [129] (Chapter 11) for a detailed explanation of these concepts.

Recall that a topological space is *simply connected* if it is path connected and if $\pi_1(X) = (0)$, which means that every closed path in X is homotopic to a point. Since we just proved that $\mathbf{Spin}(n)$ is path connected for $n \geq 2$, we just need to prove that $\pi_1(\mathbf{Spin}(n)) = (0)$ for all $n \geq 3$. The following facts are needed to prove the above assertion:

(1) The sphere S^{n-1} is simply connected for all $n \geq 3$.

(2) The group $\mathbf{Spin}(3) \simeq \mathbf{SU}(2)$ is homeomorphic to S^3, and thus, $\mathbf{Spin}(3)$ is simply connected.

(3) The group $\mathbf{Spin}(n)$ acts on S^{n-1} in such a way that we have a fibre bundle with fibre $\mathbf{Spin}(n-1)$:

$$\mathbf{Spin}(n-1) \longrightarrow \mathbf{Spin}(n) \longrightarrow S^{n-1}.$$

Fact (1) is a standard proposition of algebraic topology and a proof can found in many books. A particularly elegant and yet simple argument consists in showing that any closed curve on S^{n-1} is homotopic to a curve that omits some point. First, it is easy to see that in $\mathbb{R}^n$, any closed curve is homotopic to a piecewise linear curve (a polygonal curve), and the radial projection of such a curve on S^{n-1} provides the desired curve. Then, we use the stereographic projection of S^{n-1} from any point omitted by that curve to get another closed curve in $\mathbb{R}^{n-1}$. Since $\mathbb{R}^{n-1}$ is simply connected, that curve is homotopic to a point, and so is its preimage curve on S^{n-1}. Another simple proof uses a special version of the Seifert—van Kampen's theorem (see Gramain [64]).

Fact (2) is easy to establish directly, using (1).

To prove (3), we let $\mathbf{Spin}(n)$ act on S^{n-1} via the standard action: $x \cdot v = xvx^{-1}$. Because $\mathbf{SO}(n)$ acts transitively on S^{n-1} and there is a surjection $\mathbf{Spin}(n) \longrightarrow \mathbf{SO}(n)$, the group $\mathbf{Spin}(n)$ also acts transitively on S^{n-1}. Now, we have to show that the stabilizer of any element of S^{n-1} is $\mathbf{Spin}(n-1)$. For example, we can do this for e_1. This amounts to some simple calculations taking into account the identities among basis elements. Details of this proof can be found in Mneimné and Testard [112], Chapter 4. It is still necessary to prove that $\mathbf{Spin}(n)$ is a fibre bundle over S^{n-1} with fibre $\mathbf{Spin}(n-1)$. For this, we use the following results whose proof can be found in Mneimné and Testard [112], Chapter 4:

Lemma 21.22. *Given any topological group G, if H is a closed subgroup of G and the projection $\pi\colon G \to G/H$ has a local section at every point of G/H, then*

$$H \longrightarrow G \longrightarrow G/H$$

is a fibre bundle with fibre H.

Lemma 21.22 implies the following key proposition:

Proposition 21.23. *Given any linear Lie group G, if H is a closed subgroup of G, then*

$$H \longrightarrow G \longrightarrow G/H$$

is a fibre bundle with fibre H.

Now, a fibre bundle is a fibration (as defined in Bott and Tu [19], Chapter 3, Section 16, or in Rotman [129], Chapter 11). For a proof of this fact, see Rotman [129], Chapter 11, or Mneimné and Testard [112], Chapter 4. So, there is a homotopy sequence associated with the fibration (Bott and Tu [19], Chapter 3, Section 17, or Rotman [129], Chapter 11, Theorem 11.48), and in particular, we have the exact sequence

$$\pi_1(\mathbf{Spin}(n-1)) \longrightarrow \pi_1(\mathbf{Spin}(n)) \longrightarrow \pi_1(S^{n-1}).$$

Since $\pi_1(S^{n-1}) = (0)$ for $n \geq 3$, we get a surjection

$$\pi_1(\mathbf{Spin}(n-1)) \longrightarrow \pi_1(\mathbf{Spin}(n)),$$

and so, by induction and (2), we get

$$\pi_1(\mathbf{Spin}(n)) \approx \pi_1(\mathbf{Spin}(3)) = (0),$$

proving that $\mathbf{Spin}(n)$ is simply connected for $n \geq 3$.

We can also show that $\pi_1(\mathbf{SO}(n)) = \mathbb{Z}/2\mathbb{Z}$ for all $n \geq 3$. For this, we use Theorem 21.11 and Proposition 21.21, which imply that $\mathbf{Spin}(n)$ is a fibre bundle over $\mathbf{SO}(n)$ with fibre $\{-1, 1\}$, for $n \geq 2$:

$$\{-1, 1\} \longrightarrow \mathbf{Spin}(n) \longrightarrow \mathbf{SO}(n).$$

Again, the homotopy sequence of the fibration exists, and in particular, we get the exact sequence

$$\pi_1(\mathbf{Spin}(n)) \longrightarrow \pi_1(\mathbf{SO}(n)) \longrightarrow \pi_0(\{-1, +1\}) \longrightarrow \pi_0(\mathbf{SO}(n)).$$

Since $\pi_0(\{-1, +1\}) = \mathbb{Z}/2\mathbb{Z}$, $\pi_0(\mathbf{SO}(n)) = (0)$, and $\pi_1(\mathbf{Spin}(n)) = (0)$, when $n \geq 3$, we get the exact sequence

$$(0) \longrightarrow \pi_1(\mathbf{SO}(n)) \longrightarrow \mathbb{Z}/2\mathbb{Z} \longrightarrow (0),$$

and so, $\pi_1(\mathbf{SO}(n)) = \mathbb{Z}/2\mathbb{Z}$. Therefore, $\mathbf{SO}(n)$ is not simply connected for $n \geq 3$.

Remark: Of course, we have been rather cavalier in our presentation. Given a topological space, X, the group $\pi_1(X)$ is the *fundamental group of* X, *i.e.*, the group of homotopy classes of closed paths in X (under composition of loops). But $\pi_0(X)$ is generally *not* a group! Instead, $\pi_0(X)$ is the set of path-connected components of X. However, when X is a Lie group, $\pi_0(X)$ is indeed a group. Also, we have to make sense of what it means for the sequence to be exact. All this can be made rigorous (see Bott and Tu [19], Chapter 3, Section 17, or Rotman [129], Chapter 11).

Chapter 22

Tensor Algebras, Symmetric Algebras and Exterior Algebras

22.1 Tensors Products

We begin by defining tensor products of vector spaces over a field and then we investigate some basic properties of these tensors, in particular the existence of bases and duality. After this, we investigate special kinds of tensors, namely, symmetric tensors and skew-symmetric tensors. Tensor products of modules over a commutative ring with identity will be discussed very briefly. They show up naturally when we consider the space of sections of a tensor product of vector bundles.

Given a linear map, $f\colon E \to F$, we know that if we have a basis, $(u_i)_{i\in I}$, for E, then f is completely determined by its values, $f(u_i)$, on the basis vectors. For a multilinear map, $f\colon E^n \to F$, we don't know if there is such a nice property but it would certainly be very useful.

In many respects, tensor products allow us to define multilinear maps in terms of their action on a suitable basis. The crucial idea is to *linearize*, that is, to create a new vector space, $E^{\otimes n}$, such that the multilinear map, $f\colon E^n \to F$, is turned into a *linear map*, $f_\otimes\colon E^{\otimes n} \to F$, which is equivalent to f in a strong sense. If in addition, f is symmetric, then we can define a symmetric tensor power, $\mathrm{Sym}^n(E)$, and every symmetric multilinear map, $f\colon E^n \to F$, is turned into a *linear map*, $f_\odot\colon \mathrm{Sym}^n(E) \to F$, which is equivalent to f in a strong sense. Similarly, if f is alternating, then we can define a skew-symmetric tensor power, $\bigwedge^n(E)$, and every alternating multilinear map is turned into a *linear map*, $f_\wedge\colon \bigwedge^n(E) \to F$, which is equivalent to f in a strong sense.

Tensor products can be defined in various ways, some more abstract than others. We tried to stay down to earth, without excess!

Let K be a given field, and let $E_1, \ldots, E_n$ be $n \geq 2$ given vector spaces. For any vector space, F, recall that a map, $f\colon E_1 \times \cdots \times E_n \to F$, is *multilinear* iff it is linear in each of

its argument, that is,

$$\begin{aligned} f(u_1, \ldots u_{i_1}, v + w, u_{i+1}, \ldots, u_n) &= f(u_1, \ldots u_{i_1}, v, u_{i+1}, \ldots, u_n) \\ &\quad + f(u_1, \ldots u_{i_1}, w, u_{i+1}, \ldots, u_n) \\ f(u_1, \ldots u_{i_1}, \lambda v, u_{i+1}, \ldots, u_n) &= \lambda f(u_1, \ldots u_{i_1}, v, u_{i+1}, \ldots, u_n), \end{aligned}$$

for all $u_j \in E_j$ $(j \neq i)$, all $v, w \in E_i$ and all $\lambda \in K$, for $i = 1 \ldots, n$.

The set of multilinear maps as above forms a vector space denoted $\mathrm{L}(E_1, \ldots, E_n; F)$ or $\mathrm{Hom}(E_1, \ldots, E_n; F)$. When $n = 1$, we have the vector space of linear maps, $\mathrm{L}(E, F)$ or $\mathrm{Hom}(E, F)$. (To be very precise, we write $\mathrm{Hom}_K(E_1, \ldots, E_n; F)$ and $\mathrm{Hom}_K(E, F)$.) As usual, the *dual space, E^*, of E* is defined by $E^* = \mathrm{Hom}(E, K)$.

Before proceeding any further, we recall a basic fact about pairings. We will use this fact to deal with dual spaces of tensors.

Definition 22.1. Given two vector spaces, E and F, a map, $(-,-)\colon E \times F \to K$, is a *nondegenerate pairing* iff it is bilinear and iff $(u, v) = 0$ for all $v \in F$ implies $u = 0$ and $(u, v) = 0$ for all $u \in E$ implies $v = 0$. A nondegenerate pairing induces two linear maps, $\varphi\colon E \to F^*$ and $\psi\colon F \to E^*$, defined by

$$\begin{aligned} \varphi(u)(y) &= (u, y) \\ \psi(v)(x) &= (x, v), \end{aligned}$$

for all $u, x \in E$ and all $v, y \in F$.

Proposition 22.1. *For every nondegenerate pairing, $(-,-)\colon E \times F \to K$, the induced maps $\varphi\colon E \to F^*$ and $\psi\colon F \to E^*$ are linear and injective. Furthermore, if E and F are finite dimensional, then $\varphi\colon E \to F^*$ and $\psi\colon F \to E^*$ are bijective.*

Proof. The maps $\varphi\colon E \to F^*$ and $\psi\colon F \to E^*$ are linear because $u, v \mapsto (u, v)$ is bilinear. Assume that $\varphi(u) = 0$. This means that $\varphi(u)(y) = (u, y) = 0$ for all $y \in F$ and as our pairing is nondegenerate, we must have $u = 0$. Similarly, ψ is injective. If E and F are finite dimensional, then $\dim(E) = \dim(E^*)$ and $\dim(F) = \dim(F^*)$. However, the injectivity of φ and ψ implies that that $\dim(E) \leq \dim(F^*)$ and $\dim(F) \leq \dim(E^*)$. Consequently $\dim(E) \leq \dim(F)$ and $\dim(F) \leq \dim(E)$, so $\dim(E) = \dim(F)$. Therefore, $\dim(E) = \dim(F^*)$ and φ is bijective (and similarly $\dim(F) = \dim(E^*)$ and ψ is bijective). □

Proposition 22.1 shows that when E and F are finite dimensional, a nondegenerate pairing induces *canonical isomorphims* $\varphi\colon E \to F^*$ and $\psi\colon F \to E^*$, that is, isomorphisms that do not depend on the choice of bases. An important special case is the case where $E = F$ and we have an inner product (a symmetric, positive definite bilinear form) on E.

Remark: When we use the term "canonical isomorphism" we mean that such an isomorphism is defined independently of any choice of bases. For example, if E is a finite dimensional

vector space and $(e_1, \ldots, e_n)$ is any basis of E, we have the dual basis, $(e_1^*, \ldots, e_n^*)$, of E^* (where, $e_i^*(e_j) = \delta_{i\,j}$) and thus, the map $e_i \mapsto e_i^*$ is an isomorphism between E and E^*. This isomorphism is *not* canonical.

On the other hand, if $\langle -, - \rangle$ is an inner product on E, then Proposition 22.1 shows that the nondegenerate pairing, $\langle -, - \rangle$, induces a canonical isomorphism between E and E^*. This isomorphism is often denoted $\flat \colon E \to E^*$ and we usually write $u^\flat$ for $\flat(u)$, with $u \in E$. Given any basis, $(e_1, \ldots, e_n)$, of E (not necessarily orthonormal), if we let $g_{ij} = (e_i, e_j)$, then for every $u = \sum_{i=1}^n u_i e_i$, since $u^\flat(v) = \langle u, v \rangle$, for all $v \in V$, we get

$$u^\flat = \sum_{i=1}^n \omega_i e_i^*, \quad \text{with} \quad \omega_i = \sum_{j=1}^n g_{ij} u_j.$$

If we use the convention that coordinates of vectors are written using superscripts ($u = \sum_{i=1}^n u^i e_i$) and coordinates of one-forms (covectors) are written using subscripts ($\omega = \sum_{i=1}^n \omega_i e_i^*$), then the map, $\flat$, has the effect of lowering (flattening!) indices. The inverse of $\flat$ is denoted $\sharp \colon E^* \to E$. If we write $\omega \in E^*$ as $\omega = \sum_{i=1}^n \omega_i e_i^*$ and $\omega^\sharp \in E$ as $\omega^\sharp = \sum_{j=1}^n (\omega^\sharp)^j e_j$, since

$$\omega_i = \omega(e_i) = \langle \omega^\sharp, e_i \rangle = \sum_{j=1}^n (\omega^\sharp)^j g_{ij}, \qquad 1 \leq i \leq n,$$

we get

$$(\omega^\sharp)^i = \sum_{j=1}^n g^{ij} \omega_j,$$

where (g^{ij}) is the inverse of the matrix (g_{ij}). The inner product, $(-, -)$, on E induces an inner product on E^* also denoted $(-, -)$ and given by

$$(\omega_1, \omega_2) = (\omega_1^\sharp, \omega_2^\sharp),$$

for all $\omega_1, \omega_2 \in E^*$. Then, it is obvious that

$$(u, v) = (u^\flat, v^\flat), \quad \text{for all} \quad u, v \in E.$$

If $(e_1, \ldots, e_n)$ is a basis of E and $g_{ij} = (e_i, e_j)$, as

$$(e_i^*)^\sharp = \sum_{k=1}^n g^{ik} e_k,$$

an easy computation shows that

$$(e_i^*, e_j^*) = ((e_i^*)^\sharp, (e_j^*)^\sharp) = g^{ij},$$

that is, in the basis $(e_1^*, \ldots, e_n^*)$, the inner product on E^* is represented by the matrix (g^{ij}), the inverse of the matrix (g_{ij}).

The inner product on a finite vector space also yields a natural isomorphism between the space, $\mathrm{Hom}(E, E; K)$, of bilinear forms on E and the space, $\mathrm{Hom}(E, E)$, of linear maps from E to itself. Using this isomorphism, we can define the trace of a bilinear form in an intrinsic manner. This technique is used in differential geometry, for example, to define the divergence of a differential one-form.

Proposition 22.2. *If $\langle -, - \rangle$ is an inner product on a finite vector space, E, (over a field, K), then for every bilinear form, $f\colon E \times E \to K$, there is a unique linear map, $f^\sharp\colon E \to E$, such that*

$$f(u, v) = \langle f^\sharp(u), v \rangle, \quad \textit{for all } u, v \in E.$$

The map, $f \mapsto f^\sharp$, is a linear isomorphism between $\mathrm{Hom}(E, E; K)$ and $\mathrm{Hom}(E, E)$.

Proof. For every $g \in \mathrm{Hom}(E, E)$, the map given by

$$f(u, v) = \langle g(u), v \rangle, \quad u, v \in E,$$

is clearly bilinear. It is also clear that the above defines a linear map from $\mathrm{Hom}(E, E)$ to $\mathrm{Hom}(E, E; K)$. This map is injective because if $f(u, v) = 0$ for all $u, v \in E$, as $\langle -, - \rangle$ is an inner product, we get $g(u) = 0$ for all $u \in E$. Furthermore, both spaces $\mathrm{Hom}(E, E)$ and $\mathrm{Hom}(E, E; K)$ have the same dimension, so our linear map is an isomorphism. □

If $(e_1, \ldots, e_n)$ is an orthonormal basis of E, then we check immediately that the trace of a linear map, g, (which is independent of the choice of a basis) is given by

$$\mathrm{tr}(g) = \sum_{i=1}^{n} \langle g(e_i), e_i \rangle,$$

where $n = \dim(E)$. We define the *trace of the bilinear form*, f, by

$$\mathrm{tr}(f) = \mathrm{tr}(f^\sharp).$$

From Proposition 22.2, $\mathrm{tr}(f)$ is given by

$$\mathrm{tr}(f) = \sum_{i=1}^{n} f(e_i, e_i),$$

for any orthonormal basis, $(e_1, \ldots, e_n)$, of E. We can also check directly that the above expression is independent of the choice of an orthonormal basis.

We will also need the following Proposition to show that various families are linearly independent.

Proposition 22.3. *Let E and F be two nontrivial vector spaces and let $(u_i)_{i\in I}$ be any family of vectors $u_i \in E$. The family, $(u_i)_{i\in I}$, is linearly independent iff for every family, $(v_i)_{i\in I}$, of vectors $v_i \in F$, there is some linear map, $f\colon E \to F$, so that $f(u_i) = v_i$, for all $i \in I$.*

Proof. Left as an exercise. □

First, we define tensor products, and then we prove their existence and uniqueness up to isomorphism.

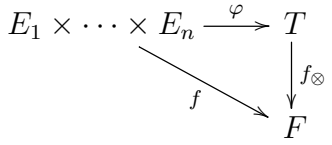
Definition 22.2. A *tensor product* of $n \geq 2$ vector spaces $E_1, \ldots, E_n$, is a vector space T, together with a multilinear map $\varphi\colon E_1 \times \cdots \times E_n \to T$, such that, for every vector space F and for every multilinear map $f\colon E_1 \times \cdots \times E_n \to F$, there is a unique linear map $f_\otimes\colon T \to F$, with

$$f(u_1, \ldots, u_n) = f_\otimes(\varphi(u_1, \ldots, u_n)),$$

for all $u_1 \in E_1, \ldots, u_n \in E_n$, or for short

$$f = f_\otimes \circ \varphi.$$

Equivalently, there is a unique linear map $f_\otimes$ such that the following diagram commutes:



First, we show that any two tensor products (T_1, φ_1) and (T_2, φ_2) for $E_1, \ldots, E_n$, are isomorphic.

Proposition 22.4. *Given any two tensor products (T_1, φ_1) and (T_2, φ_2) for $E_1, \ldots, E_n$, there is an isomorphism $h\colon T_1 \to T_2$ such that*

$$\varphi_2 = h \circ \varphi_1.$$

Proof. Focusing on (T_1, φ_1), we have a multilinear map $\varphi_2\colon E_1 \times \cdots \times E_n \to T_2$, and thus, there is a unique linear map $(\varphi_2)_\otimes\colon T_1 \to T_2$, with

$$\varphi_2 = (\varphi_2)_\otimes \circ \varphi_1.$$

Similarly, focusing now on on (T_2, φ_2), we have a multilinear map $\varphi_1\colon E_1 \times \cdots \times E_n \to T_1$, and thus, there is a unique linear map $(\varphi_1)_\otimes\colon T_2 \to T_1$, with

$$\varphi_1 = (\varphi_1)_\otimes \circ \varphi_2.$$

But then, we get

$$\varphi_1 = (\varphi_1)_\otimes \circ (\varphi_2)_\otimes \circ \varphi_1,$$

and

$$\varphi_2 = (\varphi_2)_\otimes \circ (\varphi_1)_\otimes \circ \varphi_2.$$

On the other hand, focusing on (T_1, φ_1), we have a multilinear map $\varphi_1 \colon E_1 \times \cdots \times E_n \to T_1$, but the unique linear map $h \colon T_1 \to T_1$, with

$$\varphi_1 = h \circ \varphi_1$$

is $h = \mathrm{id}$, and since $(\varphi_1)_\otimes \circ (\varphi_2)_\otimes$ is linear, as a composition of linear maps, we must have

$$(\varphi_1)_\otimes \circ (\varphi_2)_\otimes = \mathrm{id}.$$

Similarly, we must have

$$(\varphi_2)_\otimes \circ (\varphi_1)_\otimes = \mathrm{id}.$$

This shows that $(\varphi_1)_\otimes$ and $(\varphi_2)_\otimes$ are inverse linear maps, and thus, $(\varphi_2)_\otimes \colon T_1 \to T_2$ is an isomorphism between T_1 and T_2. □

Now that we have shown that tensor products are unique up to isomorphism, we give a construction that produces one.

Theorem 22.5. *Given $n \geq 2$ vector spaces $E_1, \ldots, E_n$, a tensor product $(E_1 \otimes \cdots \otimes E_n, \varphi)$ for $E_1, \ldots, E_n$ can be constructed. Furthermore, denoting $\varphi(u_1, \ldots, u_n)$ as $u_1 \otimes \cdots \otimes u_n$, the tensor product $E_1 \otimes \cdots \otimes E_n$ is generated by the vectors $u_1 \otimes \cdots \otimes u_n$, where $u_1 \in E_1, \ldots, u_n \in E_n$, and for every multilinear map $f \colon E_1 \times \cdots \times E_n \to F$, the unique linear map $f_\otimes \colon E_1 \otimes \cdots \otimes E_n \to F$ such that $f = f_\otimes \circ \varphi$, is defined by*

$$f_\otimes(u_1 \otimes \cdots \otimes u_n) = f(u_1, \ldots, u_n),$$

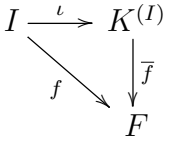
on the generators $u_1 \otimes \cdots \otimes u_n$ of $E_1 \otimes \cdots \otimes E_n$.

Proof. Given any set, I, viewed as an index set, let $K^{(I)}$ be the set of all functions, $f \colon I \to K$, such that $f(i) \neq 0$ only for finitely many $i \in I$. As usual, denote such a function by $(f_i)_{i \in I}$, it is a family of finite support. We make $K^{(I)}$ into a vector space by defining addition and scalar multiplication by

$$\begin{aligned} (f_i) + (g_i) &= (f_i + g_i) \\ \lambda(f_i) &= (\lambda f_i). \end{aligned}$$

The family, $(e_i)_{i \in I}$, is defined such that $(e_i)_j = 0$ if $j \neq i$ and $(e_i)_i = 1$. It is a basis of the vector space $K^{(I)}$, so that every $w \in K^{(I)}$ can be uniquely written as a finite linear combination of the e_i. There is also an injection, $\iota \colon I \to K^{(I)}$, such that $\iota(i) = e_i$ for every

$i \in I$. Furthermore, it is easy to show that for any vector space, F, and for any function, $f\colon I \to F$, there is a unique linear map, $\overline{f}\colon K^{(I)} \to F$, such that $f = \overline{f} \circ \iota$, as in the following diagram:



This shows that $K^{(I)}$ is the *free vector space generated by* I. Now, apply this construction to the cartesian product, $I = E_1 \times \cdots \times E_n$, obtaining the free vector space $M = K^{(I)}$ on $I = E_1 \times \cdots \times E_n$. Since every, e_i, is uniquely associated with some n-tuple $i = (u_1, \ldots, u_n) \in E_1 \times \cdots \times E_n$, we will denote e_i by $(u_1, \ldots, u_n)$.

Next, let N be the subspace of M generated by the vectors of the following type:

$$(u_1, \ldots, u_i + v_i, \ldots, u_n) - (u_1, \ldots, u_i, \ldots, u_n) - (u_1, \ldots, v_i, \ldots, u_n),$$
$$(u_1, \ldots, \lambda u_i, \ldots, u_n) - \lambda(u_1, \ldots, u_i, \ldots, u_n).$$

We let $E_1 \otimes \cdots \otimes E_n$ be the quotient M/N of the free vector space M by N, $\pi\colon M \to M/N$ be the quotient map and set

$$\varphi = \pi \circ \iota.$$

By construction, φ is multilinear, and since π is surjective and the $\iota(i) = e_i$ generate M, since i is of the form $i = (u_1, \ldots, u_n) \in E_1 \times \cdots \times E_n$, the $\varphi(u_1, \ldots, u_n)$ generate M/N. Thus, if we denote $\varphi(u_1, \ldots, u_n)$ as $u_1 \otimes \cdots \otimes u_n$, the tensor product $E_1 \otimes \cdots \otimes E_n$ is generated by the vectors $u_1 \otimes \cdots \otimes u_n$, where $u_1 \in E_1, \ldots, u_n \in E_n$.

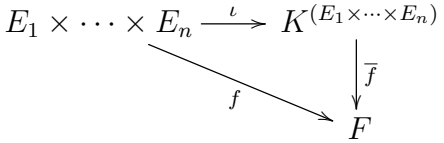
For every multilinear map $f\colon E_1 \times \cdots \times E_n \to F$, if a linear map $f_\otimes\colon E_1 \otimes \cdots \otimes E_n \to F$ exists such that $f = f_\otimes \circ \varphi$, since the vectors $u_1 \otimes \cdots \otimes u_n$ generate $E_1 \otimes \cdots \otimes E_n$, the map $f_\otimes$ is uniquely defined by

$$f_\otimes(u_1 \otimes \cdots \otimes u_n) = f(u_1, \ldots, u_n).$$

On the other hand, because $M = K^{(E_1 \times \cdots \times E_n)}$ is free on $I = E_1 \times \cdots \times E_n$, there is a unique linear map $\overline{f}\colon K^{(E_1 \times \cdots \times E_n)} \to F$, such that

$$f = \overline{f} \circ \iota,$$

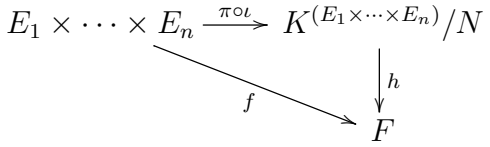
as in the diagram below:



Because f is multilinear, note that we must have $\overline{f}(w) = 0$, for every $w \in N$. But then, $\overline{f}\colon M \to F$ induces a linear map $h\colon M/N \to F$, such that

$$f = h \circ \pi \circ \iota,$$

by defining $h([z]) = \overline{f}(z)$, for every $z \in M$, where $[z]$ denotes the equivalence class in M/N of $z \in M$:



Indeed, the fact that $\overline{f}$ vanishes on N insures that h is well defined on M/N, and it is clearly linear by definition. However, we showed that such a linear map h is unique, and thus it agrees with the linear map $f_\otimes$ defined by

$$f_\otimes(u_1 \otimes \cdots \otimes u_n) = f(u_1, \ldots, u_n)$$

on the generators of $E_1 \otimes \cdots \otimes E_n$. □

What is important about Theorem 22.5 is not so much the construction itself but the fact that it produces a tensor product with the universal mapping property with respect to multilinear maps. Indeed, Theorem 22.5 yields a canonical isomorphism,

$$\mathrm{L}(E_1 \otimes \cdots \otimes E_n, F) \cong \mathrm{L}(E_1, \ldots, E_n; F),$$

between the vector space of linear maps, $\mathrm{L}(E_1 \otimes \cdots \otimes E_n, F)$, and the vector space of multilinear maps, $\mathrm{L}(E_1, \ldots, E_n; F)$, via the linear map $- \circ \varphi$ defined by

$$h \mapsto h \circ \varphi,$$

where $h \in \mathrm{L}(E_1 \otimes \cdots \otimes E_n, F)$. Indeed, $h \circ \varphi$ is clearly multilinear, and since by Theorem 22.5, for every multilinear map, $f \in \mathrm{L}(E_1, \ldots, E_n; F)$, there is a unique linear map $f_\otimes \in \mathrm{L}(E_1 \otimes \cdots \otimes E_n, F)$ such that $f = f_\otimes \circ \varphi$, the map $- \circ \varphi$ is bijective. As a matter of fact, its inverse is the map

$$f \mapsto f_\otimes.$$

Using the "Hom" notation, the above canonical isomorphism is written

$$\mathrm{Hom}(E_1 \otimes \cdots \otimes E_n, F) \cong \mathrm{Hom}(E_1, \ldots, E_n; F).$$

Remarks:

(1) To be very precise, since the tensor product depends on the field, K, we should subscript the symbol $\otimes$ with K and write

$$E_1 \otimes_K \cdots \otimes_K E_n.$$

However, we often omit the subscript K unless confusion may arise.

(2) For $F = K$, the base field, we obtain a canonical isomorphism between the vector space $\mathrm{L}(E_1 \otimes \cdots \otimes E_n, K)$, and the vector space of multilinear forms $\mathrm{L}(E_1, \ldots, E_n; K)$. However, $\mathrm{L}(E_1 \otimes \cdots \otimes E_n, K)$ is the dual space, $(E_1 \otimes \cdots \otimes E_n)^*$, and thus, the vector space of multilinear forms $\mathrm{L}(E_1, \ldots, E_n; K)$ is canonically isomorphic to $(E_1 \otimes \cdots \otimes E_n)^*$. We write

$$\mathrm{L}(E_1, \ldots, E_n; K) \cong (E_1 \otimes \cdots \otimes E_n)^*.$$

The fact that the map $\varphi\colon E_1 \times \cdots \times E_n \to E_1 \otimes \cdots \otimes E_n$ is multilinear, can also be expressed as follows:

$$u_1 \otimes \cdots \otimes (v_i + w_i) \otimes \cdots \otimes u_n = (u_1 \otimes \cdots \otimes v_i \otimes \cdots \otimes u_n) + (u_1 \otimes \cdots \otimes w_i \otimes \cdots \otimes u_n),$$
$$u_1 \otimes \cdots \otimes (\lambda u_i) \otimes \cdots \otimes u_n = \lambda(u_1 \otimes \cdots \otimes u_i \otimes \cdots \otimes u_n).$$

Of course, this is just what we wanted! Tensors in $E_1 \otimes \cdots \otimes E_n$ are also called *n-tensors*, and tensors of the form $u_1 \otimes \cdots \otimes u_n$, where $u_i \in E_i$, are called *simple (or indecomposable) n-tensors*. Those n-tensors that are not simple are often called *compound n-tensors*.

Not only do tensor products act on spaces, but they also act on linear maps (they are functors). Given two linear maps $f\colon E \to E'$ and $g\colon F \to F'$, we can define $h\colon E \times F \to E' \otimes F'$ by

$$h(u, v) = f(u) \otimes g(v).$$

It is immediately verified that h is bilinear, and thus, it induces a unique linear map

$$f \otimes g\colon E \otimes F \to E' \otimes F',$$

such that

$$(f \otimes g)(u \otimes v) = f(u) \otimes g(u).$$

If we also have linear maps $f'\colon E' \to E''$ and $g'\colon F' \to F''$, we can easily verify that the linear maps $(f' \circ f) \otimes (g' \circ g)$ and $(f' \otimes g') \circ (f \otimes g)$ agree on all vectors of the form $u \otimes v \in E \otimes F$. Since these vectors generate $E \otimes F$, we conclude that

$$(f' \circ f) \otimes (g' \circ g) = (f' \otimes g') \circ (f \otimes g).$$

The generalization to the tensor product $f_1 \otimes \cdots \otimes f_n$ of $n \geq 3$ linear maps $f_i\colon E_i \to F_i$ is immediate, and left to the reader.

22.2 Bases of Tensor Products

We showed that $E_1 \otimes \cdots \otimes E_n$ is generated by the vectors of the form $u_1 \otimes \cdots \otimes u_n$. However, there vectors are not linearly independent. This situation can be fixed when considering bases, which is the object of the next proposition.

Proposition 22.6. *Given $n \geq 2$ vector spaces $E_1, \ldots, E_n$, if $(u_i^k)_{i \in I_k}$ is a basis for E_k, $1 \leq k \leq n$, then the family of vectors*

$$(u_{i_1}^1 \otimes \cdots \otimes u_{i_n}^n)_{(i_1,\ldots,i_n) \in I_1 \times \ldots \times I_n}$$

is a basis of the tensor product $E_1 \otimes \cdots \otimes E_n$.

Proof. For each k, $1 \leq k \leq n$, every $v^k \in E_k$ can be written uniquely as

$$v^k = \sum_{j \in I_k} v_j^k u_j^k,$$

for some family of scalars $(v_j^k)_{j \in I_k}$. Let F be any nontrivial vector space. We show that for every family

$$(w_{i_1,\ldots,i_n})_{(i_1,\ldots,i_n) \in I_1 \times \ldots \times I_n},$$

of vectors in F, there is some linear map $h \colon E_1 \otimes \cdots \otimes E_n \to F$, such that

$$h(u_{i_1}^1 \otimes \cdots \otimes u_{i_n}^n) = w_{i_1,\ldots,i_n}.$$

Then, by Proposition 22.3, it follows that

$$(u_{i_1}^1 \otimes \cdots \otimes u_{i_n}^n)_{(i_1,\ldots,i_n) \in I_1 \times \ldots \times I_n}$$

is linearly independent. However, since $(u_i^k)_{i \in I_k}$ is a basis for E_k, the $u_{i_1}^1 \otimes \cdots \otimes u_{i_n}^n$ also generate $E_1 \otimes \cdots \otimes E_n$, and thus, they form a basis of $E_1 \otimes \cdots \otimes E_n$.

We define the function $f \colon E_1 \times \cdots \times E_n \to F$ as follows:

$$f(\sum_{j_1 \in I_1} v_{j_1}^1 u_{j_1}^1, \ldots, \sum_{j_n \in I_n} v_{j_n}^n u_{j_n}^n) = \sum_{j_1 \in I_1, \ldots, j_n \in I_n} v_{j_1}^1 \cdots v_{j_n}^n \, w_{j_1,\ldots,j_n}.$$

It is immediately verified that f is multilinear. By the universal mapping property of the tensor product, the linear map $f_\otimes \colon E_1 \otimes \cdots \otimes E_n \to F$ such that $f = f_\otimes \circ \varphi$, is the desired map h. □

In particular, when each I_k is finite and of size $m_k = \dim(E_k)$, we see that the dimension of the tensor product $E_1 \otimes \cdots \otimes E_n$ is $m_1 \cdots m_n$. As a corollary of Proposition 22.6, if $(u_i^k)_{i \in I_k}$ is a basis for E_k, $1 \leq k \leq n$, then every tensor $z \in E_1 \otimes \cdots \otimes E_n$ can be written in a unique way as

$$z = \sum_{(i_1,\ldots,i_n) \in I_1 \times \ldots \times I_n} \lambda_{i_1,\ldots,i_n} \, u_{i_1}^1 \otimes \cdots \otimes u_{i_n}^n,$$

for some unique family of scalars $\lambda_{i_1,\ldots,i_n} \in K$, all zero except for a finite number.

22.3 Some Useful Isomorphisms for Tensor Products

Proposition 22.7. *Given* 3 *vector spaces* E, F, G*, there exists unique canonical isomorphisms*

(1) $E \otimes F \simeq F \otimes E$

(2) $(E \otimes F) \otimes G \simeq E \otimes (F \otimes G) \simeq E \otimes F \otimes G$

(3) $(E \oplus F) \otimes G \simeq (E \otimes G) \oplus (F \otimes G)$

(4) $K \otimes E \simeq E$

such that respectively

(a) $u \otimes v \mapsto v \otimes u$

(b) $(u \otimes v) \otimes w \mapsto u \otimes (v \otimes w) \mapsto u \otimes v \otimes w$

(c) $(u,\, v) \otimes w \mapsto (u \otimes w,\, v \otimes w)$

(d) $\lambda \otimes u \mapsto \lambda u$.

Proof. These isomorphisms are proved using the universal mapping property of tensor products. We illustrate the proof method on (2). Fix some $w \in G$. The map

$$(u,\, v) \mapsto u \otimes v \otimes w$$

from $E \times F$ to $E \otimes F \otimes G$ is bilinear, and thus, there is a linear map $f_w \colon E \otimes F \to E \otimes F \otimes G$, such that $f_w(u \otimes v) = u \otimes v \otimes w$.

Next, consider the map

$$(z,\, w) \mapsto f_w(z),$$

from $(E \otimes F) \times G$ into $E \otimes F \otimes G$. It is easily seen to be bilinear, and thus, it induces a linear map

$$f \colon (E \otimes F) \otimes G \to E \otimes F \otimes G,$$

such that $f((u \otimes v) \otimes w) = u \otimes v \otimes w$.

Also consider the map

$$(u,\, v, w) \mapsto (u \otimes v) \otimes w$$

from $E \times F \times G$ to $(E \otimes F) \otimes G$. It is trilinear, and thus, there is a linear map

$$g \colon E \otimes F \otimes G \to (E \otimes F) \otimes G,$$

such that $g(u \otimes v \otimes w) = (u \otimes v) \otimes w$. Clearly, $f \circ g$ and $g \circ f$ are identity maps, and thus, f and g are isomorphisms. The other cases are similar. □

Given any three vector spaces, E, F, G, we have the canonical isomorphism

$$\mathrm{Hom}(E, F; G) \cong \mathrm{Hom}(E, \mathrm{Hom}(F, G)).$$

Indeed, any bilinear map, $f\colon E \times F \to G$, gives the linear map, $\varphi(f) \in \mathrm{Hom}(E, \mathrm{Hom}(F, G))$, where $\varphi(f)(u)$ is the linear map in $\mathrm{Hom}(F, G)$ given by

$$\varphi(f)(u)(v) = f(u, v).$$

Conversely, given a linear map, $g \in \mathrm{Hom}(E, \mathrm{Hom}(F, G))$, we get the bilinear map, $\psi(g)$, given by

$$\psi(g)(u, v) = g(u)(v),$$

and it is clear that φ and ψ and mutual inverses. Consequently, we have the important corollary:

Proposition 22.8. *For any three vector spaces, E, F, G, we have the canonical isomorphism,*

$$\mathrm{Hom}(E \otimes F, G) \cong \mathrm{Hom}(E, \mathrm{Hom}(F, G)),$$

22.4 Duality for Tensor Products

In this section, all vector spaces are assumed to have finite dimension. Let us now see how tensor products behave under duality. For this, we define a pairing between $E_1^* \otimes \cdots \otimes E_n^*$ and $E_1 \otimes \cdots \otimes E_n$ as follows: For any fixed $(v_1^*, \ldots, v_n^*) \in E_1^* \times \cdots \times E_n^*$, we have the multilinear map,

$$l_{v_1^*,\ldots,v_n^*}\colon (u_1, \ldots, u_n) \mapsto v_1^*(u_1) \cdots v_n^*(u_n),$$

from $E_1 \times \cdots \times E_n$ to K. The map $l_{v_1^*,\ldots,v_n^*}$ extends uniquely to a linear map, $L_{v_1^*,\ldots,v_n^*}\colon E_1 \otimes \cdots \otimes E_n \longrightarrow K$. We also have the multilinear map,

$$(v_1^*, \ldots, v_n^*) \mapsto L_{v_1^*,\ldots,v_n^*},$$

from $E_1^* \times \cdots \times E_n^*$ to $\mathrm{Hom}(E_1 \otimes \cdots \otimes E_n, K)$, which extends to a linear map, L, from $E_1^* \otimes \cdots \otimes E_n^*$ to $\mathrm{Hom}(E_1 \otimes \cdots \otimes E_n, K)$. However, in view of the isomorphism,

$$\mathrm{Hom}(U \otimes V, W) \cong \mathrm{Hom}(U, \mathrm{Hom}(V, W)),$$

we can view L as a linear map,

$$L\colon (E_1^* \otimes \cdots \otimes E_n^*) \otimes (E_1 \otimes \cdots \otimes E_n) \to K,$$

which corresponds to a bilinear map,

$$(E_1^* \otimes \cdots \otimes E_n^*) \times (E_1 \otimes \cdots \otimes E_n) \longrightarrow K,$$

via the isomorphism $(U \otimes V)^* \cong \mathrm{L}(U, V; K)$. It is easy to check that this bilinear map is nondegenerate and thus, by Proposition 22.1, we have a canonical isomorphism,

$$(E_1 \otimes \cdots \otimes E_n)^* \cong E_1^* \otimes \cdots \otimes E_n^*.$$

This, together with the isomorphism, $\mathrm{L}(E_1, \ldots, E_n; K) \cong (E_1 \otimes \cdots \otimes E_n)^*$, yields a canonical isomorphism

$$\mathrm{L}(E_1, \ldots, E_n; K) \cong E_1^* \otimes \cdots \otimes E_n^*.$$

We prove another useful canonical isomorphism that allows us to treat linear maps as tensors.

Let E and F be two vector spaces and let $\alpha\colon E^* \times F \to \mathrm{Hom}(E, F)$ be the map defined such that

$$\alpha(u^*, f)(x) = u^*(x)f,$$

for all $u^* \in E^*$, $f \in F$, and $x \in E$. This map is clearly bilinear and thus, it induces a linear map,

$$\alpha_\otimes\colon E^* \otimes F \to \mathrm{Hom}(E, F),$$

such that

$$\alpha_\otimes(u^* \otimes f)(x) = u^*(x)f.$$

Proposition 22.9. *If E and F are vector spaces with E of finite dimension, then the linear map, $\alpha_\otimes\colon E^* \otimes F \to \mathrm{Hom}(E, F)$, is a canonical isomorphism.*

Proof. Let $(e_j)_{1 \leq j \leq n}$ be a basis of E and, as usual, let $e_j^* \in E^*$ be the linear form defined by

$$e_j^*(e_k) = \delta_{j,k},$$

where $\delta_{j,k} = 1$ iff $j = k$ and 0 otherwise. We know that $(e_j^*)_{1 \leq j \leq n}$ is a basis of E^* (this is where we use the finite dimension of E). Now, for any linear map, $f \in \mathrm{Hom}(E, F)$, for every $x = x_1 e_1 + \cdots + x_n e_n \in E$, we have

$$f(x) = f(x_1 e_1 + \cdots + x_n e_n) = x_1 f(e_1) + \cdots + x_n f(e_n) = e_1^*(x)f(e_1) + \cdots + e_n^*(x)f(e_n).$$

Consequently, every linear map, $f \in \mathrm{Hom}(E, F)$, can be expressed as

$$f(x) = e_1^*(x)f_1 + \cdots + e_n^*(x)f_n,$$

for some $f_i \in F$. Furthermore, if we apply f to e_i, we get $f(e_i) = f_i$, so the f_i are unique. Observe that

$$(\alpha_\otimes(e_1^* \otimes f_1 + \cdots + e_n^* \otimes f_n))(x) = \sum_{i=1}^n (\alpha_\otimes(e_i^* \otimes f_i))(x) = \sum_{i=1}^n e_i^*(x)f_i.$$

Thus, $\alpha_\otimes$ is surjective. As $(e_j^*)_{1\leq j\leq n}$ is a basis of E^*, the tensors $e_j^* \otimes f$, with $f \in F$, span $E^* \otimes F$. Thus, every element of $E^* \otimes F$ is of the form $\sum_{i=1}^n e_i^* \otimes f_i$, for some $f_i \in F$. Assume

$$\alpha_\otimes(\sum_{i=1}^n e_i^* \otimes f_i) = \alpha_\otimes(\sum_{i=1}^n e_i^* \otimes f_i') = f,$$

for some $f_i, f_i' \in F$ and some $f \in \mathrm{Hom}(E, F)$. Then for every $x \in E$,

$$\sum_{i=1}^n e_i^*(x) f_i = \sum_{i=1}^n e_i^*(x) f_i' = f(x).$$

Since the f_i and f_i' are uniquely determined by the linear map, f, we must have $f_i = f_i'$ and $\alpha_\otimes$ is injective. Therefore, $\alpha_\otimes$ is a bijection. □

Note that in Proposition 22.9, the space F may have infinite dimension but E has finite dimension. In view of the canonical isomorphism

$$\mathrm{Hom}(E_1, \ldots, E_n; F) \cong \mathrm{Hom}(E_1 \otimes \cdots \otimes E_n, F)$$

and the canonical isomorphism $(E_1 \otimes \cdots \otimes E_n)^* \cong E_1^* \otimes \cdots \otimes E_n^*$, where the E_i's are finite-dimensional, Proposition 22.9 yields the canonical isomorphism

$$\mathrm{Hom}(E_1, \ldots, E_n; F) \cong E_1^* \otimes \cdots \otimes E_n^* \otimes F.$$

22.5 Tensor Algebras

The tensor product

$$\underbrace{V \otimes \cdots \otimes V}_{m}$$

is also denoted as

$$\bigotimes^m V \quad \text{or} \quad V^{\otimes m}$$

and is called the *m-th tensor power of* V (with $V^{\otimes 1} = V$, and $V^{\otimes 0} = K$). We can pack all the tensor powers of V into the "big" vector space,

$$T(V) = \bigoplus_{m\geq 0} V^{\otimes m},$$

also denoted $T^\bullet(V)$, to avoid confusion with the tangent bundle. This is an interesting object because we can define a multiplication operation on it which makes it into an *algebra* called the *tensor algebra of* V. When V is of finite dimension n, this space corresponds to the algebra of polynomials with coefficients in K in n noncommuting variables.

Let us recall the definition of an algebra over a field. Let K denote any (commutative) field, although for our purposes, we may assume that $K = \mathbb{R}$ (and occasionally, $K = \mathbb{C}$). Since we will only be dealing with associative algebras with a multiplicative unit, we only define algebras of this kind.

Definition 22.3. Given a field, K, a *K-algebra* is a K-vector space, A, together with a bilinear operation, $\cdot\colon A \times A \to A$, called *multiplication*, which makes A into a ring with unity, 1 (or 1_A, when we want to be very precise). This means that $\cdot$ is associative and that there is a multiplicative identity element, 1, so that $1 \cdot a = a \cdot 1 = a$, for all $a \in A$. Given two K-algebras A and B, a *K-algebra homomorphism,* $h\colon A \to B$, is a linear map that is also a ring homomorphism, with $h(1_A) = 1_B$.

For example, the ring, $M_n(K)$, of all $n \times n$ matrices over a field, K, is a K-algebra.

There is an obvious notion of *ideal* of a K-algebra: An ideal, $\mathfrak{A} \subseteq A$, is a linear subspace of A that is also a two-sided ideal with respect to multiplication in A. If the field K is understood, we usually simply say an algebra instead of a K-algebra.

We would like to define a multiplication operation on $T(V)$ which makes it into a K-algebra. As

$$T(V) = \bigoplus_{i \geq 0} V^{\otimes i},$$

for every $i \geq 0$, there is a natural injection $\iota_n\colon V^{\otimes n} \to T(V)$, and in particular, an injection $\iota_0\colon K \to T(V)$. The multiplicative unit, $\mathbf{1}$, of $T(V)$ is the image, $\iota_0(1)$, in $T(V)$ of the unit, 1, of the field K. Since every $v \in T(V)$ can be expressed as a finite sum

$$v = \iota_{n_1}(v_1) + \cdots + \iota_{n_k}(v_k),$$

where $v_i \in V^{\otimes n_i}$ and the n_i are natural numbers with $n_i \neq n_j$ if $i \neq j$, to define multiplication in $T(V)$, using bilinearity, it is enough to define multiplication operations,
$\cdot\colon V^{\otimes m} \times V^{\otimes n} \longrightarrow V^{\otimes(m+n)}$, which, using the isomorphisms, $V^{\otimes n} \cong \iota_n(V^{\otimes n})$, yield multiplication operations, $\cdot\colon \iota_m(V^{\otimes m}) \times \iota_n(V^{\otimes n}) \longrightarrow \iota_{m+n}(V^{\otimes(m+n)})$. More precisely, we use the canonical isomorphism,

$$V^{\otimes m} \otimes V^{\otimes n} \cong V^{\otimes(m+n)},$$

which defines a bilinear operation,

$$V^{\otimes m} \times V^{\otimes n} \longrightarrow V^{\otimes(m+n)},$$

which is taken as the multiplication operation. The isomorphism $V^{\otimes m} \otimes V^{\otimes n} \cong V^{\otimes(m+n)}$ can be established by proving the isomorphisms

$$\begin{aligned} V^{\otimes m} \otimes V^{\otimes n} &\cong V^{\otimes m} \otimes \underbrace{V \otimes \cdots \otimes V}_{n} \\ V^{\otimes m} \otimes \underbrace{V \otimes \cdots \otimes V}_{n} &\cong V^{\otimes(m+n)}, \end{aligned}$$

which can be shown using methods similar to those used to proved associativity. Of course, the multiplication, $V^{\otimes m} \times V^{\otimes n} \longrightarrow V^{\otimes(m+n)}$, is defined so that

$$(v_1 \otimes \cdots \otimes v_m) \cdot (w_1 \otimes \cdots \otimes w_n) = v_1 \otimes \cdots \otimes v_m \otimes w_1 \otimes \cdots \otimes w_n.$$

(This has to be made rigorous by using isomorphisms involving the associativity of tensor products, for details, see see Atiyah and Macdonald [9].)

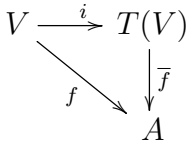
Remark: It is important to note that multiplication in $T(V)$ is **not** commutative. Also, in all rigor, the unit, $\mathbf{1}$, of $T(V)$ is **not equal** to 1, the unit of the field K. However, in view of the injection $\iota_0 \colon K \to T(V)$, for the sake of notational simplicity, we will denote $\mathbf{1}$ by 1. More generally, in view of the injections $\iota_n \colon V^{\otimes n} \to T(V)$, we identify elements of $V^{\otimes n}$ with their images in $T(V)$.

The algebra, $T(V)$, satisfies a universal mapping property which shows that it is unique up to isomorphism. For simplicity of notation, let $i \colon V \to T(V)$ be the natural injection of V into $T(V)$.

Proposition 22.10. *Given any K-algebra, A, for any linear map, $f \colon V \to A$, there is a unique K-algebra homomorphism, $\overline{f} \colon T(V) \to A$, so that*

$$f = \overline{f} \circ i,$$

as in the diagram below:



Proof. Left an an exercise (use Theorem 22.5). □

Most algebras of interest arise as well-chosen quotients of the tensor algebra $T(V)$. This is true for the *exterior algebra,* $\bigwedge(V)$ (also called *Grassmann algebra*), where we take the quotient of $T(V)$ modulo the ideal generated by all elements of the form $v \otimes v$, where $v \in V$, and for the *symmetric algebra,* $\mathrm{Sym}(V)$, where we take the quotient of $T(V)$ modulo the ideal generated by all elements of the form $v \otimes w - w \otimes v$, where $v, w \in V$.

Algebras such as $T(V)$ are graded, in the sense that there is a sequence of subspaces, $V^{\otimes n} \subseteq T(V)$, such that

$$T(V) = \bigoplus_{k \geq 0} V^{\otimes n}$$

and the multiplication, $\otimes$, behaves well w.r.t. the grading, *i.e.*, $\otimes \colon V^{\otimes m} \times V^{\otimes n} \to V^{\otimes (m+n)}$. Generally, a K-algebra, E, is said to be a *graded algebra* iff there is a sequence of subspaces, $E^n \subseteq E$, such that

$$E = \bigoplus_{k \geq 0} E^n$$

($E^0 = K$) and the multiplication, $\cdot$, respects the grading, that is, $\cdot \colon E^m \times E^n \to E^{m+n}$. Elements in E^n are called *homogeneous elements of rank (or degree) n.*

In differential geometry and in physics it is necessary to consider slightly more general tensors.

Definition 22.4. Given a vector space, V, for any pair of nonnegative integers, (r, s), the *tensor space, $T^{r,s}(V)$, of type (r, s)*, is the tensor product

$$T^{r,s}(V) = V^{\otimes r} \otimes (V^*)^{\otimes s} = \underbrace{V \otimes \cdots \otimes V}_{r} \otimes \underbrace{V^* \otimes \cdots \otimes V^*}_{s},$$

with $T^{0,0}(V) = K$. We also define the *tensor algebra*, $T^{\bullet,\bullet}(V)$, as the coproduct

$$T^{\bullet,\bullet}(V) = \bigoplus_{r,s \geq 0} T^{r,s}(V).$$

Tensors in $T^{r,s}(V)$ are called *homogeneous of degree (r, s)*.

Note that tensors in $T^{r,0}(V)$ are just our "old tensors" in $V^{\otimes r}$. We make $T^{\bullet,\bullet}(V)$ into an algebra by defining multiplication operations,

$$T^{r_1,s_1}(V) \times T^{r_2,s_2}(V) \longrightarrow T^{r_1+r_2,s_1+s_2}(V),$$

in the usual way, namely: For $u = u_1 \otimes \cdots \otimes u_{r_1} \otimes u_1^* \otimes \cdots \otimes u_{s_1}^*$ and $v = v_1 \otimes \cdots \otimes v_{r_2} \otimes v_1^* \otimes \cdots \otimes v_{s_2}^*$, let

$$u \otimes v = u_1 \otimes \cdots \otimes u_{r_1} \otimes v_1 \otimes \cdots \otimes v_{r_2} \otimes u_1^* \otimes \cdots \otimes u_{s_1}^* \otimes v_1^* \otimes \cdots \otimes v_{s_2}^*.$$

Denote by $\mathrm{Hom}(V^r, (V^*)^s; W)$ the vector space of all multilinear maps from $V^r \times (V^*)^s$ to W. Then, we have the universal mapping property which asserts that there is a canonical isomorphism

$$\mathrm{Hom}(T^{r,s}(V), W) \cong \mathrm{Hom}(V^r, (V^*)^s; W).$$

In particular,

$$(T^{r,s}(V))^* \cong \mathrm{Hom}(V^r, (V^*)^s; K).$$

For finite dimensional vector spaces, the duality of Section 22.4 is also easily extended to the tensor spaces $T^{r,s}(V)$. We define the pairing

$$T^{r,s}(V^*) \times T^{r,s}(V) \longrightarrow K$$

as follows: If

$$v^* = v_1^* \otimes \cdots \otimes v_r^* \otimes u_{r+1} \otimes \cdots \otimes u_{r+s} \in T^{r,s}(V^*)$$

and

$$u = u_1 \otimes \cdots \otimes u_r \otimes v_{r+1}^* \otimes \cdots \otimes v_{r+s}^* \in T^{r,s}(V),$$

then

$$(v^*, u) = v_1^*(u_1) \cdots v_{r+s}^*(u_{r+s}).$$

This is a nondegenerate pairing and thus, we get a canonical isomorphism,

$$(T^{r,s}(V))^* \cong T^{r,s}(V^*).$$

Consequently, we get a canonical isomorphism,

$$T^{r,s}(V^*) \cong \mathrm{Hom}(V^r, (V^*)^s; K).$$

Remark: The tensor spaces, $T^{r,s}(V)$ are also denoted $T^r_s(V)$. A tensor, $\alpha \in T^{r,s}(V)$ is said to be *contravariant* in the first r arguments and *covariant* in the last s arguments. This terminology refers to the way tensors behave under coordinate changes. Given a basis, $(e_1, \ldots, e_n)$, of V, if $(e_1^*, \ldots, e_n^*)$ denotes the dual basis, then every tensor $\alpha \in T^{r,s}(V)$ is given by an expression of the form

$$\alpha = \sum_{\substack{i_1,\ldots,i_r \\ j_1,\ldots,j_s}} a^{i_1,\ldots,i_r}_{j_1,\ldots,j_s} e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes e^*_{j_1} \otimes \cdots \otimes e^*_{j_s}.$$

The tradition in classical tensor notation is to use lower indices on vectors and upper indices on linear forms and in accordance to *Einstein summation convention* (or *Einstein notation*) the position of the indices on the coefficients is reversed. *Einstein summation convention* is to assume that a summation is performed for all values of every index that appears simultaneously once as an upper index and once as a lower index. According to this convention, the tensor α above is written

$$\alpha = a^{i_1,\ldots,i_r}_{j_1,\ldots,j_s} e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes e^{j_1} \otimes \cdots \otimes e^{j_s}.$$

An older view of tensors is that they are multidimensional arrays of coefficients,

$$\left(a^{i_1,\ldots,i_r}_{j_1,\ldots,j_s}\right),$$

subject to the rules for changes of bases.

Another operation on general tensors, contraction, is useful in differential geometry.

Definition 22.5. For all $r, s \geq 1$, the *contraction*, $c_{i,j} \colon T^{r,s}(V) \to T^{r-1,s-1}(V)$, with $1 \leq i \leq r$ and $1 \leq j \leq s$, is the linear map defined on generators by

$$\begin{aligned} &c_{i,j}(u_1 \otimes \cdots \otimes u_r \otimes v_1^* \otimes \cdots \otimes v_s^*) \\ &\qquad = v_j^*(u_i)\, u_1 \otimes \cdots \otimes \widehat{u_i} \otimes \cdots \otimes u_r \otimes v_1^* \otimes \cdots \otimes \widehat{v_j^*} \otimes \cdots \otimes v_s^*, \end{aligned}$$

where the hat over an argument means that it should be omitted.

Let us figure our what is $c_{1,1} \colon T^{1,1}(V) \to \mathbb{R}$, that is $c_{1,1} \colon V \otimes V^* \to \mathbb{R}$. If $(e_1, \ldots, e_n)$ is a basis of V and $(e_1^*, \ldots, e_n^*)$ is the dual basis, every $h \in V \otimes V^* \cong \mathrm{Hom}(V, V)$ can be expressed as

$$h = \sum_{i,j=1}^{n} a_{ij}\, e_i \otimes e_j^*.$$

As

$$c_{1,1}(e_i \otimes e_j^*) = \delta_{i,j},$$

we get

$$c_{1,1}(h) = \sum_{i=1}^{n} a_{ii} = \mathrm{tr}(h),$$

where $\mathrm{tr}(h)$ is the *trace* of h, where h is viewed as the linear map given by the matrix, (a_{ij}). Actually, since $c_{1,1}$ is defined independently of any basis, $c_{1,1}$ provides an intrinsic definition of the trace of a linear map, $h \in \mathrm{Hom}(V, V)$.

Remark: Using the Einstein summation convention, if

$$\alpha = a^{i_1,\ldots,i_r}_{j_1,\ldots,j_s} e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes e^{j_1} \otimes \cdots \otimes e^{j_s},$$

then

$$c_{k,l}(\alpha) = a^{i_1,\ldots,i_{k-1},i,i_{k+1}\ldots,i_r}_{j_1,\ldots,j_{l-1},i,j_{l+1},\ldots,j_s} e_{i_1} \otimes \cdots \otimes \widehat{e_{i_k}} \otimes \cdots \otimes e_{i_r} \otimes e^{j_1} \otimes \cdots \otimes \widehat{e^{j_l}} \otimes \cdots \otimes e^{j_s}.$$

If E and F are two K-algebras, we know that their tensor product, $E \otimes F$, exists as a vector space. We can make $E \otimes F$ into an algebra as well. Indeed, we have the multilinear map

$$E \times F \times E \times F \longrightarrow E \otimes F$$

given by $(a, b, c, d) \mapsto (ac) \otimes (bd)$, where ac is the product of a and c in E and bd is the product of b and d in F. By the universal mapping property, we get a linear map,

$$E \otimes F \otimes E \otimes F \longrightarrow E \otimes F.$$

Using the isomorphism,

$$E \otimes F \otimes E \otimes F \cong (E \otimes F) \otimes (E \otimes F),$$

we get a linear map,

$$(E \otimes F) \otimes (E \otimes F) \longrightarrow E \otimes F,$$

and thus, a bilinear map,

$$(E \otimes F) \times (E \otimes F) \longrightarrow E \otimes F,$$

which is our multiplication operation in $E \otimes F$. This multiplication is determined by

$$(a \otimes b) \cdot (c \otimes d) = (ac) \otimes (bd).$$

One immediately checks that $E \otimes F$ with this multiplication is a K-algebra.

We now turn to symmetric tensors.

22.6 Symmetric Tensor Powers

Our goal is to come up with a notion of tensor product that will allow us to treat symmetric multilinear maps as linear maps. First, note that we have to restrict ourselves to a single vector space, E, rather then n vector spaces $E_1, \ldots, E_n$, so that symmetry makes sense. Recall that a multilinear map, $f\colon E^n \to F$, is *symmetric* iff

$$f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}) = f(u_1, \ldots, u_n),$$

for all $u_i \in E$ and all permutations, $\sigma\colon \{1, \ldots, n\} \to \{1, \ldots, n\}$. The group of permutations on $\{1, \ldots, n\}$ (the *symmetric group*) is denoted $\mathfrak{S}_n$. The vector space of all symmetric multilinear maps, $f\colon E^n \to F$, is denoted by $\mathrm{S}^n(E; F)$. Note that $\mathrm{S}^1(E; F) = \mathrm{Hom}(E, F)$.

We could proceed directly as in Theorem 22.5, and construct symmetric tensor products from scratch. However, since we already have the notion of a tensor product, there is a more economical method. First, we define symmetric tensor powers.

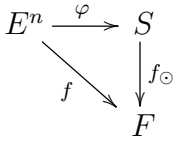
Definition 22.6. An *n-th symmetric tensor power* of a vector space E, where $n \geq 1$, is a vector space S, together with a symmetric multilinear map $\varphi\colon E^n \to S$, such that, for every vector space F and for every symmetric multilinear map $f\colon E^n \to F$, there is a unique linear map $f_\odot\colon S \to F$, with

$$f(u_1, \ldots, u_n) = f_\odot(\varphi(u_1, \ldots, u_n)),$$

for all $u_1, \ldots, u_n \in E$, or for short

$$f = f_\odot \circ \varphi.$$

Equivalently, there is a unique linear map $f_\odot$ such that the following diagram commutes:



First, we show that any two symmetric n-th tensor powers (S_1, φ_1) and (S_2, φ_2) for E, are isomorphic.

Proposition 22.11. *Given any two symmetric n-th tensor powers (S_1, φ_1) and (S_2, φ_2) for E, there is an isomorphism $h\colon S_1 \to S_2$ such that*

$$\varphi_2 = h \circ \varphi_1.$$

Proof. Replace tensor product by n-th symmetric tensor power in the proof of Proposition 22.4. □

We now give a construction that produces a symmetric n-th tensor power of a vector space E.

Theorem 22.12. *Given a vector space E, a symmetric n-th tensor power $(\mathrm{Sym}^n(E), \varphi)$ for E can be constructed ($n \geq 1$). Furthermore, denoting $\varphi(u_1, \ldots, u_n)$ as $u_1 \odot \cdots \odot u_n$, the symmetric tensor power $\mathrm{Sym}^n(E)$ is generated by the vectors $u_1 \odot \cdots \odot u_n$, where $u_1, \ldots, u_n \in E$, and for every symmetric multilinear map $f\colon E^n \to F$, the unique linear map $f_\odot\colon \mathrm{Sym}^n(E) \to F$ such that $f = f_\odot \circ \varphi$, is defined by*

$$f_\odot(u_1 \odot \cdots \odot u_n) = f(u_1, \ldots, u_n),$$

on the generators $u_1 \odot \cdots \odot u_n$ of $\mathrm{Sym}^n(E)$.

Proof. The tensor power $E^{\otimes n}$ is too big, and thus, we define an appropriate quotient. Let C be the subspace of $E^{\otimes n}$ generated by the vectors of the form

$$u_1 \otimes \cdots \otimes u_n - u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(n)},$$

for all $u_i \in E$, and all permutations $\sigma\colon \{1, \ldots, n\} \to \{1, \ldots, n\}$. We claim that the quotient space $(E^{\otimes n})/C$ does the job.

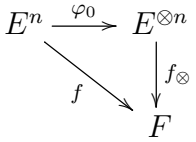
Let $p\colon E^{\otimes n} \to (E^{\otimes n})/C$ be the quotient map. Let $\varphi\colon E^n \to (E^{\otimes n})/C$ be the map

$$(u_1, \ldots, u_n) \mapsto p(u_1 \otimes \cdots \otimes u_n),$$

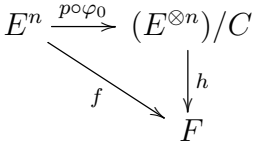
or equivalently, $\varphi = p \circ \varphi_0$, where $\varphi_0(u_1, \ldots, u_n) = u_1 \otimes \cdots \otimes u_n$.

Let us denote $\varphi(u_1, \ldots, u_n)$ as $u_1 \odot \cdots \odot u_n$. It is clear that φ is symmetric. Since the vectors $u_1 \otimes \cdots \otimes u_n$ generate $E^{\otimes n}$, and p is surjective, the vectors $u_1 \odot \cdots \odot u_n$ generate $(E^{\otimes n})/C$.

Given any symmetric multilinear map $f\colon E^n \to F$, there is a linear map $f_\otimes\colon E^{\otimes n} \to F$ such that $f = f_\otimes \circ \varphi_0$, as in the diagram below:



However, since f is symmetric, we have $f_\otimes(z) = 0$ for every $z \in E^{\otimes n}$. Thus, we get an induced linear map $h\colon (E^{\otimes n})/C \to F$, such that $h([z]) = f_\otimes(z)$, where $[z]$ is the equivalence class in $(E^{\otimes n})/C$ of $z \in E^{\otimes n}$:



However, if a linear map $f_\odot\colon (E^{\otimes n})/C \to F$ exists, since the vectors $u_1 \odot \cdots \odot u_n$ generate $(E^{\otimes n})/C$, we must have

$$f_\odot(u_1 \odot \cdots \odot u_n) = f(u_1, \ldots, u_n),$$

which shows that h and $f_\odot$ agree. Thus, $\mathrm{Sym}^n(E) = (E^{\otimes n})/C$ and φ constitute a symmetric n-th tensor power of E. □

Again, the actual construction is not important. What is important is that the symmetric n-th power has the universal mapping property with respect to symmetric multilinear maps.

Remark: The notation $\odot$ for the commutative multiplication of symmetric tensor powers is not standard. Another notation commonly used is $\cdot$. We often abbreviate "symmetric tensor power" as "symmetric power". The symmetric power, $\mathrm{Sym}^n(E)$, is also denoted $\mathrm{Sym}^n E$ or $S(E)$. To be consistent with the use of $\odot$, we could have used the notation $\bigodot^n E$. Clearly, $\mathrm{Sym}^1(E) \cong E$ and it is convenient to set $\mathrm{Sym}^0(E) = K$.

The fact that the map $\varphi\colon E^n \to \mathrm{Sym}^n(E)$ is symmetric and multinear, can also be expressed as follows:

$$\begin{aligned} u_1 \odot \cdots \odot (v_i + w_i) \odot \cdots \odot u_n &= (u_1 \odot \cdots \odot v_i \odot \cdots \odot u_n) + (u_1 \odot \cdots \odot w_i \odot \cdots \odot u_n),\\ u_1 \odot \cdots \odot (\lambda u_i) \odot \cdots \odot u_n &= \lambda(u_1 \odot \cdots \odot u_i \odot \cdots \odot u_n),\\ u_{\sigma(1)} \odot \cdots \odot u_{\sigma(n)} &= u_1 \odot \cdots \odot u_n, \end{aligned}$$

for all permutations $\sigma \in \mathfrak{S}_n$.

The last identity shows that the "operation" $\odot$ is commutative. Thus, we can view the symmetric tensor $u_1 \odot \cdots \odot u_n$ as a multiset.

Theorem 22.12 yields a canonical isomorphism

$$\mathrm{Hom}(\mathrm{Sym}^n(E), F) \cong \mathrm{S}(E^n; F),$$

between the vector space of linear maps $\mathrm{Hom}(\mathrm{Sym}^n(E), F)$, and the vector space of symmetric multilinear maps $\mathrm{S}(E^n; F)$, via the linear map $- \circ \varphi$ defined by

$$h \mapsto h \circ \varphi,$$

where $h \in \mathrm{Hom}(\mathrm{Sym}^n(E), F)$. Indeed, $h \circ \varphi$ is clearly symmetric multilinear, and since by Theorem 22.12, for every symmetric multilinear map $f \in \mathrm{S}(E^n; F)$, there is a unique linear map $f_\odot \in \mathrm{Hom}(\mathrm{Sym}^n(E), F)$ such that $f = f_\odot \circ \varphi$, the map $- \circ \varphi$ is bijective. As a matter of fact, its inverse is the map

$$f \mapsto f_\odot.$$

In particular, when $F = K$, we get a canonical isomorphism

$$(\mathrm{Sym}^n(E))^* \cong \mathrm{S}^n(E; K).$$

Symmetric tensors in $\mathrm{Sym}^n(E)$ are also called *symmetric n-tensors*, and tensors of the form $u_1 \odot \cdots \odot u_n$, where $u_i \in E$, are called *simple (or decomposable) symmetric n-tensors*. Those symmetric n-tensors that are not simple are often called *compound symmetric n-tensors*.

Given two linear maps $f\colon E \to E'$ and $g\colon E \to E'$, we can define $h\colon E \times E \to \mathrm{Sym}^2(E')$ by

$$h(u, v) = f(u) \odot g(v).$$

It is immediately verified that h is symmetric bilinear, and thus, it induces a unique linear map

$$f \odot g \colon \mathrm{Sym}^2(E) \to \mathrm{Sym}^2(E'),$$

such that

$$(f \odot g)(u \odot v) = f(u) \odot g(u).$$

If we also have linear maps $f' \colon E' \to E''$ and $g' \colon E' \to E''$, we can easily verify that

$$(f' \circ f) \odot (g' \circ g) = (f' \odot g') \circ (f \odot g).$$

The generalization to the symmetric tensor product $f_1 \odot \cdots \odot f_n$ of $n \geq 3$ linear maps $f_i \colon E \to E'$ is immediate, and left to the reader.

22.7 Bases of Symmetric Powers

The vectors $u_1 \odot \cdots \odot u_n$, where $u_1, \ldots, u_n \in E$, generate $\mathrm{Sym}^n(E)$, but they are not linearly independent. We will prove a version of Proposition 22.6 for symmetric tensor powers. For this, recall that a (finite) multiset over a set I is a function $M \colon I \to \mathbb{N}$, such that $M(i) \neq 0$ for finitely many $i \in I$, and that the set of all multisets over I is denoted as $\mathbb{N}^{(I)}$. We let $\mathrm{dom}(M) = \{i \in I \mid M(i) \neq 0\}$, which is a finite set. Then, for any multiset $M \in \mathbb{N}^{(I)}$, note that the sum $\sum_{i \in I} M(i)$ makes sense, since $\sum_{i \in I} M(i) = \sum_{i \in \mathrm{dom}(M)} M(i)$, and $\mathrm{dom}(M)$ is finite. For every multiset $M \in \mathbb{N}^{(I)}$, for any $n \geq 2$, we define the set J_M of functions $\eta \colon \{1, \ldots, n\} \to \mathrm{dom}(M)$, as follows:

$$J_M = \{\eta \mid \eta \colon \{1, \ldots, n\} \to \mathrm{dom}(M),\ |\eta^{-1}(i)| = M(i),\ i \in \mathrm{dom}(M),\ \sum_{i \in I} M(i) = n\}.$$

In other words, if $\sum_{i \in I} M(i) = n$ and $\mathrm{dom}(M) = \{i_1, \ldots, i_k\}$,[1] any function $\eta \in J_M$ specifies a sequence of length n, consisting of $M(i_1)$ occurrences of i_1, $M(i_2)$ occurrences of i_2, ..., $M(i_k)$ occurrences of i_k. Intuitively, any η defines a "permutation" of the sequence (of length n)

$$\langle \underbrace{i_1, \ldots, i_1}_{M(i_1)}, \underbrace{i_2, \ldots, i_2}_{M(i_2)}, \ldots, \underbrace{i_k, \ldots, i_k}_{M(i_k)} \rangle.$$

Given any $k \geq 1$, and any $u \in E$, we denote

$$\underbrace{u \odot \cdots \odot u}_{k}$$

as $u^{\odot k}$.

We can now prove the following Proposition.

[1] Note that must have $k \leq n$.

Proposition 22.13. *Given a vector space E, if $(u_i)_{i\in I}$ is a basis for E, then the family of vectors*

$$\left(u_{i_1}^{\odot M(i_1)} \odot \cdots \odot u_{i_k}^{\odot M(i_k)}\right)_{M\in\mathbb{N}^{(I)},\, \sum_{i\in I} M(i)=n,\, \{i_1,\ldots,i_k\}=\mathrm{dom}(M)}$$

is a basis of the symmetric n-th tensor power $\mathrm{Sym}^n(E)$.

Proof. The proof is very similar to that of Proposition 22.6. For any nontrivial vector space F, for any family of vectors

$$(w_M)_{M\in\mathbb{N}^{(I)},\, \sum_{i\in I} M(i)=n},$$

we show the existence of a symmetric multilinear map $h\colon \mathrm{Sym}^n(E) \to F$, such that for every $M \in \mathbb{N}^{(I)}$ with $\sum_{i\in I} M(i) = n$, we have

$$h(u_{i_1}^{\odot M(i_1)} \odot \cdots \odot u_{i_k}^{\odot M(i_k)}) = w_M,$$

where $\{i_1, \ldots, i_k\} = \mathrm{dom}(M)$. We define the map $f\colon E^n \to F$ as follows:

$$f(\sum_{j_1\in I} v_{j_1}^1 u_{j_1}^1, \ldots, \sum_{j_n\in I} v_{j_n}^n u_{j_n}^n) = \sum_{\substack{M\in\mathbb{N}^{(I)}\\ \sum_{i\in I} M(i)=n}} \left(\sum_{\eta\in J_M} v_{\eta(1)}^1 \cdots v_{\eta(n)}^n\right) w_M.$$

It is not difficult to verify that f is symmetric and multilinear. By the universal mapping property of the symmetric tensor product, the linear map $f_\odot\colon \mathrm{Sym}^n(E) \to F$ such that $f = f_\odot \circ \varphi$, is the desired map h. Then, by Proposition 22.3, it follows that the family

$$\left(u_{i_1}^{\odot M(i_1)} \odot \cdots \odot u_{i_k}^{\odot M(i_k)}\right)_{M\in\mathbb{N}^{(I)},\, \sum_{i\in I} M(i)=n,\, \{i_1,\ldots,i_k\}=\mathrm{dom}(M)}$$

is linearly independent. Using the commutativity of $\odot$, we can also show that these vectors generate $\mathrm{Sym}^n(E)$, and thus, they form a basis for $\mathrm{Sym}^n(E)$. The details are left as an exercise. □

As a consequence, when I is finite, say of size $p = \dim(E)$, the dimension of $\mathrm{Sym}^n(E)$ is the number of finite multisets $(j_1, \ldots, j_p)$, such that $j_1 + \cdots + j_p = n$, $j_k \geq 0$. We leave as an exercise to show that this number is $\binom{p+n-1}{n}$. Thus, if $\dim(E) = p$, then the dimension of $\mathrm{Sym}^n(E)$ is $\binom{p+n-1}{n}$. Compare with the dimension of $E^{\otimes n}$, which is p^n. In particular, when $p = 2$, the dimension of $\mathrm{Sym}^n(E)$ is $n + 1$. This can also be seen directly.

Remark: The number $\binom{p+n-1}{n}$ is also the number of homogeneous monomials

$$X_1^{j_1} \cdots X_p^{j_p}$$

of total degree n in p variables (we have $j_1 + \cdots + j_p = n$). This is not a coincidence! Symmetric tensor products are closely related to polynomials (for more on this, see the next remark).

Given a vector space E and a basis $(u_i)_{i\in I}$ for E, Proposition 22.13 shows that every symmetric tensor $z \in \mathrm{Sym}^n(E)$ can be written in a unique way as

$$z = \sum_{\substack{M\in\mathbb{N}^{(I)} \\ \sum_{i\in I} M(i)=n \\ \{i_1,\ldots,i_k\}=\mathrm{dom}(M)}} \lambda_M \, u_{i_1}^{\odot M(i_1)} \odot \cdots \odot u_{i_k}^{\odot M(i_k)},$$

for some unique family of scalars $\lambda_M \in K$, all zero except for a finite number.

This looks like a homogeneous polynomial of total degree n, where the monomials of total degree n are the symmetric tensors

$$u_{i_1}^{\odot M(i_1)} \odot \cdots \odot u_{i_k}^{\odot M(i_k)},$$

in the "indeterminates" u_i, where $i \in I$ (recall that $M(i_1) + \cdots + M(i_k) = n$). Again, this is not a coincidence. Polynomials can be defined in terms of symmetric tensors.

22.8 Some Useful Isomorphisms for Symmetric Powers

We can show the following property of the symmetric tensor product, using the proof technique of Proposition 22.7:

$$\mathrm{Sym}^n(E \oplus F) \cong \bigoplus_{k=0}^{n} \mathrm{Sym}^k(E) \otimes \mathrm{Sym}^{n-k}(F).$$

22.9 Duality for Symmetric Powers

In this section, all vector spaces are assumed to have finite dimension. We define a nondegenerate pairing, $\mathrm{Sym}^n(E^*) \times \mathrm{Sym}^n(E) \longrightarrow K$, as follows: Consider the multilinear map,

$$(E^*)^n \times E^n \longrightarrow K,$$

given by

$$(v_1^*, \ldots, v_n^*, u_1, \ldots, u_n) \mapsto \sum_{\sigma\in\mathfrak{S}_n} v_{\sigma(1)}^*(u_1) \cdots v_{\sigma(n)}^*(u_n).$$

Note that the expression on the right-hand side is "almost" the determinant, $\det(v_j^*(u_i))$, except that the sign $\mathrm{sgn}(\sigma)$ is missing (where $\mathrm{sgn}(\sigma)$ is the signature of the permutation σ, that is, the parity of the number of transpositions into which σ can be factored). Such an expression is called a *permanent*. It is easily checked that this expression is symmetric

w.r.t. the u_i's and also w.r.t. the v_j^*. For any fixed $(v_1^*, \ldots, v_n^*) \in (E^*)^n$, we get a symmetric multilinear map,

$$l_{v_1^*,\ldots,v_n^*} \colon (u_1, \ldots, u_n) \mapsto \sum_{\sigma \in \mathfrak{S}_n} v_{\sigma(1)}^*(u_1) \cdots v_{\sigma(n)}^*(u_n),$$

from E^n to K. The map $l_{v_1^*,\ldots,v_n^*}$ extends uniquely to a linear map, $L_{v_1^*,\ldots,v_n^*} \colon \mathrm{Sym}^n(E) \to K$. Now, we also have the symmetric multilinear map,

$$(v_1^*, \ldots, v_n^*) \mapsto L_{v_1^*,\ldots,v_n^*},$$

from $(E^*)^n$ to $\mathrm{Hom}(\mathrm{Sym}^n(E), K)$, which extends to a linear map, L, from $\mathrm{Sym}^n(E^*)$ to $\mathrm{Hom}(\mathrm{Sym}^n(E), K)$. However, in view of the isomorphism,

$$\mathrm{Hom}(U \otimes V, W) \cong \mathrm{Hom}(U, \mathrm{Hom}(V, W)),$$

we can view L as a linear map,

$$L \colon \mathrm{Sym}^n(E^*) \otimes \mathrm{Sym}^n(E) \longrightarrow K,$$

which corresponds to a bilinear map,

$$\mathrm{Sym}^n(E^*) \times \mathrm{Sym}^n(E) \longrightarrow K.$$

Now, this pairing in nondegenerate. This can be done using bases and we leave it as an exercise to the reader (see Knapp [90], Appendix A). Therefore, we get a canonical isomorphism,

$$(\mathrm{Sym}^n(E))^* \cong \mathrm{Sym}^n(E^*).$$

Since we also have an isomorphism

$$(\mathrm{Sym}^n(E))^* \cong \mathrm{S}^n(E, K),$$

we get a canonical isomorphism

$$\mathrm{Sym}^n(E^*) \cong \mathrm{S}^n(E, K)$$

which allows us to interpret symmetric tensors over E^* as symmetric multilinear maps.

Remark: The isomorphism, $\mu \colon \mathrm{Sym}^n(E^*) \cong \mathrm{S}^n(E, K)$, discussed above can be described explicity as the linear extension of the map given by

$$\mu(v_1^* \odot \cdots \odot v_n^*)(u_1, \ldots, u_n) = \sum_{\sigma \in \mathfrak{S}_n} v_{\sigma(1)}^*(u_1) \cdots v_{\sigma(n)}^*(u_n).$$

Now, the map from E^n to $\mathrm{Sym}^n(E)$ given by $(u_1, \ldots, u_n) \mapsto u_1 \odot \cdots \odot u_n$ yields a surjection, $\pi \colon E^{\otimes n} \to \mathrm{Sym}^n(E)$. Because we are dealing with vector spaces, this map has some section, that is, there is some injection, $\iota \colon \mathrm{Sym}^n(E) \to E^{\otimes n}$, with $\pi \circ \iota = \mathrm{id}$. If our field,

K, has characteristic 0, then there is a special section having a natural definition involving a symmetrization process defined as follows: For every permutation, σ, we have the map, $r_\sigma\colon E^n \to E^{\otimes n}$, given by

$$r_\sigma(u_1, \ldots, u_n) = u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(n)}.$$

As r_σ is clearly multilinear, r_σ extends to a linear map, $r_\sigma\colon E^{\otimes n} \to E^{\otimes n}$, and we get a map, $\mathfrak{S}_n \times E^{\otimes n} \longrightarrow E^{\otimes n}$, namely,

$$\sigma \cdot z = r_\sigma(z).$$

It is immediately checked that this is a left action of the symmetric group, $\mathfrak{S}_n$, on $E^{\otimes n}$ and the tensors $z \in E^{\otimes n}$ such that

$$\sigma \cdot z = z, \quad \text{for all} \quad \sigma \in \mathfrak{S}_n$$

are called *symmetrized* tensors. We define the map, $\iota\colon E^n \to E^{\otimes n}$, by

$$\iota(u_1, \ldots, u_n) = \frac{1}{n!} \sum_{\sigma \in \mathfrak{S}_n} \sigma \cdot (u_1 \otimes \cdots \otimes u_n) = \frac{1}{n!} \sum_{\sigma \in \mathfrak{S}_n} u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(n)}.$$

As the right hand side is clearly symmetric, we get a linear map, $\iota\colon \mathrm{Sym}^n(E) \to E^{\otimes n}$. Clearly, $\iota(\mathrm{Sym}^n(E))$ is the set of symmetrized tensors in $E^{\otimes n}$. If we consider the map, $S = \iota \circ \pi\colon E^{\otimes n} \longrightarrow E^{\otimes n}$, it is easy to check that $S \circ S = S$. Therefore, S is a projection and by linear algebra, we know that

$$E^{\otimes n} = S(E^{\otimes n}) \oplus \operatorname{Ker} S = \iota(\mathrm{Sym}^n(E)) \oplus \operatorname{Ker} S.$$

It turns out that $\operatorname{Ker} S = E^{\otimes n} \cap \mathfrak{I} = \operatorname{Ker} \pi$, where $\mathfrak{I}$ is the two-sided ideal of $T(E)$ generated by all tensors of the form $u \otimes v - v \otimes u \in E^{\otimes 2}$ (for example, see Knapp [90], Appendix A). Therefore, ι is injective,

$$E^{\otimes n} = \iota(\mathrm{Sym}^n(E)) \oplus E^{\otimes n} \cap \mathfrak{I} = \iota(\mathrm{Sym}^n(E)) \oplus \operatorname{Ker} \pi,$$

and the symmetric tensor power, $\mathrm{Sym}^n(E)$, is naturally embedded into $E^{\otimes n}$.

22.10 Symmetric Algebras

As in the case of tensors, we can pack together all the symmetric powers, $\mathrm{Sym}^n(V)$, into an algebra,

$$\mathrm{Sym}(V) = \bigoplus_{m \geq 0} \mathrm{Sym}^m(V),$$

called the *symmetric tensor algebra of* V. We could adapt what we did in Section 22.5 for general tensor powers to symmetric tensors but since we already have the algebra, $T(V)$,

we can proceed faster. If $\mathfrak{I}$ is the two-sided ideal generated by all tensors of the form $u \otimes v - v \otimes u \in V^{\otimes 2}$, we set

$$\mathrm{Sym}^\bullet(V) = T(V)/\mathfrak{I}.$$

Then, $\mathrm{Sym}^\bullet(V)$ automatically inherits a multiplication operation which is commutative and since $T(V)$ is graded, that is,

$$T(V) = \bigoplus_{m \geq 0} V^{\otimes m},$$

we have

$$\mathrm{Sym}^\bullet(V) = \bigoplus_{m \geq 0} V^{\otimes m}/(\mathfrak{I} \cap V^{\otimes m}).$$

However, it is easy to check that

$$\mathrm{Sym}^m(V) \cong V^{\otimes m}/(\mathfrak{I} \cap V^{\otimes m}),$$

so

$$\mathrm{Sym}^\bullet(V) \cong \mathrm{Sym}(V).$$

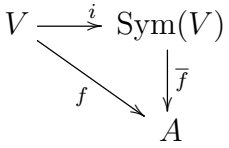
When V is of finite dimension, n, $T(V)$ corresponds to the algebra of polynomials with coefficients in K in n variables (this can be seen from Proposition 22.13). When V is of infinite dimension and $(u_i)_{i \in I}$ is a basis of V, the algebra, $\mathrm{Sym}(V)$, corresponds to the algebra of polynomials in infinitely many variables in I. What's nice about the symmetric tensor algebra, $\mathrm{Sym}(V)$, is that it provides an intrinsic definition of a polynomial algebra in any set, I, of variables.

It is also easy to see that $\mathrm{Sym}(V)$ satisfies the following universal mapping property:

Proposition 22.14. *Given any commutative K-algebra, A, for any linear map, $f\colon V \to A$, there is a unique K-algebra homomorphism, $\overline{f}\colon \mathrm{Sym}(V) \to A$, so that*

$$f = \overline{f} \circ i,$$

as in the diagram below:



Remark: If E is finite-dimensional, recall the isomorphism, $\mu\colon \mathrm{Sym}^n(E^*) \longrightarrow \mathrm{S}^n(E, K)$, defined as the linear extension of the map given by

$$\mu(v_1^* \odot \cdots \odot v_n^*)(u_1, \ldots, u_n) = \sum_{\sigma \in \mathfrak{S}_n} v_{\sigma(1)}^*(u_1) \cdots v_{\sigma(n)}^*(u_n),$$

Now, we have also a multiplication operation, $\mathrm{Sym}^m(E^*) \times \mathrm{Sym}^n(E^*) \longrightarrow \mathrm{Sym}^{m+n}(E^*)$. The following question then arises:

Can we define a multiplication, $\mathrm{S}^m(E,K) \times \mathrm{S}^n(E,K) \longrightarrow \mathrm{S}^{m+n}(E,K)$, directly on symmetric multilinear forms, so that the following diagram commutes:

$$\begin{array}{ccc} \mathrm{Sym}^m(E^*) \times \mathrm{Sym}^n(E^*) & \xrightarrow{\odot} & \mathrm{Sym}^{m+n}(E^*) \\ \big\downarrow{\scriptstyle \mu\times\mu} & & \big\downarrow{\scriptstyle \mu} \\ \mathrm{S}^m(E,K) \times \mathrm{S}^n(E,K) & \xrightarrow{\cdot} & \mathrm{S}^{m+n}(E,K). \end{array}$$

The answer is *yes*! The solution is to define this multiplication such that, for $f \in \mathrm{S}^m(E,K)$ and $g \in \mathrm{S}^n(E,K)$,

$$(f \cdot g)(u_1, \ldots, u_{m+n}) = \sum_{\sigma \in \mathrm{shuffle}(m,n)} f(u_{\sigma(1)}, \ldots, u_{\sigma(m)}) g(u_{\sigma(m+1)}, \ldots, u_{\sigma(m+n)}),$$

where $\mathrm{shuffle}(m,n)$ consists of all (m,n)-"shuffles", that is, permutations, σ, of $\{1, \ldots m+n\}$, such that $\sigma(1) < \cdots < \sigma(m)$ and $\sigma(m+1) < \cdots < \sigma(m+n)$. We urge the reader to check this fact.

Another useful canonical isomorphim (of K-algebras) is

$$\mathrm{Sym}(E \oplus F) \cong \mathrm{Sym}(E) \otimes \mathrm{Sym}(F).$$

22.11 Exterior Tensor Powers

We now consider *alternating* (also called *skew-symmetric*) multilinear maps and *exterior tensor powers* (also called *alternating tensor powers*), denoted $\bigwedge^n(E)$. In many respect, alternating multilinear maps and exterior tensor powers can be treated much like symmetric tensor powers except that the sign, $\mathrm{sgn}(\sigma)$, needs to be inserted in front of the formulae valid for symmetric powers. Roughly speaking, we are now in the world of determinants rather than in the world of permanents. However, there are also some fundamental differences, one of which being that the exterior tensor power, $\bigwedge^n(E)$, is the trivial vector space, (0), when E is finite-dimensional and when $n > \dim(E)$. As in the case of symmetric tensor powers, since we already have the tensor algebra, $T(V)$, we can proceed rather quickly. But first, let us review some basic definitions and facts.

Definition 22.7. Let $f\colon E^n \to F$ be a multilinear map. We say that f *alternating* iff $f(u_1, \ldots, u_n) = 0$ whenever $u_i = u_{i+1}$, for some i with $1 \leq i \leq n-1$, for all $u_i \in E$, that is, $f(u_1, \ldots, u_n) = 0$ whenever two adjacent arguments are identical. We say that f is *skew-symmetric* (or *anti-symmetric*) iff

$$f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}) = \mathrm{sgn}(\sigma) f(u_1, \ldots, u_n),$$

for every permutation, $\sigma \in \mathfrak{S}_n$, and all $u_i \in E$.

For $n = 1$, we agree that every linear map, $f\colon E \to F$, is alternating. The vector space of all multilinear alternating maps, $f\colon E^n \to F$, is denoted $\mathrm{Alt}^n(E;F)$. Note that $\mathrm{Alt}^1(E;F) = \mathrm{Hom}(E,F)$. The following basic proposition shows the relationship between alternation and skew-symmetry.

Proposition 22.15. *Let $f\colon E^n \to F$ be a multilinear map. If f is alternating, then the following properties hold:*

(1) For all i, with $1 \leq i \leq n-1$,

$$f(\ldots, u_i, u_{i+1}, \ldots) = -f(\ldots, u_{i+1}, u_i, \ldots).$$

(2) For every permutation, $\sigma \in \mathfrak{S}_n$,

$$f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}) = \mathrm{sgn}(\sigma) f(u_1, \ldots, u_n).$$

(3) For all i, j, with $1 \leq i < j \leq n$,

$$f(\ldots, u_i, \ldots u_j, \ldots) = 0 \quad \textit{whenever } u_i = u_j.$$

Moreover, if our field, K, has characteristic different from 2, then every skew-symmetric multilinear map is alternating.

Proof. (i) By multilinearity applied twice, we have

$$\begin{aligned} f(\ldots, u_i + u_{i+1}, u_i + u_{i+1}, \ldots) = {} & f(\ldots, u_i, u_i, \ldots) + f(\ldots, u_i, u_{i+1}, \ldots) \\ & + f(\ldots, u_{i+1}, u_i, \ldots) + f(\ldots, u_{i+1}, u_{i+1}, \ldots). \end{aligned}$$

Since f is alternating, we get

$$0 = f(\ldots, u_i, u_{i+1}, \ldots) + f(\ldots, u_{i+1}, u_i, \ldots),$$

that is, $f(\ldots, u_i, u_{i+1}, \ldots) = -f(\ldots, u_{i+1}, u_i, \ldots)$.

(ii) Clearly, the symmetric group, $\mathfrak{S}_n$, acts on $\mathrm{Alt}^n(E;F)$ on the left, *via*

$$\sigma \cdot f(u_1, \ldots, u_n) \ = f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}).$$

Consequently, as $\mathfrak{S}_n$ is generated by the transpositions (permutations that swap exactly two elements), since for a transposition, (ii) is simply (i), we deduce (ii) by induction on the number of transpositions in σ.

(iii) There is a permutation, σ, that sends u_i and u_j respectively to u_1 and u_2. As f is alternating,

$$f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}) = 0.$$

However, by (ii),

$$f(u_1, \ldots, u_n) = \mathrm{sgn}(\sigma) f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}) = 0.$$

Now, when f is skew-symmetric, if σ is the transposition swapping u_i and $u_{i+1} = u_i$, as $\mathrm{sgn}(\sigma) = -1$, we get

$$f(\ldots, u_i, u_i, \ldots) = -f(\ldots, u_i, u_i, \ldots),$$

so that

$$2f(\ldots, u_i, u_i, \ldots) = 0,$$

and in every characteristic except 2, we conclude that $f(\ldots, u_i, u_i, \ldots) = 0$, namely, f is alternating. □

Proposition 22.15 shows that in every characteristic except 2, alternating and skew-symmetric multilinear maps are identical. Using Proposition 22.15 we easily deduce the following crucial fact:

Proposition 22.16. *Let $f\colon E^n \to F$ be an alternating multilinear map. For any families of vectors, $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$, with $u_i, v_i \in E$, if*

$$v_j = \sum_{i=1}^{n} a_{ij} u_i, \qquad 1 \leq j \leq n,$$

then

$$f(v_1, \ldots, v_n) = \left(\sum_{\sigma \in \mathfrak{S}_n} \mathrm{sgn}(\sigma)\, a_{\sigma(1),1} \cdots a_{\sigma(n),n} \right) f(u_1, \ldots, u_n) = \det(A) f(u_1, \ldots, u_n),$$

where A is the $n \times n$ matrix, $A = (a_{ij})$.

Proof. Use property (ii) of Proposition 22.15. □

We are now ready to define and construct exterior tensor powers.

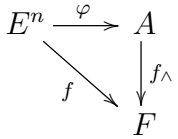
Definition 22.8. An *n-th exterior tensor power* of a vector space, E, where $n \geq 1$, is a vector space, A, together with an alternating multilinear map, $\varphi\colon E^n \to A$, such that, for every vector space, F, and for every alternating multilinear map, $f\colon E^n \to F$, there is a unique linear map, $f_\wedge\colon A \to F$, with

$$f(u_1, \ldots, u_n) = f_\wedge(\varphi(u_1, \ldots, u_n)),$$

for all $u_1, \ldots, u_n \in E$, or for short

$$f = f_\wedge \circ \varphi.$$

Equivalently, there is a unique linear map $f_\wedge$ such that the following diagram commutes:



First, we show that any two n-th exterior tensor powers (A_1, φ_1) and (A_2, φ_2) for E, are isomorphic.

Proposition 22.17. *Given any two n-th exterior tensor powers (A_1, φ_1) and (A_2, φ_2) for E, there is an isomorphism $h\colon A_1 \to A_2$ such that*

$$\varphi_2 = h \circ \varphi_1.$$

Proof. Replace tensor product by n exterior tensor power in the proof of Proposition 22.4. □

We now give a construction that produces an n-th exterior tensor power of a vector space E.

Theorem 22.18. *Given a vector space E, an n-th exterior tensor power $(\bigwedge^n(E), \varphi)$ for E can be constructed ($n \geq 1$). Furthermore, denoting $\varphi(u_1, \ldots, u_n)$ as $u_1 \wedge \cdots \wedge u_n$, the exterior tensor power $\bigwedge^n(E)$ is generated by the vectors $u_1 \wedge \cdots \wedge u_n$, where $u_1, \ldots, u_n \in E$, and for every alternating multilinear map $f\colon E^n \to F$, the unique linear map $f_\wedge\colon \bigwedge^n(E) \to F$ such that $f = f_\wedge \circ \varphi$, is defined by*

$$f_\wedge(u_1 \wedge \cdots \wedge u_n) = f(u_1, \ldots, u_n),$$

on the generators $u_1 \wedge \cdots \wedge u_n$ of $\bigwedge^n(E)$.

Proof sketch. We can give a quick proof using the tensor algebra, $T(E)$. let $\mathfrak{I}_a$ be the two-sided ideal of $T(E)$ generated by all tensors of the form $u \otimes u \in E^{\otimes 2}$. Then, let

$$\bigwedge^n(E) = E^{\otimes n}/(\mathfrak{I}_a \cap E^{\otimes n})$$

and let π be the projection, $\pi\colon E^{\otimes n} \to \bigwedge^n(E)$. If we let $u_1 \wedge \cdots \wedge u_n = \pi(u_1 \otimes \cdots \otimes u_n)$, it is easy to check that $(\bigwedge^n(E), \wedge)$ satisfies the conditions of Theorem 22.18. □

Remark: We can also define

$$\bigwedge(E) = T(E)/\mathfrak{I}_a = \bigoplus_{n \geq 0} \bigwedge^n(E),$$

the *exterior algebra* of E. This is the skew-symmetric counterpart of $\mathrm{Sym}(E)$ and we will study it a little later.

For simplicity of notation, we may write $\bigwedge^n E$ for $\bigwedge^n(E)$. We also abbreviate "exterior tensor power" as "exterior power". Clearly, $\bigwedge^1(E) \cong E$ and it is convenient to set $\bigwedge^0(E) = K$.

The fact that the map $\varphi\colon E^n \to \bigwedge^n(E)$ is alternating and multilinear, can also be expressed as follows:

$$\begin{aligned} u_1 \wedge \cdots \wedge (u_i + v_i) \wedge \cdots \wedge u_n &= (u_1 \wedge \cdots \wedge u_i \wedge \cdots \wedge u_n) \\ &\quad + (u_1 \wedge \cdots \wedge v_i \wedge \cdots \wedge u_n), \\ u_1 \wedge \cdots \wedge (\lambda u_i) \wedge \cdots \wedge u_n &= \lambda(u_1 \wedge \cdots \wedge u_i \wedge \cdots \wedge u_n), \\ u_{\sigma(1)} \wedge \cdots \wedge u_{\sigma(n)} &= \operatorname{sgn}(\sigma)\, u_1 \wedge \cdots \wedge u_n, \end{aligned}$$

for all $\sigma \in \mathfrak{S}_n$.

Theorem 22.18 yields a canonical isomorphism

$$\operatorname{Hom}(\bigwedge^n(E), F) \cong \operatorname{Alt}^n(E; F),$$

between the vector space of linear maps $\operatorname{Hom}(\bigwedge^n(E), F)$, and the vector space of alternating multilinear maps $\operatorname{Alt}^n(E; F)$, via the linear map $- \circ \varphi$ defined by

$$h \mapsto h \circ \varphi,$$

where $h \in \operatorname{Hom}(\bigwedge^n(E), F)$. In particular, when $F = K$, we get a canonical isomorphism

$$\left(\bigwedge^n(E)\right)^* \cong \operatorname{Alt}^n(E; K).$$

Tensors $\alpha \in \bigwedge^n(E)$ are called *alternating n-tensors* or *alternating tensors of degree n* and we write $\deg(\alpha) = n$. Tensors of the form $u_1 \wedge \cdots \wedge u_n$, where $u_i \in E$, are called *simple (or decomposable) alternating n-tensors*. Those alternating n-tensors that are not simple are often called *compound alternating n-tensors*. Simple tensors $u_1 \wedge \cdots \wedge u_n \in \bigwedge^n(E)$ are also called *n-vectors* and tensors in $\bigwedge^n(E^*)$ are often called *(alternating) n-forms*.

Given two linear maps $f\colon E \to E'$ and $g\colon E \to E'$, we can define $h\colon E \times E \to \bigwedge^2(E')$ by

$$h(u, v) = f(u) \wedge g(v).$$

It is immediately verified that h is alternating bilinear, and thus, it induces a unique linear map

$$f \wedge g\colon \bigwedge^2(E) \to \bigwedge^2(E'),$$

such that

$$(f \wedge g)(u \wedge v) = f(u) \wedge g(u).$$

If we also have linear maps $f'\colon E' \to E''$ and $g'\colon E' \to E''$, we can easily verify that

$$(f' \circ f) \wedge (g' \circ g) = (f' \wedge g') \circ (f \wedge g).$$

The generalization to the alternating product $f_1 \wedge \cdots \wedge f_n$ of $n \geq 3$ linear maps $f_i\colon E \to E'$ is immediate, and left to the reader.

22.12 Bases of Exterior Powers

Let E be any vector space. For any basis, $(u_i)_{i\in\Sigma}$, for E, we assume that some total ordering, $\leq$, on Σ, has been chosen. Call the pair $((u_i)_{i\in\Sigma}, \leq)$ an *ordered basis*. Then, for any nonempty finite subset, $I \subseteq \Sigma$, let

$$u_I = u_{i_1} \wedge \cdots \wedge u_{i_m},$$

where $I = \{i_1, \ldots, i_m\}$, with $i_1 < \cdots < i_m$.

Since $\bigwedge^n(E)$ is generated by the tensors of the form $v_1 \wedge \cdots \wedge v_n$, with $v_i \in E$, in view of skew-symmetry, it is clear that the tensors u_I, with $|I| = n$, generate $\bigwedge^n(E)$. Actually, they form a basis.

Proposition 22.19. *Given any vector space, E, if E has finite dimension, $d = \dim(E)$, then for all $n > d$, the exterior power $\bigwedge^n(E)$ is trivial, that is $\bigwedge^n(E) = (0)$. Otherwise, for every ordered basis, $((u_i)_{i\in\Sigma}, \leq)$, the family, (u_I), is basis of $\bigwedge^n(E)$, where I ranges over finite nonempty subsets of Σ of size $|I| = n$.*

Proof. First, assume that E has finite dimension, $d = \dim(E)$ and that $n > d$. We know that $\bigwedge^n(E)$ is generated by the tensors of the form $v_1 \wedge \cdots \wedge v_n$, with $v_i \in E$. If $u_1, \ldots, u_d$ is a basis of E, as every v_i is a linear combination of the u_j, when we expand $v_1 \wedge \cdots \wedge v_n$ using multilinearity, we get a linear combination of the form

$$v_1 \wedge \cdots \wedge v_n = \sum_{(j_1,\ldots,j_n)} \lambda_{(j_1,\ldots,j_n)}\, u_{j_1} \wedge \cdots \wedge u_{j_n},$$

where each $(j_1, \ldots, j_n)$ is some sequence of integers $j_k \in \{1, \ldots, d\}$. As $n > d$, each sequence $(j_1, \ldots, j_n)$ must contain two identical elements. By alternation, $u_{j_1} \wedge \cdots \wedge u_{j_n} = 0$ and so, $v_1 \wedge \cdots \wedge v_n = 0$. It follows that $\bigwedge^n(E) = (0)$.

Now, assume that either $\dim(E) = d$ and that $n \leq d$ or that E is infinite dimensional. The argument below shows that the u_I are nonzero and linearly independent. As usual, let $u_i^* \in E^*$ be the linear form given by

$$u_i^*(u_j) = \delta_{ij}.$$

For any nonempty subset, $I = \{i_1, \ldots, i_n\} \subseteq \Sigma$, with $i_1 < \cdots < i_n$, let l_I be the map given by

$$l_I(v_1, \ldots, v_n) = \det(u_{i_j}^*(v_k)),$$

for all $v_k \in E$. As l_I is alternating multilinear, it induces a linear map, $L_I \colon \bigwedge^n(E) \to K$. Observe that for any nonempty finite subset, $J \subseteq \Sigma$, with $|J| = n$, we have

$$L_I(u_J) = \begin{cases} 1 & \text{if } I = J \\ 0 & \text{if } I \neq J. \end{cases}$$

Note that when $\dim(E) = d$ and $n \leq d$, the forms $u_{i_1}^*, \ldots, u_{i_n}^*$ are all distinct so, the above does hold. Since $L_I(u_I) = 1$, we conclude that $u_I \neq 0$. Now, if we have a linear combination,

$$\sum_I \lambda_I u_I = 0,$$

where the above sum is finite and involves nonempty finite subset, $I \subseteq \Sigma$, with $|I| = n$, for every such I, when we apply L_I we get

$$\lambda_I = 0,$$

proving linear independence. □

As a corollary, if E is finite dimensional, say $\dim(E) = d$ and if $1 \leq n \leq d$, then we have

$$\dim(\bigwedge^n(E)) = \binom{n}{d}$$

and if $n > d$, then $\dim(\bigwedge^n(E)) = 0$.

Remark: When $n = 0$, if we set $u_\emptyset = 1$, then $(u_\emptyset) = (1)$ is a basis of $\bigwedge^0(V) = K$.

It follows from Proposition 22.19 that the family, $(u_I)_I$, where $I \subseteq \Sigma$ ranges over finite subsets of Σ is a basis of $\bigwedge(V) = \bigoplus_{n \geq 0} \bigwedge^n(V)$.

As a corollary of Proposition 22.19 we obtain the following useful criterion for linear independence:

Proposition 22.20. *For any vector space, E, the vectors, $u_1, \ldots, u_n \in E$, are linearly independent iff $u_1 \wedge \cdots \wedge u_n \neq 0$.*

Proof. If $u_1 \wedge \cdots \wedge u_n \neq 0$, then $u_1, \ldots, u_n$ must be linearly independent. Otherwise, some u_i would be a linear combination of the other u_j's (with $j \neq i$) and then, as in the proof of Proposition 22.19, $u_1 \wedge \cdots \wedge u_n$ would be a linear combination of wedges in which two vectors are identical and thus, zero.

Conversely, assume that $u_1, \ldots, u_n$ are linearly independent. Then, we have the linear forms, $u_i^* \in E^*$, such that

$$u_i^*(u_j) = \delta_{i,j} \qquad 1 \leq i, j \leq n.$$

As in the proof of Proposition 22.19, we have a linear map, $L_{u_1,\ldots,u_n} \colon \bigwedge^n(E) \to K$, given by

$$L_{u_1,\ldots,u_n}(v_1 \wedge \cdots \wedge v_n) = \det(u_j^*(v_i)),$$

for all $v_1 \wedge \cdots \wedge v_n \in \bigwedge^n(E)$. As,

$$L_{u_1,\ldots,u_n}(u_1 \wedge \cdots \wedge u_n) = 1,$$

we conclude that $u_1 \wedge \cdots \wedge u_n \neq 0$. □

Proposition 22.20 shows that, geometrically, every nonzero wedge, $u_1 \wedge \cdots \wedge u_n$, corresponds to some oriented version of an n-dimensional subspace of E.

22.13 Some Useful Isomorphisms for Exterior Powers

We can show the following property of the exterior tensor product, using the proof technique of Proposition 22.7:

$$\bigwedge^n(E \oplus F) \cong \bigoplus_{k=0}^{n} \bigwedge^k(E) \otimes \bigwedge^{n-k}(F).$$

22.14 Duality for Exterior Powers

In this section, all vector spaces are assumed to have finite dimension. We define a nondegenerate pairing, $\bigwedge^n(E^*) \times \bigwedge^n(E) \longrightarrow K$, as follows: Consider the multilinear map,

$$(E^*)^n \times E^n \longrightarrow K,$$

given by

$$(v_1^*, \ldots, v_n^*, u_1, \ldots, u_n) \mapsto \sum_{\sigma \in \mathfrak{S}_n} \operatorname{sgn}(\sigma)\, v_{\sigma(1)}^*(u_1) \cdots v_{\sigma(n)}^*(u_n) = \det(v_j^*(u_i)).$$

It is easily checked that this expression is alternating w.r.t. the u_i's and also w.r.t. the v_j^*. For any fixed $(v_1^*, \ldots, v_n^*) \in (E^*)^n$, we get an alternating multinear map,

$$l_{v_1^*,\ldots,v_n^*} \colon (u_1, \ldots, u_n) \mapsto \det(v_j^*(u_i)),$$

from E^n to K. By the argument used in the symmetric case, we get a bilinear map,

$$\bigwedge^n(E^*) \times \bigwedge^n(E) \longrightarrow K.$$

Now, this pairing in nondegenerate. This can be done using bases and we leave it as an exercise to the reader. Therefore, we get a canonical isomorphism,

$$(\bigwedge^n(E))^* \cong \bigwedge^n(E^*).$$

Since we also have a canonical isomorphism

$$(\bigwedge^n(E))^* \cong \operatorname{Alt}^n(E; K),$$

we get a canonical isomorphism

$$\bigwedge^n(E^*) \cong \operatorname{Alt}^n(E; K)$$

which allows us to interpret alternating tensors over E^* as alternating multilinear maps.

The isomorphism, $\mu\colon \bigwedge^n(E^*) \cong \mathrm{Alt}^n(E;K)$, discussed above can be described explicity as the linear extension of the map given by

$$\mu(v_1^* \wedge \cdots \wedge v_n^*)(u_1,\ldots,u_n) = \det(v_j^*(u_i)).$$

Remark: Variants of our isomorphism, μ, are found in the literature. For example, there is a version, μ', where

$$\mu' = \frac{1}{n!}\mu,$$

with the factor $\frac{1}{n!}$ added in front of the determinant. Each version has its its own merits and inconvenients. Morita [115] uses μ' because it is more convenient than μ when dealing with characteristic classes. On the other hand, when using μ', some extra factor is needed in defining the wedge operation of alternating multilinear forms (see Section 22.15) and for exterior differentiation. The version μ is the one adopted by Warner [148], Knapp [90], Fulton and Harris [58] and Cartan [29, 30].

If $f\colon E \to F$ is any linear map, by transposition we get a linear map, $f^\top\colon F^* \to E^*$, given by

$$f^\top(v^*) = v^* \circ f, \qquad v^* \in F^*.$$

Consequently, we have

$$f^\top(v^*)(u) = v^*(f(u)), \qquad \text{for all } u \in E \text{ and all } v^* \in F^*.$$

For any $p \geq 1$, the map,

$$(u_1,\ldots,u_p) \mapsto f(u_1) \wedge \cdots \wedge f(u_p),$$

from E^n to $\bigwedge^p F$ is multilinear alternating, so it induces a linear map, $\bigwedge^p f\colon \bigwedge^p E \to \bigwedge^p F$, defined on generators by

$$\Big(\bigwedge^p f\Big)(u_1 \wedge \cdots \wedge u_p) = f(u_1) \wedge \cdots \wedge f(u_p).$$

Combining $\bigwedge^p$ and duality, we get a linear map, $\bigwedge^p f^\top\colon \bigwedge^p F^* \to \bigwedge^p E^*$, defined on generators by

$$\Big(\bigwedge^p f^\top\Big)(v_1^* \wedge \cdots \wedge v_p^*) = f^\top(v_1^*) \wedge \cdots \wedge f^\top(v_p^*).$$

Proposition 22.21. *If $f\colon E \to F$ is any linear map between two finite-dimensional vector spaces, E and F, then*

$$\mu\Big(\Big(\bigwedge^p f^\top\Big)(\omega)\Big)(u_1,\ldots,u_p) = \mu(\omega)(f(u_1),\ldots,f(u_p)), \qquad \omega \in \bigwedge^p F^*,\ u_1,\ldots,u_p \in E.$$

Proof. It is enough to prove the formula on generators. By definition of μ, we have

$$\begin{aligned}\mu\Big(\Big(\bigwedge^p f^\top\Big)(v_1^* \wedge \cdots \wedge v_p^*)\Big)(u_1, \ldots, u_p) &= \mu(f^\top(v_1^*) \wedge \cdots \wedge f^\top(v_p^*))(u_1, \ldots, u_p)\\ &= \det(f^\top(v_j^*)(u_i))\\ &= \det(v_j^*(f(u_i)))\\ &= \mu(v_1^* \wedge \cdots \wedge v_p^*)(f(u_1), \ldots, f(u_p)),\end{aligned}$$

as claimed. □

The map $\bigwedge^p f^\top$ is often denoted f^*, although this is an ambiguous notation since p is dropped. Proposition 22.21 gives us the behavior of f^* under the identification of $\bigwedge^p E^*$ and $\mathrm{Alt}^p(E; K)$ *via* the isomorphism μ.

As in the case of symmetric powers, the map from E^n to $\bigwedge^n(E)$ given by $(u_1, \ldots, u_n) \mapsto u_1 \wedge \cdots \wedge u_n$ yields a surjection, $\pi\colon E^{\otimes n} \to \bigwedge^n(E)$. Now, this map has some section so there is some injection, $\iota\colon \bigwedge^n(E) \to E^{\otimes n}$, with $\pi \circ \iota = \mathrm{id}$. If our field, K, has characteristic 0, then there is a special section having a natural definition involving an antisymmetrization process.

Recall that we have a left action of the symmetric group, $\mathfrak{S}_n$, on $E^{\otimes n}$. The tensors, $z \in E^{\otimes n}$, such that

$$\sigma \cdot z = \mathrm{sgn}(\sigma)\, z, \quad \text{for all} \quad \sigma \in \mathfrak{S}_n$$

are called *antisymmetrized* tensors. We define the map, $\iota\colon E^n \to E^{\otimes n}$, by

$$\iota(u_1, \ldots, u_n) = \frac{1}{n!} \sum_{\sigma \in \mathfrak{S}_n} \mathrm{sgn}(\sigma)\, u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(n)}.$$

As the right hand side is clearly an alternating map, we get a linear map, $\iota\colon \bigwedge^n(E) \to E^{\otimes n}$. Clearly, $\iota(\bigwedge^n(E))$ is the set of antisymmetrized tensors in $E^{\otimes n}$. If we consider the map, $A = \iota \circ \pi\colon E^{\otimes n} \longrightarrow E^{\otimes n}$, it is easy to check that $A \circ A = A$. Therefore, A is a projection and by linear algebra, we know that

$$E^{\otimes n} = A(E^{\otimes n}) \oplus \mathrm{Ker}\, A = \iota(\bigwedge^n(A)) \oplus \mathrm{Ker}\, A.$$

It turns out that $\mathrm{Ker}\, A = E^{\otimes n} \cap \mathfrak{I}_a = \mathrm{Ker}\ \pi$, where $\mathfrak{I}_a$ is the two-sided ideal of $T(E)$ generated by all tensors of the form $u \otimes u \in E^{\otimes 2}$ (for example, see Knapp [90], Appendix A). Therefore, ι is injective,

$$E^{\otimes n} = \iota(\bigwedge^n(E)) \oplus E^{\otimes n} \cap \mathfrak{I} = \iota(\bigwedge^n(E)) \oplus \mathrm{Ker}\, \pi,$$

and the exterior tensor power, $\bigwedge^n(E)$, is naturally embedded into $E^{\otimes n}$.

22.15 Exterior Algebras

As in the case of symmetric tensors, we can pack together all the exterior powers, $\bigwedge^n(V)$, into an algebra,

$$\bigwedge(V) = \bigoplus_{m\geq 0} \overset{m}{\bigwedge}(V),$$

called the *exterior algebra (or Grassmann algebra) of* V. We mimic the procedure used for symmetric powers. If $\mathfrak{I}_a$ is the two-sided ideal generated by all tensors of the form $u \otimes u \in V^{\otimes 2}$, we set

$$\overset{\bullet}{\bigwedge}(V) = T(V)/\mathfrak{I}_a.$$

Then, $\bigwedge^\bullet(V)$ automatically inherits a multiplication operation, called *wedge product*, and since $T(V)$ is graded, that is,

$$T(V) = \bigoplus_{m\geq 0} V^{\otimes m},$$

we have

$$\overset{\bullet}{\bigwedge}(V) = \bigoplus_{m\geq 0} V^{\otimes m}/(\mathfrak{I}_a \cap V^{\otimes m}).$$

However, it is easy to check that

$$\overset{m}{\bigwedge}(V) \cong V^{\otimes m}/(\mathfrak{I}_a \cap V^{\otimes m}),$$

so

$$\overset{\bullet}{\bigwedge}(V) \cong \bigwedge(V).$$

When V has finite dimension, d, we actually have a finite coproduct

$$\bigwedge(V) = \bigoplus_{m=0}^{d} \overset{m}{\bigwedge}(V),$$

and since each $\bigwedge^m(V)$ has dimension, $\binom{d}{m}$, we deduce that

$$\dim(\bigwedge(V)) = 2^d = 2^{\dim(V)}.$$

The multiplication, $\wedge\colon \bigwedge^m(V) \times \bigwedge^n(V) \to \bigwedge^{m+n}(V)$, is skew-symmetric in the following precise sense:

Proposition 22.22. *For all $\alpha \in \bigwedge^m(V)$ and all $\beta \in \bigwedge^n(V)$, we have*

$$\beta \wedge \alpha = (-1)^{mn} \alpha \wedge \beta.$$

Proof. Since $v \wedge u = -u \wedge v$ for all $u, v \in V$, Proposition 22.22 follows by induction. □

Since $\alpha \wedge \alpha = 0$ for every *simple* tensor, $\alpha = u_1 \wedge \cdots \wedge u_n$, it seems natural to infer that $\alpha \wedge \alpha = 0$ for *every* tensor $\alpha \in \bigwedge(V)$. If we consider the case where $\dim(V) \leq 3$, we can indeed prove the above assertion. However, if $\dim(V) \geq 4$, the above fact is generally false! For example, when $\dim(V) = 4$, if u_1, u_2, u_3, u_4 are a basis for V, for $\alpha = u_1 \wedge u_2 + u_3 \wedge u_4$, we check that

$$\alpha \wedge \alpha = 2u_1 \wedge u_2 \wedge u_3 \wedge u_4,$$

which is nonzero.

The above discussion suggests that it might be useful to know when an alternating tensor is simple, that is, decomposable. It can be shown that for tensors, $\alpha \in \bigwedge^2(V)$, $\alpha \wedge \alpha = 0$ iff α is simple. A general criterion for decomposability can be given in terms of some operations known as *left hook* and *right hook* (also called *interior products*), see Section 22.17.

It is easy to see that $\bigwedge(V)$ satisfies the following universal mapping property:

Proposition 22.23. *Given any K-algebra, A, for any linear map, $f\colon V \to A$, if $(f(v))^2 = 0$ for all $v \in V$, then there is a unique K-algebra homomorphism, $\overline{f}\colon \bigwedge(V) \to A$, so that*

$$f = \overline{f} \circ i,$$

as in the diagram below:

$$\begin{array}{ccc} V & \xrightarrow{i} & \bigwedge(V) \\ & \searrow^{f} & \downarrow{\scriptstyle \overline{f}} \\ & & A \end{array}$$

When E is finite-dimensional, recall the isomorphism, $\mu\colon \bigwedge^n(E^*) \longrightarrow \mathrm{Alt}^n(E; K)$, defined as the linear extension of the map given by

$$\mu(v_1^* \wedge \cdots \wedge v_n^*)(u_1, \ldots, u_n) = \det(u_j^*(u_i)).$$

Now, we have also a multiplication operation, $\bigwedge^m(E^*) \times \bigwedge^n(E^*) \longrightarrow \bigwedge^{m+n}(E^*)$. The following question then arises:

Can we define a multiplication, $\mathrm{Alt}^m(E; K) \times \mathrm{Alt}^n(E; K) \longrightarrow \mathrm{Alt}^{m+n}(E; K)$, directly on alternating multilinear forms, so that the following diagram commutes:

$$\begin{array}{ccc} \bigwedge^m(E^*) \times \bigwedge^n(E^*) & \xrightarrow{\wedge} & \bigwedge^{m+n}(E^*) \\ \downarrow{\scriptstyle \mu \times \mu} & & \downarrow{\scriptstyle \mu} \\ \mathrm{Alt}^m(E; K) \times \mathrm{Alt}^n(E; K) & \xrightarrow{\wedge} & \mathrm{Alt}^{m+n}(E; K). \end{array}$$

As in the symmetric case, the answer is *yes*! The solution is to define this multiplication such that, for $f \in \mathrm{Alt}^m(E;K)$ and $g \in \mathrm{Alt}^n(E;K)$,

$$(f \wedge g)(u_1,\ldots,u_{m+n}) = \sum_{\sigma \in \mathrm{shuffle}(m,n)} \mathrm{sgn}(\sigma)\, f(u_{\sigma(1)},\ldots,u_{\sigma(m)})g(u_{\sigma(m+1)},\ldots,u_{\sigma(m+n)}),$$

where $\mathrm{shuffle}(m,n)$ consists of all (m,n)-"shuffles", that is, permutations, σ, of $\{1,\ldots m+n\}$, such that $\sigma(1) < \cdots < \sigma(m)$ and $\sigma(m+1) < \cdots < \sigma(m+n)$. For example, when $m = n = 1$, we have

$$(f \wedge g)(u,v) = f(u)g(v) - g(u)f(v).$$

When $m = 1$ and $n \geq 2$, check that

$$(f \wedge g)(u_1,\ldots,u_{m+1}) = \sum_{i=1}^{m+1} (-1)^{i-1} f(u_i)g(u_1,\ldots,\widehat{u_i},\ldots,u_{m+1}),$$

where the hat over the argument u_i means that it should be omitted.

As a result of all this, the coproduct

$$\mathrm{Alt}(E) = \bigoplus_{n\geq 0} \mathrm{Alt}^n(E;K)$$

is an algebra under the above multiplication and this algebra is isomorphic to $\bigwedge(E^*)$. For the record, we state

Proposition 22.24. *When E is finite dimensional, the maps, $\mu\colon \bigwedge^n(E^*) \longrightarrow \mathrm{Alt}^n(E;K)$, induced by the linear extensions of the maps given by*

$$\mu(v_1^* \wedge \cdots \wedge v_n^*)(u_1,\ldots,u_n) = \det(u_j^*(u_i))$$

yield a canonical isomorphism of algebras, $\mu\colon \bigwedge(E^*) \longrightarrow \mathrm{Alt}(E)$, where the multiplication in $\mathrm{Alt}(E)$ is defined by the maps, $\wedge\colon \mathrm{Alt}^m(E;K) \times \mathrm{Alt}^n(E;K) \longrightarrow \mathrm{Alt}^{m+n}(E;K)$, with

$$(f \wedge g)(u_1,\ldots,u_{m+n}) = \sum_{\sigma \in \mathrm{shuffle}(m,n)} \mathrm{sgn}(\sigma)\, f(u_{\sigma(1)},\ldots,u_{\sigma(m)})g(u_{\sigma(m+1)},\ldots,u_{\sigma(m+n)}),$$

where $\mathrm{shuffle}(m,n)$ consists of all (m,n)-"shuffles", that is, permutations, σ, of $\{1,\ldots m+n\}$, such that $\sigma(1) < \cdots < \sigma(m)$ and $\sigma(m+1) < \cdots < \sigma(m+n)$.

Remark: The algebra, $\bigwedge(E)$ is a graded algebra. Given two graded algebras, E and F, we can make a new tensor product, $E \mathbin{\widehat{\otimes}} F$, where $E \mathbin{\widehat{\otimes}} F$ is equal to $E \otimes F$ as a vector space, but with a skew-commutative multiplication given by

$$(a \otimes b) \wedge (c \otimes d) = (-1)^{\deg(b)\deg(c)}(ac) \otimes (bd),$$

where $a \in E^m, b \in F^p$, $c \in E^n, d \in F^q$. Then, it can be shown that

$$\bigwedge(E \oplus F) \cong \bigwedge(E) \mathbin{\widehat{\otimes}} \bigwedge(F).$$

22.16 The Hodge $*$-Operator

In order to define a generalization of the Laplacian that will apply to differential forms on a Riemannian manifold, we need to define isomorphisms,

$$\bigwedge^k V \longrightarrow \bigwedge^{n-k} V,$$

for any Euclidean vector space, V, of dimension n and any k, with $0 \leq k \leq n$. If $\langle -, - \rangle$ denotes the inner product on V, we define an inner product on $\bigwedge^k V$, also denoted $\langle -, - \rangle$, by setting

$$\langle u_1 \wedge \cdots \wedge u_k, v_1 \wedge \cdots \wedge v_k \rangle = \det(\langle u_i, v_j \rangle),$$

for all $u_i, v_i \in V$ and extending $\langle -, - \rangle$ by bilinearity.

It is easy to show that if $(e_1, \ldots, e_n)$ is an orthonormal basis of V, then the basis of $\bigwedge^k V$ consisting of the e_I (where $I = \{i_1, \ldots, i_k\}$, with $1 \leq i_1 < \cdots < i_k \leq n$) is an orthonormal basis of $\bigwedge^k V$. Since the inner product on V induces an inner product on V^* (recall that $\langle \omega_1, \omega_2 \rangle = \langle \omega_1^\sharp, \omega_2^\sharp \rangle$, for all $\omega_1, \omega_2 \in V^*$), we also get an inner product on $\bigwedge^k V^*$.

Recall that an *orientation* of a vector space, V, of dimension n is given by the choice of some basis, $(e_1, \ldots, e_n)$. We say that a basis, $(u_1, \ldots, u_n)$, of V is *positively oriented* iff $\det(u_1, \ldots, u_n) > 0$ (where $\det(u_1, \ldots, u_n)$ denotes the determinant of the matrix whose jth column consists of the coordinates of u_j over the basis $(e_1, \ldots, e_n)$), otherwise it is *negatively oriented*. An *oriented vector space* is a vector space, V, together with an orientation of V. If V is oriented by the basis $(e_1, \ldots, e_n)$, then V^* is oriented by the dual basis, $(e_1^*, \ldots, e_n^*)$.

If V is an oriented vector space of dimension n, then we can define a linear map,

$$*\colon \bigwedge^k V \to \bigwedge^{n-k} V,$$

called the *Hodge $*$-operator*, as follows: For any choice of a positively oriented orthonormal basis, $(e_1, \ldots, e_n)$, of V, set

$$*(e_1 \wedge \cdots \wedge e_k) = e_{k+1} \wedge \cdots \wedge e_n.$$

In particular, for $k = 0$ and $k = n$, we have

$$\begin{aligned} *(1) &= e_1 \wedge \cdots \wedge e_n \\ *(e_1 \wedge \cdots \wedge e_n) &= 1. \end{aligned}$$

It is easy to see that the definition of $*$ does not depend on the choice of positively oriented orthonormal basis.

The Hodge $*$-operators, $*\colon \bigwedge^k V \to \bigwedge^{n-k} V$, induces a linear bijection, $*\colon \bigwedge(V) \to \bigwedge(V)$. We also have Hodge $*$-operators, $*\colon \bigwedge^k V^* \to \bigwedge^{n-k} V^*$.

The following proposition is easy to show:

Proposition 22.25. *If V is any oriented vector space of dimension n, for every k, with $0 \leq k \leq n$, we have*

(i) $** = (-\mathrm{id})^{k(n-k)}$.

(ii) $\langle x, y\rangle = *(x \wedge *y) = *(y \wedge *x)$, *for all* $x, y \in \bigwedge^k V$.

If $(e_1, \ldots, e_n)$ is an orthonormal basis of V and $(v_1, \ldots, v_n)$ is any other basis of V, it is easy to see that

$$v_1 \wedge \cdots \wedge v_n = \sqrt{\det(\langle v_i, v_j\rangle)}\, e_1 \wedge \cdots \wedge e_n,$$

from which it follows that

$$*(1) = \frac{1}{\sqrt{\det(\langle v_i, v_j\rangle)}}\, v_1 \wedge \cdots \wedge v_n$$

(see Jost [84], Chapter 2, Lemma 2.1.3).

22.17 Testing Decomposability; Left and Right Hooks

In this section, all vector spaces are assumed to have finite dimension. Say $\dim(E) = n$. Using our nonsingular pairing,

$$\langle -, -\rangle \colon \bigwedge^p E^* \times \bigwedge^p E \longrightarrow K \qquad (1 \leq p \leq n),$$

defined on generators by

$$\langle u_1^* \wedge \cdots \wedge u_p^*, v_1 \wedge \cdots \wedge u_p\rangle = \det(u_i^*(v_j)),$$

we define various contraction operations,

$$\lrcorner : \bigwedge^p E \times \bigwedge^{p+q} E^* \longrightarrow \bigwedge^q E^* \qquad \text{(left hook)}$$

and

$$\llcorner : \bigwedge^{p+q} E^* \times \bigwedge^p E \longrightarrow \bigwedge^q E^* \qquad \text{(right hook)},$$

as well as the versions obtained by replacing E by E^* and E^{**} by E. We begin with the *left interior product or left hook,* $\lrcorner$.

Let $u \in \bigwedge^p E$. For any q such that $p + q \leq n$, multiplication on the right by u is a linear map

$$\wedge_R(u) \colon \bigwedge^q E \longrightarrow \bigwedge^{p+q} E,$$

given by

$$v \mapsto v \wedge u$$

where $v \in \bigwedge^q E$. The transpose of $\wedge_R(u)$ yields a linear map,

$$(\wedge_R(u))^t \colon (\bigwedge^{p+q} E)^* \longrightarrow (\bigwedge^q E)^*,$$

which, using the isomorphisms $(\bigwedge^{p+q} E)^* \cong \bigwedge^{p+q} E^*$ and $(\bigwedge^q E)^* \cong \bigwedge^q E^*$ can be viewed as a map

$$(\wedge_R(u))^t \colon \bigwedge^{p+q} E^* \longrightarrow \bigwedge^q E^*,$$

given by

$$z^* \mapsto z^* \circ \wedge_R(u),$$

where $z^* \in \bigwedge^{p+q} E^*$.

We denote $z^* \circ \wedge_R(u)$ by

$$u \lrcorner z^*.$$

In terms of our pairing, the q-vector $u \lrcorner z^*$ is uniquely defined by

$$\langle u \lrcorner z^*, v\rangle = \langle z^*, v \wedge u\rangle, \quad \text{for all } u \in \bigwedge^p E,\ v \in \bigwedge^q E \text{ and } z^* \in \bigwedge^{p+q} E^*.$$

It is immediately verified that

$$(u \wedge v) \lrcorner z^* = u \lrcorner (v \lrcorner z^*),$$

so $\lrcorner$ defines a left action

$$\lrcorner : \bigwedge^p E \times \bigwedge^{p+q} E^* \longrightarrow \bigwedge^q E^*.$$

By interchanging E and E^* and using the isomorphism,

$$(\bigwedge^k F)^* \cong \bigwedge^k F^*,$$

we can also define a left action

$$\lrcorner : \bigwedge^p E^* \times \bigwedge^{p+q} E \longrightarrow \bigwedge^q E.$$

In terms of our pairing, $u^* \lrcorner z$ is uniquely defined by

$$\langle v^*, u^* \lrcorner z\rangle = \langle v^* \wedge u^*, z\rangle, \quad \text{for all } u^* \in \bigwedge^p E^*,\ v^* \in \bigwedge^q E^* \text{ and } z \in \bigwedge^{p+q} E.$$

In order to proceed any further, we need some combinatorial properties of the basis of $\bigwedge^p E$ constructed from a basis, $(e_1, \ldots, e_n)$, of E. Recall that for any (nonempty) subset, $I \subseteq \{1, \ldots, n\}$, we let

$$e_I = e_{i_1} \wedge \cdots \wedge e_{i_p},$$

where $I = \{i_1, \ldots, i_p\}$ with $i_1 < \cdots < i_p$. We also let $e_\emptyset = 1$.

Given any two subsets $H, L \subseteq \{1, \ldots, n\}$, let

$$\rho_{H,L} = \begin{cases} 0 & \text{if } H \cap L \neq \emptyset, \\ (-1)^\nu & \text{if } H \cap L = \emptyset, \end{cases}$$

where

$$\nu = |\{(h,l) \mid (h,l) \in H \times L,\ h > l\}|.$$

Proposition 22.26. *For any basis, $(e_1, \ldots, e_n)$, of E the following properties hold:*

(1) If $H \cap L = \emptyset$, $|H| = h$, and $|L| = l$, then

$$\rho_{H,L}\rho_{L,H} = (-1)^{hl}.$$

(2) For $H, L \subseteq \{1, \ldots, m\}$, we have

$$e_H \wedge e_L = \rho_{H,L} e_{H \cup L}.$$

(3) For the left hook,

$$\lrcorner : \bigwedge^p E \times \bigwedge^{p+q} E^* \longrightarrow \bigwedge^q E^*,$$

we have

$$\begin{aligned} e_H \lrcorner e_L^* &= 0 \quad \text{if } H \not\subseteq L \\ e_H \lrcorner e_L^* &= \rho_{L-H,H} e_{L-H}^* \quad \text{if } H \subseteq L. \end{aligned}$$

Similar formulae hold for $\lrcorner : \bigwedge^p E^* \times \bigwedge^{p+q} E \longrightarrow \bigwedge^q E$. Using Proposition 22.26, we have the

Proposition 22.27. *For the left hook,*

$$\lrcorner : \bigwedge^p E \times \bigwedge^{p+q} E^* \longrightarrow \bigwedge^q E^*,$$

for every $u \in E$, we have

$$u \lrcorner (x^* \wedge y^*) = (-1)^s (u \lrcorner x^*) \wedge y^* + x^* \wedge (u \lrcorner y^*),$$

where $y \in \bigwedge^s E^*$.

Proof. We can prove the above identity assuming that x^* and y^* are of the form e_I^* and e_J^* using Proposition 22.26 but this is rather tedious. There is also a proof involving determinants, see Warner [148], Chapter 2. □

Thus, $\lrcorner$ is almost an anti-derivation, except that the sign, $(-1)^s$ is applied to the wrong factor.

It is also possible to define a *right interior product or right hook*, $\llcorner$, using multiplication on the left rather than multiplication on the right. Then, $\llcorner$ defines a right action,

$$\llcorner : \bigwedge^{p+q} E^* \times \bigwedge^{p} E \longrightarrow \bigwedge^{q} E^*,$$

such that

$$\langle z^*, u \wedge v\rangle = \langle z^* \llcorner u, v\rangle, \qquad \text{for all } u \in \textstyle\bigwedge^p E,\ v \in \bigwedge^q E, \text{ and } z^* \in \bigwedge^{p+q} E^*.$$

Similarly, we have the right action

$$\llcorner : \bigwedge^{p+q} E \times \bigwedge^{p} E^* \longrightarrow \bigwedge^{q} E,$$

such that

$$\langle u^* \wedge v^*, z\rangle = \langle v^*, z \llcorner u^*\rangle, \qquad \text{for all } u^* \in \textstyle\bigwedge^p E^*,\ v^* \in \bigwedge^q E^*, \text{ and } z \in \bigwedge^{p+q} E.$$

Since the left hook, $\lrcorner : \bigwedge^p E \times \bigwedge^{p+q} E^* \longrightarrow \bigwedge^q E^*$, is defined by

$$\langle u \lrcorner z^*, v\rangle = \langle z^*, v \wedge u\rangle, \quad \text{for all } u \in \textstyle\bigwedge^p E,\ v \in \bigwedge^q E \text{ and } z^* \in \bigwedge^{p+q} E^*,$$

the right hook,

$$\llcorner : \bigwedge^{p+q} E^* \times \bigwedge^{p} E \longrightarrow \bigwedge^{q} E^*,$$

by

$$\langle z^* \llcorner u, v\rangle = \langle z^*, u \wedge v\rangle, \qquad \text{for all } u \in \textstyle\bigwedge^p E,\ v \in \bigwedge^q E, \text{ and } z^* \in \bigwedge^{p+q} E^*,$$

and $v \wedge u = (-1)^{pq} u \wedge v$, we conclude that

$$u \lrcorner z^* = (-1)^{pq} z^* \llcorner u,$$

where $u \in \bigwedge^p E$ and $z \in \bigwedge^{p+q} E^*$.

Using the above property and Proposition 22.27 we get the following version of Proposition 22.27 for the right hook:

Proposition 22.28. *For the right hook,*

$$\llcorner : \bigwedge^{p+q} E^* \times \bigwedge^{p} E \longrightarrow \bigwedge^{q} E^*,$$

for every $u \in E$, we have

$$(x^* \wedge y^*) \llcorner u = (x^* \llcorner u) \wedge y^* + (-1)^r x^* \wedge (y^* \llcorner u),$$

where $x^* \in \bigwedge^r E^*$.

Thus, $\llcorner$ is an anti-derivation.

For $u \in E$, the right hook, $z^* \llcorner u$, is also denoted, $i(u)z^*$, and called *insertion operator* or *interior product*. This operator plays an important role in differential geometry. If we view $z^* \in \bigwedge^{n+1}(E^*)$ as an alternating multilinear map in $\mathrm{Alt}^{n+1}(E; K)$, then $i(u)z^* \in \mathrm{Alt}^n(E; K)$ is given by

$$(i(u)z^*)(v_1, \ldots, v_n) = z^*(u, v_1, \ldots, v_n).$$

Note that certain authors, such as Shafarevitch [139], denote our right hook $z^* \llcorner u$ (which is also the right hook in Bourbaki [21] and Fulton and Harris [58]) by $u \lrcorner z^*$.

Using the two versions of $\lrcorner$, we can define linear maps $\gamma\colon \bigwedge^p E \to \bigwedge^{n-p} E^*$ and $\delta\colon \bigwedge^p E^* \to \bigwedge^{n-p} E$. For any basis $(e_1, \ldots, e_n)$ of E, if we let $M = \{1, \ldots, n\}$, $e = e_1 \wedge \cdots \wedge e_n$, and $e^* = e_1^* \wedge \cdots \wedge e_n^*$, then

$$\gamma(u) = u \lrcorner e^* \quad \text{and} \quad \delta(v) = v^* \lrcorner e,$$

for all $u \in \bigwedge^p E$ and all $v^* \in \bigwedge^p E^*$. The following proposition is easily shown.

Proposition 22.29. *The linear maps $\gamma\colon \bigwedge^p E \to \bigwedge^{n-p} E^*$ and $\delta\colon \bigwedge^p E^* \to \bigwedge^{n-p} E$ are isomorphims. The isomorphisms γ and δ map decomposable vectors to decomposable vectors. Furthermore, if $z \in \bigwedge^p E$ is decomposable, then $\langle \gamma(z), z\rangle = 0$, and similarly for $z \in \bigwedge^p E^*$. If $(e'_1, \ldots, e'_n)$ is any other basis of E and $\gamma'\colon \bigwedge^p E \to \bigwedge^{n-p} E^*$ and $\delta'\colon \bigwedge^p E^* \to \bigwedge^{n-p} E$ are the corresponding isomorphisms, then $\gamma' = \lambda\gamma$ and $\delta' = \lambda^{-1}\delta$ for some nonzero $\lambda \in \Omega$.*

Proof. Using Proposition 22.26, for any subset $J \subseteq \{1, \ldots, n\} = M$ such that $|J| = p$, we have

$$\gamma(e_J) = e_J \lrcorner e^* = \rho_{M-J,J} e^*_{M-J} \quad \text{and} \quad \delta(e^*_J) = e^*_J \lrcorner e = \rho_{M-J,J} e_{M-J}.$$

Thus,

$$\delta \circ \gamma(e_J) = \rho_{M-J,J}\rho_{J,M-J} e_J = (-1)^{p(n-p)} e_J.$$

A similar result holds for $\gamma \circ \delta$. This implies that

$$\delta \circ \gamma = (-1)^{p(n-p)}\mathrm{id} \quad \text{and} \quad \gamma \circ \delta = (-1)^{p(n-p)}\mathrm{id}.$$

Thus, γ and δ are isomorphisms. If $z \in \bigwedge^p E$ is decomposable, then $z = u_1 \wedge \cdots \wedge u_p$ where $u_1, \ldots, u_p$ are linearly independent since $z \neq 0$, and we can pick a basis of E of the form $(u_1, \ldots, u_n)$. Then, the above formulae show that

$$\gamma(z) = \pm u^*_{p+1} \wedge \cdots \wedge u^*_n.$$

Clearly

$$\langle \gamma(z), z\rangle = 0.$$

If $(e'_1, \ldots, e'_n)$ is any other basis of E, because $\bigwedge^m E$ has dimension 1, we have

$$e'_1 \wedge \cdots \wedge e'_n = \lambda e_1 \wedge \cdots \wedge e_n$$

for some nonnull $\lambda \in \Omega$, and the rest is trivial. □

We are now ready to tacke the problem of finding criteria for decomposability. We need a few preliminary results.

Proposition 22.30. *Given $z \in \bigwedge^p E$, with $z \neq 0$, the smallest vector space $W \subseteq E$ such that $z \in \bigwedge^p W$ is generated by the vectors of the form*

$$u^* \lrcorner z, \qquad \text{with } u^* \in \textstyle\bigwedge^{p-1} E^*.$$

Proof. First, let W be any subspace such that $z \in \bigwedge^p(E)$ and let $(e_1, \ldots, e_r, e_{r+1}, \ldots, e_n)$ be a basis of E such that $(e_1, \ldots, e_r)$ is a basis of W. Then, $u^* = \sum_I e_I^*$, where $I \subseteq \{1, \ldots, n\}$ and $|I| = p-1$, and $z = \sum_J e_J$, where $J \subseteq \{1, \ldots, r\}$ and $|J| = p \leq r$. It follows immediately from the formula of Proposition 22.26 (3) that $u^* \lrcorner z \in W$.

Next, we prove that if W is the smallest subspace of E such that $z \in \bigwedge^p(W)$, then W is generated by the vectors of the form $u^* \lrcorner z$, where $u^* \in \bigwedge^{p-1} E^*$. Suppose not, then the vectors $u^* \lrcorner z$ with $u^* \in \bigwedge^{p-1} E^*$ span a proper subspace, U, of W. We prove that for every subspace, W', of W, with $\dim(W') = \dim(W) - 1 = r - 1$, it is not possible that $u^* \lrcorner z \in W'$ for all $u^* \in \bigwedge^{p-1} E^*$. But then, as U is a proper subspace of W, it is contained in some subspace, W', with $\dim(W') = r - 1$ and we have a contradiction.

Let $w \in W - W'$ and pick a basis of W formed by a basis $(e_1, \ldots, e_{r-1})$ of W' and w. We can write $z = z' + w \wedge z''$, where $z' \in \bigwedge^p W'$ and $z'' \in \bigwedge^{p-1} W'$, and since W is the smallest subspace containing z, we have $z'' \neq 0$. Consequently, if we write $z'' = \sum_I e_I$ in terms of the basis $(e_1, \ldots, e_{r-1})$ of W', there is some e_I, with $I \subseteq \{1, \ldots, r-1\}$ and $|I| = p - 1$, so that the coefficient λ_I is nonzero. Now, using any basis of E containing $(e_1, \ldots, e_{r-1}, w)$, by Proposition 22.26 (3), we see that

$$e_I^* \lrcorner (w \wedge e_I) = \lambda w, \qquad \lambda = \pm 1.$$

It follows that

$$e_I^* \lrcorner z = e_I^* \lrcorner (z' + w \wedge z'') = e_I^* \lrcorner z' + e_I^* \lrcorner (w \wedge z'') = e_I^* \lrcorner z' + \lambda w,$$

with $e_I^* \lrcorner z' \in W'$, which shows that $e_I^* \lrcorner z \notin W'$. Therefore, W is indeed generated by the vectors of the form $u^* \lrcorner z$, where $u^* \in \bigwedge^{p-1} E^*$. □

Proposition 22.31. *Any nonzero $z \in \bigwedge^p E$ is decomposable iff*

$$(u^* \lrcorner z) \wedge z = 0, \qquad \text{for all } u^* \in \textstyle\bigwedge^{p-1} E^*.$$

Proof. Clearly, $z \in \bigwedge^p E$ is decomposable iff the smallest vector space, W, such that $z \in \bigwedge^p W$ has dimension p. If $\dim(W) = p$, we have $z = e_1 \wedge \cdots \wedge e_p$ where $e_1, \ldots, e_p$ form a basis of W. By Proposition 22.30, for every $u^* \in \bigwedge^{p-1} E^*$, we have $u^* \lrcorner z \in W$, so each $u^* \lrcorner z$ is a linear combination of the e_i's and $(u^* \lrcorner z) \wedge z = (u^* \lrcorner z) \wedge e_1 \wedge \cdots \wedge e_p = 0$.

Now, assume that $(u^* \lrcorner z) \wedge z = 0$ for all $u^* \in \bigwedge^{p-1} E^*$ and that $\dim(W) = n > p$. If $e_1, \ldots, e_n$ is a basis of W, then we have $z = \sum_I \lambda_I e_I$, where $I \subseteq \{1, \ldots, n\}$ and $|I| = p$.

Recall that $z \neq 0$, and so, some λ_I is nonzero. By Proposition 22.30, each e_i can be written as $u^* \lrcorner z$ for some $u^* \in \bigwedge^{p-1} E^*$ and since $(u^* \lrcorner z) \wedge z = 0$ for all $u^* \in \bigwedge^{p-1} E^*$, we get

$$e_j \wedge z = 0 \quad \text{for} \quad j = 1, \ldots, n.$$

By wedging $z = \sum_I \lambda_I e_I$ with each e_j, as $n > p$, we deduce $\lambda_I = 0$ for all I, so $z = 0$, a contradiction. Therefore, $n = p$ and z is decomposable. □

In Proposition 22.31, we can let u^* range over a basis of $\bigwedge^{p-1} E^*$, and then, the conditions are

$$(e_H^* \lrcorner z) \wedge z = 0$$

for all $H \subseteq \{1, \ldots, n\}$, with $|H| = p - 1$. Since $(e_H^* \lrcorner z) \wedge z \in \bigwedge^{p+1} E$, this is equivalent to

$$e_J^* \left((e_H^* \lrcorner z) \wedge z \right) = 0$$

for all $H, J \subseteq \{1, \ldots, n\}$, with $|H| = p - 1$ and $|J| = p + 1$. Then, for all $I, I' \subseteq \{1, \ldots, n\}$ with $|I| = |I'| = p$, we can show that

$$e_J^* \left((e_H^* \lrcorner e_I) \wedge e_{I'} \right) = 0,$$

unless there is some $i \in \{1, \ldots, n\}$ such that

$$I - H = \{i\}, \quad J - I' = \{i\}.$$

In this case,

$$e_J^* \left((e_H^* \lrcorner e_{H \cup \{i\}}) \wedge e_{J - \{i\}} \right) = \rho_{\{i\},H} \rho_{\{i\},J-\{i\}}.$$

If we let

$$\epsilon_{i,J,H} = \rho_{\{i\},H} \rho_{\{i\},J-\{i\}},$$

we have $\epsilon_{i,J,H} = +1$ if the parity of the number of $j \in J$ such that $j < i$ is the same as the parity of the number of $h \in H$ such that $h < i$, and $\epsilon_{i,J,H} = -1$ otherwise.

Finally, we obtain the following criterion in terms of quadratic equations (*Plücker's equations*) for the decomposability of an alternating tensor:

Proposition 22.32. *(Grassmann-Plücker's Equations) For $z = \sum_I \lambda_I e_I \in \bigwedge^p E$, the conditions for $z \neq 0$ to be decomposable are*

$$\sum_{i \in J - H} \epsilon_{i,J,H} \lambda_{H \cup \{i\}} \lambda_{J - \{i\}} = 0,$$

for all $H, J \subseteq \{1, \ldots, n\}$ such that $|H| = p - 1$ and $|J| = p + 1$.

Using these criteria, it is a good exercise to prove that if $\dim(E) = n$, then every tensor in $\bigwedge^{n-1}(E)$ is decomposable. This can also be shown directly.

It should be noted that the equations given by Proposition 22.32 are not independent. For example, when $\dim(E) = n = 4$ and $p = 2$, these equations reduce to the single equation

$$\lambda_{12}\lambda_{34} - \lambda_{13}\lambda_{24} + \lambda_{14}\lambda_{23} = 0.$$

When the field, K, is the field of complex numbers, this is the homogeneous equation of a quadric in $\mathbb{CP}^5$ known as the *Klein quadric*. The points on this quadric are in one-to-one correspondence with the lines in $\mathbb{CP}^3$.

22.18 Vector-Valued Alternating Forms

In this section, the vector space, E, is assumed to have finite dimension. We know that there is a canonical isomorphism, $\bigwedge^n(E^*) \cong \mathrm{Alt}^n(E; K)$, between alternating n-forms and alternating multilinear maps. As in the case of general tensors, the isomorphisms,

$$\begin{aligned} \mathrm{Alt}^n(E; F) &\cong \mathrm{Hom}(\bigwedge^n(E), F) \\ \mathrm{Hom}(\bigwedge^n(E), F) &\cong (\bigwedge^n(E))^* \otimes F \\ (\bigwedge^n(E))^* &\cong \bigwedge^n(E^*) \end{aligned}$$

yield a canonical isomorphism

$$\mathrm{Alt}^n(E; F) \cong \left(\bigwedge^n(E^*)\right) \otimes F.$$

Note that F may have infinite dimension. This isomorphism allows us to view the tensors in $\bigwedge^n(E^*) \times F$ as *vector valued alternating forms*, a point of view that is useful in differential geometry. If $(f_1, \ldots, f_r)$ is a basis of F, every tensor, $\omega \in \bigwedge^n(E^*) \times F$ can be written as some linear combination

$$\omega = \sum_{i=1}^{r} \alpha_i \otimes f_i,$$

with $\alpha_i \in \bigwedge^n(E^*)$. We also let

$$\bigwedge(E; F) = \bigoplus_{n=0} \left(\bigwedge^n(E^*)\right) \otimes F = \left(\bigwedge(E)\right) \otimes F.$$

Given three vector spaces, F, G, H, if we have some bilinear map, $\Phi\colon F \otimes G \to H$, then we can define a multiplication operation,

$$\wedge_\Phi\colon \bigwedge(E; F) \times \bigwedge(E; G) \to \bigwedge(E; H),$$

as follows: For every pair, (m, n), we define the multiplication,

$$\wedge_\Phi \colon \left(\left(\bigwedge^m(E^*)\right) \otimes F\right) \times \left(\left(\bigwedge^n(E^*)\right) \otimes G\right) \longrightarrow \left(\bigwedge^{m+n}(E^*)\right) \otimes H,$$

by

$$(\alpha \otimes f) \wedge_\Phi (\beta \otimes g) = (\alpha \wedge \beta) \otimes \Phi(f, g).$$

As in Section 22.15 (following H. Cartan [30]) we can also define a multiplication,

$$\wedge_\Phi \colon \mathrm{Alt}^m(E; F) \times \mathrm{Alt}^m(E; G) \longrightarrow \mathrm{Alt}^{m+n}(E; H),$$

directly on alternating multilinear maps as follows: For $f \in \mathrm{Alt}^m(E; F)$ and $g \in \mathrm{Alt}^n(E; G)$,

$$(f \wedge_\Phi g)(u_1, \ldots, u_{m+n}) = \sum_{\sigma \in \mathrm{shuffle}(m,n)} \mathrm{sgn}(\sigma)\, \Phi(f(u_{\sigma(1)}, \ldots, u_{\sigma(m)}), g(u_{\sigma(m+1)}, \ldots, u_{\sigma(m+n)})),$$

where $\mathrm{shuffle}(m, n)$ consists of all (m, n)-"shuffles", that is, permutations, σ, of $\{1, \ldots m+n\}$, such that $\sigma(1) < \cdots < \sigma(m)$ and $\sigma(m+1) < \cdots < \sigma(m+n)$.

In general, not much can be said about $\wedge_\Phi$ unless Φ has some additional properties. In particular, $\wedge_\Phi$ is generally not associative. We also have the map,

$$\mu \colon \left(\bigwedge^n(E^*)\right) \otimes F \longrightarrow \mathrm{Alt}^n(E; F),$$

defined on generators by

$$\mu((v_1^* \wedge \cdots \wedge v_n^*) \otimes a)(u_1, \ldots, u_n) = (\det(v_j^*(u_i))a.$$

Proposition 22.33. *The map*

$$\mu \colon \left(\bigwedge^n(E^*)\right) \otimes F \longrightarrow \mathrm{Alt}^n(E; F),$$

defined as above is a canonical isomorphism for every $n \geq 0$. Furthermore, given any three vector spaces, F, G, H, and any bilinear map, $\Phi \colon F \times G \to H$, for all $\omega \in (\bigwedge^n(E^*)) \otimes F$ and all $\eta \in (\bigwedge^n(E^*)) \otimes G$,

$$\mu(\alpha \wedge_\Phi \beta) = \mu(\alpha) \wedge_\Phi \mu(\beta).$$

Proof. Since we already know that $(\bigwedge^n(E^*)) \otimes F$ and $\mathrm{Alt}^n(E; F)$ are isomorphic, it is enough to show that μ maps some basis of $(\bigwedge^n(E^*)) \otimes F$ to linearly independent elements. Pick some bases, $(e_1, \ldots, e_p)$ in E and $(f_j)_{j \in J}$ in F. Then, we know that the vectors, $e_I^* \otimes f_j$, where $I \subseteq \{1, \ldots, p\}$ and $|I| = n$ form a basis of $(\bigwedge^n(E^*)) \otimes F$. If we have a linear dependence,

$$\sum_{I,j} \lambda_{I,j} \mu(e_I^* \otimes f_j) = 0,$$

applying the above combination to each $(e_{i_1}, \ldots, e_{i_n})$ ($I = \{i_1, \ldots, i_n\}$, $i_1 < \cdots < i_n$), we get the linear combination

$$\sum_j \lambda_{I,j} f_j = 0,$$

and by linear independence of the f_j's, we get $\lambda_{I,j} = 0$, for all I and all j. Therefore, the $\mu(e_I^* \otimes f_j)$ are linearly independent and we are done. The second part of the proposition is easily checked (a simple computation). □

A special case of interest is the case where $F = G = H$ is a Lie algebra and $\Phi(a, b) = [a, b]$, is the Lie bracket of F. In this case, using a base, $(f_1, \ldots, f_r)$, of F if we write $\omega = \sum_i \alpha_i \otimes f_i$ and $\eta = \sum_j \beta_j \otimes f_j$, we have

$$[\omega, \eta] = \sum_{i,j} \alpha_i \wedge \beta_j \otimes [f_i, f_j].$$

Consequently,

$$[\eta, \omega] = (-1)^{mn+1}[\omega, \eta].$$

The following proposition will be useful in dealing with vector-valued differential forms:

Proposition 22.34. *If $(e_1, \ldots, e_p)$ is any basis of E, then every element, $\omega \in (\bigwedge^n(E^*)) \otimes F$, can be written in a unique way as*

$$\omega = \sum_I e_I^* \otimes f_I, \qquad f_I \in F,$$

where the e_I^* are defined as in Section 22.12.

Proof. Since, by Proposition 22.19, the e_I^* form a basis of $\bigwedge^n(E^*)$, elements of the form $e_I^* \otimes f$ span $(\bigwedge^n(E^*)) \otimes F$. Now, if we apply $\mu(\omega)$ to $(e_{i_1}, \ldots, e_{i_n})$, where $I = \{i_1, \ldots, i_n\} \subseteq \{1, \ldots, p\}$, we get

$$\mu(\omega)(e_{i_1}, \ldots, e_{i_n}) = \mu(e_I^* \otimes f_I)(e_{i_1}, \ldots, e_{i_n}) = f_I.$$

Therefore, the f_I are uniquely determined by f. □

Proposition can also be formulated in terms of alternating multilinear maps, a fact that will be useful to deal with differential forms.

Define the product, $\cdot\colon \mathrm{Alt}^n(E; \mathbb{R}) \times F \to \mathrm{Alt}^n(E; F)$, as follows: For all $\omega \in \mathrm{Alt}^n(E; \mathbb{R})$ and all $f \in F$,

$$(\omega \cdot f)(u_1, \ldots, u_n) = \omega(u_1, \ldots, u_n) f,$$

for all $u_1, \ldots, u_n \in E$. Then, it is immediately verified that for every $\omega \in (\bigwedge^n(E^*)) \otimes F$ of the form

$$\omega = u_1^* \wedge \cdots \wedge u_n^* \otimes f,$$

we have

$$\mu(u_1^* \wedge \cdots \wedge u_n^* \otimes f) = \mu(u_1^* \wedge \cdots \wedge u_n^*) \cdot f.$$

Then, Proposition 22.34 yields

Proposition 22.35. *If $(e_1, \ldots, e_p)$ is any basis of E, then every element, $\omega \in \mathrm{Alt}^n(E; F)$, can be written in a unique way as*

$$\omega = \sum_I e_I^* \cdot f_I, \qquad f_I \in F,$$

where the e_I^* are defined as in Section 22.12.

22.19 Tensor Products of Modules over a Commmutative Ring

If R is a commutative ring with identity (say 1), recall that a *module over R* (or *R-module*) is an abelian group, M, with a scalar multiplication, $\cdot\colon R \times M \to M$, and all the axioms of a vector space are satisfied.

At first glance, a module does not seem any different from a vector space but the lack of multiplicative inverses in R has drastic consequences, one being that unlike vector spaces, modules are generally not free, that is, have no bases. Furthermore, a module may have *torsion elements*, that is, elements, $m \in M$, such that $\lambda \cdot m = 0$, even though $m \neq 0$ and $\lambda \neq 0$.

Nevertheless, it is possible to define tensor products of modules over a ring, just as in Section 22.1 and the results of this section continue to hold. The results of Section 22.3 also continue to hold since they are based on the universal mapping property. However, the results of Section 22.2 on bases generally fail, except for free modules. Similarly, the results of Section 22.4 on duality generally fail. Tensor algebras can be defined for modules, as in Section 22.5. Symmetric tensor and alternating tensors can be defined for modules but again, results involving bases generally fail.

Tensor products of modules have some unexpected properties. For example, if p and q are relatively prime integers, then

$$\mathbb{Z}/p\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/q\mathbb{Z} = (0).$$

It is possible to salvage certain properties of tensor products holding for vector spaces by restricting the class of modules under consideration. For example, *projective modules*, have a pretty good behavior w.r.t. tensor products.

A free R-module, F, is a module that has a basis (*i.e.*, there is a family, $(e_i)_{i \in I}$, of linearly independent vectors in F that span F). Projective modules have many equivalent characterizations. Here is one that is best suited for our needs:

Definition 22.9. An R-module, P, is *projective* if it is a summand of a free module, that is, if there is a free R-module, F, and some R-module, Q, so that

$$F = P \oplus Q.$$

Given any R-module, M, we let $M^* = \mathrm{Hom}_R(M, R)$ be its *dual*. We have the following proposition:

Proposition 22.36. *For any finitely-generated projective R-modules, P, and any R-module, Q, we have the isomorphisms:*

$$\begin{aligned} P^{**} &\cong P \\ \mathrm{Hom}_R(P, Q) &\cong P^* \otimes_R Q. \end{aligned}$$

Proof sketch. We only consider the second isomorphism. Since P is projective, we have some R-modules, P_1, F, with

$$P \oplus P_1 = F,$$

where F is some free module. Now, we know that for any R-modules, U, V, W, we have

$$\mathrm{Hom}_R(U \oplus V, W) \cong \mathrm{Hom}_R(U, W) \prod \mathrm{Hom}_R(V, W) \cong \mathrm{Hom}_R(U, W) \oplus \mathrm{Hom}_R(V, W),$$

so

$$P^* \oplus P_1^* \cong F^*, \qquad \mathrm{Hom}_R(P, Q) \oplus \mathrm{Hom}_R(P_1, Q) \cong \mathrm{Hom}_R(F, Q).$$

By tensoring with Q and using the fact that tensor distributes w.r.t. coproducts, we get

$$(P^* \otimes_R Q) \oplus (P_1^* \otimes Q) \cong (P^* \oplus P_1^*) \otimes_R Q \cong F^* \otimes_R Q.$$

Now, the proof of Proposition 22.9 goes through because F is free and finitely generated, so

$$\alpha_\otimes \colon (P^* \otimes_R Q) \oplus (P_1^* \otimes Q) \cong F^* \otimes_R Q \longrightarrow \mathrm{Hom}_R(F, Q) \cong \mathrm{Hom}_R(P, Q) \oplus \mathrm{Hom}_R(P_1, Q)$$

is an isomorphism and as α_α maps $P^* \otimes_R Q$ to $\mathrm{Hom}_R(P, Q)$, it yields an isomorphism between these two spaces. □

The isomorphism $\alpha_\otimes \colon P^* \otimes_R Q \cong \mathrm{Hom}_R(P, Q)$ of Proposition 22.36 is still given by

$$\alpha_\otimes(u^* \otimes f)(x) = u^*(x) f, \qquad u^* \in P^*,\ f \in Q,\ x \in P.$$

It is convenient to introduce the *evaluation map*, $\mathrm{Ev}_x \colon P^* \otimes_R Q \to Q$, defined for every $x \in P$ by

$$\mathrm{Ev}_x(u^* \otimes f) = u^*(x) f, \qquad u^* \in P^*,\ f \in Q.$$

In Section 11.2 we will need to consider a slightly weaker version of the universal mapping property of tensor products. The situation is this: We have a commutative R-algebra, S,

where R is a field (or even a commutative ring), we have two R-modules, U and V, and moreover, U is a right S-module and V is a left S-module. In Section 11.2, this corresponds to $R = \mathbb{R}$, $S = C^\infty(B)$, $U = \mathcal{A}^i(\xi)$ and $V = \Gamma(\xi)$, where ξ is a vector bundle. Then, we can form the tensor product, $U \otimes_R V$, and we let $U \otimes_S V$ be the quotient module, $(U \otimes_R V)/W$, where W is the submodule of $U \otimes_R V$ generated by the elements of the form

$$us \otimes_R v - u \otimes_R sv.$$

As S is commutative, we can make $U \otimes_S V$ into an S-module by defining the action of S *via*

$$s(u \otimes_S v) = us \otimes_S v.$$

It is immediately verified that this S-module is isomorphic to the tensor product of U and V as S-modules and the following universal mapping property holds:

Proposition 22.37. *For every, R-bilinear map, $f\colon U \times V \to Z$, if f satisfies the property*

$$f(us, v) = f(u, sv), \qquad \textit{for all } u \in U,\ v \in V,\ s \in S,$$

then f induces a unique R-linear map, $\widehat{f}\colon U \otimes_S V \to Z$, such that

$$f(u, v) = \widehat{f}(u \otimes_S v), \qquad \textit{for all } u \in U,\ v \in V.$$

Note that the linear map, $\widehat{f}\colon U \otimes_S V \to Z$, is only R-linear, it is not S-linear in general.

22.20 The Pfaffian Polynomial

Let $\mathfrak{so}(2n)$ denote the vector space (actually, Lie algebra) of $2n \times 2n$ real skew-symmetric matrices. It is well-known that every matrix, $A \in \mathfrak{so}(2n)$, can be written as

$$A = PDP^\top,$$

where P is an orthogonal matrix and where D is a block diagonal matrix

$$D = \begin{pmatrix} D_1 & & & \\ & D_2 & & \\ & & \ddots & \\ & & & D_n \end{pmatrix}$$

consisting of 2×2 blocks of the form

$$D_i = \begin{pmatrix} 0 & -a_i \\ a_i & 0 \end{pmatrix}.$$

For a proof, see see Horn and Johnson [80], Corollary 2.5.14, Gantmacher [62], Chapter IX, or Gallier [60], Chapter 11.

Since $\det(D_i) = a_i^2$ and $\det(A) = \det(PDP^\top) = \det(D) = \det(D_1)\cdots\det(D_n)$, we get

$$\det(A) = (a_1 \cdots a_n)^2.$$

The Pfaffian is a polynomial function, $\mathrm{Pf}(A)$, in skew-symmetric $2n \times 2n$ matrices, A, (a polynomial in $(2n-1)n$ variables) such that

$$\mathrm{Pf}(A)^2 = \det(A)$$

and for every arbitrary matrix, B,

$$\mathrm{Pf}(BAB^\top) = \mathrm{Pf}(A)\det(B).$$

The Pfaffian shows up in the definition of the Euler class of a vector bundle. There is a simple way to define the Pfaffian using some exterior algebra. Let $(e_1, \ldots, e_{2n})$ be any basis of $\mathbb{R}^{2n}$. For any matrix, $A \in \mathfrak{so}(2n)$, let

$$\omega(A) = \sum_{i<j} a_{ij}\, e_i \wedge e_j,$$

where $A = (a_{ij})$. Then, $\bigwedge^n \omega(A)$ is of the form $Ce_1 \wedge e_2 \wedge \cdots \wedge e_{2n}$ for some constant, $C \in \mathbb{R}$.

Definition 22.10. For every skew symmetric matrix, $A \in \mathfrak{so}(2n)$, the *Pfaffian polynomial* or *Pfaffian* is the degree n polynomial, $\mathrm{Pf}(A)$, defined by

$$\bigwedge^n \omega(A) = n!\,\mathrm{Pf}(A)\, e_1 \wedge e_2 \wedge \cdots \wedge e_{2n}.$$

Clearly, $\mathrm{Pf}(A)$ is independent of the basis chosen. If A is the block diagonal matrix D, a simple calculation shows that

$$\omega(D) = -(a_1 e_1 \wedge e_2 + a_2 e_3 \wedge e_4 + \cdots + a_n e_{2n-1} \wedge e_{2n})$$

and that

$$\bigwedge^n \omega(D) = (-1)^n n!\, a_1 \cdots a_n\, e_1 \wedge e_2 \wedge \cdots \wedge e_{2n},$$

and so

$$\mathrm{Pf}(D) = (-1)^n a_1 \cdots a_n.$$

Since $\mathrm{Pf}(D)^2 = (a_1 \cdots a_n)^2 = \det(A)$, we seem to be on the right track.

Proposition 22.38. *For every skew symmetric matrix, $A \in \mathfrak{so}(2n)$ and every arbitrary matrix, B, we have:*

(i) $\mathrm{Pf}(A)^2 = \det(A)$

(ii) $\mathrm{Pf}(BAB^\top) = \mathrm{Pf}(A)\det(B)$.

Proof. If we assume that (ii) is proved then, since we can write $A = PDP^\top$ for some orthogonal matrix, P, and some block diagonal matrix, D, as above, as $\det(P) = \pm 1$ and $\mathrm{Pf}(D)^2 = \det(A)$, we get

$$\mathrm{Pf}(A)^2 = \mathrm{Pf}(PDP^\top)^2 = \mathrm{Pf}(D)^2\det(P)^2 = \det(A),$$

which is (i). Therefore, it remains to prove (ii).

Let $f_i = Be_i$, for $i = 1, \ldots, 2n$, where $(e_1, \ldots, e_{2n})$ is any basis of $\mathbb{R}^{2n}$. Since $f_i = \sum_k b_{ki}e_k$, we have

$$\tau = \sum_{i,j} a_{ij}\, f_i \wedge f_j = \sum_{i,j}\sum_{k,l} b_{ki}a_{ij}b_{lj}\, e_k \wedge e_l = \sum_{k,l}(BAB^\top)_{kl}\, e_k \wedge e_l,$$

and so, as $BAB^\top$ is skew symmetric and $e_k \wedge e_l = -e_l \wedge e_k$, we get

$$\tau = 2\omega(BAB^\top).$$

Consequently,

$$\bigwedge^n \tau = 2^n \bigwedge^n \omega(BAB^\top) = 2^n n!\,\mathrm{Pf}(BAB^\top)\, e_1 \wedge e_2 \wedge \cdots \wedge e_{2n}.$$

Now,

$$\bigwedge^n \tau = C\, f_1 \wedge f_2 \wedge \cdots \wedge f_{2n},$$

for some $C \in \mathbb{R}$. If B is singular, then the f_i are linearly dependent which implies that $f_1 \wedge f_2 \wedge \cdots \wedge f_{2n} = 0$, in which case,

$$\mathrm{Pf}(BAB^\top) = 0,$$

as $e_1 \wedge e_2 \wedge \cdots \wedge e_{2n} \neq 0$. Therefore, if B is singular, $\det(B) = 0$ and

$$\mathrm{Pf}(BAB^\top) = 0 = \mathrm{Pf}(A)\det(B).$$

If B is invertible, as $\tau = \sum_{i,j} a_{ij}\, f_i \wedge f_j = 2\sum_{i<j} a_{ij}\, f_i \wedge f_j$, we have

$$\bigwedge^n \tau = 2^n n!\,\mathrm{Pf}(A)\, f_1 \wedge f_2 \wedge \cdots \wedge f_{2n}.$$

However, as $f_i = Be_i$, we have

$$f_1 \wedge f_2 \wedge \cdots \wedge f_{2n} = \det(B)\, e_1 \wedge e_2 \wedge \cdots \wedge e_{2n},$$

so

$$\bigwedge^n \tau = 2^n n!\,\mathrm{Pf}(A)\det(B)\, e_1 \wedge e_2 \wedge \cdots \wedge e_{2n}$$

and as

$$\bigwedge^n \tau = 2^n n!\,\mathrm{Pf}(BAB^\top)\, e_1 \wedge e_2 \wedge \cdots \wedge e_{2n},$$

we get

$$\mathrm{Pf}(BAB^\top) = \mathrm{Pf}(A)\det(B),$$

as claimed. □

Remark: It can be shown that the polynomial, $\mathrm{Pf}(A)$, is the unique polynomial with integer coefficients such that $\mathrm{Pf}(A)^2 = \det(A)$ and $\mathrm{Pf}(\mathrm{diag}(S,\ldots,S)) = +1$, where

$$S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

see Milnor and Stasheff [111] (Appendix C, Lemma 9). There is also an explicit formula for $\mathrm{Pf}(A)$, namely:

$$\mathrm{Pf}(A) = \frac{1}{2^n n!} \sum_{\sigma \in \mathfrak{S}_{2n}} \mathrm{sgn}(\sigma) \prod_{i=1}^{n} a_{\sigma(2i-1)\,\sigma(2i)}.$$

Beware, some authors use a different sign convention and require the Pfaffian to have the value $+1$ on the matrix $\mathrm{diag}(S',\ldots,S')$, where

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

For example, if $\mathbb{R}^{2n}$ is equipped with an inner product, $\langle -,-\rangle$, then some authors define $\omega(A)$ as

$$\omega(A) = \sum_{i<j} \langle Ae_i, e_j\rangle\, e_i \wedge e_j,$$

where $A = (a_{ij})$. But then, $\langle Ae_i, e_j\rangle = a_{ji}$ and **not** a_{ij}, and this Pfaffian takes the value $+1$ on the matrix $\mathrm{diag}(S',\ldots,S')$. This version of the Pfaffian differs from our version by the factor $(-1)^n$. In this respect, Madsen and Tornehave [101] seem to have an incorrect sign in Proposition B6 of Appendix C.

We will also need another property of Pfaffians. Recall that the ring, $M_n(\mathbb{C})$, of $n \times n$ matrices over $\mathbb{C}$ is embedded in the ring, $M_{2n}(\mathbb{R})$, of $2n \times 2n$ matrices with real coefficients, using the injective homomorphism that maps every entry $z = a + ib \in \mathbb{C}$ to the 2×2 matrix

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}.$$

If $A \in M_n(\mathbb{C})$, let $A_{\mathbb{R}} \in M_{2n}(\mathbb{R})$ denote the real matrix obtained by the above process. Observe that every skew Hermitian matrix, $A \in \mathfrak{u}(n)$, (*i.e.*, with $A^* = \overline{A}^\top = -A$) yields a matrix $A_{\mathbb{R}} \in \mathfrak{so}(2n)$.

Proposition 22.39. *For every skew Hermitian matrix, $A \in \mathfrak{u}(n)$, we have*

$$\mathrm{Pf}(A_{\mathbb{R}}) = i^n \det(A).$$

Proof. It is well-known that a skew Hermitian matrix can be diagonalized with respect to a unitary matrix, U, and that the eigenvalues are pure imaginary or zero, so we can write

$$A = U \,\mathrm{diag}(ia_1, \ldots, ia_n)U^*,$$

for some reals, $a_i \in \mathbb{R}$. Consequently, we get

$$A_{\mathbb{R}} = U_{\mathbb{R}} \,\mathrm{diag}(D_1, \ldots, D_n)U_{\mathbb{R}}^{\top},$$

where

$$D_i = \begin{pmatrix} 0 & -a_i \\ a_i & 0 \end{pmatrix}$$

and

$$\mathrm{Pf}(A_{\mathbb{R}}) = \mathrm{Pf}(\mathrm{diag}(D_1, \ldots, D_n)) = (-1)^n \, a_1 \cdots a_n,$$

as we saw before. On the other hand,

$$\det(A) = \det(\mathrm{diag}(ia_1, \ldots, ia_n)) = i^n \, a_1 \cdots a_n,$$

and as $(-1)^n = i^n i^n$, we get

$$\mathrm{Pf}(A_{\mathbb{R}}) = i^n \det(A),$$

as claimed. □ □

Madsen and Tornehave [101] state Proposition 22.39 using the factor $(-i)^n$, which is wrong.

Bibliography

[1] Ralph Abraham and Jerrold E. Marsden. *Foundations of Mechanics*. Addison Wesley, second edition, 1978.

[2] George E. Andrews, Richard Askey, and Ranjan Roy. *Special Functions*. Cambridge University Press, first edition, 2000.

[3] Vincent Arsigny. *Processing Data in Lie Groups: An Algebraic Approach. Application to Non-Linear Registration and Diffusion Tensor MRI*. PhD thesis, École Polytechnique, Palaiseau, France, 2006. Thèse de Sciences.

[4] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A fast and log-euclidean polyaffine framework for locally affine registration. Technical report, INRIA, 2004, route des Lucioles, 06902 Sophia Antipolis Cedex, France, 2006. Report No. 5865.

[5] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. on Matrix Analysis and Applications*, 29(1):328–347, 2007.

[6] Vincent Arsigny, Xavier Pennec, and Nicholas Ayache. Polyrigid and polyaffine transformations: a novel geometrical tool to deal with non-rigid deformations–application to the registration of histological slices. *Medical Image Analysis*, 9(6):507–523, 2005.

[7] Michael Artin. *Algebra*. Prentice Hall, first edition, 1991.

[8] Andreas Arvanitoyeogos. *An Introduction to Lie Groups and the Geometry of Homogeneous Spaces*. SML, Vol. 22. AMS, first edition, 2003.

[9] M. F. Atiyah and I. G. Macdonald. *Introduction to Commutative Algebra*. Addison Wesley, third edition, 1969.

[10] Michael F. Atiyah. *K-Theory*. Addison Wesley, first edition, 1988.

[11] Michael F Atiyah, Raoul Bott, and Arnold Shapiro. Clifford modules. *Topology*, 3, Suppl. 1:3–38, 1964.

[12] Sheldon Axler, Paul Bourdon, and Wade Ramey. *Harmonic Function Theory*. GTM No. 137. Springer Verlag, second edition, 2001.

[13] Andrew Baker. *Matrix Groups. An Introduction to Lie Group Theory*. SUMS. Springer, 2002.

[14] Ronen Basri and David W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):228–233, 2003.

[15] Marcel Berger. *Géométrie 1*. Nathan, 1990. English edition: Geometry 1, Universitext, Springer Verlag.

[16] Marcel Berger. *A Panoramic View of Riemannian Geometry*. Springer, 2003.

[17] Marcel Berger and Bernard Gostiaux. *Géométrie différentielle: variétés, courbes et surfaces*. Collection Mathématiques. Puf, second edition, 1992. English edition: Differential geometry, manifolds, curves, and surfaces, GTM No. 115, Springer Verlag.

[18] William M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, second edition, 1986.

[19] Raoul Bott and Tu Loring W. *Differential Forms in Algebraic Topology*. GTM No. 82. Springer Verlag, first edition, 1986.

[20] Nicolas Bourbaki. *Algèbre, Chapitre 9*. Eléments de Mathématiques. Hermann, 1968.

[21] Nicolas Bourbaki. *Algèbre, Chapitres 1-3*. Eléments de Mathématiques. Hermann, 1970.

[22] Nicolas Bourbaki. *Elements of Mathematics. Lie Groups and Lie Algebras, Chapters 1–3*. Springer, first edition, 1989.

[23] Nicolas Bourbaki. *Topologie Générale, Chapitres 1-4*. Eléments de Mathématiques. Masson, 1990.

[24] Nicolas Bourbaki. *Topologie Générale, Chapitres 5-10*. Eléments de Mathématiques. CCLS, 1990.

[25] T. Bröcker and T. tom Dieck. *Representation of Compact Lie Groups*. GTM, Vol. 98. Springer Verlag, first edition, 1985.

[26] R.L. Bryant. An introduction to Lie groups and symplectic geometry. In D.S. Freed and K.K. Uhlenbeck, editors, *Geometry and Quantum Field Theory*, pages 5–181. AMS, Providence, Rhode Island, 1995.

[27] N. Burgoyne and R. Cushman. Conjugacy classes in linear groups. *Journal of Algebra*, 44:339–362, 1977.

[28] Élie Cartan. *Theory of Spinors*. Dover, first edition, 1966.

[29] Henri Cartan. *Cours de Calcul Différentiel*. Collection Méthodes. Hermann, 1990.

[30] Henri Cartan. *Differential Forms*. Dover, first edition, 2006.

[31] Roger Carter, Graeme Segal, and Ian Macdonald. *Lectures on Lie Groups and Lie Algebras*. Cambridge University Press, first edition, 1995.

[32] Sheung H. Cheng, Nicholas J. Higham, Charles Kenney, and Alan J. Laub. Approximating the logarithm of a matrix to specified accuracy. *SIAM Journal on Matrix Analysis and Applications*, 22:1112–1125, 2001.

[33] Shiing-shen Chern. *Complex Manifolds without Potential Theory*. Universitext. Springer Verlag, second edition, 1995.

[34] Claude Chevalley. *Theory of Lie Groups I*. Princeton Mathematical Series, No. 8. Princeton University Press, first edition, 1946. Eighth printing.

[35] Claude Chevalley. *The Algebraic Theory of Spinors and Clifford Algebras. Collected Works, Vol. 2*. Springer, first edition, 1997.

[36] Yvonne Choquet-Bruhat and Cécile DeWitt-Morette. *Analysis, Manifolds, and Physics, Part II: 92 Applications*. North-Holland, first edition, 1989.

[37] Yvonne Choquet-Bruhat, Cécile DeWitt-Morette, and Margaret Dillard-Bleick. *Analysis, Manifolds, and Physics, Part I: Basics*. North-Holland, first edition, 1982.

[38] Lawrence Conlon. *Differentiable Manifolds*. Birkhäuser, second edition, 2001.

[39] Morton L. Curtis. *Matrix Groups*. Universitext. Springer Verlag, second edition, 1984.

[40] James F. Davis and Paul Kirk. *Lecture Notes in Algebraic Topology*. GSM, Vol. 35. AMS, first edition, 2001.

[41] Anton Deitmar. *A First Course in Harmonic Analysis*. UTM. Springer Verlag, first edition, 2002.

[42] C. R. DePrima and C. R. Johnson. The range of $A^{-1}A^*$ in $\mathbf{GL}(n, \mathbf{C})$. *Linear Algebra and Its Applications*, 9:209–222, 1974.

[43] Jean Dieudonné. *Sur les Groupes Classiques*. Hermann, third edition, 1967.

[44] Jean Dieudonné. *Special Functions and Linear Representations of Lie Groups*. Regional Conference Series in Mathematics, No. 42. AMS, first edition, 1980.

[45] Jean Dieudonné. *Éléments d'Analyse, Tome V. Groupes de Lie Compacts, Groupes de Lie Semi-Simples*. Edition Jacques Gabay, first edition, 2003.

[46] Jean Dieudonné. *Éléments d'Analyse, Tome VI. Analyse Harmonique*. Edition Jacques Gabay, first edition, 2003.

[47] Jean Dieudonné. *Éléments d'Analyse, Tome VII. Équations Fonctionnelles Linéaires. Première partie, Opérateurs Pseudo-Différentiels*. Edition Jacques Gabay, first edition, 2003.

[48] Jean Dieudonné. *Éléments d'Analyse, Tome II. Chapitres XII à XV*. Edition Jacques Gabay, first edition, 2005.

[49] Dragomir Djokovic. On the exponential map in classical lie groups. *Journal of Algebra*, 64:76–88, 1980.

[50] Manfredo P. do Carmo. *Differential Geometry of Curves and Surfaces*. Prentice Hall, 1976.

[51] Manfredo P. do Carmo. *Riemannian Geometry*. Birkhäuser, second edition, 1992.

[52] B.A. Dubrovin, A.T. Fomenko, and S.P. Novikov. *Modern Geometry–Methods and Applications. Part I*. GTM No. 93. Springer Verlag, second edition, 1985.

[53] B.A. Dubrovin, A.T. Fomenko, and S.P. Novikov. *Modern Geometry–Methods and Applications. Part II*. GTM No. 104. Springer Verlag, first edition, 1985.

[54] J.J. Duistermaat and J.A.C. Kolk. *Lie Groups*. Universitext. Springer Verlag, first edition, 2000.

[55] Gerald B. Folland. *A Course in Abstract Harmonic Analysis*. CRC Press, first edition, 1995.

[56] Joseph Fourier. *Théorie Analytique de la Chaleur*. Edition Jacques Gabay, first edition, 1822.

[57] William Fulton. *Algebraic Topology, A first course*. GTM No. 153. Springer Verlag, first edition, 1995.

[58] William Fulton and Joe Harris. *Representation Theory, A first course*. GTM No. 129. Springer Verlag, first edition, 1991.

[59] Jean H. Gallier. Logarithms and square roots of real matrices. Technical report, University of Pennsylvania, Levine Hall, 3330 Walnut Street, Philadelphia, PA 19104, 2008. Report No. MS-CIS-08-12.

[60] Jean H. Gallier. *Geometric Methods and Applications, For Computer Science and Engineering*. TAM, Vol. 38. Springer, second edition, 2011.

[61] S. Gallot, D. Hulin, and J. Lafontaine. *Riemannian Geometry*. Universitext. Springer Verlag, second edition, 1993.

[62] F.R. Gantmacher. *The Theory of Matrices, Vol. I.* AMS Chelsea, first edition, 1977.

[63] Christopher Michael Geyer. *Catadioptric Projective Geometry: Theory and Applications.* PhD thesis, University of Pennsylvania, 200 South 33rd Street, Philadelphia, PA 19104, 2002. Dissertation.

[64] André Gramain. *Topologie des Surfaces.* Collection Sup. Puf, first edition, 1971.

[65] Robin Green. Spherical harmonic lighting: The gritty details. In *Archives of the Game Developers' Conference*, pages 1–47, 2003.

[66] Marvin J. Greenberg and John R. Harper. *Algebraic Topology: A First Course.* Addison Wesley, first edition, 1981.

[67] Phillip Griffiths and Joseph Harris. *Principles of Algebraic Geometry.* Wiley Interscience, first edition, 1978.

[68] Cindy M. Grimm. *Modeling Surfaces of Arbitrary Topology Using Manifolds.* PhD thesis, Department of Computer Science, Brown University, Providence, Rhode Island, USA, 1996. Dissertation.

[69] Cindy M. Grimm and John F. Hughes. Modeling surfaces of arbitrary topology using manifolds. In *Proceedings of the 22nd ACM Annual Conference on Computer Graphics and Interactive Techniques (SIGRAPH'95)*, pages 359–368. ACM, August 6-11 1995.

[70] Victor Guillemin and Alan Pollack. *Differential Topology.* Prentice Hall, first edition, 1974.

[71] Brian Hall. *Lie Groups, Lie Algebras, and Representations. An Elementary Introduction.* GTM No. 222. Springer Verlag, first edition, 2003.

[72] Allen Hatcher. *Algebraic Topology.* Cambridge University Press, first edition, 2002.

[73] Sigurdur Helgason. *Groups and Geometric Analysis. Integral Geometry, Invariant Differential Operators and Spherical Functions.* MSM, Vol. 83. AMS, first edition, 2000.

[74] Sigurdur Helgason. *Differential Geometry, Lie Groups, and Symmetric Spaces.* GSM, Vol. 34. AMS, first edition, 2001.

[75] Nicholas J. Higham. The scaling and squaring method of the matrix exponential revisited. *SIAM Journal on Matrix Analysis and Applications*, 26:1179–1193, 2005.

[76] D. Hilbert and S. Cohn-Vossen. *Geometry and the Imagination.* Chelsea Publishing Co., 1952.

[77] Morris W. Hirsch. *Differential Topology.* GTM No. 33. Springer Verlag, first edition, 1976.

[78] Friedrich Hirzebruch. *Topological Methods in Algebraic Geometry*. Springer Classics in Mathematics. Springer Verlag, second edition, 1978.

[79] Harry Hochstadt. *The Functions of Mathematical Physics*. Dover, first edition, 1986.

[80] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, first edition, 1990.

[81] Roger Howe. Very basic Lie theory. *American Mathematical Monthly*, 90:600–623, 1983.

[82] James E. Humphreys. *Introduction to Lie Algebras and Representation Theory*. GTM No. 9. Springer Verlag, first edition, 1972.

[83] Dale Husemoller. *Fiber Bundles*. GTM No. 20. Springer Verlag, third edition, 1994.

[84] Jürgen Jost. *Riemannian Geometry and Geometric Analysis*. Universitext. Springer Verlag, fourth edition, 2005.

[85] Charles S. Kenney and Alan J. Laub. Condition estimates for matrix functions. *SIAM Journal on Matrix Analysis and Applications*, 10:191–209, 1989.

[86] A.A. Kirillov. Spinor representations of orthogonal groups. Technical report, University of Pennsylvania, Math. Department, Philadelphia, PA 19104, 2001. Course Notes for Math 654.

[87] A.A. Kirillov. *Lectures on the Orbit Method*. GSM, Vol. 64. AMS, first edition, 2004.

[88] A.A. Kirillov (Ed.). *Representation Theory and Noncommutative Harmonic Analysis*. Encyclopaedia of Mathematical Sciences, Vol. 22. Springer Verlag, first edition, 1994.

[89] Wilhelm Klingenberg. *Riemannian Geometry*. de Gruyter & Co, second edition, 1995.

[90] Anthony W. Knapp. *Lie Groups Beyond an Introduction*. Progress in Mathematics, Vol. 140. Birkhäuser, second edition, 2002.

[91] Shoshichi Kobayashi and Katsumi Nomizu. *Foundations of Differential Geometry, II*. Wiley Classics. Wiley-Interscience, first edition, 1996.

[92] Wolfgang Kühnel. *Differential Geometry. Curves–Surfaces–Manifolds*. Student Mathematical Library, Vol. 16. AMS, first edition, 2002.

[93] Jacques Lafontaine. *Introduction Aux Variétés Différentielles*. PUG, first edition, 1996.

[94] Serge Lang. *Real and Functional Analysis*. GTM 142. Springer Verlag, third edition, 1996.

[95] Serge Lang. *Undergraduate Analysis*. UTM. Springer Verlag, second edition, 1997.

[96] Serge Lang. *Fundamentals of Differential Geometry*. GTM No. 191. Springer Verlag, first edition, 1999.

[97] Blaine H. Lawson and Marie-Louise Michelsohn. *Spin Geometry*. Princeton Math. Series, No. 38. Princeton University Press, 1989.

[98] N. N. Lebedev. *Special Functions and Their Applications*. Dover, first edition, 1972.

[99] John M. Lee. *Introduction to Smooth Manifolds*. GTM No. 218. Springer Verlag, first edition, 2006.

[100] Pertti Lounesto. *Clifford Algebras and Spinors*. LMS No. 286. Cambridge University Press, second edition, 2001.

[101] Ib Madsen and Jorgen Tornehave. *From Calculus to Cohomology. De Rham Cohomology and Characteristic Classes*. Cambridge University Press, first edition, 1998.

[102] Paul Malliavin. *Géométrie Différentielle Intrinsèque*. Enseignement des Sciences, No. 14. Hermann, first edition, 1972.

[103] Jerrold E. Marsden and T.S. Ratiu. *Introduction to Mechanics and Symmetry*. TAM, Vol. 17. Springer Verlag, first edition, 1994.

[104] William S. Massey. *Algebraic Topology: An Introduction*. GTM No. 56. Springer Verlag, second edition, 1987.

[105] William S. Massey. *A Basic Course in Algebraic Topology*. GTM No. 127. Springer Verlag, first edition, 1991.

[106] Yukio Matsumoto. *An Introduction to Morse Theory*. Translations of Mathematical Monographs No 208. AMS, first edition, 2002.

[107] John W. Milnor. *Morse Theory*. Annals of Math. Series, No. 51. Princeton University Press, third edition, 1969.

[108] John W. Milnor. On isometries of inner product spaces. *Inventiones Mathematicae*, 8:83–97, 1969.

[109] John W. Milnor. *Topology from the Differentiable Viewpoint*. The University Press of Virginia, second edition, 1969.

[110] John W. Milnor. Curvatures of left invariant metrics on lie groups. *Advances in Mathematics*, 21:293–329, 1976.

[111] John W. Milnor and James D. Stasheff. *Characteristic Classes*. Annals of Math. Series, No. 76. Princeton University Press, first edition, 1974.

[112] R. Mneimné and F. Testard. *Introduction à la Théorie des Groupes de Lie Classiques*. Hermann, first edition, 1997.

[113] Jules Molk. *Encyclopédie des Sciences Mathématiques Pures et Appliquées. Tome I (premier volume), Arithmétique*. Gauthier-Villars, first edition, 1916.

[114] Mitsuo Morimoto. *Analytic Functionals on the Sphere*. Translations of Mathematical Monographs No 178. AMS, first edition, 1998.

[115] Shigeyuki Morita. *Geometry of Differential Forms*. Translations of Mathematical Monographs No 201. AMS, first edition, 2001.

[116] James R. Munkres. *Topology, a First Course*. Prentice Hall, first edition, 1975.

[117] James R. Munkres. *Analysis on Manifolds*. Addison Wesley, 1991.

[118] Raghavan Narasimham. *Compact Riemann Surfaces*. Lecture in Mathematics ETH Zürich. Birkhäuser, first edition, 1992.

[119] Mitsuru Nishikawa. On the exponential map of the group $\mathbf{O}(p,q)_0$. *Memoirs of the Faculty of Science, Kyushu University, Ser. A*, 37:63–69, 1983.

[120] Barrett O'Neill. *Semi-Riemannian Geometry With Applications to Relativity*. Pure and Applies Math., Vol 103. Academic Press, first edition, 1983.

[121] Xavier Pennec. Intrinsic statistics on Riemannian Manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25:127–154, 2006.

[122] Peter Petersen. *Riemannian Geometry*. GTM No. 171. Springer Verlag, second edition, 2006.

[123] L. Pontryagin. *Topological Groups*. Princeton University Press, first edition, 1939.

[124] L. Pontryagin. *Topological Groups*. Gordon and Breach, second edition, 1960.

[125] Ian R. Porteous. *Topological Geometry*. Cambridge University Press, second edition, 1981.

[126] M.M. Postnikov. *Geometry VI. Riemannian Geometry*. Encyclopaedia of Mathematical Sciences, Vol. 91. Springer Verlag, first edition, 2001.

[127] Marcel Riesz. *Clifford Numbers and Spinors*. Kluwer Academic Press, first edition, 1993. Edited by E. Folke Bolinder and Pertti Lounesto.

[128] Wulf Rossmann. *Lie Groups. An Introduction Through Linear Groups*. Graduate Texts in Mathematics. Oxford University Press, first edition, 2002.

[129] Joseph J. Rotman. *Introduction to Algebraic Topology*. GTM No. 119. Springer Verlag, first edition, 1988.

[130] Arthur A. Sagle and Ralph E. Walde. *Introduction to Lie Groups and Lie Algebras*. Academic Press, first edition, 1973.

[131] Takashi Sakai. *Riemannian Geometry*. Mathematical Monographs No 149. AMS, first edition, 1996.

[132] Hans Samelson. *Notes on Lie Algebras*. Universitext. Springer, second edition, 1990.

[133] Giovanni Sansone. *Orthogonal Functions*. Dover, first edition, 1991.

[134] Hajime Sato. *Algebraic Topology: An Intuitive Approach*. Mathematical Monographs No 183. AMS, first edition, 1999.

[135] D.H. Sattinger and O.L. Weaver. *Lie Groups and Algebras with Applications to Physics, Geometry, and Mechanics*. Applied Math. Science, Vol. 61. Springer Verlag, first edition, 1986.

[136] Laurent Schwartz. *Analyse II. Calcul Différentiel et Equations Différentielles*. Collection Enseignement des Sciences. Hermann, 1992.

[137] Jean-Pierre Serre. *Lie Algebras and Lie Groups*. Lecture Notes in Mathematics, No. 1500. Springer, second edition, 1992.

[138] Jean-Pierre Serre. *Complex Semisimple Lie Algebras*. Springer Monographs in Mathematics. Springer, first edition, 2000.

[139] Igor R. Shafarevich. *Basic Algebraic Geometry 1*. Springer Verlag, second edition, 1994.

[140] Richard W. Sharpe. *Differential Geometry. Cartan's Generalization of Klein's Erlangen Program*. GTM No. 166. Springer Verlag, first edition, 1997.

[141] Marcelo Siqueira, Dianna Xu, and Jean Gallier. Construction of C^∞-surfaces from triangular meshes using parametric pseudo-manifolds. Technical report, University of Pennsylvania, Levine Hall, Philadelphia, PA 19104, 2008. pdf file available from http://repository.upenn.edu/cis_reports/877.

[142] Norman Steenrod. *The Topology of Fibre Bundles*. Princeton Math. Series, No. 14. Princeton University Press, 1956.

[143] Elias M. Stein and Guido Weiss. *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton Math. Series, No. 32. Princeton University Press, 1971.

[144] S. Sternberg. *Lectures On Differential Geometry*. AMS Chelsea, second edition, 1983.

[145] Michael E. Taylor. *Noncommutative Harmonic Analysis*. Mathematical Surveys and Monographs, No. 22. AMS, first edition, 1986.

[146] Loring W. Tu. *An Introduction to Manifolds*. Universitext. Springer Verlag, first edition, 2008.

[147] N.J. Vilenkin. *Special Functions and the Theory of Group Representations*. Translations of Mathematical Monographs No 22. AMS, first edition, 1968.

[148] Frank Warner. *Foundations of Differentiable Manifolds and Lie Groups*. GTM No. 94. Springer Verlag, first edition, 1983.

[149] André Weil. *Foundations of Algebraic Topology*. Colloquium Publications, Vol. XXIX. AMS, second edition, 1946.

[150] André Weil. *L'Intégration dans les Groupes Topologiques et ses Applications*. Hermann, second edition, 1979.

[151] R.O. Wells. *Differential Analysis on Complex Manifolds*. GTM No. 65. Springer Verlag, second edition, 1980.

[152] Hermann Weyl. *The Classical Groups. Their Invariants and Representations*. Princeton Mathematical Series, No. 1. Princeton University Press, second edition, 1946.