Universitext

Jürgen Jost

Riemannian Geometry and Geometric Analysis

Fifth Edition

Springer

Universitext

Jürgen Jost

Riemannian Geometry and Geometric Analysis

Fifth Edition

Springer

Jürgen Jost
Max Planck Institute for Mathematics in the Sciences
Inselstr. 22
04103 Leipzig
Germany
jjost@mis.mpg.de

ISBN 978-3-540-77340-5 e-ISBN 978-3-540-77341-2

Universitext

Library of Congress Control Number: 2008921366

Mathematics Subject Classification (2000): 53B21, 53L20, 32C17, 35I60, 49-XX, 58E20, 57R15

Cover design: *design & production* GmbH, Heidelberg

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

**Dedicated to Shing-Tung Yau,
for so many discussions about
mathematics and Chinese culture**

Preface

Riemannian geometry is characterized, and research is oriented towards and shaped by concepts (geodesics, connections, curvature, ...) and objectives, in particular to understand certain classes of (compact) Riemannian manifolds defined by curvature conditions (constant or positive or negative curvature, ...). By way of contrast, geometric analysis is a perhaps somewhat less systematic collection of techniques, for solving extremal problems naturally arising in geometry and for investigating and characterizing their solutions. It turns out that the two fields complement each other very well; geometric analysis offers tools for solving difficult problems in geometry, and Riemannian geometry stimulates progress in geometric analysis by setting ambitious goals.

It is the aim of this book to be a systematic and comprehensive introduction to Riemannian geometry and a representative introduction to the methods of geometric analysis. It attempts a synthesis of geometric and analytic methods in the study of Riemannian manifolds.

The present work is the fifth edition of my textbook on Riemannian geometry and geometric analysis. It has developed on the basis of several graduate courses I taught at the Ruhr-University Bochum and the University of Leipzig. The main new features of the present edition are the systematic inclusion of flow equations and a mathematical treatment of the nonlinear sigma model of quantum field theory. These new topics also led to a systematic reorganization of the other material. Naturally, I have also included several smaller additions and minor corrections (for which I am grateful to several readers).

Let me now briefly describe the contents:

In the first chapter, we introduce the basic geometric concepts, like differentiable manifolds, tangent spaces, vector bundles, vector fields and one-parameter groups of diffeomorphisms, Lie algebras and groups and in particular Riemannian metrics. We also treat the existence of geodesics with two different methods, both of which are quite important in geometric analysis in general. Thus, the reader has the opportunity to understand the basic ideas of those methods in an elementary context before moving on to more difficult versions in subsequent chapters. The first method is based on the local existence and uniqueness of geodesics and will be applied again in Chapter 8 for two-dimensional harmonic maps. The second method is the heat flow method that gained prominence through Perelman's solution of the Poincaré conjecture by the Ricci flow method.

The second chapter introduces de Rham cohomology groups and the essential tools from elliptic PDE for treating these groups. We prove the existence of harmonic forms representing cohomology classes both by a variational method, thereby introducing another of the basic schemes of geometric analysis, and by the heat flow method. The linear setting of cohomology classes allows us to understand some key ideas without the technical difficulties of nonlinear problems.

The third chapter treats the general theory of connections and curvature.

In the fourth chapter, we introduce Jacobi fields, prove the Rauch comparison theorems for Jacobi fields and apply these results to geodesics. We also develop the global geometry of spaces of nonpositive curvature.

These first four chapters treat the more elementary and basic aspects of the subject. Their results will be used in the remaining, more advanced chapters.

The fifth chapter treats Kähler manifolds symmetric spaces as important examples of Riemannian manifolds in detail.

The sixth chapter is devoted to Morse theory and Floer homology.

In the seventh chapter, we treat harmonic maps between Riemannian manifolds. We prove several existence theorems and apply them to Riemannian geometry. The treatment uses an abstract approach based on convexity that should bring out the fundamental structures. We also display a representative sample of techniques from geometric analysis.

In the eighth chapter, we treat harmonic maps from Riemann surfaces. We encounter here the phenomenon of conformal invariance which makes this two-dimensional case distinctively different from the higher dimensional one.

The ninth chapter treats variational problems from quantum field theory, in particular the Ginzburg-Landau, Seiberg-Witten equations, and a mathematical version of the nonlinear supersymmetric sigma model. In mathematical terms, the two-dimensional harmonic map problem is coupled with a Dirac field. The background material on spin geometry and Dirac operators is already developed in earlier chapters. The connections between geometry and physics will be further explored in a forthcoming monograph [144].

A guiding principle for this textbook was that the material in the main body should be self contained. The essential exception is that we use material about Sobolev spaces and linear elliptic an parabolic PDEs without giving proofs. This material is collected in Appendix A. Appendix B collects some elementary topological results about fundamental groups and covering spaces.

Also, in certain places in Chapter 6, we do not present all technical details, but rather explain some points in a more informal manner, in order to keep the size of that chapter within reasonable limits and not to loose the patience of the readers.

We employ both coordinate free intrinsic notations and tensor notations depending on local coordinates. We usually develop a concept in both notations while we sometimes alternate in the proofs. Besides not being a methodological purist, reasons for often prefering the tensor calculus to the more elegant and concise intrinsic one are the following. For the analytic aspects, one often has to employ results about (elliptic) partial differential equations (PDEs), and in order to check that the relevant assumptions like ellipticity hold and in order to make contact with the notations usually employed in PDE theory, one has to write down the differential equation in local coordinates. Also, manifold and important connections have been established between theoretical physics and our subject. In the physical literature, usually the tensor notation is employed, and therefore, familiarity with that notation is necessary for exploring those connections that have been found to be stimulating for the development of mathematics, or promise to be so in the future.

As appendices to most of the paragraphs, we have written sections with the title "Perspectives". The aim of those sections is to place the material in a broader context and explain further results and directions without detailed proofs. The material of these Perspectives will not be used in the main body of the text. Similarly, after Chapter 4, we have inserted a section entitled "A short survey on curvature and topology" that presents an account of many global results of Riemannian geometry not covered in the main text. – At the end of each chapter, some exercises for the reader are given. We assume of the reader sufficient perspicacity to understand our system of numbering and cross-references without further explanations.

The development of the mathematical subject of Geometric Analysis, namely the investigation of analytical questions arising from a geometric context and in turn the application of analytical techniques to geometric problems, is to a large extent due to the work and the influence of Shing-Tung Yau. This book, like its previous editions, is dedicated to him.

I am also grateful to Minjie Chen for dedicated help with the Tex file.

Jürgen Jost

Contents

Chapter 1

Foundational Material

1.1 Manifolds and Differentiable Manifolds

A *topological space* is a set M together with a family $\mathcal{O}$ of subsets of M satisfying the following properties:

(i) $\Omega_1, \Omega_2 \in \mathcal{O} \Rightarrow \Omega_1 \cap \Omega_2 \in \mathcal{O}$,

(ii) for any index set $A : (\Omega_\alpha)_{\alpha \in A} \subset \mathcal{O} \Rightarrow \bigcup_{\alpha \in A} \Omega_\alpha \in \mathcal{O}$,

(iii) $\emptyset, M \in \mathcal{O}$.

The sets from $\mathcal{O}$ are called *open*. A topological space is called *Hausdorff* if for any two distinct points $p_1, p_2 \in M$ there exists open sets $\Omega_1, \Omega_2 \in \mathcal{O}$ with $p_1 \in \Omega_1, p_2 \in \Omega_2, \Omega_1 \cap \Omega_2 = \emptyset$. A covering $(\Omega_\alpha)_{\alpha \in A}$ (A an arbitrary index set) is called *locally finite* if each $p \in M$ has a neighborhood that intersects only finitely many Ω_α. M is called *paracompact* if any open covering possesses a locally finite refinement. This means that for any open covering $(\Omega_\alpha)_{\alpha \in A}$ there exists a locally finite open covering $(\Omega'_\beta)_{\beta \in B}$ with

$$\forall \beta \in B \; \exists \alpha \in A : \Omega'_\beta \subset \Omega_\alpha.$$

A map between topological spaces is called *continuous* if the preimage of any open set is again open. A bijective map which is continuous in both directions is called a *homeomorphism*.

Definition 1.1.1. A *manifold* M of *dimension* d is a connected paracompact Hausdorff space for which every point has a neighborhood U that is homeomorphic to an open subset Ω of $\mathbb{R}^d$. Such a homeomorphism

$$x : U \to \Omega$$

is called a *(coordinate) chart.*

An *atlas* is a family $\{U_\alpha, x_\alpha\}$ of charts for which the U_α constitute an open covering of M.

Remarks.

1. A point $p \in U_\alpha$ is determined by $x_\alpha(p)$; hence it is often identified with $x_\alpha(p)$. Often, also the index α is omitted, and the components of $x(p) \in \mathbb{R}^d$ are called *local coordinates* of p.

2. Any atlas is contained in a maximal one, namely the one consisting of all charts compatible with the original one.

As we shall see, local coordinates yield a systematic method for locally representing a manifold in such a manner that computations can be carried out. We shall now describe a concept that will allow us to utilize the framework of linear algebra for local computations as will be explored in 1.2 and beyond.

Definition 1.1.2. An atlas $\{U_\alpha, x_\alpha\}$ on a manifold is called *differentiable* if all chart transitions

$$x_\beta \circ x_\alpha^{-1} : x_\alpha(U_\alpha \cap U_\beta) \to x_\beta(U_\alpha \cap U_\beta)$$

are differentiable of class C^∞ (in case $U_\alpha \cap U_\beta \neq \emptyset$). A maximal differentiable atlas is called a differentiable structure, and a *differentiable manifold* of dimension d is a manifold of dimension d with a differentiable structure. From now on, all atlases are supposed to be differentiable. Two atlases are called compatible if their union is again an atlas. In general, a chart is called compatible with an atlas if adding the chart to the atlas yields again an atlas. An atlas is called maximal if any chart compatible with it is already contained in it.

Remarks.

1. Since the inverse of $x_\beta \circ x_\alpha^{-1}$ is $x_\alpha \circ x_\beta^{-1}$, chart transitions are differentiable in both directions, i.e. diffeomorphisms.

2. One could also require a weaker differentiability property than C^∞.

3. It is easy to show that the dimension of a differentiable manifold is uniquely determined. For a general, not differentiable manifold, this is much harder.

4. Since any differentiable atlas is contained in a maximal differentiable one, it suffices to exhibit some differentiable atlas if one wants to construct a differentiable manifold.

Definition 1.1.3. An atlas for a differentiable manifold is called *oriented* if all chart transitions have positive functional determinant. A differentiable manifold is called *orientable* if it possesses an oriented atlas.

It is customary to write the Euclidean coordinates of $\mathbb{R}^d, \Omega \subset \mathbb{R}^d$ open, as

$$x = (x^1, \dots, x^d), \tag{1.1.1}$$

and these then are considered as local coordinates on our manifold M when $x : U \to \Omega$ is a chart.

Examples.

1. The *sphere* $S^n := \{(x^1, \dots, x^{n+1}) \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} (x^i)^2 = 1\}$ is a differentiable manifold of dimension n. Charts can be given as follows: On $U_1 := S^n \backslash \{(0, \dots, 0, 1)\}$ we put

$$\begin{aligned} f_1(x^1, \dots, x^{n+1}) &:= (f_1^1(x^1, \dots, x^{n+1}), \dots, f_1^n(x^1, \dots, x^{n+1})) \\ &:= \left(\frac{x^1}{1 - x^{n+1}}, \dots, \frac{x^n}{1 - x^{n+1}} \right) \end{aligned}$$

   and on $U_2 := S^n \backslash \{(0, \dots, 0, -1)\}$

$$\begin{aligned} f_2(x^1, \dots, x^{n+1}) &:= (f_2^1(x^1, \dots, x^{n+1}), \dots, f_2^n(x^1, \dots, x^{n+1})) \\ &:= \left(\frac{x^1}{1 + x^{n+1}}, \dots, \frac{x^n}{1 + x^{n+1}} \right). \end{aligned}$$

2. Let $w_1, w_2, \dots, w_n \in \mathbb{R}^n$ be linearly independent. We consider $z_1, z_2 \in \mathbb{R}^n$ as equivalent if there are $m_1, m_2, \dots, m_n \in \mathbb{Z}$ with

$$z_1 - z_2 = \sum_{i=1}^{n} m_i w_i.$$

   Let π be the projection mapping $z \in \mathbb{R}^n$ to its equivalence class. The *torus* $T^n := \pi(\mathbb{R}^n)$ can then be made a differentiable manifold (of dimension n) as follows: Suppose Δ_α is open and does not contain any pair of equivalent points. We put

$$\begin{aligned} U_\alpha &:= \pi(\Delta_\alpha), \\ z_\alpha &:= (\pi_{|\Delta_\alpha})^{-1}. \end{aligned}$$

3. The preceding examples are compact. Of course, there exist also noncompact manifolds. The simplest example is $\mathbb{R}^d$. In general, any open subset of a (differentiable) manifold is again a (differentiable) manifold.

4. If M and N are differentiable manifolds, the Cartesian product $M \times N$ also naturally carries the structure of a differentiable manifold. Namely, if $\{U_\alpha, x_\alpha\}_{\alpha \in A}$ and $\{V_\beta, y_\beta\}_{\beta \in B}$ are atlases for M and N, resp., then $\{U_\alpha \times V_\beta, (x_\alpha, y_\beta)\}_{(\alpha,\beta) \in A \times B}$ is an atlas for $M \times N$ with differentiable chart transitions.

Definition 1.1.4. A map $h : M \to M'$ between differentiable manifolds M and M' with charts $\{U_\alpha, x_\alpha\}$ and $\{U'_\alpha, x'_\alpha\}$ is called *differentiable* if all maps $x'_\beta \circ h \circ x_\alpha^{-1}$ are differentiable (of class C^∞, as always) where defined. Such a map is called a *diffeomorphism* if bijective and differentiable in both directions.

For purposes of differentiation, a differentiable manifold locally has the structure of Euclidean space. Thus, the differentiability of a map can be tested in local coordinates. The diffeomorphism requirement for the chart transitions then guarantees that differentiability defined in this manner is a consistent notion, i.e. independent of the choice of a chart.

Remark. We want to point out that in the context of the preceding definitions, one cannot distinguish between two homeomorphic manifolds nor between two diffeomorphic differentiable manifolds.

When looking at Definitions 1.1.2, 1.1.3, one may see a general pattern emerging. Namely, one can put any type of restriction on the chart transitions, for example, require them to be affine, algebraic, real analytic, conformal, Euclidean volume preserving,..., and thereby define a class of manifolds with that particular structure. Perhaps the most important example is the notion of a complex manifold. We shall need this, however, only at certain places in this book, namely in §5.1, §5.2.

Definition 1.1.5. A *complex manifold* of complex dimension d ($\dim_{\mathbb{C}} M = d$) is a differentiable manifold of (real) dimension $2d$ ($\dim_{\mathbb{R}} M = 2d$) whose charts take values in open subsets of $\mathbb{C}^d$ with *holomorphic* chart transitions.

In the case of a complex manifold, it is customary to write the coordinates of $\mathbb{C}^d$ as

$$z = (z^1, \dots, z^d), \qquad \text{with} \qquad z^j = x^j + iy^j \tag{1.1.2}$$

with $i := \sqrt{-1}$, that is, use $(x^1, y^1, \dots, x^d, y^d)$ as Euclidean coordinates on $\mathbb{R}^{2d}$. We then also put

$$z^{\bar{j}} := x^j - iy^j.$$

The requirement that the chart transitions $z_\beta \circ z_\alpha^{-1} : z_\alpha(U_\alpha \cap U_\beta) \to z_\beta(U_\alpha \cap U_\beta)$ be holomorphic then is expressed as

$$\frac{\partial z_\beta^j}{\partial z_\alpha^{\bar{k}}} = 0 \tag{1.1.3}$$

for all j, k where

$$\frac{\partial}{\partial z^{\bar{k}}} = \frac{1}{2}\left(\frac{\partial}{\partial x^{\bar{k}}} + i\frac{\partial}{\partial y^{\bar{k}}}\right). \tag{1.1.4}$$

We also observe that a complex manifold is always orientable because holomorphic maps always have a positive functional determinant.

We conclude this section with a useful technical result.

Lemma 1.1.1. *Let M be a differentiable manifold, $(U_\alpha)|_{\alpha \in A}$ an open covering. Then there exists a partition of unity, subordinate to (U_α). This means that there exists a locally finite refinement $(V_\beta)_{\beta \in B}$ of (U_α) and C_0^∞ (i.e. C^∞ functions φ_β with $\{x \in M : \varphi_\beta(x) \neq 0\}$ having compact closure) functions $\varphi_\beta : M \to \mathbb{R}$ with*

(i) $\operatorname{supp} \varphi_\beta \subset V_\beta$ *for all* $\beta \in B$,

(ii) $0 \leq \varphi_\beta(x) \leq 1$ *for all* $x \in M, \beta \in B$,

(iii) $\sum_{\beta \in B} \varphi_\beta(x) = 1$ *for all* $x \in M$.

Note that in (iii), there are only finitely many nonvanishing summands at each point since only finitely many φ_β are nonzero at any given point because the covering (V_β) is locally finite.

Proof. See any advanced textbook on Analysis, e.g. J. Jost, Postmodern Analysis, 3rd ed., Springer, 2005. □

Perspectives. Like so many things in Riemannian geometry, the concept of a differentiable manifold was in some vague manner implicitly contained in Bernhard Riemann's habilitation address "Über die Hypothesen, welche der Geometrie zugrunde liegen", reprinted in [262]. The first clear formulation of that concept, however, was given by H. Weyl[260].

The only one dimensional manifolds are the real line and the unit circle S^1, the latter being the only compact one. Two dimensional compact manifolds are classified by their genus and orientability character. In three dimensions, Thurston[251, 252] had proposed a program for the possible classification of compact three-dimensional manifolds. This could recently be resolved by Perel'man with techniques from geometric analysis (that were rather different from those that Thurston had developed); see the Survey on Curvature and Topology in the middle of this book for references. – In higher dimensions, the plethora of compact manifolds makes a classification useless and impossible.

In dimension at most three, each manifold carries a unique differentiable structure, and so here the classifications of manifolds and differentiable manifolds coincide. This is no longer so in higher dimensions. Milnor[189, 190] discovered exotic 7-spheres, i.e. differentiable structures on the manifold S^7 that are not diffeomorphic to the standard differentiable structure exhibited in our example. Exotic spheres likewise exist in higher dimensions. Kervaire[164] found an example of a manifold carrying no differentiable structure at all.

In dimension 4, the understanding of differentiable structures owes important progress to the work of Donaldson. He defined invariants of a differentiable 4-manifold M from the space of selfdual connections on principal bundles over it. These concepts will be discussed in more detail in §3.2.

In particular, there exist exotic structures on $\mathbb{R}^4$. A description can e.g. be found in [86].

1.2 Tangent Spaces

Let $x = (x^1, \dots, x^d)$ be Euclidean coordinates of $\mathbb{R}^d, \Omega \subset \mathbb{R}^d$ open, $x_0 \in \Omega$. The tangent space of Ω at the point x_0,

$$T_{x_0}\Omega$$

is the space $\{x_0\} \times E$, where E is the d-dimensional vector space spanned by the basis $\frac{\partial}{\partial x^1}, \dots, \frac{\partial}{\partial x^d}$. Here, $\frac{\partial}{\partial x^1}, \dots, \frac{\partial}{\partial x^d}$ are the partial derivatives at the point x_0. If $\Omega \subset \mathbb{R}^d, \Omega' \subset \mathbb{R}^c$ are open, and $f : \Omega \to \Omega'$ is differentiable, we define the *derivative* $df(x_0)$ for $x_0 \in \Omega$ as the induced linear map between the tangent spaces

$$\begin{aligned} df(x_0) : T_{x_0}\Omega &\to T_{f(x_0)}\Omega', \\ v = v^i \frac{\partial}{\partial x^i} &\mapsto v^i \frac{\partial f^j}{\partial x^i}(x_0) \frac{\partial}{\partial f^j}. \end{aligned}$$

Here and in the sequel, we use the *Einstein summation convention*: An index occuring twice in a product is to be summed from 1 up to the space dimension. Thus, $v^i \frac{\partial}{\partial x^i}$ is an abbreviation for

$$\sum_{i=1}^{d} v^i \frac{\partial}{\partial x^i},$$

$v^i \frac{\partial f^j}{\partial x^i} \frac{\partial}{\partial f^j}$ stands for

$$\sum_{i=1}^{d} \sum_{j=1}^{c} v^i \frac{\partial f^j}{\partial x^i} \frac{\partial}{\partial f^j}.$$

In the previous notations, we put

$$T\Omega := \Omega \times E \cong \Omega \times \mathbb{R}^d.$$

Thus, $T\Omega$ is an open subset of $\mathbb{R}^d \times \mathbb{R}^d$, hence in particular a differentiable manifold.

$$\begin{aligned} \pi : T\Omega &\to \Omega, \quad \text{(projection onto the first factor)} \\ (x, v) &\mapsto x \end{aligned}$$

is called a tangent bundle of Ω. $T\Omega$ is called the total space of the tangent bundle.

Likewise, we define

$$\begin{aligned} df : T\Omega &\to T\Omega', \\ \left(x, v^i \frac{\partial}{\partial x^i}\right) &\mapsto \left(f(x), v^i \frac{\partial f^j}{\partial x^i}(x) \frac{\partial}{\partial f^j}\right). \end{aligned}$$

Instead of

$$df(x, v)$$

we write

$$df(x)(v).$$

If in particular, $f : \Omega \to \mathbb{R}$ is a differentiable function, we have for $v = v^i \frac{\partial}{\partial x^i}$

$$df(x)(v) = v^i \frac{\partial f}{\partial x^i}(x) \in T_{f(x)}\mathbb{R} \cong \mathbb{R}.$$

In this case, we often write $v(f)(x)$ in place of $df(x)(v)$ when we want to express that the tangent vector v operates by differentiation on the function f.

Let now M be a differentiable manifold of dimension d, and $p \in M$. We want to define the tangent space of M at the point p. Let $x : U \to \mathbb{R}^d$ be a chart with $p \in U$, U open in M. We say that the tangent space T_pM is represented in the chart x by $T_{x(p)}x(U)$. Let $x' : U' \to \mathbb{R}^d$ be another chart with $p \in U'$, U' open in M. $\Omega := x(U), \Omega' := x'(U')$. The transition map

$$x' \circ x^{-1} : x(U \cap U') \to x'(U \cap U')$$

induces a vector space isomorphism

$$L := d(x' \circ x^{-1})(x(p)) : T_{x(p)}\Omega \to T_{x'(p)}\Omega'.$$

We say that $v \in T_{x(p)}\Omega$ and $L(v) \in T_{x'(p)}\Omega'$ represent the same tangent vector in T_pM. Thus, a tangent vector in T_pM is given by the family of its coordinate representations. This is motivated as follows: Let $f : M \to \mathbb{R}$ be a differentiable function. Assume that the tangent vector $w \in T_pM$ is represented by $v \in T_{x(p)}x(U)$. We then want to define $df(p)$ as a linear map from T_pM to $\mathbb{R}$. In the chart x, let $w \in T_pM$ be represented by $v = v^i \frac{\partial}{\partial x^i} \in T_{x(p)}x(U)$. We then say that

$$df(p)(w)$$

in this chart is represented by

$$d(f \circ x^{-1})(x(p))(v).$$

Now

$$\begin{aligned} d(f \circ x^{-1})(x(p))(v) &= d(f \circ x'^{-1} \circ x' \circ x^{-1})(x(p))(v) \\ &= d(f \circ x'^{-1})(x'(p))(L(v)) \quad \text{by the chain rule} \\ &= d(f \circ x'^{-1})(x'(p)) \circ d(x' \circ x^{-1})(x(p))(v). \end{aligned}$$

Thus, in the chart x', w is represented by $L(v)$. Here, a fundamental idea emerges that will be essential for the understanding of the sequel. T_pM is a vector space of dimension d, hence isomorphic to $\mathbb{R}^d$. This isomorphism, however, is not canonical, but depends on the choice of a chart. A change of charts changes the isomorphism, namely at the point p by the linear transformation $L = d(x' \circ x^{-1})(x(p))$. Under a change of charts, also other objects then are correspondingly transformed, for example derivatives of functions, or more generally of maps. In other words, a chart yields local representations for tangent vectors, derivatives, etc., and under a change of charts, these local representations need to be correctly transformed. Or in still other words: We know how to differentiate (differentiable) functions that are defined on open subsets of $\mathbb{R}^d$. If now a function is given on a manifold, we pull it back by a chart, to an open subset of $\mathbb{R}^d$ and then differentiate the pulled back function. In order to obtain an object that does not depend on the choice of chart, we have to know in addition the transformation behavior under chart changes. A tangent vector thus is determined by how it operates on functions by differentiation.

Likewise, for a differentiable map $F : M \to N$ between differentiable manifolds, dF is represented in local charts $x : U \subset M \to \mathbb{R}^d$, $y : V \subset N \to \mathbb{R}^c$ by

$$d(y \circ F \circ x^{-1}).$$

In the sequel, in our notation, we shall frequently drop reference to the charts and write instead of $d(y \circ F \circ x^{-1})$ simply dF, provided the choice of charts or at least the fact that charts have been chosen is obvious from the context. We can achieve this most simply as follows:
Let the local coordinates on U be

$$(x^1, \dots, x^d),$$

and those on V be $(F^1, \dots, F^c)$. We then consider $F(x)$ as abbreviation for

$$(F^1(x^1, \dots, x^d), \dots, F^c(x^1, \dots, x^d)).$$

dF now induces a linear map

$$dF : T_x M \to T_{F(x)} N,$$

which in our coordinates is represented by the matrix

$$\left(\frac{\partial F^\alpha}{\partial x^i} \right)_{\substack{\alpha=1,\dots,c \\ i=1,\dots,d}}.$$

A change of charts leads to a base change of the tangent spaces, and the transformation behavior is determined by the chain rule. If

$$\begin{aligned} (x^1, \dots, x^d) &\mapsto (\xi^1, \dots, \xi^d) \\ \text{and } (F^1, \dots, F^c) &\mapsto (\Phi^1, \dots, \Phi^c) \end{aligned}$$

are coordinate changes, then dF is represented in the new coordinates by

$$\left(\frac{\partial \Phi^\beta}{\partial \xi^j} \right) = \left(\frac{\partial \Phi^\beta}{\partial F^\alpha} \frac{\partial F^\alpha}{\partial x^i} \frac{\partial x^i}{\partial \xi^j} \right).$$

Note that the functional matrix of the coordinate change of the image N, but the inverse of the functional matrix of the coordinate change of the domain M appears here. We also remark that for a function $\varphi : N \to \mathbb{R}$ and a $v \in T_x M$,

$$(dF(v)(\varphi))(F(x)) := d\varphi(dF(v))(F(x))$$

by definition of the application of $dF(v) \in T_{F(x)}N$ to $\varphi : N \to \mathbb{R}$,

$$\begin{aligned} &= d(\varphi \circ F)(v)(x) \text{ by the chain rule} \\ &= v(\varphi \circ F)(x) \end{aligned}$$

by definition of the application of $v \in T_x M$ to $\varphi \circ F : M \to \mathbb{R}$.

Instead of applying the tangent vector $dF(v)$ to the function, one may also apply the tangent vector v to the "pulled back" function $\varphi \circ F$.

We want to collect the previous considerations in a formal definition:

Definition 1.2.1. Let $p \in M$. On $\{(x, v) : x : U \to \Omega \text{ chart with } p \in U, v \in T_{x(p)}\Omega\}$ $(x, v) \sim (y, w) :\iff w = d(y \circ x^{-1})v$. The space of equivalence classes is called the *tangent space* to M at the point p, and it is denoted by $T_p M$.

$T_p M$ naturally carries the structure of a vector space:

The equivalence class of $\lambda_1(x, v_1) + \lambda_2(x, v_2)$ $(\lambda_1, \lambda_2 \in \mathbb{R})$ is the one of $(x, \lambda_1 v_1 + \lambda_2 v_2)$. We now want to define the tangent bundle of a differentiable manifold of dimension d. TM is the disjoint union of the tangent spaces $T_p M, p \in M$, equipped with the following structure of a differentiable manifold: First let $\pi : TM \to M$ with $\pi(w) = p$ for $w \in T_p M$ be the projection onto the "base point". If $x : U \to \mathbb{R}^d$ is a chart for M, we let TU be the disjoint union of the $T_p M$ with $p \in U$ and define the chart

$$dx : TU \to Tx(U), \quad (:= \bigcup_{p \in x(U)} T_p M \ )$$

where $Tx(U)$ carries the differentiable structure of $x(U) \times \mathbb{R}^d$

$$w \mapsto dx(\pi(w))(w) \in T_{x(\pi(w))}x(U).$$

The transition maps

$$dx' \circ (dx)^{-1} = d(x' \circ x^{-1})$$

then are differentiable. π is locally represented by

$$x \circ \pi \circ dx^{-1}$$

and this map maps $(x_0, v) \in Tx(U)$ to x_0.

Definition 1.2.2. The triple (TM, π, M) is called the *tangent bundle* of M, and TM is called the *total space* of the tangent bundle.

Finally, we briefly discuss the case of a complex manifold M, to have it at our disposal in §5.2. With the previous constructions and conventions in the real case understood, we let $z^j = x^j + iy^j$ again be local holomorphic coordinates near $z \in M$, as at the end of §1.1. $T_z^{\mathbb{R}}M := T_zM$ is the ordinary (real) tangent space of M at z, and

$$T_z^{\mathbb{C}}M := T_z^{\mathbb{R}}M \otimes_{\mathbb{R}} \mathbb{C}$$

is the complexified tangent space which we then decompose as

$$T_z^{\mathbb{C}}M = \mathbb{C}\{\frac{\partial}{\partial z^j}, \frac{\partial}{\partial z^{\bar{j}}}\} =: T_z'M \oplus T_z''M,$$

where $T_z'M = \mathbb{C}\{\frac{\partial}{\partial z^j}\}$ is the holomorphic and $T_z''M = \mathbb{C}\{\frac{\partial}{\partial z^{\bar{j}}}\}$ the antiholomorphic tangent space. In $T_z^{\mathbb{C}}M$, we have a conjugation, mapping $\frac{\partial}{\partial z^j}$ to $\frac{\partial}{\partial z^{\bar{j}}}$, and so, $T_z''M = \overline{T_z'M}$. The projection $T_z^{\mathbb{R}}M \to T_z^{\mathbb{C}}M \to T_z'M$ is an $\mathbb{R}$-linear isomorphism.

Perspectives. Other definitions of the tangent space of a differentiable manifold M are possible that are more elegant and less easy to compute with.

A germ of a function at $x \in M$ is an equivalence class of smooth functions defined on neighborhoods of x, where two such functions are equivalent if they coincide on some neighborhood of x. A tangent vector at x may then be defined as a linear operator δ on the function germs at x satisfying the Leibniz rule

$$\delta(f \cdot g)(x) = (\delta f(x))g(x) + f(x)\delta g(x).$$

This definition has the obvious advantage that it does not involve local coordinates.

1.3 Submanifolds

A differentiable map $f : M \to N$ is called an *immersion*, if for any $x \in M$

$$df : T_xM \to T_{f(x)}N$$

is injective. In particular, in this case $m := \dim M \leq n := \dim N$. If an immersion $f : M \to N$ maps M homeomorphically onto its image in N, f is called differentiable embedding. The following lemma shows that locally, any immersion is a differentiable embedding:

Lemma 1.3.1. *Let $f : M \to N$ be an immersion,* $\dim M = m, \dim N = n, x \in M$. *Then there exist a neighborhood U of x and a chart (V, y) on N with $f(x) \in V$, such that*

(i) $f_{|U}$ is a differentiable embedding, and

(ii) $y^{m+1}(p) = \ldots = y^n(p) = 0$ *for all* $p \in f(U) \cap V$.

Proof. This follows from the implicit function theorem. In local coordinates $(z^1, \ldots, z^n)$ on N, $(x^1, \ldots, x^m)$ on M let, w.l.o.g. (since $df(x)$ is injective)

$$\left(\frac{\partial z^\alpha(f(x))}{\partial x^i} \right)_{i,\alpha=1,\ldots,m}$$

be nonsingular.

We consider

$$F(z,x) := (z^1 - f^1(x), \ldots, z^n - f^n(x)),$$

which has maximal rank in $x^1, \ldots, x^m, z^{m+1}, \ldots, z^n$. By the implicit function theorem, there locally exists a map

$$(z^1, \ldots, z^m) \mapsto (\varphi^1(z^1, \ldots, z^m), \ldots, \varphi^n(z^1, \ldots, z^m))$$

with

$$\begin{aligned} F(z,x) = 0 \iff x^1 &= \varphi^1(z^1, \ldots, z^m), \ldots, x^m = \varphi^m(z^1, \ldots, z^m), \\ z^{m+1} &= \varphi^{m+1}(z^1, \ldots, z^m), \ldots, z^n = \varphi^n(z^1, \ldots, z^m), \end{aligned}$$

for which $(\frac{\partial \varphi^i}{\partial z^\alpha})_{\alpha,i=1,\ldots,m}$ has maximal rank.

As new coordinates, we now choose

$$\begin{aligned} (y^1, \ldots, y^n) = (&\varphi^1(z^1, \ldots, z^m), \ldots, \varphi^m(z^1, \ldots, z^m), \\ &z^{m+1} - \varphi^{m+1}(z^1, \ldots, z^m), \ldots, z^n - \varphi^n(z^1, \ldots, z^m)). \end{aligned}$$

Then

$$\begin{aligned} z &= f(x) \\ \Leftrightarrow F(z,x) &= 0 \\ \Leftrightarrow (y^1, \ldots, y^n) &= (x^1, \ldots, x^m, 0, \ldots, 0), \end{aligned}$$

and the claim follows. $\square$

If $f : M \to N$ is a differentiable embedding, $f(M)$ is called a *differentiable submanifold* of N. A subset N' of N, equipped with the relative topology, thus is a differentiable submanifold of N, if N' is a manifold and the inclusion is a differentiable embedding.

Charts on N' then are simply given by restrictions of charts of N to N', and Lemma 1.3.1 shows that one may here always find a particularly convenient structure of the charts.

Similarly, the implicit function theorem implies

Lemma 1.3.2. *Let $f : M \to N$ be a differentiable map,* $\dim M = m$, $\dim N = n, m \geq n$, $p \in N$. *Let $df(x)$ have rank n for all $x \in M$ with $f(x) = p$. Then $f^{-1}(p)$ is a union of differentiable submanifolds of M of dimension $m - n$.*

Proof. We again represent the situation in local coordinates around $x \in M$ and $p = f(x) \in N$. Of course, in these coordinates $df(x)$ still has rank n. By the implicit function theorem, there exist an open neighborhood U of x and a differentiable map

$$g(x^{n+1}, \dots, x^m) : U_2 \subset \mathbb{R}^{m-n} \to U_1 \subset \mathbb{R}^n$$

with

$$U = U_1 \times U_2$$

and

$$f(x) = p \iff (x^1, \dots, x^n) = g(x^{n+1}, \dots, x^m).$$

With

$$\begin{aligned} y^\alpha &= x^\alpha - g(x^{n+1}, \dots, x^m) && \text{for } \alpha = 1, \dots, n, \\ y^s &= x^s && \text{for } s = n+1, \dots, m, \end{aligned}$$

we then get coordinates for which

$$f(x) = p \iff y^\alpha = 0 \quad \text{for } \alpha = 1, \dots, n.$$

$(y^{n+1}, \dots, y^m)$ thus yield local coordinates for $\{f(x) = p\}$ and this implies that in some neighborhood of x $\{f(x) = p\}$ is a submanifold of M of dimension $m - n$. $\square$

Let M be a differentiable submanifold of N, and let $i : M \to N$ be the inclusion. For $p \in M, T_pM$ can then be considered as subspace of T_pN, namely as the image $di(T_pM)$.

The standard example is the sphere

$$S^n = \{x \in \mathbb{R}^{n+1} : |x| = 1\} \subset \mathbb{R}^{n+1}.$$

By the Lemma 1.3.2, S^n is a submanifold of $\mathbb{R}^{n+1}$.

Lemma 1.3.3. *In the situation of Lemma 1.3.2, we have for the submanifold $X = f^{-1}(p)$ and $q \in X$*

$$T_qX = \ker df(q) \subset T_qM.$$

Proof. Let $v \in T_qX$, (φ, U) a chart on X with $q \in U$. Let γ be any smooth curve in $\varphi(U)$ with $\gamma(0) = \varphi(q)$, $\dot\gamma(0) := \frac{d}{dt}\gamma(t)_{|t=0} = d\varphi(v)$, for example, $\gamma(t) = \varphi(q) + td\varphi(v)$. $c := \varphi^{-1}(\gamma)$ then is a curve in X with $\dot c(0) = v$. Because of $X = f^{-1}(p)$,

$$f \circ c(t) = p \quad \forall t,$$

hence $df(q) \circ \dot{c}(0) = 0$, and consequently $v = \dot{c}(0) \in \ker df(q)$. Since also $T_qX = \dim \ker df(q) = m - n$, the claim follows. $\square$

For our example S^n, we may choose

$$f : \mathbb{R}^{n+1} \to \mathbb{R}, f(x) = |x|^2.$$

Then

$$T_xS^n = \ker df(x) = \{v \in \mathbb{R}^{n+1} : x \cdot v (= x^i v^i) = 0\}.$$

Perspectives. H. Whitney (1936) showed that any d-dimensional differentiable manifold can be embedded into $\mathbb{R}^{2d+1}$. Thus, the class of abstract differentiable manifolds is the same as the class of submanifolds of Euclidean space. Nevertheless, the abstract and intrinsic point of view offers great conceptual and technical advantages over the approach of submanifold geometry of Euclidean spaces.

1.4 Riemannian Metrics

We now want to introduce metric structures on differentiable manifolds. Again, we shall start from infinitesimal considerations. We would like to be able to measure the lengths of and the angles between tangent vectors. Then, one may, for example, obtain the length of a differentiable curve by integration. In a vector space such a notion of measurement is usually given by a scalar product. We thus define

Definition 1.4.1. A *Riemannian metric* on a differentiable manifold M is given by a scalar product on each tangent space T_pM which depends smoothly on the base point p. A *Riemannian manifold* is a differentiable manifold, equipped with a Riemannian metric.

In order to understand the concept of a Riemannian metric, we again need to study local coordinate representations and the transformation behavior of these expressions.

Thus, let $x = (x^1, \ldots, x^d)$ be local coordinates. In these coordinates, a metric is represented by a positive definite, symmetric matrix

$$(g_{ij}(x))_{i,j=1,\ldots,d}$$

(i.e. $g_{ij} = g_{ji}$ for all i, j, $g_{ij}\xi^i\xi^j > 0$ for all $\xi = (\xi^1, \ldots, \xi^d) \neq 0$), where the coefficients depend smoothly on x. The transformation formula (1.4.3) below will imply that this smoothness does not depend on the choice of coordinates. Therefore, smooth dependence on the base point as required in Definition 1.4.1 can be expressed in local coordinates.

The product of two tangent vectors $v, w \in T_pM$ with coordinate representations $(v^1, \dots, v^d)$ and $(w^1, \dots, w^d)$ (i.e. $v = v^i \frac{\partial}{\partial x^i}, w = w^j \frac{\partial}{\partial x^j}$) then is

$$\langle v, w \rangle := g_{ij}(x(p)) v^i w^j. \tag{1.4.1}$$

In particular, $\langle \frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j} \rangle = g_{ij}$.

Similarly, the length of v is given by

$$\|v\| := \langle v, v \rangle^{\frac{1}{2}}.$$

We now want to study the transformation behavior. Let $y = f(x)$ define different local coordinates. In these coordinates, v and w have representations $(\tilde{v}^1, \dots, \tilde{v}^d)$ and $(\tilde{w}, \dots, \tilde{w}^d)$ with $\tilde{v}^j = v^i \frac{\partial f^j}{\partial x^i}, \tilde{w}^j = w^i \frac{\partial f^j}{\partial x^i}$. Let the metric in the new coordinates be given by $h_{k\ell}(y)$.
It follows that

$$h_{k\ell}(f(x)) \tilde{v}^k \tilde{w}^\ell = \langle v, w \rangle = g_{ij}(x) v^i w^j, \tag{1.4.2}$$

hence

$$h_{k\ell}(f(x)) \frac{\partial f^k}{\partial x^i} \frac{\partial f^\ell}{\partial x^j} v^i w^j = g_{ij}(x) v^i w^j,$$

and since this holds for all tangent vectors v, w,

$$h_{k\ell}(f(x)) \frac{\partial f^k}{\partial x^i} \frac{\partial f^\ell}{\partial x^j} = g_{ij}(x). \tag{1.4.3}$$

Formula (1.4.3) gives the transformation behavior of a metric under coordinate changes.

The simplest example of a Riemannian metric of course is the Euclidean one. For $v = (v^1, \dots, v^d), w = (w^1, \dots, w^d) \in T_x\mathbb{R}^d$, the Euclidean scalar product is simply

$$\delta_{ij} v^i w^j = v^i w^i,$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

is the standard Kronecker symbol.

Theorem 1.4.1. *Each differentiable manifold may be equipped with a Riemannian metric.*

Proof. Let $\{(x_\alpha, U_\alpha) : \alpha \in A\}$ be an atlas, $(\varphi_\alpha)_{\alpha \in A}$ a partition of unity subordinate to $(U_\alpha)_{\alpha \in A}$ (see Lemma 1.1.1 (for simplicity of notation, we use the same index set for (φ_α) and (U_α); this may be justified by replacing the original covering (U_α) by a locally finite refinement).

For $v, w \in T_pM$ and $\alpha \in A$ with $p \in U_\alpha$ let the coordinate representations be $(v_\alpha^1, \ldots, v_\alpha^d)$ and $(w_\alpha^1, \ldots, w_\alpha^d)$. Then we put

$$\langle v, w \rangle := \sum_{\substack{\alpha \in A \\ \text{with } p \in U_\alpha}} \varphi_\alpha(p) v_\alpha^i w_\alpha^i.$$

This defines a Riemannian metric. (The metric is simply obtained by piecing the Euclidean metrics of the coordinate images together with the help of a partition of unity.) □

Let now $[a, b]$ be a closed interval in $\mathbb{R}, \gamma : [a, b] \to M$ a smooth curve, where "smooth", as always, means "of class C^∞".
The *length of* γ then is defined as

$$L(\gamma) := \int_a^b \left\| \frac{d\gamma}{dt}(t) \right\| dt$$

and the *energy* of γ as

$$E(\gamma) := \frac{1}{2} \int_a^b \left\| \frac{d\gamma}{dt}(t) \right\|^2 dt.$$

(In physics, $E(\gamma)$ is usually called "action of γ" where γ is considered as the orbit of a mass point.) Of course, these expressions can be computed in local coordinates. Working with the coordinates $(x^1(\gamma(t)), \ldots, x^d(\gamma(t)))$ we use the abbreviation

$$\dot{x}^i(t) := \frac{d}{dt}(x^i(\gamma(t))).$$

Then

$$L(\gamma) = \int_a^b \sqrt{g_{ij}(x(\gamma(t)))\dot{x}^i(t)\dot{x}^j(t)}dt$$

and

$$E(\gamma) = \frac{1}{2} \int_a^b g_{ij}(x(\gamma(t)))\dot{x}^i(t)\dot{x}^j(t)dt.$$

We also remark for later technical purposes that the length of a (continuous and) piecewise smooth curve may be defined as the sum of the lengths of the smooth pieces, and the same holds for the energy.

On a Riemannian manifold M, the *distance* between two points p, q can be defined:

$$d(p, q) := \inf\{L(\gamma) : \gamma : [a, b] \to M \text{ piecewise smooth curve with } \gamma(a) = p, \gamma(b) = q\}.$$

We first remark, that any two points $p, q \in M$ can be connected by a piecewise smooth curve, and $d(p, q)$ therefore is always defined. Namely, let

$$E_p := \{q \in M : p \text{ and } q \text{ can be connected by a piecewise smooth curve.}\}$$

With the help of local coordinates one sees that E_p is open. But then also $M \backslash E_p = \bigcup_{q \notin E_p} E_q$ is open. Since M is connected and $E_p \neq \emptyset \, (p \in E_p)$, we conclude $M = E_p$.

The distance function satisfies the usual axioms:

Lemma 1.4.1.

(i) $d(p,q) \geq 0$ *for all* $p, q,$ *and* $d(p,q) > 0$ *for all* $p \neq q$

(ii) $d(p,q) = d(q,p)$,

(iii) $d(p,q) \leq d(p,r) + d(r,q)$ *(triangle inequality) for all points* $p, q, r \in M$.

Proof. (ii) and (iii) are obvious. For (i), we only have to show $d(p,q) > 0$ for $p \neq q$. For this purpose, let $x : U \to \mathbb{R}^d$ be a chart with $p \in U$. Then there exists $\varepsilon > 0$ with

$$D_\varepsilon(x(p)) := \{y \in \mathbb{R}^d : |y - x(p)| \leq \varepsilon\} \subset x(U)$$

(the bars denote the Euclidean absolute value) and

$$q \notin x^{-1}(D_\varepsilon(x(p))). \tag{1.4.4}$$

Let the metric be represented by $(g_{ij}(x))$ in our chart. Since $(g_{ij}(x))$ is positive definite and smooth, hence continuous in x and $D_\varepsilon(x(p))$ is compact, there exists $\lambda > 0$ with

$$g_{ij}(y)\xi^i\xi^j \geq \lambda|\xi|^2 \tag{1.4.5}$$

for all $y \in D_\varepsilon(x(p)), \xi = (\xi^1, \dots, \xi^d) \in \mathbb{R}^d$. Therefore, for any curve $\gamma : [a,b] \to M$ with $\gamma(a) = p, \gamma(b) = q$

$$\begin{aligned} L(\gamma) &\geq L(\gamma \cap x^{-1}(D_\varepsilon(x(p))) \\ &\geq \lambda\varepsilon > 0, \end{aligned} \tag{1.4.6}$$

because $x(\gamma)$ by (1.4.4) has to contain a point $z \in \partial D_\varepsilon(x(p))$, i.e. a point whose Euclidean distance from $x(p)$ is ε. By (1.4.5), z then has distance from $x(p)$ at least $\lambda\varepsilon$ w.r.t. the metric (g_{ij}). □

Corollary 1.4.1. *The topology on* M *induced by the distance function* d *coincides with the original manifold topology of* M.

Proof. It suffices to show that in each chart the topology induced by d coincides with the one of $\mathbb{R}^d$, i.e. the one induced by the Euclidean distance function. Now for every x in some chart, there exists $\varepsilon > 0$ for which $D_\varepsilon(x)$ is contained in the same chart, and positive constants λ, μ with

$$\lambda^2|\xi|^2 \leq g_{ij}(y)\xi^i\xi^j \leq \mu^2|\xi|^2 \quad \text{for all } y \in D_\varepsilon(x), \xi \in \mathbb{R}^d.$$

Thus

$$\lambda|y - x| \leq d(y,x) \leq \mu|y - x| \quad \text{for all } y \in D_\varepsilon(x),$$

and thus each Euclidean distance ball contains a distance ball for d, and vice versa, (with

$$B(z,\delta) := \{y \in M : d(z,y) \le \delta\}$$

we have

$$\mathring{D}_{\lambda\delta}(x) \subset \mathring{B}(x,\delta) \subset \mathring{D}_{\mu\delta}(x),$$

if $\mu\delta \le \varepsilon$). $\square$

We now return to the length and energy functionals.

Lemma 1.4.2. *For each smooth curve $\gamma : [a,b] \to M$*

$$L(\gamma)^2 \le 2(b-a)E(\gamma), \tag{1.4.7}$$

and equality holds if and only if $\|\frac{d\gamma}{dt}\| \equiv \text{const}$.

Proof. By Hölder's inequality

$$\int_a^b \left\|\frac{d\gamma}{dt}\right\| dt \le (b-a)^{\frac{1}{2}} \left(\int_a^b \left\|\frac{d\gamma}{dt}\right\|^2 dt\right)^{\frac{1}{2}}$$

with equality precisely if $\|\frac{d\gamma}{dt}\| \equiv \text{const}$. $\square$

Lemma 1.4.3. *If $\gamma : [a,b] \to M$ is a smooth curve, and $\psi : [\alpha,\beta] \to [a,b]$ is a change of parameter, then*

$$L(\gamma \circ \psi) = L(\gamma).$$

Proof. Let $t = \psi(\tau)$.

By the chain rule,

$$L(\gamma \circ \psi) = \int_\alpha^\beta \left(g_{ij}\big(x(\gamma(\psi(\tau)))\big)\, \dot{x}^i(\psi(\tau))\, \dot{x}^j(\psi(\tau)) \Big(\frac{d\psi}{d\tau}\Big)^2 \right)^{\frac{1}{2}} d\tau$$

and by a change of variables,

$$= L(\gamma).$$

$\square$

Lemma 1.4.4. *The Euler-Lagrange equations for the energy E are*

$$\ddot{x}^i(t) + \Gamma^i_{jk}(x(t))\dot{x}^j(t)\dot{x}^k(t) = 0, \quad i = 1, \dots, d \tag{1.4.8}$$

with

$$\Gamma^i_{jk} = \frac{1}{2} g^{i\ell}(g_{j\ell,k} + g_{k\ell,j} - g_{jk,\ell}),$$

where

$$(g^{ij})_{i,j=1,\dots,d} = (g_{ij})^{-1} \quad (\textit{i.e. } g^{i\ell} g_{\ell j} = \delta_{ij})$$

and

$$g_{j\ell,k} = \frac{\partial}{\partial x^k} g_{j\ell}.$$

The expressions Γ^i_{jk} are called Christoffel symbols.

Proof. The Euler-Lagrange equations of a functional

$$I(x) = \int_a^b f(t, x(t), \dot{x}(t))dt$$

are given by

$$\frac{d}{dt}\frac{\partial f}{\partial \dot{x}^i} - \frac{\partial f}{\partial x^i} = 0, \quad i = 1, \dots, d.$$

In our case, recalling

$$E(\gamma) = \frac{1}{2}\int g_{jk}(x(t))\dot{x}^j \dot{x}^k dt,$$

we get

$$\frac{d}{dt}\left(g_{ik}(x(t))\dot{x}^k(t) + g_{ji}(x(t))\dot{x}^j(t)\right) - g_{jk,i}(x(t))\dot{x}^j(t)\dot{x}^k(t) = 0,$$

for $i = 1, \dots, d$, hence

$$g_{ik}\ddot{x}^k + g_{ji}\ddot{x}^j + g_{ik,\ell}\dot{x}^\ell \dot{x}^k + g_{ji,\ell}\dot{x}^\ell \dot{x}^j - g_{jk,i}\dot{x}^j \dot{x}^k = 0.$$

Renaming some indices and using the symmetry $g_{ik} = g_{ki}$, we get

$$2g_{\ell m}\ddot{x}^m + (g_{\ell k,j} + g_{j\ell,k} - g_{jk,\ell})\dot{x}^j \dot{x}^k = 0, \quad \ell = 1, \dots, d, \tag{1.4.9}$$

and from this

$$g^{i\ell} g_{\ell m}\ddot{x}^m + \frac{1}{2} g^{i\ell}(g_{\ell k,j} + g_{j\ell,k} - g_{jk,\ell})\dot{x}^j \dot{x}^k = 0, i = 1, \dots, d.$$

Because of

$$g^{i\ell} g_{\ell m} = \delta_{im}, \text{ and thus } g^{i\ell} g_{\ell m}\ddot{x}^m = \ddot{x}^i,$$

we obtain (1.4.8) from this. □

Definition 1.4.2. A smooth curve $\gamma = [a,b] \to M$, which satisfies (with $\dot{x}^i(t) = \frac{d}{dt}x^i(\gamma(t))$ etc.)

$$\ddot{x}^i(t) + \Gamma^i_{jk}(x(t))\dot{x}^j(t)\dot{x}^k(t) = 0, \text{ for } i = 1, \dots, d$$

is called a *geodesic*.

Thus, geodesics are the critical points of the energy functional. By Lemma 1.4.3, the length functional is invariant under parameter changes. As in the Euclidean case, one easily sees that regular curves can be parametrized by arc length. We shall attempt to minimize the length within the class of regular smooth curves, and we shall succeed and complete the program in Corollary 1.4.2 below. As the length is invariant under reparametrization by Lemma 1.4.3, therefore, if one seeks curves of shortest length, it suffices to consider curves that are parametrized by arc length. For such curves, by Lemma 1.4.2 one may minimize energy instead of length. Conversely, every critical point of the energy functional, i.e. each solution of (1.4.8), i.e. each geodesic, is parametrized proportionally to arc length.

Namely, for a solution of (1.4.8)

$$\begin{aligned}
\frac{d}{dt}\langle \dot{x}, \dot{x} \rangle &= \frac{d}{dt}(g_{ij}(x(t))\dot{x}^i(t)\dot{x}^j(t)) \\
&= g_{ij}\ddot{x}^i\dot{x}^j + g_{ij}\dot{x}^i\ddot{x}^j + g_{ij,k}\dot{x}^i\dot{x}^j\dot{x}^k \\
&= -(g_{jk,\ell} + g_{\ell j,k} - g_{\ell k,j})\dot{x}^\ell\dot{x}^k\dot{x}^j + g_{\ell j,k}\dot{x}^k\dot{x}^\ell\dot{x}^j
\end{aligned}$$

by formula (1.4.9), which is equivalent to (1.4.8)

$$= 0$$

since $g_{jk,\ell}\dot{x}^\ell\dot{x}^k\dot{x}^j = g_{\ell k,j}\dot{x}^\ell\dot{x}^k\dot{x}^j$ by interchanging the indices j and ℓ.

Consequently $\langle \dot{x}, \dot{x} \rangle \equiv \text{const}$, and hence the curve is parametrized proportionally to arc length. We have shown

Lemma 1.4.5. *Each geodesic is parametrized proportionally to arc length.* □

Theorem 1.4.2. *Let M be a Riemannian manifold, $p \in M, v \in T_pM$. Then there exist $\varepsilon > 0$ and precisely one geodesic*

$$c : [0, \varepsilon] \to M$$

with $c(0) = p, \dot{c}(0) = v$. In addition, c depends smoothly on p and v.

Proof. (1.4.8) is a system of second order ODE, and the Picard-Lindelöf Theorem yields the local existence and uniqueness of a solution with prescribed initial values and derivatives, and this solution depends smoothly on the data. □

We note that if $x(t)$ is a solution of (1.4.8), so is $x(\lambda t)$ for any constant $\lambda \in \mathbb{R}$. Denoting the geodesic of Theorem 1.4.2 with $c(0) = p, \dot{c}(0) = v$ by c_v, we obtain

$$c_v(t) = c_{\lambda v}(\tfrac{t}{\lambda}) \text{ for } \lambda > 0, t \in [0, \varepsilon].$$

In particular, $c_{\lambda v}$ is defined on $[0, \frac{\varepsilon}{\lambda}]$.

Since c_v depends smoothly on v, and $\{v \in T_pM : \|v\| = 1\}$ is compact, there exists $\varepsilon_0 > 0$ with the property that for $\|v\| = 1$ c_v is defined at least on $[0, \varepsilon_0]$. Therefore, for any $w \in T_pM$ with $\|w\| \leq \varepsilon_0$, c_w is defined at least on $[0, 1]$.

Definition 1.4.3. Let M be a Riemannian manifold, $p \in M$,

$$\begin{aligned} V_p :=& \{v \in T_pM : c_v \text{ is defined on } [0,1] \\ \exp_p :& V_p \to M \\ & v \mapsto c_v(1) \end{aligned}$$

is called the *exponential map* of M at p.

By the preceding considerations, the domain of definition of the exponential map always at least contains a small neighborhood of $0 \in T_pM$. In general, however, V_p is not all of T_pM, as is already seen in the example of a proper, open subset of $\mathbb{R}^d$, equipped with the Euclidean metric. Nevertheless, we shall see in Theorem 1.5.2 below that for a compact Riemannian manifold, $\exp_p$ can be defined on all of T_pM.

Theorem 1.4.3. *The exponential map* $\exp_p$ *maps a neighborhood of* $0 \in T_pM$ *diffeomorphically onto a neighborhood of* $p \in M$.

Proof. Since T_pM is a vector space, we may identify T_0T_pM, the tangent space of T_pM at $0 \in T_pM$, with T_pM itself. The derivative of $\exp_p$ at 0 then becomes a map from T_pM onto itself:

$$d\exp_p(0) : T_pM \to T_pM.$$

With this identification of T_0T_pM and T_pM, for $v \in T_pM$

$$\begin{aligned} d\exp_p(0)(v) &= \frac{d}{dt} c_{tv}(1)_{|t=0} \\ &= \frac{d}{dt} c_v(t)_{|t=0} \\ &= \dot{c}_v(0) \\ &= v. \end{aligned}$$

Hence

$$d\exp_p(0) = id_{|T_pM}. \tag{1.4.10}$$

In particular, $d\exp_p(0)$ has maximal rank, and by the inverse function theorem, there exists a neighborhood of $0 \in T_pM$ which is mapped diffeomorphically onto a neighborhood of $p \in M$. $\square$

Let now $e_1, e_2, \ldots, e_d$ $(d = \dim M)$ be a basis of T_pM which is orthonormal w.r.t. the scalar product on T_pM defined by the Riemannian metric. Writing for each vector $v \in T_pM$ its components w.r.t. this basis, we obtain a map

$$\begin{aligned} \Phi &: T_pM \to \mathbb{R}^d \\ v &= v^i e_i \mapsto (v^1, \ldots, v^d). \end{aligned}$$

For the subsequent construction, we identify T_pM with $\mathbb{R}^d$ via Φ. By Theorem 1.4.3, there exists a neighborhood U of p which is mapped by $\exp_p^{-1}$ diffeomorphically onto a neighborhood of $0 \in T_pM$, hence, with our identification $T_pM \cong \mathbb{R}^d$, diffeomorphically onto a neighborhood Ω of $0 \in \mathbb{R}^d$. In particular, p is mapped to 0.

Definition 1.4.4. The local coordinates defined by the chart $(\exp_p^{-1}, U)$ are called (Riemannian) *normal coordinates* with center p.

Theorem 1.4.4. *In normal coordinates, we have for the Riemannian metric*

$$g_{ij}(0) = \delta_{ij}, \tag{1.4.11}$$

$$\Gamma^i_{jk}(0) = 0, \quad \textit{(and also } g_{ij,k}(0) = 0\textit{) for all } i, j, k. \tag{1.4.12}$$

Proof. (1.4.11) directly follows from the fact that the above identification $\Phi : T_pM \cong \mathbb{R}^d$ maps an orthonormal basis of T_pM w.r.t. the Riemannian metric onto an Euclidean orthonormal basis of $\mathbb{R}^d$.

For (1.4.12), we note that in normal coordinates, the straight lines through the origin of $\mathbb{R}^d$ (or, more precisely, their portions contained in the chart image) are geodesic. Namely, the line $tv, t \in \mathbb{R}, v \in \mathbb{R}^d$, is mapped (for sufficiently small t) onto $c_{tv}(1) = c_v(t)$, where $c_v(t)$ is the geodesic, parametrized by arc length, with $\dot{c}_v(0) = v$.

Inserting now $x(t) = tv$ into the geodesic equation (1.4.8), we obtain because of $\ddot{x}(t) = 0$

$$\Gamma^i_{jk}(tv) v^j v^k = 0, \text{ for } i = 1, \ldots, d. \tag{1.4.13}$$

In particular at 0, i.e. for $t = 0$,

$$\Gamma^i_{jk}(0) v^j v^k = 0 \text{ for all } v \in \mathbb{R}^d, i = 1, \ldots, d. \tag{1.4.14}$$

We put $v = \frac{1}{2}(e_\ell + e_m)$ and obtain because of the symmetry $\Gamma^i_{jk} = \Gamma^i_{kj}$

$$\Gamma^i_{\ell m}(0) = 0 \text{ for all } i.$$

Since this holds for all ℓ, m, all $\Gamma^i_{jk}(0)$ vanish. By definition of Γ^i_{jk}, we obtain at $0 \in \mathbb{R}^d$

$$g^{i\ell}(g_{j\ell,k} + g_{k\ell,j} - g_{jk,\ell}) = 0 \ \forall\, i, j, k,$$

hence also

$$g_{jm,k} + g_{km,j} - g_{jk,m} = 0 \;\forall\, j,k,m.$$

Adding now the relation (obtained by cyclic permutation of the indices)

$$g_{kj,m} + g_{mj,k} - g_{km,j} = 0,$$

we obtain (with $g_{kj} = g_{jk}$)

$$g_{jm,k}(0) = 0, \text{ for all } j,k,m.$$

□

Later on (in Chapter 3), we shall see that in general the *second* derivatives of the metric cannot be made to vanish at a given point by a suitable choice of local coordinates. The obstruction will be given by the curvature tensor.

Further properties of Riemannian normal coordinates may best be seen by using *polar coordinates*, instead of the Euclidean ones (obtained from the map Φ). We therefore introduce on $\mathbb{R}^d$ the standard polar coordinates

$$(r, \varphi^1, \dots, \varphi^{d-1}),$$

where $\varphi = (\varphi^1, \dots, \varphi^{d-1})$ parametrizes the unit sphere S^{d-1} (the precise formula for φ will be irrelevant for our purposes), and we then obtain polar coordinates on T_pM via Φ again. We express the metric in polar coordinates and write g_{rr} instead of g_{11}, because of the special role of r. We also write $g_{r\varphi}$ instead of $g_{1\ell}, \ell \in \{2, \dots, d\}$, and $g_{\varphi\varphi}$ as abbreviation for $(g_{k\ell})_{k,\ell=2,\dots,d}$. In particular, in these coordinates at $0 \in T_pM$ (this point corresponds to $p \in M$)

$$g_{rr}(0) = 1, g_{r\varphi}(0) = 0 \tag{1.4.15}$$

by (1.4.11) and since this holds for Euclidean polar coordinates.

After these preparations, we return to the analysis of the geodesic equation (1.4.8). The lines $\varphi \equiv$ const. are geodesic when parametrized by arc length. They are given by $x(t) = (t, \varphi_0)$, φ_0 fixed, and from (1.4.8)

$$\Gamma^i_{rr} = 0 \text{ for all } i$$

(we have written Γ^i_{rr} instead of Γ^i_{11}), hence

$$g^{i\ell}(2g_{r\ell,r} - g_{rr,\ell}) = 0, \text{ for all } i,$$

thus

$$2g_{r\ell,r} - g_{rr,\ell} = 0, \text{ for all } \ell. \tag{1.4.16}$$

For $\ell = r$, we conclude

$$g_{rr,r} = 0,$$

and with (1.4.15) then

$$g_{rr} \equiv 1. \tag{1.4.17}$$

Inserting this in (1.4.16), we get

$$g_{r\varphi,r} = 0,$$

and then again with (1.4.15)

$$g_{r\varphi} \equiv 0. \tag{1.4.18}$$

We have shown

Theorem 1.4.5. *For the polar coordinates, obtained by transforming the Euclidean coordinates of* $\mathbb{R}^d$*, on which the normal coordinates with centre* p *are based, into polar coordinates, we have*

$$g_{ij} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & g_{\varphi\varphi}(r,\varphi) & \\ 0 & & & \end{pmatrix},$$

where $g_{\varphi\varphi}(r,\varphi)$ *is the* $(d-1)\times(d-1)$ *matrix of the components of the metric w.r.t. angular variables* $(\varphi^1,\dots,\varphi^{d-1}) \in S^{d-1}$. □

The polar coordinates of Theorem 1.4.5 are often called *Riemannian polar coordinates*. The situation is the same as for Euclidean polar coordinates: For example in polar coordinates on $\mathbb{R}^2$, the Euclidean metric is given by $\left(\begin{smallmatrix} 1 & 0 \\ 0 & r^2 \end{smallmatrix}\right)$. We point out once more that in contrast to Theorem 1.4.4, Theorem 1.4.5 holds not only at the origin $0 \in T_pM$, but in the whole chart.

Corollary 1.4.2. *For any* $p \in M$*, there exists* $\rho > 0$ *such that Riemannian polar coordinates may be introduced on* $B(p,\rho) := \{q \in M : d(p,q) \le \rho\}$*. For any such* ρ *and any* $q \in \partial B(p,\rho)$*, there is precisely one geodesic of shortest length* $(=\rho)$ *from* p *to* q*, and in polar coordinates, this geodesic is given by the straight line* $x(t) = (t,\varphi_0)$*,* $0 \le t \le \rho$*, where* q *is represented by the coordinates* $(\rho,\varphi_0), \varphi_0 \in S^{d-1}$*. Here, "of shortest length" means that the curve is the shortest one among all curves in* M *from* p *to* q*.*

Proof. The first claim follows from Corollary 1.4.1 (and its proof) and Theorem 1.4.3. For the second claim, let $c(t) = (r(t),\varphi(t)), 0 \le t \le T$, be an arbitrary curve from p to q. $c(t)$ need not be entirely contained in $B(p,\rho)$ and may leave our coordinate neighborhood. Let

$$t_0 := \inf\{t \le T : d(c(t),p) \ge \rho\}.$$

Then $t_0 \le T$, and the curve $c_{|[0,t_0]}$ is entirely contained in $B(p,\rho)$. We shall show $L(c_{|[0,t_0]}) \ge \rho$ with equality only for a straight line in our polar coordinates. This will then imply the second claim. The proof of this inequality goes as follows:

$$\begin{aligned} L(c_{|[0,t_0]}) &= \int_0^{t_0} (g_{ij}(c(t))\dot{c}^i\dot{c}^j)^{\frac{1}{2}}\,dt \\ &\ge \int_0^{t_0} (g_{rr}(c(t))\dot{r}\dot{r})^{\frac{1}{2}}\,dt \end{aligned}$$

by (1.4.18) and since $g_{\varphi\varphi}$ is positive definite

$$\begin{aligned}
&= \int_0^{t_0} |\dot r| dt \geq \int_0^{t_0} \dot r dt \quad \text{by (1.4.17)} \\
&= r(t_0) \\
&= \rho \quad \text{by definition of } t_0,
\end{aligned}$$

and equality holds precisely if $g_{\varphi\varphi}\dot\varphi\dot\varphi \equiv 0$, in which case $\varphi(t)$ is constant and $\dot r \geq 0$ and $c(t)$ thus is a straight line through the origin. □

In particular, under the assumptions of Corollary 1.4.2, the Euclidean ball

$$d_\rho(0) := \{y \in \mathbb{R}^d : |y| \leq \rho\} \subset T_pM$$

is mapped under $\exp_p$ diffeomorphically onto the Riemannian ball with the same radius,

$$B(p, \rho).$$

Corollary 1.4.3. *Let M be a compact Riemannian manifold. Then there exists $\rho_0 > 0$ with the property that for any $p \in M$, Riemannian polar coordinates may be introduced on $B(p, \rho_0)$.*

Proof. By Corollary 1.4.2, for any $p \in M$, there exists $\rho > 0$ with those properties. By Theorem 1.4.2, $\exp_p$ is smooth in p. If thus $\exp_p$ is injective and of maximal rank on a closed ball with radius ρ in T_pM, there exists a neighborhood U of p such that for all $q \in U, \exp_q$ is injective and of maximal rank on the closed ball with radius ρ in T_qM.

Since M is compact, it can be covered by finitely many such neighborhoods and we choose ρ_0 as the smallest such ρ. □

Corollary 1.4.4. *Let M be a compact Riemannian manifold. Then there exists $\rho_0 > 0$ with the property that any two points $p, q \in M$ with $d(p, q) \leq \rho_0$ can be connected by precisely one geodesic of shortest length. This geodesic depends continuously on p and q.*

Proof. ρ_0 from Corollary 1.4.3 satisfies the first claim by Corollary 1.4.2. Moreover, by the last claim of Corollary 1.4.2, the shortest geodesic from p to $q \in B(p, \rho_0)$ depends continuously on p. Exchanging the roles of p and q yields the continuous dependence on p as well. □

We explicitly point out that for any compact Riemannian manifold there is always more than one geodesic connection between any two points (This will be discussed in Chapter 6.). Only the shortest geodesic is unique, provided p and q are sufficiently close.

Let now M be a differentiable submanifold of the Riemannian manifold N. The Riemannian metric of N then induces a Riemannian metric on M, by restricting the former one to $T_pM \subset T_pN$ for $p \in N$. Thus, M also becomes a Riemannian manifold.

In particular, $S^n \subset \mathbb{R}^{n+1}$ obtains a Riemannian metric. We want to compute this metric in the local chart of §1.1, namely

$$\begin{aligned} f(x^1, \dots, x^{n+1}) &= \left(\frac{x^1}{1 - x^{n+1}}, \dots, \frac{x^n}{1 - x^{n+1}} \right) \quad \text{for } x^{n+1} \neq 1 \\ &=: (y^1, \dots, y^n) \in \mathbb{R}^n \,. \end{aligned}$$

In the sequel, a Latin index occuring twice in a product has to be summed from 1 to $n+1$, a Greek one from 1 to n. We compute

$$1 = x^i x^i = y^\alpha y^\alpha (1 - x^{n+1})^2 + x^{n+1} x^{n+1}$$

hence

$$x^{n+1} = \frac{y^\alpha y^\alpha - 1}{y^\alpha y^\alpha + 1}$$

and then

$$x^i = \frac{2y^i}{1 + y^\alpha y^\alpha} \quad (i = 1, \dots, n).$$

For $g := f^{-1}$ then

$$\begin{aligned} \frac{\partial g^j}{\partial y^k} &= \frac{2\delta_{jk}}{1 + y^\alpha y^\alpha} - \frac{4y^j y^k}{(1 + y^\alpha y^\alpha)^2} \quad \text{for } j = 1, \dots, n,\ k = 1, \dots, n \\ \frac{\partial g^{n+1}}{\partial y^k} &= \frac{4y^k}{(1 + y^\alpha y^\alpha)^2}. \end{aligned}$$

Let a tangent vector to S^n be represented by $w = (w^1, \dots, w^n)$ in our chart. Then

$$\begin{aligned} \langle w, w \rangle &= dg(w) \cdot dg(w), \ \text{where the point denotes the Euclidean} \\ &\qquad\qquad \text{scalar product of } \mathbb{R}^{n+1} \\ &= \frac{1}{(1 + y^\alpha y^\alpha)^4} \left\{ 4(1 + y^\alpha y^\alpha)^2 w^\beta w^\beta - 16(1 + y^\alpha y^\alpha) y^\beta w^\beta y^\gamma w^\gamma \right. \\ &\qquad \left. + 16 y^\beta y^\beta y^\gamma w^\gamma y^\delta w^\delta + 16 y^\beta w^\beta y^\gamma w^\gamma \right\} \\ &= \frac{4}{(1 + y^\alpha y^\alpha)^2} w^\beta w^\beta. \end{aligned}$$

Thus, the metric in our chart is given by

$$g_{ij}(y) = \frac{4}{(1 + |y|^2)^2} \delta_{ij}.$$

Definition 1.4.5. A diffeomorphism $h : M \to N$ between Riemannian manifolds is an *isometry* if it preserves the Riemannian metric. Thus, for $p \in M, v, w \in T_pM$, and if $\langle \cdot, \cdot \rangle_M$ and $\langle \cdot, \cdot \rangle_N$ denotes the scalar products in T_pM and $T_{h(p)}N$, resp., we have

$$\langle v, w \rangle_M = \langle dh(v), dh(w) \rangle_N.$$

A differentiable map $h : M \to N$ is a *local isometry* if for every $p \in M$ there exists a neighborhood U for which $h_{|U} : U \to h(U)$ is an isometry, and $h(U)$ is open in N.

If $(g_{ij}(p))$ and $(\gamma_{\alpha\beta}(h(p))$ are the coordinate representations of the metric, an isometry has to satisfy

$$g_{ij}(p) = \gamma_{\alpha\beta}(h(p)) \frac{\partial h^\alpha(p)}{\partial x^i} \frac{\partial h^\beta(p)}{\partial x^j}.$$

A local isometry thus has the same effect as a coordinate change. Isometries leave the lengths of tangent vectors and therefore also the lengths and energies of curves invariant. Thus, critical points, i.e. geodesics, are mapped to geodesics.

With this remark, we may easily determine the geodesics of S^n. The orthogonal group $\mathrm{O}(n+1)$ operates isometrically on $\mathbb{R}^{n+1}$, and since it maps S^n into S^n, it also operates isometrically on S^n. Let now $p \in S^n, v \in T_pS^n$. Let E be the two dimensional plane through the origin of $\mathbb{R}^{n+1}$, containing v. We claim that the geodesic c_v through p with tangent vector v is nothing but the great circle through p with tangent vector v (parametrized proportionally to arc length), i.e. the intersection of S^n with E. For this, let $S \in \mathrm{O}(n+1)$ be the reflection across that E. Together with c_v, Sc_v is also a geodesic through p with tangent vector v. The uniqueness result of Theorem 1.4.2 implies $c_v = Sc_v$, and thus the image of c_v is the great circle, as claimed.

As another example, we consider the torus T^2 introduced in §1.1. We introduce a metric on T^2 by letting the projection π be a local isometry. For each chart of the form $(U, (\pi_{|U})^{-1})$, we use the Euclidean metric on $\pi^{-1}(U)$. Since the translations

$$z \mapsto z + m_1 w_1 + m_2 w_2 \quad (m_1, m_2 \in \mathbb{Z})$$

are Euclidean isometries, the Euclidean metrics on the different components of $\pi^{-1}(U)$ (which are obtained from each other by such translations) yield the same metric on U. Hence, the Riemannian metric on T^2 is well defined.

Since π is a local isometry, Euclidean geodesics of $\mathbb{R}^2$ are mapped onto geodesics of T^2. The global behavior of geodesics on such a torus is most easily studied in the case where T^2 is generated by the two unit vectors $w_1 = (1, 0)$ and $w_2 = (0, 1)$: A straight line in $\mathbb{R}^2$ which is parallel to one of the coordinate axes then becomes a geodesic on T^2 that closes up after going around once. More generally, a straight line with rational slope becomes a closed, hence periodic geodesic on T^2, while the image of one with irrational slope lies dense in T^2.

Before ending this paragraph, we want to introduce the following important notion:

Definition 1.4.6. Let M be a Riemannian manifold, $p \in M$. The *injectivity radius* of p is

$$i(p) := \sup\{\rho > 0 : \exp_p \text{ is defined on } d_\rho(0) \subset T_pM \text{ and injective}\}.$$

The *injectivity radius* of M is

$$i(M) := \inf_{p \in M} i(p).$$

For example, the injectivity radius of the sphere S^n is π, since the exponential map of any point p maps the open ball of radius π in T_pM injectively onto the complement of the antipodal point of p.

The injectivity radius of the torus just discussed is $\frac{1}{2}$, since here the exponential map is injective on the interior of a square with centre $0 \in T_pM$ and side length 1.

Perspectives. As the name suggests, the concept of a Riemannian metric was introduced by Bernhard Riemann, in his habilitation address [262]. He also suggested to consider more generally metrics obtained by taking metrics on the tangent spaces that are not induced by a scalar product. Such metrics were first systematically investigated by Finsler and are therefore called Finsler metrics.

For a general metric space, a geodesic is defined as a curve which realizes the shortest distance between any two sufficiently close points lying on it. Those metric spaces that satisfy the conclusion of the Hopf-Rinow theorem (proved below) that any two points can be connected by a shortest geodesic are called geodesic length spaces, and they are amenable to geometric constructions as demonstrated by the school of Alexandrov. See e.g. [204], [15].

A Lorentz metric on a differentiable manifold of dimension $d+1$ is given by an inner product of signature $(1, d)$ on each tangent space T_pM depending smoothly on p. A Lorentz manifold is a differentiable manifold with a Lorentz metric. The prototype is Minkowski space, namely $\mathbb{R}^{d+1}$ equipped with the inner product

$$\langle x, y \rangle = -x^0 y^0 + x^1 y^1 + \ldots + x^d y^d$$

for $x = (x^0, x^1, \ldots, x^d), y = (y^0, y^1, \ldots, y^d)$. Lorentz manifolds are the spaces occuring in general relativity. Let us briefly discuss some concepts. Tangent vectors V with negative, positive, vanishing $\|V\|^2 = \langle V, V \rangle$ are called time-like, space-like, and light-like, resp. Length and energy of a curve may be defined formally as in the Riemannian case, and we again obtain geodesic equations. Geodesics whose tangent vectors all have norm zero are called null geodesics. They describe the paths of light rays. (Note that in our above description of the Minkowski metric, the conventions have been chosen so that the speed of light is 1.) Submanifolds of Lorentz manifolds whose tangent vectors are all space-like are ordinary Riemannian manifolds w.r.t. the induced metric. For treatments of Lorentzian geometry, an introduction is [218]. Deeper aspects are treated in Hawking and Ellis[120].

J. Nash proved that every Riemannian manifold M can be isometrically embedded into some Euclidean space $\mathbb{R}^k$. For the proof of this result, he developed an implicit function theorem in Fréchet spaces and an iteration technique that have found other important applications. A simpler proof was found by Günther[114].

Although on a conceptual level, Nash's theorem reduces the study of Riemannian manifolds to the study of submanifolds of Euclidean spaces, in practice the intrinsic point of view has proved to be preferable (see Perspectives on §1.3).

In our presentation, we only consider finite dimensional Riemannian manifolds. It is also possible, and often very useful, to introduce infinite dimensional Riemannian manifolds. Those are locally modeled on Hilbert spaces instead of Euclidean ones. The lack of local compactness leads to certain technical complications, but most ideas and constructions of

Riemannian geometry pertain to the infinite dimensional case. Such infinite dimensional manifolds arise for example naturally as certain spaces of curves on finite dimensional Riemannian manifolds. A thorough treatment is given in [168].

1.5 Existence of Geodesics on Compact Manifolds

In the preceding section, we have derived the local existence and uniqueness of geodesics on Riemannian manifolds. In this section, we address the global issue and show the existence of shortest (geodesic) connections between any two points of arbitrary distance on a given compact Riemannian manifold. In fact, we shall be able to produce a geodesic in any given homotopy class of curves with fixed endpoints, as well as in any homotopy class of closed curves.

We recall the notion of homotopy between curves (see Appendix B):

Definition 1.5.1. Two curves γ_0, γ_1 on a manifold M with common initial and end points p and q, i.e. two continuous maps

$$\gamma_0, \gamma_1 : I = [0,1] \to M$$

with $\gamma_0(0) = \gamma_1(0) = p$, $\gamma_0(1) = \gamma_1(1) = q$, are called *homotopic* if there exists a continuous map

$$\Gamma : I \times I \to M$$

with

$$\begin{aligned} \Gamma(0,s) &= p, \quad \Gamma(1,s) = q && \text{for all } s \in I \\ \Gamma(t,0) &= \gamma_0(t), \quad \Gamma(t,1) = \gamma_1(t) && \text{for all } t \in I \end{aligned}$$

Two closed curves c_0, c_1 in M, i.e. two continuous maps

$$c_0, c_1 : S^1 \to M,$$

are called *homotopic* if there exists a continuous map

$$c : S^1 \times I \to M$$

with

$$c(t,0) = c_0(t), c(t,1) = c_1(t) \quad \text{for all } t \in S^1$$

(S^1, as usual, is the unit circle parametrized by $[0, 2\pi)$.)

Lemma 1.5.1. *The concept of homotopy defines an equivalence relation on the set of all curves in M with fixed initial and end points as well as on the set of all closed curves in M.*

The *proof* is elementary. □

With the help of this concept, we now want to show the existence of geodesics:

Theorem 1.5.1. *Let M be a compact Riemannian manifold, $p, q \in M$. Then there exists a geodesic in every homotopy class of curves from p to q, and this geodesic may be chosen as a shortest curve in its homotopy class. Likewise, every homotopy class of closed curves in M contains a curve which is shortest and geodesic.*

Proof. Since the proof is the same in both cases, we shall only consider the case of closed curves. □

As a preparation, we shall first show

Lemma 1.5.2. *Let M be a compact Riemannian manifold, $\rho_0 > 0$ as in Corollary 1.4.4. Let $\gamma_0, \gamma_1 : S^1 \to M$ be curves with*

$$d(\gamma_0(t), \gamma_1(t)) \leq \rho_0 \quad \textit{for all } t \in S^1.$$

Then γ_0 and γ_1 are homotopic.

Proof. For any $t \in S^1$ let $c_t(s) : I \to M$ be the unique shortest geodesic from $\gamma_0(t)$ to $\gamma_1(t)$ (Corollary 1.4.4), as usual parametrized proportionally to arc length. Since c_t depends continuously on its end points by Corollary 1.4.4, hence on t,

$$\Gamma(t, s) := c_t(s)$$

is continuous and yields the desired homotopy. □

Proof of Theorem 1.5.1. Let $(\gamma_n)_{n\in\mathbb{N}}$ be a minimizing sequence for arc length in the given homotopy class. Here and in the sequel, all curves are parametrized proportionally to arc length. We may assume w.l.o.g. that the curves γ_n are piecewise geodesic; namely, for each curve, we may find $t_0 = 0 < t_1 < t_2 < \ldots < t_m < t_{m+1} = 2\pi$ with the property that

$$L(\gamma_{n|[t_{j-1},t_j]}) \leq \rho_0/2 (\rho_0 \text{ as in Corollary 1.4.4}),$$

$$\text{for } j = 1, \ldots, m+1 \text{ with } t_{m+1} := 2\pi).$$

Replacing $\gamma_{n|[t_{j-1},t_j]}$ by the shortest geodesic arc between $\gamma_n(t_{j-1})$ and $\gamma_n(t_j)$, we obtain a curve which is homotopic to and not longer than γ_n (the same argument also shows that each homotopy class does contain curves of finite length).

We may thus assume that for any γ_n there exist points $p_{0,n}, \ldots, p_{m,n}$ for which $d(p_{j-1,n}, p_{j,n}) \leq \rho_0$ $(p_{m+1,n} := p_{0,n}, j = 1, \ldots, m+1)$ and for which γ_n contains the shortest geodesic arc between $p_{j-1,n}$ and $p_{j,n}$. Since the lengths of the γ_n are bounded as they constitute a minimizing sequence, we may also assume that m is independent of n. After selection of a subsequence, by the compactness of M, the points $p_{0,n}, \ldots, p_{m,n}$ converge to points $p_0, \ldots, p_m$, for $n \to \infty$. The segment of γ_n between $p_{j-1,n}$ and $p_{j,n}$ then converges to the shortest geodesic arc between p_{j-1} and p_j, for example by Corollary 1.4.4. The union of these geodesic segments yields a curve γ. By Lemma 1.5.2, γ is homotopic to the γ_n, and

$$L(\gamma) = \lim_{n\to\infty} L(\gamma_n),$$

and since the curves γ_n are minimizing sequence for the length in their homotopy class, γ is a shortest curve in this class. Therefore, γ has to be geodesic. Namely, otherwise, there would exist points p and q on γ for which one of the two segments of γ between p and q would have length at most ρ_0, but would not be geodesic. By Corollary 1.4.4, γ could then be shortened by replacing this segment by the shortest geodesic arc between p and q. By the argument of Lemma 1.5.2, this does not change the homotopy class, and we obtain a contradiction to the minimizing property of γ. γ thus is the desired closed geodesic. □

Corollary 1.5.1. *On any compact Riemannian manifold M_1, any two points p, q can be connected by a curve of shortest length, and this curve is geodesic.*

Proof. Minimize over all curves between p and q (and not only over those in a fixed homotopy class) as in the proof of Theorem 1.5.1. □

We also show

Theorem 1.5.2. *Let M be a compact Riemannian manifold. Then for any $p \in M$, the exponential map $\exp_p$ is defined on all of T_pM, and any geodesic may be extended indefinitely in each direction.*

Proof. For $v \in T_pM$, let

$$\Lambda := \{t \in \mathbb{R}^+ : c_v \text{ is defined on } [-t, t]\},$$

where c_v is, as usual, the geodesic with $c_v(0) = p, \dot{c}_v(0) = v$. It follows from $c_v(-t) = c_{-v}(t)$ that c_v may also be defined for negative t, at the moment at least for those with sufficiently small absolute value. Theorem 1.4.2 implies $\Lambda \neq \emptyset$. The compactness of M implies the closedness of Λ. We shall now show openness of Λ : Let c_v be defined on $[-t, t]$; for example $\dot{c}_v(t) = w \in T_{c_v(t)}M$. By Theorem 1.4.2 there exists a geodesic $\gamma_w(s)$ with $\gamma_w(0) = c_v(t)$, $\dot{\gamma}_w(0) = \dot{c}_v(t)$, for $s \in [0, \varepsilon]$ and $\varepsilon > 0$. Putting $c_v(t+s) = \gamma_w(s)$ for $s \in [0, \varepsilon]$, we have extended c_v to $[-t, t+\varepsilon]$. Analogously, c_v may be extended in the direction of negative t. This implies openness of Λ, hence $\Lambda = \mathbb{R}^+$. The claims follow easily. □

Perspectives. For an axiomatic approach towards the construction of closed geodesics on the basis of local existence and uniqueness, see [152].

1.6 The Heat Flow and the Existence of Geodesics

In the preceding section, we have derived the global existence of geodesics from the local existence and uniqueness of geodesic connections between points. In this section, we shall present an alternative method that uses methods from partial differential equations instead. This section thus serves as a first introduction to methods of geometric analysis. A reader who wishes to understand the geometry first may therefore skip this section. Conversely, for a reader interested in analytical methods, this section should be a good starting point.

Our scheme developed here will use parabolic partial differential equations. The idea is to start with some curve (in the homotopy class under consideration) and let it evolve according to a partial differential equation that decreases its energy until the curve becomes geodesic in the limit of "time" going to infinity (in fact, this will constitute some gradient descent for the energy in an (infinite dimensional) space of curves). This is the so-called heat flow method

The methods we are going to present here can naturally prove all the statements of Theorem 1.5.1. Since we do not wish to be repetitive, however, we shall confine ourselves here to the existence of closed geodesics

Theorem 1.6.1. *Let M be a compact Riemannian manifold. Then every homotopy class of closed curves in M contains a geodesic.*

Proof. In order to conform to conventions in the theory of partial differential equations, we need to slightly change our preceding notation. The parameter on a curve $c : [0,1] \to M$ will now be called s, that is, the points on the curve are $c(s)$, because we need t for the time parameter of the evolution that we now introduce. For technical convenience, we also parametrize our closed curves on the unit circle S^1 instead of on the interval $[0,1]$ because we do not have to stipulate the closedness as an additional condition ($c(0) = c(1)$ in the preceding sections). We consider mappings

$$u : S^1 \times [0,\infty) \to M \text{ with arguments } s \in S^1, 0 \le t \tag{1.6.1}$$

and impose the partial differential equation

$$\frac{\partial}{\partial t}u^i(s,t) = \frac{\partial^2}{\partial s^2}u^i(s,t) + \Gamma^i_{jk}(u(s,t))\frac{\partial}{\partial s}u^j(s,t)\frac{\partial}{\partial s}u^k(s,t) \text{ for } s \in S^1, t \ge 0 \tag{1.6.2}$$

$$u(s,0) = \gamma(s) \text{ for } s \in S^1 \tag{1.6.3}$$

for some smooth curve $\gamma : S^1 \to M$ in the given homotopy class. (1.6.2) can also be abbreviated in obvious notation for partial derivatives as

$$u_t^i = u_{ss}^i + \Gamma_{jk}^i u_s^j u_s^k. \tag{1.6.4}$$

The proof will then consist of several steps:

1. A solution of (1.6.2) exists at least on some short time interval $[0, t_0)$ for some $t_0 > 0$. This implies more generally that the maximal interval of existence of a solution is nonempty and open.
2. For a solution $u(s,t)$, the "spatial" derivative $\frac{\partial}{\partial s} u(s,t)$ stays bounded (independently of t). We may then rewrite (1.6.4) as
$$u_t^i - u_{ss}^i = f \tag{1.6.5}$$
with some bounded function f and may apply the regularity theory for linear parabolic differential equations as presented in **??** to obtain a time-independent control of higher derivatives.
3. Therefore, when a solution exists on $[0,T)$, for $t \to T$, $u(s,t)$ will converge to a smooth curve $u(s,T)$. This curve can then be taken as new initial values to continue the solution beyond T. This implies that the maximal existence interval is also closed. Consequently, the solution will exist for all time $t > 0$.
4. $E(u(\cdot,t)$ is a decreasing function of t, in fact $\frac{d}{dt} E(u(\cdot,t) = -\int_{S^1} \|u_t(s,t)\|^2 ds$. Since this quantity is also bounded from below, because nonnegative, we can find a sequence $t_n \to \infty$ for which $u(\cdot, t_n)$ will converge to a a curve with $u_{ss}^i + \Gamma_{jk}^i u_s^j u_s^k = 0$, that is, a geodesic.
5. A convexity argument shows that this convergence not only takes place for some sequence $t_n \to \infty$, but generally for $t \to \infty$.

Step 1 is a general result from the theory of partial differential equations which follows by linearizing the equation at $t = 0$ and applying the implicit function theorem in Banach spaces, see A.3. Therefore, we shall not discuss this here any further.
For step 2, we compute, using the symmetry $g_{ij} = g_{ji}$ repeatedly,

$$\begin{aligned}&(\frac{\partial^2}{\partial s^2} - \frac{\partial}{\partial t})(g_{ij}(u(s,t))u_s^i(s,t)u_s^j(s,t)) \\ &= 2g_{ij}u_{ss}^i u_{ss}^j + 2g_{ij}(u_{sss}^i - u_{st}^i)u_s^j + 4g_{ij,k}u_s^k u_s^j u_{ss}^i - g_{ij,k}u_t^k u_s^i u_s^j + g_{ij,kl}u_s^k u_s^l u_s^i u_s^j.\end{aligned} \tag{1.6.6}$$

From (1.6.4), we obtain

$$u_{sss}^i - u_{st}^i = -\Gamma_{jk,l}^i u_s^l u_s^j u_s^k - 2\Gamma_{jk}^i u_{ss}^j u_s^k \tag{1.6.7}$$

which we can insert into (1.6.6). In order to simplify our computations, it is natural to use normal coordinates at the point under considerations so that all first derivatives of the metric g_{ij} and the Christoffel symbols Γ_{jk}^i vanish. Moreover, we then have

$$\Gamma_{jk,l}^i = \frac{1}{2}(g_{ij,kl} + g_{ik,jl} - g_{jk,il}). \tag{1.6.8}$$

Inserting this as well, we obtain altogether

$$(\frac{\partial^2}{\partial s^2} - \frac{\partial}{\partial t})(g_{ij}u_s^i u_s^j) = 2g_{ij}u_{ss}^i u_{ss}^j \tag{1.6.9}$$

because the terms with the second derivatives of g_{ij} cancel.[1] This implies

$$(\frac{\partial^2}{\partial s^2} - \frac{\partial}{\partial t})(g_{ij}u_s^i u_s^j) \geq 0, \tag{1.6.10}$$

that is, $g_{ij}u_s^i u_s^j$ is a subsolution of the heat equation. The parabolic maximum principle (Theorem A.3.1) then implies that

$$\sup_{s \in S^1} g_{ij}(u(s,t))u_s^i(s,t)u_s^j(s,t) \tag{1.6.11}$$

is a nonincreasing function of t. In particular,

$$g_{ij}(u(s,t))u_s^i(s,t)u_s^j(s,t) \leq K \tag{1.6.12}$$

for some constant that does not depend on t and s. Thus, we have (1.6.5) with some bounded function f. We also note that since M is assumed compact, our solution u will automatically stay bounded. We may therefore apply the estimates of Theorem A.3.2.[2]

By the first estimate in Theorem A.3.2, $u(s,t)$ therefore has Hölder continuous first derivatives with respect to s. Since f is given in terms of such first derivatives, f then is also Hölder continuous. By the second estimate in Theorem A.3.2, we then get higher estimates. In fact, this procedure can be iterated. Higher order estimates on u from the linear theory imply a corresponding control on f which in turn then yields even higher estimates from the linear theory. (This is the so-called bootstrapping method.) This completes step 2.

Step 3 is self-explanatory, and so, we may now turn to step 4. The computation to follow is a consequence of (1.6.9), but as it is easier than the derivation of that formula, we do it directly.

$$\begin{aligned}
\frac{d}{dt}E(u(\cdot,t)) &= \frac{1}{2}\frac{\partial}{\partial t}\int_{S^1} g_{ij}(u(s,t))u_s^i(s,t)u_s^j(s,t) \\
&= \frac{1}{2}\int_{S^1}(2g_{ij}u_{st}^i u_s^j + g_{ij,k}u_t^k u_s^i u_s^j) \\
&= \frac{1}{2}\int_{S^1}(-2g_{ij}u_{ss}^i u_t^j - 2g_{ij,k}u_s^k u_t^i u_s^j + g_{ij,k}u_t^k u_s^i u_s^j) \text{ (integrating by parts)} \\
&= -\int_{S^1} g_{ij}u_t^i u_t^j \text{ by (1.6.4)}
\end{aligned} \tag{1.6.13}$$

[1] We shall see a deeper geometric interpretation of the computation leading to (1.6.9) in 7.2B below.

[2] These estimates are local estimates on the domain, and therefore, we have to make sure that for suitable regions $\Omega \times (t_1, t_2)$ in $S^1 \times [0, \infty)$, the image of u on such a region stays in the same coordinate chart in which we write our equation (1.6.4). First of all, since we already have derived a bound on u_s, in particular u is uniformly continuous w.r.t. s. As a solution of the heat equation, u is also continuous w.r.t. to t so that we may apply the estimates locally in time. The uniform continuity w.r.t. to t will be derived shortly.

Since E is nonnegative and the integrand also satisfies pointwise estimates by step 2, we obtain the conclusion of step 4. Finally, we find by similar computations as above (again in normal coordinates) from (1.6.13)

$$\begin{aligned}\frac{d^2}{dt^2}E(u(\cdot,t)) &= -\frac{\partial}{\partial t}\int_{S^1} g_{ij}u_t^i u_t^j \\ &= -\int_{S^1} 2g_{ij}u_{tt}^i u_t^j \\ &= -\int_{S^1} 2g_{ij}u_{sst}^i u_s^j \\ &= \int_{S^1} 2g_{ij}u_{st}^i u_{st}^j \geq 0. \end{aligned} \tag{1.6.14}$$

Thus, the energy $E(u(\cdot,t))$ is a convex function of t, and since we already know that $\frac{d}{dt}E(u(\cdot,t_n)) \to 0$ for some sequence $t_n \to \infty$, we conclude that $\frac{d}{dt}E(u(\cdot,t)) \to 0$ for $t \to \infty$. Thus, again invoking our pointwise estimates, $u_t(s,t) \to 0$ for $t \to \infty$. This implies that $u(s) = \lim_{t\to\infty} u(s,t)$ exists and is geodesic.
This completes the proof. □

We remark that the closed geodesic produced by the heat flow method need not be the shortest curve in its homotopy class. The reason is simple: When the initial curve γ for the heat flow (1.6.2) happens to be a closed geodesic already, the heat flow will stay there, that is $u(s,t) = \gamma(s)$ for all $t \geq 0$. In particular, if γ is a closed geodesic that is not the shortest one in its homotopy class, the heat flow with those initial values will fail to produce a shortest one.

1.7 Existence of Geodesics on Complete Manifolds

In this section, we want to address the question whether the results of Theorem 1.5.2 continue to hold for a more general class of Riemannian manifolds than the compact ones. Obviously, they do hold for Euclidean space which is not compact, but they do not hold for any proper open subset of Euclidean space, essentially since such a set is not complete. It will turn out that completeness will be the right condition for extending Theorem 1.5.2.

Definition 1.7.1. A Riemannian manifold M is *geodesically complete* if for all $p \in M$, the exponential map $\exp_p$ is defined on all of T_pM, or, in other words, if any geodesic $c(t)$ with $c(0) = p$ is defined for all $t \in \mathbb{R}$.

We can now state the **Theorem of Hopf-Rinow.**

Theorem 1.7.1. *Let M be a Riemannian manifold. The following statements are equivalent:*

(i) *M is complete as a metric space (or equivalently, it is complete as a topological space w.r.t. its underlying topology, see Corollary 1.4.1).*

(ii) *The closed and bounded subsets of M are compact.*

(iii) *There exists $p \in M$ for which $\exp_p$ is defined on all of T_pM.*

(iv) *M is geodesically complete, i.e. for every $p \in M, \exp_p$ is defined on all of T_pM.*

Furthermore, each of the statements (i) – (iv) implies

(v) *Any two points $p, q \in M$ can be joined by a geodesic of length $d(p, q)$, i.e. by a geodesic of shortest length.*

Proof. We shall first prove that if $\exp_p$ is defined on all of T_pM, then any $q \in M$ can be connected with p by a shortest geodesic. In particular, this will show the implication (iv) $\Rightarrow$ (v).

For this purpose, let
$$r := d(p, q),$$
and let $\rho > 0$ be given by Corollary 1.4.2, let $p_0 \in \partial B(p, \rho)$ be a point where the continuous function $d(q, \cdot)$ attains its minimum on the compact set $\partial B(p, \rho)$. Then $p_0 = \exp_p \rho V$, for some $V \in T_pM$. We consider the geodesic
$$c(t) := \exp_p tV,$$
and we want to show that
$$c(r) = q. \tag{1.7.1}$$
$c_{|[0,r]}$ will then be a shortest geodesic from p to q.

For this purpose, let
$$I := \{t \in [0, r] : d(c(t), q) = r - t\}.$$
(1.7.1)) means $r \in I$, and we shall show $I = [0, r]$ for that purpose. I is not empty, as it contains 0 by definition of r, and it is closed for continuity reasons. $I = [0, r]$ will therefore follow if we can show openness of I.

Let $t_0 \in I$. Let $\rho_1 > 0$ be the radius of Corollary 1.4.2 corresponding to the point $c(t_0) \in M$. W.l.o.g. $\rho_1 \leq r - t_0$. Let $p_1 \in \partial B(c(t_0), \rho_1)$ be a point where the continuous function $d(q, \cdot)$ assumes its minimum on the compact set $\partial B(c(t_0), \rho_1)$. Then
$$d(p, p_1) \geq d(p, q) - d(q, p_1). \tag{1.7.2}$$
Now for every curve γ from $c(t_0)$ to q, there exists some
$$\gamma(t) \in \partial B(c(t_0), \rho_1).$$

Hence

$$\begin{aligned} L(\gamma) &\geq d(c(t_0), \gamma(t)) + d(\gamma(t), q) \\ &= \rho_1 + d(\gamma(t), q) \\ &\geq \rho_1 + d(p_1, q) \quad \text{because of the minimizing property of } p_1. \end{aligned}$$

Hence also

$$d(q, c(t_0)) \geq \rho_1 + d(p_1, q) \tag{1.7.3}$$

and by the triangle inequality, we then actually must have equality. Inserting (1.7.3) into (1.7.2) and recalling $d(q, c(t_0)) = r - t_0$ gives

$$d(p, p_1) \geq r - (r - t_0 - \rho_1) = t_0 + \rho_1.$$

On the other hand, there exists a curve from p to p_1 of length $t_0 + \rho_1$; namely one goes from p to $c(t_0)$ along c and then takes the geodesic from $c(t_0)$ to p_1 of length ρ_1. That curve thus is shortest and therefore has to be geodesic as shown in the proof of Theorem 1.5.1. By uniqueness of geodesics with given initial values, it has to coincide with c, and then

$$p_1 = c(t_0 + \rho_1).$$

Since we observed that equality has to hold in (1.7.3), we get

$$d(q, c(t_0 + \rho_1)) = r - (t_0 + \rho_1),$$

hence

$$t_0 + \rho_1 \in I,$$

and openness of I follows, proving our claim.

It is now easy to complete the proof of Theorem 1.7.1:

(iv) $\Rightarrow$ (iii) is trivial.

(iii) $\Rightarrow$ (ii) Let $K \subset M$ be closed and bounded. Since bounded, $K \subset B(p, r)$ for some $r > 0$. By what we have shown in the beginning, any point in $B(p, r)$ can be connected with p by a geodesic (of length $\leq r$). Hence, $B(p, r)$ is the image of the compact ball in T_pM of radius r under the continuous map $\exp_p$. Hence, $B(p, r)$ is compact itself. Since K is assumed to be closed and shown to be contained in a compact set, it must be compact itself.

(ii) $\Rightarrow$ (i) Let $(p_n)_{n\in\mathbb{N}} \subset M$ be a Cauchy sequence. It then is bounded, and, by (ii), its closure is compact. It therefore contains a convergent subsequence, and being Cauchy, it has to converge itself. This shows completeness of M.

(i) $\Rightarrow$ (iv) Let c be a geodesic in M, parametrized by arc length, and being defined on a maximal interval I. I then is nonempty, and by Theorem 1.4.2, it is also open. To show closedness, let $(t_n)_{n\in\mathbb{N}} \subset I$ converge to t.

Since

$$d(c(t_n), c(t_m)) \leq |t_n - t_m|$$

as c is parametrized by arc length, $c(t_n)$ is a Cauchy sequence, hence has a limit $p \in M$, because we assume M to be complete. Let $\rho > 0$ be as in Corollary 1.4.2.

Then $B(p,\rho)$ is compact, being the image of the compact ball of radius r in T_pM under the continuous map $\exp_p$. Therefore, the argument of Corollary 1.4.3 and Corollary 1.4.4 applies to show that there exists $\rho_0 > 0$ with the property that for any point $q \in B(p,\rho)$ any geodesic starting from q can be extended at least up to length ρ_0.

Since $c(t_n)$ converges to p, for all sufficiently large m, n

$$d(c(t_n), c(t_m)) \le |t_n - t_m| \le \rho_0/2$$

and

$$d(c(t_n), p), d(c(t_m), p) \le \rho_0.$$

Therefore, the shortest geodesic from $c(t_n)$ to $c(t_m)$ can be defined at least on the interval $[-\rho_0, \rho_0]$. This shortest geodesic, however, has to be a subarc of c, and c thus can be defined up to the parameter value $t_n + \rho_0$, in particular for t, showing closedness of I. □

1.8 Vector Bundles

Definition 1.8.1. A *(differentiable) vector bundle* of rank n consists of a total space E, a base M, and a projection $\pi : E \to M$, where E and M are differentiable manifolds, π is differentiable, each "fiber" $E_x := \pi^{-1}(x)$ for $x \in M$, carries the structure of an n-dimensional (real) vector space, and the following local triviality requirement is satisfied: For each $x \in M$, there exist a neighborhood U and a diffeomorphism

$$\varphi : \pi^{-1}(U) \to U \times \mathbb{R}^n$$

with the property that for every $y \in U$

$$\varphi_y := \varphi_{|E_y} : E_y \to \{y\} \times \mathbb{R}^n$$

is a vector space isomorphism, i.e. a bijective linear map. Such a pair (φ, U) is called a *bundle chart.*

In the sequel, we shall omit the word "differentiable" for a vector bundle. Often, a vector bundle will simply be denoted by its total space.

It is important to point out that a vector bundle is by definition locally, but not necessarily globally a product of base and fiber. A vector bundle which is isomorphic to $M \times \mathbb{R}^n$ ($n =$ rank) is called trivial.

A vector bundle may be considered as a family of vector spaces (all isomorphic to a fixed model $\mathbb{R}^n$) parametrized (in a locally trivial manner) by a manifold.

Let now (E, π, M) be a vector bundle of rank n, $(U_\alpha)_{\alpha \in A}$ a covering of M by open sets over which the bundle is trivial, and $\varphi_\alpha : \pi^{-1}(U_\alpha) \to U_\alpha \times \mathbb{R}^n$ be the corresponding local trivializations. For nonempty $U_\alpha \cap U_\beta$,we obtain *transition maps*

$$\varphi_{\beta\alpha} : U_\alpha \cap U_\beta \to \mathrm{Gl}\,(n, \mathbb{R})$$

by

$$\varphi_\beta \circ \varphi_\alpha^{-1}(x,v) = (x, \varphi_{\beta\alpha}(x)v) \quad \text{for}\, x \in U_\alpha \cap U_\beta, v \in \mathbb{R}^n, \tag{1.8.1}$$

where $\mathrm{Gl}(n,\mathbb{R})$ is the general linear group of bijective linear self maps of $\mathbb{R}^n$. The transition maps express the transformation behavior of a vector in the fiber under a change of local trivialization.

The transition maps satisfy

$$\varphi_{\alpha\alpha}(x) = \mathrm{id}_{\mathbb{R}^n} \quad \text{for} \quad x \in U_\alpha \tag{1.8.2}$$

$$\varphi_{\alpha\beta}(x)\varphi_{\beta\alpha}(x) = \mathrm{id}_{\mathbb{R}^n} \quad \text{for} \quad x \in U_\alpha \cap U_\beta \tag{1.8.3}$$

$$\varphi_{\alpha\gamma}(x)\varphi_{\gamma\beta}(x)\varphi_{\beta\alpha}(x) = \mathrm{id}_{\mathbb{R}^n} \quad \text{for} \quad x \in U_\alpha \cap U_\beta \cap U_\gamma\,. \tag{1.8.4}$$

These properties are direct consequences of (1.8.1).

A vector bundle can be reconstructed from its transition maps.

Theorem 1.8.1.

$$E = \coprod_{\alpha \in A} U_\alpha \times \mathbb{R}^n / \sim\,,$$

where $\amalg$ *denotes disjoint union, and the equivalence relation* $\sim$ *is defined by*

$$(x,v) \sim (y,w) :\Longleftrightarrow\ x = y \text{ and } w = \varphi_{\beta\alpha}(x)v \quad (x \in U_\alpha, y \in U_\beta,\, v, w \in \mathbb{R}^n)$$

Proof. This is a straightforward verification of the properties required in Definition 1.8.1. A reader who does not want to carry this out him/herself may consult [129]. □

Definition 1.8.2. Let G be a subgroup of $\mathrm{Gl}(n,\mathbb{R})$, for example $\mathrm{O}(n)$ or $\mathrm{SO}(n)$, the orthogonal or special orthogonal group. We say that a vector bundle has the *structure group* G if there exists an atlas of bundle charts for which all transition maps have their values in G.

Definition 1.8.3. Let (E,π,M) be a vector bundle. A *section* of E is a differentiable map $s : M \to E$ with $\pi \circ s = \mathrm{id}_M$. The space of sections of E is denoted by $\Gamma(E)$.

We have already seen an example of a vector bundle above, namely the tangent bundle TM of a differentiable manifold M.

Definition 1.8.4. A section of the tangent bundle TM of M is called a *vector field* on M.

Let now $f : M \to N$ be a differentiable map, (E,π,N) a vector bundle over N. We want to pull back the bundle via f, i.e. construct a bundle f^*E, for which the fiber over $x \in M$ is $E_{f(x)}$, the fiber over the image of x.

Definition 1.8.5. The *pulled back bundle* f^*E is the bundle over M with bundle charts $(\varphi \circ f, f^{-1}(U))$, where (φ, U) are bundle charts of E.

We now want to extend some algebraic concepts and constructions from vector spaces to vector bundles by performing them fiberwise. For example:

Definition 1.8.6. Let (E_1, π_1, M) and (E_2, π_2, M) be vector bundles over M. Let the differentiable map $f : E_1 \to E_2$ be fiber preserving, i.e.

$$\pi_2 \circ f = \pi_1,$$

and let the fiber maps $f_x : E_{1,x} \to E_{2,x}$ be linear, i.e. vector space homomorphisms. Then f is called a *bundle homomorphism.*

Definition 1.8.7. Let (E, π, M) be a vector bundle of rank n. Let $E' \subset E$, and suppose that for any $x \in M$ there exists a bundle chart (φ, U) with $x \in U$ and

$$\varphi(\pi^{-1}(U) \cap E') = U \times \mathbb{R}^m (\subset U \times \mathbb{R}^n, m \leq n).$$

The resulting vector bundle $(E', \pi_{|E'}, M)$ is called *subbundle* of E of rank m.

Let us discuss an example: $S^1 = \{x \in \mathbb{R}^2 : |x|^2 = 1\}$ is a submanifold of $\mathbb{R}^2$. If we restrict the tangent bundle $T\mathbb{R}^2$ of $\mathbb{R}^2$ to S^1, we obtain a bundle E over S^1 that is isomorphic to $S^1 \times \mathbb{R}^2$. The tangent bundle of S^1 has fiber $T_x S^1 = \{y \in \mathbb{R}^2 : x \cdot y = 0\} \subset \mathbb{R}^2$ (where the dot $\cdot$ denotes the Euclidean scalar product). TS^1 is a subbundle of $T\mathbb{R}^2|S^1$; the reader is invited to write down explicit bundle charts.

Definition 1.8.8. Let (E_1, π_1, M) and (E_2, π_2, M) be vector bundles over M. The *Cartesian product* of E_1 and E_2 is the vector bundle over M with fiber $E_{1,x} \times E_{2,x}$ and bundle charts $(\varphi_\alpha \times \psi_\beta, U_\alpha \cap V_\beta)$, where $(\varphi_\alpha, U_\alpha)$ and (ψ_β, V_β) are bundle charts for E_1 and E_2 resp., and

$$(\varphi_\alpha \times \psi_\beta)(x, (v, w)) := (\varphi_\alpha(x, v), \psi_\beta(x, w)) \quad (v \in E_{1,x}, w \in E_{2,x}).$$

Thus, the *product bundle* is simply the bundle with fiber over $x \in M$ being the product of the fibers of E_1 and E_2 over x. By this pattern, all constructions for vector spaces can be extended to vector bundles. Of particular importance for us will be dual space, exterior and tensor product. Let us briefly recall the definition of the latter:

Let V and W be vector spaces (as always over $\mathbb{R}$) of dimension m and n, resp., and let $(e_1, \ldots, e_m)$ and $(f_1, \ldots, f_n)$ be bases. Then $V \otimes W$ is the vector space of dimension mn spanned by the basis $(e_i \otimes f_j)_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}}$. There exists a canonical bilinear map

$$L : V \times W \to V \otimes W$$

mapping $(a^i e_i, b^j f_j)$ onto $a^i b^j e_i \otimes f_j$

One may then also define the *tensor product* of more than two vector spaces in an associative manner.

Definition 1.8.9. Let M be a differentiable manifold, $x \in M$. The vector space dual to the tangent space $T_x M$ to $\mathbb{R}$ is called the *cotangent space* of M at the point x and denoted by $T_x^* M$. The vector bundle over M whose fibers are the cotangent spaces of M is called the *cotangent bundle* of M and denoted by $T^* M$. Elements of $T^* M$ are called *cotangent vectors,* sections of $T^* M$ are 1-*forms.*

We now want to study the transformation behavior of cotangent vectors. Let $(e_i)_{i=1,\dots,d}$ be a basis of T_xM and $(\omega^j)_{j=1,\dots,d}$ the dual basis of T_x^*M, i.e.

$$\omega^j(e_i) = \delta_i^j = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j. \end{cases}$$

Moreover, let $v = v^i e_i \in T_xM$, $\eta = \eta_j \omega^j \in T_x^*M$. We have $\eta(v) = \eta_i v^i$. Let the bases (e_i) and (ω^j) be given by local coordinates, i.e.

$$e_i = \frac{\partial}{\partial x^i}, \quad \omega^j = dx^j.$$

Let now f be a coordinate change. v is transformed to

$$f_*(v) := v^i \frac{\partial f^\alpha}{\partial x^i} \frac{\partial}{\partial f^\alpha}.$$

η then has to be transformed to

$$f^*(\eta) := \eta_j \frac{\partial x^j}{\partial f^\beta} df^\beta$$

because in this case

$$f^*(\eta)(f_*(v)) = \eta_j \frac{\partial x^j}{\partial f^\alpha} v^i \frac{\partial f^\alpha}{\partial x^i} = \eta_i v^i = \eta(v).$$

Thus a tangent vector transforms with the functional matrix of the coordinate change whereas a cotangent vector transforms with the transposed inverse of this matrix. This different transformation behavior is expressed by the following definition:

Definition 1.8.10. A p times *contravariant* and q times *covariant* tensor on a differentiable manifold M is a section of

$$\underbrace{TM \otimes \ldots \otimes TM}_{p \text{ times}} \quad \otimes \quad \underbrace{T^*M \otimes \ldots \otimes T^*M}_{q \text{ times}}.$$

Actually, one should speak of a *tensor field,* because "tensor" often also means an element of the corresponding fibers, in the same manner, as a (tangent) vector is an element of T_xM and a vector field a section of TM.

If f is a coordinate change, a p times contravariant and q times covariant tensor is transformed p times by the matrix (df) and q times by the matrix $(df^{-1})^t$.

Lemma 1.8.1. *A Riemannian metric on a differentiable manifold M is a two times covariant (and symmetric and positive definite) tensor on M.*

Proof. From the formula (1.4.3) for the transformation behavior of a Riemannian metric. □

A Riemannian metric thus is a section of $T^*M \otimes T^*M$. We consequently write the metric in local coordinates as

$$g_{ij}(x)dx^i \otimes dx^j.$$

Theorem 1.8.2. *The tangent bundle of a Riemannian manifold M of dimension d has structure group* $\mathrm{O}(d)$.

Proof. Let (f, U) be a bundle chart for TM,

$$f : \pi^{-1}(U) \to U \times \mathbb{R}^d.$$

Let $e_1, \dots, e_d$ be the canonical basis vectors of $\mathbb{R}^d$, and let $v_1, \dots, v_d$ be the sections of $\pi^{-1}(U)$ with $f(v_i) = e_i, i = 1, \dots, d$. Applying the Gram-Schmidt orthogonalization procedure to $v_1(x), \dots, v_d(x)$ for each $x \in U$ we obtain sections $w_1, \dots, w_d$ of $\pi^{-1}(U)$ for which $w_1(x), \dots, w_d(x)$ are an orthonormal basis w.r.t. the Riemannian metric on T_xM, for each $x \in U$. By

$$\begin{aligned} f' : \pi^{-1}(U) &\to U \times \mathbb{R}^d \\ \lambda^i w_i(x) &\mapsto (x, \lambda^1, \dots, \lambda^d) \end{aligned}$$

we then get a bundle chart which maps the basis $w_1(x), \dots, w_d(x)$, i.e. an orthonormal basis w.r.t. the Riemannian metric, for each $x \in U$ onto an Euclidean orthonormal basis of $\mathbb{R}^d$. We apply this orthonormalization process for each bundle chart and obtain a new bundle atlas whose transition maps always map an Euclidean orthonormal basis of $\mathbb{R}^d$ into another such basis, and are hence in $\mathrm{O}(d)$. □

We want to point out, however, that in general there do not exist local coordinates for which $w_i(x) = \frac{\partial}{\partial x^i}$ for $i = 1, \dots, d$.

Corollary 1.8.1. *The tangent bundle of an oriented Riemannian manifold of dimension d has structure group* $\mathrm{SO}(d)$.

Proof. The orientation allows to select an atlas for which all transition maps have positive functional determinant. From this, one sees that we also may obtain transition functions for the tangent bundle with positive determinant. The orthonormalization process of Theorem 1.8.2 preserves the positivity of the determinant, and thus, in the oriented case, we obtain a new bundle atlas with transition maps in $\mathrm{SO}(d)$. □

Definition 1.8.11. Let (E, π, M) be a a vector bundle. A *bundle metric* is given by a family of scalar products on the fibers E_x, depending smoothly on $x \in M$.

In the same manner as Theorem 1.8.2, one shows

Theorem 1.8.3. *Each vector bundle (E,π,M) of rank n with a bundle metric has structure group $\mathrm{O}(n)$. In particular, there exist bundle charts $(f,U), f:\pi^{-1}(U)\to U\times\mathbb{R}^n$, for which for all $x\in U$, $f^{-1}(x,(e_1,\dots,e_n))$ is an orthonormal basis of E_x $(e_1,\dots,e_n$ is an orthonormal basis of $\mathbb{R}^n)$.* □

Definition 1.8.12. The bundle charts of Theorem 1.8.3 are called *metric.*

In the same manner as Theorem 1.4.1, one shows

Theorem 1.8.4. *Each vector bundle can be equipped with a bundle metric.* □

It will be more important for us, however, that a Riemannian metric automatically induces bundle metrics on all tensor bundles over M. The metric of the cotangent bundle is given in local coordinates by

$$\langle\omega,\eta\rangle = g^{ij}\omega_i\eta_j \quad \text{for } \omega=\omega_i dx^i, \eta=\eta_i dx^i. \tag{1.8.5}$$

(We recall that (g^{ij}) is the matrix inverse to (g_{ij})).

Namely, this expression has the correct transformation behavior under coordinate changes: If $w\mapsto x(w)$ is a coordinate change, we get

$$\omega_i dx^i = \omega_i \frac{\partial x^i}{\partial w^\alpha} dw^\alpha =: \tilde{\omega}_\alpha dw^\alpha,$$

while g^{ij} is transformed into

$$h^{\alpha\beta} = g^{ij}\frac{\partial w^\alpha}{\partial x^i}\frac{\partial w^\beta}{\partial x^j},$$

and

$$h^{\alpha\beta}\tilde{\omega}_\alpha\tilde{\eta}_\beta = g^{ij}\omega_i\eta_j.$$

Moreover, we get

$$\|\omega(x)\| = \sup\{\omega(x)(v) : v\in T_xM, \|v\|=1\}.$$

A Riemannian metric also induces an identification between TM and T^*M :

$$v = v^i\frac{\partial}{\partial x^i} \text{ corresponds to } \omega=\omega_j dx^j$$

$$\begin{aligned} \text{with} \quad \omega_j &= g_{ij}v^i \\ \text{or} \quad v^i &= g^{ij}\omega_j\,. \end{aligned}$$

(1.8.5) may also be justified as follows:

Under this identification, to $v\in T_xM$ there corresponds a 1-form $\omega\in T_x^*M$ via

$$\omega(w) := \langle v,w\rangle \quad \text{for all } w$$

and (1.8.5) means then that

$$\|\omega\| = \|v\|.$$

For example, on $TM \otimes TM$, the metric is given by

$$\langle v \otimes w, \xi \otimes \eta \rangle = g_{ij} v^i \xi^j g_{k\ell} w^k \eta^l \tag{1.8.6}$$

($v = v^i \frac{\partial}{\partial x^i}$ etc. in local coordinates).

Definition 1.8.13. A local orthonormal basis of $T_x M$ of the type obtained in Theorem 1.8.3 is called an *(orthonormal) frame field.*

We put

$$\Lambda^p(T_x^* M) := \underbrace{T_x^* M \wedge \ldots \wedge T_x^* M}_{p \text{ times}} \quad \text{(exterior product).}$$

On $\Lambda^p(T_x^* M)$, we have two important operations: First, the exterior product by $\eta \in T_x^* M = \Lambda^1(T_x^* M)$:

$$\begin{aligned} \Lambda^p(T_x^* M) &\to \Lambda^{p+1}(T_x^* M) \\ \omega &\longmapsto \epsilon(\eta)\omega := \eta \wedge \omega. \end{aligned}$$

Second, the interior product or contraction by an element $v \in T_x M$:

$$\begin{aligned} \Lambda^p(T_x^* M) &\to \Lambda^{p-1}(T_x^* M) \\ \omega &\longmapsto \iota(v)\omega \end{aligned}$$

with

$$(\iota(v)\omega(v_1, \ldots, v_{p-1}) := \omega(v, v_1, \ldots, v_{p-1})$$
$$\text{for } v, v_1, \ldots, v_{p-1} \in T_x M.$$

In fact, such constructions may be carried out with any vector space W and its dual W^* in place of $T_x^* M$ and $T_x M$. This will be relevant in §1.11.

The vector bundle over M with fiber $\Lambda^p(T_x^* M)$ over x is then denoted by $\Lambda^p(M)$.

Definition 1.8.14. The space of sections of $\Lambda^p(M)$ is denoted by $\Omega^p(M)$, i.e. $\Omega^p(M) = \Gamma(\Lambda^p(M))$. Elements of $\Omega^p(M)$ are called *(exterior) p-forms.*

A p-form thus is a sum of terms of the form

$$\omega(x) = \eta(x) dx^{i_1} \wedge \ldots \wedge dx^{i_p}$$

where $\eta(x)$ is a smooth function and $(x^1, \ldots, x^d)$ are local coordinates.

Definition 1.8.15. The *exterior derivative* $d : \Omega^p(M) \to \Omega^{p+1}(M)$ $(p = 0, \ldots, d = \dim M)$ is defined through the formula

$$d(\eta(x) dx^{i_1} \wedge \ldots \wedge dx^{i_p}) = \frac{\partial \eta(x)}{\partial x^j} dx^j \wedge dx^{i_1} \wedge \ldots \wedge dx^{i_p}$$

and extended by linearity to all of $\Omega^p(M)$.

Lemma 1.8.2. *If* $\omega \in \Omega^p(M), \vartheta \in \Omega^q(M)$, *then* $d(\omega \wedge \vartheta) = d\omega \wedge \vartheta + (-1)^p \omega \wedge d\vartheta$.

Proof. This easily follows from the formula $\omega \wedge \vartheta = (-1)^{pq} \vartheta \wedge \omega$ and the definition of d. □

Let $f : M \to N$ be a differentiable map,

$$\omega(z) = \eta(z) dz^{i_1} \wedge \ldots \wedge dz^{i_p} \in \Omega^p(N).$$

We then define

$$f^*(\omega)(x) = \eta(f(x)) \frac{\partial f^{i_1}}{\partial x^{\alpha_1}} dx^{\alpha_1} \wedge \ldots \wedge \frac{\partial f^{i_p}}{\partial x^{\alpha_p}} dx^{\alpha_p}.$$

This obviously is the correct transformation formula for p-forms.

Lemma 1.8.3.

$$d(f^*(\omega)) = f^*(d\omega).$$

Proof. This easily follows from the transformation invariance

$$\frac{\partial \eta(z)}{\partial z^j} dz^j = \frac{\partial \eta(f(x))}{\partial z^j} \frac{\partial f^j}{\partial x^\alpha} dx^\alpha = \frac{\partial \eta(f(x))}{\partial x^\alpha} dx^\alpha.$$

□

Corollary 1.8.2. *d is independent of the choice of coordinates.*

Proof. Apply Lemma 1.8.3 to a coordinate transformation f. □

Theorem 1.8.5.

$$d \circ d = 0.$$

Proof. By linearity of d, it suffices to check the asserted identity on forms of the type

$$\omega(x) = f(x) dx^{i_1} \wedge \ldots \wedge dx^{i_p}.$$

Now

$$\begin{aligned} d \circ d(\omega(x)) &= d\Big(\frac{\partial f}{\partial x^j} dx^j \wedge dx^{i_1} \wedge \ldots \wedge dx^{i_p} \Big) \\ &= \frac{\partial^2 f}{\partial x^j \partial x^k} dx^k \wedge dx^j \wedge dx^{i_1} \wedge \ldots \wedge dx^{i_p} \\ &= 0, \end{aligned}$$

since $\frac{\partial^2 f}{\partial x^j \partial x^k} = \frac{\partial^2 f}{\partial x^k \partial x^j}$ (f is assumed to be smooth) and

$$dx^j \wedge dx^k = -dx^k \wedge dx^j.$$

$\square$

Let now M be a differentiable submanifold of the Riemannian manifold N; $\dim M = m, \dim N = n$. We saw already that M then also carries a Riemannian metric. For $x \in M$, we define

$$T_x^\perp M \subset T_x N$$

by

$$T_x^\perp M := \{v \in T_x N : \forall\, w \in T_x M : \langle v, w \rangle = 0\},$$

where $\langle .,. \rangle$, as usual, is the scalar product given by the Riemannian metric.

The spaces $T_x^\perp M$ are the fibers of a vector bundle $T^\perp M$ over M, and TM and $T^\perp M$ are both subbundles of $TN_{|M}$, the restriction of TN to M (in a more complicated manner: $TN_{|M} = i^* TN$, where $i : M \to N$ is the differentiable embedding of M as a submanifold of N). In order to see this, one may choose the first m basis vectors $v_1, \ldots, v_m$ of $TN_{|M}$ in the orthonormalization procedure of the proof of Theorem 1.8.2 in such a manner that they locally span TM.

Then TM is also locally spanned by $w_1, \ldots, w_m$ (notation as in the proof of Theorem 1.8.2), and the remaining basis vectors then span $T^\perp M$, and we have

$$\langle w_i, w_\alpha \rangle = 0 \text{ for } i = 1, \ldots, m, \alpha = m+1, \ldots, n.$$

Thus, $T^\perp M$ is the orthogonal complement of TM in $TN_{|M}$.

Definition 1.8.16. $T^\perp M$ is called the *normal bundle* of M in N.

For our example of the submanifold S^1 of $\mathbb{R}^2$, $T^\perp S^1$ is the subbundle of $T\mathbb{R}^2_{|S^1}$, the restriction of $T\mathbb{R}^2$ to S^1, with fiber $T_x^\perp S^1 = \{\lambda x : \lambda \in \mathbb{R}\} \subset \mathbb{R}^2$.

We conclude this section with a consideration of the complex case – again, we remind the reader that is needed only in particular places, like §5.2.

Definition 1.8.17. A vector bundle E over a differentiable manifold M is called a *complex vector bundle* if each fiber $E_z = \pi^{-1}(z)$ is a complex vector space, i.e., isomorphic to $z \times \mathbb{C}^k$, and if that complex structure varies smoothly, that is, the local trivializations are of the form

$$\varphi : \pi^{-1}(U) \to U \times \mathbb{C}^k.$$

We thus have transition maps

$$\varphi_{\beta\alpha} : U_\alpha \cap U_\beta \to \mathrm{Gl}\,(k, \mathbb{C}).$$

Here, in contrast to the Definition 1.1.5 of a complex manifold, we neither require that the base M be complex nor that these transition maps be holomorphic. If, however, these conditions are satisfied, that is, M is a complex manifold and the transition maps are ho lomorphic, then we have a *holomorphic vector bundle* . On a complex manifold M, in local holomorphic coordinates, we have the 1-forms

$$dz^j := dx^j + idy^j,\ dz^{\bar{k}} := dx^j - idy^j$$

(recall (1.1.2)). We can then decompose the space Ω^k of k-forms into subspaces $\Omega^{p,q}$ with $p+q=k$. Namely, $\Omega^{p,q}$ is locally spanned by forms of the type

$$\omega(z) = \eta(z)dz^{i_1} \wedge \ldots \wedge dz^{i_p} \wedge dz^{\bar{j}_1} \wedge \ldots \wedge dz^{\bar{j}_q}.$$

Thus

$$\Omega^k(M) = \sum_{p+q=k} \Omega^{p,q}(M). \tag{1.8.7}$$

We can then let the differential operators

$$\partial = \frac{1}{2}\left(\frac{\partial}{\partial x^j} - i\frac{\partial}{\partial y^j}\right)(dx^j + idy^j) \quad \text{and} \quad \bar{\partial} = \frac{1}{2}\left(\frac{\partial}{\partial x^j} + i\frac{\partial}{\partial y^j}\right)(dx^j - idy^j) \tag{1.8.8}$$

operate on such a form by

$$\partial\omega = \frac{\partial\eta}{\partial z^i} dz^i \wedge dz^{i_1} \wedge \ldots \wedge dz^{i_p} \wedge dz^{\bar{j}_1} \wedge \ldots \wedge dz^{\bar{j}_q} \tag{1.8.9}$$

and

$$\bar{\partial}\omega = \frac{\partial\eta}{\partial z^{\bar{j}}} dz^{\bar{j}} \wedge dz^{i_1} \wedge \ldots \wedge dz^{i_p} \wedge dz^{\bar{j}_1} \wedge \ldots \wedge dz^{\bar{j}_q}. \tag{1.8.10}$$

The following important relations link them with the exterior derivative d:

Lemma 1.8.4. *The exterior derivative d satisfies*

$$d = \partial + \bar{\partial}. \tag{1.8.11}$$

Moreover,

$$\partial\partial = 0,\ \bar{\partial}\bar{\partial} = 0, \tag{1.8.12}$$

$$\partial\bar{\partial} = -\bar{\partial}\partial\,. \tag{1.8.13}$$

Proof. We have

$$\begin{aligned} \partial + \bar{\partial} =& \frac{1}{2}\left(\frac{\partial}{\partial x^j} - i\frac{\partial}{\partial y^j}\right)(dx^j + idy^j) + \frac{1}{2}\left(\frac{\partial}{\partial x^j} + i\frac{\partial}{\partial y^j}\right)(dx^j - idy^j) \\ =& \frac{\partial}{\partial x^j}dx^j + \frac{\partial}{\partial y^j}dy^j = d. \end{aligned}$$

Therefore,

$$0 = d^2 = (\partial + \bar{\partial})(\partial + \bar{\partial}) = \partial^2 + \partial\bar{\partial} + \bar{\partial}\partial + \bar{\partial}^2$$

and decomposing this into types yields (1.8.12) and (1.8.13). One may verify these relations also by direct computation, e.g.

$$\partial\bar{\partial} = \frac{\partial^2}{\partial z^j \partial z^{\bar{k}}} dz^j \wedge dz^{\bar{k}} = -\frac{\partial^2}{\partial z^{\bar{k}} \partial z^j} dz^{\bar{k}} \wedge dz^j = -\bar{\partial}\partial.$$

□

1.9 Integral Curves of Vector Fields. Lie Algebras

Let M be a differentiable manifold, X a vector field on M, i.e. a (smooth) section of the tangent bundle TM. X then defines a first order differential equation (or, more precisely, if $\dim M > 1$, a system of differential equations):

$$\dot{c} = X(c). \tag{1.9.1}$$

This means the following: For each $p \in M$, one wants to find an open interval $I = I_p$ around $0 \in \mathbb{R}$ and a solution of the following differential equation for $c : I \to M$

$$\begin{aligned} \frac{dc}{dt}(t) &= X(c(t)) \text{ for } t \in I \\ c(0) &= p\,. \end{aligned} \tag{1.9.2}$$

One checks in local coordinates that this is indeed a system of differential equations: in such coordinates, let $c(t)$ be given by

$$(c^1(t), \ldots, c^d(t)) \quad (d = \dim M)$$

and let X be represented by

$$X^i \frac{\partial}{\partial x^i}.$$

Then (1.9.2) becomes

$$\frac{dc^i}{dt}(t) = X^i(c(t)) \quad \text{for } i = 1, \ldots, d. \tag{1.9.3}$$

Since (1.9.3) has a unique solution for given initial value $c(0) = p$ by the Picard-Lindelöf theorem, we obtain

Lemma 1.9.1. *For each $p \in M$, there exist an open interval $I_p \subset \mathbb{R}$ with $0 \in I_p$ and a smooth curve*

$$c_p : I_p \to M$$

with

$$\frac{dc_p}{dt}(t) = X(c_p(t))$$
$$c_p(0) = p.$$

□

Since the solution also depends smoothly on the initial point p by the theory of ODE, we furthermore obtain

Lemma 1.9.2. *For each $p \in M$, there exist an open neighborhood U of p and an open interval I with $0 \in I$, with the property that for all $q \in U$, the curve c_q ($\dot{c}_q(t) = X(c_q(t)), c_q(0) = q$) is defined on I. The map $(t,q) \mapsto c_q(t)$ from $I \times U$ to M is smooth.* □

Definition 1.9.1. The map $(t,q) \mapsto c_q(t)$ is called the *local flow* of the vector field X. The curve c_q is called the *integral curve* of X through q.

For fixed q, one thus seeks a curve through q whose tangent vector at each point coincides with the value of X at this point, i.e. a curve which is always tangent to the vector field X. Now, however, we want to fix t and vary q; we put

$$\varphi_t(q) := c_q(t).$$

Theorem 1.9.1. *We have*

$$\varphi_t \circ \varphi_s(q) = \varphi_{t+s}(q), \quad \text{if } s, t, t+s \in I_q, \tag{1.9.4}$$

and if φ_t is defined on $U \subset M$, it maps U diffeomorphically onto its image.

Proof. We have

$$\dot{c}_q(t+s) = X(c_q(t+s)),$$

hence

$$c_q(t+s) = c_{c_q(s)}(t).$$

Starting from q, at time s one reaches the point $c_q(s)$, and if one proceeds a time t further, one reaches $c_q(t+s)$. One therefore reaches the same point if one walks from q on the integral curve for a time $t+s$, or if one walks a time t from $c_q(s)$. This shows (1.9.4). Inserting $t = -s$ into (1.9.4) for $s \in I_q$, we obtain

$$\varphi_{-s} \circ \varphi_s(q) = \varphi_0(q) = q.$$

Thus, the map φ_{-s} is the inverse of φ_s, and the diffeomorphism property follows. □

Corollary 1.9.1. *Each point in M is contained in precisely one integral curve for (1.9.1).*

Proof. Let $p \in M$. Then $p = c_p(0)$, and so, it is trivially contained in an integral curve. Assume now that $p = c_q(t)$. Then, by Theorem 1.9.1, $q = c_p(-t)$. Thus, any point whose flow line passes through p is contained in the same flow line, namely the one starting at p. Therefore, there is precisely one flow line going through p. □

We point out, however, that flow lines can reduce to single points; this happens for those points for which $X(p) = 0$. Also, flow lines in general are not closed even if the flow exists for all $t \in \mathbb{R}$. Namely, the points $\lim_{t \to \pm\infty} c_p(t)$ (assuming that these limits exist) need not be contained in the flow line through p.

Definition 1.9.2. A family $(\varphi_t)_{t \in I}$ (I open interval with $0 \in I$) of diffeomorphisms from M to M satisfying (1.9.4) is called a *local* 1*-parameter group of diffeomorphisms.*

In general, a local 1-parameter group need not be extendable to a group, since the maximal interval of definition I_q of c_q need not be all of $\mathbb{R}$. This is already seen by easy examples, e.g. $M = \mathbb{R}$, $X(\tau) = \tau^2 \frac{d}{d\tau}$, i.e. $\dot{c}(t) = c^2(t)$ as differential equation.

However

Theorem 1.9.2. *Let X be a vector field on M with compact support. Then the corresponding flow is defined for all $q \in M$ and all $t \in \mathbb{R}$, and the local* 1*-parameter group becomes a group of diffeomorphisms.*

Proof. By Lemma 1.9.2, for every $p \in M$ there exist a neighborhood U and $\varepsilon > 0$ such that for all $q \in U$, the curve c_q is defined on $(-\varepsilon, \varepsilon)$. Let now $\operatorname{supp} X \subset K$, K compact. K can then be covered by finitely many such neighborhoods, and we choose ε_0 as the smallest such ε.

Since for $q \notin K$ $X(q) = 0$,

$$\varphi_t(q) = c_q(t)$$

is defined on $(-\varepsilon_0, \varepsilon_0) \times M$, and for $|s|, |t| < \varepsilon_0/2$, we have the semigroup property (1.9.4).

Since the interval of existence $(-\varepsilon_0, \varepsilon_0)$ may be chosen uniformly for all q, one may iteratively extend the flow to all of $\mathbb{R}$. For this purpose, we write $t \in \mathbb{R}$ as

$$t = m\frac{\varepsilon_0}{2} + \rho \quad \text{with } m \in \mathbb{Z}, 0 \le \rho < \varepsilon_0/2$$

and put

$$\varphi_t := (\varphi_{\varepsilon_0/2})^m \circ \varphi_\rho$$

$(\varphi_t)_{t \in \mathbb{R}}$ then is the desired 1-parameter group. □

Corollary 1.9.2. *On a compact differentiable manifold, any vector field generates a* 1*-parameter group of diffeomorphisms.* □

The preceding is a geometric interpretation of systems of first order ODE on manifolds. However, also higher order systems of ODE may be reduced to first order systems by introducing additional independent variables. As an example, we want to study the system for geodesics, i.e. in local coordinates

$$\ddot{x}^i(t) + \Gamma^i_{jk}(x(t))\dot{x}^j(t)\dot{x}^k(t) = 0, \quad i = 1, \dots, d. \tag{1.9.5}$$

We want to transform this second order system into a first order system on the cotangent bundle T^*M. As usual, we locally trivialize T^*M by a chart

$$T^*M_{|U} \simeq U \times \mathbb{R}^d$$

with coordinates $(x^1, \dots, x^d, p_1, \dots, p_d)$.

We also put

$$H(x,p) = \frac{1}{2} g^{ij}(x) p_i p_j \quad (g^{ij}(x) g_{jk}(x) = \delta^i_k)\,. \tag{1.9.6}$$

(The transformation behavior of g^{ij} and p_k implies that H does not depend on the choice of coordinates.)

Theorem 1.9.3. *(1.9.5) is equivalent to the following system on T^*M :*

$$\begin{aligned} \dot{x}^i &= \frac{\partial H}{\partial p_i} = g^{ij}(x) p_j \\ \dot{p}_i &= -\frac{\partial H}{\partial x^i} = -\frac{1}{2} g^{jk}{}_{,i}(x) p_j p_k \quad (g^{jk}{}_{,i} := \frac{\partial}{\partial x^i} g^{jk})\,. \end{aligned} \tag{1.9.7}$$

Proof. From the first equation

$$\begin{aligned} \ddot{x}^i &= g^{ij}(x)\dot{p}_j + g^{ij}{}_{,k}(x)\dot{x}^k p_j \\ &= g^{ij}\dot{p}_j + g^{ij}{}_{,k}\dot{x}^k g_{j\ell}\dot{x}^\ell \end{aligned}$$

and with the second equation then

$$\begin{aligned} \ddot{x}^i &= -\frac{1}{2} g^{ij} g^{\ell k}{}_{,j} p_\ell p_k + g^{ij}{}_{,k} g_{j\ell}\dot{x}^k \dot{x}^\ell, \\ &= \frac{1}{2} g^{ij} g^{\ell m} g_{mn,j} g^{nk} g_{\ell r}\dot{x}^r g_{ks}\dot{x}^s - g^{im} g_{mn,k} g^{nj} g_{j\ell}\dot{x}^k \dot{x}^\ell \end{aligned}$$

using $g^{ij}{}_{,\ell} = -g^{im} g_{mn,\ell} g^{nj}$ (which follows from $g^{ij} g_{jk} = \delta^i_k$),

$$\begin{aligned} &= \frac{1}{2} g^{ij} g_{mn,j}\dot{x}^m \dot{x}^n - g^{im} g_{mn,k}\dot{x}^k \dot{x}^n \\ &= \frac{1}{2} g^{ij} (g_{mn,j} - g_{jn,m} - g_{jm,n})\dot{x}^m \dot{x}^n \end{aligned}$$

since $g_{mn,k}\dot{x}^k\dot{x}^n = \frac{1}{2}g_{mn,k}\dot{x}^k\dot{x}^n + \frac{1}{2}g_{mk,n}\dot{x}^k\dot{x}^n$ and after renumbering some indices,

$$= -\Gamma^i_{mn}\dot{x}^m\dot{x}^n.$$

□

Definition 1.9.3. The flow determined by (1.9.7) is called the *cogeodesic flow.* The *geodesic flow* on TM is obtained from the cogeodesic flow by the first equation of (1.9.7).

Thus, the geodesic lines are the projections of the integral curves of the geodesic flow onto M.

The reason for considering the cogeodesic instead of the geodesic flow is that the former is a Hamiltonian flow for the Hamiltonian H from (1.9.6).

We remark that by (1.9.7), we have along the integral curves

$$\frac{dH}{dt} = H_{x^i}\dot{x}^i + H_{p_i}\dot{p}_i = -\dot{p}_i\dot{x}^i + \dot{x}^i\dot{p}_i = 0.$$

Thus, the cogeodesic flow maps the set $E_x := \{(x,p) \in T^*M : H(x,p) = \lambda\}$ onto itself for every $\lambda \geq 0$. If M is compact, so are all E_λ. Hence, by Corollary 1.9.2, the geodesic flow is defined on all of E_λ, for every λ. Since $M = \cup_{\lambda \geq 0} E_\lambda$, Theorem 1.9.3 yields a new proof of Theorem 1.5.2. If $\psi : M \to N$ is a diffeomorphism between differentiable manifolds, and if X is a vector field on M, we define a vector field

$$Y = \psi_* X$$

on N by

$$Y(p) = d\psi(X(\psi^{-1}(p))). \tag{1.9.8}$$

Then

Lemma 1.9.3. *For any differentiable function $f : N \to \mathbb{R}$*

$$(\psi_* X)(f)(p) = X(f \circ \psi)(\psi^{-1}(p))\,. \tag{1.9.9}$$

Proof.

$$\begin{aligned}(\psi_* X)(f)(p) &= (d\psi \circ X)(f)(p)\\ &= (df \circ d\psi \circ X)(\psi^{-1}(p))\\ &= X(f \circ \psi)(\psi^{-1}(p)).\end{aligned}$$

□

If $\varphi : N \to P$ is another diffeomorphism, obviously

$$(\varphi \circ \psi)_* X = \varphi_*(\psi_*(X)). \tag{1.9.10}$$

Lemma 1.9.4. *Let X be a vector field on M, $\psi : M \to N$ a diffeomorphism. If the local 1-parameter group generated by X is given by φ_t, the local group generated by $\psi_* X$ is*

$$\psi \circ \varphi_t \circ \psi^{-1}.$$

Proof. $\psi \circ \varphi_t \circ \psi^{-1}$ is a local 1-parameter group, and therefore, by uniqueness of solutions of ODE, it suffices to show the claim near $t = 0$. Now

$$\begin{aligned}
\frac{d}{dt}(\psi \circ \varphi_t \circ \psi^{-1}(p))_{|t=0} &= d\psi\left(\frac{d}{dt}\varphi_t \circ \psi^{-1}(p)_{|t=0}\right) \\
&\quad (\text{where} \quad d\psi \text{ is evaluated at } \varphi_0 \circ \psi^{-1}(p)) = \psi^{-1}(p)) \\
&= d\psi X(\psi^{-1}(p)) \\
&= \psi_* X(p).
\end{aligned}$$

□

Definition 1.9.4. For vector fields X, Y on M, the *Lie bracket*

$$[X, Y]$$

is defined as the vector field

$$X^j \frac{\partial Y^i}{\partial x^j}\frac{\partial}{\partial x^i} - Y^j \frac{\partial X^i}{\partial x^j}\frac{\partial}{\partial x^i} \quad (X = X^i \frac{\partial}{\partial x^i}, Y = Y^i \frac{\partial}{\partial x^i}).$$

We say that the *vector fields* X and Y *commute,* if

$$[X, Y] = 0.$$

Lemma 1.9.5. *$[X, Y]$ is linear (over $\mathbb{R}$) in X and Y. For a differentiable function $f : M \to \mathbb{R}$, we have $[X, Y]f = X(Y(f)) - Y(X(f))$. Furthermore, the Jacobi identity holds:*

$$[[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0$$

for any three vector fields X, Y, Z.

Proof. In local coordinates with $X = X^i \frac{\partial}{\partial x^i}, Y = Y^i \frac{\partial}{\partial x^i}$, we have

$$[X, Y]f = X^j \frac{\partial Y^i}{\partial x^j}\frac{\partial f}{\partial x^i} - Y^j \frac{\partial X^i}{\partial x^j}\frac{\partial f}{\partial x^i} = X(Y(f)) - Y(X(f)) \tag{1.9.11}$$

and this is linear in f, X, Y. This implies the first two claims. The Jacobi identity follows by direct computation. □

Definition 1.9.5. A *Lie algebra* (over $\mathbb{R}$) is a real vector space V equipped with a bilinear map $[\cdot, \cdot] : V \times V \to V$, the *Lie bracket,* satisfying:

(i) $[X, X] = 0$ for all $X \in V$.

(ii) $[X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] = 0$ for all $X, Y, Z \in V$.

Corollary 1.9.3. *The space of vector fields on M, equipped with the Lie bracket, is a Lie algebra.* □

Lemma 1.9.6. *Let $\psi : M \to N$ be a diffeomorphism, X, Y vector fields on M. Then*

$$[\psi_* X, \psi_* Y] = \psi_* [X, Y]. \tag{1.9.12}$$

Thus, ψ_* induces a Lie algebra isomorphism.

Proof. Directly from Lemma 1.9.3. □

We now want to investigate how one might differentiate tensor fields. A function $f : M \to \mathbb{R}$, if smooth, may simply be differentiated at a point x by comparing its values at x with those at neighboring points. For a tensor field S, this is not possible any more, because the values of S at different points lie in different spaces, and it is not clear how to compare elements of different fibers. For this purpose, however, one might use a map F of one fiber onto another one, and an element v of the first fiber may then be compared with an element w of the second fiber by comparing $F(v)$ and w. One possibility to obtain such a map at least between neighboring fibers (which is sufficient for purposes of differentiation) is to use a local 1-parameter group $(\psi_t)_{t \in I}$ of diffeomorphisms. If for example $X = X^i \frac{\partial}{\partial x^i}$ is a vector field, we consider $(\psi_{-t})_* X(\psi_t(x))$. This yields a curve X_t in $T_x M$ (for $t \in I$), and such a curve may be differentiated. In particular,

$$(\psi_{-t})_* \frac{\partial}{\partial x^i}(\psi_t(x)) = \frac{\partial \psi_{-t}^k}{\partial x^i} \frac{\partial}{\partial x^k} \text{ (evaluated at } \psi_t(x)). \tag{1.9.13}$$

(In general, one has for $\varphi : M \to N, \varphi_* \frac{\partial}{\partial x^i} = \frac{\partial \varphi^k}{\partial x^i} \frac{\partial}{\partial \varphi^k}$, but in case $M = N$ and x and $\varphi(x)$ are contained in the same coordinate neighborhood, of course $\frac{\partial}{\partial \varphi^k} = \frac{\partial}{\partial x^k}$).

If $\omega = \omega_i dx^i$ is a 1-form, we may simply consider

$$(\psi_t^*)(\omega)(x) = \omega_i(\psi_t(x)) \frac{\partial \psi_t^i}{\partial x^k} dx^k, \tag{1.9.14}$$

which is a curve in $T_x^* M$.

In general for a smooth map $\varphi : M \to N$ and a 1-form $\omega = \omega_i dz^i$ on N,

$$\varphi^* \omega := \omega_i(\varphi(x)) \frac{\partial z^i}{\partial x^k} dx^k; \tag{1.9.15}$$

note that φ need not be a diffeomorphism here.

Analogously, for a section $h = h_{ij} dz^i \otimes dz^j$, of $T^* N \otimes T^* N$

$$(\varphi^*) h = h_{ij} \frac{\partial z^i}{\partial x^k} \frac{\partial z^j}{\partial x^\ell} dx^k \otimes dx^\ell. \tag{1.9.16}$$

Finally, for a function $f : N \to \mathbb{R}$ of course

$$\varphi^* f = f \circ \varphi. \tag{1.9.17}$$

If $\varphi : M \to N$ is a diffeomorphism, and Y is a vector field on N, we put

$$\varphi^* Y := (\varphi^{-1})_* Y. \tag{1.9.18}$$

in order to unify our notation.

φ^* is then defined analogously for other contravariant tensors.

In particular, for a vector field X on M and a local group $(\psi_t)_{t\in I}$ as above:

$$(\psi_t^*)X = (\psi_{-t})_* X. \tag{1.9.19}$$

Definition 1.9.6. Let X be a vector field with a local 1-parameter group $(\psi_t)_{t\in I}$ of local diffeomorphisms, S a tensor field on M. The *Lie derivative* of S in the direction X is defined as

$$L_X S := \frac{d}{dt}(\psi_t^* S)_{|t=0}.$$

Theorem 1.9.4.

(i) Let $f : M \to \mathbb{R}$ be a (differentiable) function. Then

$$L_X(f) = df(X) = X(f).$$

(ii) Let Y be a vector field on M. Then

$$L_X Y = [X, Y].$$

(iii) Let $\omega = \omega_j dx^j$ be a 1-form on M. Then for $X = X^i \frac{\partial}{\partial x^i}$

$$L_X \omega = \left(\frac{\partial \omega_j}{\partial x^i} X^i + \frac{\partial X^i}{\partial x^j} \omega_i \right) dx^j.$$

Proof.

(i) $L_X(f) = \frac{d}{dt}\psi_t^* f_{|t=0} = \frac{d}{dt} f \circ \psi_{t_{|t=0}} = \frac{\partial f}{\partial x^i} X^i = X(f)$ (cf. (1.9.17).

(ii) $Y = Y^i \frac{\partial}{\partial x^i}$.

$$\begin{aligned}
L_X Y &= \frac{d}{dt}\psi_t^*(Y^i \frac{\partial}{\partial x^i})_{|t=0} \\
&= \frac{d}{dt}(\psi_{-t})_*(Y^i \frac{\partial}{\partial x^i})_{|t=0} \quad \text{by (1.9.19)} \\
&= \frac{d}{dt}(Y^i(\psi_t)\frac{\partial \psi^j_{-t}}{\partial x^i}\frac{\partial}{\partial x^j})_{|t=0} \quad \text{by (1.9.13), Lemma 1.9.3} \\
&= \frac{\partial Y^i}{\partial x^k} X^k \delta^j_i \frac{\partial}{\partial x^j} + Y^i(-\frac{\partial X^j}{\partial x^i})\frac{\partial}{\partial x^j}, \text{ since } \psi_0 = \text{id}, \frac{d}{dt}\psi_{-t_{|t=0}} = -X \\
&= (X^k \frac{\partial Y^j}{\partial x^k} - Y^k \frac{\partial X^j}{\partial x^k})\frac{\partial}{\partial x^j} \\
&= [X, Y].
\end{aligned}$$

(iii)

$$\begin{aligned}
L_X\omega &= \frac{d}{dt}(\psi_t^*\omega)_{|t=0} \\
&= \frac{d}{dt}(\omega_j(\psi_t)\frac{\partial\psi_t^j}{\partial x^k}dx^k)_{|t=0} \quad \text{by (1.9.14)} \\
&= \frac{\partial\omega^j}{\partial x^i}X^i\delta_k^j dx^k + \omega_j\frac{\partial X^j}{\partial x^k}dx^k, \text{ since } \psi_0 = \text{ id}, \ \frac{d}{dt}\psi_{t_{|t=0}} = X \\
&= (\frac{\partial\omega^j}{\partial x^i}X^i + \frac{\partial X^i}{\partial x^j}\omega_i)dx^j.
\end{aligned}$$

□

In this manner, also Lie derivatives of arbitrary tensor fields may be computed. For example for $h = h_{ij}dx^i \otimes dx^j$

$$\begin{aligned}
L_X h &= h_{ij,k}X^k dx^i \otimes dx^j + h_{ij}\frac{\partial X^i}{\partial x^k}dx^k \otimes dx^j + h_{ij}\frac{\partial X^j}{\partial x^k}dx^i \otimes dx^k \\
&= (h_{ij,k}X^k + h_{kj}\frac{\partial X^k}{\partial x^i} + h_{ik}\frac{\partial X^k}{\partial x^j})dx^i \otimes dx^j
\end{aligned} \tag{1.9.20}$$

Remark. For vector fields X, Y, Z and $\psi = \psi_t$, the local flow of X, Lemma 1.9.6 yields by differentiation at $t = 0$

$$L_X[Y, Z] = [L_X Y, Z] + [Y, L_X Z],$$

and with Theorem 1.9.4 (ii), we then obtain the Jacobi identity

$$\begin{aligned}
[X, [Y, Z]] &= [[X, Y], Z] + [Y, [X, Z]] \\
&= -[Z, [X, Y]] - [Y, [Z, X]].
\end{aligned}$$

Definition 1.9.7. Let M carry a Riemannian metric

$$g = g_{ij}dx^i \otimes dx^j.$$

A vector field X on M is called a *Killing field* or an *infinitesimal isometry* if

$$L_X(g) = 0. \tag{1.9.21}$$

Lemma 1.9.7. *A vector field X on a Riemannian manifold M is a Killing field if and only if the local 1-parameter group generated by X consists of local isometries.*

Proof. From (1.9.21)

$$\frac{d}{dt}(\psi_t^* g)_{|t=0} = 0. \tag{1.9.22}$$

Since this holds for every point of M, we obtain

$$\psi_t^* g = g \quad \text{for all } t \in I.$$

Therefore, the diffeomorphisms ψ_t are isometries. Conversely, if the ψ_t are isometries, (1.9.22) holds, hence also (1.9.21). □

Lemma 1.9.8. *The Killing fields of a Riemannian manifold constitute a Lie algebra.*

Proof. The space of all vector fields on a differentiable manifold constitute a Lie algebra by Corollary 1.9.3. The claim then follows if we show that the space of Killing fields is closed under the Lie bracket $[.,.]$, i.e. that for any two Killing fields X and $Y, [X,Y]$ is again a Killing field. This, however, follows from the following identity which was derived in the proof of Theorem 1.9.4 (ii):

$$[X,Y] = L_X Y = \frac{d}{dt} d\psi_{-t} Y(\psi_t)_{|t=0},$$

where $(\psi_t)_{t\in I}$ is the local group of isometries generated by X. Namely, for any fixed t,

$$\psi_{-t} \circ \varphi_s \circ \psi_t,$$

is the local group for $d\psi_{-t}Y(\psi_t)$, where $(\varphi_s)_{s\in I}$ is the local group generated by Y. Since ψ_t and φ_s are isometries, so are $\psi_{-t} \circ \varphi_s \circ \psi_t$.

It follows that

$$L_{[X,Y]} g = \frac{\partial^2}{\partial s \partial t} (\psi_{-t} \varphi_s \psi_t)^* g_{|s=t=0} = 0.$$

Thus, $[X,Y]$ indeed is a Killing field. □

1.10 Lie Groups

Definition 1.10.1. A *Lie group* is a group G carrying the structure of a differentiable manifold or, more generally, of a disjoint union of finitely many differentiable manifolds for which the following maps are differentiable:

$$\begin{aligned} G \times G &\to G \quad \text{(multiplication)} \\ (g,h) &\mapsto g \cdot h \end{aligned}$$

and

$$\begin{aligned} G &\to G \quad \text{(inverse)} \\ g &\mapsto g^{-1}. \end{aligned}$$

We say that G acts on a differentiable manifold M from the left if there is a differentiable map

$$\begin{aligned} G \times M &\to M \\ (g, x) &\mapsto gx \end{aligned}$$

that respects the Lie group structure of G in the sense that

$$g(hx) = (g \cdot h)x \quad \text{for all } g, h \in G,\ x \in M.$$

An action from the right is defined analogously.

The Lie groups we shall encounter will mostly be linear algebraic groups. In order to describe the most important ones, let V be a vector space over $\mathbb{R}$ of dimension n. We put

$$\mathrm{Gl}(V) := \{A : V \to V \text{ linear and bijective}\},$$

the vector space isomorphisms of V.

If V is equipped with a scalar product $\langle \cdot, \cdot \rangle$, we put

$$\mathrm{O}(V) := \{A \in \mathrm{Gl}(V) : \langle Av, Av \rangle = \langle v, v \rangle \text{ for all } v \in V.\}$$

and

$$\begin{aligned} \mathrm{SO}(V) := \{A \in \mathrm{O}(V) : &\text{ the matrix } \langle Ae_i, e_j \rangle_{i,j=1,\ldots,n} \text{ has positive} \\ &\text{determinant for some (and hence any) basis } e_1, \ldots, e_n \text{ of } V\}. \end{aligned}$$

(In the terminology of §2.1 below, one might express the last condition as: A transforms positive bases into positive bases.) Clearly $\mathrm{SO}(V) \subset \mathrm{O}(V)$. $\mathrm{Gl}(V)$, $\mathrm{SO}(V)$ and $\mathrm{O}(V)$ become Lie groups w.r.t. composition of linear maps. Since bijectivity is an open condition, the tangent space to $\mathrm{Gl}(V)$, for example at the identity linear map, i.e. the Lie algebra of $\mathrm{Gl}(V)$, can be identified with

$$\mathfrak{gl}(V) := \{X : V \to V \text{ linear}\},$$

the space of endomorphisms of V. The Lie algebra bracket is simply given by

$$[X, Y] = XY - YX.$$

The Lie algebra of $\mathrm{SO}(V)$ then is obtained by differentiating the relation $\langle Av, Aw \rangle = \langle v, w \rangle$, i.e. as

$$\mathfrak{so}(V) := \{X \in \mathfrak{gl}(V) : \langle Xv, w \rangle + \langle v, Xw \rangle = 0 \text{ for all } v, w \in V\},$$

the skew symmetric endomorphisms of V. (Of course, this is also the Lie algebra of $\mathrm{O}(V)$, and therefore in the sequel, we shall sometimes write $\mathfrak{o}(V)$ in place of $\mathfrak{so}(V)$.)

The relation between a Lie algebra and its Lie group is given by the exponential map which in the present case is simply

$$\mathrm{e}^X = \mathrm{Id} + X + \frac{1}{2}X^2 + \frac{1}{3!}X^3 + \cdots$$

For $t \in \mathbb{R}$, we have

$$\mathrm{e}^{tX} = \mathrm{Id} + tX + \frac{t^2}{2}X^2 + \dots$$

As the ordinary exponential map converges, this series converges for all $t \in \mathbb{R}$, and e^{tX} is continuous in t.

For $s, t \in \mathbb{R}$, we have

$$\mathrm{e}^{(s+t)X} = \mathrm{e}^{sX}\mathrm{e}^{tX}.$$

In particular

$$\mathrm{e}^{X}\mathrm{e}^{-X} = \mathrm{Id}$$

Therefore, e^X is always invertible, i.e. in $\mathrm{Gl}(V)$, with inverse given by e^{-X}. Thus, for each $X \in \mathfrak{gl}(V)$,

$$t \longmapsto \mathrm{e}^{tX}$$

yields a group homomorphism from $\mathbb{R}$ to $\mathrm{Gl}(V)$.

We assume that $\langle \cdot, \cdot \rangle$ is nondegenerate. Every $X \in \mathfrak{gl}(V)$ then has a adjoint X^* characterized by the relation

$$\langle Xv, w \rangle = \langle v, X^* w \rangle \quad \text{for all } v, w \in V.$$

With this notation

$$X \in \mathfrak{so}(V) \iff X = -X^*.$$

For $X \in \mathfrak{so}(V)$, then

$$\begin{aligned} \left(\mathrm{e}^X\right)^* &= \mathrm{Id} + X^* + \frac{1}{2}(X^*)^2 + \dots \\ &= \mathrm{Id} - X + \frac{1}{2}X^2 - \dots = \mathrm{e}^{-X} = (\mathrm{e}^X)^{-1}, \end{aligned}$$

hence $\mathrm{e}^X \in \mathrm{SO}(V)$.

In fact, the exponential map maps $\mathfrak{so}(V)$ onto $\mathrm{SO}(V)$. However, the exponential map from $\mathfrak{gl}(V)$ is not surjective; its image does not even contain all elements of $\mathrm{Gl}_+(V)$, the subgroup of automorphisms of V with positive determinant (w.r.t. some basis).

Typically, $(V, \langle \cdot, \cdot \rangle)$ will be the Euclidean space of dimension n, i.e. $\mathbb{R}^n$ with its standard Euclidean scalar product. For that purpose, we shall often use the notation $\mathrm{Gl}(n, \mathbb{R})$ in place of $\mathrm{Gl}(V)$, $\mathfrak{gl}(n)$, $\mathrm{O}(n)$, $\mathrm{SO}(n)$, $\mathfrak{o}(n)$, $\mathfrak{so}(n)$ in place of $\mathfrak{gl}(V)$, $\mathrm{O}(V)$, $\mathrm{SO}(V)$, $\mathfrak{o}(V)$, $\mathfrak{so}(V)$ etc.

Sometimes, we shall also need complex vector spaces. Let $V_{\mathbb{C}}$ be a vector space over $\mathbb{C}$ of complex dimension m. We put

$$\mathrm{Gl}(V_{\mathbb{C}}) := \{A : V_{\mathbb{C}} \to V_{\mathbb{C}} \text{ complex linear and bijective}\}.$$

If $V_{\mathbb{C}}$ is equipped with a Hermitian product $\langle \cdot, \cdot \rangle$, we put

$$\begin{aligned} \mathrm{U}(V_{\mathbb{C}})(:= \mathrm{U}(V_{\mathbb{C}}, \langle \cdot, \cdot \rangle)) &:= \{A \in \mathrm{Gl}(V_{\mathbb{C}}) : \langle Av, Aw \rangle = \langle v, w \rangle \text{ for all } v, w \in V_{\mathbb{C}}\} \\ \mathrm{SU}(V_{\mathbb{C}}) &:= \{A \in \mathrm{U}(V_{\mathbb{C}}) : \det A = 1\}. \end{aligned}$$

The associated Lie algebras are

$$\begin{aligned}\mathfrak{gl}(V_{\mathbb{C}}) &:= \{X : V_{\mathbb{C}} \to V_{\mathbb{C}} \text{ complex linear}\}\\ \mathfrak{u}(V_{\mathbb{C}}) &:= \{X \in \mathfrak{gl}(V_{\mathbb{C}}) : \langle Xv, w\rangle + \langle v, Xw\rangle = 0 \text{ for all } v, w \in V_{\mathbb{C}}\},\end{aligned}$$

(the skew Hermitian endomorphisms of $V_{\mathbb{C}}$), and

$$\mathfrak{su}(V_{\mathbb{C}}) := \{X \in \mathfrak{u}(V_{\mathbb{C}}) : \operatorname{tr} X = 0\}$$

(the skew Hermitian endomorphisms with vanishing trace), where the trace tr is defined using a unitary basis $e_1, \dots, e_m$ of $V_{\mathbb{C}}$, i.e. $\langle e_i, e_j\rangle = \delta_{ij}$.

If V is $\mathbb{C}^m$ with its standard Hermitian product, we write $\mathrm{Gl}(m, \mathbb{C})$, $\mathrm{U}(m)$, $\mathrm{SU}(m)$ etc. in place of $\mathrm{Gl}(V_{\mathbb{C}})$, $\mathrm{U}(V_{\mathbb{C}})$, $\mathrm{SU}(V_{\mathbb{C}})$ etc.

For $A, B \in \mathrm{Gl}(V)$, we have the conjugation by A.

$$\operatorname{Int}(A)B = ABA^{-1}. \tag{1.10.1}$$

For $X \in \mathfrak{gl}(V)$, then the induced action of A is given by

$$(\operatorname{Ad} A)X = AXA^{-1},$$

and for $Y \in \mathfrak{gl}(V)$, we obtain the infinitesimal version

$$(\operatorname{ad} Y)X = YX - XY = [Y, X]$$

as follows by writing $B = \mathrm{e}^{tX}$, $A = \mathrm{e}^{sY}$ and differentiating (1.10.1) w.r.t. t and s and $s = t = 0$.

Thus, Ad and ad associate to each element in $\mathrm{Gl}(V)$ resp. $\mathfrak{gl}(V)$ a linear endomorphism of the vector space $\mathfrak{gl}(V)$. Thus, Ad and ad yield representations of the Lie group $\mathrm{Gl}(V)$ and the Lie algebra $\mathfrak{gl}(V)$, resp., on the vector space $\mathfrak{gl}(V)$. These representations are called adjoint representations.

The *unit element* of a Lie group G will be denoted by e.

For $g \in G$, we have the *left translation*

$$\begin{aligned}L_g : G &\to G\\ h &\mapsto gh\end{aligned}$$

and the *right translation*

$$\begin{aligned}R_g : G &\to G\\ h &\mapsto hg.\end{aligned}$$

L_g and R_g are diffeomorphisms of $G, (L_g)^{-1} = L_{g^{-1}}$.

A vector field X on G is called *left invariant* if for all $g, h \in G$

$$L_{g*}X(h) = X(gh),$$

(see (1.9.8) for the definition of L_{g*}; note that we should write $(L_g)_*$ for L_{g*})

i.e.

$$L_{g*}X = X \circ L_g\,. \tag{1.10.2}$$

Theorem 1.10.1. *Let G be a Lie group. For every $V \in T_eG$,*

$$X(g) := L_{g*}V \tag{1.10.3}$$

defines a left invariant vector field on G, and we thus obtain an isomorphism between T_eG and the space of left invariant vector fields on G.

Proof.

$$X(gh) = L_{(gh)*}V = L_{g*}L_{h*}V = L_{g*}X(h)$$

which is left invariance.

Since a left invariant vector field is determined by its value at any point of G, for example at e, we obtain an isomorphism between T_eG and the space of left invariant vector fields. □

By Lemma 1.9.6, for $g \in G$ and vector fields X, Y

$$[L_{g*}X, L_{g*}Y] = L_{g*}[X, Y]. \tag{1.10.4}$$

Consequently, the Lie bracket of left invariant vector fields is left invariant itself, and the space of left invariant vector fields is closed under the Lie bracket and hence forms a Lie subalgebra of the Lie algebra of all vector fields on G (cf. Corollary 1.9.3). From Theorem 1.10.1, we obtain

Corollary 1.10.1. *T_eG carries the structure of a Lie algebra.* □

Definition 1.10.2. The *Lie algebra* $\mathfrak{g}$ of G is the vector space T_eG equipped with the Lie algebra structure of Corollary 1.10.1.

We may easily construct so-called *left invariant Riemannian metrics* on a Lie group G by the following procedure:

We select a scalar product $\langle \cdot, \cdot \rangle$ on the Lie algebra T_eG. For $h \in G, V \in T_hG$, there exists a unique $V_e \in T_eG$ with

$$V = L_{h*}V_e, \tag{1.10.5}$$

since L_h is a diffeomorphism. We then put for $V, W \in T_hG$

$$\langle V, W \rangle := \langle V_e, W_e \rangle. \tag{1.10.6}$$

This defines a Riemannian metric on G which is left invariant. In analogy to the definition of a vector bundle (Definition 1.8.1) where the fiber is a vector space we now define a principal bundle as one where the fiber is a Lie group.

Definition 1.10.3. Let G be a Lie group. A *principal G-bundle* consists of a base M, which is a differentiable manifold, and a differentiable manifold P, the total space of the bundle, and a differentiable projection $\pi : P \to M$, with an action of G on P satisfying:

(i) G acts freely on P from the right: $(q, g) \in P \times G$ is mapped to $qg \in P$, and $qg \neq q$ for $g \neq e$.

The G-action then defines an equivalence relation on $P : p \sim q :\Longleftrightarrow \exists g \in G : p = qg$.

(ii) M is the quotient of P by this equivalence relation, and $\pi : P \to M$ maps $q \in P$ to its equivalence class. By (i), each fiber $\pi^{-1}(x)$ can then be identified with G.

(iii) P is locally trivial in the following sense:

For each $x \in M$, there exist a neighborhood U of x and a diffeomorphism

$$\varphi : \pi^{-1}(U) \to U \times G$$

of the form $\varphi(p) = (\pi(p), \psi(p))$ which is G-equivariant, i.e. $\varphi(pg) = (\pi(p), \psi(p)g)$ for all $g \in G$.

As in Definition 1.8.2, a subgroup H of G is called the *structure group* of the bundle P if all transition maps take their values in H. Here, the structure group operates on G by left translations.

The notions of vector and principal bundle are closely associated with each other as we now want to explain briefly. Given a principal G-bundle $P \to M$ and a vector space V on which G acts from the left, we construct the associated vector bundle $E \to M$ with fiber V as follows:
We have a free action of G on $P \times V$ from the right:

$$\begin{aligned} P \times V \times G &\to P \times V \\ (p, v) \cdot g &= (p \cdot g, g^{-1}v)\,. \end{aligned}$$

If we divide out this G-action, i.e. identify (p, v) and $(p, v) \cdot g$, the fibers of $(P \times V)/G \to P/G$ become vector spaces isomorphic to V, and

$$E := P \times_G V := (P \times V)_{/G} \to M$$

is a vector bundle with fiber $G \times_G V := (G \times V)_{/G} = V$ and structure group G. The transition functions for P also give transition functions for E via the left action of G on V. Conversely, given a vector bundle E with structure group G, we construct a principal G-bundle as

$$\coprod_\alpha U_\alpha \times G / \sim$$

with

$$(x_\alpha, g_\alpha) \sim (x_\beta, g_\beta) :\Longleftrightarrow x_\alpha = x_\beta \in U_\alpha \cap U_\beta \quad \text{and} \quad g_\beta = \varphi_{\beta\alpha}(x) g_\alpha$$

where $\{U_\alpha\}$ is a local trivialization of E with transition functions $\varphi_{\beta\alpha}$, as in Theorem 1.8.1.

P can be considered as the *bundle of admissible bases* of E. In a local trivialization, each fiber of E is identified with $\mathbb{R}^n$, and each admissible basis is represented by a matrix contained in G. The transition functions describe a base change.

For example, if we have an $\mathrm{SO}(n)$ vector bundle E, i.e. a vector bundle with structure group $\mathrm{SO}(n)$, then the associated principal $\mathrm{SO}(n)$ bundle is the bundle of oriented orthonormal bases (frames) for the fibres of E.

Perspectives. Lie groups, while only treated relatively briefly in the present text book, form a central object of mathematical study. An introduction to their geometry and classification may be found in [123]. As symmetry groups of physical systems, they also play an important role in modern physics, in particular in quantum mechanics and quantum field theory.

We shall encounter Lie groups again in Chapter 5 as isometry groups of symmetric spaces. A theorem of Myers-Steenrod says that the isometry group of a Riemannian manifold is a Lie group. For a generic Riemannian manifold, the isometry group is discrete or even trivial. A homogeneous space is a Riemannian manifold with a transitive group G of isometries. It may thus be represented as G/H where $H := \{g \in G : gx_0 = x_0\}$ is the isotropy group of an arbitrarily selected $x_0 \in M$. Homogeneous spaces form important examples of Riemannian manifolds and include the symmetric spaces discussed in Chapter 5.

1.11 Spin Structures

For the definition of the Dirac operator in §3.4 and its applications in Chapter 9, we need a compact Lie group, $\mathrm{Spin}\,(n)$, which is not a subgroup of $\mathrm{Gl}(n, \mathbb{R})$, but rather a two-fold covering of $\mathrm{SO}(n)$ for $n \geq 3$. The case $n = 4$ will be particularly important for our applications. In order to define $\mathrm{Spin}\,(n)$, we start by introducing Clifford algebras.

We let V be a vector space of dimension n over $\mathbb{R}$, equipped with a positive definite inner product $\langle \cdot, \cdot \rangle$. We put $\|v\| := \langle v, v \rangle^{\frac{1}{2}}$, for every $v \in V$. For a substantial part of the algebraic constructions to follow in fact a not necessarily nondegenerate quadratic form on V would suffice, but here we have no need to investigate the most general possible construction. On the contrary, for our purposes it suffices to take $\mathbb{R}^n$ with its standard Euclidean scalar product. An orthonormal basis will be denoted by $e_1, \ldots, e_n$.

Definition 1.11.1. The *Clifford algebra* $\mathrm{Cl}(V)$, also denoted $\mathrm{Cl}(n)$, is the quotient of the tensor algebra $\bigoplus\limits_{k \geq 0} V \otimes \ldots \otimes V$ generated by V by the two sided ideal generated by all elements of the form $v \otimes v + \|v\|^2$ for $v \in V$.

Thus, the multiplication rule for the Clifford algebra $\mathrm{Cl}(V)$ is

$$vw + wv = -2\langle v, w \rangle \tag{1.11.1}$$

In particular, in terms of our orthonormal basis $e_1, \dots, e_n$, we have

$$e_i{}^2 = -1 \text{ and } e_i e_j = -e_i e_j \text{ for } i \neq j. \tag{1.11.2}$$

From this, one easily sees that a basis of $\mathrm{Cl}(V)$ as a real vector space is given by

$$e_0 := 1, \quad e_{\boldsymbol{\alpha}} := e_{\alpha_1} e_{\alpha_2} \dots e_{\alpha_k}$$

with $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_k\} \subset \{1, \dots, n\}$ and $\alpha_1 < \alpha_2 \dots < \alpha_k$. For such an $\boldsymbol{\alpha}$, we shall put $|\boldsymbol{\alpha}| := k$ in the sequel. Thus, as a vector space, $\mathrm{Cl}(V)$ is isomorphic to $\Lambda^*(V)$ (as algebras, these two spaces are of course different). In particular, the dimension of $\mathrm{Cl}(V)$ as a vector space is 2^n. Also, declaring this basis as being orthonormal, we obtain a scalar product on $\mathrm{Cl}(V)$ extending the one on V.

We define the degree of $e_{\boldsymbol{\alpha}}$ as being $|\boldsymbol{\alpha}|$. The $e_{\boldsymbol{\alpha}}$ of degree k generate the subset $\mathrm{Cl}^k(V)$ of elements of degree k. We have

$$\mathrm{Cl}^0 = \mathbb{R}$$
$$\mathrm{Cl}^1 = V.$$

Finally, we let $\mathrm{Cl}^{\mathrm{ev}}(V)$ and $\mathrm{Cl}^{\mathrm{odd}}(V)$ be the subspaces of elements of even, resp. odd degree. The former is a subalgebra of $\mathrm{Cl}(V)$, but not the latter.

Lemma 1.11.1. *The center of* $\mathrm{Cl}(V)$ *consists of those elements that commute with all* $v \in \mathrm{Cl}^1(V) = V$*. For n even, the center is* $\mathrm{Cl}^0(V)$*, while for n odd, it is* $\mathrm{Cl}^0(V) \oplus \mathrm{Cl}^n(V)$*.*

Proof. It suffices to consider basis vectors $e_{\boldsymbol{\alpha}} = e_{\alpha_1} \dots e_{\alpha_k}$ as above. For $j \notin \boldsymbol{\alpha}$, we have

$$e_{\boldsymbol{\alpha}} e_j = (-1)^{|\boldsymbol{\alpha}|} e_j e_{\boldsymbol{\alpha}},$$

and thus $|\boldsymbol{\alpha}|$ has to be even for $e_{\boldsymbol{\alpha}}$ to commute with e_j, while

$$e_{\boldsymbol{\alpha}} e_{\alpha_j} = (-1)^{|\boldsymbol{\alpha}|-1} e_{\alpha_j} e_{\boldsymbol{\alpha}},$$

so that $|\boldsymbol{\alpha}|$ needs to be odd for a commutation.

The conclusion follows easily for monomials and with a little algebra also in the general case. □

We next observe that

$$\mathrm{Cl}^2 =: \mathfrak{spin}(V) \quad \text{(or simply } \mathfrak{spin}(n))$$

is a Lie algebra with the bracket

$$[a, b] = ab - ba \tag{1.11.3}$$

For that, note that $[a, b] \in \mathrm{Cl}^2(V)$ if $a, b \in \mathrm{Cl}^2(V)$ as an easy consequence of (1.11.2).

To verify this, let us first consider the case

$$a = e_i e_j, \quad b = e_k e_l.$$

with the indices i, j, k, l all different. In this case

$$\begin{aligned} e_i e_j e_k e_l - e_k e_l e_i e_j &= e_i e_k e_l e_j - e_k e_l e_i e_j \\ &= e_k e_l e_i e_j - e_k e_l e_i e_j = 0 \quad \text{by (1.11.2)} \end{aligned}$$

Another case is

$$a = e_i e_j, \quad b = e_j e_k.$$

Then, using (1.11.2)

$$\begin{aligned} e_i e_j e_j e_k - e_j e_k e_i e_j &= -e_i e_k - e_j e_j e_k e_i \\ &= -e_i e_k + e_k e_i \\ &= -2 e_i e_k \in \mathrm{Cl}^2(V). \end{aligned}$$

From these two cases, the general pattern should be clear.

In a similar manner, the bracket defines an action τ of $\mathrm{Cl}^2(V)$ on $\mathrm{Cl}^1(V) = V$:

$$\tau(a)v := [a, v] := av - va \tag{1.11.4}$$

Again, by (1.11.2) $[a, v] \in \mathrm{Cl}^1(V)$ if $a \in \mathrm{Cl}^2(V)$, $v \in \mathrm{Cl}^1(V)$.

Let us consider the two typical cases as before, first

$$a = e_i e_j, \quad v = e_k,$$

with i, j, k all different. Then

$$e_i e_j e_k - e_k e_i e_j = e_i e_j e_k - e_i e_j e_k = 0.$$

The second case is

$$a = e_i e_j, \quad v = e_i,$$

Then

$$e_i e_j e_i - e_i e_i e_j = -e_i e_i e_j - e_i e_i e_j = 2 e_j \in \mathrm{Cl}^1(V).$$

Lemma 1.11.2. *τ defines a Lie algebra isomorphism between $\mathfrak{spin}(V)$ and $\mathfrak{so}(V)$.*

Proof. Since, as noted, $\tau(a)$ preserves V, and since one readily checks that $\tau[a, b] = [\tau(a), \tau(b)]$, τ defines a Lie algebra homomorphism from $\mathfrak{spin}(V) = \mathrm{Cl}^2(V)$ to $\mathfrak{gl}(V)$. For $a \in \mathrm{Cl}^2(V)$,

$$\begin{aligned} \langle \tau(a)v, w \rangle + \langle v, \tau(a)w \rangle &= -\frac{1}{2}[[a, v], w] - \frac{1}{2}[v, [a, w]] \quad \text{by (1.11.1)} \\ &= 0, \end{aligned} \tag{1.11.5}$$

as one easily checks by employing (1.11.2), after the same pattern as above.

Therefore, $\tau(a) \in \mathfrak{so}(V)$ for all $a \in \mathrm{Cl}^2(V)$. It follows from Lemma 1.11.1 that τ is injective on $\mathrm{Cl}^2(V)$. Since $\mathrm{Cl}^2(V)$ and $\mathfrak{so}$ both are vector spaces of dimension $\frac{n(n-1)}{2}$, and τ is an injective linear map between them, τ in fact has to be bijective. $\square$

In the Clifford algebra $\mathrm{Cl}(V)$, one can now define an exponential series as in $\mathfrak{gl}(V)$, and one may define the group $\mathrm{Spin}\,(V)$ as the exponential image of the Lie algebra $\mathfrak{spin}(V)$. $\mathrm{Spin}\,(V)$ then becomes a Lie group. This follows from general properties of the exponential map. Here, however, we rather wish to define $\mathrm{Spin}\,(v)$ directly, as this may be more instructive from a geometric point of view.

For that purpose, let us first introduce an anti-automorphism $a \mapsto a^t$ of $\mathrm{Cl}(V)$, defined on a basis vector $e_{\alpha_1} e_{\alpha_2} \dots e_{\alpha_k}$ as above by

$$(e_{\alpha_1} e_{\alpha_2} \dots e_{\alpha_k})^t = e_{\alpha_k} \dots e_{\alpha_2} e_{\alpha_1} \ (= (-1)^{\frac{k(k-1)}{2}} e_{\alpha_1} e_{\alpha_2} \dots e_{\alpha_k}). \tag{1.11.6}$$

In particular

$$e_{\alpha_1} e_{\alpha_2} \dots e_{\alpha_k} (e_{\alpha_1} \dots e_{\alpha_k})^t = \begin{cases} 1, & \text{if } k \text{ is even} \\ -1, & \text{if } k \text{ is odd} \end{cases} \tag{1.11.7}$$

Also, for all $a, b \in \mathrm{Cl}(V)$

$$(ab)^t = b^t a^t. \tag{1.11.8}$$

Definition 1.11.2. $\mathrm{Pin}(V)$ is the group of elements of $\mathrm{Cl}(V)$ of the form

$$a = a_1 \dots a_k \text{ with } a_i \in V, \|a_i\| = 1 \text{ for } i = 1, \dots, k$$

$\mathrm{Spin}\,(V)$ is the group $\mathrm{Pin}(V) \cap \mathrm{Cl}^{\mathrm{ev}}(V)$, i.e. the group of elements of $\mathrm{Cl}(v)$ of the form

$$a = a_1 \dots a_{2m} \text{ with } a_i \in V, \|a_i\| = 1 \text{ for } i = 1, \dots, 2m \ (m \in \mathbb{N}).$$

We shall often write $\mathrm{Pin}(n)$, $\mathrm{Spin}\,(n)$ in place of $\mathrm{Pin}(\mathbb{R}^n)$, $\mathrm{Spin}\,(\mathbb{R}^n)$, resp.

From (1.11.7), we see that $\mathrm{Spin}\,(V)$ is the group of all elements $a \in \mathrm{Pin}(V)$ with

$$aa^t = 1. \tag{1.11.9}$$

Theorem 1.11.1. *Putting*

$$\rho(a)v := ava^t$$

defines a surjective homomorphism $\rho : \mathrm{Pin}(V) \to \mathrm{O}(V)$ *with* $\rho(\mathrm{Spin}\,(V)) = \mathrm{SO}(V)$.

In particular, $\mathrm{Pin}(V) \subset \mathrm{Cl}(V)$ acts on V. This is the so-called vector representation, not to be confused with the spinor representation introduced below.

Proof. We start with $a \in V$, $\|a\| = 1$. In that case, every $v \in V$ decomposes as

$$v = \lambda a + a^{\perp}, \text{ with } \langle a, a^{\perp} \rangle = 0, \ \lambda \in \mathbb{R}.$$

Then, since $a = a^t$ for $a \in V$

$$\begin{aligned}
\rho(a)v &= a(\lambda a + a^{\perp})a \\
&= -\lambda a - aaa^{\perp}, \qquad && \text{since } aa = aa^t = -1 \qquad && \text{by (1.11.7)} \\
& && \text{and } a^{\perp}a + aa^{\perp} = 0 && \text{by (1.11.2)} \\
&= -\lambda a + a^{\perp}.
\end{aligned}$$

Consequently $\rho(a)$ is the reflection across the hyperplane orthogonal to a. This is an element of $\mathrm{O}(V)$. Then also for a general $a = a_1 \dots a_k \in \mathrm{Pin}(V)$, $\rho(a)$ is a product of reflections across hyperplanes, hence in $\mathrm{O}(V)$. The preceding construction also shows that all reflections across hyperplanes are contained in the image of $\rho(\mathrm{Pin}(V))$. Since every element in $\mathrm{O}(V)$ can be represented as a product of such reflections[3], it follows that $\rho(\mathrm{Pin}(V)) = \mathrm{O}(V)$. If now $a \in \mathrm{Spin}\,(V)$, then $\rho(a)$ is a product of an even number of reflections, hence in $\mathrm{SO}(V)$. Since every element $\mathrm{SO}(V)$ can conversely be represented as a product of an even number of reflections, it follows that $\rho(\mathrm{Spin}\,(V)) = \mathrm{SO}(V)$.

From (1.11.8), it is clear that $\rho(ab) = \rho(a)\rho(b)$, and so ρ defines a homomorphism. □

Let us now determine the kernel of

$$\rho : \mathrm{Spin}\,(V) \to \mathrm{SO}(V).$$

If $a \in \ker \rho$, then $\rho(a)v = v$ for all $v \in V$.

From the definition of ρ and $aa^t = 1$ for $a \in \mathrm{Spin}\,(V)$, we obtain that this is equivalent to

$$av = va \text{ for all } v \in V,$$

i.e. a commutes with all elements of V. Since all elements in $\mathrm{Spin}\,(V)$ are even, Lemma 1.11.1 implies $a \in \mathbb{R}$. Since $aa^t = 1$, we conclude that

$$a = \pm 1.$$

We next claim that $\mathrm{Spin}\,(V)$ is connected for $\dim_{\mathbb{R}} V \geq 2$. Let

$$a = a_1 \dots a_{2m} \in \mathrm{Spin}\,(V), \text{ with } a_i \text{ in the unit sphere of } V. \tag{1.11.10}$$

Since that sphere is connected, we may connect every a_i by a path $a_i(t)$ to e_1. Hence, a can be connected to $e_1 \dots e_1$ ($2m$ times), which is ± 1. Thus we need to connect 1 and -1. We use the path

$$\begin{aligned}
\gamma(t) &= \left(\cos\left(\frac{\pi}{2}t\right) e_1 + \sin\left(\frac{\pi}{2}t\right) e_2\right)\left(\cos\left(\frac{\pi}{2}t\right) e_1 - \sin\left(\frac{\pi}{2}t\right) e_2\right) \\
&= -\cos^2\left(\frac{\pi}{2}t\right) + \sin^2\left(\frac{\pi}{2}t\right) - 2\sin\left(\frac{\pi}{2}t\right)\cos\left(\frac{\pi}{2}t\right) e_1 e_2, \\
&\qquad \text{since } e_1 e_1 = e_2 e_2 = -1.
\end{aligned}$$

[3]Every rotation of a plane is a product of two reflections, and the normal form of an orthogonal matrix shows that it can be represented as a product of rotations and reflections in mutually orthogonal planes.

This path is contained in $\mathrm{Spin}\,(V)$ and satisfy $\gamma(0) = -1$, $\gamma(1) = 1$, and we have shown connectedness of $\mathrm{Spin}\,(V)$ for $\dim_{\mathbb{R}} V \geq 2$.

(1.11.10) also easily implies that $\mathrm{Spin}\,(V)$ is compact. If we finally use the information that $\pi_1(\mathrm{SO}(V)) = \mathbb{Z}_2$ for $n = \dim_{\mathbb{R}} V \geq 3$, we obtain altogether

Theorem 1.11.2. $\rho : \mathrm{Spin}\,(V) \to \mathrm{SO}(V)$ *is a nontrivial double covering.* $\mathrm{Spin}\,(V)$ *is compact and connected, and for* $\dim_{\mathbb{R}} V \geq 3$*, it is also simply connected. Thus, for* $\dim_{\mathbb{R}} V \geq 3$*,* $\mathrm{Spin}\,(V)$ *is the universal cover of* $\mathrm{SO}(V)$*.* □

Let us briefly return to the relation between $\mathfrak{spin}(V)$ and $\mathrm{Spin}\,(V)$. If we differentiate the relation characterizing $\mathrm{Spin}\,(V)$, i.e.

$$aa^t = 1 \quad \text{and} \quad ava^t \in V \text{ for all } v \in V,$$

(differentiating means that we consider $a = 1 + \epsilon b + \mathrm{O}(\epsilon^2)$ and take the derivative w.r.t. ϵ at $\epsilon = 0$), we obtain the infinitesimal relations

$$b + b^t = 0 \quad \text{and} \quad bv + vb^t = bv - vb \text{ for all } v \in V,$$

which were the relations satisfied by elements of $\mathfrak{spin}(V) = \mathrm{Cl}^2(V)$. Since the preceding implies that $\mathrm{Spin}\,(V)$ and $\mathfrak{spin}(V)$ have the same dimension, namely the one of $\mathrm{SO}(V)$ and $\mathfrak{so}(V)$, i.e. $\frac{n(n-1)}{2}$, $\mathfrak{spin}(V)$ indeed turns out to be the Lie algebra of the Lie group $\mathrm{Spin}\,(V)$.

Let us also discuss the induced homomorphism

$$d\rho : \mathfrak{spin}(V) \to \mathfrak{so}(V),$$

the infinitesimal version of ρ. The preceding discussion implies that $d\rho$ coincides with the Lie algebra isomorphism τ of Lemma 1.11.2. In order to obtain a more explicit relation, we observe that a basis for $\mathfrak{so}(n)$, the Lie algebra of skew symmetric $n \times n$-matrices is given by the matrices $e_i \wedge e_j$, $1 \leq i < j \leq n$, (denoting the skew symmetric matrix that has -1 at the intersection of the i^{th} row and the j^{th} column, $+1$ at the intersection of the j^{th} row and the i^{th} column, and 0 entries elsewhere)[4]. $e_i \wedge e_j$ is the tangent vector at the identity of $\mathrm{SO}(n)$ for the one parameter subgroup of rotations through an angle ϑ in the $e_i e_j$ plane from e_i towards e_j. In $\mathrm{Spin}\,(n)$, we may consider the one parameter subgroup

$$\vartheta \mapsto e_i(-\cos(\vartheta)e_i + \sin(\vartheta)e_j) = \cos(\vartheta) + \sin(\vartheta)e_i e_j.$$

Its tangent vector at 1, i.e. at $\vartheta = 0$, is $e_i e_j$.

Lemma 1.11.3.

$$d\rho(e_i e_j) = 2(e_i \wedge e_j)$$

Proof. We have seen in the proof of Theorem 1.11.1, that $\rho(a)$ is the reflection across the hyperplane perpendicular to a, for a unit vector $a \in \mathbb{R}^n$. Thus, $\rho(\cos(\vartheta) +$

[4] For the sake of the present discussion, we identify V with $\mathbb{R}^n$ ($n = \dim_{\mathbb{R}} V$).

$\sin(\vartheta)e_ie_j)$ is the reflection across the hyperplane orthogonal to $-\cos(\vartheta)e_i+\sin(\vartheta)e_j$ followed by the one across the hyperplane orthogonal to e_i. This, however, is the rotation in the e_i, e_j plane through an angle of 2ϑ from e_i towards e_j. $\square$

Examples.

1. From its definition, the Clifford algebra $\mathrm{Cl}(\mathbb{R})$ is $\mathbb{R}[x]/(x^2+1)$, the algebra generated by x with the relation $x^2=-1$. In order to make contact with our previous notation, we should write e_1 in place of x. Of course this algebra can be identified with $\mathbb{C}$, and we identify the basis vector e_1 with i $\mathrm{Cl}^{\mathrm{ev}}(\mathbb{R}) = \mathrm{Cl}^0(\mathbb{R})$ then are the reals, while $\mathrm{Cl}^{\mathrm{odd}}(\mathbb{R}) = \mathrm{Cl}^1(\mathbb{R})$ is identified with the purely imaginary complex numbers. $\mathrm{Pin}(\mathbb{R})$ then is the subgroup of $\mathbb{C}$ generated by $\pm i$, and $\mathrm{Spin}\,(\mathbb{R})$ is the group with elements ± 1.

2. $\mathrm{Cl}(\mathbb{R}^2)$ is the algebra generated by x and y with the relations

$$x^2=-1, \quad y^2=-1, \quad xy=-yx.$$

Again, we write e_1, e_2 in place of x, y. This algebra can be identified with the quaternion algebra $\mathbb{H}$, by putting

$$i=e_1, \quad j=e_2, \quad k=e_1e_2.$$

Since $i^2=j^2=k^2=-1$, $ij+ji=ik+ki=jk+kj=0$ the relations (1.11.2) are indeed satisfied.
In fact, we have a natural linear embedding

$$\gamma:\mathbb{H}\to\mathbb{C}^{2\times 2} \quad \text{(2 by 2 matrices with complex coefficients)} \tag{1.11.11}$$

by writing $w\in\mathbb{H}$ as

$$w=(w_0+kw_1)-i(w_2+kw_3)=\omega-i\psi$$

with $w_0,w_1,w_2,w_3\in\mathbb{R}$ while we consider ω and ψ as elements of $\mathbb{C}$, and putting

$$w \longmapsto \begin{pmatrix}\omega & -\overline{\psi}\\ \psi & \overline{\omega}\end{pmatrix}.$$

Then

$$\gamma(i)=\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}, \quad \gamma(j)=\begin{pmatrix}0 & i\\ i & 0\end{pmatrix}, \quad \gamma(k)=\begin{pmatrix}i & 0\\ 0 & -i\end{pmatrix}.$$

These matrices satisfy the same commutation relations as i,j,k, and

$$\gamma(ww')=\gamma(w)\gamma(w'), \quad \gamma(\overline{w})=\gamma(w)^*$$

for all $w,w'\in\mathbb{H}$. Thus, γ is an algebra homomorphism.
The subalgebra $\mathrm{Cl}^{\mathrm{ev}}(\mathbb{R}^2)$ is generated by k, and thus it is isomorphic to $\mathbb{C}\subset\mathbb{H}$,

where the purely imaginary complex numbers correspond to multiples of k. Under the embedding γ, it corresponds to the diagonal elements in $\mathbb{C}^{2\times 2}$, that is, the ones of the form $w \longmapsto \begin{pmatrix} \omega & 0 \\ 0 & \overline{\omega} \end{pmatrix}$, i.e. those with $\psi = 0$.

$\mathrm{Pin}(\mathbb{R}^2)$ is generated by the circle $\cos(\vartheta)i + \sin(\vartheta)j$ through i and j ($\vartheta \in S^1$). $\mathrm{Spin}\,(\mathbb{R}^2)$ then is the group consisting of products $(\cos(\vartheta_1)i+\sin(\vartheta_1)j)(\cos(\vartheta_2)i+\sin(\vartheta_2)j)$ $(\vartheta_1, \vartheta_2 \in S^1)$ $= -\cos\vartheta_1\cos\vartheta_2 - \sin\vartheta_1\sin\vartheta_2 + (\cos\vartheta_1\sin\vartheta_2 - \cos\vartheta_2 \sin\vartheta_1)k$, i.e. the unit circle in the above subspace $\mathbb{C} \subset \mathbb{H}$. (So, while $\mathrm{Pin}(V)$ is generated by $1, i, j, k$, $\mathrm{Spin}\,(V)$ is generated by $1, k$. i and j act on $\mathbb{R}^2$ by reflection while k acts as a rotation.) Thus, $\mathrm{Spin}\,(\mathbb{R}^2)$ is isomorphic to $\mathrm{U}(1) \cong S^1$. We should note, however, that it is a double cover of $\mathrm{SO}(2)$ as ± 1 both are mapped to the trivial element of $\mathrm{SO}(2)$.

3. Similarly, we identify $\mathrm{Cl}(\mathbb{R}^3)$ with $\mathbb{H} \oplus \mathbb{H}$ by putting

$$e_0 = (1,1), \quad e_1 = (i,-i), \quad e_2 = (j,-j) \quad e_3 = (k,-k).$$

Then

$$e_1e_2 = (k,k), \quad e_2e_3 = (i,i), \quad e_3e_1 = (j,j),$$

and $\mathrm{Cl}^{\mathrm{ev}}(\mathbb{R}^3)$ is identified with the diagonal embedding of $\mathbb{H}$ into $\mathbb{H} \oplus \mathbb{H}$. Since $\mathrm{Cl}^1(\mathbb{R}^3) = \mathbb{R}^3$ is identified with the pairs $(\alpha, -\alpha)$ of purely imaginary quaternions α, $\mathrm{Pin}(\mathbb{R}^3)$ is generated by such elements of length 1. $\mathrm{Spin}\,(\mathbb{R}^3)$ then is the group of pairs (β, β) of unit quaternions β, as every such pair can be obtained as a product $(\alpha_1, -\alpha_1)(\alpha_2, -\alpha_2)$ where α_1, α_2 are purely imaginary unit quaternions themselves. Thus, $\mathrm{Spin}\,(\mathbb{R}^3)$ is isomorphic to the group $\mathrm{Sp}(1)$ of unit quaternions in $\mathbb{H}$. One also knows that this group is isomorphic to $\mathrm{SU}(2)$. The above embedding $\gamma : \mathbb{H} \to \mathbb{C}^{2\times 2}$ (1.11.11) induces an isomorphism between $\mathrm{Sp}(1)$ and $\mathrm{SU}(2)$.

4. $\mathrm{Cl}(\mathbb{R}^4)$ is identified with $\mathbb{H}^{2\times 2}$, the space of two by two matrices with quaternionic coefficients, by putting

$$e_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad e_1 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad e_2 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix},$$
$$e_3 = \begin{pmatrix} 0 & j \\ j & 0 \end{pmatrix}, \qquad e_4 = \begin{pmatrix} 0 & k \\ k & 0 \end{pmatrix}.$$

$\mathrm{Pin}(\mathbb{R}^4)$ is generated by the unit sphere in $\mathrm{Cl}^1(\mathbb{R}^4) = \mathbb{R}^4$, i.e. in our identification by all linear combinations of e_1 e_2, e_3, e_4 of unit length. $\mathrm{Spin}\,(\mathbb{R}^4)$ then is the group of products of two such elements, i.e. the group of all elements of the form $\begin{smallmatrix} \alpha & 0 \\ 0 & \beta \end{smallmatrix}$ where α and β are unit quaternions. Thus, $\mathrm{Spin}\,(\mathbb{R}^4)$ is homeomorphic to $S^3 \times S^3 \cong \mathrm{Sp}(1) \times \mathrm{Sp}(1) \cong \mathrm{SU}(2) \times \mathrm{SU}(2)$. From Theorem 1.11.2, we then infer that

$$\mathrm{SO}(4) \cong \mathrm{Spin}\,(4)/\mathbb{Z}_2 \cong (\mathrm{SU}(2) \times \mathrm{SU}(2))/\mathbb{Z}_2.$$

In the sequel, we shall also need the complex Clifford algebra and the corresponding spin group. For V as before, we denote the complexified Clifford algebra by

$$\mathrm{Cl}^{\mathbb{C}}(V) = \mathrm{Cl}(V) \otimes_{\mathbb{R}} \mathbb{C}.$$

Thus, the $e_{\boldsymbol{\alpha}}$ again form a basis, and the only difference is that we now admit complex coefficients.

For the sequel, we need to choose an orientation of V, i.e. select an (orthonormal) basis $e_1, \dots, e_n$ of V being positive. (Any other basis of V obtained from this particular one by an element of $\mathrm{SO}(V)$ then is also called positive.)

Definition 1.11.3. Let $e_1, \dots, e_n$ be a positive orthonormal basis of V. The *chirality operator* is

$$\Gamma = i^m e_1, \dots, e_n \in \mathrm{Cl}^{\mathbb{C}}(V)$$

with $m = \frac{n}{2}$ for even n, $m = \frac{n+1}{2}$ for odd n.

It is easy to check that Γ is independent of the chosen positive orthonormal basis. To see the mechanism, let us just consider the case $n = 2$, and the new basis $f_1 = \cos\vartheta e_1 + \sin\vartheta e_2$, $f_2 = -\sin\vartheta e_1 + \cos\vartheta e_2$. Then

$$\begin{aligned} f_1 f_2 &= -\sin\vartheta\cos\vartheta e_1 e_1 + \sin\vartheta\cos\vartheta e_2 e_2 + \cos^2\vartheta e_1 e_2 - \sin^2\vartheta e_2 e_1 \\ &= e_1 e_2 \quad \text{by (1.11.2)} \end{aligned}$$

Lemma 1.11.4.

$$\Gamma^2 = 1.$$

For odd n,	$\Gamma v = v\Gamma,$	*for all $v \in V$.*
For even n,	$\Gamma v = -v\Gamma,$	*for all $v \in V$.*

Proof. A simple computation based on (1.11.2). □

Thus, we may use Γ to obtain a decomposition

$$\mathrm{Cl}^{\mathbb{C}}(V)^{\pm}$$

of $\mathrm{Cl}^{\mathbb{C}}(V)$ into the eigenspaces with eigenvalue ± 1 under multiplication by Γ. This is particularly interesting for even n, because we have

$$v\mathrm{Cl}^{\mathbb{C}}(V)^{\pm} = \mathrm{Cl}^{\mathbb{C}}(V)^{\mp} \quad \text{for every } v \in V \setminus \{0\}, \tag{1.11.12}$$

i.e. Clifford multiplication by v interchanges these eigenspaces. This is a simple consequence of Lemma 1.11.4, namely if e.g.

$$\Gamma a = a$$

then

$$\Gamma v a = -v\Gamma a = -va.$$

Definition 1.11.4. $\mathrm{Spin}^c(V)$ is the subgroup of the multiplicative group of units of $\mathrm{Cl}^{\mathbb{C}}(V) = \mathrm{Cl}(V) \otimes \mathbb{C}$ generated by $\mathrm{Spin}(V)$ and the unit circle in $\mathbb{C}$.

Lemma 1.11.5. $\mathrm{Spin}^c(V)$ *is isomorphic to* $\mathrm{Spin}\, V \times_{\mathbb{Z}_2} S^1$*, where the* $\mathbb{Z}_2$ *action identifies* (a, z) *with* $(-a, -z)$.

Proof. By Lemma 1.11.1, the unit complex scalars are in the center of $\mathrm{Cl}^{\mathbb{C}}(V)$, and hence commute with $\mathrm{Spin}(V)$. Therefore, we obtain a map

$$\mathrm{Spin}(V) \times S^1 \to \mathrm{Spin}^c(V), \tag{1.11.13}$$

which is surjective. The kernel of this mapping are the elements (a, z) with $az = 1$, which means $a = z^{-1} \in \mathrm{Spin}(V) \cap S^1$. We have already seen in the preparations for Theorem 1.11.2 that this latter set consists precisely of ± 1. □

By Lemma 1.11.5, changing (a, z) to $(-a, z)$ amounts to the same as changing (a, z) to $(a, -z)$, and thus we obtain an action of $\mathbb{Z}_2$ on $\mathrm{Spin}^c(V)$. The quotient of $\mathrm{Spin}^c(V)$ by this action yields a double covering

$$\mathrm{Spin}^c(V) \to \mathrm{SO}(V) \times S^1 \tag{1.11.14}$$

that is nontrivial on both factors.

The maps given in (1.11.13), (1.11.14) allow to determine the fundamental group $\pi_1(\mathrm{Spin}^c(V))$. Namely, a homotopically nontrivial loop γ in S^1 induces a loop in $\mathrm{Spin}^c(V)$ that is mapped to the loop 2γ in S^1 by (1.11.14) (2γ means the loop γ traversed twice) which again is nontrivial. Thus, $\pi_1(\mathrm{Spin}^c(V))$ contains $\pi_1(S^1) = \mathbb{Z}$ as a subgroup. On the other hand, if we have a loop in $\mathrm{Spin}^c(V)$ that is mapped to a homotopically trivial one in S^1 when we compose (1.11.14) with the projection on the second factor, it is homotopic to a loop in the kernel of that composition. That kernel can be identified with $\mathrm{Spin}(V)$ by (1.11.13), and since $\mathrm{Spin}(V)$ is simply connected by Theorem 1.11.2 for $\dim V \geq 3$, such a loop is homotopically trivial for $\dim V \geq 3$. Thus

Theorem 1.11.3. *For* $\dim V \geq 3$

$$\pi_1(\mathrm{Spin}^c(V)) = \mathbb{Z}$$

□

Examples. The treatment here will be based on the above discussion of examples in the real case.

1. $\mathrm{Cl}^{\mathbb{C}}(\mathbb{R}) = \mathrm{Cl}(\mathbb{R}) \otimes_{\mathbb{R}} \mathbb{C} = \mathbb{C} \oplus \mathbb{C}$, and $\mathrm{Spin}^c(\mathbb{R}) \cong S^1$ sits diagonally in this space.
2. $\mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^2) = \mathrm{Cl}(\mathbb{R}^2) \otimes_{\mathbb{R}} \mathbb{C} = \mathbb{H} \otimes_{\mathbb{R}} \mathbb{C}$. We want to identify $\mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^2)$ with $\mathbb{C}^{2\times 2}$, the space of two by two matrices with complex coefficients. We consider the

above homomorphism of algebras $\mathbb{H} \to \mathbb{C}^{2\times 2}$, and extending scalars, we obtain an isomorphism of $\mathbb{C}$-algebras

$$\mathbb{H} \otimes \mathbb{C} \to \mathbb{C}^{2\times 2}.$$

Thus, we identify $\mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^2)$ with $\mathbb{C}^{2\times 2}$. Under this identification, $\mathrm{Spin}\,(\mathbb{R}^2)$ corresponds to the elements

$$\begin{pmatrix} \alpha & 0 \\ 0 & \overline{\alpha} \end{pmatrix} \quad \text{with } \alpha \in S^1 = \mathrm{U}(1) \subset \mathbb{C}.$$

$\mathrm{Spin}^c\,(\mathbb{R}^2)$ then consists of the unitary diagonal matrices, i.e. $\mathrm{Spin}^c\,(\mathbb{R}^2) = \mathrm{U}(1) \times \mathrm{U}(1) = S^1 \times S^1$.

3. $\mathrm{Cl}^c(\mathbb{R}^3) = \mathrm{Cl}(\mathbb{R}^3) \otimes \mathbb{C} = (\mathbb{H} \oplus \mathbb{H}) \otimes \mathbb{C} = \mathbb{C}^{2\times 2} \oplus \mathbb{C}^{2\times 2}$ from the preceding example. We have identified $\mathrm{Spin}\,(\mathbb{R}^3)$ with $\mathrm{SU}(2)$, and so

$$\mathrm{Spin}^c\,(\mathbb{R}^3) \cong \left\{ \mathrm{e}^{\mathrm{i}\vartheta} U : \vartheta \in \mathbb{R}, U \in \mathrm{SU}(2) \right\} = \mathrm{U}(2)$$

4. Similarly, $\mathrm{Cl}^c(\mathbb{R}^4) = \mathrm{Cl}(\mathbb{R}^3) \otimes_{\mathbb{R}} \mathbb{C} = \mathbb{H}^{2\times 2} \otimes \mathbb{C} = \mathbb{C}^{4\times 4}$. We have identified $\mathrm{Spin}\,(\mathbb{R}^4)$ with $\mathrm{SU}(2) \times \mathrm{SU}(2)$, and so

$$\begin{aligned} \mathrm{Spin}^c\,(\mathbb{R}^4) &= \mathrm{Spin}\,(\mathbb{R}^4) \times_{\mathbb{Z}_2} S^1 \\ &\cong \{(U, V) \in \mathrm{U}(2) \times \mathrm{U}(2) : \det U = \det V\} \end{aligned}$$

In order to describe the isomorphism $\mathrm{Cl}^c(\mathbb{R}^4) \cong \mathbb{C}^{4\times 4}$ more explicitly, we recall the homomorphism $\gamma : \mathbb{H} \to \mathbb{C}^{2\times 2}$ from the description of $\mathrm{Cl}(\mathbb{R}^3)$. We define

$$\Gamma : \mathbb{H} \to \mathbb{C}^{4\times 4}$$

via

$$\Gamma(w) = \begin{pmatrix} 0 & \gamma(w) \\ -\gamma(w)^* & 0 \end{pmatrix}.$$

We recall

$$\gamma(1) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \gamma(i) = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix},$$
$$\gamma(j) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \qquad \gamma(k) = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}.$$

We identify $\mathbb{R}^4$ with $\mathbb{H}$, putting $e_1 = 1$, $e_2 = i$, $e_3 = j$, $e_4 = k$. Then

$$\Gamma(e_1)\Gamma(e_2) = \begin{pmatrix} 0 & i & & \\ i & 0 & & \\ & & 0 & -i \\ & & -i & 0 \end{pmatrix} = -\,\Gamma(e_2)\Gamma(e_1),$$

$$\Gamma(e_1)\Gamma(e_3) = \begin{pmatrix} 0 & -1 & & \\ 1 & 0 & & \\ & & 0 & -1 \\ & & 1 & 0 \end{pmatrix} = -\,\Gamma(e_3)\Gamma(e_1),$$

$$\Gamma(e_1)\Gamma(e_4) = \begin{pmatrix} i & 0 & & \\ 0 & -i & & \\ & & -i & 0 \\ & & 0 & i \end{pmatrix} = -\,\Gamma(e_4)\Gamma(e_1),$$

$$\Gamma(e_2)\Gamma(e_3) = \begin{pmatrix} i & 0 & & \\ 0 & -i & & \\ & & i & 0 \\ & & 0 & -i \end{pmatrix} = -\,\Gamma(e_3)\Gamma(e_2),$$

$$\Gamma(e_2)\Gamma(e_4) = \begin{pmatrix} 0 & 1 & & \\ -1 & 0 & & \\ & & 0 & 1 \\ & & -1 & 0 \end{pmatrix} = -\,\Gamma(e_4)\Gamma(e_2),$$

$$\Gamma(e_3)\Gamma(e_4) = \begin{pmatrix} 0 & i & & \\ i & 0 & & \\ & & 0 & -i \\ & & -i & 0 \end{pmatrix} = -\,\Gamma(e_4)\Gamma(e_3),$$

(always with 0's in the off diagonal blocks). One also easily checks that

$$\Gamma(e_\alpha)\Gamma(e_\alpha) = -\mathrm{Id}, \quad \text{for } \alpha = 1, 2, 3, 4.$$

Thus, Γ preserves the relations in the Clifford algebra, and it is not hard to verify that Γ in fact extends to the desired isomorphism between $\mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^4)$ and $\mathbb{C}^{4\times 4}$.

The preceding examples seem to indicate a general pattern that we now wish to demonstrate by induction on the basis of

Lemma 1.11.6. *For any vector space V as above*

$$\mathrm{Cl}^{\mathbb{C}}(V \oplus \mathbb{R}^2) \cong \mathrm{Cl}^{\mathbb{C}}(V) \otimes_{\mathbb{C}} \mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^2).$$

Proof. We choose orthonormal bases $v_1, \ldots, v_n$ of V and e_1, e_2 of $\mathbb{R}^2$. In order to define a map that is linear over $\mathbb{R}$,

$$l : V \oplus \mathbb{R}^2 \to \mathrm{Cl}^{\mathbb{C}}(V) \otimes_{\mathbb{C}} \mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^2)$$

we put

$$
\begin{aligned}
l(v_j) &:= iv_j \otimes e_1 e_2, && \text{for } j = 1, \dots, n, \\
l(e_\alpha) &:= 1 \otimes e_\alpha, && \text{for } \alpha = 1, 2.
\end{aligned}
$$

Since for example

$$
\begin{aligned}
l(v_j v_k + v_k v_j) &= (-v_j v_k - v_k v_j) \otimes e_1 e_2 e_1 e_2 = v_j v_k + v_k v_j \otimes 1 \\
l(v_j e_\alpha + e_\alpha v_j) &= iv_j \otimes (e_1 e_2 e_\alpha + e_\alpha e_1 e_2) = 0 \quad \text{for } \alpha = 1, 2
\end{aligned}
$$

we have an extension of l as an algebra homomorphism

$$l : \mathrm{Cl}(V \oplus \mathbb{R}^2) \to \mathrm{Cl}^{\mathbb{C}}(V) \otimes_{\mathbb{C}} \mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^2).$$

Extending scalars from $\mathbb{R}$ to $\mathbb{C}$, we obtain an algebra homomorphism

$$l : \mathrm{Cl}^{\mathbb{C}}(V \oplus \mathbb{R}^2) \to \mathrm{Cl}^{\mathbb{C}}(V) \otimes_{\mathbb{C}} \mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^2).$$

Now l has become a homomorphism between two algebras of the same dimension, and it is injective (and surjective) on the generators, hence an isomorphism. □

Corollary 1.11.1.

(i) If $\dim_{\mathbb{R}} V = 2n, \qquad \mathrm{Cl}^{\mathbb{C}}(V) \cong \mathbb{C}^{2^n \times 2^n},$

(ii) If $\dim_{\mathbb{R}} V = 2n + 1, \qquad \mathrm{Cl}^{\mathbb{C}}(V) \cong \mathbb{C}^{2^n \times 2^n} \oplus \mathbb{C}^{2^n \times 2^n}.$

Proof. By Example 2, $\mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^2) \cong \mathbb{C}^{2\times 2}$, and the proof follows from Lemma 1.11.6 by induction, starting with Example 2 in the even and Example 1 in the odd dimensional case, and using

$$\mathbb{C}^{m \times m} \otimes_{\mathbb{C}} \mathbb{C}^{2\times 2} \cong \mathbb{C}^{2m \times 2m}.$$

□

We now wish to identify $\mathrm{Cl}^{\mathbb{C}}(V)$ for even dimensional V as the algebra of endomorphisms of some other vector space in a more explicit manner than in Corollary 1.11.1. We thus assume that $n = \dim_{\mathbb{R}} V$ is even, $n = 2m$. We also choose an orientation of V, i.e. select a positive orthonormal basis $e_1, \dots, e_n$.

In $V \otimes \mathbb{C}$, we consider the subspace W spanned by the basis vectors

$$\eta_j := \frac{1}{\sqrt{2}}(e_{2j-1} - ie_{2j}), \quad j = 1, \dots, m. \tag{1.11.15}$$

If we extend the scalar product $\langle \cdot, \cdot \rangle$ to $V \otimes \mathbb{C}$ by complex linearity, we have

$$\langle \eta_i, \eta_j \rangle_{\mathbb{C}} = 0 \quad \text{for all } j, \tag{1.11.16}$$

hence

$$\langle w, w\rangle_{\mathbb{C}} = 0 \quad \text{for all } w \in W \tag{1.11.17}$$

(One expresses this by saying that W is isotropic w.r.t. $\langle\cdot,\cdot\rangle_{\mathbb{C}}$)

We have

$$V \otimes \mathbb{C} = W \oplus \overline{W},$$

with $\overline{W}$ spanned by the vectors $\overline{\eta_j} = \frac{1}{\sqrt{2}}(e_{2j-1} + ie_{2j})$, $j = 1, \dots, m$. Because of (1.11.17), $\overline{W}$ is the dual space W^* of W w.r.t. $\langle\cdot,\cdot\rangle_{\mathbb{C}}$, i.e. for every $w \in W \setminus \{0\}$, there exists a unique $w' \in \overline{W}$ with $\|w'\| = 1$ and

$$\langle w, w'\rangle_{\mathbb{C}} = \|w\|.$$

Definition 1.11.5. The *spinor space* S is defined as the exterior algebra ΛW of W. If we want to emphasize the dimension n of V, we write S_n in place of S.

We may then identify $\mathrm{Cl}^{\mathbb{C}}(V)$ as $\mathrm{End}_{\mathbb{C}}(S)$ as follows: We write $v \in V \otimes \mathbb{C}$ as

$$v = w + w' \quad \text{with } w \in W,\ w' \in \overline{W},$$

and for $s \in S = \Lambda W$, we put

$\rho(w)s := \sqrt{2}\epsilon(w)s$ (= $\sqrt{2}w \wedge s$, as ϵ denotes the exterior product)
$\rho(w')s := -\sqrt{2}\iota(w')s$ (where $\iota(w')$ denotes the interior product; note that we identify $\overline{W}$ with the dual space W^* of W, c.f. §1.8)

ρ obviously extends to all of $\mathrm{Cl}^{\mathbb{C}}(V)$ by the rule $\rho(vw) = \rho(v)\rho(w)$.

We have the following explicit rules for $\epsilon(w)$ and $\iota(w')$: If $s = \eta_{j_1} \wedge \dots \eta_{j_k}$, with $1 \le j_1 < \dots < j_k \le m$, then

$$\epsilon(\eta_j)s = \eta_j \wedge \eta_{j_1} \wedge \dots \wedge \eta_{j_k} \quad (= 0 \text{ if } j \in \{j_1, \dots, j_k\}), \tag{1.11.18}$$

and

$$\iota(\overline{\eta}_j)s = \begin{cases} 0 & \text{if } j \notin \{j_1, \dots, j_k\}, \\ (-1)^{\mu-1}\eta_{j_1} \wedge \dots \wedge \widehat{\eta_{j_\mu}} \wedge \dots \wedge \eta_{j_k} & \text{if } j = j_\mu. \end{cases} \tag{1.11.19}$$

In particular

$$\epsilon(\eta_j)i\iota(\overline{\eta}_j)s = \begin{cases} 0 & \text{if } j \notin \{j_1, \dots, j_k\}, \\ s & \text{if } j \in \{j_1, \dots, j_k\}. \end{cases} \tag{1.11.20}$$

$$\iota(\overline{\eta}_j)\epsilon(\eta_j)s = \begin{cases} s & \text{if } j \notin \{j_1, \dots, j_k\}, \\ 0 & \text{if } j \in \{j_1, \dots, j_k\}. \end{cases} \tag{1.11.21}$$

Thus, we have for all s and all j

$$(\epsilon(\eta_j)\iota(\overline{\eta}_j) + \iota(\overline{\eta}_j)\epsilon(\eta_j))s = s. \tag{1.11.22}$$

For subsequent use in §5.2, we also record that in the same manner, one sees that

$$(\epsilon(\eta_j)\iota(\overline{\eta}_\ell) + \iota(\overline{\eta}_\ell)\epsilon(\eta_j))s = 0 \text{ for } j \neq \ell. \tag{1.11.23}$$

In order to verify that the claimed identification is possible, we need to check first that ρ preserves the relations in the Clifford algebra. The following examples will bring out the general pattern:

$$\begin{aligned}
\rho\left({e_1}^2\right) &= 2\left(\frac{1}{\sqrt{2}}\epsilon(\eta_1) - \frac{1}{\sqrt{2}}\iota(\overline{\eta}_1)\right)\left(\frac{1}{\sqrt{2}}\epsilon(\eta_1) - \frac{1}{\sqrt{2}}\iota(\overline{\eta}_1)\right) \\
&= -(\epsilon(\eta_1)\iota(\overline{\eta}_1) + \iota(\overline{\eta}_1)\epsilon(\eta_1)) \quad \text{since } \epsilon(\eta_1)^2 = 0 = \iota(\overline{\eta}_1)^2 \\
&= -1 \quad \text{by (1.11.22)},
\end{aligned}$$

and

$$\begin{aligned}
\rho(e_1e_2) + \rho(e_2e_1) &= (\epsilon(\eta_1) - \iota(\overline{\eta}_1))i(\epsilon(\eta_1) + \iota(\overline{\eta}_1)) \\
&\quad + i(\epsilon(\eta_1) + \iota(\overline{\eta}_1))(\epsilon(\eta_1) - \iota(\overline{\eta}_1)) \\
&= 0, \\
\rho(e_1e_3) + \rho(e_3e_1) &= (\epsilon(\eta_1) - \iota(\overline{\eta}_1))(\epsilon(\eta_2) - \iota(\overline{\eta}_2)) \\
&\quad + (\epsilon(\eta_2) - \iota(\overline{\eta}_2))(\epsilon(\eta_1) - \iota(\overline{\eta}_1)) \\
&= (\epsilon(\eta_1)\epsilon(\eta_2) + \epsilon(\eta_2)(\epsilon(\eta_1) \\
&\quad + (\iota(\overline{\eta}_1)\iota(\overline{\eta}_2) + \iota(\overline{\eta}_2)\iota(\overline{\eta}_1)) \\
&\quad - (\epsilon(\eta_1)\iota(\overline{\eta}_2) + \iota(\overline{\eta}_2)\epsilon(\eta_1)) \\
&\quad - (\epsilon(\eta_2)\iota(\overline{\eta}_1) + \iota(\overline{\eta}_1)\epsilon(\eta_2)) \\
&= 0,
\end{aligned}$$

since the $\epsilon(\eta_1), \ldots, \iota(\overline{\eta}_2)$ all anticommute, e.g.

$$\begin{aligned}
\epsilon(\eta_1)\iota(\overline{\eta}_2)\eta_2 \wedge \eta_3 &= \qquad\qquad \epsilon(\eta_1)\eta_3 = \quad \eta_1 \wedge \eta_3, \\
\iota(\overline{\eta}_2)\epsilon(\eta_1)\eta_2 \wedge \eta_3 &= \iota(\overline{\eta}_2)\eta_1 \wedge \eta_2 \wedge \eta_3 = -\,\eta_1 \wedge \eta_3.
\end{aligned}$$

Now $\dim_{\mathbb{C}} \mathrm{Cl}^{\mathbb{C}}(V) = 2^n = (\dim_{\mathbb{C}}(\Lambda W))^2 = \dim_{\mathbb{C}}(\mathrm{End}_{\mathbb{C}}(S))$, and since ρ has nontrivial kernel, we conclude

Theorem 1.11.4. *If $n = \dim_{\mathbb{R}} V$ is even, $\mathrm{Cl}^{\mathbb{C}}(V)$ is isomorphic to the algebra of complex linear endomorphisms of the spinor space S.* □

(Later on, we shall omit the symbol ρ and simply say that $\mathrm{Cl}^{\mathbb{C}}(V)$ operates on the spinor space S via Clifford multiplication, denoted by "$\cdot$"). Now since

$$\eta_j\overline{\eta}_j - \overline{\eta}_j\eta_j = 2ie_{2j-1}e_{2j},$$

we have

$$\Gamma = 2^{-m}\left(\eta_1\overline{\eta}_1 - \overline{\eta}_1\eta_1\right)\ldots\left(\eta_m\overline{\eta}_m - \overline{\eta}_m\eta_m\right)$$

and so Γ acts on the spinor space $S = \Lambda W$ via

$$\rho(\Gamma) = (-1)^m(\epsilon(\eta_1)\iota(\overline{\eta}_1) - \iota(\overline{\eta}_1)\epsilon(\eta_1)) \dots (\epsilon(\eta_m)\iota(\overline{\eta}_m) - \iota(\overline{\eta}_m)\epsilon(\eta_m)),$$

and for the same reasons as in the computation of $\rho({e_1}^2)$, we see that $\rho(\Gamma)$ equals $(-1)^k$ on $\Lambda^k W$. As above, any representation of $\mathrm{Cl}^{\mathbb{C}}(V)$, in particular ρ, decomposes into the eigenspaces of Γ for the eigenvalues ± 1, and so in the present case we have the decomposition

$$S^{\pm} := \Lambda^{\pm} W$$

where the + (−) sign on the right hand side denotes elements of even (odd) degree.

Since $\mathrm{Spin}\,(V)$ sits in $\mathrm{Cl}(V)$, hence in $\mathrm{Cl}^{\mathbb{C}}(V)$, any representation of the Clifford algebra $\mathrm{Cl}^{\mathbb{C}}(V)$ restricts to a representation of $\mathrm{Spin}\,(V)$, and we thus have a representation

$$\rho : \mathrm{Spin}\,(V) \to \mathrm{End}_{\mathbb{C}}(S).$$

Since $\mathrm{Spin}\,(V) \subset \mathrm{Cl}^+(V)$, $\mathrm{Spin}\,(V)$ leaves the spaces S^+ and S^- invariant, and thus the representation is not irreducible, but decomposes into the ones on S^+ and S^-. (The latter are in fact irreducible.) As in (1.11.12), multiplication by an element of $\mathrm{Cl}^-(V)$, in particular by a vector $v \in V$, exchanges S^+ and S^-.

Definition 1.11.6. The above representation ρ of $\mathrm{Spin}\,(V)$ on the spinor space S is called the *spinor representation*, and the representations on S^+ and S^- are called *half spinor representations*.

Note that the spinor space $S = \Lambda W$ is different from the Clifford space $\mathrm{Cl}(V)$ ($= \Lambda^*(V)$ as a vector space). $\mathrm{Cl}(V)$, and therefore also V, acts on both of them by Clifford multiplication.

We now want to extend the representation of $\mathrm{Spin}\,(V)$ to $\mathrm{Spin}^c\,(V)$.

Lemma 1.11.7. *Let* $\sigma : \mathrm{Spin}\,(V) \to \mathrm{End}_{\mathbb{C}}(T)$ *be a complex representation of* $\mathrm{Spin}\,(V)$ *on some vector space* T*, satisfying*

$$\sigma(-1) = -1.$$

Then σ *extends in a unique manner to a representation*

$$\widetilde{\sigma} : \mathrm{Spin}^c\,(V) \to \mathrm{End}_{\mathbb{C}}(T).$$

Proof. Since σ is complex linear, it commutes with multiplication by complex scalars, in particular with those of unit length. Thus, σ extends to $\sigma' : \mathrm{Spin}\,(V) \times S^1 \to \mathrm{End}_{\mathbb{C}}(T)$. Since $\sigma(-1) = -1$, it descends to $\mathrm{Spin}^c\,(V)$. □

Corollary 1.11.2. *The spinor and half spinor representations of* $\mathrm{Spin}\,(V)$ *possess unique extensions to* $\mathrm{Spin}^c\,(V)$. □

Of course, this is also clear from the fact that these representations of Spin (V) come from $\mathrm{Cl}^{\mathbb{C}}(V)$.

For $\mathrm{Cl}^c(\mathbb{R}^2)$, the spinor space is isomorphic to $\mathbb{C}^2$ and generated by $v_1 := 1$ and $v_2 := \eta_1 = \frac{1}{\sqrt{2}}(e_1 - ie_2)$, see (1.11.15). Since $e_1 = \frac{1}{\sqrt{2}}(\eta_1 + \bar{\eta}_1)$ and $e_2 = \frac{i}{\sqrt{2}}(\eta_1 - \bar{\eta}_1)$, we have

$$e_1 v_1 = v_2, \quad e_1 v_2 = -v_1, \quad e_2 v_1 = iv_2, \quad e_2 v_2 = iv_1,$$

that is, the action of $\mathrm{Cl}^c(\mathbb{R}^2)$ on its spinor space is given by the above representation (1.11.11) of $\mathbb{H}$ as $\mathbb{C}^{2\times 2}$ acting on $\mathbb{C}^2$.

Let us also discuss the example of $\mathrm{Cl}^c(\mathbb{R}^4)$ once more. We recall the isomorphism

$$\Gamma : \mathrm{Cl}^c(\mathbb{R}^4) \to \mathbb{C}^{4\times 4}$$

Γ in fact is the representation described in Theorem 1.11.4, and $\mathbb{C}^4$ is isomorphic to S_4. The formulas given above for the products $\Gamma(e_\alpha)\Gamma(e_\beta)$ also show that the representation admits a decomposition into two copies of $\mathbb{C}^2$ that is preserved by the elements of even order of $\mathrm{Cl}^c(\mathbb{R}^4)$. In fact, these yield the half spinor representations $S_4^\pm$ in dimension 4. In the above formulas, the upper left block corresponds to S^+, the lower right one to S^-.

In dimension 4, we also have a decomposition

$$\Lambda^2 = \Lambda^{2,+} \oplus \Lambda^{2,-} \quad (\Lambda^2 = \Lambda^2 V^*,\ \dim V = 4)$$

of exterior two forms. Namely, we have the Hodge $*$ operator (to be discussed in §2.1 for arbitrary dimensions) determined by

$$\begin{aligned}
*(e^1 \wedge e^2) &= e^3 \wedge e^4,\\
*(e^1 \wedge e^3) &= -e^2 \wedge e^4,\\
*(e^1 \wedge e^4) &= e^2 \wedge e^3,\\
*(e^2 \wedge e^3) &= e^1 \wedge e^4,\\
*(e^2 \wedge e^4) &= -e^1 \wedge e^3,\\
*(e^3 \wedge e^4) &= e^1 \wedge e^2
\end{aligned}$$

and linear extensions, where $e^1, \dots, e^4$ is an orthonormal frame in V^*.

We have

$$** = 1,$$

and $*$ thus has eigenvalues ± 1, and $\Lambda^{2,\pm}$ then are defined as the corresponding eigenspaces. Both these spaces are three dimensional. $\Lambda^{2,+}$ is spanned by $e^1 \wedge e^2 + e^3 \wedge e^4$, $e^1 \wedge e^3 - e^2 \wedge e^4$, $e^1 \wedge e^4 + e^2 \wedge e^3$, while $\Lambda^{2,-}$ is spanned by $e^1 \wedge e^2 - e^3 \wedge e^4$, $e^1 \wedge e^3 + e^2 \wedge e^4$, $e^1 \wedge e^4 - e^2 \wedge e^3$. Elements of $\Lambda^{2,+}$ are called selfdual, those of $\Lambda^{2,-}$ antiselfdual.

We have a bijective linear map $\Lambda V^* \to \mathrm{Cl}^2(V)$, given by $e^i \wedge e^j \to e_i \cdot e_j$ (where e^i is the orthonormal frame in V^* dual to the frame e_i in V).

Therefore, Γ induces a map $\Gamma^1 : \Lambda^2 V^* \to \mathrm{End}\,(\mathbb{C}^4)$. In the above decomposition of the representation of $\mathrm{Cl}^{c,ev}(\mathbb{R}^4)$, the selfdual forms then act only on $\mathbb{C}^2 \oplus \{0\}$, while the antiselfdual ones act only on $\{0\} \oplus \mathbb{C}^2$, as one directly sees from the formulae for $\Gamma(e_\alpha)\Gamma(e_\beta)$ and the description of the bases of $\Lambda^{2,\pm}$.

Finally, let us briefly summarize the situation in the odd dimensional case. Here, according to Corollary 1.11.1, $\mathrm{Cl}^{\mathbb{C}}(V)$ is a sum of two endomorphism algebras, and we therefore obtain two representations of $\mathrm{Cl}^{\mathbb{C}}(V)$. When restricted to $\mathrm{Spin}\,(V)$, these representations become isomorphic and irreducible. This yields the spinor representation in the odd dimensional case. We omit the details.

We also observe that the spinor representation is a unitary representation in a natural manner. For that purpose, we now extend the scalar product $\langle \cdot, \cdot \rangle$ from V to $V \otimes \mathbb{C}$ as a Hermitian product, i.e.

$$\Big\langle \sum_{i=1}^n \alpha_i e_i, \sum_{j=1}^n \beta_j e_j \Big\rangle = \sum_{i=1}^n \alpha_i \overline{\beta_i} \quad \text{for } \alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n \in \mathbb{C}.$$

Note that this is different from the above complex linear extensions $\langle \cdot, \cdot \rangle_{\mathbb{C}}$. This product extends to ΛV by letting the monomials $e_{i_1} \wedge \ldots e_{i_k}$, $1 \le i_1 < \ldots < \ldots i_k \le n$, constitute an orthonormal basis. From the above computations for the $\rho(e_j)$, one checks that each $\rho(e_j)$ preserves $\langle \cdot, \cdot \rangle$, i.e.

$$\langle \rho(e_j)s, \rho(e_j)s' \rangle = \langle s, s' \rangle \quad \text{for all } s,\ s' \in \Lambda W.$$

Of course, this then holds more generally for every $v \in V$ with $\|v\| = 1$, and then also for products $v_1 \ldots v_k$ with $\|v_j\| = 1$ for $j = 1, \ldots, k$. This implies

Corollary 1.11.3. *The induced representation of* $\mathrm{Pin}(V)$ *and* $\mathrm{Spin}\,(V)$ *on* $\mathrm{End}_{\mathbb{C}}(S)$ *preserves the Hermitian product* $\langle \cdot, \cdot \rangle$. □

Corollary 1.11.4.

$$\langle \rho(v)s, s' \rangle = -\langle s, \rho(v)s' \rangle \quad \textit{for all } s,\ s' \in \Lambda W,\ v \in V.$$

Proof. We may assume $\|v\| = 1$. Then $\rho(v)^2 = -1$, hence

$$\langle \rho(v)s, s' \rangle = -\langle \rho(v)s, \rho(v)\rho(v)s' \rangle = -\langle s, \rho(v)s' \rangle \quad \text{by Corollary 1.11.3.}$$

□

After these algebraic preparations, we may now define spin structures on an oriented Riemannian manifold M. At each point $x \in M$, we may take the tangent space T_xM as the vector space V for the definition of the Clifford algebra $\mathrm{Cl}(V)$, and we want to to construct vector bundles with fibers carrying the above constructions of spin groups and spinors.

We let TM be the tangent bundle of M. The Riemannian metric allows to reduce the structure group of TM to $\mathrm{SO}(n)$ ($n = \dim M$), and we obtain an associated principal bundle P over M with fiber $\mathrm{SO}(n)$, the so-called frame bundle of M.

Definition 1.11.7. A *spin structure* on M is a principal bundle $\widetilde{P}$ over M with fiber $\mathrm{Spin}\,(n)$ for which the quotient of each fiber by the center ± 1 is isomorphic to the above frame bundle of M. A Riemannian manifold with a fixed spin structure is called a *spin manifold*.

In other words, we require that the following diagram commutes,

$$\begin{tikzcd} \widetilde{P} \arrow[r, "\rho"] \arrow[dr, "\pi"'] & P \arrow[d, "\pi"] \\ & M \end{tikzcd}$$

where π denotes the projection onto the base point, and ρ is the nontrivial double covering $\rho : \mathrm{Spin}\,(n) \to \mathrm{SO}(n)$ on each fiber as described in Theorem 1.11.2. This is also expressed by saying that the frame bundle is lifted to a $\mathrm{Spin}\,(n)$ bundle. It is important to note that such a lift need not always be possible. One way to realize this is by considering the corresponding transition functions. We recall from §1.8 that the frame bundle P for each trivializing covering $(U_\alpha)_{\alpha \in A}$ of M induces transition functions

$$\varphi_{\beta\alpha} : U_\alpha \cap U_\beta \to \mathrm{SO}(n)$$

satisfying

$$\begin{aligned} \varphi_{\alpha\alpha}(x) &= \mathrm{id} \quad \text{for } x \in U_\alpha \\ \varphi_{\alpha\beta}(x)\varphi_{\beta\alpha} &= \mathrm{id} \quad \text{for } x \in U_\alpha \cap U_\beta \\ \varphi_{\alpha\gamma}(x)\varphi_{\gamma\beta}(x)\varphi_{\beta\alpha}(x) &= \mathrm{id} \quad \text{for } x \in U_\alpha \cap U_\beta \cap U_\gamma. \end{aligned}$$

Lifting the frame bundle to a $\mathrm{Spin}\,(n)$ bundle then requires finding transition functions

$$\widetilde{\varphi}_{\beta\alpha} : U_\alpha \cap U_\beta \to \mathrm{Spin}\,(n)$$

with

$$\rho\,(\widetilde{\varphi}_{\beta\alpha}) = \varphi_{\beta\alpha} \quad \text{for all } \beta,\ \alpha \tag{1.11.24}$$

and satisfying the same relations as the $\varphi_{\beta\alpha}$. By making the U_α sufficiently small, in particular simply connected, lifting the $\varphi_{\beta\alpha}$ to $\widetilde{\varphi}_{\beta\alpha}$ satisfying (1.11.24), is no problem, but the problem arises with the third relation, i.e.

$$\widetilde{\varphi}_{\alpha\beta}(x)\widetilde{\varphi}_{\beta\gamma}(x)\widetilde{\varphi}_{\gamma\alpha}(x) = \mathrm{id} \quad \text{for } x \in U_\alpha \cap U_\beta \cap U_\gamma. \tag{1.11.25}$$

Namely, it may happen that $\widetilde{\varphi}_{\alpha\beta}(x)\widetilde{\varphi}_{\beta\gamma}(x)$ and $\widetilde{\varphi}_{\gamma\alpha}(x)$ differ by the nontrivial deck transformation of the covering $\rho : \mathrm{Spin}\,(n) \to \mathrm{SO}(n)$.

In fact, the existence of a spin structure, i.e. the possibility of such a lift, depends on a topological condition, the vanishing of the so-called Stiefel-Whitney class $w_2(M) \in H^2(M, \mathbb{Z}_2)$. Here, however, we cannot define these topological concepts. Furthermore, if a spin structure exists, it need not to be unique. For example, a compact oriented two-dimensional Riemannian manifold of genus[5] g carries 2^{2g} different

[5] The genus is a basic topological invariant of a compact surface. There are several different ways of defining or characterizing it, see [147]. For instance, it equals the first Betti number b_1, the dimension of the first cohomology, that will be defined in the next chapter.

spin structures. In particular, the two-dimensional sphere S^2 has genus 0 and hence carries a unique spin structure.

Let us assume that M possesses a spin structure $\widetilde{P} \to M$. Since the fiber Spin (n) of $\widetilde{P}$ operates on the spinor space S_n and for even n also on the half spinor spaces $S_n^{\pm}$ via the (half) spinor representations, we obtain associated vector bundles $\mathcal{S}_n$, $\mathcal{S}_n^{\pm}$ over M with structure group Spin (n),

$$\mathcal{S}_n := \widetilde{P} \times_{\mathrm{Spin}\,(n)} S_n, \quad \mathcal{S}_n^{\pm} := \widetilde{P} \times_{\mathrm{Spin}\,(n)} S_n^{\pm},$$

with

$$\mathcal{S}_n = \mathcal{S}_n^{+} \oplus \mathcal{S}_n^{-} \quad \text{for even } n.$$

Definition 1.11.8. $\mathcal{S}_n$ is called the *spinor bundle*, $\mathcal{S}_n^{\pm}$ the *half spinor bundles associated with the spin structure* $\widetilde{P}$. Sections are called *(half) spinor fields*.

From Corollary 1.11.3, we infer that these bundles carry Hermitian products that are invariant under the action of Spin (n), and even of Pin(n), on each fiber. In particular, Clifford multiplication by a unit vector in $\mathbb{R}^n \subset \mathrm{Cl}(\mathbb{R}^n)$ is an isometry on each fiber.

We may also consider Spin^c (n) in place of of Spin (n) and ask for a lift of the frame bundle P over M to a principal Spin^c (n) bundle $\widetilde{P}^c$. Of course, the requirement here is that the map from a fiber of $\widetilde{P}^c$ to the corresponding one of P is given by the homomorphism

$$\mathrm{Spin}^c\,(n) \to \mathrm{SO}(n)$$

obtained from (1.11.14) by projecting onto the first factor.

Definition 1.11.9. Such a principal Spin^c (n) bundle $\widetilde{P}^c$ (if it exists) is called a *spin^c structure on M*. An oriented Riemannian manifold M equipped with a fixed spin^c structure is called a *spin^c manifold*.

Again, the existence of a spin^c structure depends on a topological condition, namely that $w_2(M)$ lifts to an integral class in $H^2(M, \mathbb{Z}_2)$. Again, however, we cannot explain this here any further. We point out, however, that the required condition is satisfied for all oriented Riemannian manifolds of dimension 4. Thus, each oriented four-manifold possesses a spin^c structure.

Given a spin^c structure, we may also consider the homomorphism

$$\mathrm{Spin}^c\,(n) \to S^1$$

obtained from (1.11.14) by projecting on the second factor. Identifying S^1 with U(1), we see that a spin^c structure induces a set of transition functions for a vector bundle L with fiber $\mathbb{C}$, a so called (complex) line bundle.

Definition 1.11.10. The line bundle L is called the *determinant line bundle of the spin^c structure*.

As in the case of a spin structure, a spin^c structure induces (half) spinor bundles $\mathcal{S}_n^{\pm}$, cf. Corollary 1.11.2.

We return to the frame bundle P over M with fiber $\mathrm{SO}(n)$. $\mathrm{SO}(n)$ acts on $\mathrm{Cl}(\mathbb{R}^n)$ and $\mathrm{Cl}^c(\mathbb{R}^n)$ simply by extending the action of $\mathrm{SO}(n)$ on $\mathbb{R}^n$. Thus, P induces bundles

$$\mathrm{Cl}(P) = P \times_{\mathrm{SO}(n)} \mathrm{Cl}(\mathbb{R}^n)$$
$$\mathrm{Cl}^{\mathbb{C}}(P) = P \times_{\mathrm{SO}(n)} \mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^n)$$

of Clifford algebras.

Definition 1.11.11. The bundles $\mathrm{Cl}(P)$ and $\mathrm{Cl}^{\mathbb{C}}(P)$ are called the *Clifford bundles*.

Again, these Clifford bundles can be decomposed into bundles of elements of even and of odd degree. The chirality operator Γ (cf. Definition 1.11.3) is invariant under the action of $\mathrm{SO}(n)$,and it therefore defines a section of $\mathrm{Cl}^{\mathbb{C}}(P)$ of norm 1.

The definition of the Clifford bundles did not need a spin or spin^c structure on M. But suppose now that we do have such a structure, a spin structure, say. $\mathrm{Spin}\,(n)$ acts on $\mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^n)$ by conjugation.

$$\rho(a)v = ava^{-1} \quad \text{for } a \in \mathrm{Spin}\,(n), v \in \mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^n) \tag{1.11.26}$$

(cf. Theorem 1.11.1 (note that $a^t = a^{-1}$ for $a \in \mathrm{Spin}\,(n)$ by (1.11.9)) for the action of $\mathrm{Spin}\,(n)$ on $\mathbb{R}^n$, and extend this action to $\mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^n)$; this is of course induced by the above action of $\mathrm{SO}(n)$ on $\mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^n)$). This action commutes with the action of $\mathrm{Spin}\,(n)$ on $\mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^n)$ given by (1.11.26) and the action of $\mathrm{Spin}\,(n)$ on S_n; namely for $a \in \mathrm{Spin}\,(n)$, $v \in \mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^n)$, $s \in S_n$

$$(ava^{-1})(as) = a(vs). \tag{1.11.27}$$

This compatibility with the $\mathrm{Spin}\,(n)$ actions ensures that we get a global action

$$\mathrm{Cl}^{\mathbb{C}}(\mathbb{R}^n) \times \mathcal{S}_n \to \mathcal{S}_n, \tag{1.11.28}$$

which is the above action by Clifford multiplication on each fiber. Recalling that the space $\mathbb{R}^n$ here is a tangent space T_xM, we thus can Clifford multiply a tangent vector $v \in T_xM$ at x with a spinor $s \in \mathcal{S}_{n,x}$ at x. In fact, since a vector is an odd element in the Clifford algebra, we have the action

$$T_xM \times \mathcal{S}_{n,x}^{\pm} \to \mathcal{S}_{n,x}^{\mp}. \tag{1.11.29}$$

According to Corollary 1.11.4, this Clifford multiplication is skew-symmetric w.r.t. the Hermitian product on $\mathcal{S}_{n,x}$, that is,

$$\langle vs, s' \rangle = -\langle s, vs' \rangle \quad \text{for all } s,\ s' \in \mathcal{S}_{n,x},\ v \in T_xM. \tag{1.11.30}$$

Perspectives. References for this section are [7], [176], [267], [18], [221], [196].

Exercises for Chapter 1

1. Give five more examples of differentiable manifolds besides those discussed in the text.

2. Determine the tangent space of S^n. (Give a concrete description of the tangent bundle of S^n as a submanifold of $S^n \times \mathbb{R}^{n+1}$.)

3. Let M be a differentiable manifold, $\tau : M \to M$ an involution without fixed points, i.e. $\tau \circ \tau = \mathrm{id}, \tau(x) \neq x$ for all $x \in M$. We call points x and y in M equivalent if $y = \tau(x)$. Show that the space M/τ of equivalence classes possesses a unique differentiable structure for which the projection $M \to M/\tau$ is a local diffeomorphism.

   Discuss the example $M = S^n \subset \mathbb{R}^{n+1}, \tau(x) = -x$. M/τ is real projective space $\mathbb{RP}^n$.

4. a: Let N be a differentiable manifold, $f : M \to N$ a homeomorphism. Introduce a structure of a differentiable manifold on M such that f becomes a diffeomorphism. Show that such a differentiable structure is unique.

   b: Can the boundary of a cube, i.e. the set $\{x \in \mathbb{R}^n ; \max\{|x_i| : i = 1, \ldots, n\} = 1\}$ be equipped with a structure of a differentiable manifold?

5. We equip $\mathbb{R}^{n+1}$ with the inner product

$$\langle x, y \rangle := -x^0 y^0 + x^1 y^1 + \ldots + x^n y^n$$

   for $x = (x^0, x^1, \ldots, x^n), y = (y^0, y^1, \ldots, y^n)$. We put

$$H^n := \{x \in \mathbb{R}^{n+1} : \langle x, x \rangle = -1, x_0 > 0\}.$$

   Show that $\langle \cdot, \cdot \rangle$ induces a Riemannian metric on the tangent spaces $T_p H^n \subset T_p \mathbb{R}^{n+1}$ for $p \in H^n$. H^n is called hyperbolic space.

6. In the notations of Exercise 5, let

$$s = (-1, 0, \ldots, 0) \in \mathbb{R}^{n+1}$$
$$f(x) := s - \frac{2(x - s)}{\langle x - s, x - s \rangle}$$

   Show that $f : H^n \to \{\xi \in \mathbb{R}^n : |\xi| < 1\}$ is a diffeomorphism (here, $\mathbb{R}^n = \{(0, x^1, \ldots, x^n)\} \subset \mathbb{R}^{n+1}$). Show that in this chart, the metric assumes the form

$$\frac{4}{(1 - |\xi|^2)^2} d\xi^i \otimes d\xi^i.$$

7. Determine the geodesics of H^n in the chart given in Exercise 6. (The geodesics through 0 are the easiest ones.)

   Hint for Exercises 5, 6, 7: Consult §4.4.

8. Determine the exponential map of the sphere S^n, for example at the north pole p. Write down normal coordinates. Compute the supremum of the radii of balls in T_pS^n on which $\exp_p$ is injective. Where does $\exp_p$ have maximal rank?

9. Same as 8. for the flat torus generated by $(1, 0)$ and $(0, 1) \in \mathbb{R}^2$.

10. What is the transformation behavior of the Christoffel symbols under coordinate changes? Do they define a tensor?

11. Let $c_0, c_1 : [0, 1] \to M$ be smooth curves in a Riemannian manifold.

    If $d(c_0(t), c_1(t)) < i(c_0(t))$ for all t, there exists a smooth map $c : [0, 1] \times [0, 1] \to M$ with $c(t, 0) = c_0(t), c(t, 1) = c_1(t)$ for which the curves $c(t, \cdot)$ are geodesics for all t.

12. Consider the surface S of revolution obtained by rotating the curve $(x, y = e^x, z = 0)$ in the plane, i.e. the graph of the exponential function, about the $x - ax$ in Euclidean 3-space, equiped with the induced Riemannian metric from that Euclidean space. Show that X is complete and compute its injectivity radius.

13. Show that the structure group of the tangent bundle of an oriented d-dimensional Riemannian manifold can be reduced to $\mathrm{SO}(d)$.

14. Can one define the normal bundle of a differentiable submanifold of a differentiable manifold in a meaningful manner without introducing a Riemannian metric?

15. Let M be a differentiable submanifold of the Riemannian manifold N. M then receives an induced Riemannian metric, and this metric defines a distance function and a topology on M, as explained in §1.4. Show that this topology coincides with the topology on M that is induced from the topology of N.

16. We consider the constant vector field $X(x) = a$ for all $x \in \mathbb{R}^{n+1}$. We obtain a vector field $\tilde{X}(x)$ on S^n by projecting $X(x)$ onto T_xS^n for $x \in S^n$. Determine the corresponding flow on S^n.

17. Let T be the flat torus generated by $(1, 0)$ and $(0, 1) \in \mathbb{R}^2$, with projection $\pi : \mathbb{R}^2 \to T$. For which vector fields X on $\mathbb{R}^2$ can one define a vector field π_*X on T in a meaningful way? Determine the flow of π_*X on T for a constant vector field X.

18. Compute a formula for the Lie derivative (in the direction of a vector field) for a p-times contravariant and q-times covariant tensor.

19. Show that for arbitrary vector fields X, Y, the Lie derivative satisfies

$$L_X \circ L_Y - L_Y \circ L_X = L_{[X,Y]}.$$

20. Prove Corollaries 4.2.3 and 4.2.4 below with the arguments used in the proofs of Theorem 1.4.5 and Corollary 1.4.2.

Chapter 2

De Rham Cohomology and Harmonic Differential Forms

2.1 The Laplace Operator

We need some preparations from linear algebra. Let V be a real vector space with a scalar product $\langle \cdot, \cdot \rangle$, and let $\Lambda^p V$ be the p-fold exterior product of V. We then obtain a scalar product on $\Lambda^p V$ by

$$\langle v_1 \wedge \ldots \wedge v_p, w_1 \wedge \ldots \wedge w_p \rangle = \det(\langle v_i, w_j \rangle) \tag{2.1.1}$$

and bilinear extension to $\Lambda^p(V)$. If $e_1, \ldots, e_d$ is an orthonormal basis of V,

$$e_{i_1} \wedge \ldots \wedge e_{i_p} \quad \text{with}\, 1 \leq i_1 < i_2 < \ldots < i_p \leq d \tag{2.1.2}$$

constitute an orthonormal basis of $\Lambda^p V$.

An *orientation* on V is obtained by distinguishing a basis of V as positive. Any other basis that is obtained from this basis by a base change with positive determinant then is likewise called positive, and the remaining bases are called negative.

Let now V carry an orientation. We define the linear star operator

$$* : \Lambda^p(V) \to \Lambda^{d-p}(V) \quad (0 \leq p \leq d)$$

by

$$*(e_{i_1} \wedge \ldots \wedge e_{i_p}) = e_{j_1} \wedge \ldots \wedge e_{j_{d-p}}, \tag{2.1.3}$$

where $j_1, \ldots, j_{d-p}$ is selected such that $e_{i_1}, \ldots, e_{i_p}, e_{j_1}, \ldots, e_{j_{d-p}}$ is a positive basis of V. Since the star operator is supposed to be linear, it is determined by its values on some basis (2.1.3).

In particular,

$$*(1) = e_1 \wedge \ldots \wedge e_d \tag{2.1.4}$$
$$*(e_1 \wedge \ldots \wedge e_d) = 1, \tag{2.1.5}$$

if $e_1, \ldots, e_d$ is a positive basis.

From the rules of multilinear algebra, it easily follows that if A is a $d \times d$-matrix, and if $f_1, \ldots, f_p \in V$, then

$$*(Af_1 \wedge \ldots \wedge Af_p) = (\det A) * (f_1 \wedge \ldots \wedge f_p).$$

In particular, this implies that the star operator does not depend on the choice of positive orthonormal basis in V, as any two such bases are related by a linear transformation with determinant 1.

For a negative basis instead of a positive one, one gets a minus sign on the right hand sides of (2.1.3), (2.1.4), (2.1.5).

Lemma 2.1.1. $** = (-1)^{p(d-p)} : \Lambda^p(V) \to \Lambda^p(V)$.

Proof. $**$ maps $\Lambda^p(V)$ onto itself. Suppose

$$*(e_{i_1} \wedge \ldots \wedge e_{i_p}) = e_{j_1} \wedge \ldots \wedge e_{j_{d-p}} \quad \text{(cf. (2.1.3))}.$$

Then

$$* * (e_{i_1} \wedge \ldots \wedge e_{i_p}) = \pm e_{i_1} \wedge \ldots \wedge e_{i_p},$$

depending on whether $e_{j_1}, \ldots, e_{j_{d-p}}, e_{i_1}, \ldots, e_{i_p}$ is a positive or negative basis of V. Now

$$\begin{aligned} &e_{i_1} \wedge \ldots \wedge e_{i_p} \wedge e_{j_1} \wedge \ldots \wedge e_{j_{d-p}} \\ &\quad = (-1)^{p(d-p)} e_{j_1} \wedge \ldots \wedge e_{j_{d-p}} \wedge e_{i_1} \wedge \ldots \wedge e_{i_p}, \end{aligned}$$

and $(-1)^{p(d-p)}$ thus is the determinant of the base change from $e_{i_1}, \ldots, e_{j_{d-p}}$ to $e_{j_1}, \ldots, e_{i_p}$. □

Lemma 2.1.2. *For* $v, w \in \Lambda^p(V)$

$$\langle v, w \rangle = *(w \wedge *v) = *(v \wedge *w). \tag{2.1.6}$$

Proof. It suffices to show (2.1.6) for elements of the basis (2.1.2). For any two different such basis vectors, $w \wedge *v = 0$, whereas

$$\begin{aligned} *(e_{i_1} \wedge \ldots \wedge e_{i_p} \wedge *(e_{i_1} \wedge \ldots \wedge e_{i_p})) &= *(e_1 \wedge \ldots \wedge e_d), \quad \text{where } e_1, \ldots, e_d \\ &\qquad \text{is an orthonormal basis (2.1.3)} \\ &= 1 \quad \text{by (2.1.5)}, \end{aligned}$$

and the claim follows. □

Remark. We may consider $\langle \cdot, \cdot \rangle$ as a scalar product on

$$\Lambda(V) := \bigoplus_{p=0}^{d} \Lambda^p(V)$$

with $\Lambda^p(V)$ and $\Lambda^q(V)$ being orthogonal for $p \neq q$.

Lemma 2.1.3. *Let $v_1, \dots, v_d$ be an arbitrary positive basis of V. Then*

$$*(1) = \frac{1}{\sqrt{\det(\langle v_i, v_j \rangle)}} v_1 \wedge \dots \wedge v_d. \tag{2.1.7}$$

Proof. Let $e_1, \dots, e_d$ be a positive orthonormal basis as before. Then

$$v_1 \wedge \dots \wedge v_d = (\det(\langle v_i, v_j \rangle))^{\frac{1}{2}} e_1 \wedge \dots \wedge e_d,$$

and the claim follows from (2.1.4). □

Let now M be an oriented Riemannian manifold of dimension d. Since M is oriented, we may select an orientation on all tangent spaces $T_x M$, hence also on all cotangent spaces $T_x^* M$ in a consistent manner. We simply choose the Euclidean orthonormal basis $\frac{\partial}{\partial x^1}, \dots, \frac{\partial}{\partial x^d}$ of $\mathbb{R}^d$ as being positive. Since all chart transitions of an oriented manifold have positive functional determinant, calling the basis $d\varphi^{-1}(\frac{\partial}{\partial x^1}), \dots, d\varphi^{-1}(\frac{\partial}{\partial x^d})$ of $T_x M$ positive, will not depend on the choice of the chart.

Since M carries a Riemannian structure, we have a scalar product on each $T_x^* M$. We thus obtain a star operator

$$* : \Lambda^p(T_x^* M) \to \Lambda^{d-p}(T_x^* M),$$

i.e. a base point preserving operator

$$* : \Omega^p(M) \to \Omega^{d-p}(M) \quad (\Omega^p(M) = \Gamma(\Lambda^p(M))).$$

We recall that the metric on $T_x^* M$ is given by $(g^{ij}(x)) = (g_{ij}(x))^{-1}$. Therefore, by Lemma 2.1.3 we have in local coordinates

$$*(1) = \sqrt{\det(g_{ij})} dx^1 \wedge \dots \wedge dx^d. \tag{2.1.8}$$

This expression is called the *volume form.*
In particular

$$\mathrm{Vol}\,(M) := \int_M *(1) \tag{2.1.9}$$

(provided this is finite).

For $\alpha, \beta \in \Omega^p(M)$ with compact support, we define the L^2-product as

$$\begin{aligned}(\alpha, \beta) :&= \int_M \langle \alpha, \beta \rangle * (1) \\ &= \int_M \alpha \wedge * \beta \quad \text{by Lemma 2.1.2}.\end{aligned}$$

This product on $\Omega^p(M)$ is obviously bilinear and positive definite.
We shall also use the L^2-norm

$$\|\alpha\| := (\alpha, \alpha)^{1/2}. \tag{2.1.10}$$

(In 2.2 below, we shall also introduce another norm, the Sobolev norm $\| \cdot \|_{H^{1,2}}$.) So far, we have considered only smooth sections of vector bundles, in particular only smooth p-forms. For later purposes, we shall also need L^p- and Sobolev spaces of sections of vector bundles. For this aim, from now on, we deviate from Definition 1.8.3 and don't require sections to be smooth anymore. We let E be a vector bundle over M, $s : M \to E$ a section of E with compact support. We say that s is contained in the *Sobolev space* $H^{k,r}(E)$, if for any bundle atlas with the property that on compact sets all coordinate changes and all their derivatives are bounded (it is not difficult to obtain such an atlas, by making coordinate neighborhoods smaller if necessary), and for any bundle chart from such an atlas,

$$\varphi : E_{|U} \to U \times \mathbb{R}^n$$

we have that $\varphi \circ s_{|U}$ is contained in $H^{k,r}(U)$. We note the following consistency property: If $\varphi_1 : E_{|U_1} \to U_1 \times \mathbb{R}^n$, $\varphi_2 : E_{|U_2} \to U_2 \times \mathbb{R}^n$ are two such bundle charts, then $\varphi_1 \circ s_{|U_1 \cap U_2}$ is contained in $H^{k,r}(U_1 \cap U_2)$ if and only if $\varphi_2 \circ s_{|U_1 \cap U_2}$ is contained in this space. The reason is that the coordinate change $\varphi_2 \circ \varphi_1^{-1}$ is of class C^∞, and all derivatives are bounded on the support of s which was assumed to be compact.

We can extend our product $(\cdot, \cdot)$ to $L^2(\Omega^p(M))$. It remains bilinear, and also positive definite, because as usual, in the definition of L^2, functions that differ only on a set of measure zero are identified.

We now make the assumption that M is compact, in order not to always have to restrict our considerations to compactly supported forms.

Definition 2.1.1. d^* is the operator which is (formally) adjoint to d on $\bigoplus_{p=0}^{d} \Omega^p(M)$ w.r.t. $(\cdot, \cdot)$. This means that for $\alpha \in \Omega^{p-1}(M), \beta \in \Omega^p(M)$

$$(d\alpha, \beta) = (\alpha, d^* \beta); \tag{2.1.11}$$

d^* therefore maps $\Omega^p(M)$ to $\Omega^{p-1}(M)$.

Lemma 2.1.4. $d^* : \Omega^p(M) \to \Omega^{p-1}(M)$ *satisfies*

$$d^* = (-1)^{d(p+1)+1} * d * . \tag{2.1.12}$$

Proof. For $\alpha \in \Omega^{p-1}(M), \beta \in \Omega^p(M)$

$$\begin{aligned} d(\alpha \wedge *\beta) &= d\alpha \wedge *\beta + (-1)^{p-1}\alpha \wedge d * \beta \\ &= d\alpha \wedge *\beta + (-1)^{p-1}(-1)^{(p-1)(d-p+1)}\alpha \wedge * * (d * \beta) \\ &\qquad \text{by Lemma 2.1.1 } (d * \beta \text{ is a } (d-p+1)\text{-form}) \\ &= d\alpha \wedge *\beta - (-1)^{d(p+1)+1}\alpha \wedge * * d * \beta \\ &= \pm * (\langle d\alpha, \beta\rangle - (-1)^{d(p+1)+1}\langle \alpha, *d * \beta\rangle). \end{aligned}$$

We integrate this formula. By Stokes' theorem, the integral of the left hand side vanishes, and the claim results. □

Definition 2.1.2. The *Laplace(-Beltrami) operator* on $\Omega^p(M)$ is

$$\Delta = dd^* + d^*d : \Omega^p(M) \to \Omega^p(M).$$

$\omega \in \Omega^p(M)$ is called *harmonic* if

$$\Delta\omega = 0\,.$$

Remark. Since two stars appear on the right hand side of (2.1.12), d^* and hence also Δ may also be defined by (2.1.12) on nonorientable Riemannian manifolds. We just define it locally, hence globally up to a choice of sign which then cancels in (2.1.12). Similarly, the L^2-product can be defined on nonorientable Riemannian manifolds, because the ambiguity of sign of the $*$ involved cancels with the one coming from the integration.

More precisely, one should write

$$\begin{aligned} d^p &: \Omega^p(M) \to \Omega^{p+1}(M) \\ d^* &: \Omega^p(M) \to \Omega^{p-1}(M). \end{aligned}$$

Then

$$\Delta^p = d^{p-1}d^* + d^*d^p : \Omega^p(M) \to \Omega^p(M).$$

Nevertheless, we shall usually omit the index p.

Corollary 2.1.1. Δ *is (formally) selfadjoint, i.e.*

$$(\Delta\alpha, \beta) = (\alpha, \Delta\beta) \quad \textit{for } \alpha, \beta \in \Omega^p(M).$$

Proof. Directly from the definition of Δ. □

Lemma 2.1.5. $\Delta\alpha = 0 \iff d\alpha = 0$ *and* $d^*\alpha = 0$.

Proof.
" $\Leftarrow$ ": obvious.
" $\Rightarrow$ ": $(\Delta\alpha, \alpha) = (dd^*\alpha, \alpha) + (d^*d\alpha, \alpha) = (d^*\alpha, d^*\alpha) + (d\alpha, d\alpha)$.
Since both terms on the right hand side are nonnegative and vanish only if $d\alpha = 0 = d^*\alpha$, $\Delta\alpha = 0$ implies $d\alpha = 0 = d^*\alpha$. □

Corollary 2.1.2. *On a compact Riemannian manifold, every harmonic function is constant.* □

Lemma 2.1.6. $*\Delta = \Delta *$.

Proof. Direct computation. □

We want to compare the Laplace operator as defined here with the standard one on $\mathbb{R}^d$. For this purpose, let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. We have

$$df = \frac{\partial f}{\partial x^i} dx^i$$

and for $\varphi = \varphi_i dx^i$ with compact support, and $*\varphi = \sigma_{i=1}^d (-1)^{i-1} \varphi_i dx^1 \wedge \ldots \wedge \widehat{dx^i} \wedge \ldots \wedge dx^d$

$$\begin{aligned}(df, \varphi) &= \int_{\mathbb{R}^d} \frac{\partial f}{\partial x^i} \varphi_i dx^1 \wedge \ldots \wedge dx^d \\ &= -\int_{\mathbb{R}^d} f \frac{\partial \varphi^i}{\partial x^i} dx^1 \wedge \ldots \wedge dx^d, \text{ since } \varphi \text{ is compactly supported.}\end{aligned}$$

It follows that $d^*\varphi = -\frac{\partial \varphi^i}{\partial x^i} = -\operatorname{div} \varphi$, and

$$\Delta f = d^*df = -\sum_{i=1}^d \frac{\partial^2 f}{(\partial x^i)^2} = -\mathrm{div}\,(\mathrm{grad}\, f).$$

This Laplace operator therefore differs from the usual one on $\mathbb{R}^d$ by a minus sign. This is regrettable, but cannot be changed any more since the notation has been established too thoroughly. With our definition above, Δ is a positive operator.

More generally, for a differentiable function, the Laplace-Beltrami operator is $f : M \to \mathbb{R}$

$$\Delta f = -\frac{1}{\sqrt{g}} \frac{\partial}{\partial x^j} \Big(\sqrt{g} g^{ij} \frac{\partial f}{\partial x^i} \Big), \tag{2.1.13}$$

with $g := \det(g_{ij})$. This is seen as follows:

Since for functions, i.e. 0-forms, we have $d^* = 0$, we get for $\varphi : M \to \mathbb{R}$

(differentiable with compact support)

$$\begin{aligned}\int \Delta f\cdot\varphi\sqrt{g}dx^1\wedge\ldots\wedge dx^d &= (\Delta f,\varphi) = (df,d\varphi)\\ &= \int \langle df, d\varphi\rangle * (1)\\ &= \int g^{ij}\frac{\partial f}{\partial x^i}\frac{\partial\varphi}{\partial x^j}\sqrt{g}dx^1\ldots dx^d\\ &= -\int \frac{1}{\sqrt{g}}\frac{\partial}{\partial x^j}\left(\sqrt{g}g^{ij}\frac{\partial f}{\partial x^i}\right)\varphi\sqrt{g}dx^1\ldots dx^d,\end{aligned}$$

and since this holds for all $\varphi\in C_0^\infty(M,\mathbb{R})$, (2.1.13) follows.

For a function f, we may define its gradient as

$$\nabla f := \operatorname{grad} f := g^{ij}\frac{\partial f}{\partial x^i}\frac{\partial}{\partial x^j}. \tag{2.1.14}$$

We thus have for any vector field X

$$\langle \operatorname{grad} f, X\rangle = X(f) = df(X). \tag{2.1.15}$$

The divergence of a vector field $Z = Z^i\frac{\partial}{\partial x^i}$ is defined as

$$\operatorname{div} Z := \frac{1}{\sqrt{g}}\frac{\partial}{\partial x^j}(\sqrt{g}Z^j) = \frac{1}{\sqrt{g}}\frac{\partial}{\partial x^j}\left(\sqrt{g}g^{ij}\left\langle Z,\frac{\partial}{\partial x^i}\right\rangle\right). \tag{2.1.16}$$

(2.1.13) then becomes

$$\Delta f = -\operatorname{div}\operatorname{grad} f. \tag{2.1.17}$$

In particular, if M is compact, and $f : M\to\mathbb{R}$ is a smooth function, then as a consequence of (2.1.17) and (2.1.16) or (2.1.13) and the Gauss theorem, we have

$$\int_M \Delta f * (1) = 0. \tag{2.1.18}$$

We now want to compute the Euclidean Laplace operator for p-forms. It is denoted by Δ_e; likewise, the star operator w.r.t. the Euclidean metric is denoted by $*_e$, and d^* is the operator adjoint to d w.r.t. the Euclidean scalar product.

Let now

$$\omega = \omega_{i_1\ldots i_p}dx^{i_1}\wedge\ldots\wedge dx^{i_p}$$

be a p-form on an open subset of $\mathbb{R}^d$, as usual with an increasing p-tuple $1\le i_1 < i_2 < \ldots < i_p \le d$. We choose $j_1,\ldots,j_{d-p}$ such that $\frac{\partial}{\partial x^{i_1}},\ldots,\frac{\partial}{\partial x^{i_p}},\frac{\partial}{\partial x^{j_1}},\ldots,\frac{\partial}{\partial x^{i_{d-p}}}$ is a positive orthonormal basis of $\mathbb{R}^d$. In the sequel always

$$\ell\in\{1,\ldots,p\}, k\in\{1,\ldots,d-p\}.$$

Now

$$
\begin{aligned}
d\omega &= \sum_{k=1}^{d-p} \frac{\partial \omega_{i_1 \dots i_p}}{\partial x^{j_k}} dx^{j_k} \wedge dx^{i_1} \wedge \dots \wedge dx^{i_p} \\
*_e d\omega &= \sum_{k=1}^{d-p} (-1)^{p+k-1} \frac{\partial \omega_{i_1 \dots i_p}}{\partial x^{j_k}} dx^{j_1} \wedge \dots \wedge \widehat{dx^{j_k}} \wedge \dots \wedge dx^{j_{d-p}} \qquad (2.1.19)\\
d *_e d\omega &= \sum_{k=1}^{d-p} (-1)^{p+k-1} \frac{\partial^2 \omega_{i_1 \dots i_p}}{(\partial x^{j_k})^2} dx^{j_k} \wedge dx^{j_1} \wedge \dots \wedge \widehat{dx^{j_k}} \wedge \dots \wedge dx^{i_{d-p}} \\
&\quad + \sum_{k=1}^{d-p} \sum_{\ell=1}^{p} (-1)^{p+k-1} \frac{\partial^2 \omega_{i_1 \dots i_p}}{\partial x^{j_k} \partial x^{i_\ell}} dx^{i_\ell} \wedge dx^{j_1} \wedge \dots \wedge \widehat{dx^{j_k}} \wedge \dots \wedge dx^{i_{d-p}} \qquad (2.1.20)\\
*_e d *_e d\omega &= \sum_{k=1}^{d-p} (-1)^{p+p(d-p)} \frac{\partial^2 \omega_{i_1 \dots i_p}}{(\partial x^{j_k})^2} dx^{i_1} \wedge \dots \wedge dx^{i_p} \\
&\quad + \sum_{k=1}^{d-p} \sum_{\ell=1}^{p} (-1)^{pd+\ell} \frac{\partial^2 \omega_{i_1 \dots i_p}}{\partial x^{j_k} \partial x^{i_\ell}} dx^{j_k} \wedge dx^{i_1} \wedge \dots \wedge \widehat{dx^{i_\ell}} \wedge \dots \wedge dx^{i_p}. \qquad (2.1.21)
\end{aligned}
$$

Hence with (2.1.12)

$$
\begin{aligned}
d^* d\omega &= \sum_{k=1}^{d-p} (-1) \frac{\partial^2 \omega_{i_1 \dots i_p}}{(\partial x^{j_k})^2} dx^{i_1} \wedge \dots \wedge dx^{i_p} \\
&\quad + \sum_{k=1}^{d-p} \sum_{\ell=1}^{p} (-1)^{\ell+1} \frac{\partial^2 \omega_{i_1 \dots i_p}}{\partial x^{j_k} \partial x^{i_\ell}} dx^{j_k} \wedge dx^{i_1} \wedge \dots \wedge \widehat{dx^{i_\ell}} \wedge \dots \wedge dx^{i_p}\,. \qquad (2.1.22)
\end{aligned}
$$

Analogously

$$
\begin{aligned}
*_e\omega &= \omega_{i_1 \dots i_p} dx^{j_1} \wedge \dots \wedge dx^{j_{d-p}} \qquad (2.1.23)\\
d *_e \omega &= \sum_{\ell=1}^{p} \frac{\partial \omega_{i_1 \dots i_p}}{\partial x^{i_\ell}} dx^{i_\ell} \wedge dx^{j_1} \wedge \dots \wedge dx^{i_{d-p}} \qquad (2.1.24)\\
*_e d *_e \omega &= \sum_{\ell=1}^{p} (-1)^{p(d-p)+d-p+\ell-1} \frac{\partial \omega_{i_1 \dots i_p}}{\partial x^{i_\ell}} dx^{i_1} \wedge \dots \wedge \widehat{dx^{i_\ell}} \wedge \dots \wedge dx^{i_p} \qquad (2.1.25)\\
d *_e d *_e \omega &= \sum_{\ell=1}^{p} (-1)^{p(d-p)+d-p+\ell-1} \frac{\partial^2 \omega_{i_1 \dots i_p}}{(\partial x^{i_\ell})^2} dx^{i_\ell} \wedge dx^{i_1} \wedge \dots \wedge \widehat{dx^{i_\ell}} \wedge \dots \wedge dx^{i_p} \\
&\quad + \sum_{\ell=1}^{p} \sum_{k=1}^{d-p} (-1)^{p(d-p)+d-p+\ell-1} \frac{\partial^2 \omega_{i_1 \dots i_p}}{\partial x^{i_\ell} \partial x^{j_k}} dx^{j_k} \wedge dx^{i_1} \wedge \dots \wedge \widehat{dx^{i_\ell}} \wedge \dots \wedge dx^{i_p}, \\
&\qquad (2.1.26)
\end{aligned}
$$

hence with (2.1.25)

$$
\begin{aligned}
dd^*\omega = &\sum_{\ell=1}^{p}(-1)\frac{\partial^2\omega_{i_1\dots i_p}}{(\partial x^{i_\ell})^2}dx^{i_1}\wedge\dots\wedge dx^{i_p}\\
&+\sum_{\ell=1}^{p}\sum_{k=1}^{d-p}(-1)^{\ell}\frac{\partial^2\omega_{i_1\dots i_p}}{\partial x^{i_\ell}\partial x^{j_k}}dx^{j_k}\wedge dx^{i_1}\wedge\dots\wedge\widehat{dx^{i_\ell}}\wedge\dots\wedge dx^{i_p}. \qquad (2.1.27)
\end{aligned}
$$

(2.1.22) and (2.1.27) yield

$$
\Delta_e\omega = d^*d\omega + dd^*\omega = (-1)\sum_{m=1}^{d}\frac{\partial^2\omega_{i_1\dots i_p}}{(\partial x^m)^2}dx^{i_1}\wedge\dots\wedge dx^{i_p}\,. \qquad (2.1.28)
$$

Some more formulae:

We write

$$
\eta := \sqrt{g}dx^1\wedge\dots\wedge dx^d =: \eta_{i_1\dots i_d}dx^{i_1}\wedge\dots\wedge dx^{i_d}\,. \qquad (2.1.29)
$$

For $\beta = \beta_{j_1\dots j_p}dx^{j_1}\wedge\dots\wedge dx^{j_p}$

$$
\beta^{i_1\dots i_p} := g^{i_1j_1}g^{i_2j_2}\dots g^{i_pj_p}\beta_{j_1\dots j_p}. \qquad (2.1.30)
$$

With these conventions, for $\alpha = \alpha_{i_1\dots i_p}dx^{i_1}\wedge\dots\wedge dx^{i_p}$

$$
(*\alpha)_{i_{p+1}\dots i_d} = \frac{1}{p!}\eta_{i_1\dots i_p}\alpha^{i_1\dots i_p} \qquad (2.1.31)
$$

and

$$
(d^*\alpha)_{i_1\dots i_{p-1}} = -g^{k\ell}\big(\frac{\partial\alpha_{ki_1\dots i_{p-1}}}{\partial x^\ell} - \Gamma^j_{k\ell}\alpha_{ji_1\dots i_{p-1}}\big). \qquad (2.1.32)
$$

Further

$$
\begin{aligned}
(\alpha,\beta) &= \alpha_{i_1\dots i_p}\beta^{i_1\dots i_p} && (2.1.33)\\
(d\alpha,d\beta) &= \frac{\partial\alpha_{i_1\dots i_p}}{\partial x^k}\frac{\partial\beta_{j_1\dots j_p}}{\partial x^\ell}g^{k\ell}g^{i_1j_1}\dots g^{i_pj_p} && (2.1.34)\\
(d^*\alpha,d^*\beta) &= \Bigg(g^{k\ell}\Big(\frac{\partial\alpha_{ki_1\dots i_{p-1}}}{\partial x^\ell} - \Gamma^j_{k\ell}\alpha_{ji_1\dots i_{p-1}}\Big)e_{i_1}\wedge\dots\wedge e_{i_{p-1}},\\
&\qquad g^{mn}\Big(\frac{\partial\beta_{mj_1\dots j_{p-1}}}{\partial x^n} - \Gamma^r_{mn}\beta_{rj_1\dots j_{p-1}}\Big)e_{j_1}\wedge\dots\wedge e_{j_{p-1}}\Bigg) && (2.1.35)\\
&= \frac{\partial\alpha_{ki_1\dots i_{p-1}}}{\partial x^\ell}\frac{\partial\beta_{mj_1\dots j_{p-1}}}{\partial x^n}g^{k\ell}g^{mn}g^{i_1j_1}\dots g^{i_{p-1}j_{p-1}}\\
&\quad - \frac{\partial\alpha_{ki_1\dots i_{p-1}}}{\partial x^\ell}\Gamma^i_{mn}\beta_{ij_1\dots j_{p-1}}g^{k\ell}\dots g^{i_{p-1}j_{p-1}}\\
&\quad - \frac{\partial\beta_{mj_1\dots j_{p-1}}}{\partial x^n}\Gamma^j_{mn}\alpha_{ji_1\dots i_{p-1}}g^{k\ell}g^{mn}g^{i_1j_1}\dots g^{i_{p-1}j_{p-1}}.
\end{aligned}
$$

Formula (2.1.31) is clear. (2.1.32) may be verified by a straightforward, but somewhat lengthy computation. We shall see a different proof in §3.3 as a consequence of Lemma 3.3.4. The remaining formulae then are clear again.

2.2 Representing Cohomology Classes by Harmonic Forms

We first recall the definition of the de Rham cohomology groups. Let M be a differentiable manifold. The operator $d : \Omega^p(M) \to \Omega^{p+1}(M)$ satisfies (Theorem 1.8.5)

$$d \circ d = 0 \quad (d \circ d : \Omega^p(M) \to \Omega^{p+2}(M)). \tag{2.2.1}$$

$\alpha \in \Omega^p(M)$ is called *closed* if $d\alpha = 0$, *exact,* if there exists $\eta \in \Omega^{p-1}(M)$ with $d\eta = \alpha$. Because of (2.2.1), exact forms are always closed. Two closed forms $\alpha, \beta \in \Omega^p(M)$ are called *cohomologous* if $\alpha - \beta$ is exact. This property determines an equivalence relation on the space of closed forms in $\Omega^p(M)$, and the set of equivalence classes is a vector space over $\mathbb{R}$, called the p-th *de Rham cohomology group* and denoted by

$$H^p_{dR}(M, \mathbb{R}).$$

Usually, however, we shall simply write

$$H^p(M).$$

In this Paragraph, we want to show the following fundamental result:

Theorem 2.2.1 (**Hodge**). *Let M be a compact Riemannian manifold. Then every cohomology class in $H^p(M)$ $(0 \leq p \leq d = \dim M)$ contains precisely one harmonic form.*

Here, we shall demonstrate the Hodge theorem by a variational method. An alternative proof, by the heat flow method, as well as some important extensions, will be given in 2.4 below.

Proof. Uniqueness is easy: Let $\omega_1, \omega_2 \in \Omega^p(M)$ be cohomologous and both harmonic.

Then either $p = 0$ (in which case $\omega_1 = \omega_2$ anyway) or

$$\begin{aligned}(\omega_1 - \omega_2, \omega_1 - \omega_2) &= (\omega_1 - \omega_2, d\eta) \\ &\qquad \text{for some}\, \eta \in \Omega^{p-1}(M), \text{ since} \\ &\qquad \omega_1 \text{ and } \omega_2 \text{ are cohomologous} \\ &= (d^*(\omega_1 - \omega_2), \eta) \\ &= 0, \text{ since}\, \omega_1 \text{ and}\, \omega_2 \text{ are harmonic,} \\ &\qquad \text{hence satisfy } d^*\omega_1 = 0 = d^*\omega_2\,.\end{aligned}$$

Since $(\cdot, \cdot)$ is positive definite, we conclude $\omega_1 = \omega_2$, hence uniqueness. $\square$

For the proof of existence, which is much harder, we shall use Dirichlet's principle.

Let ω_0 be a (closed) differential form, representing the given cohomology class in $H^p(M)$.

All forms cohomologous to ω_0 then are of the form

$$\omega = \omega_0 + d\alpha \quad (\alpha \in \Omega^{p-1}(M)).$$

We now minimize the L^2-norm

$$D(\omega) := (\omega, \omega)$$

in the class of all such forms.

The essential step consists in showing that the infimum is achieved by a smooth form η. Such an η then has to satisfy the Euler-Lagrange equations for D, i.e.

$$\begin{aligned} 0 &= \frac{d}{dt}(\eta + td\beta, \eta + td\beta)_{|t=0} \quad \text{for all}\, \beta \in \Omega^{p-1}(M) \\ &= 2(\eta, d\beta). \end{aligned} \tag{2.2.2}$$

This implies $\delta\eta = 0$. Since $d\eta = 0$ anyway, η is harmonic.

In order to make Dirichlet's principle precise, we shall need some results and constructions from the calculus of variations. Some of them will be merely sketched (see §A.1, A.2 of the Appendices), and for details, we refer to our textbook [143]. First of all, we have to work with the space of L^2-forms instead of the one of C^∞-forms, since we want to minimize the L^2-norm and therefore certainly need a space that is complete w.r.t. L^2-convergence. For technical purposes, we shall also need Sobolev spaces which we now want to define in the present context (see also §A.1).

On $\Omega^p(M)$, we introduce a new scalar product

$$((\omega, \omega)) := (d\omega, d\omega) + (\delta\omega, \delta\omega) + (\omega, \omega) \tag{2.2.3}$$

and put

$$\|\omega\|_{H^{1,2}(M)} := ((\omega, \omega))^{\frac{1}{2}}. \tag{2.2.4}$$

(This norm is to be distinguished from the L^2-norm of (2.1.10).) We complete the space $\Omega^p(M)$ of smooth p-forms w.r.t. the $\|\cdot\|_{H^{1,2}(M)}$-norm. The resulting Hilbert space will be denoted by $H^{1,2}_p(M)$ or simply by $H^{1,2}(M)$, if the index p is clear from the context.

Let now $V \subset \mathbb{R}^d$ be open. For a smooth map $f : V \to \mathbb{R}^n$, the Euclidean Sobolev norm is given by

$$\|f\|_{H^{1,2}_{\text{eucl.}}(V)} := \left(\int_V f \cdot f + \int_V \frac{\partial f}{\partial x^i} \cdot \frac{\partial f}{\partial x^i} \right)^{\frac{1}{2}},$$

the dot $\cdot$ denoting the Euclidean scalar product.

With the help of charts for M and bundle charts for $\Lambda^p(M)$ for every $x_0 \in M$, there exist an open neighborhood U and a diffeomorphism

$$\varphi : \Lambda^p(M)_{|U} \to V \times \mathbb{R}^n$$

where V is open in $\mathbb{R}^d$, $n = \binom{d}{p}$ is the dimension of the fibers of $\Lambda^p(M)$, and the fiber over $x \in U$ is mapped to a fiber $\{\pi(\varphi(x))\} \times \mathbb{R}^n$, where $\pi : V \times \mathbb{R}^n \to V$ is the projection onto the first factor.

Lemma 2.2.1. *On any $U' \Subset U$, the norms*

$$\|\omega\|_{H^{1,2}(U')} \quad \text{and} \quad \|\varphi(\omega)\|_{H^{1,2}_{\text{eucl.}}(V')}$$

(with $V' := \pi(\varphi(U'))$) are equivalent.

Proof. As long as we restrict ourselves to relatively compact subsets of U, all coordinate changes lead to equivalent norms. Furthermore, by a covering argument, it suffices to find for every x in the closure of U' a neighborhood U'' on which the claimed equivalence of norms holds.

After these remarks, we may assume that first of all $\pi \circ \varphi$ is the map onto normal coordinates with center x_0, and that secondly for the metric in our neighborhood of x_0, we have

$$|g_{ij}(x) - \delta_{ij}| < \varepsilon \text{ and } |\Gamma^i_{jk}(x)| < \varepsilon \text{ for } i, j, k = 1, \ldots, d. \tag{2.2.5}$$

The formulae (2.1.33) - (2.1.35) then imply that the claim holds for sufficiently small $\varepsilon > 0$, i.e. for a sufficiently small neighborhood of x_0. Since $\bar{U}' \subset U$ is compact by assumption, the claim for U' follows by a covering argument. $\square$

Lemma 2.2.1 implies that the Sobolev spaces defined by the norms $\|\cdot\|_{H^{1,2}(M)}$ and $\|\cdot\|_{H^{1,2}_{\text{eucl.}}}$ coincide. Hence all results for Sobolev spaces in the Euclidean setting may be carried over to the Riemannian situation. In particular, we have Rellich's theorem (cf. Theorem A.1.8):

Lemma 2.2.2. *Let $(\omega_n)_{n\in\mathbb{N}} \subset H^{1,2}_p(M)$ be bounded, i.e.*

$$\|\omega_n\|_{H^{1,2}(M)} \le K.$$

Then a subsequence of (ω_n) converges w.r.t. the L^2-norm

$$\|\omega\|_{L^2(M)} := (\omega, \omega)^{\frac{1}{2}}$$

to some $\omega \in H^{1,2}_p(M)$. $\square$

Corollary 2.2.1. *There exists a constant c, depending only on the Riemannian metric of M, with the property that for all closed forms β that are orthogonal to the kernel of d^*,*

$$(\beta, \beta) \le c(d^*\beta, d^*\beta). \tag{2.2.6}$$

Proof. Otherwise, there would exist a sequence of closed forms β_n orthogonal to the kernel of d^*, with

$$(\beta_n, \beta_n) \geq n(d^*\beta_n, d^*\beta_n). \tag{2.2.7}$$

We put

$$\lambda_n := (\beta_n, \beta_n)^{-\frac{1}{2}}.$$

Then

$$1 = (\lambda_n\beta_n, \lambda_n\beta_n) \geq n(d^*(\lambda_n\beta_n), d^*(\lambda_n\beta_n)). \tag{2.2.8}$$

Since $d\beta_n = 0$, we have

$$\|\lambda_n\beta_n\|_{H^{1,2}} \leq 1 + \frac{1}{n}.$$

By Lemma 2.2.2, after selection of a subsequence, $\lambda_n\beta_n$ converges in L^2 to some form ψ. By (2.2.8), $d^*(\lambda_n\beta_n)$ converges to 0 in L^2. Hence $d^*\psi = 0$; this is seen as follows:

For all φ

$$\begin{aligned} 0 &= \lim_{n\to\infty}(d^*(\lambda_n\beta_n), \varphi) = \lim(\lambda_n\beta_n, d\varphi) \\ &= (\psi, d\varphi) = (d^*\psi, \varphi) \, \text{and hence} \, d^*\psi = 0. \end{aligned}$$

(With the same argument, $d\beta_n = 0$ for all n implies $d\psi = 0$.)

Now, since $d^*\psi = 0$ and β_n is orthogonal to the kernel of d^*,

$$(\psi, \lambda_n\beta_n) = 0. \tag{2.2.9}$$

On the other hand, $(\lambda_n\beta_n, \lambda_n\beta_n) = 1$ and the L^2-convergence of $\lambda_n\beta_n$ to ψ imply

$$\lim_{n\to\infty}(\psi, \lambda_n\beta_n) = 1.$$

This is a contradiction, and (2.2.7) is impossible. □

We can now complete the *proof of Theorem 2.2.1:*

Let $(\omega_n)_{n\in\mathbb{N}}$ be a minimizing sequence for $D(\omega)$ in the given cohomology class, i.e.

$$\begin{aligned} \omega_n &= \omega_0 + d\alpha_n \\ D(\omega_n) &\to \inf_{\omega=\omega_0+d\alpha} D(\omega) =: \kappa. \end{aligned} \tag{2.2.10}$$

By (2.2.10), w.l.o.g.

$$(\omega_n, \omega_n) = D(\omega_n) \leq \kappa + 1. \tag{2.2.11}$$

As with Dirichlet's principle in $\mathbb{R}^d$, ω_n converges weakly to some ω, after selection of a subsequence.

We have

$$(\omega - \omega_0, \varphi) = 0 \, \text{for all} \, \varphi \in \Omega^p(M) \, \text{with} \, d^*\varphi = 0, \tag{2.2.12}$$

because

$$(\omega_n - \omega_0, \varphi) = (d\alpha_n, \varphi) = (\alpha_n, d^*\varphi) = 0 \text{ for all such } \varphi.$$

(2.2.12) means that $\omega - \omega_0$ is weakly exact.

We want to study this condition more closely and put

$$\eta := \omega - \omega_0.$$

We define a linear functional on $d^*(\Omega^p(M))$ by

$$\ell(\delta\varphi) := (\eta, \varphi), \tag{2.2.13}$$

ℓ is well defined; namely if $d^*\varphi_1 = d^*\varphi_2$, then

$$(\eta, \varphi_1 - \varphi_2) = 0 \text{ by (2.2.12).}$$

For $\varphi \in \Omega^p(M)$ let $\pi(\varphi)$ be the orthogonal projection onto the kernel of d^*, and $\psi := \varphi - \pi(\varphi)$; in particular $d^*\psi = d^*\varphi$.

Then

$$\ell(d^*\varphi) = \ell(d^*\psi) = (\eta, \psi). \tag{2.2.14}$$

Since ψ is orthogonal to the kernel of δ, by Corollary 2.2.1,

$$\|\psi\|_{L^2} \le c\|d^*\psi\|_{L^2} = c\|d^*\varphi\|_{L^2}. \tag{2.2.15}$$

(2.2.14) and (2.2.15) imply

$$|\ell(d^*\varphi)| \le c\|\eta\|_{L^2}\|d^*\varphi\|_{L^2}.$$

Therefore, the function ℓ on $d^*(\Omega^p(M))$ is bounded and can be extended to the L^2-closure of $d^*(\Omega^p(M))$. By the Riesz representation theorem, any bounded linear functional on a Hilbert space is representable as the scalar product with an element of the space itself. Consequently, there exists α with

$$(\alpha, d^*\varphi) = (\eta, \varphi) \tag{2.2.16}$$

for all $\varphi \in \Omega^p(M)$.

Thus, we have weakly

$$d\alpha = \eta. \tag{2.2.17}$$

Therefore, $\omega = \omega_0 + \eta$ is contained in the closure of the considered class. Instead of minimizing among the ω cohomologous to ω_0, we could have minimized as well in the closure of this class, i.e., in the space of all ω for which there exists some α with

$$(\alpha, d^*\varphi) = (\omega - \omega_0, \varphi) \text{ for all } \varphi \in \Omega^p(M).$$

Then ω, as weak limit of a minimizing sequence, is contained in this class. Namely, suppose $\omega_n = \omega_0 + d\alpha_n$ weakly, i.e.

$$\ell_n(d^*\varphi) := (\alpha_n, d^*\varphi) = (\omega_n - \omega_0, \varphi)\ \forall\, \varphi \in \Omega^p(M).$$

By the same estimate as above, the linear functionals ℓ_n converge to some functional ℓ, again represented by some α. Since D also is weakly lower semicontinuous w.r.t. weak convergence, it follows that

$$\kappa \leq D(\omega) \leq \lim_{n\to\infty} \inf D(\omega_n) = \kappa,$$

hence

$$D(\omega) = \kappa.$$

Furthermore, by (2.2.2),

$$0 = (\omega, d\beta) \text{ for all } \beta \in \Omega^{p-1}(M). \tag{2.2.18}$$

In this sense, ω is weakly harmonic.

We still need the regularity theorem implying that solutions of (2.2.18) are smooth. This can be carried out as in the Euclidean case. If one would be allowed to insert $\beta = d^*\omega$ in (2.2.18) and integrate by parts, it would follow that

$$0 = (d^*\omega, d^*\omega),$$

i.e. $d^*\omega = 0$.

Iteratively, also higher derivatives would vanish, and the Sobolev embedding theorem would imply regularity. However, we cannot yet insert $\beta = d^*\omega$, since we do not know yet whether $dd^*\omega$ exists. This difficulty, however, may be overcome as usual by replacing derivatives by difference quotients (See §A.2 of the Appendix.). In this manner, one obtains regularity and completes the proof. □

Corollary 2.2.2. *Let M be a compact, oriented, differentiable manifold. Then all cohomology groups $H^p_{dR}(M, \mathbb{R})$ $(0 \leq p \leq d := \dim M)$ are finite dimensional.*

Proof. By Theorem 1.4.1, a Riemannian metric may be introduced on M. By Theorem 2.2.1 any cohomology class may be represented by a form which is harmonic w.r.t. this metric. We now assume that $H^p(M)$ is infinite dimensional. Then, there exists an orthonormal sequence of harmonic forms $(\eta_n)_{n\in\mathbb{N}} \subset H^p(M)$, i.e.

$$(\eta_n, \eta_m) = \delta_{nm} \text{ for } n, m \in \mathbb{N}. \tag{2.2.19}$$

Since the η_n are harmonic, $d^*\eta_n = 0$, and $d\eta_n = 0$. By Rellich's theorem (Lemma 2.2.2), after selection of a subsequence, (η_n) converges in L^2 to some η. This, however, is not compatible with (2.2.19), because (2.2.19) implies

$$\|\eta_n - \eta_m\|_{L^2} \geq 1 \text{ for } n \neq m,$$

so that (η_n) cannot be a Cauchy sequence in L^2.

This contradiction proves the finite dimensionality. □

Let now M be a compact, oriented, differentiable manifold of dimension d. We define a bilinear map

$$H^p_{dR}(M, \mathbb{R}) \times H^{d-p}_{dR}(M, \mathbb{R}) \to \mathbb{R}$$

by

$$(\omega, \eta) \mapsto \int_M \omega \wedge \eta \tag{2.2.20}$$

for representatives ω, η of the cohomology classes considered. It remains to show that (2.2.20) depends only on the cohomology classes of ω and η, in order that the map is indeed defined on the cohomology groups. If, however, ω' and ω are cohomologous, there exists a $(p-1)$ form α with $\omega' = \omega + d\alpha$, and

$$\begin{aligned} \int_M \omega' \wedge \eta &= \int_M (\omega + d\alpha) \wedge \eta \\ &= \int_M \omega \wedge \eta + \int_M d(\alpha \wedge \eta) \quad \text{since } \eta \text{ is closed} \\ &= \int_M \omega \wedge \eta \quad \text{by Stokes' theorem.} \end{aligned}$$

Therefore, (2.2.20) indeed depends only on the cohomology class of ω, and likewise only on the cohomology class of η.

Let us now recall a simple result of linear algebra. Let V and W be finite dimensional real vector spaces, and let

$$(\cdot, \cdot) : V \times W \to \mathbb{R}$$

be bilinear and nondegenerate in the sense that for any $v \in V, v \neq 0$, there exists $w \in W$ with $(v, w) \neq 0$, and conversely. Then V can be identified with the dual space W^* of W, and W may be identified with V^*. Namely,

$$\begin{aligned} i_1 &: V \to W^* \qquad \text{with } i_1(v)(w) := (v, w), \\ i_2 &: W \to V^* \qquad \text{with } i_2(w)(v) := (v, w), \end{aligned}$$

are two injective linear maps. Then V and W must be of the same dimension, and i_1 and i_2 are isomorphisms.

Theorem 2.2.2. *Let M be a compact, oriented, differentiable manifold of dimension d. The bilinear form (2.2.20) is nondegenerate, and hence $H^p_{dR}(M, \mathbb{R})$ is isomorphic to $(H^{d-p}_{dR}(M, \mathbb{R}))^*$.*

Proof. For each nontrivial cohomology class in $H^p(M)$, represented by some ω (i.e. $d\omega = 0$, but not $\omega = d\alpha$ for any $(p-1)$-form α), we have to find some cohomology class in $H^{d-p}(M)$ represented by some η, such that

$$\int_M \omega \wedge \eta \neq 0.$$

For this purpose, we introduce a Riemannian metric on M which is possible by Theorem 1.4.1. By Theorem 2.2.1, we may assume that ω is harmonic (w.r.t. this metric). By Lemma 2.1.6

$$\Delta * \omega = *\Delta\omega,$$

and therefore, $*\omega$ is harmonic together with ω. Now

$$\int_M \omega \wedge *\omega = (\omega, \omega) \neq 0, \text{ since } \ \omega \ \text{ does not vanish identically.}$$

Therefore, $*\omega$ represents a cohomology class in $H^{d-p}(M)$ with the desired property. Thus the bilinear form is nondegenerate, and the claim follows. □

Definition 2.2.1. The p-th *homology group* $H_p(M,\mathbb{R})$ of a compact, differentiable manifold M is defined to be $(H^p_{dR}(M,\mathbb{R}))^*$. The p-th *Betti number* of M is $b_p(M) := \dim H^p(M,\mathbb{R})$.

With this definition, Theorem 2.2.2 becomes

$$H_p(M,\mathbb{R}) \cong H^{d-p}_{dR}(M,\mathbb{R}). \tag{2.2.21}$$

This statement is called *Poincaré duality.*

Corollary 2.2.3. *Let M be a compact, oriented, differentiable manifold of dimension d. Then*

$$H^d_{dR}(M,\mathbb{R}) \cong \mathbb{R}. \tag{2.2.22}$$

and

$$b_p(M) = b_{d-p}(M) \quad \text{for} \quad 0 \le p \le d. \tag{2.2.23}$$

Proof. $H^0_{dR}(M,\mathbb{R}) \cong \mathbb{R}$. This follows e.g. from Corollary 2.1.2 and Theorem 2.2.1, but can also be seen in an elementary fashion.
Theorem 2.2.2 then implies (2.2.22), as well as (2.2.23). □

As an example, let us consider an n-dimensional torus T^n. As shown in §1.4, it can be equipped with a Euclidean metric for which the covering $\pi : \mathbb{R}^n \to T^n$ is a local isometry.

By (2.1.28), we have for the Laplace operator of the Euclidean metric

$$\Delta(\omega_{i_1,\dots,i_p} dx^{i_1} \wedge \ldots \wedge dx^{i_p}) = (-1) \sum_{m=1}^{n} \frac{\partial^2 \omega_{i_1 \dots i_p}}{(\partial x^m)^2} dx^{i_1} \wedge \ldots \wedge dx^{i_p}$$

($x^1, \dots, x^n$ Euclidean coordinates of $\mathbb{R}^n$.) Thus, a p-form is harmonic if and only if all coefficients w.r.t. the basis $dx^{i_1} \wedge \ldots \wedge dx^{i_p}$ are harmonic. Since T^n is compact, by Corollary 2.1.2, they then have to be constant. Consequently

$$b_p(T^n) = \dim H^p(T^n) = \dim \Lambda^p(\mathbb{R}^n) = \binom{n}{p} \qquad (0 \le p \le n).$$

Perspectives. The results of this Paragraph were found in the 1940s by Weyl, Hodge, de Rham and Kodaira.

2.3 Generalizations

The constructions of this chapter may easily be generalized. Here, we only want to indicate some such generalizations.

Let E and F be vector bundles over the compact, oriented, differentiable manifold M. Let $\Gamma(E)$ and $\Gamma(F)$ be the spaces of differentiable sections. Sobolev spaces of sections can be defined with the help of bundle charts: Let (f, U) be a bundle chart for E, f then identifies $E_{|U}$ with $U \times \mathbb{R}^n$. A section s of E is then contained in the Sobolev space $H^{k,p}(E)$ if for any such bundle chart and any $U' \Subset U$, we have $p_2 \circ f \circ s_{|U'} \in H^{k,p}(U', \mathbb{R}^n)$, where $p_2 : U' \times \mathbb{R}^n \to \mathbb{R}^n$ is the projection onto the second factor.

A linear map $L : \Gamma(E) \to \Gamma(F)$ is called (linear) differential operator of order ℓ from E to F if in any bundle chart, L defines such an operator. For the Laplace operator, of course $E = F = \Lambda^p(T^*M)$, $\ell = 2$.

In a bundle chart, we write L as

$$L = P_\ell(D) + \ldots + P_0(D),$$

where each $P_j(D)$ is an $(m \times n)$-matrix ($m, n =$ fiber dimensions of E and F, resp.), whose components are differential operators of the form

$$\sum_{|\alpha|=j} a_\alpha(x) D^\alpha$$

where α is a multi index, and D^α is a homogeneous differential operator of degree $|\alpha| = j$. Let us assume that the $a_\alpha(x)$ are differentiable.

For $\xi = (\xi^1, \ldots, \xi^m) \in \mathbb{R}^m$, let $P_j(\xi)$ be the matrix obtained for $P_j(D)$ by replacing D^α by ξ^α.
$P_j(\xi)$ thus has components

$$\sum_{|\alpha|=j} a_\alpha(x) \xi^\alpha.$$

L is called *elliptic* at the point x, if $P_\ell(\xi)$ ($\ell =$ degree of L) is nonsingular at x for all $\xi \in \mathbb{R}^m \backslash \{0\}$. Note that in this case necessarily $n = m$.

L is called *elliptic* if it is elliptic at every point. Let now $\langle \cdot, \cdot \rangle_E$ and $\langle \cdot, \cdot \rangle_F$ be bundle metrics on E and F, resp. (those always exist by Theorem 1.8.3), let M carry a Riemannian metric (existing by Theorem 1.4.1) and an orientation. Integrating the bundle metrics, for example

$$(\cdot, \cdot)_E := \int_M \langle \cdot, \cdot \rangle_E d\mathrm{Vol}_g \quad (d\mathrm{Vol}_g = \sqrt{\det(g_{ij})} dx^1 \wedge \ldots \wedge dx^d),$$

we obtain L^2-metrics on $\Gamma(E)$ and $\Gamma(F)$. Let L^* be the operator formally adjoint to L, i.e.

$$(Lv, w)_F = (v, L^*w)_E \quad \text{for } v \in \Gamma(E), w \in \Gamma(F).$$

L is elliptic if L^* is.

The importance of the ellipticity condition rests on the fact that solutions of elliptic differential equations are regular, and the space of solutions has finite dimension.

Here, however, this shall not be pursued any further.

2.4 The Heat Flow and Harmonic Forms

In this section, we shall present an alternative proof of Theorem 2.2.1. This proof will procede by solving a parabolic equation, the so-called heat flow. The idea is to let the objects involved, here p-forms, depend not only on the position x in the manifold M, but also on another variable, the "time" $t \in [0, \infty)$, and to replace the elliptic equation that one wishes to solve by a parabolic equation that one can solve for given starting values at time $t = 0$. In our case of differential forms, this heat equation is

$$\frac{\partial \beta(x,t)}{\partial t} + \Delta \beta(x,t) = 0 \tag{2.4.1}$$

$$\beta(x,0) = \beta_0(x) \tag{2.4.2}$$

where β_0 is a p-form in the cohomology class that we wish to study.

The strategy then consists in showing that (2.4.1) can be uniquely solved for all positive t (this is called global or long time existence) and that, as $t \to \infty$, the solution $\beta(x,t)$ converges to a harmonic p-form in the same cohomology class.

(2.4.1) is a linear parabolic differential equation (or more precisely, a system of linear differential equations since the dimension of the fibers Λ^p is larger than 1 except for trivial cases). Therefore, the global existence and existence of solutions follows from the general theory of linear parabolic differential equations. Since we consider this equation as a prototype of other, typically nonlinear, parabolic differential equations arising in geometric analysis, we shall only use the short time existence here (which also holds for nonlinear equations by linearization) and deduce the long time existence from differential inequalities for the geometric objects involved.

The short time existence is contained in

Lemma 2.4.1. *Let $\beta_0 \in \Omega^p$ be of class $C^{2,\alpha}$ for some $0 < \alpha < 1$. Then, for some $0 < \epsilon$, (2.4.1) has a solution $\beta(x,t)$ for $0 \le t < \epsilon$, and this solution is also of class $C^{2,\alpha}$.*

In order to procede to the global existence, we shall consider the L^2-norm

$$\|\beta(\cdot,t)\|^2 = \int_M \beta(x,t) \wedge *\beta(x,t) \tag{2.4.3}$$

and the energy

$$E(\beta(\cdot,t)) := \frac{1}{2}\|d\beta(\cdot,t)\| + \frac{1}{2}\|d^*\beta(\cdot,t)\|. \tag{2.4.4}$$

(Note that $(\|\beta(\cdot,t)\|^2 + 2E(\beta(\cdot,t)))^{1/2}$ is the Sobolev norm of $\beta(\cdot,t)$ as introduced in (2.2.4).)

Lemma 2.4.2.

$$\frac{d}{dt}\|\beta(\cdot,t)\|^2 \le 0 \tag{2.4.5}$$

$$\frac{d^2}{dt^2}\|\beta(\cdot,t)\|^2 \ge 0 \tag{2.4.6}$$

$$\frac{d}{dt}E(\beta(\cdot,t)) \le 0. \tag{2.4.7}$$

Proof.

$$\begin{aligned}\frac{d}{dt}\|\beta(\cdot,t)\|^2 &= 2(\tfrac{\partial}{\partial t}\beta(\cdot,t),\beta(\cdot,t))\\ &= -2(\Delta\beta(\cdot,t),\beta(\cdot,t))\\ &= -2(d\beta(\cdot,t),d\beta(\cdot,t)) - 2(d^*\beta(\cdot,t),d^*\beta(\cdot,t))\\ &= -4E(\beta(\cdot,t)) \qquad (2.4.8)\\ &\le 0\end{aligned}$$

which shows (2.4.5). Next

$$\begin{aligned}\frac{d}{dt}E(\beta(\cdot,t)) &= (d\tfrac{\partial}{\partial t}\beta(\cdot,t),d\beta(\cdot,t)) + (d^*\tfrac{\partial}{\partial t}\beta(\cdot,t),d^*\beta(\cdot,t))\\ &= (\tfrac{\partial}{\partial t}\beta(\cdot,t),\Delta\beta(\cdot,t))\\ &= -(\tfrac{\partial}{\partial t}\beta(\cdot,t),\tfrac{\partial}{\partial t}\beta(\cdot,t))\\ &\le 0\end{aligned}$$

which shows (2.4.7). (2.4.6) follows from this and (2.4.8). □

In particular, when $\beta(x,0) \equiv 0$, then, by (2.4.5), $\beta(x,t) \equiv 0$ for all t for which the solution exists. From this, we deduce

Corollary 2.4.1. *Solutions of (2.4.1) are unique*
(if $\beta_1(x,t)$ and $\beta_2(x,t)$ are solutions of (2.4.1) for $0 \le t \le T$ with the same initial values, i.e., $\beta_1(x,0) = \beta_2(x,0)$, then they also coincide for $0 \le t \le T$)
and satisfy a semigroup property
(if $\beta(\cdot,t)$ solves (2.4.1), then $\beta(\cdot,t+s) = \beta_s(\cdot,t)$ where $\beta_s(\cdot,t)$ is the solution of (2.4.1) with initial values $\beta_s(\cdot,0) = \beta(\cdot,s)$).

In fact, we have a more general stability result

Corollary 2.4.2. *For a family $\beta(x,t,s)$ of solutions of (2.4.1) that depends differentiably on the parameter $s \in \mathbb{R}$,*

$$\frac{d}{dt}\|\frac{\partial}{\partial s}\beta(\cdot,t,s)\|^2 \leq 0. \tag{2.4.9}$$

Proof. $\frac{\partial}{\partial s}\beta(x,t,s)$ also solves (2.4.1), and (2.4.9) therefore follows from (2.4.5). □

We now need some apriori estimates:

Lemma 2.4.3. *A solution $\beta(x,t)$ of (2.4.1) defined for $0 \leq t \leq T$ with initial values $\beta_0(x) \in L^2$ satisfies for $\tau \leq t \leq T$, for any $\tau > 0$, estimates of the form*

$$\|\beta(\cdot,t)\|_{C^{2,\alpha}(M)} + \|\frac{\partial}{\partial t}\beta(\cdot,t)\|_{C^\alpha(M)} \leq c_1 \tag{2.4.10}$$

with a constant c_1 depending only on $\|\beta_0\|_{L^2(M)}$, τ and the geometry of M (but not on the particular solution $\beta(x,t)$).

Remark. An important consequence of this lemma that we shall use repeatedly in the sequel is that from the estimates we can infer convergence results. In fact, the Arzela-Ascoli Theorem implies that any sequence (f_n) that is bounded in the Hölder space $C^\alpha(M)$ for some $0 < \alpha < 1$ contains a subsequence that converges in $C^{\alpha'}(M)$, for any $\alpha' < \alpha$. See [143] for details.
Proof. From (2.4.5),

$$\|\beta(\cdot,t)\|_{L^2(M)} \leq \|\beta_0\|_{L^2(M)} \tag{2.4.11}$$

See ... □

We can now deduce the global existence of solutions of (2.4.1):

Corollary 2.4.3. *Let $\beta_0 \in C^{2,\alpha}$ for some $0 < \alpha < 1$. Then the solution $\beta(x,t)$ of (2.4.1) with those initial values exists for all $t \geq 0$.*

Proof. By local existence (Lemma 2.4.1), the solution exists on some positive time interval $0 \leq t < \epsilon$. Whenever it exists on some interval $0 \leq t \leq T$, for $t \to T$, by Lemma 2.4.3, $\beta(x,t)$ converges to some form $beta(x,T)$ in $C^{2,\alpha'}$ for $0 < \alpha' < \alpha$. Applying the semigroup property (Corollary 2.4.1) and local existence (Lemma 2.4.1) again, the solution can be continued to some time interval beyond T, that is, it exists for $0 \leq t < T + \epsilon$. Thus, the existence interval is open and closed and nonempty and therefore consists of the entire positive real line. □

The final step in the program is the asymptotic behavior of solutions as $t \to \infty$. With this, we shall complete the proof of

Theorem 2.4.1 (Milgram-Rosenbloom). *Given a p-form $\beta_0(x)$ on M of class $C^{2,\alpha}$, for some $0<\alpha<1$, there exists a unique solution of*

$$\frac{\partial \beta(x,t)}{\partial t} + \Delta\beta(x,t) = 0 \text{ for all } 0 \le t < \infty \tag{2.4.12}$$

$$\text{with } \beta(x,0) = \beta_0(x) \tag{2.4.13}$$

As $t \to \infty$, $\beta(\cdot,t)$ converges in $C^{2,\alpha}$ to a harmonic form $H\beta$.
If β_0 is closed, i.e., $d\beta_0 = 0$, then all the forms $\beta(\cdot,t)$ are closed as well, $d\beta(\cdot,t) = 0$. Also, in this case, if ω is a coclosed $(d-p)$-form, i.e. $d^*\omega = 0$, then $\int_M \beta(x,t)\wedge\omega(x)$ does not depend on t, and we have $\int_M H\beta(x)\wedge\omega(x) = \int_M \beta_0(x)\wedge\omega(x)$.

This result obviously contains the Hodge Theorem 2.2.1 and provides an alternative proof of it.
Proof. Since $E(\beta(\cdot,t)) \ge 0$, (2.4.5) implies that there exists at least some sequence $t_n \to \infty$ for which

$$\|\frac{\partial}{\partial t}\beta(\cdot,t_n)\| \to 0. \tag{2.4.14}$$

The control of the higher norms of $\beta(\cdot,t_n)$ of Lemma 2.4.3 then implies that $\Delta\beta(\cdot,t_n) = -\frac{\partial}{\partial t}\beta(\cdot,t_n)$ converges to 0 in some Hölder space $C^{2,\alpha'}$, that is, $\beta(\cdot,t_n)$ converges in $C^{2,\alpha'}$ to a harmonic form $H\beta$. The difference

$$\beta_1(x,t) := \beta(x,t) - H\beta(x)$$

then also solves (2.4.12). Using (2.4.14) and (2.4.5) once more, we see that $\|\beta(\cdot,t) - H\beta(\cdot)\| \to 0$ as $t\to\infty$, and by Lemma 2.4.3, $\beta(x,t)$ converges to $H\beta(x)$ in $C^{2,\alpha'}$.
Uniqueness was already deduced in Corollary 2.4.1.
Since the exterior derivative d commutes with the Laplacian Δ as is clear from the definition of the latter and obviously also with $\frac{\partial}{\partial t}$, if $\beta(x,t)$ solves (2.4.12), then so does $d\beta(x,t)$. Thus, using e.g. (2.4.5) again, if $d\beta_0 = 0$, then also $d\beta(\cdot,t) = 0$. Finally, if also $d^*\omega = 0$, then

$$\begin{aligned}
&\frac{\partial}{\partial t}\int_M \beta(x,t)\wedge\omega(x) = -\int_M \Delta\beta(x,t)\wedge\omega(x) \\
&= -\int_M dd^*\beta(x,t)\wedge\omega(x) = -\int_M d^*\beta(x,t)\wedge d^*\omega(x) = 0.
\end{aligned}$$

□

The heat flow method can also conveniently deduce some refinements of this theorem. We observe

Lemma 2.4.4. *Under the assumptions of Theorem 2.4.1, the solution $\beta(x,t)$ of (2.4.1) converges exponentially towards the harmonic form $H\beta_0(x)$, that is,*

$$\|\beta(\cdot,t) - H\beta_0(\cdot)\| \le ce^{-\lambda t} \tag{2.4.15}$$

for some positive constants c, λ. Here, λ is independent of β.

Proof. Given $t > 0$, we seek β with $\|\beta\| = 1$ and $H\beta = 0$ for which for the solution $\beta(x,t)$ of (2.4.1) with initial values $\beta(x,0) = \beta(x)$,

$$\|\beta(\cdot,t)\|$$

is maximal. Since, by Lemma 2.4.3, the $C^{1,\alpha}$-norm of $\beta(\cdot,t)$ is bounded in terms of $\|\beta(\cdot,0)\|$, this maximum is attained. Let this maximal value be $b(t)$. Since $H\beta = 0$, (2.4.5) must be strictly negative. This implies $b(t) < 1$. The semigroup property of Corollary 2.4.1 then implies

$$b(nt) \leq b(t)^n \text{ for } n \in \mathbb{N},$$

from which

$$b(t) \leq e^{-\lambda t} \text{ for some } \lambda > 0.$$

Therefore, for general $\beta(x,0) \in L^2$, we obtain (2.4.15). □

We can then show

Corollary 2.4.4. *The equation*

$$\Delta\nu = \eta \tag{2.4.16}$$

for a p-form η of class L^2 is solvable iff

$$(\eta,\omega) = 0 \textit{ for all } \omega \textit{ with } \Delta\omega = 0. \tag{2.4.17}$$

This solution then is unique up to addition of a harmonic form.
Therefore, the space of p-forms of class L^2 admits the decomposition

$$\Omega^p_{L^2}(M) = \ker\Delta \bigoplus \operatorname{image}\Delta \tag{2.4.18}$$

(note that the first summand, the kernel of Δ, is finite dimensional).

Proof. We consider

$$\begin{aligned}\frac{\partial}{\partial t}\mu + \Delta\mu &= \gamma \\ \mu(\cdot,t) &= \mu_0.\end{aligned} \tag{2.4.19}$$

We put

$$T_t\mu_0 = \beta(\cdot,t)$$

for the solution of

$$\begin{aligned}\frac{\partial}{\partial t}\beta + \Delta\beta &= 0 \\ \beta(\cdot,t) &= \mu_0.\end{aligned} \tag{2.4.20}$$

We then have

$$\mu(x,t) = T_t\mu_0(x) + \int_0^t T_{t-s}\gamma(x)ds = T_t\mu_0(x) + \int_0^t T_s\gamma(x)ds \qquad (2.4.21)$$

as γ does not depend on t.
By (2.4.15), we have

$$\|T_s\gamma - H\gamma\| \le e^{-\lambda s}$$

whence

$$\|\mu - tH\gamma - T_t\mu_0\| \le \int_0^t e^{-\lambda s}ds.$$

We conclude that

$$\nu(x) := \lim_{t\to\infty}(\mu(x,t) - tH\gamma(x))$$

exists, in L^2 and then also in $C^{2,\alpha}$, by the estimates. Since $\Delta H\gamma = 0$, we have

$$(\frac{\partial}{\partial t} + \Delta)(\mu(x,t) - tH\gamma(x)) = \eta(x) - H\eta(x).$$

Therefore,

$$\Delta\nu = \eta - H\eta.$$

This implies the solvability of (2.4.16) under the condition (2.4.17) because $\eta - H\eta$ is the projection onto the L^2-orthogonal complement of the kernel of Δ. $\square$

Exercises for Chapter 2

1. Compute the Laplace operator of S^n on p-forms ($0 \le p \le n$) in the coordinates given in §1.1.

2. Let $\omega \in \Omega^1(S^2)$ be a 1-form on S^2. Suppose

$$\varphi^*\omega = \omega$$

for all $\varphi \in \mathrm{SO}(3)$. Show that $\omega \equiv 0$.

Formulate and prove a general result for invariant differential forms on S^n.

3. Give a detailed proof of the formula

$$*\Delta = \Delta * .$$

4. Let M be a two dimensional Riemannian manifold. Let the metric be given by $g_{ij}(x)dx^i \otimes dx^j$ in local coordinates (x^1, x^2). Compute the Laplace operator on 1-forms in these coordinates. Discuss the case where

$$g_{ij}(x) = \lambda^2(x)\delta_{ij}$$

with a positive function $\lambda^2(x)$.

5. Suppose that $\alpha \in H_p^{1,2}(M)$ satisfies

$$(d^*\alpha, d^*\varphi) + (d\alpha, d\varphi) = (\eta, \varphi) \quad \text{for all} \quad \varphi \in \Omega^p(M),$$

with some given $\eta \in \Omega^p(M)$. Show $\alpha \in \Omega^p(M)$, i.e. smoothness of α.

6. Compute a relation between the Laplace operators on functions on $\mathbb{R}^{n+1}$ and the one on $S^n \subset \mathbb{R}^{n+1}$.

7. Eigenvalues of the Laplace operator:

Let M be a compact oriented Riemannian manifold, and let Δ be the Laplace operator on $\Omega^p(M)$. $\lambda \in \mathbb{R}$ is called eigenvalue if there exists some $u \in \Omega^p(M), u \neq 0$, with

$$\Delta u = \lambda u.$$

Such a u is called eigenform or eigenvector corresponding to λ. The vector space spanned by the eigenforms for λ is denoted by V_λ and called eigenspace for λ.

Show:

a: All eigenvalues of Δ are nonnegative.

b: All eigenspaces are finite dimensional.

c: The eigenvalues have no finite accumulation point.

d: Eigenvectors for different eigenvalues are orthogonal.

The next results need a little more analysis (cf. e.g. [143])

e: There exist infinitely many eigenvalues

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \leq \ldots$$

f: All eigenvectors of Δ are smooth.

g: The eigenvectors of Δ constitute an L^2-orthonormal basis for the space of p-forms of class L^2.

8. Here is another long exercise:

Let M be a compact oriented Riemannian manifold with boundary $\partial M \neq \emptyset$. For $x \in \partial M, V \in T_x M$ is called tangential if it is contained in $T_x \partial M \subset T_x M$ and $W \in T_x M$ is called normal if

$$\langle V, W \rangle = 0 \quad \text{for all tangential} \quad V.$$

An arbitrary $Z \in T_x M$ can then be decomposed into a tangential and a normal component:

$$Z = Z_{\tan} + Z_{\text{nor}}.$$

Analogously, $\eta \in \Gamma^p(T^x, M)$ can be decomposed into

$$\eta = \eta_{\tan} + \eta_{\text{nor}}$$

where $\eta_{\tan}$ operates on tangential p-vectors and η_{nor} on normal ones. For p-forms ω on M, we may impose the so-called absolute boundary conditions

$$\begin{aligned} \omega_{\tan} &= 0, \\ (\delta\omega)_{\text{nor}} &= 0, \end{aligned} \qquad \text{on } \partial M,$$

or the relative boundary conditions

$$\begin{aligned} \omega_{\text{nor}} &= 0, \\ (d\omega)_{\text{nor}} &= 0, \end{aligned} \qquad \text{on } \partial M.$$

(These two boundary conditions are interchanged by the $*$-operator.)

Develop a Hodge theory under either set of boundary conditions.

Chapter 3

Parallel Transport, Connections, and Covariant Derivatives

3.1 Connections in Vector Bundles

Let X be a vector field on $\mathbb{R}^d$, V a vector at $x_0 \in \mathbb{R}^d$. We want to analyze how one takes the derivative of X at x_0 in the direction V. For this derivative, one forms

$$\lim_{t \to 0} \frac{X(x_0 + tV) - X(x_0)}{t}.$$

Thus, one first adds the vector tV to the point x_0. Next, one compares the vector $X(x_0 + tV)$ at the point $x_0 + tV$ and the vector $X(x_0)$ at x_0; more precisely, one subtracts the second vector from the first one. Division by t and taking the limit then are obvious steps.

A vector field on $\mathbb{R}^d$ is a section of the tangent bundle $T(\mathbb{R}^d)$. Thus, $X(x_0+tV)$ lies in $T_{x_0+tV}(\mathbb{R}^d)$, while $X(x_0)$ lies in $T_{x_0}(\mathbb{R}^d)$. The two vectors are contained in different spaces, and in order to subtract the second one from the first one, one needs to identify these spaces. In $\mathbb{R}^d$, this is easy. Namely, for each $x \in \mathbb{R}^d$, $T_x\mathbb{R}^d$ can be canonically identified with $T_0\mathbb{R}^d \cong \mathbb{R}^d$. For this, one uses Euclidean coordinates and identifies the tangent vector $\frac{\partial}{\partial x^i}$ at x with $\frac{\partial}{\partial x^i}$ at 0. This identification is even expressed by the notation. The reason why it is canonical is simply that the Euclidean coordinates of $\mathbb{R}^d$ can be obtained in a geometric manner. For this, let $c(t) = tx, t \in [0,1]$ the straight line joining 0 and x. For a vector X_1 at x, let X_t be the vector at $c(t)$ parallel to X_1; in particular, X_t has the same length as X_1 and forms the same angle with $\dot{c}$. X_0 then is the vector at 0 that gets identified with X_1. The advantage of the preceding geometric description lies in the fact that X_1 and X_0 are connected through a continuous geometric process. Again, this process in $\mathbb{R}^d$ has to be considered as canonical.

On a manifold, in general there is no canonical method anymore for identifying tangent spaces at different points, or, more generally fibers of a vector bundle at different points. For example, on a general manifold, we don't have canonical coordinates. Thus, we have to expect that a notion of derivative for sections of a vector bundle, for example for vector fields, has to depend on certain choices.

Definition 3.1.1. Let M be a differentiable manifold, E a vector bundle over M. A *covariant derivative*, or equivalently, a *(linear) connection* is a map

$$D : \Gamma(E) \to \Gamma(E) \otimes \Gamma(T^*M)$$

with the properties subsequently listed:

By property (i) below, we may also consider D as a map from $\Gamma(TM) \otimes \Gamma(E)$ to $\Gamma(E)$ and write for $\sigma \in \Gamma(E), V \in T_xM$

$$D\sigma(V) =: D_V\sigma.$$

We then require:

(i) D is tensorial in V :

$$D_{V+W}\sigma = D_V\sigma + D_W\sigma \qquad \text{for } V, W \in T_xM, \sigma \in \Gamma(E), \tag{3.1.1}$$

$$D_{fV}\sigma = fD_V\sigma \qquad \text{for } f \in C^\infty(M, \mathbb{R}), V \in \Gamma(TM). \tag{3.1.2}$$

(ii) D is $\mathbb{R}$-linear in σ :

$$D_V(\sigma + \tau) = D_V\sigma + D_V\tau \quad \text{for } V \in T_xM, \sigma, \tau \in \Gamma(E) \tag{3.1.3}$$

and it satisfies the following product rule:

$$D_V(f\sigma) = V(f) \cdot \sigma + fD_V\sigma \qquad \text{for } f \in C^\infty(M, \mathbb{R}). \tag{3.1.4}$$

Of course, all these properties are satisfied for the differentiation of a vector field in $\mathbb{R}^d$ as described; in that case, we have $D_VX = dX(V)$.

Let $x_0 \in M$, and let U be an open neighborhood of x_0 such that a chart for M and a bundle chart for E are defined on U. We thus obtain coordinate vector fields $\frac{\partial}{\partial x^1}, \dots, \frac{\partial}{\partial x^d}$, and through the identification

$$E_{|U} \cong U \times \mathbb{R}^n \quad (n = \text{fiber dimension of } E),$$

a basis of $\mathbb{R}^n$ yields a basis $\mu_1, \dots, \mu_n$ of sections of $E_{|U}$. For a connection D, we define the so-called *Christoffel symbols* Γ^k_{ij} $(j, k = 1, \dots, n, i = 1, \dots, d)$ by

$$D_{\frac{\partial}{\partial x^i}}\mu_j =: \Gamma^k_{ij}\mu_k. \tag{3.1.5}$$

We shall see below that the Christoffel symbols as defined here are a generalization of those introduced in §1.4.

Let now $\mu \in \Gamma(E)$; locally, we write $\mu(y) = a^k(y)\mu_k(y)$. Also let $c(t)$ be a smooth curve in U. Putting $\mu(t) := \mu(c(t))$, we define a section of E along c. Furthermore, let $V(t) = \dot{c}(t)(:= \frac{d}{dt}c(t)) = \dot{c}^i(t)\frac{\partial}{\partial x^i}$.

Then by (3.1.1) – (3.1.5)

$$\begin{aligned} D_{V(t)}\mu(t) &= \dot{a}^k(t)\mu_k(c(t)) + \dot{c}^i(t)a^k(t)D_{\frac{\partial}{\partial x^i}}\mu_k \\ &= \dot{a}^k(t)\mu_k(c(t)) + \dot{c}^i(t)a^k(t)\Gamma^j_{ik}(c(t))\mu_j(c(t)). \end{aligned} \tag{3.1.6}$$

(In particular, $D_X\mu$ depends only on the values of μ along a curve c with $\dot{c}(0) = X$, and not on all the values of μ in a neighborhood of the base point of X.)

$D_{V(t)}\mu(t) = 0$ thus represents a linear system of first order ODEs for the coefficients $a^1(t), \ldots, a^n(t)$ of $\mu(t)$. Therefore, for given initial values $\mu(0) \in E_{c(0)}$, there exists a unique solution of

$$D_{V(t)}\mu(t) = 0\,. \tag{3.1.7}$$

Definition 3.1.2. The solution $\mu(t)$ of (3.1.7) is called the *parallel transport* of $\mu(0)$ along the curve c.

Thus, if x_0 and x_1 are points in M, the fibers of E above x_0 and x_1, E_{x_0} and E_{x_1}, resp., can be identified by choosing a curve c from x_0 to x_1 ($x_0 = c(0), x_1 = c(1)$) and moving each $\mu_0 \in E_{x_0}$ along c to E_{x_1} by parallel transport. This identification depends only on the choice of the curve c. One might now try to select geodesics w.r.t. a Riemannian metric as canonical curves, but those are in general not uniquely determined by their endpoints.

From parallel transport on a Riemannian manifold, i.e. the identification of the fibers of a vector bundle along curves, one may obtain a notion of covariant derivative. For this purpose, given $V \in T_xM$, let c be a curve in M with $c(0) = x, \dot{c}(0) = V$. For $\mu \in \Gamma(E)$, we then put

$$D_V\mu := \lim_{t\to 0} \frac{P_{c,t}(\mu(c(t))) - \mu(c(0))}{t},$$

where $P_{c,t} : E_{c(t)} \to E_{c(0)}$ is the identification by parallel transport along c. In order to see that the two processes of covariant derivative and parallel transport are equivalent, we select a basis of *parallel* sections $\mu_1(t), \ldots, \mu_n(t)$ of E along c,

i.e.

$$D_{\dot{c}(t)}\mu_j(t) = 0 \quad \text{for}\, j = 1, \ldots, n. \tag{3.1.8}$$

An arbitrary section μ of E along c is then written as

$$\mu(t) = a^k(t)\mu_k(t),$$

and for $X = \dot{c}(0)$, we have

$$D_X\mu(t) = \dot{a}^k(t)\mu_k(t) \quad \text{by (3.1.6), (3.1.8)} \tag{3.1.9}$$

and consequently,

$$\begin{aligned}(D_X\mu)(c(0)) &= \lim_{t\to 0}\frac{a^k(t)-a^k(0)}{t}\mu_k(0)\\ &= \lim_{t\to 0}\frac{P_{c,t}(\mu(t))-\mu(0)}{t}\ .\end{aligned}$$

It is important to remark that this does not depend on the choice of the curve c, as long as $\dot{c}(0)=X$.

We want to explain the name "connection". We consider the tangent space at the point ψ to the total space E of a vector bundle, $T_\psi E$. Inside $T_\psi E$, there is a distinguished subspace, namely the tangent space to the fiber E_x containing ψ ($x=\pi(\psi)$). This space is called vertical space V_ψ. However, there is no distinguished "horizontal space" H_ψ complementary to V_ψ, i.e. satisfying $T_\psi E = V_\psi \oplus H_\psi$. If we have a covariant derivative D, however, we can parallely transport ψ for each $X \in T_x M$ along a curve $c(t)$ with $c(0)=x, \dot{c}(0)=X$. Thus, for each X, we obtain a curve $\psi(t)$ in E. The subspace of $T_\psi E$ spanned by all tangent vectors to E at ψ of the form

$$\frac{d}{dt}\psi(t)_{|t=0}$$

then is the horizontal space H_ψ. In this manner, one obtains a rule how the fibers in neighboring points are "connected" with each other.

We return to (3.1.6), i.e.

$$\begin{aligned}&D_{\dot{c}^i(t)\frac{\partial}{\partial x^i}}(a^j(t)\mu_j(c(t)))\\ &\quad = \dot{a}^j(t)\mu_j(c(t)) + \dot{c}^i(t)a^j(t)\Gamma^k_{ij}(c(t))\mu_k(c(t)).\end{aligned} \tag{3.1.10}$$

Here,

$$\dot{a}^j(t) = \dot{c}^i(t)\frac{\partial a^j}{\partial x^i}(c(t)). \tag{3.1.11}$$

This part thus is completely independent of D.

Γ^k_{ij} now has indices j and k, running from 1 to n, and an index running from 1 to d. The index i describes the application of the tangent vector $\dot{c}^i(t)\frac{\partial}{\partial x^i}$. We thus consider $(\Gamma^k_{ij})_{i,j,k}$ as an $(n\times n)$-matrix valued 1-form on U :

$$(\Gamma^k_{ij})_{i,j,k} \in \Gamma(\mathfrak{gl}(n,\mathbb{R})\otimes T^*M_{|U}) \tag{3.1.12}$$

(Here, $\mathfrak{gl}(n,\mathbb{R})$ is the space of $(n\times n)$-matrices with real coefficients.) In a more abstract manner, we now write on U

$$D = d + A, \tag{3.1.13}$$

where d is exterior derivative and $A\in\Gamma(\mathfrak{gl}(n,\mathbb{R})\otimes T^*M_{|U})$. Of course, A can also be considered as an $(n\times n)$-matrix with values in sections of the cotangent bundle of M; A, applied to the tangent vector $\frac{\partial}{\partial x^i}$, becomes $(\Gamma^k_{ij})_{j,k=1,\dots,n}$. By (3.1.10), the application of A to $a^j\mu_j$ is given by ordinary matrix multiplication. Once more:

$$D(a^j\mu_j) = d(a^j)\mu_j + a^j A\mu_j, \tag{3.1.14}$$

where A is a matrix with values in T^*M.

We now want to study the transformation behavior of A. As in §1.8, let $(U_\alpha)_{\alpha\in A}$ be a covering of M by open sets over which the bundle is trivial, with transition maps

$$\varphi_{\beta\alpha} : U_\alpha \cap U_\beta \to \mathrm{Gl}\,(n, \mathbb{R}).$$

D then defines a T^*M-valued matrix A_α on U_α. Let the section μ be represented by μ_α on U_α. Here, a Greek index is not a coordinate index, but refers to the chosen covering (U_α). Thus,

$$\mu_\beta = \varphi_{\beta\alpha}\, \mu_\alpha \text{ on } U_\alpha \cap U_\beta. \tag{3.1.15}$$

But then we must also have

$$\varphi_{\beta\alpha}(d + A_\alpha)\mu_\alpha = (d + A_\beta)\mu_\beta \text{ on}\, U_\alpha \cap U_\beta; \tag{3.1.16}$$

on the left hand side we have first computed $D\mu$ in the trivialization defined by the U_α and then transformed the result to the trivialization defined by U_β, while on the right hand side, we have directly expressed $D\mu$ in the latter trivialization.

We obtain

$$A_\alpha = \varphi_{\beta\alpha}^{-1} d\varphi_{\beta\alpha} + \varphi_{\beta\alpha}^{-1} A_\beta \varphi_{\beta\alpha}. \tag{3.1.17}$$

This formula gives the desired *transformation behavior*. Thus, A_α does not transform as a tensor (see the discussion following Definition 1.8.10), because of the term $\varphi_{\beta\alpha}^{-1} d\varphi_{\beta\alpha}$. However, the *difference* of two connections transforms as a tensor. The space of all connections on a given vector bundle E thus is an affine space. The difference of two connections D_1, D_2 is a $\mathfrak{gl}(n, \mathbb{R})$-valued 1-form, i.e. $D_1 - D_2 \in \Gamma(\mathrm{End}\, E \otimes T^*M)$, considering $\mathfrak{gl}(n, \mathbb{R})$ as the space of linear endomorphisms of the fibers.

We return to our fixed neighborhood U and thus drop the index α.

We want to extend D from E to other bundles associated with E, in particular to E^* and $\mathrm{End}(E) = E \otimes E^*$.

We now write

$$A\mu_j = A_j^k \mu_k, \tag{3.1.18}$$

where each A_j^k now is a 1-form, $A_j^k = \Gamma_{ij}^k dx^i$. Let $\mu_1^*, \ldots, \mu_n^*$ be the basis dual to $\mu_1, \ldots, \mu_n$ on the bundle E^* dual to E, i.e.

$$(\mu_i, \mu_j^*) = \delta_{ij}, \tag{3.1.19}$$

where $(\cdot, \cdot) : E \otimes E^* \to \mathbb{R}$ is the bilinear pairing between E and E^*.

Definition 3.1.3. Let D be a connection on E. The connection D^* dual to D on the dual bundle E^* is defined by the requirement

$$d(\mu, \nu^*) = (D\mu, \nu^*) + (\mu, D^*\nu^*) \tag{3.1.20}$$

for any $\mu \in \Gamma(E), \nu^* \in \Gamma(E^*)$.

($D\mu \in \Gamma(E \otimes T^*M)$, and $(D\mu, \nu^*)$ pairs the E-factor of $D\mu$ with ν^*. Thus $(D\mu, \nu^*)$, and similarly $(\mu, D^*\nu^*)$, is a 1-form.)

As usual, we write $D = d + A$ on U and compute

$$\begin{aligned} 0 = d(\mu_i, \mu_j^*) &= (A_i^k \mu_k, \mu_j^*) + (\mu_i, A_j^{*\ell} \mu_\ell^*) \\ &= A_i^j + A_j^{*i} \quad \text{by (3.1.19)}, \end{aligned}$$

i.e.

$$A^* = -A^t. \tag{3.1.21}$$

Recalling (3.1.15), we get

$$D^*_{\frac{\partial}{\partial x^i}} \mu_j^* = -\Gamma_{ik}^j \mu_k^*.$$

Definition 3.1.4. Let E_1, E_2 be vector bundles over M with connections D_1, D_2, resp. The induced connection D on $E := E_1 \otimes E_2$ is defined by the requirement

$$D(\mu_1 \otimes \mu_2) = D_1\mu_1 \otimes \mu_2 + \mu_1 \otimes D_2\mu_2 \tag{3.1.22}$$

for $\mu_i \in \Gamma(E_i), i = 1, 2$.

In particular, we obtain an induced connection on $\mathrm{End}\,(E) = E \otimes E^*$, again denoted by D. Let $\sigma = \sigma_j^i \mu_i \otimes \mu_j^*$ be a section of $\mathrm{End}\,(E)$. We compute

$$\begin{aligned} D(\sigma_j^i \mu_i \otimes \mu_j^*) &= d\sigma_j^i \mu_i \otimes \mu_j^* + \sigma_j^i A_i^k \mu_k \otimes \mu_j^* - \sigma_j^i A_k^j \mu_i \otimes \mu_k^* \\ &= d\sigma + [A, \sigma]. \end{aligned} \tag{3.1.23}$$

The induced connection on $\mathrm{End}\,(E)$ thus operates by taking the Lie bracket.

We next want to extend the operation of a connection D from $\Gamma(E)$ to $\Gamma(E) \otimes \Omega^p(M)$ $(0 \le p \le d)$. Since, on $\Omega^p(M)$, we have the exterior derivative d, we define in analogy with Definition 3.1.4 for $\mu \in \Gamma(E), \omega \in \Omega^p(M)$

$$D(\mu \otimes \omega) = D\mu \wedge \omega + \mu \otimes d\omega. \tag{3.1.24}$$

(Here, we have employed a wedge product of forms with values in vector bundles, as $D\mu$ is an element of $\Gamma(E) \otimes \Omega^1(M)$: If $\sigma \in \Gamma(E), \omega_1 \in \Omega^1(M), \omega_2 \in \Omega^p(M)$, then

$$(\sigma \otimes \omega_1) \wedge \omega_2 := \sigma \otimes (\omega_1 \wedge \omega_2),$$

and the general case is defined by linear extension.)

As an abbreviation, we write

$$\Omega^p(E) := \Gamma(E) \otimes \Omega^p(M), \quad \Omega^p := \Omega^p(M).$$

Thus

$$D : \Omega^p(E) \to \Omega^{p+1}(E), \quad 0 \le p \le d.$$

We want to compare this with the exterior derivative

$$d : \Omega^p \to \Omega^{p+1}.$$

Here, we have

$$d \circ d = 0.$$

Such a relation, however, in general does not hold anymore for D.

Definition 3.1.5. The *curvature* of a connection D is the operator

$$F := D \circ D : \Omega^0(E) \to \Omega^2(E).$$

The connection is called *flat*, if its curvature satisfies $F = 0$.

The exterior derivative d thus yields a flat connection on the trivial bundle $M \times \mathbb{R}$.

We compute for $\mu \in \Gamma(E)$

$$\begin{aligned} F(\mu) &= (d + A) \circ (d + A)\mu \\ &= (d + A)(d\mu + A\mu) \\ &= (dA)\mu - Ad\mu + Ad\mu + A \wedge A\mu \end{aligned}$$

(the minus sign occurs, because A is a 1-form).

Thus

$$F = dA + A \wedge A. \tag{3.1.25}$$

If we write $A = A_j dx^j$, (3.1.24) becomes

$$\begin{aligned} F &= \left(\frac{\partial A_j}{\partial x^i} + A_i A_j \right) dx^i \wedge dx^j \\ &= \frac{1}{2} \left(\frac{\partial A_j}{\partial x^i} - \frac{\partial A_i}{\partial x^j} + [A_i, A_j] \right) dx^i \wedge dx^j \end{aligned} \tag{3.1.26}$$

(note that each A_j is an $(n \times n)$-matrix).

We now want to compute DF. F is a map from $\Omega^0(E)$ to $\Omega^2(E)$, i.e.

$$F \in \Omega^2(E) \otimes (\Omega^0(E))^* = \Omega^2(\operatorname{End} E).$$

We thus consider F as a 2-form with values in $\operatorname{End} E$. By (3.1.23) then

$$\begin{aligned} DF &= dF + [A, F] \\ &= dA \wedge A - A \wedge dA + [A, dA + A \wedge A] \text{ by (3.1.23)} \\ &= dA \wedge A - A \wedge dA + A \wedge dA - dA \wedge A + [A, A \wedge A] \\ &= [A, A \wedge A] \\ &= [A_i dx^i, A_j dx^j \wedge A_k dx^k] \\ &= A_i A_j A_k (dx^i \wedge dx^j \wedge dx^k - dx^j \wedge dx^k \wedge dx^i) \\ &= 0. \end{aligned}$$

This is the so-called *second Bianchi identity.*

Theorem 3.1.1. *The curvature F of a connection D satisfies*

$$DF = 0.$$

□

We now want to study the transformation behavior of F. We use the same covering $(U_\alpha)_{\alpha\in A}$ as above, and on U_α, we write again $D = d + A_\alpha$, $A_\alpha = A_{\alpha,i}dx^i$. F then has the corresponding representation

$$F_\alpha = \frac{1}{2}\left(\frac{\partial A_{\alpha,j}}{\partial x^i} - \frac{\partial A_{\alpha,i}}{\partial x^j} + [A_{\alpha,i}, A_{\alpha,j}]\right) dx^i \wedge dx^j \tag{3.1.27}$$

by (3.1.26). Using the transformation formula (3.1.16) for A_α, we see that in the transformation formula for F_α, all derivatives of $\varphi_{\beta\alpha}$ cancel, and we have

$$F_\alpha = \varphi_{\beta\alpha}^{-1} F_\beta \varphi_{\beta\alpha} . \tag{3.1.28}$$

Thus, in contrast to A, F transforms as a tensor.

We now want to express F in terms of the Christoffel symbols. In order to make contact with the classical notation, we denote the curvature operator, considered as an element of $\Omega^2(\operatorname{End} E)$ by R :

$$\begin{aligned} F : \Omega^0(E) &\to \Omega^2(E) \\ \mu &\mapsto R(\cdot,\cdot)\mu, \end{aligned}$$

and we define the components $R^k_{\ell ij}$ by

$$R\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right)\mu_\ell = R^k_{\ell ij}\mu_k \tag{3.1.29}$$

($k, \ell \in \{1, \dots, n\}, i, j \in \{1, \dots, d\}$). By (3.1.26)

$$\begin{aligned} R(\cdot,\cdot)\mu_\ell &= F\mu_\ell \\ &= \frac{1}{2}\left(\frac{\partial \Gamma^k_{j\ell}}{\partial x^i} - \frac{\partial \Gamma^k_{i\ell}}{\partial x^j} + \Gamma^k_{im}\Gamma^m_{j\ell} - \Gamma^k_{jm}\Gamma^m_{i\ell}\right) dx^i \wedge dx^j \otimes \mu_k, \end{aligned} \tag{3.1.30}$$

i.e.

$$R^k_{\ell ij} = \frac{\partial \Gamma^k_{j\ell}}{\partial x^i} - \frac{\partial \Gamma^k_{i\ell}}{\partial x^j} + \Gamma^k_{im}\Gamma^m_{j\ell} - \Gamma^k_{jm}\Gamma^m_{i\ell} . \tag{3.1.31}$$

Theorem 3.1.2. *The curvature tensor R of a connection D satisfies*

$$R(X, Y)\mu = D_X D_Y \mu - D_Y D_X \mu - D_{[X,Y]}\mu \tag{3.1.32}$$

for all vector fields X, Y on M, and all $\mu \in \Gamma(E)$.

Proof. A direct computation is possible. However, one may also argue more abstractly as follows: First, (3.1.32) holds for $X = \frac{\partial}{\partial x^i}, Y = \frac{\partial}{\partial x^j}$. Namely, in this case $[X, Y] = 0$, and (3.1.32) follows from (3.1.26).

We have seen already that R transforms as a tensor (the tensorial transformation behavior w.r.t. X, Y also follows from (3.1.26), for example), and thus the value of $R(X,Y)\mu$ at the point x depends only on the values of X and Y at x. Now for $X = \xi^i \frac{\partial}{\partial x^i}, Y = \eta^j \frac{\partial}{\partial x^j}$

$$\begin{aligned} D_X D_Y \mu - D_Y D_X \mu &= \xi^i \frac{\partial \eta^j}{\partial x^i} D_{\frac{\partial}{\partial x^j}} \mu - \eta^j \frac{\partial \xi^i}{\partial x^j} D_{\frac{\partial}{\partial x^i}} \mu \\ &\quad + \xi^i \eta^j \left(D_{\frac{\partial}{\partial x^i}} D_{\frac{\partial}{\partial x^j}} - D_{\frac{\partial}{\partial x^j}} D_{\frac{\partial}{\partial x^i}} \right) \mu \end{aligned}$$

and

$$D_{[X,Y]}\mu = D_{(\xi^i \frac{\partial \eta^j}{\partial x^i} \frac{\partial}{\partial x^j} - \eta^j \frac{\partial \xi^i}{\partial x^j} \frac{\partial}{\partial x^i})} \mu,$$

hence

$$D_X D_Y \mu - D_Y D_X \mu - D_{[X,Y]}\mu = \xi^i \eta^j \left(D_{\frac{\partial}{\partial x^i}} D_{\frac{\partial}{\partial x^j}} - D_{\frac{\partial}{\partial x^j}} D_{\frac{\partial}{\partial x^i}} \right) \mu,$$

and this has the desired tensorial form. □

In order to develop the geometric intuition for the curvature tensor, we want to consider vector fields X, Y with $[X,Y] = 0$, e.g. coordinate vector fields $\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}$. Then

$$R(X,Y) = D_X D_Y - D_Y D_X.$$

When forming $D_X D_Y \mu$, we first move μ by infinitesimal parallel transport in the direction Y and then in the direction X; when forming $D_Y D_X \mu$, the order is reversed. $R(X,Y)\mu$ then expresses the difference in the results of these two operations, or, in other words, the dependence of parallel transport on the chosen path.

Corollary 3.1.1. *We have*

$$R(X,Y) = -R(Y,X). \tag{3.1.33}$$

Proof. From (3.1.32). □

Corollary 3.1.2.

$$R^k_{\ell ij} = -R^k_{\ell ji} \ \forall\, i, j, k, \ell\,.$$

Proof. This reformulation of (3.1.33) also follows from (3.1.30). □

Connections on the tangent bundle TM are particularly important:

Definition 3.1.6. Let ∇ be a connection on the tangent bundle TM of a differentiable manifold M. A curve $c : I \to M$ is called *autoparallel* or *geodesic* w.r.t. ∇ if

$$\nabla_{\dot c} \dot c \equiv 0, \tag{3.1.34}$$

i.e. if the tangent field of c is parallel along c.

In local coordinates, $\dot{c} = \dot{c}^i \frac{\partial}{\partial x^i}$, and

$$\nabla_{\dot{c}} \dot{c} = (\ddot{c}^k + \Gamma^k_{ij} \dot{c}^i \dot{c}^j) \frac{\partial}{\partial x^k}, \tag{3.1.35}$$

and the equation for geodesics has the same form as the one in §1.4. The difference is that the Christoffel symbols now have been defined differently. We shall clarify the relation between these two definitions below in §3.3. According to (3.1.35), (3.1.34) is a system of 2nd order ODE, and thus, as in §1.4, for each $x \in M, X \in T_x M$, there exist a maximal interval $I = I_X \subset \mathbb{R}$ with $0 \in I_X$ and a geodesic $c = c_X$

$$c : I \to M$$

with $c(0) = x, \dot{c}(0) = X$.

$C := \{X \in TM : 1 \in I_X\}$ is a star-shaped neighborhood of the zero section of TM, and as in §1.4, we define an exponential map by

$$\begin{aligned} \exp : C &\to M \\ X &\mapsto c_X(1). \end{aligned}$$

If $X \in C, 0 \le t \le 1$, then $\exp(tX) = c_X(t)$.

Definition 3.1.7. The *torsion tensor* of a connection ∇ on TM is defined as

$$T(X,Y) := T_\nabla(X,Y) := \nabla_X Y - \nabla_Y X - [X,Y] \quad (X, Y \in \Gamma(TM)). \tag{3.1.36}$$

∇ is called *torsion free* if

$$T \equiv 0. \tag{3.1.37}$$

Remark. It is not difficult to verify that T is indeed a tensor, i.e. that the value of $T(X,Y)(x)$ only depends on the values of X and Y at the point x.

In terms of our local coordinates, the components of the torsion tensor T are given by

$$T_{ij} = T\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) = \nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} - \nabla_{\frac{\partial}{\partial x^j}} \frac{\partial}{\partial x^i} = (\Gamma^k_{ij} - \Gamma^k_{ji}) \frac{\partial}{\partial x^k}. \tag{3.1.38}$$

We conclude

Lemma 3.1.1. *The connection ∇ on TM is torsion free if and only if*

$$\Gamma^k_{ij} = \Gamma^k_{ji} \quad \textit{for all } i, j, k. \tag{3.1.39}$$

□

Definition 3.1.8. A connection ∇ on TM is called *flat* if each point in M possesses a neighborhood U with local coordinates for which all the coordinate vector fields $\frac{\partial}{\partial x^i}$ are parallel, that is,

$$\nabla \frac{\partial}{\partial x^i} = 0. \tag{3.1.40}$$

Theorem 3.1.3. *A connection ∇ on TM is flat if and only if its curvature and torsion vanish identically.*

Proof. When the connection is flat, all $\nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} = 0$, and so, all Christoffel symbols $\Gamma_{ij}^k = 0$, and therefore, also T and R vanish, as they can be expressed in terms of the Γ_{ij}^k.

For the converse direction, we need to find local coordinates for which $0 = \nabla_{\frac{\partial}{\partial x^j}} \frac{\partial}{\partial x^i}$ for all i, j. Putting $\mu_i := \frac{\partial}{\partial x^i}$, we obtain the system

$$\frac{\partial}{\partial x^j} \mu_i + \Gamma_{ji}^k \mu_k = 0 \quad \text{for all}\, i, j. \tag{3.1.41}$$

In vector notation, this becomes

$$\frac{\partial}{\partial x^j} \mu + \Gamma_j \mu = 0, \tag{3.1.42}$$

and by the theorem of Frobenius, this can be solved if and only if the integrability condition

$$[\Gamma_i, \Gamma_j] + \frac{\partial}{\partial x^i} \Gamma_j - \frac{\partial}{\partial x^j} \Gamma_i = 0 \tag{3.1.43}$$

holds for all i, j. With indices, this is

$$\frac{\partial \Gamma_{j\ell}^k}{\partial x^i} - \frac{\partial \Gamma_{i\ell}^k}{\partial x^j} + \Gamma_{im}^k \Gamma_{j\ell}^m - \Gamma_{jm}^k \Gamma_{i\ell}^m = 0 \quad \text{for all}\, i, j, \tag{3.1.44}$$

which by equation (3.1.31) means that the curvature tensor vanishes. We can thus solve (3.1.41) for the μ_i. In order that these μ_i are coordinate vector fields $\frac{\partial}{\partial x^i}$, the necessary and sufficient condition (again, by the theorem of Frobenius) is

$$\frac{\partial}{\partial x^i} \mu_j = \frac{\partial}{\partial x^j} \mu_i \quad \text{for all}\, i, j, \tag{3.1.45}$$

which by (3.1.41) in turn is equivalent to the condition $\Gamma_{ij}^k = \Gamma_{ji}^k \quad \text{for all}\, i, j, k$, that is, by Lemma 3.1.1, the vanishing of the torsion T. This completes the proof. □

Perspectives. Ehresmann was the first to arrive at the correct concept of a connection in a vector bundle. Equivalently, the concept may also be introduced in a principal bundle (see the discussion at the end of §1.10). The theory of connections is systematically explored in [169, 170].

The curvature tensor introduced here generalizes the Riemann curvature tensor derived from a Riemannian metric in §3.3 below.

The Bianchi identity (Theorem 3.1.1) may be derived in a more conceptual way as the infinitesimal version of the equivariance of the curvature form F with respect to certain transformations in horizontal directions, see[223].

For a more detailed and elementary discussion of integrability conditions and the Frobenius theorem, we refer to [76].

3.2 Metric Connections. The Yang-Mills Functional

Definition 3.2.1. Let E be a vector bundle on the differentiable manifold M with bundle metric $\langle \cdot, \cdot \rangle$. A connection D on E is called *metric* if

$$d\langle \mu, \nu \rangle = \langle D\mu, \nu \rangle + \langle \mu, D\nu \rangle \text{ for all } \mu, \nu \in \Gamma(E). \tag{3.2.1}$$

A metric connection thus has to respect an additional structure, namely the metric.

We want to interpret condition (3.2.1). Let $X \in T_x M$; (3.2.1) then means

$$X\langle \mu, \nu \rangle = \langle D_X \mu, \nu \rangle + \langle \mu, D_X \nu \rangle. \tag{3.2.2}$$

Let now $c : I \to M$ be a smooth curve, and let $\mu(t)$ and $\nu(t)$ be parallel along c, i.e. $D_{\dot{c}}\mu = 0 = D_{\dot{c}}\nu$. Then from (3.2.2)

$$\frac{d}{dt}\langle \mu(t), \nu(t) \rangle = 0. \tag{3.2.3}$$

This can be interpreted as follows:

Lemma 3.2.1. *The parallel transport induced by a metric connection on a vector bundle preserves the bundle metric in the sense that parallel transport constitutes an isometry of the corresponding fibers.*

Namely, (3.2.3) means that the scalar product is preserved under parallel transport. □

Lemma 3.2.2. *Let D be a metric connection on the vector bundle E with bundle metric $\langle \cdot, \cdot \rangle$. Assume that w.r.t. a metric bundle chart (cf. Definition 1.8.12 and Theorem 1.8.3), we have the decomposition*

$$D = d + A.$$

Then for any $X \in TM$, the matrix $A(X)$ is skew symmetric, i.e.

$$A(X) \in \mathfrak{o}(n) \quad (= \textit{Lie algebra of } \mathrm{O}(n)) \quad (n = \textit{ rank of } E).$$

Proof. As described in Theorem 1.8.3, a metric bundle chart (f, U) generates sections $\mu_1, \ldots, \mu_n$ on U that form an orthonormal basis of the fiber E_x at each $x \in U$, i.e.

$$\langle \mu_i(x), \mu_j(x) \rangle = \delta_{ij}.$$

Moreover, since the μ_i are constant in the bundle chart, we have for the exterior derivative d defined by the chart

$$d\mu_i \equiv 0 \quad (i = 1, \ldots, n).$$

Let now $X \in T_x M, x \in U$.

It follows that

$$\begin{aligned} 0 = X\langle \mu_i, \mu_j \rangle &= \langle A(X)\mu_i, \mu_j \rangle + \langle \mu_i, A(X)\mu_j \rangle \\ &= \langle A(X)_i^k \mu_k, \mu_j \rangle + \langle \mu_i, A(X)_j^k \mu_k \rangle \\ &= A(X)_i^j + A(X)_j^i \,. \end{aligned}$$

□

By

$$\Omega^p(\mathrm{Ad}\, E),$$

we denote the space of those elements of $\Omega^p(\mathrm{End}\, E)$ for which the endomorphism of each fiber is skew symmetric. Thus, if $D = d + A$ is a metric connection, we have

$$A \in \Omega^1(\mathrm{Ad}\, E).$$

We define

$$D^* : \Omega^p(\mathrm{Ad}\, E) \to \Omega^{p-1}(\mathrm{Ad}\, E)$$

as the operator dual to

$$D : \Omega^{p-1}(\mathrm{Ad}\, E) \to \Omega^p(\mathrm{Ad}\, E)$$

w.r.t. $(\cdot, \cdot)$; thus

$$(D^*\nu, \mu) = (\nu, D\mu) \quad \text{for all}\, \mu \in \Omega^{p-1}(\mathrm{Ad}\, E), \nu \in \Omega^p(\mathrm{Ad}\, E). \tag{3.2.4}$$

This is in complete analogy with the definition of d^* in §2.1. Indeed, for $D = d + A$ ($A \in \Omega^1(\mathrm{Ad}\, E)$), $A = A_i dx^i$

$$(\nu, d\mu + A_i dx^i \wedge \mu) = (d^*\nu, \mu) - (A_i \nu, dx^i \wedge \mu), \quad \text{since}\ A_i \ \text{is skew symmetric.} \tag{3.2.5}$$

By Lemma 2.1.1, in this case

$$** = (-1)^{p(d-p)}.$$

$* : \Omega^p(\mathrm{Ad}\, E) \to \Omega^{d-p}(\mathrm{Ad}\, E)$ operates on the differential form part as described in §2.1 and leaves the $\mathrm{Ad}\, E$-part as it is:

$$*(\mu \otimes \omega) = \mu \otimes *\omega \quad \text{for}\, \mu \in \Gamma(\mathrm{Ad}\, E), \omega \in \Omega^p,$$

and by Lemma 2.1.4

$$d^* = (-1)^{d(p+1)+1} * d * .$$

Moreover, A_i and $*$ commute, since A_i operates on the $\operatorname{Ad} E$-part and $*$ on the form part. In particular,

$$*A_i* = A_i .$$

Thus, from (3.2.5)

$$D^* = (-1)^{d(p+1)+1} * (d + A)* = (-1)^{d(p+1)+1} * D * . \tag{3.2.6}$$

(Note, however, that A operates on the form part by contraction and not by multiplication with dx^i).

In Chapter 9, we shall need to compute expressions of the form

$$\Delta\langle\varphi,\varphi\rangle$$

where φ is a section of a vector bundle E with a metric connection D. We obtain

$$\begin{aligned}\Delta\langle\varphi,\varphi\rangle &= d^* d\langle\varphi,\varphi\rangle\\ &= (-1) * d * d\langle\varphi,\varphi\rangle\\ &= 2(-1) * d * \langle D\varphi,\varphi\rangle \quad \text{since } D \text{ is metric}\\ &= 2(-1) * d\langle *D\varphi,\varphi\rangle\end{aligned}$$

since $D\varphi$ is a 1-form with values in E, and $*$ operates on the form part, whereas $\langle\cdot,\cdot\rangle$ multiplies the vector parts, and so $*$ and $\langle\cdot,\cdot\rangle$ commute

$$\begin{aligned}&= 2(-1) * (\langle D * D\varphi,\varphi\rangle + \langle *D\varphi, D\varphi\rangle) \quad \text{since } D \text{ is metric}\\ &= 2(\langle D^* D\varphi,\varphi\rangle - \langle D\varphi, D\varphi\rangle)\end{aligned}$$

by (3.2.6), and since $** = 1$ on 2-forms.
Thus, we obtain the formula

$$\Delta\langle\varphi,\varphi\rangle = 2(\langle D^* D\varphi,\varphi\rangle - \langle D\varphi, D\varphi\rangle). \tag{3.2.7}$$

We now study the curvature of a metric connection and observe first

Corollary 3.2.1. *Let $D = d + A$ be a metric connection on E. Then the curvature F of D satisfies*

$$F \in \Omega^2(\operatorname{Ad} E).$$

Proof. We consider (3.1.25). Under the conditions of Lemma 3.2.2,

$$\frac{\partial A_i}{\partial x^j} - \frac{\partial A_j}{\partial x^i} + [A_i, A_j]$$

is a skew symmetric matrix for each pair (i,j), because the Lie bracket of two skew symmetric matrices is skew symmetric again, since $\mathfrak{o}(n)$ is a Lie algebra. □

Note that $F_{ij} = \frac{1}{2}\left(\frac{\partial A_i}{\partial x^j} - \frac{\partial A_j}{\partial x^i} + [A_i, A_j]\right)$ is always skew symmetric in i and j. This is also expressed by Corollary 3.1.1. By way of contrast, Corollary 3.2.1 expresses the skew symmetry of the matrix

$$R^k_{\ell ij}$$

w.r.t. the indices k and ℓ :

Corollary 3.2.2. *For a metric connection,*

$$R^k_{\ell ij} = -R^\ell_{kij} \quad \textit{for all } i, j \in \{1, \ldots, d\}, k, \ell \in \{1, \ldots, n\} \tag{3.2.8}$$

($d = \dim M, n =$ rank of E). □

For $A, B \in \mathfrak{o}(n)$, we put

$$A \cdot B = -\operatorname{tr}(AB). \tag{3.2.9}$$

This is the negative of the Killing form of the Lie algebra $\mathfrak{o}(n)$. (3.2.9) defines a (positive definite) scalar product on $\mathfrak{o}(n)$. (3.2.9) then also defines a scalar product on $\operatorname{Ad} E$. We now recall that we also have a pointwise scalar product for p-forms: For $\omega_1, \omega_2 \in \Lambda^p T_x^* M$ we have

$$\langle \omega_1, \omega_2 \rangle = *(\omega_1 \wedge *\omega_2), \tag{3.2.10}$$

cf. Lemma 2.1.2. Thus, we also have a scalar product for $\mu_1 \otimes \omega_1, \mu_2 \otimes \omega_2 \in \operatorname{Ad} E_x \otimes \Lambda^p T_x^* M$, namely

$$\langle \mu_1 \otimes \omega_1, \mu_2 \otimes \omega_2 \rangle := \mu_1 \cdot \mu_2 \langle \omega_1, \omega_2 \rangle. \tag{3.2.11}$$

Thus, by linear extension, we also obtain a scalar product on $\operatorname{Ad} E_x \otimes \Lambda^p T_x^* M$. This in turn yields an L^2-scalar product on $\Omega^p(\operatorname{Ad} E)$:

$$(\mu_1 \otimes \omega_1, \mu_2 \otimes \omega_2) := \int_M \langle \mu_1 \otimes \omega_1, \mu_2 \otimes \omega_2 \rangle * (1), \tag{3.2.12}$$

assuming again that M is compact and oriented.

Definition 3.2.2. Let M be a compact, oriented Riemannian manifold, E a vector bundle with a bundle metric over M, D a metric connection on E with curvature $F_D \in \Omega^2(\operatorname{Ad} E)$. The *Yang-Mills functional* applied to D is

$$YM(D) := (F_D, F_D) = \int_M \langle F_D, F_D \rangle * (1).$$

We now recall that the space of all connections on E is an affine space; the difference of two connections is an element of $\Omega^1(\operatorname{End} E)$. Likewise, the space of all metric connections on E is an affine space; the difference of two metric connections is an element of $\Omega^1(\operatorname{Ad} E)$. If we want to determine the Euler-Lagrange equations for the Yang-Mills functional, we may thus use variations of the form

$$D + tB \quad \text{with } B \in \Omega^1(\operatorname{Ad} E).$$

For $\sigma \in \Gamma(E) = \Omega^0(E)$,

$$\begin{aligned} F_{D+tB}(\sigma) &= (D+tB)(D+tB)\sigma \\ &= D^2\sigma + tD(B\sigma) + tB \wedge D\sigma + t^2(B \wedge B)\sigma \\ &= (F_D + t(DB) + t^2(B \wedge B))\sigma, \end{aligned} \tag{3.2.13}$$

since $D(B\sigma) = (DB)\sigma - B \wedge D\sigma$ (compare the derivation of (3.1.25)).

Consequently

$$\begin{aligned} \frac{d}{dt} YM(D+tB)_{|t=0} &= \frac{d}{dt}\int \langle F_{D+tB}, F_{D+tB}\rangle * (1)_{|t=0} \\ &= 2\int \langle DB, F_D\rangle * (1). \end{aligned} \tag{3.2.14}$$

Recalling the definition of D^* (3.2.4), (3.2.14) becomes

$$\frac{d}{dt} YM(D+tB)_{|t=0} = 2(B, D^* F_D).$$

Thus, D is a critical point of the Yang-Mills functional if and only if

$$D^* F_D = 0. \tag{3.2.15}$$

Definition 3.2.3. A metric connection D on the vector bundle E with a bundle metric over the oriented Riemannian manifold M is called a *Yang-Mills connection* if

$$D^* F_D = 0.$$

We write $F_D = F_{ij} dx^i \wedge dx^j$, and we want to interpret (3.2.15) in local coordinates with $g_{ij}(x) = \delta_{ij}$. In such coordinates,

$$d^*(F_{ij} dx^i \wedge dx^j) = -\frac{\partial F_{ij}}{\partial x^i} dx^j,$$

and from (3.2.5) hence

$$D^* F_D = \left(-\frac{\partial F_{ij}}{\partial x^i} - [A_i, F_{ij}] \right) dx^j.$$

(3.2.15) thus means

$$\frac{\partial F_{ij}}{\partial x^i} + [A_i, F_{ij}] = 0 \quad \text{for } j = 1, \dots, d. \tag{3.2.16}$$

We now discuss gauge transformations.

Let E again be a vector bundle with a bundle metric. Aut (E) then is the bundle with fiber over $x \in M$ the group of orthogonal self transformations of the fiber E_x.

Definition 3.2.4. A *gauge transformation* is a section of $\mathrm{Aut}(E)$. The group $\mathcal{G}$ of gauge transformations is called the *gauge group* of the metric bundle E.

The group structure here is given by fiberwise matrix multiplication. $s \in \mathcal{G}$ operates on the space of metric connections D on E via

$$s^*(D) := s^{-1} \circ D \circ s,$$

i.e.

$$s^*(D)\mu = s^{-1} D(s\mu) \tag{3.2.17}$$

for $\mu \in \Gamma(E)$. For $D = d + A$, we obtain as in the proof of (3.1.16)

$$s^*(A) = s^{-1} ds + s^{-1} A s. \tag{3.2.18}$$

Subsequently, this notion will also be applied in somewhat greater generality. Namely, if the structure group of E is not necessarily $\mathrm{SO}(n)$, but any subgroup of $\mathrm{Gl}(,\mathbb{R})$, we let $\mathrm{Aut}\,(E)$ the bundle with fiber given by G, and operating on E again by conjugation. The group of sections of $\mathrm{Aut}\,(E)$ will again be called the gauge group.

Given $x_0 \in M$, we may always find a neighborhood of U of x_0 and a section s of $\mathrm{Aut}\,(E)$ over U, i.e. a gauge transformation defined on U, such that

$$s^*(A)(x_0) = 0$$

Namely, according to (3.2.18), we just have to solve

$$s(x_0) = \mathrm{id}, \quad ds(x_0) = -A(x_0).$$

This is possible since $A \in \Omega^1(\mathrm{Ad}\,E)$, and the fiber of $\mathrm{Ad}\,E$ is the Lie algebra of the fiber of $\mathrm{Aut}\,(E)$, a section of which s has to be. Thus,

Lemma 3.2.3. *Let D be a connection on the vector bundle E over M. For any $x_0 \in M$, there exists a gauge transformation s defined on some neighborhood of x_0 such that the gauge transformed connection $s^*(D)$ satisfies*

$$s^*(D) = d \qquad \text{at } x_0.$$

Of course, the gauge transformation can always be chosen to be compatible with any structure preserved by D, in particular a metric. □

In the same notation as in the derivation of (3.1.16), s as a section of $\mathrm{Aut}(E)$ transforms as

$$s_\beta = \varphi_{\beta\alpha} s_\alpha \varphi_{\beta\alpha}^{-1}. \tag{3.2.19}$$

The curvature F of D transforms as in (3.2.17):

$$s^* F = s^{-1} \circ F \circ s. \tag{3.2.20}$$

An orthogonal self map of E is an isometry of $\langle \cdot, \cdot \rangle$, and hence

$$\langle s^* F, s^* F \rangle = \langle F, F \rangle. \tag{3.2.21}$$

We conclude:

Theorem 3.2.1. *The Yang-Mills functional is invariant under the operation of the gauge group* $\mathcal{G}$*. Hence also the set of critical points of* YM*, i.e. the set of Yang-Mills connections, is invariant. Thus, if* D *is a Yang-Mills connection, so is* s^*D *for* $s \in \mathcal{G}$*.* □

Corollary 3.2.3. *The space of Yang-Mills connections on a given metric vector bundle* E *of rank* ≥ 2 *is infinite dimensional, unless empty.* □

For $n > 2$, $\mathfrak{o}(n)$ is nonabelian. Thus, by (3.2.18), in general not only $s^{-1}As \neq A$, but by (3.2.20) also

$$s^*F \neq F.$$

It is nevertheless instructive to consider the case $n = 2$. $\mathfrak{o}(2)$ is a trivial Lie algebra in the sense that the Lie bracket vanishes identically. $\operatorname{Ad} E$ thus is the trivial bundle $M \times \mathbb{R}$. Consequently for $D = d + A$

$$F = dA. \tag{3.2.22}$$

Similarly, the Bianchi identity (Theorem 3.1.1) becomes

$$dF = 0, \tag{3.2.23}$$

and the Yang-Mills equation (3.2.15) becomes

$$d^*F = 0. \tag{3.2.24}$$

(3.2.22) does not mean that the 2-form F is exact, because (3.2.22) depends on the local decomposition $D = d + A$ which in general is not global. That F, as the curvature of a connection, satisfies the Bianchi identity, does mean, however, that F is closed. F then is harmonic if and only if D is a Yang-Mills connection, cf. Lemma 2.1.5. Thus, existence and uniqueness of the curvature of a Yang-Mills connection are consequences of Hodge theory as in §2.2. Thus, Yang-Mills theory is a generalization (nonlinear in general) of Hodge theory.

We now write (for $n = 2$) $s \in \mathcal{G}$ as $s = e^u$. Then

$$s^*(A) = A + du \quad \text{by (3.2.18).} \tag{3.2.25}$$

(3.2.24) becomes

$$d^*dA = 0. \tag{3.2.26}$$

If we require in addition to $d^*dA = 0$ the gauge condition

$$d^*A = 0, \tag{3.2.27}$$

we obtain the equation

$$\Delta A = (d^*d + dd^*)A = 0. \tag{3.2.28}$$

Without the gauge fixing (3.2.27), if A is a solution of the Yang-Mills equation, so is

$$A + a \text{ with } a \in \Omega^1,\ da = 0,$$

and conversely, this way, knowing one solution, one obtains every other one; namely, if $A + a$ with $a \in \Omega^1$ is a solution, we get $d^*a = 0$, hence as in §2.1 $da = 0$. If $H^1(M, \mathbb{R}) = 0$, for each such a, there exists a function u with $a = du$. With $s = e^u$, we put

$$s^*(A) = A + a,$$

and thus, in this case $\mathcal{G}$ operates transitively on the space of Yang-Mills connections.

We now consider the case $d = 4$ which is of special interest for the Yang-Mills equations. As always, M is compact and oriented and carries a Riemannian metric. $*$ then maps $\Lambda^2 T_x^* M$ into itself:

$$* : \Lambda^2 T_x^* M \to \Lambda^2 T_x^* M \quad (x \in M).$$

Since by Lemma 2.1.1, $** = 1$, we obtain a decomposition

$$\Lambda^2 T_x^* M = \Lambda^+ \oplus \Lambda^-$$

into the eigenspaces of $*$ corresponding to the eigenvalues ± 1. $\Lambda^2 T_x^* M$ is of dimension 6, and Λ^+ and Λ^- are both of dimension 3. Choosing normal coordinates with center x, Λ^+ is generated by

$$\begin{gathered} dx^1 \wedge dx^2 + dx^3 \wedge dx^4 \\ dx^1 \wedge dx^4 + dx^2 \wedge dx^3 \\ dx^1 \wedge dx^3 - dx^2 \wedge dx^4 \end{gathered}$$

and Λ^- by

$$\begin{gathered} dx^1 \wedge dx^3 + dx^2 \wedge dx^4 \\ dx^1 \wedge dx^2 - dx^3 \wedge dx^4 \\ dx^1 \wedge dx^4 - dx^2 \wedge dx^3. \end{gathered}$$

The elements of Λ^+ are called selfdual, those of Λ^- antiselfdual.

Definition 3.2.5. A connection D on a vector bundle over an oriented four dimensional Riemannian manifold is called *(anti)selfdual* or an *(anti)instanton* if its curvature F_D is an (anti)selfdual 2-form.

Theorem 3.2.2. *Each (anti)selfdual metric connection is a solution of the Yang-Mills equations.*

Proof. The Yang-Mills equation is

$$D^* F = 0.$$

By (3.2.6), this is equivalent to

$$D * F = 0. \tag{3.2.29}$$

Let now F be (anti)selfdual. Then

$$F = \pm * F. \tag{3.2.30}$$

(3.2.29) then becomes

$$D * *F = 0,$$

hence by $** = 1$,

$$DF = 0.$$

This, however, is precisely the Bianchi identity, which is satisfied by Theorem 3.1.1. □

In order to find a global interpretation of Theorem 3.2.2 in terms of the Yang-Mills functional, it is most instructive to consider the case of $\mathrm{U}(m)$ or $\mathrm{SU}(m)$ connections instead of $\mathrm{SO}(n)$ connections. The preceding theory carries over with little changes from $\mathrm{SO}(n)$ to an arbitrary compact subgroup of the general linear group, in particular $\mathrm{U}(m)$ or $\mathrm{SU}(m)$. We shall also need the concept of Chern classes. For that purpose, let E now be a *complex* vector bundle of Rank m over the compact manifold M, D a connection in E with curvature $F = D^2 : \Omega^0 \to \Omega^2(E)$. We also recall the transformation rule (3.2.28):

$$F_\alpha = \varphi_{\beta\alpha}^{-1} F_\beta \varphi_{\alpha\beta} \tag{3.2.31}$$

which allows to consider F as an element of $\mathrm{Ad}\, E$; at the moment, the structure group is $\mathrm{Gl}(m, \mathbb{C})$ (as E is an arbitrary complex vector bundle), and so $\mathrm{Ad}\, E = \mathrm{End}\, E = \mathrm{Hom}_{\mathbb{C}}(E, E)$. We let M_m denote the space of complex $m \times m$-matrices, and we call a polynomial function, homogenous of degree k in its entries,

$$P : M_m \to \mathbb{C},$$

invariant if for all $B \in M_m$, $\varphi \in \mathrm{Gl}(m, \mathbb{C})$

$$P(B) = P(\varphi^{-1} B \varphi).$$

Examples are the elementary symmetric polynomials $P^j(B)$ of the eigenvalues of B. Those satisfy

$$\det(B + t\mathrm{Id}) = \sum_{k=0}^{m} P^{m-k}(B) t^k. \tag{3.2.32}$$

Similary, a k-linear form

$$\widetilde{P} : M_m \times \ldots \times M_m \to \mathbb{C}$$

is called invariant if for $B_1, \ldots, B_k \in M_m$, $\varphi \in \mathrm{Gl}(m, \mathbb{C})$

$$\widetilde{P}(B_1, \ldots, B_k) = \widetilde{P}(\varphi^{-1} B_1 \varphi, \ldots, \varphi^{-1} B_k \varphi).$$

The infinitesimal version of this property is that for all $B_1, \ldots, B_k \in M_m$, $A \in \mathfrak{gl}(m, \mathbb{C})$

$$\sum_{i=1}^{k} \tilde{P}(B_1, \ldots, [A, B_i], \ldots, B_k) = 0. \tag{3.2.33}$$

Restricting an invariant k-form to the diagonal defines an invariant polynomial

$$P(B) = \widetilde{P}(B, \ldots, B)$$

Conversely, given an invariant polynomial, we may obtain an invariant k-form by polarization:

$$\widetilde{P}(B_1, \ldots, B_k) := \frac{(-1)^k}{k!} \sum_{j=1}^{k} \sum_{i_1 < \ldots < i_j} (-1)^j P(B_{i_1} + \ldots + B_{i_j}).$$

Given an invariant polynomial P of degree k, we may use the transformation rule (3.2.31) for the curvature F of a connection D to define

$$P(F) := P(F_\alpha),$$

using any local trivialization. $P(F)$ then is a globally defined differential form of degree $2k$. In particular, $P(F)$ remains invariant under gauge transformations, as those transform F into $s^{-1} \circ F \circ s$, cf. (3.2.20)

Lemma 3.2.4. *For an invariant polynomial of degree k, we have $dP(F) = 0$. Consequently, $P(F)$ defines a cohomology class $[P(F)] \in H^{2k}(M)$, and this class does not depend on the chosen connection.*

Proof. Let $\widetilde{P}$ be an invariant k-form with $\widetilde{P}(B, \ldots, B) = P(B)$ as above. As explained in §3.1, we may extend D as

$$D : \Omega^p(\operatorname{End} E) \to \Omega^{p+1}(\operatorname{End} E).$$

Since $\widetilde{P}$ is linear, we have

$$d\widetilde{P}(B_1, \ldots, B_k) = \sum_i (-1)^{P_1 + \ldots, P_{i-1}} \widetilde{P}(B_1, \ldots, dB_i, \ldots, B_k)$$

By assumption

$$P(F) = \widetilde{P}(F, \ldots, F),$$

is invariant under gauge transformations. For any $x_0 \in M$, Lemma 3.2.3 means that after applying a local gauge transformation, we may assume that at x_0, we have

$$d = D.$$

Thus, at x_0,

$$dP(F) = \sum_i \widetilde{P}(F, \ldots, \underset{\substack{\uparrow \\ i^{\text{th}} \text{ entry}}}{DF}, \ldots F).$$

As x_0 was arbitrary, this holds for all M.
(Alternatively, this may also be derived from (3.2.33), without using Lemma 3.2.3).

The Bianchi identy $DF = 0$ thus implies

$$dP(F) = 0.$$

If D_0, D_1 are connections on E, then $\eta := D_1 - D_0 \in \Omega^1(\operatorname{End} E)$. We write locally

$$D_0 = d + A,$$

and we put

$$D_t := D_0 + t\eta = d + A + t\eta.$$

The curvatures thus are given by

$$F_t = d(A + t\eta) + (A + t\eta) \wedge (A + t\eta),$$

and

$$\frac{\partial}{\partial t} F_t = D_t \eta.$$

We obtain

$$\begin{aligned} \frac{\partial}{\partial t} P(F_t) &= k\widetilde{P}\left(\frac{\partial}{\partial t} F_t, F_t, \dots, F_t\right) \\ &= k\widetilde{P}\left(D_t \eta, F_t, \dots, F_t\right) \\ &= d(k\widetilde{P}(\eta, F_t, \dots, F_t)) \quad \text{as } D_t F_t = 0 \text{ by the Bianchi identiy.} \end{aligned}$$

Therefore

$$P(F_1) - P(F_0) = \int_0^1 \frac{\partial}{\partial t} P(F_t)\, dt$$

is cohomologous to zero. □

Definition 3.2.6. The *Chern classes* of E are defined as

$$c_j(E) = \left[P^j\left(\frac{i}{2\pi} F\right)\right] \in H^{2j}(M)$$

where P^j is the j^{th} elementary symmetric polynomial, and F is the curvature of an arbitrary connection on E.

Recalling (3.2.32), we have

$$\det\left(\frac{i}{2\pi} F + t\mathrm{Id}\right) = \sum_{k=0}^{m} c_{m-k}(E) t^k,$$

or with the eigenvalues λ_α of $\frac{i}{2\pi}F$ (the λ_α are 2-forms) and $\tau := t^{-1}$,

$$\sum_{j=0}^{m} c_j(E)\tau^j = \det\left(\frac{i}{2\pi}\tau F + \mathrm{Id}\right) = \prod_{\alpha=1}^{m}(1+\lambda_\alpha\tau). \tag{3.2.34}$$

In particular, we have

$$c_1(E) = \frac{i}{2\pi}\mathrm{tr}\, F, \tag{3.2.35}$$

$$c_2(E) - \frac{m-1}{2m}c_1(E)\wedge c_1(E) = \frac{1}{8\pi^2}\mathrm{tr}\,(F_0\wedge F_0), \tag{3.2.36}$$

where

$$F_0 := F - \frac{1}{m}\mathrm{tr}\, F\cdot \mathrm{Id}_E \quad \text{is the trace free part of } F. \tag{3.2.37}$$

We now return to the situation of a $\mathrm{U}(m)$ vector bundle E over a four dimensional oriented Riemannian manifold M. We let D be a unitary connection on E with curvature $F = D^2$ as usual. We decompose F_0 into its selfdual and antiselfdual components

$$F_0 = F_0^+ + F_0^-. \tag{3.2.38}$$

Then, since the $\wedge$ product of a selfdual 2-form with an antiselfdual one always vanishes (this can be seen from the above generators of Λ^+ and Λ^-),

$$\begin{aligned}\mathrm{tr}\,(F_0\wedge F_0) &= \mathrm{tr}\,(F_0^+\wedge F_0^+) + \mathrm{tr}\,(F_0^-\wedge F_0^-)\\ &= \mathrm{tr}\,(F_0^+\wedge *F_0^+) - \mathrm{tr}\,(F_0^-\wedge *F_0^-) \quad \text{since } *F_0^\pm = \pm F_0^\pm\\ &= -|F_0^+|^2 + |F_0^-|^2 \quad \text{cf. (3.2.9).}\end{aligned} \tag{3.2.39}$$

Recalling (3.2.36), we conclude that integrating over M yields

$$(c_2(E) - \frac{m-1}{2m}c_1(E)^2)[M] = -\frac{1}{8\pi^2}\int\left(|F_0^+|^2 - |F_0^-|^2\right) * (1). \tag{3.2.40}$$

The Yang-Mills functional decomposes as

$$\begin{aligned}YM(D) &= \int_M \left(\frac{1}{m}|\mathrm{tr}\, F|^2 + |F_0|^2\right) * (1)\\ &= \int_M \left(\frac{1}{m}|\mathrm{tr}\, F|^2 + |F_0^+|^2 + |F_0^-|^2\right) * (1).\end{aligned} \tag{3.2.41}$$

Since $\mathrm{tr}\, F$ represents the cohomology class $-2\pi i c_1(E)$, the cohomology class of $\mathrm{tr}\, F$ is fixed, and

$$\int_M |\mathrm{tr}\, F|^2 * (1)$$

becomes minimal if $\operatorname{tr} F$ is a harmonic 2-form in this class, see §2.1. $\int |\operatorname{tr} F|^2$ and $\int |F_0|^2$ may be minimized independently, and because of the constraint (3.2.40), $\int |F_0|^2$ becomes minimal if, depending on the sign of $(c_2(E) - \frac{m-1}{m} c_1(E)^2)[M]$,

$$F_0^+ = 0 \quad \text{or} \quad F_0^- = 0, \tag{3.2.42}$$

i.e. if F_0 is antiselfdual or selfdual.

If D is a $\mathrm{SU}(m)$ connection, then the fiber of $\operatorname{Ad} E$ is $\mathfrak{su}(m)$ which is tracefree, and thus $F \in \Omega^2(\operatorname{Ad} E)$ satisfies

$$\operatorname{tr} F = 0. \tag{3.2.43}$$

Hence, by (3.2.35)

$$c_1(E) = 0,$$

and by (3.2.36), (3.2.40)

$$c_2(E)[M] = -\frac{1}{8\pi^2} \int\limits_M \left(|F^+|^2 - |F^-|^2\right) * (1)$$

where $F^{\pm}$ are the (anti)selfdual parts of F.

Also,

$$YM(D) = \int\limits_M \left(|F^+|^2 + |F^-|^2\right) * (1)$$

then is minimized if F is (anti)selfdual, again depending on the sign of $c_2(E)[M]$. In conclusion we obtain

Theorem 3.2.3. *Let E be an $\mathrm{SU}(m)$ vector bundle over the compact oriented four dimensional manifold M. Then an $\mathrm{SU}(m)$ connection D on E yields an absolute minimum for YM if F is antiselfdual or selfdual (depending on the sign $c_2(E)[M]$), i.e if it satisfies the first order equation $F = \pm * F$.*

Remark. Here, we do not address the question when the lower bound for the Yang-Mills functional just derived is achieved, i.e. when there exist (anti) selfdual connections.

The Yang-Mills functional exhibits special features in dimension 4, as we have seen. There is also a functional that is well adapted to 3-dimensional manifolds, namely the Chern-Simons functional that we shall now briefly discuss.

Let M be a compact 3-dimensional differentiable manifold, and let E be a vector bundle over M with structure group a compact subgroup G of $\mathrm{Sl}(n, \mathbb{R})$, with Lie algebra $\mathfrak{g}$ as usual. We consider G-connections D, i.e. connections that can locally be written as

$$D = d + A, \quad \text{with } A \in \Omega^1(\mathfrak{g}).$$

(As before, we identify $\mathfrak{g}$ with the fibers of $\operatorname{Ad} E$, the endomorphisms of the fibers of E that are given by elements of $\mathfrak{g}$. The discussion here is a little more general than the one we presented in the 4-dimensional case, but the latter can easily be extended to the present level of generality as well.)

We also suppose that E is a trivial G-bundle, i.e. as a vector bundle, E is isomorphic to $M \times \mathbb{R}^n$, and the connection on E given by the exterior derivative d preserves the G-structure (e.g. if $G = \mathrm{SO}(n)$, and $\langle \cdot, \cdot \rangle$ is the corresponding metric on the fibers, then for any two sections σ_1, σ_2 of E (that are considered as functions $\sigma_1, \sigma_2 : M \to \mathbb{R}^n$ under the above isomorphism), we have

$$d\langle \sigma_1, \sigma_2 \rangle = \langle d\sigma_1, \sigma_2 \rangle + \langle \sigma_1, d\sigma_2 \rangle).$$

In this case, for any other G-connection

$$D = d + A$$

on E, A is a globally defined 1-form with values in $\mathfrak{g}$.

Definition 3.2.7. The *Chern-Simons functional of A* is defined as

$$CS(A) = \int_M \mathrm{tr}\left(A \wedge dA + \frac{2}{3} A \wedge A \wedge A\right). \tag{3.2.44}$$

(Here, tr of course is the trace in $\mathfrak{g}$, or in more general terms, the negative of the Killing form of $\mathfrak{g}$. In fact, one may take any Ad invariant scalar product on $\mathfrak{g}$ here.)

Remark. Without the assumption that E is a trivial G-bundle, we need to choose a base connection $D_0 = d + A_0$. For $D = d + A$, $A - A_0$ then is a globally defined 1-form with values in $\mathfrak{g}$, and we may thus insert $A - A_0$ in place of A in the definition of CS.

An important observation is that for the definition of CS, we do not need to specify a Riemannian metric on M as the integrand is a 3-form on a 3-dimensional manifold. Thus, any invariants constructed from the Chern-Simons functional will automatically be topological invariants of the differentiable manifold M.

In order to compute the Euler-Lagrange equations for CS, we consider variations $A + tB$, $B \in \Omega^1(\mathfrak{g})$, as in the derivation of the Yang-Mills equations. Using (3.2.13) and, with $A = A_i dx^i$, $B = B_i dx^i$,

$$\begin{aligned} \mathrm{tr}\,(A \wedge B \wedge A) &= \mathrm{tr}\,(A_k dx^k \wedge B_i dx^i \wedge A_j dx^j) \\ &= \mathrm{tr}\,(B_i dx^i \wedge A_j dx^j \wedge A_k dx^k) = \mathrm{tr}\,(B \wedge A \wedge A) \end{aligned}$$

and similary for $\mathrm{tr}\,(A \wedge A \wedge B)$, as the trace is invariant under cyclic permutations, we have

$$\frac{d}{dt} CS(A + tB)_{|t=0} = \int \mathrm{tr}\,(B \wedge dA + A \wedge dB + 2B \wedge A \wedge A)$$

and using $\int \mathrm{tr}\,(A_i dx^i \wedge \frac{\partial B_k}{\partial dx^j} dx^j \wedge dx^k) = \int \mathrm{tr}\,(B_k dx^k \wedge \frac{\partial A_i}{\partial x^j} dx^j \wedge dx^i)$,

$$\begin{aligned} &= 2 \int \mathrm{tr}\,(B \wedge (dA + A \wedge A)) \\ &= 2 \int \mathrm{tr}\,(B \wedge F_A), \end{aligned} \tag{3.2.45}$$

where $F_A = dA + A \wedge A$ is the curvature of the connection $D = d + A$. If this expression vanishes for all variations $B \in \Omega^1(\mathfrak{g})$, then $F_A = 0$. Consequently, the Euler-Lagrange equations for CS are

$$F_A = 0, \tag{3.2.46}$$

i.e. A is a flat G-connection on E.

Like the Yang-Mills equation, the equation (3.2.46) obviously remains invariant under gauge transformations. The equation (3.2.46) also arises as a reduction of the (anti)selfduality equations to 3 dimensions. Namely, suppose that M is a 3-dimensional oriented Riemannian manifold, and that we have a selfdual connection $D = d + A$ on the 4-dimensional manifold

$$N = M \times \mathbb{R}$$

with the product metric, and that $D = d + A$ can be written locally as

$$d + A_1 dx^1 + A_2 dx^2 + A_3 dx^3,$$

where x^1, x^2, x^3 are coordinates on M and where A_1, A_2, A_3 are functions of the x^1, x^2, x^3 only, and independent of the $\mathbb{R}$-direction. Thus, we assume that D is trivial in the direction of the factor $\mathbb{R}$. We denote the coordinate in that direction by x^4. We write, in our coordinates, the curvature of D as

$$F = F_{ij} dx^i \wedge dx^j = \left(\frac{\partial A_j}{\partial x^i} - \frac{\partial A_i}{\partial x^j} + [A_i, A_j] \right) dx^i \wedge dx^j.$$

Our assumption implies that

$$F_{i4} = 0 = F_{4j} \quad \text{for all } i,\ j. \tag{3.2.47}$$

On the other hand, if x^1, x^2, x^3 now are normal coordinates at the point of M under consideration, the selfduality equations become

$$F_{12} = F_{34}, \quad F_{13} = -F_{24}, \quad F_{14} = F_{23}. \tag{3.2.48}$$

(3.2.47) and (3.2.48) imply

$$F = 0$$

i.e. $D = d + A$ is flat.

Perspectives. In the work of Donaldson, detailed accounts of which can be found in [86], [66], instantons were introduced as important tools for the study of the differential topology of four-dimensional manifolds. Let M be a compact differentiable four-manifold. As explained in §2.2, one has a natural pairing

$$\begin{aligned} \Gamma : H^2(M) \times H^2(M) &\to \mathbb{R} \\ (\alpha, \beta) &\mapsto \int_M \alpha \wedge \beta. \end{aligned}$$

Γ is called intersection form of M.

Donaldson showed that if M is simply connected ($\pi_1(M) = \{1\}$) and if Γ is definite, then for a suitable basis of $H^2(M)$, Γ is represented by $\pm$ identity matrix. Since by the work of M. Freedman, there exist simply connected compact four-dimensional manifolds with definite intersection form not equivalent to $\pm$ identity matrix, it follows that such manifolds cannot carry a differentiable structure, or in other words that there exist restrictions on the topology of compact, simply connected differentiable four-dimensional manifolds that are not present for nondifferentiable ones. The crucial ingredient in the proof of Donaldson's theorem is the moduli space $\mathfrak{M}$ of instantons on a vector bundle over M with structure group SU (2) and with so-called topological charge

$$\frac{-1}{8\pi^2} \int_M \operatorname{tr}(F \wedge F) = 1$$

for the curvature F of a SU(2)-connection. As explained, the topological charge is a topological invariant of the bundle and does not depend on the choice of SU (2)-connection (it is the negative of the second Chern class of the bundle). In order to construct the moduli space of instantons, one identifies instantons that are gauge equivalent, i.e. differ only by a gauge transformation (see Theorem 3.2.1). Donaldson then showed that under the stated assumptions, $\mathfrak{M}$ is an oriented five-dimensional manifold with point singularities, at least for generic Riemannian metrics on M. Neighborhoods of the singular points are cones over complex projective space $\mathbb{CP}^2$ (see §5.1 below), and M itself is the boundary of $\mathfrak{M}$. Deleting neighborhoods of the singular points, one obtains a smooth oriented five-dimensional manifold with boundary consisting of M and some copies of $\mathbb{CP}^2$. Therefore, in the terminology of algebraic topology, M is cobordant to a union of $\mathbb{CP}^2$'s, and one knows that M then has the same intersection form as this union of $\mathbb{CP}^2$'s. As will be demonstrated in 5.1, $H^2(\mathbb{CP}^2, \mathbb{R}) = \mathbb{R}$, and the intersection form of $\mathbb{CP}^2$ is 1. These facts then imply Donaldson's theorem. The main work in the proof goes into deriving the stated properties of the moduli space $\mathfrak{M}$. In particular, one uses a theorem of Taubes on the existence of self-dual connections over four-manifolds with definite intersection form.

Donaldson then went on to use the topology and geometry of these moduli spaces to define new invariants for differentiable four-manifolds, the so-called Donaldson polynomials. These invariants greatly enhanced the understanding of the topology of differentiable four-manifolds. Subsequently, however, there has been found a simpler approach to this theory that is based on coupled equations for a section of a spinor bundle and a connection on an auxiliary bundle with an abelian gauge group, namely U(1). This will be explained in Chapter 9.

3.3 The Levi-Civita Connection

Let M be a Riemannian manifold with metric $\langle \cdot, \cdot \rangle$.

Theorem 3.3.1. *On each Riemannian manifold M, there is precisely one metric and torsion free connection ∇ (on TM). It is determined by the formula*

$$\begin{aligned}\langle \nabla_X Y, Z\rangle = \frac{1}{2}\big\{&X\langle Y, Z\rangle - Z\langle X, Y\rangle + Y\langle Z, X\rangle \\ &- \langle X, [Y, Z]\rangle + \langle Z, [X, Y]\rangle + \langle Y, [Z, X]\rangle\big\}.\end{aligned} \tag{3.3.1}$$

Definition 3.3.1. The connection ∇ determined by (3.3.1) is called the *Levi-Civita connection* of M.

In the sequel, ∇ will always denote the Levi-Civita connection.

Proof of Theorem 3.3.1. We shall first prove that each metric and torsion free connection ∇ on TM has to satisfy (3.3.1). This will imply uniqueness.

Since ∇ should be metric, it has to satisfy:

$$\begin{aligned}X\langle Y, Z\rangle &= \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle\,, \\ Y\langle Z, X\rangle &= \langle \nabla_Y Z, X\rangle + \langle Z, \nabla_Y X\rangle\,, \\ Z\langle X, Y\rangle &= \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle\,.\end{aligned}$$

Since ∇ should also be torsion free, this implies

$$\begin{aligned}&X\langle Y, Z\rangle - Z\langle X, Y\rangle + Y\langle Z, X\rangle \\ &\quad = 2\langle \nabla_X Y, Z\rangle - \langle [X, Y], Z\rangle + \langle Y, [X, Z]\rangle + \langle X, [Y, Z]\rangle\,,\end{aligned}$$

i.e. (3.3.1).

For the existence proof, for fixed X, Y, we consider the 1-form ω assigning the right hand side of (3.3.1) to each Z. $\omega(Z)$ is tensorial in Z, because for $f \in C^\infty(M)$

$$\begin{aligned}\omega(fZ) &= f\omega(Z) + \frac{1}{2}((Xf)\langle Y, Z\rangle + (Yf)\langle Z, X\rangle \\ &\qquad - (Xf)\langle Y, Z\rangle - (Yf)\langle X, Z\rangle) \\ &= f\omega(Z),\end{aligned} \tag{3.3.2}$$

and the additivity in Z is obvious.

Therefore, there exists precisely one vector field A with

$$\omega(Z) = \langle A, Z\rangle\,,$$

since $\langle \cdot, \cdot \rangle$ is nondegenerate. We thus put $\nabla_X Y := A$. It remains to show that this defines a metric and torsion free connection. Let us first verify that ∇ defines a connection: Additivity w.r.t. X and Y is clear, the tensorial behavior w.r.t. X follows as in (3.3.2), and the derivation property $\nabla_X fY = f\nabla_X Y + X(f)$ is verified in the same manner. That ∇ is metric follows from (3.3.1) by adding $\langle \nabla_X Y, Z\rangle$ and $\langle \nabla_X Z, Y\rangle$. Likewise (3.3.1) implies $\langle \nabla_X Y, Z\rangle - \langle \nabla_Y X, Z\rangle = \langle [X,Y], Z\rangle$, i.e. that ∇ is torsion free. $\square$

As in §1.4, let the metric in a local chart be given by $(g_{ij})_{i,j=1,\ldots,d}$. The Christoffel symbols of the Levi-Civita connection ∇ then are

$$\nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} = \Gamma_{ij}^k \frac{\partial}{\partial x^k}, \quad i,j = 1,\ldots,d. \tag{3.3.3}$$

From (3.1.21), we then get

$$\nabla_{\frac{\partial}{\partial x^i}} dx^j = -\Gamma_{ik}^j dx^k \tag{3.3.4}$$

Corollary 3.3.1. *For the Levi-Civita connection, we have*

$$\Gamma_{ij}^k = \frac{1}{2} g^{k\ell}(g_{i\ell,j} + g_{j\ell,i} - g_{ij,\ell}).$$

Thus, the Christoffel symbols coincide with those defined in §1.4. Likewise, the two concepts of geodesics (from §1.4 and §3.1) coincide. In particular,

$$\Gamma_{ij}^k = \Gamma_{ji}^k \quad \text{for all } i,j,k\,.$$

Proof.

$$\begin{aligned}
\Gamma_{ij}^k &= g^{k\ell}\Gamma_{ij}^m \Big\langle \frac{\partial}{\partial x^m}, \frac{\partial}{\partial x^\ell} \Big\rangle \\
&= g^{k\ell}\Big\langle \nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^\ell} \Big\rangle \\
&= \frac{1}{2} g^{k\ell}\Big\{ \frac{\partial}{\partial x^i} g_{j\ell} - \frac{\partial}{\partial x^\ell} g_{ij} + \frac{\partial}{\partial x^j} g_{i\ell} \Big\} \qquad \text{by (3.3.1)},
\end{aligned}$$

since the Lie brackets of coordinate vector fields vanish. $\square$

We now want to exhibit some formulae for the curvature tensor R of the Levi-Civita connection ∇. R is given by

$$R(X,Y)Z = \nabla_X\nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$$

(cf. (3.1.32)). In local coordinates, as in (3.1.29),

$$R\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right)\frac{\partial}{\partial x^\ell} = R_{\ell ij}^k \frac{\partial}{\partial x^k}. \tag{3.3.5}$$

We put

$$R_{k\ell ij} := g_{km} R_{\ell ij}^m,$$

i.e.

$$R_{k\ell ij} = \left\langle R\Big(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\Big)\frac{\partial}{\partial x^\ell}, \frac{\partial}{\partial x^k}\right\rangle.^1 \tag{3.3.6}$$

Lemma 3.3.1. *For vector fields X, Y, Z, W, we have*

$$R(X,Y)Z = -R(Y,X)Z, \quad \textit{i.e. } R_{k\ell ij} = -R_{k\ell ji}, \tag{3.3.7}$$

$$R(X,Y)Z + R(Y,Z)X + R(Z,X)Y = 0, \quad \textit{i.e. } R_{k\ell ij} + R_{kij\ell} + R_{kj\ell i} = 0, \tag{3.3.8}$$

$$\langle R(X,Y)Z, W\rangle = -\langle R(X,Y)W, Z\rangle, \quad \textit{i.e. } R_{k\ell ij} = -R_{\ell kij}, \tag{3.3.9}$$

$$\langle R(X,Y)Z, W\rangle = \langle R(Z,W)X, Y\rangle, \quad \textit{i.e. } R_{k\ell ij} = R_{ijk\ell}. \tag{3.3.10}$$

Proof. It suffices to verify all claims for coordinate vector fields $\frac{\partial}{\partial x^i}$. We may thus assume that all Lie brackets of X, Y, Z and W vanish. (3.3.7) then is Corollary 3.1.1. For (3.3.8), we observe

$$\begin{aligned} &R(X,Y)Z + R(Y,Z)X + R(Z,X)Y \\ &\quad = \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z + \nabla_Y\nabla_Z X \\ &\quad - \nabla_Z\nabla_Y X + \nabla_Z\nabla_X Y - \nabla_X\nabla_Z Y \\ &\quad = 0, \end{aligned}$$

since $\nabla_Y Z = \nabla_Z Y$ etc. because ∇ is torsion free.

For (3.3.9) it suffices to show $\langle R(X,Y)Z, Z\rangle = 0$ for all X, Y, Z, i.e. $R_{kkij} = 0$. This follows from Corollary 3.2.2. (3.3.10) is proved as follows:
From (3.3.7), (3.3.8)

$$\begin{aligned} \langle R(X,Y)Z, W\rangle &= -\langle R(Y,X)Z, W\rangle \\ &= \langle R(X,Z)Y, W\rangle + \langle R(Z,Y)X, W\rangle, \end{aligned} \tag{3.3.11}$$

and from (3.3.8), (3.3.9)

$$\begin{aligned} \langle R(X,Y)Z, W\rangle &= -\langle R(X,Y)W, Z\rangle \\ &= \langle R(Y,W)X, Z\rangle + \langle R(W,X)Y, Z\rangle. \end{aligned} \tag{3.3.12}$$

From (3.3.11) and (3.3.12)

$$\begin{aligned} 2\langle R(X,Y)Z, W\rangle &= \langle R(X,Z)Y, W\rangle + \langle R(Z,Y)X, W\rangle \\ &\quad + \langle R(Y,W)X, Z\rangle + \langle R(W,X)Y, Z\rangle. \end{aligned} \tag{3.3.13}$$

[1] We point out that the indices k and l appear in different orders at the two sides of (3.3.6). This somewhat unusual convention has been adopted in order to achieve as much conformity as possible with the - often conflicting - sign conventions that occur in Riemannian geometry. Differing sign conventions often lead to considerable confusion, and we hope that the convention adopted here does not add too much to that problem.

Analogously,

$$\begin{aligned}2\langle R(Z,W)X,Y\rangle &= \langle R(Z,X)W,Y\rangle + \langle R(X,W)Z,Y\rangle \\ &\quad + \langle R(W,Y)Z,X\rangle + \langle R(Y,Z)W,X\rangle \\ &= 2\langle R(X,Y)Z,W\rangle\end{aligned}$$

by applying (3.3.7) and (3.3.9) to all terms. □

Remark. (3.3.7) holds for any connection, (3.3.8) for a torsion free one, and (3.3.9) for a metric one.

(3.3.8) is called the *first Bianchi identity.*

Lemma 3.3.2 (Second Bianchi Identity).

$$\frac{\partial}{\partial x^h} R_{k\ell ij} + \frac{\partial}{\partial x^k} R_{\ell hij} + \frac{\partial}{\partial x^\ell} R_{hkij} = 0. \tag{3.3.14}$$

Proof. This is a special case of Theorem 3.1.1. We want to exhibit a different method of proof, however. Since all expressions are tensors, in order to prove (3.3.14) at a point $x_0 \in M$, we may choose arbitrary coordinates around x_0. We thus choose normal coordinates with center x_0, i.e. $g_{ij}(x_0) = \delta_{ij}, g_{ij,k}(x_0) = 0 = \Gamma^k_{ij}(x_0)$ for all i, j, k.
From (3.1.30), we obtain at x_0

$$\begin{aligned}R_{k\ell ij} &= \frac{1}{2}(g_{jk,\ell i} + g_{\ell k,ij} - g_{j\ell,ki} - g_{ik,\ell j} - g_{\ell k,ij} + g_{i\ell,kj}) \\ &= \frac{1}{2}(g_{jk,\ell i} + g_{i\ell,kj} - g_{j\ell,ki} - g_{ik,\ell j}),\end{aligned} \tag{3.3.15}$$

hence also

$$R_{k\ell ij,h} = \frac{1}{2}(g_{jk,\ell ih} + g_{i\ell,kjh} - g_{j\ell,kih} - g_{ik,\ell jh}),$$

since all other terms contain certain first derivatives of g_{ij}, hence vanish at x_0. Thus

$$\begin{aligned}R_{k\ell ij,h} + R_{\ell hij,k} + R_{hkij,\ell} &= \frac{1}{2}\big(g_{jk,\ell ih} + g_{i\ell,kjh} - g_{j\ell,kih} - g_{ik,\ell jh} \\ &\quad + g_{j\ell,hik} + g_{ih,\ell jk} - g_{jh,\ell ik} - g_{i\ell,hjk} \\ &\quad + g_{jh,ki\ell} + g_{ik,hj\ell} - g_{jk,hi\ell} - g_{ih,kj\ell}\big) \\ &= 0.\end{aligned}$$

□

Formula (3.3.15) is often useful.

Definition 3.3.2. The *sectional curvature* of the plane spanned by the (linearly independent) tangent vectors $X = \xi^i \frac{\partial}{\partial x^i}, Y = \eta^i \frac{\partial}{\partial x^i} \in T_x M$ of the Riemannian manifold M is

$$\begin{aligned} K(X \wedge Y) :&= \langle R(X,Y)Y, X\rangle \frac{1}{|X \wedge Y|^2} \\ &= \frac{R_{ijk\ell}\xi^i \eta^j \xi^k \eta^\ell}{g_{ik}g_{j\ell}(\xi^i\xi^k\eta^j\eta^\ell - \xi^i\xi^j\eta^k\eta^\ell)} \\ &= \frac{R_{ijk\ell}\xi^i \eta^j \xi^k \eta^\ell}{(g_{ik}g_{j\ell} - g_{ij}g_{k\ell})\xi^i\eta^j\xi^k\eta^\ell} \end{aligned} \tag{3.3.16}$$

($|X \wedge Y|^2 = \langle X, X\rangle\langle Y, Y\rangle - \langle X, Y\rangle^2$).

Definition 3.3.3. The Ricci curvature in the direction $X = \xi^i \frac{\partial}{\partial x^i} \in T_x M$ is

$$\mathrm{Ric}\,(X,X) = g^{j\ell}\Big\langle R\Big(X, \frac{\partial}{\partial x^j}\Big)\frac{\partial}{\partial x^\ell}, X\Big\rangle. \tag{3.3.17}$$

The Ricci tensor is

$$R_{ik} = g^{j\ell}R_{ijk\ell}. \tag{3.3.18}$$

From (3.3.10) and (3.3.18) we get the symmetry

$$R_{ik} = R_{ki}. \tag{3.3.19}$$

Finally, the scalar curvature is

$$R = g^{ik}R_{ik}.$$

Thus, the Ricci curvature is the average of the sectional curvatures of all planes in $T_x M$ containing X, and the scalar curvature is the average of the Ricci curvatures of all unit vectors, i.e. of the sectional curvatures of all planes in $T_x M$.

Lemma 3.3.3. *With $K(X,Y) := K(X \wedge Y)|X \wedge Y|^2 (= \langle R(X,Y)Y, X\rangle)$, we have*

$$\begin{aligned} \langle R(X,Y)Z, W\rangle = &+ K(X+W, Y+Z) - K(X+W, Y) - K(X+W, Z) \\ &- K(X, Y+Z) \qquad - K(W, Y+Z) + K(X,Z) + K(W,Y) \\ &- K(Y+W, X+Z) + K(Y+W, X) + K(Y+W, Z) \\ &+ K(Y, X+Z) \qquad + K(W, X+Z) - K(Y,Z) - K(W,X). \end{aligned}$$

Thus, the sectional curvature determines the whole curvature tensor.

Proof. Direct computation from Lemma 3.3.1. □

For $d = \dim M = 2$, the curvature tensor is simply given by

$$R_{ijk\ell} = K(g_{ik}g_{j\ell} - g_{ij}g_{k\ell}), \tag{3.3.20}$$

since $T_x M$ contains only one plane, namely $T_x M$ itself. The function $K = K(x)$ is called the *Gauss curvature.*

Definition 3.3.4. The Riemannian manifold M is called a space of constant sectional curvature, or a *space form* if $K(X \wedge Y) = K \equiv$ const. for all linearly independent $X, Y \in T_x M$ and all $x \in M$. A space form is called *spherical*, *flat*, or *hyperbolic*, depending on whether $K > 0, = 0, < 0$.
M is called an *Einstein manifold* if

$$R_{ik} = c g_{ik}, \quad c \equiv \text{ const.}$$

(note that c does not depend on the choice of local coordinates).

From Lemma 3.3.3 and Theorem 3.1.3, we see that the Riemannian manifolds of vanishing sectional curvature, the *flat* ones, are those that are locally isometric to Euclidean space, that is, possess local coordinates for which the coordinate vector fields $\frac{\partial}{\partial x^i}$ are parallel and by a linear transformation can then be chosen to satisfy

$$g_{ij} = \langle \frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j} \rangle \equiv \delta_{ij}.$$

Theorem 3.3.2 (Schur). *Let $d = \dim M \geq 3$. If the sectional curvature of M is constant at each point, i.e.*

$$K(X \wedge Y) = f(x) \quad \textit{for } X, Y \in T_x M,$$

then $f(x) \equiv$ const and M is a space form.

Likewise, if the Ricci curvature is constant at each point, i.e.

$$R_{ik} = c(x) g_{ik},$$

then $c(x) \equiv$ const and M is Einstein.

Proof. Let K be constant at every point, i.e. $K(X \wedge Y) = f(x)$. From Lemma 3.3.3, we obtain with $f_h = \frac{\partial}{\partial x^h}(f)$

$$R_{ijk\ell} = f(x)(g_{i\ell} g_{jk} - g_{ik} g_{j\ell}).$$

By Lemma 3.3.2, with normal coordinates at x, we obtain

$$\begin{aligned} 0 = R_{ijk\ell,h} + R_{jhk\ell,i} + R_{hik\ell,j} &= f_h(\delta_{i\ell}\delta_{jk} - \delta_{ik}\delta_{j\ell}) \\ &+ f_i(\delta_{j\ell}\delta_{hk} - \delta_{jk}\delta_{h\ell}) + f_j(\delta_{h\ell}\delta_{ik} - \delta_{hk}\delta_{i\ell}). \end{aligned}$$

Since we assume $\dim M \geq 3$, for each h, we can find h, i, j, k, ℓ with $i = \ell, j = k, h \neq i$, $h \neq j, i \neq j$. It follows that $0 = f_h$. Since this holds for all $x \in M$ and all h, we recall that M is connected by our general convention and conclude $f \equiv$ const.

The second claim follows in the same manner. □

Schur's theorem says that the isotropy of a Riemannian manifold, i.e. the property that at each point all directions are geometrically indistinguishable, implies the homogeneity, i.e. that all points are geometrically indistinguishable. In particular, a pointwise property implies a global one.

Example. We shall show that S^n has constant sectional curvature, when equipped with the metric of §1.4, induced by the ambient Euclidean metric of $\mathbb{R}^{n+1}$. The reason is simply that the group of orientation preserving isometries of $S^n, \mathrm{SO}(n+1)$, operates transitively on the set of planes in TS^n, i.e. can map any plane in TS^n into any other one. This is geometrically obvious and also easily derived formally: First of all, we have already seen that $\mathrm{SO}(n+1)$ operates transitively on S^n. It thus suffices to show that for any point p, e.g. $p = (1, 0, \ldots 0), \mathrm{SO}(n+1)$ maps any plane in T_pS^n onto any other one. The isotropy group of $p = (1, 0 \ldots 0)$ is

$$\begin{pmatrix} 1 & 0 \\ 0 & A \end{pmatrix} \quad \text{with } A \in \mathrm{SO}(n)$$

(here, the zeroes are $(1, n)$ and $(n, 1)$ matrices).

W.r.t. the Euclidean metric, T_pS^n is orthogonal to p, and $\mathrm{SO}(n + 1)$ thus operates by $X \mapsto AX$ on T_pS^n, and this operation is transitive on the 2-dimensional planes in T_pS^n. Since curvature is preserved by isometries it indeed follows that S^n has constant sectional curvature.

We want to consider the operation of the covariant derivative ∇ of Levi-Civita on tensor fields once more. For a 1-form ω and vector fields X, Y, as in §3.1

$$X(\omega(Y)) = (\nabla_X \omega)(Y) + \omega(\nabla_X Y). \tag{3.3.21}$$

Next, as in §3.1, for arbitrary tensors S, T

$$\nabla_X (S \otimes T) = \nabla_X S \otimes T + S \otimes \nabla_X T. \tag{3.3.22}$$

If e.g. S is a p-times covariant tensor, and $Y_1, \ldots, Y_p$ are vector fields,

$$\begin{aligned} (\nabla_X S)(Y_1, \ldots, Y_p) &= X(S(Y_1, \ldots, Y_p)) \\ &\quad - \sum_{i=1}^{p} S(Y_1, \ldots, Y_{i-1}, \nabla_X Y_i, Y_{i+1}, \ldots, Y_p). \end{aligned} \tag{3.3.23}$$

If in particular $S = g_{ij} dx^i \otimes dx^j =: g$ is the metric tensor, we get

$$\nabla_X g = 0 \text{ for all vectorfields } X. \tag{3.3.24}$$

This, of course, simply expresses the fact that ∇ is a metric connection.

We also want to compare ∇ with the Lie derivative of §1.9. From Theorem 1.9.4 (notations as there), we obtain

$$\begin{aligned} (L_X S)(Y_1, \ldots, Y_p) &= X(S(Y_1, \ldots, Y_p)) \\ &\quad - \sum_{i=1}^{p} S(Y_1, \ldots, Y_{i-1}, [X, Y_i], Y_{i+1}, \ldots, Y_p). \end{aligned} \tag{3.3.25}$$

Since ∇ is torsion free, $[X, Y_i] = \nabla_X Y_i - \nabla_{Y_i} X$, and with (3.3.23), we obtain

$$\begin{aligned}(L_X S)(Y_1, \ldots, Y_p) &= (\nabla_X S)(Y_1, \ldots, Y_p) \\ &\quad + \sum_{i=1}^{p} S(Y_1, \ldots, Y_{i-1}, \nabla_{Y_i} X, \ldots, Y_p)\end{aligned} \tag{3.3.26}$$

For example, for $g = g_{ij} dx^i \otimes dx^j$, we get

$$\begin{aligned}(L_X g)(Y, Z) &= g(\nabla_Y X, Z) + g(Y, \nabla_Z X) \\ &(= \langle \nabla_Y X, Z\rangle + \langle Y, \nabla_Z X\rangle \ )\end{aligned} \tag{3.3.27}$$

From (3.3.25), we obtain for a p-form ω

$$\begin{aligned}d\omega(Y_0, \ldots, Y_p) &= \sum_{i=0}^{p} (-1)^i L_{Y_i}(\omega(Y_0, \ldots, \hat{Y}_i, \ldots, Y_p)) \\ &+ \sum_{0 \le i < j \le p} (-1)^{i+j} \omega([Y_i, Y_j], Y_0, \ldots, \hat{Y}_i, \ldots, \hat{Y}_j, \ldots, Y_p),\end{aligned} \tag{3.3.28}$$

and hence

$$d\omega(Y_0, \ldots, Y_p) = \sum_{i=0}^{p} (-1)^i \nabla_{Y_i} \omega(Y_0, \ldots, Y_{i-1}, \bar{Y}_i, Y_{i+1}, \ldots, Y_p). \tag{3.3.29}$$

Lemma 3.3.4. *Let $e_1, \ldots, e_d$ $(d = \dim M)$ be a local orthonormal frame field (i.e. $e_1(y), \ldots, e_d(y)$ constitute an orthonormal basis of $T_y M$ for all y in some open subset of M). Let $\eta^1, \ldots, \eta^d$ be the dual coframe field (i.e. $\eta^j(e_i) = \delta_i^j$).*

The exterior derivative satisfies

$$d = \eta^j \wedge \nabla_{e_j} \tag{3.3.30}$$

and its adjoint (cf. Definition 2.1.1) is given by

$$d^* = -\iota(e_j) \nabla_{e_j} \tag{3.3.31}$$

where ι denotes the interior product ($\iota : \Omega^p(M) \to \Omega^{p-1}(M)$, and for $\omega \in \Omega^p(M)$, $Y_0, \ldots, Y_{p-1} \in T_y M$, we have

$$(\iota(Y_0)\omega)(Y_1, \ldots, Y_{p-1}) = \omega(Y_0, Y_1, \ldots, Y_{p-1})). \tag{3.3.32}$$

Proof. (3.3.30) is the same as (3.3.29). We are going to give a different method of proof, however, that does not use the Lie derivative and that also gives (3.3.31).

We put

$$\tilde{d} := \eta^j \wedge \nabla_{e_j}.$$

In order to show that $d = \tilde{d}$, i.e. (3.3.30), we proceed in several steps:

1) $\tilde{d}$ does not depend on the choice of the frame field $e_1, \ldots, e_d$.
Let $f_1, \ldots, f_d$ be another local frame field, with dual coframe field $\xi^1, \ldots, \xi^d$. Then
$$f_j = \alpha_j^k e_k \tag{3.3.33}$$
for some coefficients α_j^k, and
$$\xi^j = \beta_k^j \eta^k,$$
with
$$\alpha_j^k \beta_\ell^j = \delta_\ell^k$$
from the standard transformation rules.

Consequently
$$\begin{aligned} \xi^j \wedge \nabla_{f_j} &= \beta_\ell^j \eta^\ell \wedge \nabla_{\alpha_j^k e_k} \\ &= \alpha_j^k \beta_\ell^i \eta^\ell \wedge \nabla_{e_k} \\ &= \eta^k \wedge \nabla_{e_k}. \end{aligned}$$
$\tilde{d}$ is independent of the choice of frame field, indeed.

2) Since d does not depend on a choice of frame field either (see Lemma 1.8.2 and Corollary 1.8.1), it therefore suffices to check (3.3.30) for one particular choice of frame field. The independence on the choice of frame field of both sides of (3.3.30) will then imply that (3.3.30) will hold for any choice of frame field.

3) We now choose normal coordinates $(x^1, \ldots, x^d)$ centered at $x_0 \in M$ (Corollary 1.4.2) and the frame field $e_j = \frac{\partial}{\partial x^j}$ which is orthonormal at x_0. Then $\eta^k = dx^k$. We are now going to verify (3.3.30) at the point x_0 for those choices of e_j and η^k. By 2), and since $x_0 \in M$ is arbitrary, that suffices.
At x_0, the center of our normal coordinates, we have for all j, k
$$\begin{aligned} \nabla_{\frac{\partial}{\partial x^j}} \frac{\partial}{\partial x^k} &= 0, \\ \nabla_{\frac{\partial}{\partial x^j}} dx^k &= 0 \end{aligned} \tag{3.3.34}$$
(Theorem 1.4.4 and Corollary 3.3.1).
Since d and $\tilde{d}$ are both linear operators, it also suffices to verify the claim on forms of the type $\varphi(y)dx^{i_1} \wedge \ldots \wedge dx^{i_p}$. Renumbering indices, it even suffices to consider the form
$$\varphi(y)dx^1 \wedge \ldots \wedge dx^p.$$
Using (3.3.34), we have at x_0
$$\begin{aligned} \tilde{d}(\varphi(x_0)dx^1 \wedge \ldots \wedge dx^p) &= dx^j \wedge \big(\nabla_{\frac{\partial}{\partial x^j}} \varphi\big)(x_0)dx^1 \wedge \ldots \wedge dx^p \\ &= \frac{\partial \varphi}{\partial x^j} dx^j \wedge dx^1 \wedge \ldots dx^p \\ &= d(\varphi(x_0)dx^1 \wedge \ldots \wedge dx^p), \end{aligned}$$

which is the desired formula.

In order to verify (3.3.31), we use the same method.
We put

$$\tilde{d}^* = -\iota(e_j)\nabla_{e_j}.$$

1) Independence of the choice of frame field:

   Since both $(f_j)_{j=1,\ldots,d}$ and $(e_k)_{k=1,\ldots,d}$ constitute an orthonormal basis of T_yM, the matrix $(\alpha_j^k)_{j,k=1,\ldots,d}$ of (3.3.33) is orthogonal, i.e.

$$\alpha_j^k \alpha_j^\ell = \delta^{k\ell}.$$

   Thus

$$-\iota(f_j)\nabla_{f_j} = -\iota(\alpha_j^k e_k)\nabla_{\alpha_j^\ell e_\ell} = -\alpha_j^k\alpha_j^\ell \iota(e_k)\nabla_{e_\ell} = -\iota(e_k)\nabla_{e_k}. \tag{3.3.35}$$

2) By 1), it again suffices to verify (3.3.31) for one particular choice of frame field.

3) We choose normal coordinates centered at x_0 as before, and $e_j = \frac{\partial}{\partial x^j}, \eta^k = dx^k$. Then again at x_0

$$\begin{aligned}\tilde{d}^*(\varphi(x_0)dx^1 \wedge \ldots \wedge dx^p) &= -\iota(\frac{\partial}{\partial x^j})(\frac{\partial}{\partial x^j}\varphi)(x_0)dx^1 \wedge \ldots \wedge dx^p \\ &= (-1)^j(\frac{\partial}{\partial x^j}\varphi)(x_0)dx^1 \wedge \ldots \wedge \widehat{dx^j} \wedge \ldots \wedge dx^p,\end{aligned}$$

   where in the last expression, j only runs from 1 to p. We compare this with

$$\begin{aligned}&d^*(\varphi(x_0)dx^1 \wedge \ldots \wedge dx^p)\\ &= (-1)^{d(p+1)+1} * d * (\varphi(x_0)dx^1 \ldots \wedge dx^p) \quad \text{by Lemma 2.1.4}\\ &= (-1)^{d(p+1)+1} * d(\varphi(x_0)dx^{p+1} \wedge \ldots \wedge dx^d) \quad \text{by definition of } *\\ &= (-1)^{d(p+1)+1} * dx^j \wedge \big(\nabla_{\frac{\partial}{\partial x^j}}\varphi\big)(x_0)dx^{p+1} \wedge \ldots \wedge dx^d\\ &\quad \text{by (3.3.30) and (3.3.34)}\\ &= (-1)^{d(p+1)+1}(-1)^{(p-1)(d-p+1)+(p-j)}\nabla_{\frac{\partial}{\partial x^j}}\varphi dx^1 \wedge \ldots \wedge \widehat{dx^j} \wedge \ldots \wedge dx^d\\ &\quad \text{by definition of } *\\ &= (-1)^j \nabla_{\frac{\partial}{\partial x^j}}\varphi dx^1 \wedge \ldots \wedge \widehat{dx^j} \wedge \ldots \wedge dx^d.\end{aligned}$$

   Thus, $d^* = \tilde{d}^*$.

□

Remarks.

1. For (3.3.30), we do not need to assume that the frame field is orthonormal. It suffices that the vectors $e_1(y), \ldots, e_d(y)$ constitute a basis of T_yM. Of course, this is to be expected from the fact that the definition of the exterior derivative does not involve a choice of metric. By way of contrast, in (3.3.31) the e_j have to be orthonormal, and of course, the definition of d^* does depend on the choice of a metric.

2. We may now give a proof of formula (2.1.32):

   We recall from formula (3.3.35) that we have for arbitrary (not necessarily orthonormal) bases of T_yM with
$$f_j = \alpha_j^k e_k$$
   that
$$-\iota(f_j)\nabla_{f_j} = -\alpha_j^k \alpha_j^\ell \iota(e_k)\nabla_{e_k}. \tag{3.3.36}$$
   We now choose $(f_j)_{j=1,\ldots,d}$ to be orthonormal and $e_k = \frac{\partial}{\partial x^k}$ w.r.t. local coordinates. Then of course
$$\langle e_k, e_\ell\rangle = \langle \frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^\ell}\rangle = g_{k\ell}$$
   and hence
$$\delta_{ij} = \langle f_i, f_j\rangle = \langle \alpha_i^k e_k, \alpha_j^\ell e_\ell\rangle = \alpha_i^k \alpha_j^\ell g_{k\ell},$$
   and thus
$$\alpha_i^k \alpha_j^\ell = \delta_{ij} g^{k\ell}. \tag{3.3.37}$$
   From (3.3.31), (3.3.36), (3.3.37) (since (f_j) is orthonormal)
$$d^* = -g^{k\ell}\iota(\frac{\partial}{\partial x^k})\nabla_{\frac{\partial}{\partial x^\ell}}\,. \tag{3.3.38}$$
   Then for $\alpha = \alpha_{i_1\ldots i_p} dx^{i_1} \wedge \ldots \wedge dx^{i_p}$
$$d^*\alpha = -g^{k\ell}\iota\left(\frac{\partial}{\partial x^k}\right)\left(\frac{\partial \alpha_{i_1\ldots i_p}}{\partial x^\ell} - \alpha_{i_1\ldots i_p}\Gamma^j_{\ell m} dx^m \wedge dx^{i_1} \wedge \widehat{dx^{i_j}} \wedge \ldots \wedge dx^{i_p}\right), \tag{3.3.39}$$
   using (3.3.4) and thus
$$d^*\alpha_{i_1\ldots i_{p-1}} = -g^{k\ell}\left(\frac{\partial \alpha_{k i_1\ldots i_{p-1}}}{\partial x^\ell} - \Gamma^j_{\ell k}\alpha_{j i_1\ldots i_{p-1}}\right),$$
   which is (2.1.32).

We next want to express the Laplace-Beltrami operator Δ (cf. Definition 2.1.2) in terms of the Levi-Civita connection ∇. For that purpose, we define the second covariant derivative as
$$\nabla^2_{XY} = \nabla_X \nabla_Y - \nabla_{\nabla_X Y}. \tag{3.3.40}$$

Theorem 3.3.3 (Weitzenböck Formula). *Let $e_1, \dots, e_d$ ($d = \dim M$) be a local orthonormal frame field as in Lemma 3.3.4, with the dual coframe field $\eta^1, \dots, \eta^d$. Then the Laplace-Beltrami operator acting on p-forms ($p = 0, 1, \dots, d$) is given by*

$$\Delta = -\nabla^2_{e_i e_i} - \eta^i \wedge \iota(e_j) R(e_i, e_j). \tag{3.3.41}$$

Proof. We shall use invariance arguments as in the proof of Lemma 3.3.4. The right hand side of (3.3.41) is independent of the choice of our orthonormal frame field v_i. Therefore, if we want to verify (3.3.41) at an arbitrary point $x_0 \in M$, we choose normal coordinates centered at x_0 and put at x_0,

$$e_i = \frac{\partial}{\partial x^i}.$$

Then, always at x_0,

$$\nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} = 0,$$

hence

$$\nabla^2_{e_i e_i} = \nabla_{e_i} \nabla_{e_i} \tag{3.3.42}$$

and also $[\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}] = 0$, hence

$$R(e_i, e_j) = \nabla_{e_i} \nabla_{e_j} - \nabla_{e_j} \nabla_{e_i} \quad \text{(cf. (3.1.32))}. \tag{3.3.43}$$

Using Lemma 3.3.4, we then have at x_0

$$\begin{aligned} d^* d &= -\iota(e_j) \nabla_{e_j} (\eta^i \wedge \nabla_{e_i}) \\ &= -\iota(e_j)(\eta^i \wedge \nabla_{e_j} \nabla_{e_i}) \qquad \text{since } \nabla_{e_j} \eta^i = 0 \text{ at } x_0 \\ &= -\nabla_{e_k} \nabla_{e_k} + \eta^i \wedge \iota(e_j) \nabla_{e_j} \nabla_{e_i} \, . \end{aligned} \tag{3.3.44}$$

Next,

$$\begin{aligned} d d^* &= -\eta^i \wedge \nabla_{e_i} (\iota(e_j) \nabla_{e_j}) \\ &= -\eta^i \wedge \iota(e_j) \nabla_{e_i} \nabla_{e_j}, \end{aligned} \tag{3.3.45}$$

since at $x_0, \iota(e_j) \nabla_{e_i} = \nabla_{e_i} \iota(e_j)$ because of $\nabla_{e_k} \eta^j = 0$.

(3.3.42) – (3.3.45) imply (3.3.41). □

Remark. On functions, i.e. 0-forms f, we have

$$R(e_i, e_j) f = f R(e_i, e_j) 1 = 0$$

because of the tensorial property of R.

Hence for a function $f : M \to \mathbb{R}$,

$$\Delta f = -\nabla^2_{e_i e_i} f. \tag{3.3.46}$$

Definition 3.3.5. The *Hessian* of a differentiable function $f : M \to \mathbb{R}$ on a Riemannian manifold M is

$$\nabla df.$$

We have $df = \frac{\partial f}{\partial x^i} dx^i$ in local coordinates, hence

$$\nabla_{\frac{\partial}{\partial x^j}} df = \frac{\partial^2 f}{\partial x^i \partial x^j} dx^i - \frac{\partial f}{\partial x^i} \Gamma^i_{jk} dx^k,$$

i.e.

$$\nabla df = \left(\frac{\partial^2 f}{\partial x^i \partial x^j} - \frac{\partial f}{\partial x^k} \Gamma^k_{ij} \right) dx^i \otimes dx^j. \tag{3.3.47}$$

We also have

$$\nabla df(X, Y) = \langle \nabla_X \operatorname{grad} f, Y \rangle, \tag{3.3.48}$$

since $Y(f) = \langle \operatorname{grad} f, Y \rangle$ and thus

$$\begin{aligned} X(Y(f)) &= X \langle \operatorname{grad} f, Y \rangle \\ &= \langle \nabla_X \operatorname{grad} f, Y \rangle + \langle \operatorname{grad} f, \nabla_X Y \rangle \\ &= \langle \nabla_X \operatorname{grad} f, Y \rangle + (\nabla_X Y)(f), \end{aligned}$$

and applying (3.3.47) to X and Y yields

$$\nabla df(X, Y) = X(Y(f)) - (\nabla_X Y)(f). \tag{3.3.49}$$

This formula can be given the following geometric interpretation: Let $X \in T_p M$ and take a geodesic $c : [0, \epsilon) \to M$ (for some $\epsilon > 0$) with $c(0) = p$, $\dot{c}(0) = X$. Then at p

$$\nabla df(X, X) = \frac{d^2}{dt^2} f(c(t))_{|t=0}. \tag{3.3.50}$$

Namely

$$\begin{aligned} X(X(f)) &= \dot{c} \langle \operatorname{grad} f(p), \dot{c} \rangle \\ &= \dot{c} \left(\frac{d}{dt} f(c(t))_{|t=0} \right) \\ &= \frac{d^2}{dt^2} f(c(t))_{|t=0} \end{aligned}$$

and

$$\nabla_{\dot{c}} \dot{c} = 0,$$

since c is geodesic (see (3.1.34) and Corollary 3.3.1 so that (3.3.50) follows from (3.3.49).

Definition 3.3.6. The differentiable function $f : M \to \mathbb{R}$ is called (strictly) convex if the Hessian ∇df is positive semidefinite (definite).

Theorem 3.3.4. *Let M be a compact Riemannian manifold with metric tensor g. There then exists a constant c (depending on the geometry of M) such that for any (smooth) vector field X on M*

$$\begin{aligned}\int_M \|\nabla X\|^2\, d\mathrm{Vol} &+ \int |\mathrm{div}\, X|^2\, d\mathrm{Vol} \\ &\leq c\left(\int_M \|X\|^2\, d\mathrm{Vol} + \int_M \|L_X g\|^2\, d\mathrm{Vol}\right),\end{aligned} \tag{3.3.51}$$

where $L_X g$ is the Lie derivative of g in the direction of X (see (1.9.20)*).*

Proof. In local coordinates, by (1.9.20),

$$L_X g = \Big(g_{kj}\frac{\partial X^k}{\partial x^i} + g_{ik}\frac{\partial X^k}{\partial x^j} + g_{ij,k}X^k\Big)dx^i \otimes dx^j$$

Thus,

$$\|L_X g\|^2 = 2\, g_{km} g^{i\ell}\frac{\partial X^k}{\partial x^i}\frac{\partial X^m}{\partial x^\ell} + 2\,\frac{\partial X^k}{\partial x^i}\frac{\partial X^i}{\partial x^k} + P(X, \nabla X), \tag{3.3.52}$$

where, here and in the sequel, $P(X, \nabla X)$ stands for any terms that are bounded by

$$\mathrm{const}\cdot(\|X\|\,\|\nabla X\| + \|X\|^2).$$

Now

$$\frac{\partial X^k}{\partial x^i}\frac{\partial X^i}{\partial x^k} = \frac{\partial}{\partial x^i}\Big(X^k\frac{\partial X^i}{\partial x^k} - X^i\frac{\partial X^k}{\partial x^k}\Big) + \frac{\partial X^k}{\partial x^k}\frac{\partial X^i}{\partial x^i}. \tag{3.3.53}$$

Also

$$\|\nabla X\|^2 = g_{km} g^{i\ell}\frac{\partial X^k}{\partial x^i}\frac{\partial X^m}{\partial x^\ell} + P(X, \nabla X), \tag{3.3.54}$$

$$|\mathrm{div}\, X|^2 = \frac{\partial X^k}{\partial x^k}\frac{\partial X^i}{\partial x^i} + P(X, \nabla X). \tag{3.3.55}$$

From (3.3.52) – (3.3.55),

$$\int \|\nabla X\|^2 + \int |\mathrm{div}\, X|^2 \leq \frac{1}{2}\int \|L_X g\|^2 + \int P(X, \nabla X). \tag{3.3.56}$$

Using the inequality

$$\|X\|\,\|\nabla X\| \leq \frac{\delta}{2}\|\nabla X\|^2 + \frac{2}{\delta}\|X\|^2 \quad \text{for any } \delta > 0,$$

we can estimate

$$\int P(X, \nabla X) \leq \varepsilon\int \|\nabla X\|^2 + c(\varepsilon)\int \|X\|^2, \tag{3.3.57}$$

where $c(\varepsilon)$ depends on $\varepsilon > 0$ and on the constants involved in the terms $P(X, \nabla X)$, i.e. on bounds for the metric tensor g and its first derivatives. Using (3.3.57) with $\varepsilon = \frac{1}{2}$ in (3.3.56), we easily obtain (3.3.51). □

Corollary 3.3.2. *Let M be a compact Riemannian manifold. Then the vector space of Killing fields (cf. Definition 1.9.7) on M is finite dimensional.*

Proof. By definition of a Killing field X,

$$L_X g = 0.$$

Inserting this into (3.3.51), we obtain

$$\int_M \|\nabla X\|^2 + \int_M |\operatorname{div} X|^2 \leq c \int_M \|X\|^2. \tag{3.3.58}$$

If $(X_n)_{n \in \mathbb{N}}$ then is a sequence of Killing fields with $\int \|X_n\|^2 = 1$ for all n, we bound their Sobolev $H^{1,2}$-norm by (3.3.58), apply Rellich's theorem (Theorem A.1.8 in the Appendix), and conclude that the X_n contain a subsequence that converges in L^2 . This implies that the space of Killing fields is a finite dimensional subspace of the space of L^2-vector fields on M. □

Perspectives. The sectional curvature as an invariant of a Riemannian metric was introduced by Riemann in his habilitation address (quoted in the Perspectives on §1.1). The tensor calculus for Riemannian manifolds was developed by Christoffel, Ricci, and others. It also played an important role in the development of Einstein's theory of general relativity.

Levi-Civita introduced the notion of parallel transport for a Riemannian manifold. (Similar concepts were also developed by other mathematicians at about the time.) The concept was expanded and clarified by Weyl, see [261]. For a historical account, see also [228, 229].

Space forms are quotients of the sphere S^n, Euclidean space $\mathbb{R}^n$, or hyperbolic space H^n (see §4.4). They can be classified, cf. Wolf[266].

Einstein manifolds form an important class of Riemannian manifolds. Every two dimensional manifold carries a metric of constant curvature, i.e. is a space form, by the uniformization theorem. In higher dimensions, some necessary topological conditions have been found for the existence of Einstein metrics. The question which manifolds admit Einstein metrics is far from being solved. Even in three dimensions where a metric is Einstein if and only if it has constant sectional curvature, the question is not yet fully solved. See however [251], [252]. A comprehensive account of Einstein manifolds is given in the monograph [19].

Theorem 3.3.4 is a Riemannian version of Korn's inequality. This result, and the proof of Corollary 3.3.2 given here, are taken from [51]. One may also identify the terms $P(X, \nabla X)$ in (3.3.52) in terms of the Ricci curvature to obtain the Bochner–Yano formula, see [25].

3.4 Connections for Spin Structures and the Dirac Operator

Let ∇ be the Levi-Civita connection of the oriented manifold M of dimension n, according to Theorem 3.3.1. By Lemma 3.2.2, it admits a local decomposition

$$\nabla = d + A \tag{3.4.1}$$

with $A \in \Omega^1(\mathrm{Ad}\, TM)$, i.e. a one form with values in $\mathfrak{so}(n)$ that transforms according to (3.1.17). Conversely, given a vector bundle E with bundle metric $\langle \cdot, \cdot \rangle$ on which $\mathrm{SO}(n)$ acts by isometries, and a one form A with values in $\mathfrak{so}(n)$ that transforms by (3.1.17), then (3.4.1) can be used to define a metric connection on E according to the discussion in §3.2. Consequently, for any such bundle E on which $\mathrm{SO}(n)$ acts with the same transition functions as for the action on TM, the Levi-Civita connection induces a connection. Applying this observation to the Clifford bundles $\mathrm{Cl}(P)$ and $\mathrm{Cl}^{\mathbb{C}}(P)$ from Definition 1.11.11, we conclude that the Levi-Civita connection induces a connection, again denoted by ∇, on each Clifford bundle.

Lemma 3.4.1. *For smooth sections μ, ν of $\mathrm{Cl}(P)$ (or $\mathrm{Cl}^{\mathbb{C}}(P)$) we have*

$$\nabla(\mu\nu) = \nabla(\mu)\nu + \mu\nabla(\nu). \tag{3.4.2}$$

Proof. It is clear that the exterior derivative d satisfies the product rule, and we recall that A in the decomposition (3.4.1) is in $\mathfrak{so}(n)$, i.e. acts by the infinitesimal version of the $\mathrm{SO}(n)$ action on $\mathrm{Cl}(P)$. Since this $\mathrm{SO}(n)$ action extends to the one on the tangent bundle TM, $B \in \mathrm{SO}(n)$ acts via

$$B(\mu\nu) = B(\mu)B(\nu), \tag{3.4.3}$$

and differentiating (3.4.3) yields the product rule for A. □

Corollary 3.4.1. *∇ leaves the decomposition of the Clifford bundles into elements of even and odd degree invariant.*

Proof. It is clear from the definition, that subbundles of degree 0 and 1 are preserved, and the claim then easily follows from (3.4.2). □

Since the chirality operator Γ of Definition 1.11.3 defines a section of $\mathrm{Cl}^{\mathbb{C}}(P)$ that is invariant under the action of $\mathrm{SO}(n)$, it must be covariantly constant, i.e.

Lemma 3.4.2.

$$\nabla(\Gamma) = 0.$$

□

Similarly, since the Lie algebra $\mathfrak{spin}(n)$ can be identified with $\mathfrak{so}(n)$ (see Lemma 1.11.2, in the case of a spin structure $\widetilde{P}$ over M (cf. Definition 1.11.7), we may use the same procedure to obtain induced connections on the associated spinor bundles. We denote them again by ∇. The action of $\mathrm{Cl}^{\mathbb{C}}(P)$ on the spinor bundle $\mathbb{S}_n$ via Clifford multiplication on each fiber (see (1.11.27)) is compatible with these connections; more precisely

Lemma 3.4.3. *For smooth sections μ of $\mathrm{Cl}^{\mathbb{C}}(P)$, σ of $\mathbb{S}_n$*

$$\nabla(\mu\sigma) = \nabla(\mu)\sigma + \mu\nabla(\sigma) \tag{3.4.4}$$

(where the products of course are given by Clifford multiplication).

Proof. Similar to the one of Lemma 3.4.1. □

Suppose that in a local trivialization of TM, A from (3.4.1) is given by the (skew symmetric) matrix Ω_{ij}. We write

$$A = \sum_{i<j} \Omega_{ij} e_i \wedge e_j,$$

where $e_i \wedge e_j$ denotes the matrix with (-1) at the place (i,j), $+1$ at (j,i), and 0 otherwise. According to Lemma 1.11.3, $e_i \wedge e_j$ in $\mathfrak{so}(n)$ corresponds to $\frac{1}{2}e_ie_j$ in $\mathfrak{spin}(n)$. Thus, the connection on the spinor bundle w.r.t. the induced local trivialization is given by

$$d + \frac{1}{2}\sum_{i<j} \Omega_{ij} e_i e_j. \tag{3.4.5}$$

Here, e_ie_j of course operates by Clifford multiplication on spinors.

We next consider the case of a spin^c structure $\widetilde{P}$ over M (cf. Definition 1.11.9). Here, the Levi-Civita connection ∇ does not suffice to determine a unique connection on bundles on which Spin^c acts. Namely, since the Lie algebra of Spin$^c(n)$ is $\mathfrak{spin}(n) \oplus \mathfrak{u}(1)$, we need to specify in addition a connection on the $\mathfrak{u}(1)$ part, i.e. on the determinant line bundle L of the spin^c structure (Definition 1.11.10). We identify the Lie algebra $\mathfrak{u}(1)$ of U(1) with $i\mathbb{R}$, and thus, a unitary connection on L is locally represented by a function iA with imaginary values. Given the Levi-Civita connection and such a connection on L, we represent the induced spin^c connection ∇_A locally as

$$\nabla_A = d + \frac{1}{2}\left(\sum_{i<j} \Omega_{ij} e_i e_j + iA\right) \tag{3.4.6}$$

as in (3.4.5).

Definition 3.4.1.

(i) Let $\widetilde{P} \to M$ be a spin structure on the oriented Riemannian manifold M, with Levi-Civita connection ∇ as explained above. The *Dirac operator* $\not{\partial}$ operates on sections σ of the spinor bundle $\mathbb{S}_n$ via

$$\not{\partial}\sigma(x) = e_i \nabla_{e_i}(\sigma)(x) \tag{3.4.7}$$

where e_i, $i = 1, \dots, n$, is an orthonormal basis of T_xM $(x \in M)$. The product on the right hand side of (3.4.7) is given by Clifford multiplication.

(ii) Let $\widetilde{P}^c \to M$ be a spin^c structure on M, and let A represent a unitary connection on the associated determinant line bundle L. The *Dirac operator* $\not{\partial}_A$ operating on $\mathbb{S}_n$ is given by

$$\not{\partial}_A\sigma(x) = e_i \nabla_{A,e_i}(\sigma)(x).$$

Example. We consider the case of $\mathbb{R}^2$ with coordinates x, y. Recalling the discussion in 1.11, the spinor space then is $\mathbb{C}^2$, and the vectors e_1 and e_2 act on spinors via

$$\gamma(e_1) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \gamma(e_2) = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}.$$

Writing a spinor field $\sigma : \mathbb{R}^2 \to \mathbb{C}^2$ in components as $\begin{pmatrix} \sigma^1 \\ \sigma^2 \end{pmatrix}$, we then have

$$\not{\partial}\sigma = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \frac{\partial\sigma^1}{\partial x} \\ \frac{\partial\sigma^2}{\partial x} \end{pmatrix} + \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \begin{pmatrix} \frac{\partial\sigma^1}{\partial y} \\ \frac{\partial\sigma^2}{\partial y} \end{pmatrix} = 2 \begin{pmatrix} \frac{\partial\sigma^2}{\partial \bar{z}} \\ -\frac{\partial\sigma^1}{\partial z}. \end{pmatrix} \tag{3.4.8}$$

Thus, in this case, the Dirac operator is simply the Cauchy-Riemann operator.

Remark. Since V also operates on the Clifford space $\mathrm{Cl}(V)$, $= \Lambda^*(V)$ as a vector space, we can also define a Dirac operator on the Clifford bundle instead of the spinor bundle, namely,

$$D := d + d^*, \quad \text{with } d = \eta_j \wedge \nabla_{e_j},\ d^* = -\iota(e_j)\nabla_{e_j} \tag{3.4.9}$$

as in Lemma 3.3.4 that then satisfies

$$D^2 = \Delta, \quad \text{the Laplacian.} \tag{3.4.10}$$

These two Dirac operators should not be confused.

Lemma 3.4.4. *The Dirac operators $\not{\partial}$ and $\not{\partial}_A$ do not depend on the choice of an orthonormal frame e_i.*

Proof. Any other such frame f_j, $j = 1, \dots, n$, can be obtained as

$$f_j = b_{ij} e_i$$

for some $B = (b_{ij})_{j,i=1,\dots,n} \in \mathrm{O}(n)$. Then

$$\begin{aligned} f_j \nabla_{f_j} &= b_{ji} e_i \nabla_{b_{jk} e_k} \\ &= b_{ji} b_{jk} e_i \nabla_{e_k} \\ &= \delta_{ik} e_i \nabla_{e_k} \qquad \text{since } B \in \mathrm{O}(n) \\ &= e_i \nabla_{e_i}, \end{aligned}$$

which is the invariance of $\partial\!\!\!/$, and the same computation works for $\partial\!\!\!/_A$. □

A more abstract way to express the Dirac operator is the following. Let

$$\begin{aligned} cl : TM \otimes \mathbb{S} &\to \mathbb{S} \\ v \otimes \sigma &\to v \cdot \sigma \end{aligned}$$

denote the Clifford multiplication. Thus, tangent vectors of M act on spinors by Clifford multiplication. Denote the space of smooth sections of a vector bundle E over M by $\Gamma(E)$. Then

$$\partial\!\!\!/ = cl \circ \nabla : \Gamma(\mathbb{S}) \xrightarrow{\nabla} \Gamma(T^*M \otimes \mathbb{S}) \cong \Gamma(TM \otimes \mathbb{S}) \xrightarrow{cl} \Gamma(\mathbb{S})$$

where the identification between $\Gamma(T^*M \otimes \mathbb{S})$ and $\Gamma(TM \otimes \mathbb{S})$ uses the Riemannian metric of M.

Lemma 3.4.5. *Let M be even dimensional, and let $\mathbb{S}_n^\pm$ be the half spinor bundles for a spin or a* spin^c *structure on M. Then the Dirac operator $\partial\!\!\!/$ ($\partial\!\!\!/_A$) maps $\Gamma(\mathbb{S}_n^\pm)$ to $\Gamma(\mathbb{S}_n^\mp)$.*

Proof. By Corollary 3.4.1, ∇, and similarly ∇_A, leaves the decomposition into sections of even and odd degree invariant, while Clifford multiplication by e_i interchanges sections of even and odd degree. □

We recall from Corollary 1.11.3 that on the bundle $\mathbb{S}_n$ of spinors, we have a pointwise Hermitian product $\langle \cdot, \cdot \rangle$ (invariant under $\mathrm{Spin}(n)$). We suppose now that M is compact. We may then form the associated L^2 product

$$(\sigma_1, \sigma_2) := \int_M \langle \sigma_1(x), \sigma_2(x) \rangle * (1)$$

where $*(1)$ is the volume form of M (see (2.1.19)).

Lemma 3.4.6. *Let M be a compact Riemannian manifold with a spin structure. Then the corresponding Dirac operator $\partial\!\!\!/$ is formally selfadjoint, i.e.*

$$(\partial\!\!\!/ \sigma_1, \sigma_2) = (\sigma_1, \partial\!\!\!/ \sigma_2) \tag{3.4.11}$$

for all spinor fields σ_1, σ_2.

Proof. Let $x \in M$, and choose normal coordinates centered at x. With $e_i := \frac{\partial}{\partial x^i}$, we then have at x

$$\nabla_{e_i}(e_j) = 0 \quad \text{for all } i, j \text{ (cf. Theorem 1.4.4 and Corollary 3.3.1).} \tag{3.4.12}$$

We then have

$$\begin{aligned}\langle \partial\!\!\!/ \sigma_1(x), \sigma_2(x)\rangle &= \langle e_i \nabla_{e_i} \sigma_1(x), \sigma_2(x)\rangle \\ &= -\langle \nabla_{e_i}\sigma_1(x), e_i\sigma_2(x)\rangle\end{aligned}$$

since $\langle \cdot, \cdot \rangle$ is invariant under Clifford multiplication by the unit vector e_i

$$= -e_i\langle \sigma_1(x), e_i\sigma_2(x)\rangle + \langle \sigma_1(x), \nabla_{e_i}(e_i\sigma_2)(x)\rangle$$

since ∇ is a metric connection

$$\begin{aligned}&= -e_i\langle \sigma_1(x), e_i\sigma_2(x)\rangle + \langle \sigma_1(x), e_i\nabla_{e_i}\sigma_2(x)\rangle \text{ by (3.4.12)} \\ &= -e_i\langle \sigma_1(x), e_i\sigma_2(x)\rangle + \langle \sigma_1(x), \partial\!\!\!/ \sigma_2(x)\rangle.\end{aligned}$$

We now consider $V^i = \langle \sigma_1(x), e_i\sigma_2(x)\rangle$ as the i^{th} component of a vector field V (in fact V is a complexified vector field, i.e. a section of $TM \otimes \mathbb{C}$). The preceding formula then becomes

$$\langle \partial\!\!\!/ \sigma_1(x), \sigma_2(x)\rangle = -\operatorname{div} V(x) + \langle \sigma_1(x), \partial\!\!\!/ \sigma_2(x)\rangle. \tag{3.4.13}$$

Since all terms in (3.4.13) are independent of the particular choice of coordinates, they continue to hold regardless of whether (3.4.12) is satisfied. (This point has been discussed in §3.3, e.g. in the derivation of Lemma 3.3.4, but since this an important computational trick, we repeat it here). Since

$$\int_M \operatorname{div} V(x) * (1) = 0$$

by the Gauss theorem (see the discussion in §2.1), (3.4.11) follows by integrating (3.4.13). □

Corollary 3.4.2. *On a compact spin manifold M, $\partial\!\!\!/ \sigma = 0$ for a spinor field iff $\partial\!\!\!/$ $^2\sigma = 0$.*

Proof. This follows from

$$\left(\partial\!\!\!/^2 \sigma, \sigma\right) = (\partial\!\!\!/ \sigma, \partial\!\!\!/ \sigma)$$

by Lemma 3.4.6. □

Definition 3.4.2. A spinor field satisfying $\partial\!\!\!/ \sigma = 0$ is called *harmonic*.

We shall now introduce another type of spinors that will be used in 9.3 below. Let $M = S^2$, the two-dimensional sphere. As always, we choose a local orthonormal basis $e_i, i = 1, 2$ of the tangent bundle TS^2. S^2 carries a unique spin structure (see 1.11); let $\mathcal{S}S^2$ be the corresponding spinor bundle. The spinor σ is called a twistor spinor if

$$\nabla_V \sigma + \frac{1}{2} V \cdot \partial\!\!\!/ \sigma = 0 \tag{3.4.14}$$

for any vector field V on S^2.

We now come to Weitzenböck formulas that constitute analogues of Theorem 3.3.3.

Theorem 3.4.1. *Let M be a spin manifold with a local orthonormal frame field $e_1, \ldots, e_n$ (as in Lemma 3.3.4, $n = \dim M$). Then the Dirac operator $\partial\!\!\!/$ satisfies*

$$\partial\!\!\!/^2 = -\nabla^2_{e_i e_i} + \frac{1}{4} R, \tag{3.4.15}$$

where R is the scalar curvature of M.

Proof. As in (3.4.12), we assume

$$\nabla_{e_i}(e_j) = 0 \quad \text{at the point } x \in M \text{ under consideration, for all } i, j, \tag{3.4.16}$$

as well as

$$[e_i, e_j] = 0 \quad \text{since this holds for all coordinate vector fields } e_i = \frac{\partial}{\partial x^i}. \tag{3.4.17}$$

We compute, for a spinor field σ, at x,

$$\begin{aligned} \partial\!\!\!/^2 \sigma &= e_j \nabla_{e_j} ((e_i \nabla_{e_i}) \sigma) \\ &= e_j e_i \nabla_{e_j} \nabla_{e_i} \sigma \quad \text{by (3.4.16)} \\ &= -\nabla_{e_i} \nabla_{e_i} \sigma + \sum_{i<j} e_j e_i \left(\nabla_{e_j} \nabla_{e_i} - \nabla_{e_i} \nabla_{e_j} \right) \sigma \text{ , because of } e_j e_i + e_i e_j = -2\delta_{ij}, \\ &= -\nabla^2_{e_i e_i} \sigma + \sum_{i<j} e_j e_i R(e_j, e_i) \sigma \end{aligned} \tag{3.4.18}$$

by (3.3.42) and where $R(\cdot, \cdot)$ is the curvature tensor of the Levi-Civita connection ∇ and where we have used Theorem 3.1.2 and (3.4.17).

$R(e_j, e_i)$ here acts on spinor fields, and if we express this operator w.r.t. our local frame field e_k, we obtain a factor $\frac{1}{2}$ as in (3.4.5), coming from Lemma 1.11.3:

$$R(e_i, e_j) = \frac{1}{2} \sum_{k<l} \langle R(e_i, e_j) e_k, e_l \rangle e_k e_l, \tag{3.4.19}$$

where $e_k e_l$ again operates by Clifford multiplication. In order to derive (3.4.15) from (3.4.18), it thus remains to evaluate

$$\frac{1}{2}\sum_{j<i}\sum_{k<l}\langle R(e_i,e_j)e_k,e_l\rangle e_ie_je_ke_l = \frac{1}{8}\sum_{i,j,k,l}\langle R(e_i,e_j)e_k,e_l\rangle e_ie_je_ke_l. \tag{3.4.20}$$

If i, j, k are all distinct,

$$e_ie_je_k = e_je_ke_i = e_ke_ie_j,$$

and the first Bianchi identity (see Lemma 3.3.1) implies in this case that

$$R(e_i,e_j)e_k + R(e_j,e_k)e_i + R(e_k,e_i)e_j = 0.$$

The remaining terms are

$$\begin{aligned}
&\frac{1}{8}\sum_{i,j,k,l}\left(\langle R(e_i,e_k)e_k,e_l\rangle e_ie_ke_ke_l + \langle R(e_k,e_i)e_k,e_l\rangle e_ke_ie_ke_l\right)\\
&= -\frac{1}{4}\sum_{i,k,l}\langle R(e_i,e_k)e_k,e_l\rangle e_ie_l \quad \text{by (3.1.33) and } e_k^2=-1\\
&= -\frac{1}{4}R_{il}e_ie_l \quad \text{where } R_{il} \text{ is the Ricci tensor}\\
&= \quad \frac{1}{4}R_{ii} \quad \text{since } R_{il}=R_{li} \text{ (see (3.3.19))}, \quad e_ie_l+e_le_i=-2\delta_{ij}\\
&= \quad \frac{1}{4}R.
\end{aligned}$$

□

Theorem 3.4.2. *Let M be a* spin^c *manifold with a local orthonormal frame field $e_1,\dots,e_n$ and a* spin^c *connection ∇_A. The Dirac operator $\not{\partial}_A$ satisfies*

$$\not{\partial}_A^2 = -\nabla^2_{A,e_ie_i} + \frac{1}{4}R + \frac{1}{2}F_A, \tag{3.4.21}$$

where F_A, an imaginary valued two-form, is the curvature of A. (F_A acts on spinors by Clifford multiplication; in our frame field, $\sum_{i<j} F_{A,ij}e_i\wedge e_j$ becomes $\frac{1}{2}\sum_{i<j} F_{A,ij}e_ie_j$ as usual.)

Proof. The proof is the same as the one of Theorem 3.4.1, except for the additional $\mathfrak{u}(1)$ part A of the connection that leads to the additional F_A in the formula. □

Perspectives. See the references given in the perspectives on §1.11. The Dirac operator on the spinor bundle (Definition 3.4.1) was introduced by Atiyah and Singer [8] in their investigation of the index of elliptic operators. The simpler Dirac operator on the Clifford bundle had been studied earlier by Kähler[160].

3.5 The Bochner Method

Lemma 3.5.1. *Let $(e_i)_{i=1,\ldots,d}$ be a local orthonormal frame field on M, with dual coframe field $(\eta^i)_{i=1,\ldots,d}$, as in Lemma 3.3.4.*

If ω is a harmonic form, then

$$-\Delta\langle\omega,\omega\rangle = 2\langle\nabla_{e_i}\omega,\nabla_{e_i}\omega\rangle \; - \; 2\langle\omega,\eta^i\wedge\iota(e_j)R(e_i,e_j)\omega\rangle. \tag{3.5.1}$$

Proof. Let x_0 be a point in M where we perform the computations, and choose normal coordinates centered at x_0 and $e_i = \frac{\partial}{\partial x^i}$. Again, the formulae will not depend on the choice of a local orthonormal frame. Then, by the remark after Theorem 3.3.3 and (3.3.42)

$$\begin{aligned} -\Delta\langle\omega,\omega\rangle &= \nabla_{e_i}\nabla_{e_i}\langle\omega,\omega\rangle \\ &= 2\langle\nabla_{e_i}\omega,\nabla_{e_i}\omega\rangle + 2\langle\omega,\nabla_{e_i}\nabla_{e_i}\omega\rangle\,. \end{aligned} \tag{3.5.2}$$

(3.3.42) and (3.3.41) then yield (3.5.1), since $\Delta\omega = 0$ by assumption. □

Lemma 3.5.2. *With the notation of Lemma 3.5.1, we have for a harmonic 1-form ω on M*

$$-\Delta\langle\omega,\omega\rangle \;= 2|\nabla\omega|^2 + 2\,\mathrm{Ric}\,(\omega,\omega) \tag{3.5.3}$$

with $|\nabla\omega|^2 := \langle\nabla_{e_i}\omega,\nabla_{e_i}\omega\rangle$ and writing $\omega = f_i\eta^i$,

$$\mathrm{Ric}\,(\omega,\omega) := \mathrm{Ric}\,(f_ie_i,f_je_j) = f_if_j\,\mathrm{Ric}\,(e_i,e_j).$$

Proof. We compute the curvature term in (3.5.1) for a 1-form ω :

$$\begin{aligned} \langle\omega,\eta^i\wedge\iota(e_j)R(e_i,e_j)\omega\rangle &= \langle f_\ell\eta^\ell,\eta^i\wedge\iota(e_j)R(e_i,e_j)f_k\eta^k\rangle \\ &= -f_\ell f_k\langle\eta^\ell,\eta^i\wedge\iota(e_j)R_{kmij}\eta^m\rangle \\ &= -f_\ell f_k\langle\eta^\ell,R_{kjij}\eta^i\rangle \\ &= -f_\ell f_kR_{kj\ell j} \\ &= -f_\ell f_kR_{k\ell} \\ &= -\mathrm{Ric}\,(\omega,\omega), \end{aligned}$$

where we have used the tensor notation of §3.3, (e.g. (3.3.6) and (3.3.18)). □

Theorem 3.5.1 (Bochner).

(i) *Let M be a compact Riemannian manifold with nonnegative Ricci curvature. Then every harmonic 1-form ω is parallel (i.e. $\nabla\omega \equiv 0$). In particular, the first de Rham cohomology group satisfies*

$$\dim H^1_{dR}(M,\mathbb{R}) \le d(=\dim M).$$

(ii) If M is a compact Riemannian manifold of positive Ricci curvature, then M has no nontrivial harmonic 1-form. Thus,

$$H^1_{dR}(M, \mathbb{R}) = \{0\}.$$

Proof. We integrate formula (3.5.3). Then from (2.1.18)

$$0 = -\int_M \Delta\langle\omega, \omega\rangle * (1) = 2\int_M (|\nabla\omega|^2 + \mathrm{Ric}\,(\omega, \omega)) * (1). \tag{3.5.4}$$

By our assumption, the integrand on the right hand side is pointwise nonnegative. It therefore has to vanish identically. This implies in particular

$$\nabla\omega \equiv 0, \tag{3.5.5}$$

and ω is parallel.

A parallel 1-form is determined by its value at one point of M (cf. the discussion before Definition 3.1.2).

Therefore, the dimension of the vector space of parallel 1-forms is at most the dimension of the cotangent space T_x^*M, i.e. d. Likewise, (3.5.4) implies

$$\mathrm{Ric}\,(\omega, \omega) \equiv 0. \tag{3.5.6}$$

Thus, if M has positive Ricci curvature, we must have $\omega \equiv 0$. □

Remark. In (ii) of the preceding theorem, it suffices to assume that M has nonnegative Ricci curvature, and that there exists some point x_0 where the Ricci curvature is positive. Namely, from

$$\mathrm{Ric}\,(\omega, \omega) \equiv 0,$$

we then conclude that, $\omega(x_0) = 0$, and since ω is parallel, it then vanishes everywhere.

Below, we shall derive a stronger result (Corollary 4.3.1, Theorem of Bonnet-Myers) on the topology of Riemannian manifolds of positive Ricci curvature by a different method. Nevertheless, the Bochner method is an important tool in Riemannian geometry because it has a rather general range of applicability. It also applies to harmonic sections of bundles (suitably defined), harmonic mappings (see Chapter 7) etc. The harmonicity of the object under consideration will imply a formula of the type of (3.5.1). The essential point of (3.5.1) is that instead of third order derivatives that would appear for a general, nonharmonic object, one only has a commutator term given by a curvature expression. The other term on the right hand side is a square, hence nonnegative. If one then assumes that the curvature is such that the curvature term also is nonnegative, both terms have to vanish identically, because the integral of the left hand side vanishes. The vanishing of the square term then implies that the object is parallel. If the curvature is even positive, the vanishing of the curvature term implies that the object itself vanishes.

We shall see another instance of the Bochner method in §7.2.

When combining the preceding reasoning with the Weitzenböck formula of Theorem 3.4.1, we get

Theorem 3.5.2 (**Lichnerowicz**). *Let M be a compact spin manifold. If M has nonnegative scalar curvature, then every harmonic spinor field is parallel. If the scalar curvature is positive, then every harmonic spinor field vanishes.*

Proof. As in the proof of Lemma 3.5.1, we compute for a harmonic spinor field σ

$$\begin{aligned} -\Delta\langle\sigma,\sigma\rangle &= 2\langle\nabla_{e_i}\sigma,\nabla_{e_i}\sigma\rangle + 2\langle\sigma,\nabla^2_{e_ie_i}\sigma\rangle \\ &= 2\langle\nabla_{e_i}\sigma,\nabla_{e_i}\sigma\rangle + \frac{1}{2}R\langle\sigma,\sigma\rangle \quad \text{by (3.4.15).} \end{aligned}$$

As in the proof of Theorem 3.5.1, we integrate this formula to get

$$2\int\langle\nabla_{e_i}\sigma,\nabla_{e_i}\sigma\rangle * (1) + \frac{1}{2}\int R\langle\sigma,\sigma\rangle * (1) = 0.$$

If $R \geq 0$, both integrands have to vanish identically; in particular $\nabla_{e_i} \equiv 0$ meaning that σ is parallel. If $R > 0$, $\sigma \equiv 0$. □

Perspectives. Further applications of the Bochner method may be found in the monograph [267]. See also the Perspectives on §7.2.

3.6 The Geometry of Submanifolds. Minimal Submanifolds

Let M be an m-dimensional submanifold of the n-dimensional Riemannian manifold N. The metric $\langle .,. \rangle$ on N induces a metric on M, as described in §1.4. The question arises how to compute the Levi-Civita connection ∇^M of M from the one on N, ∇^N.

Theorem 3.6.1. *We have*

$$\nabla^M_X Y = (\nabla^N_X Y)^\top \quad \textit{for } X, Y \in \Gamma(TM), \tag{3.6.1}$$

where $\top : T_xN \to T_xM$ for $x \in M$ denotes the orthogonal projection.

Proof. In order that the right hand side of (3.6.1) is defined, we have to extend X and Y locally to a neighborhood of M in N. This is most easily done in local coordinates around $x \in M$ that locally map M to $\mathbb{R}^m \subset \mathbb{R}^n$. The extension of $X = \xi^i(x)\frac{\partial}{\partial x^i}$ then for example is

$$\tilde{X}(x^1, \ldots, x^n) = \sum_{i=1}^{m} \xi^i(x^1, \ldots, x^n)\frac{\partial}{\partial x^i}.$$

We then have

$$\begin{aligned}\langle \tilde{X}, \tilde{Y}\rangle(x) &= \langle X, Y\rangle(x)\\ [\tilde{X}, \tilde{Y}](x) &= [X, Y](x).\end{aligned}$$

Since (3.3.1) has to hold for ∇^M as well as for ∇^N, (3.6.1) follows. (It follows from the representation of ∇^N by Christoffel symbols, that $(\nabla^N_X Y)^\top$ does not depend on the chosen extensions. It is also clear that $(\nabla^N_X Y)^T$ defines a torsion free connection on M because ∇^N is a torsion free connection on M, and since $\nabla^N_X Y - \nabla^N_Y X - [X, Y]$ vanishes, also the part of this expression that is tangential to M has to vanish.) $\square$

With the help of Theorem 3.6.1, we may easily determine the Levi-Civita connection of $S^n \subset \mathbb{R}^{n+1}$. Let $\nu(x)$ be a vector field in a neighborhood of $x_0 \in M \subset N$, that is orthogonal to M, i.e.

$$\langle \nu(x), X\rangle = 0 \quad \text{for all } X \in T_xM. \tag{3.6.2}$$

We denote the orthogonal complement of T_xM in T_xN $T_xM^\perp$. The bundle $TM^\perp$ with fiber $T_xM^\perp$ at $x \in M$ is called normal bundle of M in N. (3.6.2) thus means

$$\nu(x) \in T_xM^\perp.$$

Lemma 3.6.1. $(\nabla^N_X \nu)^\top(x)$ *only depends on* $\nu(x)$, *the value of* ν *at* x.

Proof. For a real valued function f in a neighborhood of x

$$\begin{aligned}(\nabla^N_X f\nu)^\top(x) &= (X(f)(x)\nu(x))^\top + f(x)(\nabla_X \nu)^\top(x)\\ &= f(x)(\nabla_X \nu)^\top(x),\end{aligned}$$

since $\nu(x) \in T_xM^\perp$. $\square$

Lemma 3.6.1 makes the following definition possible.

Definition 3.6.1. *The second fundamental tensor* of M at the point x is the map

$$S : T_xM \times T_xM^\perp \to T_xM, \tag{3.6.3}$$

defined by $S(X, \nu) = (\nabla^N_X \nu)^\top$.

Lemma 3.6.2. *For* $X, Y \in T_xM$,

$$\ell_\nu(X, Y) := \langle S(X, \nu), Y\rangle \tag{3.6.4}$$

is symmetric in X *and* Y.

Proof.

$$
\begin{aligned}
\ell_\nu(X,Y) &= \langle (\nabla^N_X \nu)^\top, Y\rangle \\
&= \langle \nabla^N_X \nu, Y\rangle && \text{since } Y \in T_xM \\
&= -\langle \nu, \nabla^N_X Y\rangle && \text{since } \langle \nu, Y\rangle = 0 \text{ and } \nabla^N \text{ is metric} \\
&= -\langle \nu, \nabla^N_Y X + [X,Y]\rangle && \text{since } \nabla^N \text{ is torsion free} \\
&= -\langle \nu, \nabla^N_Y X\rangle && \text{since } [X,Y] \in T_xM, \nu \in T_xM^\perp \\
&= \langle \nabla^N_Y \nu, X\rangle && \text{since } \langle \nu, X\rangle = 0 \text{ and } \nabla^N \text{ is metric} \\
&= \langle (\nabla^N_Y \nu)^\top, X\rangle && \text{since } X \in T_xM \\
&= \ell_\nu(Y,X)\,. && (3.6.5)
\end{aligned}
$$

□

Definition 3.6.2. $\ell_\nu(\cdot,\cdot)$ is called the *second fundamental form* of M w.r.t. N.

Remark. The *first fundamental form* is the metric, applied to X and $Y \in T_xM$, i.e. $\langle X, Y\rangle$. For a fixed normal field ν, we write $S_\nu(X) = S(X,\nu)$. $S_\nu : T_xM \to T_xM$ then is selfadjoint w.r.t. the metric $\langle .,. \rangle$, by Lemma 3.6.2. Suppose now $\langle \nu,\nu\rangle \equiv 1$; i.e. ν is a unit normal field. The m eigenvalues of S_ν which are all real by self adjointness are called the *principal curvatures* of M in the direction ν, and the corresponding eigenvectors are called principal curvature vectors.

The *mean curvature* of M in the direction ν is

$$H_\nu := \frac{1}{m}\operatorname{tr} S_\nu.$$

The *Gauss-Kronecker curvature* of M in the direction ν is

$$K_\nu := \det S_\nu.$$

For an orthonormal basis $e_1,\dots,e_m$ of T_xM,

$$K_\nu = \det(\ell_\nu(e_i,e_j)).$$

We now consider the case where M has codimension 1, i.e. $n = m+1$. In this case, for each $x \in M$, there are precisely two normal vectors $\nu \in T_xM^\perp$ with $\langle \nu,\nu\rangle = 1$. We locally fix such a normal field and drop the subscript ν. If we would choose the opposite normal field instead, ℓ and S would change their sign, and the mean curvature M as well. For even m, however, the Gauss-Kronecker curvature does not depend on the choice of the direction of ν.

Furthermore, because of $\langle \nu,\nu\rangle \equiv 1$ $\nabla^N_X \nu$ is always tangential to M, and geometrically, it measures the "tilting velocity" with which ν is tilted (relative to a fixed parallel vector field in N) when moving on M in the direction X.

We now want to compare the curvature tensors of M and N, R^M and R^N. It turns out that their difference is given by the second fundamental tensor; namely

Theorem 3.6.2 (**Gauss Equations**). *Let M be a submanifold of the Riemannian manifold N, $m = \dim M, n = \dim N$, $k = n - m, x \in M$, $\nu_1, \ldots, \nu_k$ an orthonormal basis for $(T_xM)^\perp, S_\alpha := S_{\nu_\alpha}, \ell_\alpha := \ell_{\nu_\alpha}(\alpha = 1, \ldots, k)$. With the convention that a Greek minuscule occuring twice is summed from 1 to k, for $X, Y, Z, W \in T_xM$*

$$R^M(X,Y)Z - (R^N(X,Y)Z)^\top = \ell_\alpha(Y,Z)S_\alpha X - \ell_\alpha(X,Z)S_\alpha Y \tag{3.6.6}$$

and hence also

$$\begin{aligned} &\langle R^M(X,Y)Z, W\rangle - \langle R^N(X,Y)Z, W\rangle \\ &\quad = \ell_\alpha(Y,Z)\ell_\alpha(X,W) - \ell_\alpha(X,Z)\ell_\alpha(Y,W). \end{aligned} \tag{3.6.7}$$

Proof. Since everything is tensorial, we extend $X, Y, Z, W, \nu_1, \ldots, \nu_k$ to vector fields in TM and $TM^\perp$, resp., with the ν_α always being orthonormal.

$$\nabla^N_Y Z = (\nabla^N_Y Z)^\top + (\nabla^N_Y Z)^\perp = \nabla^M_Y Z + \langle \nu_\alpha, \nabla^N_Y Z\rangle \nu_\alpha,$$

since the ν_α form an orthonormal basis of $TM^\perp$.

Hence

$$\nabla^N_X \nabla^N_Y Z = \nabla^N_X \nabla^M_Y Z + X(\langle \nu_\alpha, \nabla^N_Y Z\rangle)\nu_\alpha + \langle \nu_\alpha, \nabla^N_Y Z\rangle \nabla^N_X \nu_\alpha,$$

i.e.

$$\begin{aligned} (\nabla^N_X \nabla^N_Y Z)^\top &= \nabla^M_X \nabla^M_Y Z + \langle \nu_\alpha, \nabla^N_Y Z\rangle (\nabla^N_X \nu_\alpha)^\top \\ &= \nabla^M_X \nabla^M_Y Z - \ell_\alpha(Y,Z) S_\alpha(X) \quad \text{by (3.6.5)}. \end{aligned} \tag{3.6.8}$$

Analogously

$$(\nabla^N_Y \nabla^N_X Z)^\top = \nabla^M_Y \nabla^M_X Z - \ell_\alpha(X,Z) S_\alpha(Y). \tag{3.6.9}$$

Moreover,

$$(\nabla^N_{[X,Y]} Z)^\top = \nabla^M_{[X,Y]} Z \quad \text{by Theorem (3.6.1)}. \tag{3.6.10}$$

(3.6.6) follows from (3.6.8) – (3.6.10), and (3.6.7) follows from (3.6.6). □

The "theorema egregium" of Gauss is the following special case of Theorem 3.6.2:

Corollary 3.6.1. *For a surface M in $\mathbb{R}^3$ (i.e. $m = 2, n = 3$) the Gauss curvature, defined as the determinant of the second fundamental form, hence defined through the embedding of M in $\mathbb{R}^3$, coincides with the Riemannian curvature of M which is determined by the metric, hence independent of the embedding. Thus, the Gauss curvature does not depend on the embedding of M into $\mathbb{R}^3$ either.* □

Definition 3.6.3. A Riemannian submanifold M of a Riemannian manifold N is called *totally geodesic* if all geodesics in M are also geodesics in N.

Theorem 3.6.3. *M is totally geodesic in N if and only if all second fundamental forms of M vanish identically.*

Proof. Let $c : I \to M$ be geodesic in M, i.e. $\nabla^M_{\dot c}\dot c = 0$. Because of $(\nabla^N_{\dot c}\dot c)^\top = \nabla^M_{\dot c}\dot c$ (Theorem 3.6.1), c is geodesic in N if and only if $(\nabla^N_{\dot c}\dot c)^\perp = 0$, i.e.

$$\langle \nabla^N_{\dot c}\dot c, \nu\rangle = 0 \quad \text{for all } \nu \in TM^\perp.$$

Now

$$\begin{aligned}\langle \nabla^N_{\dot c}\dot c, \nu\rangle &= -\langle \dot c, \nabla^N_{\dot c}\nu\rangle, \quad \text{since } \langle \dot c, \nu\rangle = 0 \text{ and } \nabla^N \text{ is metric}\\ &= -\ell_\nu(\dot c, \dot c).\end{aligned}$$

The claim directly follows. □

For example, each closed geodesic in a Riemannian manifold defines a 1-dimensional compact totally geodesic submanifold.

The totally geodesic submanifolds of Euclidean space are precisely the affine linear subspaces (and their open subsets). The closed totally geodesic subspaces of the sphere $S^n \subset \mathbb{R}^{n+1}$ are precisely the intersections of S^n with linear subspaces of $\mathbb{R}^{n+1}$, hence spheres themselves. This follows directly from the description of the geodesics on S^n in §1.4. A generic Riemannian manifold, however, does not have any totally geodesic submanifolds of dimension > 1.

We want to briefly discuss a global aspect.

Let M be an oriented submanifold of the oriented Riemannian manifold N. This means that M itself is an oriented manifold whose orientation coincides with the one induced by N. If thus for $x \in M$ $e_1, \dots, e_n$ is a positive basis of T_xN for which $e_1, \dots, e_m$ are tangential to M, then $e_1, \dots, e_m$ constitute a positive basis of T_xM.

If under this assumption, we have $n = m+1$, we may also determine the sign of the unit normal field ν by requiring that if $e_1, \dots, e_m$ is a positive basis of T_xM, then $e_1, \dots, e_m, \nu$ is a positive basis of T_xN. Suppose now that $N = \mathbb{R}^n$, i.e. that M is an oriented hypersurface of $\mathbb{R}^n$. Let $p : T\mathbb{R}^n \to \mathbb{R}^n$ map each fiber of $T\mathbb{R}^n$ isomorphically onto $\mathbb{R}^n$, in the usual canonical manner, i.e. by parallel transport into the origin.

Definition 3.6.4. $p \circ \nu : M \to S^{n-1}$ is called the *Gauss map* of M.

The Gauss-Kronecker curvature, i.e. the Jacobian of $d\nu(x) : T_xM \to T_xM$, then becomes the Jacobian of the Gauss map. It thus measures the infinitesimal volume distortion of M by the Gauss map. Theorem 3.6.2 allows an easy computation of the curvature of the sphere $S^n \subset \mathbb{R}^{n+1}$. Namely, for $x = (x^1, \dots, x^{n+1}) \in S^n$, a unit normal vector $\nu(x)$ is given by

$$\nu(x) = x^i \frac{\partial}{\partial x^i}.$$

Furthermore,

$$\nabla^{\mathbb{R}^{n+1}}_{\frac{\partial}{\partial x^j}} \nu(x) = \frac{\partial}{\partial x^j}(x^i)\frac{\partial}{\partial x^i} = \frac{\partial}{\partial x^j}.$$

Since we have already seen that the isometry group of S^n operates transitively on S^n, we may consider w.l.o.g. the north pole $(0, 0, \dots, 0, 1)$. $\frac{\partial}{\partial x^1}, \dots, \frac{\partial}{\partial x^n}$ are tangential to

S^n at this point. It follows that

$$\begin{aligned} S\Big(\frac{\partial}{\partial x^j}\Big) &= \frac{\partial}{\partial x^j} - \Big\langle \nu(x), \frac{\partial}{\partial x^j}\Big\rangle \nu(x) \\ &= \frac{\partial}{\partial x^j} \qquad \text{for } j = 1, \dots, n \end{aligned}$$

and

$$\ell\Big(\frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^k}\Big) = \Big\langle \frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^k}\Big\rangle = \delta_{jk}.$$

We conclude

$$\Big\langle R^{S^n}\Big(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\Big)\frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^\ell}\Big\rangle = \delta_{jk}\delta_{i\ell} - \delta_{ik}\delta_{j\ell}. \tag{3.6.11}$$

In particular, the sectional curvature is 1.

We also obtain the formula

$$R^{S^n}(X,Y)Z = \langle Y, Z\rangle X - \langle X, Z\rangle Y. \tag{3.6.12}$$

We want to consider a particular class of submanifolds in more detail, namely those that are critical points of the volume functional.

Let $\tilde{M}$ be an m-dimensional submanifold of N, with frame $\tilde{e}_1, \dots, \tilde{e}_m$, coframe $\tilde{\eta}^1, \dots, \tilde{\eta}^m$ and volume form $\tilde{\eta}$ as before, and let

$$\Phi : M \to \tilde{M}$$

be a diffeomorphism. Let $e_1, \dots, e_m$ be a frame on M, η the volume form.

Then

$$\begin{aligned} \operatorname{Vol}(\tilde{M}) &= \Big| \int_{\tilde{M}} \tilde{\eta} \Big| \\ &= \Big| \int_M \Phi^* \tilde{\eta} \Big| \\ &= \Big| \int_M \Phi^*\tilde{\eta}^1 \wedge \ldots \wedge \Phi^*\tilde{\eta}^m \Big| \\ &= \Big| \int_M |\Phi_* e_1 \wedge \ldots \wedge \Phi_* e_m| \eta \Big| \\ &= \Big| \int_M \langle \Phi_* e_1 \wedge \ldots \wedge \Phi_* e_m, \Phi_* e_1 \wedge \ldots \wedge \Phi_* e_m \rangle^{\frac{1}{2}} \eta \Big| . \end{aligned} \tag{3.6.13}$$

We now consider a more special situation. We define a local variation of M to be a smooth map

$$F : M \times (-\varepsilon, \varepsilon) \to N \quad (\varepsilon > 0)$$

with

$$\operatorname{supp} F := \overline{\{x \in M : F(x,t) \neq x \quad \text{for some } t \in (-\varepsilon, \varepsilon)\}} \tag{3.6.14}$$

being a compact subset of M and

$$F(x,0) = x \quad \text{for all } x \in M.$$

For small enough $|t|$, $\Phi_t(\cdot) := F(\cdot, t)$ then is a diffeomorphism from M onto a submanifold M_t of N, by the implicit function theorem. We assume that $\varepsilon > 0$ is chosen so small, that this is the case for all $t \in (-\varepsilon, \varepsilon)$. Since the subsequent computations are local, we also assume that $\{x \in M : F(x,t) \neq x\}$ is orientable and that $e_1, \ldots, e_m$ is a positively oriented orthonormal basis.

The variation of volume then is (by (3.6.13))

$$\begin{aligned}
&\frac{d}{dt}\operatorname{Vol}(\Phi_t(M))_{|t=0} \\
&\quad = \frac{d}{dt}\int_M \langle \Phi_{t*}e_1 \wedge \ldots \wedge \Phi_{t*}e_m, \Phi_{t*}e_1 \wedge \ldots \wedge \Phi_{t*}e_m\rangle^{\frac{1}{2}} \eta_{|t=0} \\
&\quad = \sum_{\alpha=1}^m \int_M \frac{\langle \Phi_{t*}e_1 \wedge \ldots \wedge \frac{\partial}{\partial t}\Phi_{t*}e_\alpha \wedge \ldots \wedge \Phi_{t*}e_m, \Phi_{t*}e_1 \wedge \ldots \wedge \Phi_{t*}e_m\rangle}{|\Phi_{t*}e_1 \wedge \ldots \wedge \Phi_{t*}e_m|}\eta_{|t=0}\,.
\end{aligned}$$

Putting

$$X := \frac{\partial}{\partial t}\Phi_{t_{|t=0}},$$

we obtain

$$\begin{aligned}
&\frac{d}{dt}\operatorname{Vol}(\Phi_t(M))_{|t=0} \\
&\quad = \sum_{\alpha=1}^m \int_M \frac{\langle e_1 \wedge \ldots \wedge \nabla^N_{e_\alpha}X \wedge \ldots \wedge e_m, e_1 \wedge \ldots \wedge e_m\rangle}{|e_1 \wedge \ldots \wedge e_m|}\eta\,.
\end{aligned}$$

Namely, if $c_\alpha(s)$ is a curve on M with $c_\alpha(0) = x, c'_\alpha(0) = e_\alpha$, and $c_\alpha(s,t) = \Phi_t(c_\alpha(s))$, then

$$\Phi_{t*}e_\alpha = \frac{\partial}{\partial s}c_\alpha(s,t)_{|s=0}$$

and

$$\begin{aligned}
\frac{\partial}{\partial t}\Phi_{t*}e_{\alpha_{|t=0}} &= \frac{\partial}{\partial t}\frac{\partial}{\partial s}c_\alpha(s,t)_{|s=t=0} \\
&= \frac{\partial}{\partial s}\frac{\partial}{\partial t}c_\alpha(s,t)_{|s=t=0} \\
&= \nabla^N_{\frac{\partial}{\partial s}}X_{|s=0} \\
&= \nabla^N_{e_\alpha}X\,.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{d}{dt}\operatorname{Vol}(\Phi_t(M))_{|t=0} &= \int_M \langle \nabla^N_{e_\alpha}X, e_\alpha\rangle\eta \\
&= \int_M \{e_\alpha\langle X, e_\alpha\rangle - \langle X, \nabla^N_{e_\alpha}e_\alpha\rangle\}\eta\,.
\end{aligned} \tag{3.6.15}$$

Now $e_\alpha \langle X, e_\alpha \rangle = \operatorname{div} X^T$, and since X vanishes outside a compact subset of M (see (3.6.14)), we have by Gauss' theorem

$$\int_M e_\alpha \langle X, e_\alpha \rangle = 0.$$

As in the proof of Lemma 3.3.4 3), we may assume that at the point under consideration

$$\nabla^M_{e_\alpha} e_\alpha = 0.$$

We then obtain from (3.6.15)

$$\frac{d}{dt} \operatorname{Vol}\left(\Phi_t(M)\right)_{|t=0} = -\int_M \langle X^\perp, \nabla^N_{e_\alpha} e_\alpha \rangle \cdot \eta \,. \tag{3.6.16}$$

We conclude

Theorem 3.6.4. *A submanifold M of the Riemannian manifold N is a critical point of the volume function, i.e.*

$$\frac{d}{dt} \operatorname{Vol}\left(\Phi_t(M)\right)_{|t=0} = 0 \tag{3.6.17}$$

for all local variations of M if and only if the mean curvature H_ν of M vanishes for all normal directions ν.

Proof. We choose an orthonormal basis $\nu_1, \ldots, \nu_k$ $(k = n - m)$ of $T_x M^\perp$ for $x \in M$ and write

$$X^\perp = \xi^j \nu_j. \tag{3.6.18}$$

Then

$$\langle X^\perp, \nabla^N_{e_\alpha} e_\alpha \rangle = \xi^j \operatorname{tr} S_{\nu_j} = m \xi^j H_{\nu_j}. \tag{3.6.19}$$

Since every section X of $TM^\perp$ over M with compact support on M defines a local variation

$$F(x,t) := \exp_x tX(x)$$

of M, (3.6.17) holds if and only if (3.6.19) vanishes for all choices of ξ^j, and the conclusion follows. □

Definition 3.6.5. A submanifold M of the Riemannian manifold N is called *minimal* if its mean curvature H_ν vanishes for all normal directions ν.

We want to consider a somewhat more general situation. We let M and N be Riemannian manifolds of dimension m and n, resp., and we let

$$f : M \to N$$

be an *isometric immersion*. This means that for each $p \in M$, there exists a neighborhood U for which

$$f : U \to f(U)$$

is an isometry ($f(U)$ is equipped with the metric induced from N). The point here is that $f(M)$ need not be an embedded submanifold of N but may have self-intersections or may even be dense in N. We may then define local variations $F(x,t) : M \to N$ with $F(x,0) = f(x)$ as before, and $f(M)$ is critical for the volume functional if and only if its mean curvature vanishes, in the sense that for all U as above, $f(U)$ has vanishing mean curvature in all normal directions. Such an $f(M)$ then is called an *immersed minimal submanifold* of N. We now want to write the condition for the vanishing of the mean curvature, namely

$$(\nabla^N_{e_\alpha} e_\alpha)^\perp = 0 \tag{3.6.20}$$

in terms of f.

For that purpose, we introduce normal coordinates at the point $x \in M$ under consideration, i.e. at x

$$\begin{aligned} \left\langle \frac{\partial}{\partial x^\alpha}, \frac{\partial}{\partial x^\beta} \right\rangle &= \delta_{\alpha\beta}, \\ \nabla^M_{\frac{\partial}{\partial x^\alpha}} \frac{\partial}{\partial x^\beta} &= 0, \end{aligned} \tag{3.6.21}$$

for $\alpha, \beta = 1, \dots, m$.

Here , ∇^M is the Levi-Civita connection of M, and because f is an isometric immersion, for all X and $Y \in T_x M$,

$$\nabla^M_X Y = \nabla^{f(M)}_{f_* X} f_* Y = (\nabla^N_{f_* X} f_* Y)^\top \quad \text{by Theorem 3.6.1.} \tag{3.6.22}$$

(This fact may also be expressed by saying that ∇^M is the connection in the pull back bundle $f^*(Tf(M))$ induced by the Levi-Civita connection of N.)

$$e_\alpha := f_* \left(\frac{\partial}{\partial x^\alpha} \right) = \frac{\partial f^i}{\partial x^\alpha} \frac{\partial}{\partial f^i}$$

where $(f^1, \dots, f^n)$ now are local coordinates for N near $f(x)$. Thus, for a function $\varphi : N \to \mathbb{R}, (e_\alpha(\varphi))(f(x)) = \frac{\partial}{\partial x^\alpha} \varphi \circ f(x)$.

Then, computing at x,

$$\begin{aligned} (\nabla^N_{e_\alpha} e_\alpha)^\perp &= \nabla^N_{e_\alpha} e_\alpha \quad \text{by (3.6.21), (3.6.22)} \\ &= \nabla^N_{\frac{\partial f^i}{\partial x^\alpha} \frac{\partial}{\partial f^i}} \frac{\partial f^j}{\partial x^\alpha} \frac{\partial}{\partial f^j} \\ &= \frac{\partial^2 f^f}{(\partial x^\alpha)^2} \frac{\partial}{\partial f^j} + \frac{\partial f^i}{\partial x^\alpha} \frac{\partial f^k}{\partial x^\alpha} \Gamma^j_{ik} \frac{\partial}{\partial f^j} \, . \end{aligned}$$

Here, Γ^j_{ik} are the Christoffel symbols of N.

We conclude that $f(M)$ has vanishing mean curvature, i.e. (3.6.20) holds if and only if

$$\frac{\partial^2 f^j}{(\partial x^\alpha)^2} + \Gamma^j_{ik}(f(x)) \frac{\partial f^i}{\partial x^\alpha} \frac{\partial f^k}{\partial x^\alpha} = 0 \quad \text{for } j = 1, \ldots, n. \tag{3.6.23}$$

(3.6.23) requires that the coordinates are normal at x. In arbitrary coordinates, (3.6.23) is transformed into

$$-\Delta_M f^j + \gamma^{\alpha\beta}(x) \Gamma^j_{ik}(f(x)) \frac{\partial f^i}{\partial x^\alpha} \frac{\partial f^k}{\partial x^\beta} = 0 \quad \text{for } j = 1, \ldots, n, \tag{3.6.24}$$

where Δ_M is the Laplace-Beltrami operator of M (see §2.1) and $(\gamma_{\alpha\beta})_{\alpha,\beta=1,\ldots,m}$ is the metric tensor of M.

In §7.1, solutions of (3.6.24) will be called *harmonic maps*. Thus, an isometric immersion is minimal if and only if it is harmonic.

A consequence of (3.6.24) is

Corollary 3.6.2. *The one dimensional immersed minimal submanifolds of N are the geodesics in N.* □

We now consider the case where N is Euclidean space $\mathbb{R}^n$. In Euclidean coordinates, all Christoffel symbols Γ^j_{ik} vanish, and we obtain

Corollary 3.6.3. *An immersed submanifold of $\mathbb{R}^n$ is minimal if and only if all coordinate functions are harmonic (w.r.t. the Laplace-Beltrami operator of the submanifold induced by the ambient Euclidean metric). In particular, there are no nontrivial compact minimal submanifolds of Euclidean space.*

Proof. The first claim follows from (3.6.24). The second one follows from the fact that, on a compact manifold, every harmonic function is constant by Corollary 2.1.2. And a manifold whose coordinate functions are all constant is a point, hence trivial. □

There is, however, a multitude of noncompact, but complete minimal surfaces in $\mathbb{R}^3$. Besides the trivial example of a plane, we mention:

1) The catenoid, given by the coordinate representation

$$f(s,t) = (\cosh s \cos t, \cosh s \sin t, s).$$

2) The helicoid, given by the coordinate representation

$$f(s,t) = (t \cos s, t \sin s, s).$$

3) Enneper's surface, given by the coordinate representation

$$f(s,t) = \left(\frac{s}{2} - \frac{s^3}{6} + \frac{st^2}{2}, -\frac{t}{2} + \frac{t^3}{6} - \frac{s^2 t}{2}, \frac{s^2}{2} - \frac{t^2}{2} \right).$$

We leave it as an exercise to the reader to verify that these have vanishing mean curvature and hence are minimal surfaces indeed.

In order to obtain a further slight generalization of the concept of a minimal surface in a Riemannian manifold, we observe that (3.6.24) is not affected if the operator occuring in that formula is multiplied by some (non-vanishing) function. In order to elaborate on that observation, we assume that Σ is a two dimensional Riemannian manifold and that coordinates x^1, x^2 are chosen on Σ for which $\frac{\partial}{\partial x^1}$ and $\frac{\partial}{\partial x^2}$ are always orthogonal and of the same length w.r.t. the metric $\langle \cdot, \cdot \rangle_\gamma$ of Σ, i.e.

$$\begin{aligned} \Big\langle \frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^1} \Big\rangle_\gamma &= \Big\langle \frac{\partial}{\partial x^2}, \frac{\partial}{\partial x^2} \Big\rangle_\gamma, \\ \Big\langle \frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^2} \Big\rangle_\gamma &= 0. \end{aligned} \tag{3.6.25}$$

This is equivalent to the metric γ being represented by

$$\lambda^2(x)(dx^1 \otimes dx^1 + dx^2 \otimes dx^2) \tag{3.6.26}$$

with some positive function $\lambda^2(x)\,(x = (x^1, x^2))$. Moreover, the precise value of $\lambda^2(x)$ is irrelevant for (3.6.25).

In those coordinates, (3.6.24) becomes, for an isometric immersion $f : \Sigma \to N$,

$$\frac{1}{\lambda^2(x)} \Big(\frac{\partial^2 f^i}{(\partial x^1)^2} + \frac{\partial^2 f^i}{(\partial x^2)^2} + \Gamma^i_{jk}(f(x)) \Big(\frac{\partial f^j}{\partial x^1} \frac{\partial f^k}{\partial x^1} + \frac{\partial f^j}{\partial x^2} \frac{\partial f^k}{\partial x^2} \Big) \Big) = 0,$$

and since as observed the factor $\frac{1}{\lambda^2(x)}$ is irrelevant, this becomes

$$\frac{\partial^2 f^i}{(\partial x^1)^2} + \frac{\partial^2 f^i}{(\partial x^2)^2} + \Gamma^i_{jk}(f(x)) \Big(\frac{\partial f^j}{\partial x^1} \frac{\partial f^k}{\partial x^1} + \frac{\partial f^j}{\partial x^2} \frac{\partial f^k}{\partial x^2} \Big) = 0. \tag{3.6.27}$$

Since f is required to be an isometric immersion, (3.6.25) becomes

$$\begin{aligned} \Big\langle \frac{\partial f}{\partial x^1}, \frac{\partial f}{\partial x^1} \Big\rangle &= \Big\langle \frac{\partial f}{\partial x^2}, \frac{\partial f}{\partial x^2} \Big\rangle, \\ \Big\langle \frac{\partial f}{\partial x^1}, \frac{\partial f}{\partial x^2} \Big\rangle &= 0, \end{aligned} \tag{3.6.28}$$

where now the metric is the one of N.

In order to provide a conceptual context for a reformulation of the preceding insights, we state

Definition 3.6.6. A surface Σ with a *conformal structure* is a two dimensional differentiable manifold with an atlas of so-called *conformal coordinates* whose transition functions $z = \varphi(x)$ satisfy

$$dz^1 \otimes dz^1 + dz^2 \otimes dz^2 = \mu^2(x)(dx^1 \otimes dx^1 + dx^2 \otimes dx^2) \tag{3.6.29}$$

($z = (z^1, z^2), x = (x^1, x^2)$), for some positive function $\mu^2(x)$. A map $f : \Sigma \to N$ from a surface Σ with a conformal structure into a Riemannian manifold N is called *conformal* if in conformal coordinates always

$$\Big\langle \frac{\partial f}{\partial x^1}, \frac{\partial f}{\partial x^1} \Big\rangle = \Big\langle \frac{\partial f}{\partial x^2}, \frac{\partial f}{\partial x^2} \Big\rangle \quad \text{and} \quad \Big\langle \frac{\partial f}{\partial x^1}, \frac{\partial f}{\partial x^2} \Big\rangle = 0. \tag{3.6.30}$$

In order to interpret (3.6.29), we compute

$$dz^1 \otimes dz^1 + dz^2 \otimes dz^2 = (\varphi^1_{x^i}\varphi^1_{x^j} + \varphi^2_{x^i}\varphi^2_{x^j})dx^i \otimes dx^j.$$

(3.6.29) then implies

$$\frac{\partial \varphi^1}{\partial x^1}\frac{\partial \varphi^1}{\partial x^1} + \frac{\partial \varphi^2}{\partial x^1}\frac{\partial \varphi^2}{\partial x^1} = \frac{\partial \varphi^1}{\partial x^2}\frac{\partial \varphi^1}{\partial x^2} + \frac{\partial \varphi^2}{\partial x^2}\frac{\partial \varphi^2}{\partial x^2}$$

and

$$\frac{\partial \varphi^1}{\partial x^1}\frac{\partial \varphi^1}{\partial x^2} + \frac{\partial \varphi^2}{\partial x^1}\frac{\partial \varphi^2}{\partial x^2} = 0.$$

Thus, the coordinate transformations are conformal in the Euclidean sense. A special case of a surface with a conformal structure is a Riemann surface as defined in Definition 8.1.1 below.

We also observe that (3.6.30) is independent of a particular choice of conformal coordinates, by a computation analogous to the one just performed.

Definition 3.6.7. Let Σ be a surface with conformal structure, N a Riemannian manifold.

A (parametric) *minimal surface* in N is a nonconstant map $f : \Sigma \to N$ satisfying (3.6.27) and (3.6.28).

This definition includes the previous definition of a minimal surface, i.e. a twodimensional minimal submanifold of N. Namely, the pull back $(f^*g)_{\alpha\beta}$ of the metric tensor g_{ij} of N is given by

$$\gamma_{\alpha\beta}(x) = g_{ij}(fx))\frac{\partial f^i}{\partial x^\alpha}\frac{\partial f^j}{\partial x^\beta} = \Big\langle \frac{\partial f}{\partial x^\alpha}, \frac{\partial f}{\partial x^\beta} \Big\rangle,$$

and if f is conformal, i.e. satisfies (3.6.28), then

$$\gamma_{\alpha\beta}(x) = \lambda^2(x)\delta_{\alpha\beta}$$

for some function $\lambda^2(x)$.

If $\lambda^2(x) \neq 0$, this is the situation previously discussed, and the vanishing of the mean curvature of $f(\Sigma)$ was shown to be equivalent to (3.6.27). $\lambda^2(x) \neq 0$ means that the derivative of f has maximal rank at x, and thus is a local immersion. Therefore, the only generalization of our previous concept admitted by Definition 3.6.7 is that we now include the degenerate case where

$$\Big\langle \frac{\partial f}{\partial x^1}, \frac{\partial f}{\partial x^1} \Big\rangle = 0 = \Big\langle \frac{\partial f}{\partial x^2}, \frac{\partial f}{\partial x^2} \Big\rangle \tag{3.6.31}$$

at some (but not all) points of Σ.

It may actually be shown that this can happen at most at a discrete set of points.

Perspectives. The theorema egregium of Gauss was the starting point of modern differential geometry. It provided the first instance of a nontrivial intrinsic differential invariant of a metric, and it motivated Riemann's definition of sectional curvature. For more details, we refer to [76]. In that textbook, also parametric minimal surfaces in $\mathbb{R}^3$ are treated. For a comprehensive treatment of minimal surfaces, we refer to the monographs [60], [205]. A good reference for minimal submanifolds of arbitrary dimension and codimension is [269].

Some further discussions about minimal surfaces may be found in Chapter 7.

Exercises for Chapter 3

1. Compute the transformation behaviour of the Christoffel symbols of a connection under coordinate transformations.

2. Let E be a vector bundle with fiber $\mathbb{C}^n$ and a Hermitian bundle metric. Develop a theory of unitary connections, i.e. of connections respecting the bundle metric.

3. Show that each vector bundle with a bundle metric admits a metric connection.

4. Let $x_0 \in M$, D a flat metric connection on a vector bundle E over M. Show that D induces a map $\pi_1(M, x_0) \to \mathrm{O}(n)$, considering $\mathrm{O}(n)$ as the isometry group of the fiber E_{x_0}.

5. Let $S_r^n := \{x \in \mathbb{R}^{n+1} : |x| = r\}$ be the sphere of radius r. Compute its curvature tensor and volume.

6. Consider the hyperboloid in $\mathbb{R}^3$ defined by the equation
$$x^2 + y^2 - z^2 = -1, z > 0$$
and compute its curvature.

7. Verify that the catenoid, the helicoid, and Enneper's surface are minimal surfaces.

8. Determine all surfaces of revolution in $\mathbb{R}^3$ that are minimal. (Answer: The catenoid is the only one.)

9. Let $F : M^m \to \mathbb{R}^{m+1}$ be an isometric immersion ($m = \dim M$). Give a complete derivation of the formula
$$\Delta F = m\eta$$
where Δ is the Laplace-Beltrami operator of M and η is the mean curvature vector of $F(M)$.

10. Let $F : M^m \to S^n \subset \mathbb{R}^{n+1}$ be an isometric immersion. Show that $F(M)$ is minimal in S^n if and only if there exists a function φ on M with $\Delta F = \varphi F$ and that in this case necessarily $\varphi \equiv m$.

11. Show that for $n \geq 4$, there exists no hypersurface (i.e. a submanifold of codimension 1) in $\mathbb{R}^n$ with negative sectional curvature.

12. Verify the formula $\partial\!\!\!/ = cl \circ \nabla$ given in §3.4.

Chapter 4

Geodesics and Jacobi Fields

4.1 1st and 2nd Variation of Arc Length and Energy

We start with a preliminary technical remark:

Let M be a d-dimensional Riemannian manifold with Levi-Civita connection ∇. Let H be a differentiable manifold, and let $f : H \to M$ be smooth. In the sequel, H will be an interval I or a square $I \times I$ in $\mathbb{R}^2$. Since f is not necessarily injective, it is not always possible to speak in an unambiguous way about the tangent space to $f(H)$ at a point $p \in f(H)$, even, if f is an immersion. Let for example $p = f(x) = f(y)$ with $x \neq y$. If f is an immersion, we may restrict f to sufficiently small neighborhoods U and V of x and y such that $f(U)$ and $f(V)$ have well defined tangent spaces at p. Thus, in a double point of $f(H)$, the tangent space can be specified by specifying the preimage (x or y). This can be formalized as follows: We consider the bundle $f^*(TM)$ over H, pulled back by f. The fiber over $x \in H$ here is $T_{f(x)}M$. This process already has been treated in a more general context in Definition 1.8.5. We now introduce a connection $f^*(\nabla)$ on $f^*(TM)$ by putting for $X \in T_xH,\ Y$ a section of $f^*(TM)$,

$$(f^*\nabla)_X Y := \nabla_{df(X)} Y \tag{4.1.1}$$

(here, $f^*(TM)_x$ is identified with $T_{f(x)}M$).

As in §3.4, in order that the right hand side is well defined, Y first has to be extended to a neighborhood of $f(H)$; as in §3.4, however, it turns out that the result will not depend on the choice of extension. In the sequel, instead of $(f^*\nabla)$, we shall simply write ∇, since the map f will be clear from the context.

A section of $f^*(TM)$ is called a *vector field along* f. An important rôle will be played by vector fields along curves $c : I \to M$, i.e. sections of $c^*(TM)$.

Let now $c : [a, b] \to M$ be a smooth curve, $\varepsilon > 0$. A *variation of* c is a differentiable map $F : [a, b] \times (-\varepsilon, \varepsilon) \to M$ with $F(t, 0) = c(t)$ for all $t \in [a, b]$. The variation is called proper if the endpoints stay fixed, i.e. $F(a, s) = c(a), F(b, s) = c(b)$ for all $s \in (-\varepsilon, \varepsilon)$. We also put $c_s(t) := c(t, s) := F(t, s)$, $\dot{c}(t, s) := \frac{\partial}{\partial t} c(t, s)$ (more precisely, $dF(\frac{\partial}{\partial t})c(t, s))$, $c'(t, s) = \frac{\partial}{\partial s} c(t, s)$ (more precisely $dF(\frac{\partial}{\partial s})c(t, s))$.

As in §1.4, let $L(\gamma)$ and $E(\gamma)$ denote the length and the energy of a curve γ. The following lemma is a reformulation of formulae from §1.4. Here, we want to give an intrinsic proof. For simplicity, we shall write $L(s), E(s)$ in place of $L(c_s), E(c_s)$ resp.

Lemma 4.1.1. *$L(s)$ and $E(s)$ are differentiable w.r.t. s, and we have*

$$L'(0) = \int_a^b \left(\frac{\frac{\partial}{\partial t}\langle c', \dot{c}\rangle}{\langle \dot{c}, \dot{c}\rangle^{\frac{1}{2}}} - \frac{\langle c', \nabla_{\frac{\partial}{\partial t}} \dot{c}\rangle}{\langle \dot{c}, \dot{c}\rangle^{\frac{1}{2}}} \right) dt, \tag{4.1.2}$$

$$E'(0) = \langle c'(b, 0), \dot{c}(b, 0)\rangle - \langle c'(a, 0), \dot{c}(a, 0)\rangle - \int_a^b \left\langle \frac{\partial c}{\partial s}, \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial t}(t, s) \right\rangle dt\,. \tag{4.1.3}$$

Proof.

$$\begin{aligned}
E(s) &= \frac{1}{2} \int_a^b \left\langle \frac{\partial c}{\partial t}(t, s), \frac{\partial c}{\partial t}(t, s) \right\rangle dt\,, \text{and then} \\
\frac{d}{ds} E(s) &= \frac{1}{2} \int_a^b \frac{\partial}{\partial s} \left\langle \frac{\partial c}{\partial t}(t, s), \frac{\partial c}{\partial t}(t, s) \right\rangle dt \\
&= \int_a^b \left\langle \nabla_{\frac{\partial}{\partial s}} \frac{\partial c}{\partial t}(t, s), \frac{\partial c}{\partial t}(t, s) \right\rangle dt \quad \text{since } \nabla \text{ preserves the metric}^1 \\
&= \int_a^b \left\langle \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial s}(t, s), \frac{\partial c}{\partial t}(t, s) \right\rangle dt \quad \text{since } \nabla \text{ is torsion free} \\
&= \int_a^b \left(\frac{\partial}{\partial t} \left\langle \frac{\partial c}{\partial s}(t, s), \frac{\partial c}{\partial t}(t, s) \right\rangle - \left\langle \frac{\partial c}{\partial s}, \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial t}(t, s) \right\rangle \right) dt \\
&= \left\langle \frac{\partial c}{\partial s}, \frac{\partial c}{\partial t} \right\rangle \Big|_{t=a}^{t=b} - \int_a^b \left\langle \frac{\partial c}{\partial s}, \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial t}(t, s) \right\rangle dt\,,
\end{aligned}$$

and similarly,

$$\begin{aligned}
L(s) &= \int_a^b \left\langle \frac{\partial c}{\partial t}(t, s), \frac{\partial c}{\partial t}(t, s) \right\rangle^{\frac{1}{2}} dt \\
\frac{d}{ds} L(s) &= \int_a^b \frac{\langle \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial s}(t, s), \frac{\partial c}{\partial t}(t, s)\rangle}{\langle \frac{\partial c}{\partial t}(t, s), \frac{\partial c}{\partial t}(t, s)\rangle^{\frac{1}{2}}} dt \\
&= \int_a^b \left(\frac{\frac{\partial}{\partial t}\langle c', \dot{c}\rangle}{\langle \dot{c}, \dot{c}\rangle^{\frac{1}{2}}} - \frac{\langle c', \nabla_{\frac{\partial}{\partial t}} \dot{c}\rangle}{\langle \dot{c}, \dot{c}\rangle^{\frac{1}{2}}} \right) dt\,.
\end{aligned}$$

□

In the special case where $c = c_0$ is parametrized proportionally to arclength, i.e. $\|\dot{c}(t,0)\| \equiv$ const., (4.1.2) becomes

$$L'(0) = \frac{1}{\langle \dot{c}, \dot{c} \rangle^{\frac{1}{2}}} \left(\langle c', \dot{c} \rangle |_{t=a,s=0}^{t=b,s=0} - \int_a^b \langle c', \nabla_{\frac{\partial}{\partial t}} \dot{c} \rangle dt \right). \tag{4.1.4}$$

Lemma 4.1.1 implies that c is stationary for E (w.r.t. variations that keep the end-points fixed) and if parametrized proportionally to arc length, also stationary for L if and only if

$$\nabla_{\frac{\partial}{\partial t}} \dot{c}(t,0) \equiv 0. \tag{4.1.5}$$

We recall that $\nabla_{\frac{\partial}{\partial t}}$ stands for $\nabla_{dF(\frac{\partial}{\partial t})}$; now $dF(\frac{\partial}{\partial t}) = \frac{\partial}{\partial t} c(t,s) = \dot{c}$, and (4.1.5), as to be expected is the equation for c being geodesic.

For the case where $c = c_0$ is geodesic, we now want to compute the second derivatives of E and L at $s = 0$:

Theorem 4.1.1. *Let $c : [a,b] \to M$ be geodesic. Then*

$$E''(0) = \int_a^b \langle \nabla_{\frac{\partial}{\partial t}} c'(t,0), \nabla_{\frac{\partial}{\partial t}} c'(t,0) \rangle dt - \int_a^b \langle R(\dot{c}, c')c', \dot{c} \rangle dt_{|s=0} + \langle \nabla_{\frac{\partial}{\partial s}} c', \dot{c} \rangle |_{t=a,s=0}^{t=b,s=0} \tag{4.1.6}$$

and with $c'^{\perp} := c' - \langle \frac{\dot{c}}{\|\dot{c}\|}, c' \rangle \frac{\dot{c}}{\|\dot{c}\|}$ (the component of c' orthogonal to $\dot{c}$),

$$L''(0) = \frac{1}{\|\dot{c}\|} \left\{ \int_a^b (\langle \nabla_{\frac{\partial}{\partial t}} c'^{\perp}, \nabla_{\frac{\partial}{\partial t}} c'^{\perp} \rangle - \langle R(\dot{c}, c'^{\perp})c'^{\perp}, \dot{c} \rangle) \, dt + \langle \nabla_{\frac{\partial}{\partial s}} c', \dot{c} \rangle |_{t=a}^{t=b} \right\} \Big|_{s=0}. \tag{4.1.7}$$

An important point is that for a geodesic c, the second variation depends only on the first derivative $\frac{\partial}{\partial s} c(t,s)_{|s=0}$ of the variation, but not on higher derivatives. This fact will allow the definition of the index form I below.

Proof. According to the formulae of the proof of lemma 4.1.1,

$$\begin{aligned}
\frac{d^2}{ds^2}E(s) &= \int_a^b \frac{\partial}{\partial s}\Big\langle \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial s}(t,s), \frac{\partial c}{\partial t}(t,s)\Big\rangle\, dt \\
&= \int_a^b \Big\langle \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial s}(t,s), \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial s}(t,s)\Big\rangle\, dt \\
&\quad + \int_a^b \Big\langle \nabla_{\frac{\partial}{\partial s}} \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial s}(t,s), \frac{\partial c}{\partial t}(t,s)\Big\rangle\, dt \qquad \text{again,} \\
&\qquad\qquad \text{since } \nabla \text{ is metric and torsion free} \\
&= \int_a^b \Big\langle \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial s}(t,s), \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial s}(t,s)\Big\rangle\, dt \\
&\quad + \int_a^b \Big\langle \nabla_{\frac{\partial}{\partial t}} \nabla_{\frac{\partial}{\partial s}} \frac{\partial c}{\partial s}(t,s), \frac{\partial c}{\partial t}(t,s)\Big\rangle\, dt \\
&\quad - \int_a^b \Big\langle R\Big(\frac{\partial c}{\partial t}, \frac{\partial c}{\partial s}\Big)\frac{\partial c}{\partial s}, \frac{\partial c}{\partial t}\Big\rangle\, dt \qquad \text{by definition of } R.
\end{aligned}$$

Since c is geodesic, we have $\nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial t}(t,0) = 0$, and conclude

$$\begin{aligned}
\frac{d^2}{ds^2}E(0) &= \int_a^b \Big\langle \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial s}(t,0), \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial s}(t,0)\Big\rangle\, dt \\
&\quad - \int_a^b \Big\langle R\Big(\frac{\partial c}{\partial t}, \frac{\partial c}{\partial s}\Big), \frac{\partial c}{\partial s}\frac{\partial c}{\partial t}\Big\rangle\, dt\Big|_{s=0} \\
&\quad + \Big\langle \nabla_{\frac{\partial}{\partial s}} \frac{\partial c}{\partial s}, \frac{\partial c}{\partial t}\Big\rangle\Big|_{t=a,s=0}^{t=b,s=0}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{d^2}{ds^2}L(0) = &\quad \int_a^b \frac{\partial}{\partial s}\left(\frac{\langle \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial s}(t,s), \frac{\partial c}{\partial t}(t,s)\rangle}{\langle \frac{\partial c}{\partial t}(t,s), \frac{\partial c}{\partial t}(t,s)\rangle^{\frac{1}{2}}}\right) dt\Big|_{s=0} \\
= &\quad \frac{1}{\|\dot c\|}\Bigg\{\int_a^b \Big\langle \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial s}(t,0), \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial s}(t,0)\Big\rangle\, dt - \int_a^b \Big\langle R\Big(\frac{\partial c}{\partial t}, \frac{\partial c}{\partial s}\Big)\frac{\partial c}{\partial s}, \frac{\partial c}{\partial t}\Big\rangle\, dt\Big|_{s=0} \\
&\qquad + \Big\langle \nabla_{\frac{\partial}{\partial s}} \frac{\partial c}{\partial s}, \frac{\partial c}{\partial t}\Big\rangle\Big|_{t=a,s=0}^{t=b,s=0}\Bigg\} \\
&- \frac{1}{\|\dot c\|^3}\int_a^b \Big(\Big\langle \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial s}(t,0), \frac{\partial c}{\partial t}(t,0)\Big\rangle\Big)^2 dt \\
= &\quad \frac{1}{\|\dot c\|}\Bigg\{\int_a^b \Big\langle \nabla_{\frac{\partial}{\partial t}}\Big(c' - \big\langle \tfrac{\dot c}{\|\dot c\|}, c'\big\rangle \tfrac{\dot c}{\|\dot c\|}\Big), \nabla_{\frac{\partial}{\partial t}}\Big(c' - \big\langle \tfrac{\dot c}{\|\dot c\|}, c'\big\rangle \tfrac{\dot c}{\|\dot c\|}\Big)\Big\rangle\, dt \\
&\qquad - \int_a^b \Big\langle R\Big(\frac{\partial c}{\partial t}, \frac{\partial c}{\partial s}\Big)\frac{\partial c}{\partial s}, \frac{\partial c}{\partial t}\Big\rangle\, dt \; + \Big\langle \nabla_{\frac{\partial}{\partial s}} \frac{\partial c}{\partial s}, \frac{\partial c}{\partial t}\Big\rangle\Big|_{t=a}^{t=b}\Bigg\}\Bigg|_{s=0}.
\end{aligned}$$

Also
$$\langle R(\dot c, c')c', \dot c\rangle = \left\langle R(\dot c, c' - \langle \tfrac{\dot c}{\|\dot c\|}, c'\rangle \tfrac{\dot c}{\|\dot c\|})(c' - \left\langle \tfrac{\dot c}{\|\dot c\|}, c'\right\rangle \tfrac{\dot c}{\|\cdot c\|}),\ \dot c\right\rangle,$$
so that for the second variation of L through a proper variation, only the component of the variation vector field $\frac{\partial c}{\partial s}$ orthogonal to $\dot c$ appears. □

In the same manner, we may consider closed geodesics $c : S^1 \to M$. The formulae for the second variations of E and L then of course do not contain any boundary terms anymore. Otherwise, they remain the same.

We can already draw some consequences:

If the sectional curvature of M is nonpositive, the curvature term in the second variation formula is always nonnegative, because of the negative sign in front of it. The first term only vanishes for parallel variations and is positive otherwise. If we consider a proper variation that is nontrivial, i.e. $c' \neq 0$, we get $\frac{d^2}{ds^2}E(0) > 0$, hence $E(c_s) > E(c_0)$ for sufficiently small $|s|$. We conclude

Corollary 4.1.1. *On a manifold with nonpositive sectional curvature, geodesics with fixed endpoints are always locally minimizing.*

(Here, "locally minimizing" means that there exists some $\delta > 0$ such that for any (smooth) curve $\gamma : [a, b] \to M$ with $\gamma(a) = c(a), \gamma(b) = c(b)$ and $d(\gamma(t), c(t)) \leq \delta$ for all $t \in [a, b]$, we have $E(\gamma) \geq E(c)$.)

Proof. Let $c : [a, b] \to M$ be a smooth geodesic, and let $\gamma : [a, b] \to M$ be another curve with $\gamma(a) = c(a), \gamma(b) = c(b)$, and such that for no $t \in [a, b]$, the distance between $\gamma(t)$ and $c(t)$ exceeds the injectivity radius of $c(t)$. We may then find a smooth geodesic interpolation between c and γ, namely the family $c(t, s) := \exp_{c(t)} s \exp_{c(t)}^{-1} \gamma(t)$, i.e. a family that satisfies $c(t, 0) = c(t), c(t, 1) = \gamma(t)$ for all $t \in [a, b]$, and for which all the curves $c(t, s)$ for fixed t and s varying in $[0, 1]$ are geodesic. Thus, $\frac{\nabla_\partial}{\partial s} \frac{\partial c}{\partial s}(t, s) = 0$ for all t and s, and from the proof of Theorem 4.1.1 $\frac{d^2}{ds^2}E(s) \geq 0$ for all $s \in [0, 1]$, not only for $s = 0$. Since $\frac{d}{ds}E(s)_{|s=0} = 0$ as c is geodesic, we conclude $E(\gamma) \geq E(c)$. (Since we may assume that γ is parametrized proportionally to arclength, we also get $L(\gamma) \geq L(c)$.) □

Although it is a general fact that sufficiently short geodesics are minimizing (cf. §1.4), on a positively curved manifold, longer geodesics need not be minimizing anymore, as is already seen on S^2.

Similarly,

Corollary 4.1.2. *On a manifold with negative sectional curvature, closed geodesics are strict local minima of E (and L) (except for reparametrizations).*

Proof. For each variation normal to $\dot c$ the curvature term is positive, because of the negative sign in front of it. □

On a manifold with vanishing curvature, geodesics are still minimizing, but not necessarily strictly so anymore, as the example of a flat torus or cylinder shows. On a manifold with positive curvature, closed geodesics in general do not minimize anymore, see S^2 again. We want to derive a global consequence of this fact.

Theorem 4.1.2 (Synge). *Any compact oriented even-dimensional Riemannian manifold with positive sectional curvature is simply connected.*

Proof. Otherwise, there exists a nontrivial element of $\pi_1(M, x_0)$ (let $x_0 \in M$ be the base point). Let this element be represented by a closed curve $\gamma : S^1 \to M$. γ cannot be homotopic to a constant curve even if we do not keep the base point fixed. On the other hand, by Theorem 1.5.1, γ is homotopic to a closed geodesic c of shortest length (and smallest energy) in this free homotopy class. Thus, $c : S^1 \to M$ cannot be a constant curve.

Parallel transport P along c from $c(0)$ to $c(2\pi) = c(0)$ is orientation preserving and leaves the orthogonal complement E of $\dot{c}(0)$ invariant. Since E has odd dimension (since M has an even one), there exists a vector $v \in E$ with $Pv = v$.

Let now X be the parallel vector field along c with $X(0) = v$. We consider a variation $c : S^1 \times (-\varepsilon, \varepsilon) : (t, s) \mapsto c(t, s)$ of c with $c'(t, 0) = X(t)$ for all t.

Since c is geodesic, $E'(0) = 0$. Since X is parallel and $X(0) = X(2\pi)$,

$$\begin{aligned} E''(0) &= \int_0^{2\pi} \langle \nabla_{\frac{\partial}{\partial t}} X(t), \nabla_{\frac{\partial}{\partial t}} X(t) \rangle dt - \int_0^{2\pi} \langle R(\dot{c}, X)X, \dot{c} \rangle dt \\ &= -\int_0^{2\pi} \langle R(\dot{c}, X)X, \dot{c} \rangle dt \\ &< 0. \end{aligned}$$

Hence

$$E(c_s) < E(c) \quad \text{for sufficiently small } s,$$

and c cannot have least energy in its homotopy class.

This contradiction proves the claim. □

Remark. The previous reasoning would have applied to L instead of E as well.

Let now X be a vector field along c, i.e. a section of $c^*(TM)$; in the sequel, c will always be geodesic. There exists a variation $c : [a, b] \times (-\varepsilon, \varepsilon) \to M$ of $c(t)$ with $\frac{\partial c}{\partial s}_{|s=0} = X$.

We put

$$I(X, X) := \int_a^b \Big(\langle \nabla_{\frac{\partial}{\partial t}} X, \nabla_{\frac{\partial}{\partial t}} X \rangle - \langle R(\dot{c}, X)X, \dot{c} \rangle \Big) dt,$$

i.e.

$$I(X, X) = \frac{d^2}{ds^2} E(0), \text{ if } X(a) = 0 = X(b).$$

Instead of a 1-parameter variation $c(t,s)$, we may also consider a 2-parameter variation and put ($Y := \frac{\partial c}{\partial r}$)

$$I(X,Y) := \int_a^b (\langle \nabla_{\frac{\partial}{\partial t}} X, \nabla_{\frac{\partial}{\partial t}} Y\rangle - \langle R(\dot{c}, X)Y, \dot{c}\rangle)dt. \tag{4.1.8}$$

$I(X,Y)$ is bilinear and symmetric in X and Y (by (3.3.10)).

Definition 4.1.1. I is called the *index form* of the geodesic c.

For a vector field X along c that is only piecewise differentiable, we define $I(X,X)$ as the sum of the respective expressions on those subintervals where X is differentiable. Each piecewise smooth vector field X along c may be approximated by smooth vector fields X_n in such a manner that $I(X_n, X_n)$ converges to $I(X,X)$. For technical purposes, it is useful, however, to consider piecewise smooth vector fields. A variation that is piecewise C^2 gives rise to a piecewise C^1 vector field, and vice versa.

4.2 Jacobi Fields

Definition 4.2.1. Let $c : I \to M$ be geodesic. A vector field X along c is called a *Jacobi field* if

$$\nabla_{\frac{d}{dt}} \nabla_{\frac{d}{dt}} X + R(X, \dot{c})\dot{c} = 0\,. \tag{4.2.1}$$

As an abbreviation, we shall sometimes write

$$\dot{X} = \nabla_{\frac{d}{dt}} X, \quad \ddot{X} = \nabla_{\frac{d}{dt}} \nabla_{\frac{d}{dt}} X,$$

(4.2.1) then becomes

$$\ddot{X} + R(X, \dot{c})\dot{c} = 0\,. \tag{4.2.2}$$

Lemma 4.2.1. *A vector field X along a geodesic $c : [a,b] \to M$ is a Jacobi field if and only if the index form of c satisfies*

$$I(X,Y) = 0$$

for all vector fields Y along c with $Y(a) = Y(b) = 0$.

Proof.

$$\begin{aligned} I(X,Y) &= \int_a^b (\langle \nabla_{\frac{d}{dt}} X, \nabla_{\frac{d}{dt}} Y\rangle - \langle R(X,\dot{c})\dot{c}, Y\rangle)\, dt, \\ &\qquad \text{using the symmetries of the curvature tensor} \\ &= \int_a^b (\langle -\nabla_{\frac{d}{dt}} \nabla_{\frac{d}{dt}} X, Y\rangle - \langle R(X,\dot{c})\dot{c}, Y\rangle)\, dt, \\ &\qquad \text{since } \nabla \text{ is metric and } Y(a) = 0 = Y(b), \end{aligned}$$

and this vanishes for all Y if

$$\nabla_{\frac{d}{dt}}\nabla_{\frac{d}{dt}}X + R(X,\dot{c})\dot{c} = 0$$

holds (by the fundamental lemma of the calculus of variations). □

Lemma 4.2.2. *A vector field X along the geodesic $c : [a,b] \to M$ is a Jacobi field if and only if it is a critical point of $I(X,X)$ w.r.t. all variations with fixed endpoints, i.e.*

$$\frac{d}{ds}I(X+sY, X+sY)_{|s=0} = 0$$

for all vector fields Y along c with $Y(a) = 0 = Y(b)$.

Proof. We compute

$$\frac{d}{ds}I(X+sY, X+sY)_{|s=0} =$$
$$2\int_a^b (-\langle\nabla_{\frac{\partial}{\partial t}}\nabla_{\frac{\partial}{\partial t}}X, Y\rangle - \langle R(X,\dot{c})\dot{c}, Y\rangle)\, dt \quad \text{by the proof of Lemma (4.2.1).}$$

□

The Jacobi equation thus is the Euler-Lagrange equation for

$$I(X) := I(X,X).$$

More generally, one can consider the second variation for each critical point of a variational problem. The second variation then is a quadratic integral in the variation vector fields, and the second variation may hence be considered as a new variational problem. This new variational problem is called accessory variational problem of the original one. Most of the considerations of this Paragraph may be generalized to such accessory variational problems.

We now want to prove existence and uniqueness of Jacobi fields with given initial values. For this purpose, we shall simply interpret the Jacobi equation as a system of $d(=\dim M)$ linear second order ODEs.

Lemma 4.2.3. *Let $c : [a,b] \to M$ be geodesic. For any $v, w \in T_{c(a)}M$, there exists a unique Jacobi field X along c with*

$$X(a) = v, \dot{X}(a) = w.$$

Proof. Let $v_1, \ldots, v_d$ be an orthonormal basis of $T_{c(a)}M$. Let $X_1, \ldots, X_d$ be parallel vector fields along c with $X_i(a) = v_i,\ \ i = 1, \ldots, d$. Then, for each $t \in [a,b]$, $X_1(t), \ldots, X_d(t)$ is an orthonormal base of $T_{c(t)}M$. An arbitrary vector field X along c is written as

$$X = \xi^i X_i \quad (\xi^i(t) = \langle X(t), X_i(t)\rangle).$$

Since the vector fields X_i are parallel, we have

$$\nabla_{\frac{d}{dt}} X = \frac{d\xi^i}{dt} X_i, \nabla_{\frac{d}{dt}} \nabla_{\frac{d}{dt}} X = \frac{d^2\xi^i}{dt^2} X_i.$$

We likewise write the curvature term in (4.2.1) as a linear combination of the X_k :

$$R(X_i, \dot{c})\dot{c} = \rho_i^k X_k;$$

and then also

$$R(X, \dot{c})\dot{c} = \xi^i \rho_i^k X_k.$$

The Jacobi equation (4.2.1) now becomes

$$\left(\frac{d^2\xi^k}{dt^2} + \xi^i \rho_i^k \right) X_k = 0,$$

i.e. a system of d linear 2nd order ODE

$$\frac{d^2\xi^k(t)}{dt^2} + \xi^i(t)\rho_i^k(t) = 0, \quad k = 1, \ldots, d,$$

and for such systems, the desired existence and uniqueness result is valid. □

It is easy to describe those Jacobi fields that are tangential to c.

Lemma 4.2.4. *Let $c : [a, b] \to M$ be geodesic, $\lambda, \mu \in \mathbb{R}$. Then the Jacobi field X along c with $X(a) = \lambda\dot{c}(a), \dot{X}(a) = \mu\dot{c}(a)$ is given by*

$$X(t) = (\lambda + (t - a)\mu)\dot{c}(t).$$

Proof. Directly from (4.2.1), since $R(\dot{c}, \dot{c}) = 0$ because of the skew symmetry of R. □

Thus, tangential Jacobi fields do not depend at all on the geometry of M, and hence, they cannot yield any information about the geometry of M. Consequently, they are without any interest for us. We shall see in the sequel, however, that normal Jacobi fields are extremely useful tools for studying the geometry of Riemannian manifolds.

Examples.

1. In Euclidean space $\mathbb{R}^n$, geodesics are straight lines. Jacobi fields are linear: Namely, the Jacobi field X along a straight line c with c mit $X(a) = v, \dot{X}(a) = w$ is given by

$$X(t) = V(t) + (t - a)W(t), \tag{4.2.3}$$

where $V(t)$ and $W(t)$ are parallel fields along c with $V(a) = v, W(a) = w$.

2. $S^n \subset \mathbb{R}^{n+1}$. Let $c : [0,T] \to S^n$ be geodesic with $\|\dot c\| \equiv 1$, $v, w \in T_{c(0)}S^n$, V, W parallel vector fields along c with $V(0) = v, W(0) = w$. Assume $\langle v, \dot c(0)\rangle = 0 = \langle w, \dot c(0)\rangle$. We claim that the Jacobi field X with $X(0) = v, \dot X(0) = w$ along c is given by
$$X(t) = V(t)\cos t + W(t)\sin t. \tag{4.2.4}$$
Namely, since V and W are parallel,
$$\dot X(t) = -V(t)\sin t + W(t)\cos t$$
$$\ddot X(t) = -V(t)\cos t - W(t)\sin t.$$
By (3.6.12),
$$R(X, \dot c)\dot c = \langle \dot c, \dot c\rangle X - \langle X, \dot c\rangle \dot c = X, \quad \text{since } \langle \dot c, \dot c\rangle = 1$$
and since v and w, hence also V and W are orthogonal to $\dot c$.
Hence,
$$\ddot X + R(X, \dot c)\dot c = 0,$$
and X indeed is a Jacobi field.

Arbitrary initial values that are not necessarily orthogonal to $\dot c$ may be split into a tangential and a normal part. The desired Jacobi field then is the sum of the corresponding tangential and normal ones, because as (4.2.1) is linear the sum of two solutions of (4.2.1) is a solution again.

If more generally $\|\dot c\| = \mu$, the Jacobi field with initial values v, w normal to $\dot c$ is given by
$$X(t) = V(t)\cos(\mu t) + W(t)\sin(\mu t). \tag{4.2.5}$$
If we consider more generally the sphere
$$S^n_\rho := \{x \in \mathbb{R}^{n+1} : |x| = \rho\}$$
of radius ρ, then the curvature is given by
$$R(X,Y)Z = \frac{1}{\rho^2}(\langle Y, Z\rangle X - \langle X, Z\rangle Y)$$
and the Jacobi field with initial values v, w normal to $\dot c$ with $\|\dot c\| = 1$
$$X(t) = V(t)\cos\frac{t}{\rho} + \rho W(t)\sin\frac{t}{\rho}. \tag{4.2.6}$$

Theorem 4.2.1. *Let $c : [0,T] \to M$ be geodesic. Let $c(t,s)$ be a variation of $c(t)$ ($c(\cdot,\cdot) : [0,T] \times (-\varepsilon, \varepsilon) \to M$), for which all curves $c(\cdot, s) =: c_s(\cdot)$ are geodesics, too. Then,*
$$X(t) := \frac{\partial}{\partial s} c(t,s)_{|s=0}$$
is a Jacobi field along $c(t) = c_0(t)$. Conversely, every Jacobi field along $c(t)$ may be obtained in this way, i.e. by a variation of $c(t)$ through geodesics.

Proof.

$$\begin{aligned}
\ddot{X}(t) &= \nabla_{\frac{\partial}{\partial t}} \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial s}_{|s=0} \\
&= \nabla_{\frac{\partial}{\partial t}} \nabla_{\frac{\partial}{\partial s}} \frac{\partial c}{\partial t}_{|s=0} \\
&= \nabla_{\frac{\partial}{\partial s}} \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial t}_{|s=0} - R\left(\frac{\partial c}{\partial s}, \frac{\partial c}{\partial t}\right) \frac{\partial c}{\partial t}_{|s=0} \quad \text{by definition of } R \\
&= -R\left(\frac{\partial c}{\partial s}, \frac{\partial c}{\partial t}\right) \frac{\partial c}{\partial t}_{|s=0}, \quad \text{since all curves } c_s \text{ are geodesic} \\
&= -R\left(X, \frac{\partial c}{\partial t}\right) \frac{\partial c}{\partial t} \quad \text{by definition of } X.
\end{aligned}$$

Thus, X indeed is a Jacobi field.

Conversely, let X be a Jacobi field along $c(t)$. Let γ be the geodesic $\gamma : (-\varepsilon, \varepsilon) \to M$ with $\gamma(0) = c(0), \gamma'(0) = X(0)$.

Let V and W be parallel vector fields along γ with

$$V(0) = \dot{c}(0), W(0) = \dot{X}(0).$$

We put

$$c(t,s) := \exp_{\gamma(s)}(t(V(s) + sW(s))). \tag{4.2.7}$$

Then all curves $c(\cdot, s) = c_s(\cdot)$ are geodesic (by definition of the exponential map), and $c(t,0) = \exp_{c(0)} t\dot{c}(0) = c(t)$. Thus, $c(t,s)$ is a variation of $c(t)$ through geodesics. By the first part of the proof,

$$Y(t) := \frac{\partial}{\partial s} c(t,s)_{|s=0}$$

then is a Jacobi field along c_0. Finally,

$$\begin{aligned}
Y(0) &= \frac{\partial}{\partial s}(\exp_{\gamma(s)} 0)_{|s=0} \\
&= \frac{\partial}{\partial s}\gamma(s)_{|s=0} \\
&= X(0), && \text{by definition of } \gamma \\
\dot{Y}(0) &= \nabla_{\frac{\partial}{\partial t}} \frac{\partial}{\partial s} c(t,s)_{|s=0} \\
&= \nabla_{\frac{\partial}{\partial s}} \frac{\partial}{\partial t} c(t,s)_{|s=0}, && \text{since } \nabla \text{ is torsion free} \\
&= \nabla_{\frac{\partial}{\partial s}} (V(s) + sW(s))_{|s=0} \\
&= W(0) && \text{since } V \text{ and } W \text{ are parallel along } \gamma \\
&= \dot{X}(0)\,.
\end{aligned}$$

Thus, Y is a Jacobi field along c_0 with the same initial values $Y(0), \dot{Y}(0)$ as X. The uniqueness result of Lemma 4.2.3 implies $X = Y$. We have thus shown that X may be obtained from a variation of $c(t)$ through geodesics. □

The computation at the beginning of the previous proof reveals the geometric origin of the Jacobi equation:

Let $c(t,s) = c_s(t)$ be a family of geodesics parametrized by s, i.e.

$$\nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial t}(t,s) = 0 \quad \text{for all}\, s.$$

Then also

$$\nabla_{\frac{\partial}{\partial s}} \nabla_{\frac{\partial}{\partial t}} \frac{\partial c}{\partial t}(t,s) = 0,$$

and this implies that $X(t) = \frac{\partial c}{\partial s}(t,s)_{|s=0}$ satisfies the Jacobi equation. Consequently, the Jacobi equation is the linearization of the equation for geodesic curves. This also illuminates the relation between Jacobi fields and the index form. If one has in particular a *proper* variation of a geodesic through *geodesics,* then also the 2nd derivative of the length and energy functionals w.r.t. the family parameter vanish.

As an example, consider the family of great semicircles on S^n through two fixed antipodal points, e.g. north pole and south pole. Here, the length is even constant on the whole family.

The theory of Jacobi fields can be generalized to other variational problems, and actually, this theory was already conceived by Jacobi in general form.

Corollary 4.2.1. *Every Killing field X on M is a Jacobi field along any geodesic c in M.*

Proof. By Lemma 1.9.7, a Killing field X generates a local 1-parameter group of isometries. Isometries map geodesics to geodesics. Thus, X generates a variation of c through geodesics. Theorem 4.2.1 then implies the claim. □

Corollary 4.2.2. *Let $c : [0,T] \to M$ be a geodesic, $p = c(0)$, i.e.*

$$c(t) = \exp_p t\dot{c}(0).$$

For $w \in T_pM$, the Jacobi field X along c with $X(0) = 0, \dot{X}(0) = w$ then is given by

$$X(t) = (D\exp_p)(t\dot{c}(0))(tw) \quad \text{or, in different notation,} \quad D_{t\dot{c}(0)} \exp_p(tw) \tag{4.2.8}$$

(the derivative of the exponential map $\exp_p : T_pM \to M$, evaluated at the point $t\dot{c}(0) \in T_pM$ and applied to tw).

Proof. $c(t,s) := \exp_p t(\dot{c}(0) + sw)$ is a variation of $c(t)$ through geodesics, and by Theorem 4.2.1, the corresponding Jacobi field is

$$X(t) = \frac{\partial}{\partial s} c(t,s)_{|s=0} = (D\exp_p)(t\dot{c}(0))(tw),$$

and

$$X(0) = (D\exp_p)(0)(0) = 0,$$
$$\dot{X}(0) = w \qquad \text{(as in the proof of Theorem 4.2.1).}$$

□

Consequently, the derivative of the exponential map can be computed from Jacobi fields along radial geodesics.

Corollary 4.2.2 yields an alternative method for a quick computation of the curvature tensor of S^n. Let $x_0 \in S^n, z \in T_{x_0}S^n$ with $\|z\| = 1$. The geodesic $c : \mathbb{R} \to S^n$ with $c(0) = x_0, \dot{c}(0) = z$ then is given by

$$c(t) = (\cos t)x_0 + (\sin t)z\,.$$

Let $w \in T_{x_0}S^n$, $\|w\| = 1, \langle w, z\rangle = 0$,

$$c(t, s) = (\cos t)x_0 + (\sin t)((\cos s)z + (\sin s)w)$$

then is a variation of $c(t)$ through geodesics. Furthermore, the vector field along $c(t)$ defined by $W(t) = w$ is parallel (cf. Theorem 3.4.1). Hence, the corresponding Jacobi field is

$$X(t) = \frac{\partial}{\partial s}c(t, s)_{|s=0} = (\sin t)W(t) \qquad \text{(cf. (4.2.4))}.$$

We have

$$\ddot{X}(t) + X(t) = 0.$$

The Jacobi equation then implies

$$X(t) = R(X(t), \dot{c})\dot{c},$$

and in particular

$$\langle R(w, z)z, w\rangle = 1 = \langle w, w\rangle\langle z, z\rangle - \langle w, z\rangle^2.$$

Lemma 3.3.3 implies

$$\langle R(u, v)w, z\rangle = \langle u, z\rangle\langle v, w\rangle - \langle u, w\rangle\langle v, z\rangle, \quad \text{i.e. (3.4.13).}$$

Another consequence is the so called *Gauss lemma:*

Corollary 4.2.3. *Let* $p \in M, v \in T_pM, c(t) := \exp_p tv$ *the geodesic with* $c(0) = p, \dot{c}(0) = v$ $(t \in [0, 1])$, *assuming that* v *is contained in the domain of definition of* $\exp_p$. *Then for any* $w \in T_pM$

$$\langle v, w\rangle = \langle (D_v \exp_p)v, (D_v \exp_p)w\rangle, \tag{4.2.9}$$

where $D_v \exp_p$, *the derivative of* $\exp_p$ *at the point* v, *is applied to the vectors* v *and* w *considered as vectors tangent to* T_pM *at the point* v.

Proof. By Corollary 4.2.2,
$$X(t) = D_{tv} \exp_p(tw) \tag{4.2.10}$$
is a Jacobi field along c, and
$$\dot{X}(0) = w,$$
and hence
$$\langle v, w\rangle = \langle \dot{c}(0), \dot{X}(0)\rangle. \tag{4.2.11}$$
We split $X(t)$ into a part $X^{\tan}$ tangential to c and a part X^{nor} normal to c.

By Lemma 4.2.4
$$X^{\tan}(t) = t\mu\dot{c}(t), \text{ with } \dot{X}^{\tan}(0) = \mu\dot{c}(0). \tag{4.2.12}$$
Hence
$$\begin{aligned}
\langle v, w\rangle &= \langle \dot{c}(0), \dot{X}^{\tan}(0)\rangle && \text{with (4.2.11) and since}\\
&&& \langle \dot{c}(t), X^{\text{nor}}(t)\rangle \equiv 0\\
&= \langle \dot{c}(1), X^{\tan}(1)\rangle && \text{with (4.2.12)}\\
&= \langle \dot{c}(1), X(1)\rangle && \text{since } \langle \dot{c}(t), X^{\text{nor}}(t)\rangle \equiv 0\\
&= \langle (D_v \exp_p)v, (D_v \exp_p)w\rangle && \text{with (4.2.10).}
\end{aligned}$$
□

(4.2.9) means that $\exp_p$ is a radial isometry in the sense that the length of the radial component of any vector tangent to T_pM is preserved. If a curve $\gamma(s)$ in T_pM intersects the radius orthogonally, then the curve $\exp_p \gamma(s)$ in M intersects the geodesic $c(t) = \exp_p tv$ orthogonally as well. In particular, $c(t) = \exp_p tv$ is orthogonal to the images of all distance spheres in T_pM.

Moreover, we may repeat Corollary 1.4.2:

Corollary 4.2.4. *Let $p \in M$, and let $v \in T_pM$ be contained in the domain of definition of $\exp_p$, and let $c(t) = \exp_p tv$. Let the piecewise smooth curve $\gamma : [0,1] \to T_pM$ be likewise contained in the domain of definition of $\exp_p$, and assume $\gamma(0) = 0, \gamma(1) = v$. Then*
$$\|v\| = L(\exp_p tv_{|t\in[0,1]}) \leq L(\exp_p \circ\gamma), \tag{4.2.13}$$
and equality holds if and only if γ differs from the curve $tv, t \in [0,1]$ only by reparametrization.

Proof. We shall show that any piecewise smooth curve $\gamma : [0,1] \to T_pM$ with $\gamma(0) = 0$ satisfies
$$L(\exp_p \gamma) \geq \|\gamma(1)\|, \tag{4.2.14}$$
with equality precisely for those curves whose image under $\exp_p$ is the radius $t\gamma(1), 0 \leq t \leq 1$. This will then imply (4.2.13).

We write
$$\gamma(t) = r(t)\varphi(t) \quad (r(t) \in \mathbb{R}, \varphi(t) \in T_pM)$$

with $\|\varphi(t)\| \equiv 1$ (polar coordinates in T_pM). Applying the subsequent estimates on any subinterval of $[0,1]$ on which γ is differentiable, we may assume from the onset that γ is smooth everywhere.

We have

$$\dot{\gamma}(t) = \dot{r}(t)\varphi(t) + r(t)\dot{\varphi}(t) \quad \text{with } \langle \varphi(t), \dot{\varphi}(t) \rangle \equiv 0.$$

Thus, by Corollary (4.2.2), also

$$\langle D_{\gamma(t)} \exp \varphi(t), D_{\gamma(t)} \exp \dot{\varphi}(t) \rangle = 0, \quad \|D_{\gamma(t)} \exp \varphi(t)\| = \|\varphi(t)\| = 1,$$

and it follows that

$$\begin{aligned} \|(\exp_p \circ \gamma)^{\cdot}(t)\| &= \|(D_{\gamma(t)} \exp_p)(\dot{\gamma}(t))\| \\ &\geq |\dot{r}(t)|, \end{aligned}$$

hence

$$L(\exp_p \gamma) = \int_0^1 \|(\exp_p \circ \gamma)^{\cdot}(t)\| dt \geq \int_0^1 |\dot{r}(t)| dt \geq r(1) - r(0) = \|\gamma(1)\|,$$

with equality only, if $\dot{\varphi}(t) \equiv 0$ and $r(t)$ is monotone, i.e. if $\gamma(t)$ coincides with the radial curve $t\gamma(1), 0 \leq t \leq 1$ up to reparametrization. □

We point out that alternatively, one can also prove Corollaries 4.2.3 and 4.2.4 with the arguments of the proofs of Theorem 1.4.5 and Corollary 1.4.2.

Corollary 4.2.4 by no means implies that the geodesic $c(t) = \exp_p tv$ is the shortest connection between its end points. It only is shorter than any other curve that is the exponential image of a curve with the same initial and end points as the ray $tv, 0 \leq t \leq 1$.

4.3 Conjugate Points and Distance Minimizing Geodesics

Definition 4.3.1. Let $c : I \to M$ be geodesic. For $t_0, t_1 \in I, t_0 \neq t_1$, $c(t_0)$ and $c(t_1)$ are called *conjugate along* c if there exists a Jacobi field $X(t)$ along c that does not vanish identically, but satisfies

$$X(t_0) = 0 = X(t_1).$$

Of course, such a Jacobi field X is always normal to c (Lemma 4.2.4). If $t_0, t_1 \in I, t_0 \neq t_1$, are not conjugate along c, then for $v \in T_{c(t_0)}M, w \in T_{c(t_1)}M$, there exists a unique Jacobi field $Y(t)$ along c with $Y(t_0) = v, Y(t_1) = w$. Namely, let J_c be the vector space of Jacobi fields along c ($\dim J_c = 2 \dim M$ by Lemma 4.2.3). We define a linear map

$$A : J_c \to T_{c(t_0)}M \times T_{c(t_1)}M$$

by

$$A(Y) = (Y(t_0), Y(t_1)).$$

Since t_0 and t_1 are not conjugate along c, the kernel of A is trivial, and A is injective, hence bijective as domain and range of A have the same dimension.

Theorem 4.3.1. *Let $c : [a, b] \to M$ be geodesic.*

(i) If there does not exist a point conjugate to $c(a)$ along c, then there exists $\varepsilon > 0$ with the property that for any piecewise smooth curve

$$g : [a, b] \to M$$

with $g(a) = c(a), g(b) = c(b), d(g(t), c(t)) < \varepsilon$ for all $t \in [a, b]$, we have

$$L(g) \geq L(c) \tag{4.3.1}$$

with equality if and only if g is a reparametrization of c.

(ii) If there does exist $\tau \in (a, b)$ for which $c(a)$ and $c(\tau)$ are conjugate along c, then there exists a proper variation

$$c(t, s) : [a, b] \times (-\varepsilon, \varepsilon) \to M$$

with

$$L(c_s) < L(c) \quad \text{for } 0 < |s| < \varepsilon \quad (c_s(t) := c(t, s)). \tag{4.3.2}$$

Proof.

(i) We want to apply Corollary 4.2.4. We therefore have to show that in the absence of conjugate points, for each curve as in (i), there exists a curve γ as described in Corollary 4.2.4. W.l.o.g. $a = 0, b = 1$. We put $v := \dot{c}(0)$.

By Corollary 4.2.2, since there are no conjugate points along c, the exponential map $\exp_p$ is of maximal rank along any radial curve $tv, 0 \leq t \leq 1$. Thus, by the inverse function theorem, for each such t, $\exp_p$ is a diffeomorphism in a suitable neighborhood of tv. We cover $\{tv, 0 \leq t \leq 1\}$ by finitely many such neighborhoods $\Omega_i, i = 1, \ldots, k; U_i := \exp_p \Omega_i$.

Let us assume

$$tv \in \Omega_i \quad \text{for } t_{i-1} \leq t \leq t_i \quad (t_0 = 0, t_k = 1).$$

If $\varepsilon > 0$ is sufficiently small, we have for any curve $g : [0, 1] \to M$ satisfying the assumptions of (i),

$$g([t_{i-1}, t_i]) \subset U_i. \tag{4.3.3}$$

We now claim that for any g satisfying (4.3.3), there exists a curve γ in T_pM with $\exp_p \gamma = g, \gamma(0) = 0, \gamma(1) = v$.

For this, we simply put

$$\gamma(t) = (\exp_{p|\Omega_i})^{-1}(g(t)) \quad \text{for } t_{i-1} \leq t \leq t_i$$

γ then satisfies the assumption of Corollary 4.2.4, and we obtain (i).

(ii) Again, w.l.o.g. $a = 0, b = 1$. Let X be a nontrivial Jacobi field along c with $X(0) = 0 = X(\tau)$. We have $\dot{X}(\tau) \neq 0$, since otherwise $X \equiv 0$ by the uniqueness result of Lemma 4.2.3. Let now $Z(t)$ be an arbitrary vector field along c with

$$Z(0) = 0 = Z(1), Z(\tau) = -\dot{X}(\tau).$$

For $\eta > 0$, we put

$$\begin{aligned} Y_\eta^1(t) &:= X(t) + \eta Z(t) && \text{for } 0 \leq t \leq \tau, \\ Y_\eta^2(t) &:= \eta Z(t) && \text{for } \tau \leq t \leq 1. \end{aligned}$$

$$Y_\eta(t) := \begin{cases} Y_\eta^1(t) & \text{for } 0 \leq t \leq \tau, \\ Y_\eta^2(t) & \text{for } \tau \leq t \leq 1. \end{cases}$$

With $Z^1 := Z_{|[0,\tau]}, Z^2 := Z_{|[\tau,1]}$ we have

$$\begin{aligned} I(Y_\eta^1, Y_\eta^1) &= \langle \dot{X}(\tau), 2\eta Z(\tau)\rangle + \eta^2 I(Z^1, Z^1) \\ &= -2\eta \|\dot{X}(\tau)\|^2 + \eta^2 I(Z^1, Z^1) \\ I(Y_\eta^2, Y_\eta^2) &= \eta^2 I(Z^2, Z^2). \end{aligned}$$

Hence

$$I(Y_\eta, Y_\eta) = I(Y_\eta^1, Y_\eta^1) + I(Y_\eta^2, Y_\eta^2) = -2\eta \|\dot{X}(\tau)\|^2 + \eta^2 I(Z, Z)$$

for sufficiently small $\eta > 0$. The variation $c(t, s) := \exp_{c(t)} sY_\eta(t)$ then satisfies (with $L(s) := L(c_s)$)

$$L'(0) = 0, L''(0) = I(Y_\eta, Y_\eta) < 0,$$

and the claim follows from Taylor's theorem.

□

Theorem 4.3.1 (i) implies only that in the absence of conjugate points, a geodesic is length minimizing when compared with sufficiently close curves. As is seen by considering geodesics on a flat cylinder or torus that wind around more than once, even when there are no conjugate points, a geodesic need not be the global shortest connection between its end points.

On the sphere S^n, on any geodesic the first point conjugate to the initial point is reached precisely after travelling a semi circle (see (4.2.4)). By Theorem 4.3.1 consequently each geodesic arc shorter than a great semi circle, i.e. shorter than π, is locally length minimizing, whereas any geodesic arc on S^n longer than π is not even locally the shortest connection of its end points.

For a curve $c : [a, b] \to M$ let $\mathcal{V}_c$ be the space of vector fields along c, i.e.

$$\mathcal{V}_c = \Gamma(c^* TM),$$

and let $\overset{\circ}{\mathcal{V}}_c$ be the space of vector fields along c satisfying $V(a) = V(b) = 0$.

Lemma 4.3.1. *Let $c : [a, b] \to M$ be geodesic. Then there is no pair of conjugate points along c if and only if the index form I of c is positive definite on $\overset{\circ}{\mathcal{V}}_c$.*

Proof. Assume that c has no conjugate points. Theorem 4.3.1 (i) implies

$$I(X, X) \geq 0 \quad \text{for all } X \in \overset{\circ}{\mathcal{V}}_c, \tag{4.3.4}$$

because otherwise $c(t, s) := \exp_{c(t)} sX(t)$ would be a locally length decreasing deformation. If $I(Y, Y) = 0$ for some $Y \in \overset{\circ}{\mathcal{V}}_c$, then by (4.3.4) for all $Z \in \overset{\circ}{\mathcal{V}}_c, \lambda \in \mathbb{R}$,

$$0 \leq I(Y - \lambda Z, Y - \lambda Z) = -2\lambda I(Y, Z) + \lambda^2 I(Z, Z),$$

and hence $I(Y, Z) = 0$ for all $Z \in \overset{\circ}{\mathcal{V}}_c$. Lemma 4.2.1 then implies that Y is a Jacobi field. Since there are no conjugate points along c, we get $Y = 0$. Hence, I is positive definite.

Now assume that for $t_0, t_1 \in [a, b]$ (w.l.o.g. $t_0 < t_1$), $c(t_0)$ and $c(t_1)$ are conjugate along c. Then there exists a nontrivial Jacobi field X along c with $X(t_0) = 0 = X(t_1)$. We put

$$Y(t) := \begin{cases} 0 & \text{for } a \leq t \leq t_0, \\ X(t) & \text{for } t_0 \leq t \leq t_1, \\ 0 & \text{for } t_1 \leq t \leq b. \end{cases}$$

Then $I(Y, Y) = 0$, and I is not positive definite. □

We now introduce the following norm on $\overset{\circ}{\mathcal{V}}_c$:

$$\|X\| := \Big(\int_a^b (\langle \dot{X}, \dot{X} \rangle + \langle X, X \rangle) dt \Big)^{\frac{1}{2}}. \tag{4.3.5}$$

Let $\overset{\circ}{H}{}^1_c$ be the completion of $\overset{\circ}{\mathcal{V}}_c$ w.r.t. $\| \cdot \|$.

Introducing an orthonormal basis $\{V_i\}$ of parallel vector fields ($i = 1, \dots, d = \dim M$) and writing

$$X = \xi^i V_i,$$

we have $\dot{X} = \dot{\xi}^i V_i$, and

$$\|X\| = \Big(\int_a^b (\dot{\xi}^i \dot{\xi}^i + \xi^i \xi^i) dt \Big)^{\frac{1}{2}}.$$

Hence, $\overset{\circ}{H}{}^1_c$ can be identified with the Sobolev space $\overset{\circ}{H}{}^{1,2}(I, \mathbb{R}^d)$. We now consider the index form of c as a quadratic form on $\overset{\circ}{H}{}^1_c$:

$$\begin{aligned} &I : \overset{\circ}{H}{}^1_c \times \overset{\circ}{H}{}^1_c \to \mathbb{R}, \\ &I(X,Y) = \int_a^b (\langle \dot{X}, \dot{Y} \rangle - \langle R(\dot{c}, X)Y, \dot{c} \rangle)\, dt\,. \end{aligned} \tag{4.3.6}$$

Definition 4.3.2. The *index* of c, $\operatorname{Ind}(c)$, is the dimension of the largest subspace of $\overset{\circ}{H}{}^1_c$, on which I is negative definite, and the *extended index of* c, $\operatorname{Ind}_0(c)$, is the dimension of the largest subspace of $\overset{\circ}{H}{}^1_c$, on which I is negative semidefinite. Finally, the *nullity of* c is

$$N(c) := \operatorname{Ind}_0(c) - \operatorname{Ind}(c).$$

Lemma 4.3.2. $\operatorname{Ind}(c)$ *and* $N(c)$ *are finite.*

Proof. Otherwise, there exists a sequence $(X_n)_{n\in\mathbb{N}}$ with

$$I(X_n, X_n) \le 0 \tag{4.3.7}$$

and

$$\int_a^b \langle X_n, X_m \rangle\, dt = \delta_{nm} \tag{4.3.8}$$

for all $n, m \in \mathbb{N}$. ((4.3.8) means that (X_n) is an orthonormal sequence w.r.t. the L^2-product.)

(4.3.7) and (4.3.8) imply

$$\int_a^b \langle \dot{X}_n, \dot{X}_n \rangle \le \sup |R|\, E(c) \tag{4.3.9}$$

(where R is the curvature tensor of M).

By (4.3.8) and (4.3.9)

$$\|X_n\| \le \text{ const.} \tag{4.3.10}$$

By Rellich's Theorem (Corollary A.1.3), a subsequence converges in L^2. This, however, is not compatible with (4.3.8), since an orthonormal sequence cannot be a Cauchy sequence. $\square$

For $t \in (a, b]$ let J_c^t be the space of Jacobi fields X along c with $X(a) = 0 = X(t)$.

Lemma 4.3.3. $N(c) = \dim J_c^b$.

Proof. From Lemma 4.2.1. □

We now want to derive the **Morse Index Theorem.**

Theorem 4.3.2. *Let $c : [a, b] \to M$ be geodesic. Then there are at most finitely many points conjugate to $c(a)$ along c, and*

$$\operatorname{Ind}(c) = \sum_{t \in (a,b)} \dim J_c^t\,, \tag{4.3.11}$$

$$\operatorname{Ind}_0(c) = \sum_{t \in (a,b]} \dim J_c^t\,. \tag{4.3.12}$$

Proof. For each $t_i \in (a, b]$, for which $c(t_i)$ is conjugate to $c(a)$, there exists a Jacobi field X_i along c with $X_i(a) = 0 = X_i(t_i)$. We put

$$Y_i(t) := \begin{cases} X_i(t) & \text{for } a \le t \le t_i, \\ 0 & \text{otherwise.} \end{cases}$$

The Y_i are linearly independent, and $I(Y_i, Y_i) = 0$ for all i. Therefore, the number of conjugate points is at most $\operatorname{Ind}_0(c)$, hence finite by Lemma 4.3.2.

For $\tau \in (a, b]$, we put

$$\varphi(\tau) := \operatorname{Ind}(c_{|[a,\tau]}),\ \varphi_0(\tau) = \operatorname{Ind}_0(c_{|[a,\tau]}).$$

(i) $\varphi(\tau)$ is left continuous.

(ii) $\varphi_0(\tau)$ is right continuous.

Proof of (i). For $\tau \in (a, b]$ let I_τ be the index form of $c_{|[a,\tau]}$. Let the vector field X along $c_{|[a,\tau]}$ satisfy $I_\tau(X, X) < 0, \|X\| = 1$. We consider the vector field $\tilde{X}$ defined by $\tilde{X}(t) := X(\frac{\tau}{\sigma} t)$ on $[a, \sigma]$. Then

$$\begin{aligned}\int_0^\sigma \langle \dot{\tilde{X}}(t), \dot{\tilde{X}}(t) \rangle\, dt &= \int_0^\sigma \left(\frac{\tau}{\sigma}\right)^2 \left\langle \dot{X}\left(\frac{\tau}{\sigma} t\right), \dot{X}\left(\frac{\tau}{\sigma} t\right) \right\rangle dt \\ &= \left(\frac{\tau}{\sigma}\right) \int_0^\tau \langle \dot{X}(s), \dot{X}(s) \rangle\, ds,\end{aligned}$$

hence

$$\int_0^\sigma \langle \dot{\tilde{X}}(t), \dot{\tilde{X}}(t) \rangle\, dt \to \int_0^\tau \langle \dot{X}(t), \dot{X}(t) \rangle\, dt \quad \text{for } \sigma \to \tau.$$

Moreover, because of $\|X\| = 1$, X is continuous by the Sobolev embedding theorem (Theorem A.1.7). Hence, $\tilde{X}$ also converges pointwise to X as $\sigma \to \tau$, hence also

$$\int_0^\sigma \langle R(\dot{c}, \tilde{X})\tilde{X}, \dot{c} \rangle\, dt \to \int_0^\tau \langle R(\dot{c}, X)X, \dot{c} \rangle\, dt \quad \text{for } \sigma \to \tau.$$

We conclude
$$I_\sigma(\tilde{X},\tilde{X}) \to I_\tau(X,X) \quad \text{for}\, \sigma \to \tau.$$
In particular,
$$I_\sigma(\tilde{X},\tilde{X}) < 0, \ \text{if}\, \sigma \,\text{is sufficiently close to}\, \tau.$$
For each orthonormal basis of a space on which I_τ is negative definite, we may thus find a basis of some space on which I_σ is negative definite, provided σ is sufficiently close to τ.

Since φ is monotonically increasing, this implies the left continuity of φ.

Proof of (ii). Let $(\tau_n)_{n\in\mathbb{N}} \subset (a,b]$ converge to $\tau \in (a,b]$. For each $n \in \mathbb{N}$, let X_n be a vector field along $c_{|[a,\tau_n]}$ with $\|X_n\| = 1$ and $I_{\tau_n}(X_n,X_n) \le 0$. After selecting a subsequence, X_n converges weakly in the Sobolev $H^{1,2}$ topology to some vector field X along $c_{|[a,\tau]}$ (cf. Theorem A.1.9). Then
$$\int_0^\tau \langle \dot{X},\dot{X}\rangle\, dt \le \liminf_{n\to\infty} \int_0^{\tau_n} \langle \dot{X}_n,\dot{X}_n\rangle\, dt.$$

Furthermore, by Rellich's theorem (Corollary A.1.3), X_n also converges (strongly) in L^2, hence
$$\int_0^\tau \langle R(\dot{c},X)X,\dot{c}\rangle\, dt = \lim_{n\to\infty} \int_0^{\tau_n} \langle R(\dot{c},X)X,\dot{c}\rangle\, dt.$$
We conclude
$$I_\tau(X,X) \le \liminf_{n\to\infty} I_{\tau_n}(X_n,X_n) \le 0.$$

We also need to check that X does not vanish identically. Since $I(X_n,X_n) \le 0$, we have
$$\int_0^{\tau_n} \langle \dot{X}_n,\dot{X}_n\rangle\, dt \le \kappa \int_0^{\tau_n} \langle X_n,X_n\rangle\, dt,$$
where the constant κ depends on the norm of $\dot{c}$ and the curvature tensor R. Since the Sobolev norm $\|X_n\| = 1$, this implies that the right hand side cannot go to 0 as otherwise so would the left hand side, and then also $\|X_n\|$ would go to 0. Since X_n converges strongly to X in L^2, by Relich's theorem, the L^2-norm of X is positive as well. Moreover, by a similar argument, if we have two such sequences $(X_n^1),(X_n^2)$, with $\int < X_n^1, X_n^2 > dt = 0$ for all n, then the same holds for the limits X^1, X^2.

Since φ_0 is monotonically increasing, this implies the right continuity of φ_0.

We can now easily conclude the proof of Theorem 4.3.2:

Let $a < t_1 < t_2 < \ldots < t_k \le b$ be the points for which $c(t_i)$ is conjugate to $c(a)$. Lemma 4.3.3 implies
$$\varphi_0(t) - \varphi(t) = 0 \quad \text{for}\, t \in (a,b]\backslash\{t_1,\ldots,t_k\}. \tag{4.3.13}$$

Hence
$$\sum_{t\in(a,b]} \dim J_c^t = \sum_{t\in(a,b]} (\varphi_0(t) - \varphi(t)) = \sum_{i=1}^k (\varphi_0(t_i) - \varphi(t_i)).$$

Since φ is left continuous and φ_0 is right continuous, we have

$$\varphi_0(t_i) = \varphi(t_{i+1}) \quad (i = 1, \ldots, k-1).$$

Hence

$$\sum_{i=1}^{k} (\varphi_0(t_i) - \varphi(t_i)) = \varphi_0(t_k) - \varphi(t_1).$$

Since φ is left continuous, Lemma 4.3.1 implies $\varphi(t_1) = 0$. The continuity properties of φ and φ_0 and (4.3.13) imply that φ and φ_0 can jump only at those points τ where $\varphi_0(\tau) \neq \varphi(\tau)$, i.e. at the conjugate points. In particular, φ_0 is constant on $[t_k, b]$, hence $\varphi_0(t_k) = \varphi_0(b)$. Altogether, we conclude $\varphi_0(b) = sum_{t\in(a,b]} \dim J_c^t$, i.e. (4.3.12). (4.3.11) then follows with the help of Lemma 4.3.3. □

As an application of the second variation, we now present the *Theorem of Bonnet-Myers:*

Corollary 4.3.1. *Let M be a Riemannian manifold of dimension n with Ricci curvature $\geq \lambda > 0$, i.e.*

$$\mathrm{Ric}\,(X, X) \geq \lambda \langle X, X\rangle \quad \text{for all } X \in TM.$$

Let M be complete in the sense that it is closed and any two points can be joined by a shortest geodesic (cf. the Hopf-Rinow Theorem 1.7.1). Then the diameter of M is less or equal to $\pi\sqrt{\frac{n-1}{\lambda}}$. In particular, M is compact. Also, M has finite fundamental group $\pi_1(M)$.

Remark. The diameter is defined as

$$\mathrm{diam}(M) := \sup_{p,q\in M} d(p, q),$$

where $d(\cdot, \cdot)$ denotes the distance function of the Riemannian metric.

The sphere

$$S^n(r) := \{x \in \mathbb{R}^{n+1} : |x| = r\}$$

of radius r has curvature $\frac{1}{r^2}$, hence Ricci curvature $\frac{n-1}{r^2}$ and diameter πr. We choose r such that $\lambda = \frac{n-1}{r^2}$. Corollary 4.3.1 then means that if M has Ricci curvature not less than the one of $S^n(r)$, then the diameter of M is at most the one of $S^n(r)$.

Proof. For each $\rho <$ diam(M), there exist $p, q \in M$ with $d(p, q) = \rho$ and then by the completeness assumption a shortest geodesic arc $c : [0, \rho] \to M$ with $c(0) = p, c(\rho) = q$. Let $e_1, \ldots, e_n$ be an orthonormal basis of T_pM, $e_1 = \dot{c}(0)$. As usual, from this, we may construct a parallel orthonormal basis

$$\{\dot{c}(t), X_2(t), \ldots, X_n(t)\}$$

along c. With $Y_i(t) := (\sin \frac{\pi t}{\rho}) X_i(t), i = 2, \ldots, n$ we have

$$\begin{aligned} I(Y_i, Y_i) &= \int_0^\rho (-\langle \ddot{Y}_i, Y_i \rangle - \langle R(Y_i, \dot{c})\dot{c}, Y_i \rangle) dt \\ &= \int_0^\rho \Big(\sin^2 \frac{\pi t}{\rho} \Big) \Big(\frac{\pi^2}{\rho^2} - \big\langle R(X_i, \dot{c})\dot{c}, X_i \big\rangle \Big)\, dt\,. \end{aligned}$$

Since c is the shortest connection of its end points, by Theorem 4.3.1 (ii), there is no pair of conjugate points in the interior of c, and Lemma 4.3.1 implies

$$I(Y_i, Y_i) \geq 0 \quad \text{for all } i,$$

hence also

$$\begin{aligned} 0 \leq \sum_{i=2}^n I(Y_i, Y_i) &= \int_0^\rho \Big(\sin^2 \frac{\pi t}{\rho} \Big) \Big(\frac{n^2}{\rho^2}(n-1) - \operatorname{Ric}(\dot{c}, \dot{c}) \Big)\, dt \\ &\leq \Big(\frac{\pi^2}{\rho^2}(n-1) - \lambda \Big) \int_0^\rho \sin^2 \frac{\pi t}{\rho}\, dt, \end{aligned}$$

since the Y_i form an orthonormal basis of the subspace of $T_{c(t)}M$ normal to $\dot{c}$. Consequently, $\rho \leq \pi \sqrt{\frac{n-1}{\lambda}}$, and since this holds for any $\rho < \operatorname{diam}(M)$, we obtain the estimate for the diameter. The universal cover of M satisfies the same assumption on the Ricci curvature. Hence, it is compact as well. This implies that the group of covering transformations, i.e. $\pi_1(M)$, is finite. □

4.4 Riemannian Manifolds of Constant Curvature

We have already met Euclidean spaces and spheres as Riemannian manifolds of vanishing and constant positive sectional curvature, resp. We now want to discuss hyperbolic space as an example of a Riemannian manifold with constant negative sectional curvature.

For this purpose, we equip $\mathbb{R}^{n+1}$ with the quadratic form

$$\langle x, x \rangle := -(x^0)^2 + (x^1)^2 + \ldots + (x^n)^2 \quad (x = (x^0, \ldots, x^n)).$$

We define

$$H^n := \{x \in \mathbb{R}^{n+1} : \langle x, x \rangle = -1, x^0 > 0\}.$$

Thus, H^n is a hyperboloid of revolution; the condition $x^0 > 0$ ensures that H^n is connected.

The symmetric bilinear form

$$I := -(dx^0)^2 + (dx^1)^2 + \ldots + (dx^n)^2$$

induces a positive definite symmetric bilinear form on H^n. Namely, if $p \in H^n$, T_pH^n is orthogonal to p w.r.t. $\langle \cdot, \cdot \rangle$. Therefore, the restriction of I to T_pH^n is positive definite by Sylvester's theorem. We thus obtain a Riemannian metric $\langle \cdot, \cdot \rangle$ on H^n. The resulting Riemannian manifold is called *hyperbolic space.*

Let $\mathrm{O}(n, 1)$ be the group of those linear self maps of $\mathbb{R}^{n+1}$ that leave the form $\langle .,. \rangle$ invariant. Those elements of $\mathrm{O}(n, 1)$ that map the positive x^0-axis onto itself, then also leave H^n invariant and operate on H^n by isometries. This is completely analogous to the isometric operation of $\mathrm{O}(n + 1)$ on $S^n \subset \mathbb{R}^{n+1}$. As we have seen in §1.4 for S^n, we see here that the geodesics of H^n are precisely the intersections of H^n with twodimensional linear subspaces of $\mathbb{R}^{n+1}$.

If $p \in H^n, v \in T_pH^n$ with $\|v\| = 1$, the geodesic $c : \mathbb{R} \to H^n$ with $c(0) = p, \dot{c}(0) = v$ is given by

$$c(t) = (\cosh\, t)p + (\sinh\, t)v;$$

indeed,

$$\langle c(t), c(t) \rangle = -\cosh^2 t + \sinh^2 t = -1,$$

since

$$\langle p, p \rangle = -1, \quad \langle p, v \rangle = 0, \quad \langle v, v \rangle = 1$$

and

$$\langle \dot{c}(t), \dot{c}(t) \rangle = -\sinh^2 t + \cosh^2 t = 1.$$

As on S^n, we may now compute the curvature with the help of Jacobi fields.

For this, let $w \in T_pH^n$, $\langle w, w \rangle = 1, \langle w, v \rangle = 0$. We then obtain a family of geodesics

$$c(t, s) := (\cosh t)p + \sinh\, t(\cos s\, v + \sin s\, w).$$

The corresponding Jacobi field

$$X(t) = \frac{\partial}{\partial s} c(t, s)_{|s=0} = (\sinh t)w$$

then satisfies

$$\ddot{X}(t) = X(t).$$

The Jacobi equation implies $R(X, \dot{c})\dot{c} = -X$, and so, the sectional curvature is -1.

We may then also obtain a space $H^n(\rho)$ of constant sectional curvature $-\rho$ by scaling the metric with factor ρ and considering

$$\langle \cdot, \cdot \rangle_\rho := \rho \langle \cdot, \cdot \rangle\,.$$

4.5 The Rauch Comparison Theorems and Other Jacobi Field Estimates

We first compare the three model spaces $S^n, \mathbb{R}^n, H^n$ of curvature $1, 0, -1$. Let $c(t)$ be a geodesic with $\|\dot{c}\| = 1$, $v \in T_{c(0)}M, M \in \{S^n, \mathbb{R}^n, H^n\}$ with $\|v\| = 1$. The Jacobi field $J(t)$ along c with

$$J(0) = 0, \dot{J}(0) = v$$

is given by $(\sin t)v, tv$, $(\sinh t)v$, resp.

According to our geometric interpretation of Jacobi fields as infinitesimal families of geodesics (Theorem 4.2.1) this means, that on S^n, geodesics with the same initial point initially diverge, but then converge again, whereas such geodesics diverge linearly on $\mathbb{R}^n$ and even exponentially on H^n.

Let now M be a Riemannian manifold with curvature K satisfying

$$\lambda \le K \le \mu$$

and suppose initially $\lambda \le 0, \mu \ge 0$. We shall estimate a Jacobi field in M from above by the Jacobi field in $H^n(-\lambda)$ with initial values of same lengths, and from below by the corresponding one in $S^n(\mu)$. This implies that the distance between geodesics and also the derivative of the exponential map of M can be controlled by the geometry of the model spaces $H^n(-\lambda)$ and $S^n(\mu)$. Since tangential Jacobi fields are always linear (Lemma 4.2.4), hence independent of the geometry of M, for our curvature bounds $\lambda \le K \le \mu$, we shall need to assume in the sequel $\lambda \le 0$ and $\mu \ge 0$, or else, we shall have to restrict attention to Jacobi fields whose tangential component $J^{\tan}$ vanishes identically.

For abbreviation, we put for $\rho \in \mathbb{R}$

$$c_\rho(t) := \begin{cases} \cos(\sqrt{\rho}\,t) & \text{if } \rho > 0, \\ 1 & \text{if } \rho = 0, \\ \cosh(\sqrt{-\rho}\,t) & \text{if } \rho < 0, \end{cases}$$

and

$$s_\rho(t) := \begin{cases} \frac{1}{\sqrt{\rho}} \sin(\sqrt{\rho}\,t) & \text{if } \rho > 0, \\ t & \text{if } \rho = 0, \\ \frac{1}{\sqrt{-\rho}} \sinh(\sqrt{-\rho}\,t) & \text{if } \rho < 0. \end{cases}$$

These functions are solutions of the Jacobi equation for constant sectional curvature ρ, namely

$$\ddot{f}(t) + \rho f(t) = 0 \tag{4.5.1}$$

with initial values $f(0) = 1, \dot{f}(0) = 0$, resp. $f(0) = 0, \dot{f}(0) = 1$. $c(t)$ will always be a geodesic on M parametrized by arc length, i.e. satisfying

$$\|\dot{c}\| \equiv 1. \tag{4.5.2}$$

Let $J(t)$ be a Jacobi field along $c(t)$.

Theorem 4.5.1. *Suppose $K \le \mu$, and as always, $\|\dot{c}\| \equiv 1$. Assume either $\mu \ge 0$ or $J^{\tan} \equiv 0$. Let $f_\mu := |J(0)|c_\mu + |J|^\cdot(0)s_\mu$ solve*

$$\ddot{f} + \mu f = 0$$

with $f(0) = |J(0)|, \dot{f}(0) = |J|^\cdot(0)$, i.e. $f_\mu = |J(0)|c_\mu + |J|^\cdot(0)s_\mu$.

If

$$f_\mu(t) > 0 \quad \text{for } 0 < t < \tau, \tag{4.5.3}$$

then

$$\langle J, \dot{J}\rangle f_\mu \ge \langle J, J\rangle \dot{f}_\mu \quad \text{on } [0, \tau], \tag{4.5.4}$$

$$1 \le \frac{|J(t_1)|}{f_\mu(t_1)} \le \frac{|J(t_2)|}{f_\mu(t_2)}, \quad \text{if } 0 < t_1 \le t_2 < \tau, \tag{4.5.5}$$

$$|J(0)|c_\mu(t) + |J|^\cdot(0)s_\mu(t) \le |J(t)| \qquad \text{for } 0 \le t \le \tau. \tag{4.5.6}$$

We point out that the assumption (4.5.3), i.e.

$$f_\mu(t) > 0 \quad \text{on } (0, \tau)$$

is indeed necessary. To see this, let $M = S^n(\mu - \varepsilon), J(0) = 0;$ $f_\mu(t)$ then has a zero at $t = \frac{\pi}{\sqrt{\mu}}$, $J(t)$ one at $t = \frac{\pi}{\sqrt{\mu-\varepsilon}}$. In particular, for small positive ε and any t which is only a little larger than $\frac{\pi}{\sqrt{\mu-\varepsilon}}$, we have $\frac{|J(t)|}{f(t)} < 1$, and for example, (4.5.5) does not hold anymore.

Proof.

$$\begin{aligned} |J|^{\cdot\cdot} + \mu|J| &= \frac{1}{|J|}(-\langle R(J, \dot{c})\dot{c}, J\rangle + \mu\langle J, J\rangle) \\ &\quad + \frac{1}{|J|^3}(|\dot{J}|^2\,|J|^2 - \langle J, \dot{J}\rangle^2) \\ &\ge 0, \end{aligned}$$

because $K \le \mu$, for $0 < t < \tau$, provided J has no zero on $(0, \tau)$.

We then also have

$$(|J|^\cdot f_\mu - |J|\dot{f}_\mu)^\cdot = |J|^{\cdot\cdot} f_\mu - |J|\ddot{f}_\mu \ge 0,$$

since $\ddot{f}_\mu + \mu f_\mu = 0$, provided $f_\mu(t) \ge 0$.

Because of $|J|(0) = f_\mu(0), |J|^\cdot(0) = \dot{f}_\mu(0)$, we conclude

$$|J|^\cdot f_\mu - |J|\dot{f}_\mu \ge 0,$$

i.e. (4.5.4).

Next

$$\left(\frac{|J|}{f_\mu}\right)^\cdot = \frac{1}{f_\mu^2}(|J|^\cdot f_\mu - |J|\dot{f}_\mu) \ge 0,$$

and from this and the initial conditions, we get (4.5.5). In particular, the first zero of J cannot occur before the first zero of f_μ, and the preceding considerations are valid on $(0, \tau)$.

(4.5.5) implies (4.5.6). □

Corollary 4.5.1. *Assume $K \leq \mu, c_\mu \geq 0$ on $(0, \tau)$, and in addition either $\mu \geq 0$ or $J^{\tan} \equiv 0$. Furthermore, let $\|\dot{c}\| \equiv 1, J(0) = 0, |R| \leq \Lambda$ where R stands for the curvature tensor.*

Then

$$|J(t) - t\dot{J}(t)| \leq |J(\tau)|\frac{1}{2}\Lambda t^2. \tag{4.5.7}$$

Proof. Let P be a parallel vector field of length 1 along c, $t \in (0, \tau)$

$$\begin{aligned}
|\langle J(t) - t\dot{J}(t), P(t)\rangle^{\cdot}| &= |t\langle R(J, \dot{c})\dot{c}, P\rangle(t)| \\
&\leq \Lambda t|J(t)| \\
&\leq \Lambda t|J(\tau)|\frac{s_\mu(t)}{s_\mu(\tau)} \quad \text{by (4.5.5), because of} J(0) = 0 \\
&\leq \Lambda t|J(\tau)|, \qquad \text{since } c_\mu \geq 0 \text{ on } [0, \tau].
\end{aligned}$$

Integrating this yields (4.5.7), as $J(0) = 0$. □

We now want to study the influence of lower curvature bounds. It will turn out that this is more complicated than for upper curvature bounds.

Theorem 4.5.2. *Assume $\lambda \leq K \leq \mu$ and either $\lambda \leq 0$ or $J^{\tan} \equiv 0$; $\|\dot{c}\| \equiv 1$. Moreover, let $J(0)$ and $\dot{J}(0)$ be linearly dependent.*

Assume

$$s_{\frac{1}{2}(\lambda+\mu)} > 0 \quad \text{on } (0, \tau). \tag{4.5.8}$$

Then for $0 \leq t \leq \tau$,

$$|J(t)| \leq |J(0)|c_\lambda(t) + |J|^{\cdot}(0)s_\lambda(t). \tag{4.5.9}$$

Proof. Let $\rho \in \mathbb{R}, \eta := \max(\mu - \rho, \rho - \lambda)$. Let A be the vector field along c with

$$\begin{aligned}
\ddot{A} + \rho A &= 0, \\
A(0) &= J(0), \\
\dot{A}(0) &= \dot{J}(0).
\end{aligned} \tag{4.5.10}$$

((4.5.10) is a system of linear 2nd order ODEs, and hence, for given initial value and initial derivative, there is a unique solution.) Let $a : I \to \mathbb{R}$ be the solution of

$$\begin{aligned}
\ddot{a} + (\rho - \eta)a &= \eta|A|, \\
a(0) = \dot{a}(0) &= 0,
\end{aligned} \tag{4.5.11}$$

and let $b : I \to \mathbb{R}$ be the solution of

$$\begin{aligned} \ddot{b} + \rho b &= \eta |J|, \\ b(0) &= \dot{b}(0) = 0 \end{aligned} \tag{4.5.12}$$

(since (4.5.11) and (4.5.12) are linear 2nd order ODE, too, again there exist unique solutions).

For each vector field P along c with $\|P\| \equiv 1$, we then have by (4.5.10)

$$|\langle J - A, P\rangle^{\cdot\cdot} + \rho\langle J - A, P\rangle| = |\langle \ddot{J} + \rho J, P\rangle| \leq \eta |J|$$

by choice of η and since J solves the Jacobi equation.

Therefore, by (4.5.12) for $d := (\langle J - A, P\rangle - b)^{\cdot} s_\rho - (\langle J - A, P\rangle - b)\dot{s}_\rho$,

$$\dot{d} = (\langle J - A, P\rangle - b)^{\cdot\cdot} s_\rho - (\langle J - A, P\rangle - b)\ddot{s}_\rho \leq 0,$$

and hence, if $s_\rho > 0$ on $(0, t]$, because $d(0) = 0$,

$$\left(\frac{1}{s_\rho}(\langle J - A, P\rangle - b)\right)^{\cdot}(t) = \frac{d(t)}{s_\rho^2(t)} \leq 0. \tag{4.5.13}$$

Note that $\langle J - A, P\rangle - b$ has a second order zero at $t = 0$, and hence $\frac{1}{s_\rho}(\langle J - A, P\rangle - b)$ vanishes for $t = 0$.

Therefore, we obtain from (4.5.13)

$$\frac{1}{s_\rho}(\langle J - A, P\rangle - b) \leq 0 \quad \text{on } (0, \tau). \tag{4.5.14}$$

If $s_\rho > 0$ on $(0, \tau)$, this implies

$$|J - A| \leq b \quad \text{on } (0, \tau) \tag{4.5.15}$$

and by (4.5.12) then

$$\ddot{b} + (\rho - \eta)b \leq \eta |A|. \tag{4.5.16}$$

From (4.5.12) and (4.5.16) we conclude with the same argument as the one leading to (4.5.14),

$$\frac{1}{s_{\rho-\eta}}(b - a) \leq 0,$$

i.e.

$$b \leq a \tag{4.5.17}$$

provided

$$s_{\rho-\eta} > 0 \quad \text{on } (0, \tau).$$

From (4.5.15) and (4.5.17)

$$|J - A| \leq a. \tag{4.5.18}$$

Now by (4.5.10)

$$(\langle \dot{A}, \dot{A}\rangle\langle A, A\rangle - \langle A, \dot{A}\rangle\langle A, \dot{A}\rangle)^{\cdot} = 0, \tag{4.5.19}$$

and hence,

$$\langle \dot{A}, \dot{A}\rangle\langle A, A\rangle - \langle A, \dot{A}\rangle\langle A, \dot{A}\rangle \equiv 0, \tag{4.5.20}$$

because this expression vanishes for $t = 0$, since $A(0) = J(0)$ and $\dot{A}(0) = \dot{J}(0)$ are linearly dependent by assumption. This implies

$$|A|^{\cdot\cdot} + \rho|A| = 0,$$

i.e. putting

$$f_\sigma = |J(0)|c_\sigma + |J|^{\cdot}(0)s_\sigma, \tag{4.5.21}$$

we have

$$|A| = f_\rho. \tag{4.5.22}$$

This implies in turn in conjunction with (4.5.11)

$$a = f_{\rho-\eta} - f_\rho. \tag{4.5.23}$$

(4.5.18), (4.5.22), (4.5.23) yield

$$|J| \le f_{\rho-\eta}.$$

Putting $\rho = \frac{1}{2}(\mu+\lambda)$, i.e. $\rho-\eta = \lambda$, we get (4.5.9). (Note that then $\eta = \frac{1}{2}(\mu-\lambda) \ge 0$, and hence $s_\rho > 0$ implies $s_{\rho-\eta} > 0$ on $(0,\tau)$.) □

Theorem 4.5.3. *Suppose* $\|\dot{c}\| \equiv 1, |K| \le \Lambda$. *Let* $J(0)$ *and* $\dot{J}(0)$ *be linearly dependent. Let* P_t *denote parallel transport along* c *from* $c(0)$ *to* $c(t)$.

Then

$$\begin{aligned} |J(t) - P_t(J(0) + t\dot{J}(0))| &\le |J(0)|(\cosh(\sqrt{\Lambda}t) - 1) \\ &\quad + |J|^{\cdot}(0)\left(\frac{1}{\sqrt{\Lambda}}\sinh(\sqrt{\Lambda}t) - t\right). \end{aligned} \tag{4.5.24}$$

Proof. From (4.5.20)

$$\left(\frac{A}{|A|}\right)^{\cdot} = 0.$$

This means that $\frac{A}{|A|}$ is a parallel vector field. In the proof of Theorem 4.5.2, we now put $\rho = 0$. We then get $|A| = \rho_0$ (cf. (4.5.22)), i.e.

$$A(t) = P_t(J(0) + t\dot{J}(0)).$$

With $\rho = 0$, we have $\eta = \Lambda$, and hence s_ρ and $s_{\rho-\eta} > 0$ for $t > 0$, as required in the proof of Theorem 4.5.2. (4.5.18) and (4.5.23) then yield the claim. $\square$

Remark. If we do not assume $\|\dot{c}\| \equiv 1$, in all the preceding estimates, t has to be replaced by $t\|\dot{c}\|$ as argument of s_τ, c_τ, f_τ etc.

Namely, let

$$\tilde{c}(t) = c\left(\frac{t}{\|\dot{c}\|}\right)$$

be the reparametrization of c by arc length, i.e. $\|\dot{\tilde{c}}\| = 1$.

Then

$$\tilde{J}(t) = J\left(\frac{t}{\|\dot{c}\|}\right)$$

is the Jacobi field along $\tilde{c}$ with $\tilde{J}(0) = J(0), \dot{\tilde{J}}(0) = \frac{\dot{J}(0)}{\|\dot{c}\|}$; namely, since J satisfies the Jacobi equation, $\tilde{J}$ satisfies

$$\ddot{\tilde{J}} + R(\tilde{J}, \dot{\tilde{c}})\dot{\tilde{c}} = 0.$$

Thus, estimates for $\tilde{J}$ yield corresponding estimates for J.

Remark. The derivation of the Jacobi field estimates of the present paragraph follows P. Buser and H. Karcher, Gromov's almost flat manifolds, Astérisque 81, 1981.

Perspectives. The Rauch comparison theorems are infinitesimal comparison results for the geometry of a Riemannian manifold in terms of the geometry of spaces of constant curvature.

A global comparison result is Toponogov's theorem:

Let M be a Riemannian manifold with sectional curvature $K \geq \lambda$. Let Δ be a triangle in M with corners p, q, r and distance minimizing geodesic edges c_{pq}, c_{qr}, c_{pr}. Then there exists a geodesic triangle Δ_0 in the simply connected space M_λ of curvature λ with the same side lengths as Δ and with angles at its corners not larger than the ones of Δ at the corresponding corners. In case $\lambda > 0$, we have in particular

$$L(\partial\Delta) \leq \frac{2\pi}{\sqrt{\lambda}}.$$

4.6 Geometric Applications of Jacobi Field Estimates

We first recall Corollary 4.2.2: Let $c(t) = \exp_p t\dot{c}(0)$ be geodesic, $w \in T_pM$, J the Jacobi field along c with $J(0) = 0, \dot{J}(0) = w$. $J(t)$ then yields the derivative of the exponential map

$$J(t) = (D_{t\dot{c}(0)} \exp_p)(tw). \tag{4.6.1}$$

We obtain

Corollary 4.6.1. *Let the sectional curvature of M satisfy $\lambda \le K \le \mu$. Furthermore, let $\langle w, \dot{c}(0)\rangle = 0$. Then, provided $t\|\dot{c}(0)\| \le \frac{\pi}{\sqrt{\mu}}$ in case $\mu > 0$,*

$$|w|\frac{s_\mu(t\|\dot{c}(0)\|)}{t\|\dot{c}(0)\|} \le |(D_{t\dot{c}(0)} \exp_p)w| \le |w|\frac{s_\lambda(t\|\dot{c}(0)\|)}{t\|\dot{c}(0)\|}. \tag{4.6.2}$$

(Of course, if w is a multiple of $\dot{c}(0)$, we have $(D_{t\dot{c}(0)} exp_p)w = w$.)

Proof. For $\|\dot{c}(0)\| = 1$, this follows from (4.5.6) and (4.5.9).

We now put $\tilde{c}(t) := \exp_p t\frac{\dot{c}(0)}{\|\dot{c}(0)\|}$. $\tilde{c}$ thus is a reparametrization of c, and $\|\dot{\tilde{c}}\| \equiv 1$. Let $\tilde{J}$ be the Jacobi field along $\tilde{c}$ with $\tilde{J}(0) = 0, \dot{\tilde{J}}(0) = w$. Finally,

$$\begin{aligned}(D_{t\dot{c}(0)} \exp_p)(tw) &= \frac{1}{\|\dot{c}(0)\|}(D_{t\|\dot{c}(0)\|\dot{\tilde{c}}(0)} \exp_p)(t\|\dot{c}(0)\|w)\\ &= \frac{1}{\|\dot{c}(0)\|}\tilde{J}(t\|\dot{c}(0)\|)\,,\end{aligned}$$

and $\tilde{J}(t\|\dot{c}(0)\|)$ is controlled by $s_\mu(t\|\dot{c}(0)\|)$ and $s_\lambda(t\|\dot{c}(0)\|)$ from below and above, resp. □

Theorem 4.6.1. *Let the exponential map $\exp_p : T_pM \to M$ be a diffeomorphism on $\{v \in T_pM : \|v\| \le \rho\}$. Let the curvature of M in the ball*

$$B(p, \rho) := \{q \in M : d(p, q) \le \rho\}$$

satisfy

$$\lambda \le K \le \mu, \ \text{with}\, \lambda \le 0, \mu \ge 0,$$

and suppose

$$\rho < \frac{\pi}{2\sqrt{\mu}} \quad \text{in case } \mu > 0. \tag{4.6.3}$$

Let $r(x) := d(x, p), k(x) := \frac{1}{2}d^2(x, p)$. Then k is smooth on $B(p, \rho)$ and satisfies

$$\operatorname{grad} k(x) = -\exp_x^{-1} p, \tag{4.6.4}$$

and therefore

$$|\operatorname{grad} k(x)| = r(x). \tag{4.6.5}$$

$$\begin{aligned}\sqrt{\mu}r(x)\operatorname{ctg}(\sqrt{\mu}r(x))\|v\|^2 &\le \nabla dk(v, v)\\ &\le \sqrt{-\lambda}r(x)\operatorname{ctgh}(\sqrt{-\lambda}r(x))\|v\|^2\end{aligned} \tag{4.6.6}$$

for $x \in B(p, \rho), v \in T_xM$.

Proof. We have

$$\operatorname{grad} k(x) = -\exp_x^{-1} p,$$

because the gradient of k is orthogonal to the level surfaces of k, and those are the spheres $S(p,r) := \{q \in M : d(p,q) = r\} = \exp_p\{v \in T_pM : \|v\| = r\}$ $(r \leq \rho)$; in particular, the gradient of k has length $d(x,p)$, proving (4.6.5).

The Hessian ∇dk of k is symmetric, and can hence be diagonalized. It thus suffices to show (4.6.6) for each eigen direction v of ∇dk. Let $\gamma(s)$ be the curve in M with $\gamma(0) = x, \gamma'(0) = v$.

$$c(t,s) := \exp_{\gamma(s)}(t \exp_{\gamma(s)}^{-1} p), \tag{4.6.7}$$

in particular $c(0,s) = \gamma(s), c(1,s) \equiv p$.

Then by (4.6.4)

$$(\operatorname{grad} k)(\gamma(s)) = -\frac{\partial}{\partial t} c(t,s)_{|t=0},$$

hence

$$\begin{aligned}(\nabla_v \operatorname{grad} k)(x) &= -\nabla_{\frac{\partial}{\partial s}} \frac{\partial}{\partial t} c(t,s)_{|t=0,s=0} \\ &= -\nabla_{\frac{\partial}{\partial t}} \frac{\partial}{\partial s} c(t,s)_{|t=0,s=0}.\end{aligned} \tag{4.6.8}$$

$J(t) = \frac{\partial}{\partial s} c(t,s)_{|s=0}$ is a Jacobi field along the geodesic from x to p with $J(0) = \dot\gamma(0) = v, J(1) = 0 \in T_pM$ (by (4.6.7)). (4.6.8) thus implies

$$\nabla_v \operatorname{grad} k(x) = -\dot J(0),$$

i.e.

$$\nabla dk(v,v) = \langle \nabla_v \operatorname{grad} k, v\rangle = -\langle \dot J(0), J(0)\rangle. \tag{4.6.9}$$

Since v is an eigen direction of ∇dk, $\nabla_v \operatorname{grad} k$ and v, i.e. $\dot J(0)$ and $J(0)$ are linearly dependent. (4.5.6) and (4.5.9) imply for $t = 1\,(J(1) = 0)$ (recall the remark at the end of §4.5)

$$|v| c_\mu(r(x)) + |J|^{\cdot}(0) s_\mu(r(x)) \leq 0 \leq |v| c_\lambda(r(x)) + |J|^{\cdot}(0) s_\lambda(r(x))$$

and with (4.6.9), this gives (4.6.6). $\square$

We want to briefly describe the relation between Jacobi fields and the 2nd fundamental form of the distance spheres

$$\partial B(p,r) = \{q \in M, d(p,q) = r\}.$$

Assume the hypotheses of Theorem 4.6.1; in particular, assume that $\exp_p$ is a diffeomorphism of $\{\|v\| \leq \rho\}$ onto $B(p,\rho)$, and that $r \leq \rho$.

We have

$$N(x) = \operatorname{grad} k(x) = -\exp_x^{-1} p \quad \text{(by (4.6.4))}; \tag{4.6.10}$$

where $N(x)$ is the exterior normal vector of the distance up here containing x. For the second fundamental form S of the distance sphere and for X tangential to this sphere, we then have

$$\begin{aligned} S(X,N) &= \nabla_X N \qquad \text{since } N(x) \text{ has constant length } r \\ &\text{on } \partial B(p,r), \text{the part of } \nabla_x N \text{ normal to } \partial B(p,r) \text{ vanishes} \\ &= \nabla_X \operatorname{grad} k. \end{aligned} \tag{4.6.11}$$

We now obtain a diffeomorphism from $\partial B(p,r)$ onto $\partial B(p,r+t)$ (assuming $r+t \leq \rho$) by

$$E_t(x) := \exp_x tN(x) \quad (x \in \partial B(p,r)).$$

Let $\gamma(s)$ be a curve in $\partial B(p,r)$ with $\dot\gamma(0) = v, \gamma(0) = x$. Then

$$J(t) = \frac{\partial}{\partial s} E_t(\gamma(s))_{|s=0} \tag{4.6.12}$$

is a Jacobi field along $E_t(x)$ with

$$\begin{aligned} J(0) &= \dot\gamma(0) \\ &= v, \end{aligned}$$

and

$$\begin{aligned} \dot J(0) &= \nabla_{\frac{\partial}{\partial t}} \frac{\partial}{\partial s} \exp_{\gamma(s)}(tN(\gamma(s))) \\ &= \nabla_{\frac{\partial}{\partial s}} \frac{\partial}{\partial t} \exp_{\gamma(s)}(tN(\gamma(s)))_{|\substack{t=0\\s=0}} \\ &= \nabla_{\frac{\partial}{\partial s}} N(\gamma(s))_{|s=0} \\ &= S(v,N) \\ &= S(J(0),N). \end{aligned}$$

Since $E_t(\gamma(s))$ is a curve in $\partial B(p, r+t)$, we likewise have

$$\dot J(t) = S(J(t),N). \tag{4.6.13}$$

We put $S_t(\cdot) = S(\cdot, N(t))$.

From this, we get

$$\ddot J(t) = \nabla_{\frac{\partial}{\partial t}}(S_t(J(t)) = \dot S_t(J(t)) + S_t(\dot J(t)).$$

The Jacobi equation $\ddot J + R(J,N)N = 0$ thus implies a *Riccati equation* for S_t :

$$\dot S_t(\cdot) = -R(\cdot,N)N - S_t \circ S_t(\cdot). \tag{4.6.14}$$

Thus, on one hand, (4.6.13) describes the geometry of distance spheres through Jacobi fields. On the other hand, solutions of the Riccati equation satisfy a 1st order ODE and hence are easy to estimate, and from such estimates one may then obtain Jacobi field estimates. In order to explain this last point, let P be a vector field parallel along $E_t(x)$ with $\|P\| = 1$. Then

$$\langle S_t(P), P\rangle^{\cdot} = -\langle R(P,N)N, P\rangle - \langle S_t^2(P), P\rangle. \tag{4.6.15}$$

Since the 2nd fundamental tensor is symmetric,

$$\langle S_t^2(P), P\rangle = \langle S_t(P), S_t(P)\rangle \quad \text{(cf. Lemma 3.4.2).} \tag{4.6.16}$$

We put $\Sigma(\cdot) = \frac{1}{\|N\|} S_t(\cdot)$. Since all expressions in (4.6.15) are quadratically homogeneous in $\|N\|$, we obtain

$$\begin{aligned} \langle \Sigma(P), P\rangle^{\cdot} &= -\Big\langle R\Big(P, \frac{N}{\|N\|}\Big)\frac{N}{\|N\|}, P\Big\rangle - \langle \Sigma(P), \Sigma(P)\rangle \\ &\leq -\Big\langle R\Big(P, \frac{N}{\|N\|}\Big)\frac{N}{\|N\|}, P\Big\rangle - \langle \Sigma(P), P\rangle^2. \end{aligned} \tag{4.6.17}$$

If the sectional curvature satisfies $\lambda \leq K$, because of $\|P\| = 1$,

$$\varphi := \langle \Sigma(P), P\rangle$$

then satisfies the differential equation

$$\dot{\varphi} \leq -\lambda - \varphi^2. \tag{4.6.18}$$

Now

$$ct_\lambda(t) := \frac{\dot{s}_\lambda(t)}{s_\lambda(t)} = \frac{c_\lambda(t)}{s_\lambda(t)}$$

satisfies the differential equation

$$\dot{ct}_\lambda = -\lambda - ct_\lambda^2,$$

and it easily follows that

$$\varphi(t) \leq ct_\lambda(t), \text{ provided } \varphi(s) > -\infty \text{ for all } s \text{ with } 0 < s < t.$$

With (4.6.13), we conclude from this for a Jacobi field J along E_t with $J(0) = 0$

$$\left(\frac{|J(t)|}{s_\lambda(t)}\right)^{\cdot}(t) \leq 0,$$

provided in $(0, t]$ there is no point conjugate to 0.

In particular

$$|J(t)| \leq |J|^{\cdot}(0) s_\lambda(t), \tag{4.6.19}$$

i.e. a special case of (4.5.9), up to the first conjugate point.

Perspectives. Let M_ρ be the simply connected space form of curvature ρ. Let $V^\rho(r)$ denote the volume of a ball in M_ρ with radius r. Let M be a Riemannian manifold, $p \in M, r < i(p)$ (= injectivity radius of p) (i.e. $B(p,r)$ is disjoint from the cut locus of p.) We then have the volume comparison theorems of R. Bishop:

If $\mathrm{Ric}\,(M) \geq \mathrm{Ric}\,(M_\rho)$, then

$$\mathrm{Vol}\,(B(p,r)) \leq V^\rho(r)$$

and P. Günther:

If $K(M) \leq \rho \quad$ (K is the sectional curvature), then

$$\mathrm{Vol}\,(B(p,r)) \geq V^\rho(r).$$

These estimates are also proved with the help of Jacobi field estimates.

4.7 Approximate Fundamental Solutions and Representation Formulae

Lemma 4.7.1. *Suppose* $\exp_p : T_pM \to M$ *is a diffeomorphism on the ball* $\{v \in T_pM : \|v\| \leq \rho\}$, *and suppose the sectional curvature in* $B(p,\rho)$ *satisfies*

$$\lambda \leq K \leq \mu \quad \textit{with } \lambda \leq 0, \mu \geq 0,$$

put $\Lambda := \max(-\lambda, \mu)$, *and assume*

$$\rho < \frac{\pi}{\sqrt{\mu}} \quad \textit{in case } \mu > 0.$$

Then, with $r(x) = d(x,p)$, *for* $x \neq p$

$$|\Delta \log r(x)| \leq 2\Lambda \qquad \textit{if } n = \dim M = 2, \tag{4.7.1}$$

$$|\Delta(r(x)^{2-n})| \leq \frac{n-2}{2}\Lambda r^{2-n}(x) \qquad \textit{if } n = \dim M \geq 3. \tag{4.7.2}$$

Proof. We prove only (4.7.2) as (4.7.1) is similar.

$$\begin{aligned} -\Delta r(x)^{2-n} &= -\Delta(d^2(x,p))^{\frac{2-n}{2}} \\ &= \frac{2-n}{2}\Big(-\frac{n}{2}\Big)(d^2(x,p))^{-\frac{n+2}{2}}\|\mathrm{grad}\, d^2(x,p)\|^2 \\ &+ \frac{2-n}{2}(d^2(x,p))^{-\frac{n}{2}}(-\Delta)d^2(x,p). \end{aligned}$$

Now by Theorem 4.6.1

$$\begin{aligned}\|\operatorname{grad} d^2(x,p)\|^2 &= 4d^2(x,p),\\ 2n(1-\mu r^2(x)) &\le -\Delta d^2(x,p)\\ &\le 2n(1-\lambda r^2(x)) \quad \text{noting } -\Delta = \operatorname{trace}\nabla d,\end{aligned}$$

and (4.7.2) follows. □

Lemma 4.7.2. *Suppose $B(p,\rho)$ is as in Lemma 4.7.1. Let ω_n be the volume of the unit sphere in $\mathbb{R}^n, n = \dim M$. For $h \in C^2(B(p,\rho),\mathbb{R})$ then (with Λ as in Lemma 4.7.1)*

if $n = 2$,

$$\left|\omega_2 h(p) - \int_{B(p,\rho)} (\Delta h)\log\frac{r(x)}{\rho} - \frac{1}{\rho}\int_{\partial B(p,\rho)} h\right| \le 2\Lambda \int_{B(p,\rho)} |h|, \tag{4.7.3}$$

if $n \ge 3$,

$$\begin{aligned}&\left|(n-2)\omega_n h(p) - \int_{B(p,\rho)} (\Delta h)\Big(\frac{1}{r(x)^{n-2}} - \frac{1}{\rho^{n-2}}\Big) - \frac{n-2}{\rho^{n-1}}\int_{\partial B(p,\rho)} h\right|\\ &\qquad \le \frac{n-2}{2}\Lambda\int_{B(p,\rho)} \frac{|h|}{r(x)^{n-2}}\ .\end{aligned} \tag{4.7.4}$$

Proof. We prove only (4.7.4) as (4.7.3) is similar.

We put

$$g(x) := r(x)^{2-n} - \rho^{2-n}.$$

Then for $\varepsilon > 0$

$$\int_{B(p,\rho)\backslash B(p,\varepsilon)} (g\Delta h - h\Delta g) = \int_{\partial(B(p,\rho)\backslash B(p,\varepsilon))} \langle h\operatorname{grad} g - g\operatorname{grad} h, d\overrightarrow{\nu}\rangle\ .$$

($\overrightarrow{\nu}$ denotes the outer unit normal of $\partial(B(p,\rho)\backslash B(p,\varepsilon))$.)

Now

$$\begin{aligned}
\int_{B(p,\rho)\setminus B(p,\varepsilon)} |h\Delta g| &\leq \frac{n-2}{2}\Lambda \int_{B(p,\rho)} \frac{|h|}{r^{n-2}(x)} \quad \text{by (4.7.2)}, \\
g_{|\partial B(p,\rho)} &= 0, \\
\int_{\partial B(p,\rho)} h\langle \operatorname{grad} g, d\overrightarrow{\nu}\rangle &= \frac{n-2}{\rho^{n-1}} \int_{\partial B(p,\rho)} h\,, \\
\lim_{\varepsilon\to 0} \int_{\partial B(p,\varepsilon)} g\langle \operatorname{grad} h, d\overrightarrow{\nu}\rangle &= 0, \\
\lim_{\varepsilon\to 0} \int_{\partial B(p,\varepsilon)} \langle \operatorname{grad} g, d\overrightarrow{\nu}\rangle &= -(n-2)\omega_n h(p),
\end{aligned}$$

and (4.7.4) follows. □

For the interpretation of the preceding formulae, we observe that in the Euclidean case

$$\Delta r(x)^{2-n} = 0 \quad \text{for } x \neq p, \tag{4.7.5}$$

whereas individual second derivatives of $r(x)^{2-n}$ grow like $r(x)^{-n}$ for $x \to p$. Therefore, in the Riemannian case, although (4.7.5) is not an identity anymore it holds up to an error term which gains two orders of magnitude against the crude growth estimate $r(x)^{-n}$. The same holds for the representation formulae in Lemma 4.7.2. The error terms on the right hand side are two orders better than the other integrands.

Perspectives. The results of this paragraph are from [148]. Extensions of these results can be found in [132].

4.8 The Geometry of Manifolds of Nonpositive Sectional Curvature

In this section, we shall present some results that apply to compact or complete Riemannian manifolds of nonpositive sectional curvature. It is very instructive to see how strongly an infinitesimal geometric condition, namely that the sectional curvature is nonpositive, influences the global geometry and topology of the manifold in question.

At one place, we shall refer to a subsequent chapter for a proof ingredient. This is done for the sake of conciseness although the result in question can also be given an elementary - but not entirely trivial - proof with the tools already developed, and an ambitious reader may wish to find such a proof.

From §4.6, we obtain

Lemma 4.8.1. *Let N be a Riemannian manifold with sectional curvature ≤ 0. Let $p \in M$. Then the exponential map*

$$\exp_p : T_pN \to N$$

has everywhere maximal rank.

Furthermore, for

$$k(x) := \frac{1}{2}d^2(x,p)$$

if $\exp_p$ is a diffeomorphism on the ball $B(p,\rho), x \in B(p,\rho), v \in T_xN$, we have

$$\nabla dk(v,v) \geq \|v\|^2. \tag{4.8.1}$$

Proof. Corollary 4.6.1 and Theorem 4.6.1. □

These are local results. We shall now state a fundamental global result:

Theorem 4.8.1. *Let N be a complete Riemannian manifold of nonpositive sectional curvature, $p, q \in N$. Then in any homotopy class of curves from p to q, there is precisely one geodesic arc from p to q, and this arc minimizes length in its class.*

Proof. There exists a sequence (γ_n) of curves from p to q with

$$\lim_{n\to\infty} L(\gamma_n) = r := \inf\{\text{ lengths in given homotopy class }\}$$

(L denoting length).

W.l.o.g., for all n

$$\gamma_n \subset B(p, r+1)$$

in particular

$$\gamma_n \cap B(p, r+2)\backslash B(p, r+1) = \emptyset.$$

The proof of Theorem 1.5.1 therefore works with $B(p, r+1)$ instead of the Riemannian manifold M considered there to show the existence of a shortest geodesic arc γ from p to q in the given homotopy class.

To show uniqueness, we first observe that by Theorem 4.1.1, every geodesic arc γ from p to q is a strict local minimum of energy among all arcs with endpoints p and q, because $I_\gamma(W,W) > 0$ for all $W \not\equiv 0$ with $W(p) = 0 = W(q)$. (Here, W is a section along γ. The index form I_γ was defined in (4.1.8).)

Let now $\gamma_i : [0,1] \to N, i = 1,2$, be homotopic geodesic arcs from p to q, with $\gamma_1 \neq \gamma_2$, and let

$$\Gamma : [0,1] \times [0,1] \to N$$

be a homotopy, i.e. with

$$\begin{aligned} &\Gamma(t,0) = \gamma_1(t), \quad \Gamma(t,1) = \gamma_2(t), \qquad \text{for all } t,\\ &\Gamma(0,s) = p, \qquad\ \ \Gamma(1,s) = q, \qquad\quad\ \ \text{for all } s. \end{aligned}$$

Let

$$R := \max_{s\in[0,1]} E(\Gamma(\cdot, s)). \tag{4.8.2}$$

As in Theorem 6.11.3 below, one shows that there exists another geodesic arc γ_3, different from γ_1 and γ_2, with

$$\max(E(\gamma_1), E(\gamma_2)) < E(\gamma_3) \leq R. \tag{4.8.3}$$

Again, by Theorem 4.1.1, γ_3 is a strict local minimum of E, and so, replacing e.g. γ_2 by γ_3 in the previous argument, we obtain a fourth geodesic arc γ_4 with

$$E(\gamma_3) < E(\gamma_4) \leq R.$$

(It is not hard to see from the proof of Theorem 6.11.3, that γ_3 may be connected with γ_1 or γ_2 through arcs of energy $\leq R$ so that the maximum in (4.8.2) will not be increased.) We therefore obtain a sequence $(\gamma_n)_{n\in\mathbb{N}}$ of geodesic arcs from p to q with

$$E(\gamma_n) \leq R \quad \text{for all } n.$$

Let $\gamma_n(t) = \exp_p tv_n$ with $v_n \in T_pN$, $\|v_n\|^2 \leq 2R$. After selection of a subsequence, $(v_n)_{n\in\mathbb{N}}$ converges to some $v \in T_pM$ with $\|v\|^2 \leq 2R$. Since all v_n are different from each other, but $\exp_p v_n = q$ for all n, $\exp_p$ cannot have maximal rank at v. This is a contradiction, since by Lemma 4.8.1, the exponential map of a manifold of nonpositive curvature has everywhere maximal rank. Thus, $\gamma_1 = \gamma_2$, proving uniqueness. □

As a corollary, we have the following result of Hadamard-Cartan

Corollary 4.8.1. *Let Y be a simply connected complete Riemannian manifold of nonpositive sectional curvature. Then Y is diffeomorphic to $\mathbb{R}^n$ $(n = \dim Y)$, and such a diffeomorphism can be obtained from the exponential map*

$$\exp_p : T_pY (= \mathbb{R}^n) \to Y$$

of any $p \in Y$. This exponential map is distance nondecreasing, i.e.

$$\|v - w\| \leq d(\exp_p v, \exp_p w) \quad \textit{for all } v, w \in T_pY.$$

Proof. Theorem 4.8.1 implies that for every $p, q \in Y$, there exists precisely one geodesic arc from p to q because there is only one homotopy class of such arcs as Y is simply connected. One easily concludes that for every $p \in Y$, $\exp_p : T_pY \to Y$ is injective and surjective. (It is defined on all of T_pY because Y is complete.) Since it is of maximal rank everywhere by Lemma 4.8.1, it follows that Y is diffeomorphic to T_pY. The distance increasing property of the exponential map follows from Corollary 4.6.1. □

Lemma 4.8.2. *Let Y be a simply connected complete manifold of nonpositive curvature, $p \in Y$. Then, with $k(x) = \frac{1}{2}d^2(x,p)$, for every $v \in T_xY$, $x \in Y$*

$$\nabla dk(v,v) \geq \|v\|^2. \tag{4.8.4}$$

Proof. From Corollary 4.8.1 and Lemma 4.8.1. □

We also have

Theorem 4.8.2. *Let $c_1(t)$ and $c_2(t)$ be geodesics in Y, a simply connected complete manifold of nonpositive sectional curvature. Then*

$$d^2(c_1(t), c_2(t))$$

is a convex function of t.

Proof. Since the geodesic arc from $c_1(t)$ to $c_2(t)$ is uniquely determined by Theorem 4.8.1, it depends smoothly on t. Hence $d^2(c_1(t), c_2(t))$ is a smooth function of t. For each t, we denote this geodesic arc from $c_1(t)$ to $c_2(t)$ by $\gamma(s,t)$, with s the arc length parameter. Then

$$d^2(c_1(t), c_2(t)) = 2E(\gamma(\cdot,t)). \tag{4.8.5}$$

Now by Theorem 4.1.1 (exchanging the roles of s and t in that theorem)

$$\begin{aligned}\frac{d^2}{dt^2}E(\gamma(\cdot,t)) = &\int_0^{d(c_1(t),c_2(t))} \langle \nabla_{\frac{\partial}{\partial s}} \frac{\partial}{\partial t}\gamma(s,t), \nabla_{\frac{\partial}{\partial s}} \frac{\partial}{\partial t}\gamma(s,t)\rangle ds \\ &- \int_0^{d(c_1(t),c_2(t))} \langle R\left(\frac{\partial \gamma}{\partial s}, \frac{\partial \gamma}{\partial t}\right)\frac{\partial \gamma}{\partial t}, \frac{\partial \gamma}{\partial s}\rangle ds,\end{aligned} \tag{4.8.6}$$

where R denotes the curvature tensor of Y. Since Y has nonpositive sectional curvature, (4.8.6) implies

$$\frac{d^2}{dt^2}E(\gamma(\cdot,t)) \geq 0$$

and with (4.8.5) the claim follows. □

A reformulation of the preceding result is

Corollary 4.8.2. *Let Y be a simply connected complete manifold of nonpositive (sectional) curvature. Then*

$$d^2 : Y \times Y \to \mathbb{R}$$

is a convex function. (Note that here, d^2 is considered as a function of two variables.)

Proof. According to Definition 3.3.6, we have to show that the Hessian of d^2 is positive semidefinite.

By (3.3.50), we have to compute the second derivative of d^2 along geodesics in $Y \times Y$. Such geodesics c are given as (c_1, c_2) where c_1, c_2 are geodesics in Y. We thus have to show that $d^2(c_1(t), c_2(t)$ is a convex function of the arc length parameter t. This is Theorem (4.8.2). □

Remark. On a not necessarily simply connected Riemannian manifold N of nonpositive sectional curvature, the results of Lemma 4.8.2 and Theorem 4.8.2 hold locally: If

$$\exp_p : T_pN \to N$$

is a diffeomorphism on the ball $\{v \in T_pN : \|v\| \leq \rho) \subset T_pN$ for some $\rho > 0$, then (4.8.4) holds for $x \in B(p, \rho) \subset N$, and d^2 is convex on $B(p, \rho) \times B(p, \rho)$, i.e. for any geodesics $c_1, c_2 : [0, 1] \to B(p, \rho)$, $d^2(c_1(t), c_2(t))$ is a convex function of t.

Building upon Lemma 4.8.2, we shall now derive some quantitative versions of the preceding convexity results

Lemma 4.8.3. *As always in this §, let N be a Riemannian manifold of nonpositive sectional curvature, $p \in N$, and suppose that*

$$\exp_p : T_pN \to N$$

is a diffeomorphism on the ball $\{v \in T_pN : \|v\| \leq \rho\}$ (here, $\rho > 0$, and if N is complete and simply connected, we may take $\rho = \infty$ by Corollary 4.8.1).

Then

$$\begin{aligned} d^2(p, \gamma(t)) \leq (1-t)d^2(p, \gamma(0)) + td^2(p, \gamma(1)) \\ - t(1-t)d^2(\gamma(0)), \gamma(1)). \end{aligned} \tag{4.8.7}$$

Proof. Let $k_0 : [0, 1] \to \mathbb{R}$ be the function with

$$\begin{aligned} k_0(0) &= d^2(p, \gamma(0)), \\ k_0(1) &= d^2(p, \gamma(1)), \\ k_0''(t) &= 2\|\gamma'(t)\|^2. \end{aligned}$$

Then

$$d^2(p, \gamma(t)) \leq k_0$$

as a consequence of (4.8.4). Since

$$k_0(t) = (1-t)k_0(0) + tk_0(1) - t(1-t)d^2(\gamma(0), \gamma(1))$$

(note $\|\gamma'(t)\| = d(\gamma(0), \gamma(1))$), the claim follows. □

Corollary 4.8.3. *Under the assumptions of Lemma 4.8.3, let*

$$\gamma_1, \gamma_2 : [0,1] \to B(\gamma, \rho) \subset N$$

be geodesics with

$$\gamma_1(0) = p = \gamma_2(0).$$

Then, for $0 \le t \le 1$,

$$d(\gamma_1(t), \gamma_2(t)) \le t d(\gamma_1(1), \gamma_2(1)). \tag{4.8.8}$$

Proof. Applying (4.8.7) to $\gamma_1(1)$ in place of $p, \gamma_2(t)$ in place of $\gamma(t)$,

$$\begin{aligned} d^2(\gamma_1(1), \gamma_2(t)) \le t d^2(\gamma_1(1), \gamma_2(1)) &+ (1-t) d^2(\gamma_1(1), p) \\ &- t(1-t) d^2(\gamma_2(1), p), \end{aligned}$$

applying (4.8.7) to $\gamma_2(t)$ in place of $p, \gamma_1(t)$ in place of $\gamma(t)$,

$$\begin{aligned} d^2(\gamma_1(t), \gamma_2(t)) \le t d^2(\gamma_1(1), \gamma_2(t)) &+ (1-t) d^2(p, \gamma_2(t)) \\ &- t(1-t) d^2(\gamma_1(1), p). \end{aligned}$$

Noting $d^2(p, \gamma_2(t)) = t^2 d^2(p, \gamma_2(1))$ and inserting the first inequality into the second one yields the result. □

Remark. It is also easy to give a direct proof of Lemma 4.8.3 based on the Jacobi field estimate (4.5.5).

We now come to Reshetnyak's quadrilateral comparison theorem:

Theorem 4.8.3. *As in the preceding lemma, let*

$$\exp_p : T_p N \to N$$

be a diffeomorphism on the ball of radius ρ *in* $T_p N$, N *a Riemannian manifold of nonpositive sectional curvature.*

Let

$$\gamma_1, \gamma_2 : [0,1] \to B(p, \rho) \subset N$$

be geodesics. For $0 \le t \le 1$, *and a parameter* $0 \le s \le 1$ *then*

$$\begin{aligned} &d^2(\gamma_1(0), \gamma_2(t)) + d^2(\gamma_1(1), \gamma_2(1-t)) \\ &\quad \le d^2(\gamma_1(0), \gamma_2(0)) + d^2(\gamma_1(1), \gamma_2(1)) + 2t^2 d^2(\gamma_2(0), \gamma_2(1)) \\ &\qquad + t(d^2(\gamma_1(0), \gamma_1(1)) - d^2(\gamma_2(0), \gamma_2(1))) \\ &\qquad - ts(d(\gamma_1(0), \gamma_1(1)) - d(\gamma_2(0), \gamma_2(1)))^2 \\ &\qquad - t(1-s)(d(\gamma_1(0), \gamma_2(0)) - d(\gamma_1(1), \gamma_2(1)))^2. \end{aligned} \tag{4.8.9}$$

Note that this inequality is sharp for certain quadrilaterals in the Euclidean plane.

Proof. We first consider the case $t = 1$, $s = 0$. For simplicity of notation, we define

$$\begin{aligned} a_i &:= d(\gamma_i(0), \gamma_i(1)), & &\text{for } i = 1, 2, \\ b_1 &:= d(\gamma_1(0), \gamma_2(0)), & b_2 &:= d(\gamma_1(1), \gamma_2(1)), \\ d_1 &:= d(\gamma_2(0), \gamma_1(1)), & d_2 &:= d(\gamma_1(0), \gamma_2(1)). \end{aligned}$$

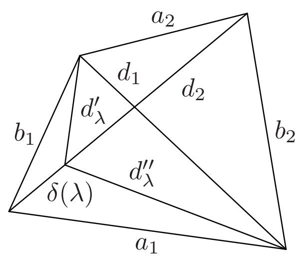

Figure 4.8.1:

Also, we let $\delta : [0,1] \to B(p,\rho) \subset N$ be the geodesic arc from $\gamma_1(0)$ to $\gamma_2(1)$, as always parametrized proportionally to arclength. Its length is d_2. We also put for $0 < \lambda < 1$

$$d'_\lambda := d(\gamma_2(0), \delta(\lambda)), \quad d''_\lambda := d(\gamma_1(1), \delta(\lambda)).$$

Then by (4.8.7)

$$\begin{aligned} d'^2_\lambda &\le (1-\lambda) b_1^2 + \lambda a_2^2 - \lambda(1-\lambda) d_2^2, \\ d''^2_\lambda &\le \lambda b_2^2 + (1-\lambda) a_1^2 - \lambda(1-\lambda) d_2^2. \end{aligned}$$

Therefore, for $0 < \varepsilon$,

$$\begin{aligned} d_1^2 &\le (d'_\lambda + d''_\lambda)^2 \\ &\le (1+\varepsilon) d'^2_\lambda + \left(1 + \frac{1}{\varepsilon}\right) d''^2_\lambda \\ &\le (1+\varepsilon)(1-\lambda) b_1^2 + (1+\varepsilon)\lambda a_2^2 \\ &\quad + \left(1 + \frac{1}{\varepsilon}\right) \lambda b_2^2 + \left(1 + \frac{1}{\varepsilon}\right)(1-\lambda) a_1^2 \\ &\quad - \left(2 + \varepsilon + \frac{1}{\varepsilon}\right) \lambda(1-\lambda) d_2^2. \end{aligned}$$

We choose $\varepsilon = \frac{1-\lambda}{\lambda}$ so that the coefficient in front of d_2^2 becomes 1. This yields

$$d_2^2 + d_1^2 \le a_1^2 + a_2^2 + \frac{1-\lambda}{\lambda} b_1^2 + \frac{\lambda}{1-\lambda} b_2^2.$$

With

$$\lambda = \frac{b_1}{b_1 + b_2},$$

we obtain

$$d_1^2 + d_2^2 \leq a_1^2 + a_2^2 + b_1^2 + b_2^2 - (b_1 - b_2)^2.$$

This is the required inequality for $t = 1, s = 0$. For symmetry reasons, we also obtain the inequality for $t = 1, s = 1$, namely

$$d_1^2 + d_2^2 \leq a_1^2 + a_2^2 + b_1^2 + b_2^2 - (a_1 - a_2)^2,$$

and taking convex combinations yields the inequality for $t = 1,\ 0 \leq s \leq 1$:

$$d_1^2 + d_2^2 \leq a_1^2 + a_2^2 + b_1^2 + b_2^2 - s(a_1 - a_2)^2 - (1-s)(b_1 - b_2)^2. \tag{4.8.10}$$

We therefore obtain the inequality for $0 \leq t \leq 1$ from (4.8.7) and (4.8.10)

$$\begin{aligned} &d^2(\gamma_1(0), \gamma_2(t)) + d^2(\gamma_1(1), \gamma_2(1-t)) \\ &\quad \leq (1-t)b_1^2 + td_2^2 - t(1-t)a_2^2 + (1-t)b_2^2 + td_1^2 - t(1-t)a_2^2 \\ &\quad \leq b_1^2 + b_2^2 + 2t^2 a_2^2 - t(a_2^2 - a_1^2) - ts(a_1 - a_2)^2 - t(1-s)(b_1 - b_2)^2. \end{aligned}$$

□

Theorem 4.8.3 allows us to derive the following quantitative version of the convexity of the distance between geodesics.

Corollary 4.8.4. *Let $\gamma_1, \gamma_2 : [0,1] \to N$ be geodesics as in Theorem 4.8.3. Then we have for $0 \leq t \leq 1$, $0 \leq s \leq 1$,*

$$\begin{aligned} d^2(\gamma_1(t), \gamma_2(t)) \leq{}& (1-t)d^2(\gamma_1(0), \gamma_2(0)) + td^2(\gamma_1(1), \gamma_2(1)) \\ &- t(1-t)\{s(d(\gamma_1(0), \gamma_1(1)) - d(\gamma_2(0), \gamma_2(1)))^2 \\ &+ (1-s)(d(\gamma_1(0), \gamma_2(0)) - d(\gamma_1(1), \gamma_2(1)))^2\}. \end{aligned} \tag{4.8.11}$$

Proof. We shall show the inequality for $t = \frac{1}{2}$. It is then straightforward to deduce the inequality for arbitrary t.

We keep the notations of the preceding proof, and we also put

$$e_1 := d\Big(\gamma_1(0), \gamma_2\Big(\frac{1}{2}\Big)\Big), \quad e_2 := d\Big(\gamma_1(1), \gamma_2\Big(\frac{1}{2}\Big)\Big).$$

Then by (4.8.7)

$$d^2\Big(\gamma_1\Big(\frac{1}{2}\Big), \gamma_2\Big(\frac{1}{2}\Big)\Big) \leq \frac{1}{2}e_1^2 + \frac{1}{2}e_2^2 - \frac{1}{4}a_1^2.$$

By (4.8.8)

$$e_1^2 + e_2^2 \leq b_1^2 + b_2^2 + \frac{1}{2}a_1^2 - \frac{1}{2}s(a_1 - a_2)^2 - \frac{1}{2}(1-s)(b_1 - b_2)^2.$$

Thus

$$d^2\Big(\gamma_1\Big(\frac{1}{2}\Big), \gamma_2\Big(\frac{1}{2}\Big)\Big) \leq \frac{1}{2}b_1^2 + \frac{1}{2}b_2^2 - \frac{1}{4}s(a_1 - a_2)^2 - \frac{1}{4}(1-s)(b_1 - b_2)^2$$

which yields the inequality for $t = \frac{1}{2}$. □

As an application of Theorem 4.8.3, let us consider the following Pythagoras inequality

Corollary 4.8.5. *Let the assumptions of Lemma 4.8.3 hold.*
Suppose
$$d(\gamma(0), p) = \min_{0 \le t \le 1} d(\gamma(t), p)$$
(i.e. $\gamma(0)$ is the point on γ closest to p). Then
$$d^2(\gamma(s), p) \ge d^2(\gamma(0), p) + s^2 d^2(\gamma(0), \gamma(1)) \quad \text{for } 0 \le s \le 1. \tag{4.8.12}$$

Proof. It suffices to treat the case $s = 1$.
By (4.8.7),
$$d^2(\gamma(t), p) \le (1-t) d^2(\gamma(0), p) + t d^2(\gamma(1), p) - t(1-t) d^2(\gamma(0), \gamma(1)).$$
Since by assumption
$$d^2(\gamma(0), p) \le d^2(\gamma(t), p),$$
we get
$$t d^2(\gamma(1), p) \ge t d^2(\gamma(0), p) + t d^2(\gamma(0), \gamma(1)) - t^2 d^2(\gamma(0), \gamma(1)).$$
Dividing by t and letting $t \to 0$ yields the desired inequality. □

We now turn to Karcher's center of mass constructions and their applications. While such constructions are meaning- and useful under more general conditions, here we only consider nonpositively curved manifolds, because in that case, the geometry is most favorable to them.

Thus, let Y be a complete, simply connected, nonpositively curved Riemannian manifold. We recall that by Corollary 4.8.1, $\exp_p : T_pY \to Y$ is a global diffeomorphism. This will be used implicitly below at several places.

Let μ be a probability measure on Y, i.e. a nonnegative measure with
$$\mu(Y) = \int d\mu = 1.$$

Definition 4.8.1. $q \in Y$ is called a center of mass for μ if
$$\int d^2(q, y)\, d\mu(y) = \inf_{p \in Y} \int d^2(p, y)\, d\mu(y) < \infty. \tag{4.8.13}$$

In the sequel we shall always assume that the infimum in (4.8.13) is finite. This is satisfied if, for example, the support of the measure μ is bounded.

Examples.

1. If μ is a Dirac measure δ_q supported at $q \in Y$, then q is its center of mass.
2. If $\mu = \frac{1}{2}(\delta_{q_1} + \delta_{q_2})$ for $q_1, q_2 \in Y$, then the center of mass is $\gamma(\frac{1}{2})$ where $\gamma : [0,1] \to Y$ is the unique geodesic from q_1 to q_2.

Lemma 4.8.4.
$$F(p) := \frac{1}{2} \int d^2(p,y)\, d\mu(y)$$
is a differentiable function of p, with
$$\operatorname{grad} F(p) = -\int \exp_p^{-1}(y)\, d\mu(y). \tag{4.8.14}$$
(Here, $\exp_p^{-1} : Y \to T_pY$ is considered as a vector valued function.)

Thus, q is a center of mass of μ if
$$\int \exp_q^{-1}(y)\, d\mu(y) = 0. \tag{4.8.15}$$

Proof. (4.8.14) follows from (4.6.4). Thus, F is differentiable, and a minimizer has to satisfy $\operatorname{grad} F(p) = 0$, i.e. (4.8.15). □

We now use the nonpositive curvature of Y in an essential manner:

Lemma 4.8.5.
$$F(p) = \frac{1}{2} \int d^2(p,y)\, d\mu(y)$$
is a strictly convex function of p.

Proof. From Lemma 4.8.2 by integration, because μ is nonnegative. □

We deduce

Theorem 4.8.4. *There exists a unique center of mass for μ, i.e. a unique $q \in Y$ with*
$$\int d^2(q,y)\, d\mu(y) = \inf_{p \in Y} \int d^2(p,y)\, d\mu(y).$$

Proof. This follows from the strict convexity and the fact that $F(p)$ is coercive, i.e.
$$F(p_n) \to \infty \;\text{ if } d^2(p_n, p_0) \to \infty \;\text{ for some fixed } p_0 \text{ and a sequence } (p_n)_{n \in \mathbb{N}} \subset Y.$$
□

Remark. Up to this point, we have not used the normalization
$$\mu(Y) = 1.$$

Thus, Theorem 4.8.4 holds for any nonnegative measure (provided the infimum in (4.8.13) is finite, of course). This will be applied in §7.3 below. The subsequent estimates, however, will use this normalization; without that normalization, additional factors will occur.

Lemma 4.8.6. *Let q be the center of mass of μ. Then*

for every $p \in Y$,

$$d(p,q) \leq \|\mathrm{grad}\, F(p)\|, \tag{4.8.16}$$

and for every $v \in T_qY$,

$$\|\nabla_v \mathrm{grad}\, F(q)\| \geq \|v\|. \tag{4.8.17}$$

Proof. Let $\gamma : [0,1] \to Y$ be the geodesic from q to p.

Thus,

$$\|\dot\gamma(t)\| = d(p,q) \quad \text{for all } t \in [0,1].$$

We have

$$\begin{aligned}\langle \mathrm{grad}\, F(p), \dot\gamma(1)\rangle &= -\int \langle \exp_p^{-1} y, \dot\gamma(1)\rangle \, d\mu(y) \\ &= -\int \left(\int_0^1 \frac{d}{dt} \langle \exp_{\gamma(t)}^{-1} y, \dot\gamma(t)\rangle dt \right) d\mu(y) \\ &\quad - \int \langle \exp_q^{-1} y, \dot\gamma(0)\rangle \, d\mu(y).\end{aligned}$$

The last integral vanishes by (4.8.15), since q is the center of mass for μ. By the proof of Corollary 4.6.1, since Y has nonpositive curvature (and since $D_{\frac{d}{dt}} \dot\gamma(t) = 0$ as γ is geodesic)

$$-\frac{d}{dt} \langle \exp_{\gamma(t)}^{-1} Y, \dot\gamma(t)\rangle \geq \|\dot\gamma(t)\|^2.$$

Thus

$$\|\mathrm{grad}\, F(p)\| d(p,q) \geq \langle \mathrm{grad}\, F(p), \dot\gamma(1)\rangle \geq d(p,q)^2,$$

which implies (4.8.16). (4.8.17) is the infinitesimal version of (4.8.16) (of course, (4.8.17) can also be derived directly from the proof of Corollary 4.6.1). □

Lemma 4.8.7. *Let μ_1, μ_2 be two probability measures on Y, with centers of mass q_1, q_2 resp. Then*

$$d(q_1, q_2) \leq \int d(q_2, y)\, |d\mu_1 - d\mu_2|(y). \tag{4.8.18}$$

Proof. By (4.8.16), with $F_i(p) = \frac{1}{2}\int d^2(y,p)d\mu_i(y)$, for $i = 1, 2$,

$$\begin{aligned} d(q_1, q_2) &\leq \|\operatorname{grad} F_1(q_2)\| \\ &\leq \left| \int \exp_{q_2}^{-1}(y)\, d\mu_1(y) \right| \\ &= \left| \int (exp_{q_2}^{-1} y)(d\mu_1 - d\mu_2)(y) \right| \quad \text{since } \operatorname{grad} F_2(q_2) = 0. \end{aligned}$$

We use $|\exp_{q_2}^{-1} y| = d(q_2, y)$ to get (4.8.18). □

We now consider the situation where

$$\mu = f_* \nu,$$

for some measurable map $f : A \to Y$ for a set A with a probability measure ν.

Then

$$\int d^2(q,y)\, d\mu(y) = \int d^2(q, f(x))\, d\nu(x). \tag{4.8.19}$$

For the moment, ν will be fixed, and so we shall call a minimizer a center of mass for the map f.

Lemma 4.8.8. *Let $f_1, f_2 \colon A \to Y$ be measurable maps with centers of mass q_1, q_2, resp. Then*

$$d(q_1, q_2) \leq \int d(f_1(x), f_2(x))\, d\nu(x). \tag{4.8.20}$$

Proof. By Lemma 4.8.6 and (4.8.14)

$$\begin{aligned} d(q_1, q_2) &\leq \left| \int \exp_{q_2}^{-1} f_1(x)\, d\nu(x) \right| \\ &= \left| \int \left(\exp_{q_2}^{-1} f_1(x) - \exp_{q_2}^{-1} f_2(x) \right) d\nu(x) \right| \end{aligned}$$

because q_2 is the center of mass for f_2,

$$\leq \int d(f_1(x), f_2(x))\, d\nu(x),$$

because the exponential map into a space of nonpositive curvature is distance nondecreasing by Corollary 4.8.1. □

Corollary 4.8.6. *Let $f : A \to Y$ be measurable with center of mass q. Then, for all $x \in A$,*

$$d(f(x), q) \leq \int d(f(x), f(y))\, d\nu(y). \tag{4.8.21}$$

Proof. We consider the map $f_1(y) = f(y)$ and the constant map $f_2(y) = f(x)$, for all $y \in A$; the former has center of mass q, the latter center of mass $f(x)$. We apply (4.8.20). □

The next result will be applied in §7.6 below only:

Corollary 4.8.7. *Let $f_1 \colon (A_1, \nu_1) \to Y, f_2 \colon (A_2, \nu_2) \to Y$ be measurable maps from probability measure spaces into Y.*

Let q_1, q_2 be the corresponding centers of mass.

Let $\varphi : (A_1, \nu_1) \to (A_2, \nu_2)$ be measurable, with $f_2 = f_1 \circ \varphi$.

Then

$$d(q_1, q_2) \leq \int d(f_1(x), f_2(\varphi(x)))\, d\nu_1(x) + \int d(f_2(x), q_2)\, |d\nu_2 - \varphi_* d\nu_1|(x). \tag{4.8.22}$$

Proof. Let q_2' be the center of mass for $f_2 \circ \varphi$ w.r.t. ν_1.

By Lemma 4.8.8

$$d(q_1, q_2') \leq \int d(f_1(x), f_2 \circ \varphi(x))\, d\nu_1(x).$$

By Lemma 4.8.7, since q_2' is the center of mass for f_2 w.r.t. $\varphi_* \nu_1$

$$d(q_2, q_2') \leq \int d(f_2(x), q_2)\, |d\nu_2 - \varphi_* d\nu_1|(x).$$

□

We now turn to the smoothing or mollification of maps with values in spaces of nonpositive curvature; this generalizes the standard construction for functions ("Friedrichs mollification").

We consider any C_0^∞ function

$$\rho : \mathbb{R} \to \mathbb{R}$$

with $\rho(s) \geq 0$ for all s and $\rho(s) = 0$ for $|s| \geq 1$, for example

$$\rho(s) := \begin{cases} \exp \frac{1}{s^2 - 1} & \text{for } |s| < 1, \\ 0 & \text{for } |s| \geq 1. \end{cases}$$

Given a ball $B(x, h) \subset M$ in some Riemannian manifold M, with $0 < h <$ injectivity radius of M at x, we put

$$\rho_{x,h}(y) = \frac{\rho\left(\frac{d(x,y)}{h}\right)}{\int_{B(x,h)} \rho\left(\frac{d(x,z)}{h}\right) dz}\,. \tag{4.8.23}$$

Here, $d(x, y)$ is the distance from x to $y \in B(x, h)$ w.r.t. the Riemannian metric of M.

To simplify the presentation, and in particular to eliminate an additional dependence on x, here, we do not work with the Riemannian volume form on $B(x, h)$ but rather with the Euclidean one, dz, induced via the exponential map $\exp_x : T_xM \to M$. Because of the denominator in (4.8.23),

$$\rho_{x,h}(y)\,dy$$

defines a probability measure on $B(x, h)$ (which we may extend by 0 to the rest of M).

Definition 4.8.2. Given a map

$$\begin{aligned} &f : M \to N, \\ &N \text{ a Riemannian manifold of nonpositive sectional curvature} \end{aligned}$$

its mollification with parameter h ($<$ injectivity radius of M) is defined by

$$f_h(x) := \text{ center of mass of } f \text{ w.r.t. the measure}$$

$$\rho_{x,h}(x)\,dy \text{ on } B(x, h)$$

Thus, $f_h(x)$ is the unique minimizer of

$$F(p) = \frac{1}{2} \int\limits_{B(x,h)} d^2(f(z), p)\rho_{x,h}(z)\,dz$$

Here, we do not need to assume that N is simply connected because on the simply connected ball, we can lift f to a map $f : B(x, h) \to \tilde{N}$ into the universal cover of N, apply the center of mass construction there and project back to N.

Lemma 4.8.9. *If f is locally integrable, then $f_h : M \to N$ is continuous for $h > 0$.*

Proof. Let $x_1, x_2 \in M$; we denote the above measures defined by ρ_h on the balls $B(x_1, h), B(x_2, h)$ by ν_1 and ν_2, resp. By Lemma 4.8.7

$$d(f_h(x_1), f_h(x_2)) \leq \int d(f(x), f_h(x_2))\,|d\nu_1 - d\nu_2|(x),$$

and the difference measure $d\nu_1 - d\nu_2$ goes to 0 if the distance between x_1 and x_2 goes to 0. $\square$

In fact, f_h is even smooth for $h > 0$. To see this, recall that $f_h(x)$ as a center of mass is characterized by (4.8.15), i.e.

$$\operatorname{grad} F(f_h(x)) = -\int\limits_{B(x,h)} \exp^{-1}_{f_h(x)}(f(z))\rho_{x,h}(z)\,dz = 0$$

Thus, in order to compute the derivative of f_h w.r.t. x, by the implicit function theorem, we must show that the derivative of grad $F(p)$ w.r.t. p is non zero.

This, however, follows from (4.8.17).

Theorem 4.8.5. *Let $f : M \to N$ be locally integrable. Then, for $0 < h <$ injectivity radius of M, the mollification f_h of f is smooth.*

Proof. We have just seen how the first derivative of f_h w.r.t. $x \in M$ can be computed from the implicit function theorem. Because of the smoothness of $\rho_{x,h}(z)$ w.r.t. x, higher derivatives then also exist. □

Lemma 4.8.10. *Let f be continuous at $x \in M$. Then*

$$\lim_{h \to 0} f_h(x) = f(x). \tag{4.8.24}$$

If f is uniformly continuous, then it is the uniform limit of the maps f_h for $h \to 0$.

Proof. Since f is continuous at x, given $\varepsilon > 0$, we may find $\delta > 0$ such that

$$f(B(x, \delta)) \subset B(f(x), \varepsilon).$$

Because the ball $B(f(x), \varepsilon)$ is convex, therefore also

$$f_h(x) \subset B(f(x), \varepsilon)$$

for $0 < h \leq \delta$. This implies (4.8.24). The remaining statement also follows from these considerations. □

We close this section with some constructions and results about the asymptotic geometry of complete simply connected Riemannian manifolds of nonpositive sectional curvature. Let Y be such a manifold for the rest of this section.

Definition 4.8.3. Two geodesic rays $c_1(t), c_2(t)(t \geq 0)$ in Y (i.e. $c_1, c_2 : [0, \infty) \to Y$) parametrized by arc length are called *asymptotic* if there exists $k \in \mathbb{R}$ with

$$d(c_1(t), c_2(t)) \leq k$$

for all $t \geq 0$. This defines an equivalence relation on the space of geodesic rays parametrized by arc length, and the set of equivalence classes is denoted by $Y(\infty)$. ($Y(\infty)$ is sometimes called the sphere at infinity of Y.)

Example. In Euclidean space, two geodesic rays, i.e. straight half lines, are equivalent iff they are parallel.

Lemma 4.8.11. *For each pair $p \in Y, x \in Y(\infty)$, there exists a unique geodesic ray $c = c_{px}$ parametrized by arc length in the equivalence class defined by x with $c(0) = p$.*

Proof. Existence: Let c_0 be a geodesic ray representing x. For $n \in \mathbb{N}$, let $c_n(t)$ be the geodesic arc from $p = c_n(0)$ to $c_0(n)$, parametrized by arc length as usual, $t_n := d(p, c_0(n))$, i.e. $c_n(t_n) = c_0(n)$, and $v_n := \frac{d}{dt}c_n(t)_{|t=0} \in T_pY$ the tangent vector to c_n at p. Since c_n is parametrized by arc length, v_n has length 1, hence converges towards some $v \in T_pY$ after selecting a subsequence. We put

$$c(t) := \exp_p tv, t \geq 0.$$

Because of the convexity of $d^2(c_0(t), c_n(t))$ (Theorem 4.8.2), for $0 \leq t \leq t_n$

$$d^2(c_n(t), c_0(t)) \leq \max(d^2(c_n(0), c_0(0)), d^2(c_n(t_n), c_0(t_n))). \tag{4.8.25}$$

We have

$$\begin{aligned} t_n &= d(c_n(0), c_n(t_n)) && \text{since } c_n \text{ is parametrized by arc length} \\ &\leq d(c_n(0), c_0(0)) + d(c_0(0), c_0(n)) && \text{since } c_0(n) = c_n(t_n) \\ &= d(p, c_0(0)) + n. \end{aligned}$$

and likewise

$$\begin{aligned} n &= d(c_0(0), c_0(n)) \\ &\leq d(c_0(0), c_n(0)) + d(c_n(0), c_0(n)) \\ &= d(c_0(0), p) + d(c_n(0), c_n(t_n)) \\ &= d(p, c_0(0)) + t_n, \end{aligned}$$

hence altogether

$$d(c_0(t_n), c_0(n)) = |n - t_n| \leq d(p, c_0(0)).$$

This implies in conjunction with (4.8.25) for $0 \leq t \leq t_n$

$$\begin{aligned} d(c_n(t), c_0(t)) &\leq \max(d(p, c_0(0)), d(c_0(n), c_0(t_n))) \\ &= d(p, c_0(0)). \end{aligned}$$

For $n \to \infty$, we therefore also get

$$d(c(t), c_0(t)) \leq d(p, c_0(0)). \tag{4.8.26}$$

(4.8.26) means that c_0 and c are asymptotic. This proves the existence of $c_{px} = c$.

Uniqueness: Let c_1, c_2 be rays asymptotic to c_0 with $c_1(0) = p = c_2(0)$. Then for all $t \geq 0$

$$d^2(c_1(t), c_2(t)) \leq \text{ const.}$$

Since $d^2(c_1(t), c_2(t))$ is convex in t by Theorem 4.8.2 and vanishes for $t = 0$, it vanishes identically, hence $c_1(t) = c_2(t)$, proving uniqueness. □

Lemma 4.8.11 implies that for each $p, Y(\infty)$ can be identified with the unit sphere $S_pY := \{v \in T_pY : \|v\| = 1\}$ in T_pY. Namely, each unit tangent vector uniquely determines an equivalence class of asymptotic geodesic rays. It is also not difficult to realize that the topology on $Y(\infty)$ defined through this identification is independent of the choice of p. We thus obtain a natural topology on $\bar{Y} = Y \cup Y(\infty)$, the so-called cone topology. $\bar{Y}$ thus becomes a compact space. We call $v \in T_pY, w \in T_qY$ asymptotic if the geodesic rays $\exp_p tv, \exp_p tw (t \geq 0)$ are asymptotic.

Since any isometry of Y maps geodesics onto geodesics and classes of asymptotic geodesic rays onto classes of asymptotic geodesic rays, each isometry of Y induces an operation on $Y(\infty)$, hence on $\bar{Y}$, too.

Perspectives. Corollary 4.8.1 goes back to the work of von Mangoldt, Hadamard, and E. Cartan.

The center of mass has been likewise instroduced by E. Cartan. The constructions and applications presented here are due to Karcher[162]. In fact, Karcher's constructions are more general than presented here and also apply to the case where the manifold can have positive curvature. Then, however, one has to work with local constructions, and one needs to assume that the measures are supported in some convex ball, more precisely in a ball of a radius that is smaller than min (injectivity radius, $\pi/2\sqrt{\kappa}$), $\kappa \geq 0$ being an upper bound for the sectional curvature. Inspite of this restriction, of course the mollifications are quite useful, for example for creating or investigating Lipschitz maps.

More generally, using triangle comparison properties as in this section, one can also introduce and investigate metric spaces with any upper and/or lower curvature bounds. For a general treatment, we refer to [15].

The theory of spaces with lower curvature bounds in the sense of Alexandrov has been systematically developed by Yu. Burago, M. Gromov, G. Perel'man[34].

Spaces with both upper and lower curvature bounds naturally arise as limits of Riemannian manifolds with those same curvature bounds, as will be discussed in the following Survey.

Theorem 4.8.3 is a special case of a result of Y. G. Reshetnyak[215]. The proof given here is taken from [142].

If X is a complete, simply connected Riemannian manifold of nonpositive curvature, then by Theorem 4.8.2 the squared distance between any two geodesics is a convex function of the arclength parameter. One may then abstract this property and call a complete metric space (Y, d) that is a geodesic length space, i.e. for which any two points can be joined by a length minimizing curve - such curves then again are called geodesics - a metric space of nonpositive curvature if that convexity property holds. These spaces have been named after Busemann as he was the first to systematically investigate this property. A stronger property - which is still satisfied by all complete, simply connected Riemannian manifolds of nonpositive curvature as shown in Lemma 4.8.3 - is the one introduced by Alexandrov that the distances between any two points on a geodesic triangle are always less than or equal to the ones in a Euclidean triangle with the same side lengths. In fact, in the Riemannian case, both Busemann's and Alexandrov's property are equivalent to nonpositive sectional curvature. In the context of metric spaces, however, Busemann's property is more general. A reference for these theories is [142]. For applications of these concepts, see the Perspectives on §7.7.

The compactification $\bar{Y} = Y \cup Y(\infty)$ of a complete simply connected Riemannian manifold of nonpositive curvature, sometimes called a Hadamard manifold, through asymptotic equivalence classes of geodesic rays is due to Eberlein and O'Neill[71].

Anticipating some of the Perspectives for Chapter 7, the following monographs explore the geometry of nonpositive curvature: [12], [11], [69].

Exercises for Chapter 4

1. Let M_1, M_2 be submanifolds of the Riemannian manifold M. Let the curve $c : [a, b] \to M$ satisfy $c(a) \in M_1, c(b) \in M_2$. A variation $c : [a, b] \times (-\varepsilon, \varepsilon) \to M$ is called variation of $c(t)$ w.r.t. M_1, M_2 if $c(a, s) \in M_1, c(b, s) \in M_2$ for all $s \in (-\varepsilon, \varepsilon)$.

   What are the conditions for c to be an extremal of L or E w.r.t. such variations?

   Compute the second variation of E for such an extremal and express the boundary terms by the second fundamental forms of M_1 and M_2.

2. Let M be a submanifold of the Riemannian manifold N, $c : [a, b] \to N$ geodesic with $c(a) \in M, \dot{c}(a) \in (T_{c(a)}M)^{\perp}$. For $\tau \in (a, b], c(\tau)$ is called a focal point of M along c if there exists a nontrivial Jacobi field X along c with $X(a) \in T_{c(a)}M, X(\tau) = 0$.

   Show:

   a: If M has no focal point along c, then for each $\tau \in (a, b)$, c is the unique shortest connection to $c(\tau)$ when compared with all sufficiently close curves with initial point on M.

   b: Beyond a focal point, a geodesic is no longer the shortest connection to M.

3. Let $S^{n-1} := \{(x^1, \dots, x^n, 0) \in \mathbb{R}^{n+1}, \Sigma x^i x^i = 1\} \subset S^n$ be the equator sphere. Determine all focal points of S^{n-1} in S^n, and also all focal points of S^n in $\mathbb{R}^{n+1}$.

4. Let p, q be relatively prime integers. We represent S^3 as

   $$S^3 = \{(z_1, z_2) \in \mathbb{C}^2 : |z_1|^2 + |z_2|^2 = 1\}.$$

   $\mathbb{Z}_q$ operates on S^3 via

   $$(z_1, z_2) \mapsto \left(z_1 e^{\frac{2\pi i m}{q}}, z_2 e^{\frac{2\pi i m p}{q}}\right) \qquad \text{with } 0 \le m \le q - 1.$$

   Show that this operation is isometric and free. The quotient $L(q, p) := S^3/\mathbb{Z}_q$ is a so-called lens space. Compute its curvature and diameter.

5. Show that any compact odd-dimensional Riemannian manifold with positive sectional curvature is orientable. (Hint: Use the argument of the proof of Synge's theorem 4.1.2.)

6. Show that the real projective space $\mathbb{RP}^n$ (cf. Exercise 3 of Chapter 1) is orientable for odd n and nonorientable for even n. (Hint: Use Synge's theorem 4.1.2 and the preceding exercise.)

7. Show that Synge's theorem does not hold in odd dimensions. (Hint: Use the preceding exercise or Exercise 4 to give a counterexample.)

8. Try to generalize the theory of Jacobi fields to other variational problems.

9. Here is a more difficult exercise:

   Compute the second variation of volume for a minimal submanifold of a Riemannian manifold.

10. Give examples to show that a curve $c(t) = \exp_p tv$ as in Corollary 4.2.4 need not be the shortest connection of its endpoints. (Hint: Consider for example a flat torus.)

11. Let $c : [0, \infty) \to S^n$ be a geodesic parametrized by arc length. For $t > 0$, compute the dimension of the space J_c^t of Jacobi fields X along c with $X(0) = 0 = X(t)$. Use the Morse index theorem 4.3.2 to compute the indices and nullities of geodesics on S^n.

12. Show that if under the assumptions of Theorem 4.5.1 we have equality in (4.5.6) for some t with $0 < t \le \tau$, then the sectional curvature of the plane spanned by $\dot{c}(s)$ and $J(s)$ is equal to μ for all s with $0 \le s \le t$.

13. Let $p \in M, n = \dim M, r(x) = d(x, p)$,

    $$w(x,t) := \frac{1}{t^{\frac{n}{2}}} \exp\Big(-\frac{r^2(x)}{4t}\Big).$$

    In the Euclidean case, $w(x,t)$ is fundamental solution of the heat operator, i.e. for $(x,t) \ne (p,0)$

    $$\Big(\frac{\partial}{\partial t} + \Delta\Big) w(x,t) = 0.$$

    Under the assumptions of Lemma 4.7.1, derive the estimate

    $$\Big|\Big(\frac{\partial}{\partial t} + \Delta\Big) w(x,t)\Big| \le 2\Lambda^2 \frac{r^2(x)}{4t} w(x,t)$$

    for $(x,t) \ne (p,0)$.

A Short Survey on Curvature and Topology

We have now covered half of the chapters of the present textbook and the more elementary aspects of the subject. Before penetrating into more advanced topics, a short survey on some directions of global Riemannian geometry may be a useful orientation guide. Because of the size and scope of the present book, this survey needs to be selective.

A basic question, formulated in particular by H. Hopf, is to what extent the existence of a Riemannian metric with particular curvature properties restricts the topology of the underlying differentiable manifold.

The classical example is the

Gauss-Bonnet Theorem. *Let M be a compact oriented, two-dimensional Riemannian manifold with curvature K. Then its Euler characteristic is determined by*

$$\chi(M) = \frac{1}{2\pi} \int_M K \, d\mathrm{Vol}\, M.$$

We have also seen some higher dimensional examples already, namely the Theorem 4.1.2 of Synge on manifolds with positive sectional curvature, the Theorem 3.5.1 of Bochner and the Bonnet-Myers Theorem (Corollary 4.3.1) on manifolds of positive Ricci curvature. We have already seen a result for nonpositive sectional curvature, namely the Hadamard-Cartan Theorem (Corollary 4.8.1) that a simply-connected, complete manifold of nonpositive sectional curvature is diffeomorphic to some $\mathbb{R}^n$, and in Chapter 7, we shall prove the Preissmann Theorem (Corollary 7.7.2) that any abelian subgroup of the fundamental group of a compact manifold of negative sectional curvature is infinite cyclic, i.e. isomorphic to $\mathbb{Z}$. In order to put these results in a better perspective, we want to discuss the known implications of curvature properties for the topology more systematically.

We start with the implications of positive sectional curvature. Here, we have the

Sphere Theorem. *Let M be a compact, simply connected Riemannian manifold whose sectional curvature K satisfies*

$$0 < \frac{1}{4}\kappa < K \leq \kappa$$

for some fixed number κ. Then M is homeomorphic to the sphere S^n ($n = \dim M$).

This was shown by Berger[16] and Klingenberg[166]. Recently, Brendle-Schoen [31] strengthened the result by showing that M is even diffeomorphic to a sphere, using the Ricci flow method of Hamilton described below. Thus, exotic spheres cannot carry such 1/4-pinched as in the theorem.

The pinching number 1/4 is optimal in even dimensions ≥ 4, because $\mathbb{CP}^m$ (see §5.1) is simply connected, has sectional curvature between 1/4 and 1 for its Fubini-Study metric and is not homeomorphic to S^{2m} for $m > 1$. In odd dimensions, the pinching number can be decreased below 1/4, as shown by Abresch and Meyer[2, 3], but the optimal value of the pinching constant is unknown at present.

For $n = 2$ or 3, the conclusion is valid already if M has positive sectional curvature. For $n = 2$, this follows from the Gauss-Bonnet Theorem. For $n = 3$, Hamilton[116] showed that any simply connected compact manifold of positive Ricci curvature is diffeomorphic to S^3. Hamilton studied the so-called Ricci flow, i.e. he considered the evolution problem for a time dependent family of metrics g_{ij} on M with Ricci curvature R_{ij}.

$$\frac{\partial}{\partial t} g_{ij}(x,t) = \frac{2}{n} r(t) g_{ij}(x,t) - 2R_{ij}(x,t),$$

with initial metric $g_{ij}(x,0) = g^0_{ij}(x)$, where

$$r(t) = \frac{\int R(x,t)\, d\mathrm{Vol}\,(g(\cdot,t))}{\int\, d\mathrm{Vol}\,(g(\cdot,t))}$$

is the average of the scalar curvature of the metric $g_{ij}(\cdot,t)$. He showed that if g^0_{ij} is a metric with positive Ricci curvature on a compact 3-manifold, then a solution of this evolution problem exists for all time, the Ricci curvature stays positive for all t, and as $t \to \infty$, $g_{ij}(\cdot,t)$ converges to a metric of constant (positive) sectional curvature.

This method has since become important in Riemannian geometry, although in general without suitable curvature assumptions on the initial metric, singularities will develop in finite time. The analysis was carried further in [117]. For expositions, see [52, 53]. In dimension 3, the complete understanding of the formation of singularities and the continuation of the flow past such singularities was achieved by Perel'man, with profound implications for the structure and classification of 3-manifolds, see [206, 208, 207]. In particular, a consequence of Perel'man's work is the solution of the Poincaré conjecture that any compact, simply connected, 3-dimensional differentiable manifold is diffeomorphic to the 3-sphere S^3. More generally, Perel'man's work leads to a proof of Thurston's geometrization conjecture that for any compact, orientable and prime three-manifold M, there exists an embedding of a finite number of disjoint unions (possibly empty) of incompressible two-tori in M such that every component of the complement admits a locally homogeneous Riemannian metric of finite volume. Here, M is called prime if it is not diffeomorphic to S^3 and if every (topological) two-sphere that separates M into two pieces has the property that one of the two pieces is diffeomorphic to a three-ball. The possible eight homogeneous 3-manifolds that can occur in this decomposition had been identified by Thurston [251] and are

1. the three-sphere S^3
2. the Euclidean space $\mathbb{R}^3$
3. the three-dimensional hyperbolic space H^3
4. $S^2 \times \mathbb{R}$
5. $H^2 \times \mathbb{R}$
6. the three-dimensional nilpotent Heisenberg group Nil that consists of upper triangular 3×3 matrices with diagonal entries 1
7. $\mathrm{PSL}(2, \mathbb{R})$, the universal cover of the unit sphere bundle of H^2
8. the three-dimensional solvable Lie group Sol.

Kleiner-Lott [165] wrote a useful set of notes on Perel'man's papers. The first proof of Perel'man's results that contained all details, including the Poincaré and geometrization conjectures, was presented by Cao-Zhu [38] (see also [39] for a slightly modified version). Another exposition of these results was given by Morgan-Tian [198].

It is not known whether an exotic sphere can carry a metric of positive sectional curvature. Also, the problem of H. Hopf whether $S^2 \times S^2$ can carry a metric of positive sectional curvature is unsolved. The essential question is to understand compact, simply connected Riemannian manifolds of positive sectional curvature. Only very few examples of such manifolds are known. In fact, besides the general series of compact rank one symmetric spaces (spheres, complex projective spaces (see §5.1 below) in all even dimensions, quaternionic projective spaces in all dimensions that are multiples of 4, and the Cayley projective plane in dimension 16), one only knows the family of Allof-Wallach spaces in dimension 7 and the isolated examples of Eschenburg and Bazaikin.

In recent years, however, the first indications of a general structure theory seem to emerge, in the work of Petrunin, Tuschmann, Rong, Fang [212], [213], [79]. For a comprehensive treatment, see [257]. Essential points of this approach are that one studies the more general class of Alexandrov spaces of positive curvature which allows to study sequences of positively curved spaces and use compactness arguments by the result of Nikolaev quoted below, and in particular to utilize collapsing techniques and that the role of the second homotopy group becomes more prominent in determining the topological possibilities of positively curved spaces. (So, one might speculate that the theory of minimal 2-spheres developed in §8.2 might furnish useful tools for understanding the topology of positively curved spaces.)

We also mention that Wilking[263] showed that in general, a metric of positive curvature outside a finite number of points on a compact manifold cannot be deformed into a metric of positive curvature everywhere.

For positive Ricci curvature, we have already exhibited some results. An important generalization of these results is Gromov's [105, 108]

First Betti Number Theorem. *Let M be a compact Riemannian manifold of dimension n, with diameter $\leq D$ and Ricci curvature $\geq \lambda$ (i.e. $(R_{ij} - \lambda g_{ij})_{i,j}$ is a positive semidefinite tensor). Then the first Betti number satisfies*

$$b_1(M) \leq f(n, \lambda, D)$$

with an explicit function $f(n, \lambda, D)$,

$$f(n, 0, D) = n, \quad f(n, \lambda, D) = 0 \quad \text{for } \lambda > 0.$$

Finally, it has been determined which simply connected manifolds admit metrics of positive scalar curvature and which ones don't, in the work of Schoen and Yau[227], Gromov and Lawson[109] and S. Stolz[241].

In the non simply-connected case, also restrictions for positive scalar curvature are known. For example, for dimension ≤ 7, a torus cannot admit a metric of positive scalar curvature, see Schoen and Yau[226]. Such a result for any n and other restrictions on metrics of positive scalar curvature were given by Gromov and Lawson[110].

The preceding results all apply to compact manifolds. For noncompact manifolds, let us only quote the splitting theorem of Cheeger and Gromoll[46].

T Theorem.*he universal covering $\tilde{M}$ of a compact Riemannian manifold with nonnegative Ricci curvature splits isometrically as a product $\tilde{M} = N \times \mathbb{R}^k, 0 \leq k \leq \dim M$, where N is a compact manifold.*

For a more detailed survey of manifolds of nonnegative curvature, we refer to the survey article [102].

For manifolds of negative or nonpositive sectional curvature, much more is known than for those of positive curvature. Some discussion can be found in the Perspectives on §7.7. We also refer to the survey article [70].

Lohkamp[180, 181] proved that any differentiable manifold of dimension ≥ 3 admits a complete metric of negative Ricci curvature. As a consequence, negative

Ricci curvature does not imply any topological restrictions.

Riemannian manifolds of vanishing sectional curvature are called flat. The compact ones are classified by the

Bieberbach Theorem. *Let M be a compact flat Riemannian manifold of dimension n. Then its fundamental group contains a free abelian normal subgroup of rank n and finite index. Thus, M is a finite quotient of a flat torus.*

In analogy to the sphere theorem, one may ask about the structure of Riemannian manifolds that are almost flat in the sense that their curvature is close to zero. Since the curvature of a Riemannian metric may always be made arbitrarily small by rescaling the metric, the appropriate curvature condition has to be more carefully formulated in a scaling invariant manner. Let us look at the typical example:

We consider the nilpotent Lie group H of upper triangular matrices with 1's on the diagonal. Its Lie algebra is

$$\mathfrak{h} = \left\{ A = \begin{pmatrix} 0 & & a_{ij} \\ & \ddots & \\ 0 & & 0 \end{pmatrix} : a_{ij} \in \mathbb{R}, 1 \leq i < j \leq n \right\}.$$

On $\mathfrak{h}$, we may introduce a family of scalar products via

$$\|A\|_q^2 := \sum_{i<j} a_{ij}^2 q^{2(j-i)}$$

for $q > 0$. These scalar products induce left invariant Riemannian metrics on H whose curvature can be estimated as

$$\|R_q(A,B)C\|_q \leq 24(n-2)^2 \|A\|_q^2 \|B\|_q^2 \|C\|_q^2.$$

This bound is independent of q. By a q-independent rescaling, we may therefore assume that the sectional curvature satisfies $|K| \leq 1$. We let $H(\mathbb{Z})$ be the subgroup of H with integer entries, and one may thus construct left invariant metrics on H which induce on the quotient $H/H(\mathbb{Z})$ metrics with $|K| \leq 1$ and diam $< \varepsilon$, for every $\varepsilon > 0$, simply by choosing q sufficiently small.

Conversely,

Theorem. *For every n, there exists $\varepsilon(n) > 0$ with the property that any compact n-dimensional Riemannian manifold M with*

$$|K| \, (\mathrm{diam})^2 < \varepsilon(n)$$

is diffeomorphic to a finite quotient of a nilmanifold. (A nilmanifold is by definition a compact homogeneous space of a nilpotent Lie group.)

This is due to Gromov, see [35] for an exposition, and for the refinement that M as above is actually an infranilmanifold by Ruh[216].

In order to place this result in a broader context, we introduce the notions of convergence and collapse of manifolds. For compact subsets A_1, A_2 of a metric space Z, we define

$$d_H^Z(A_1, A_2) := \inf\{r : A_1 \subset \cup_{x \in A_2} \overset{\circ}{B}(x, r),\ A_2 \subset \cup_{x \in A_1} \overset{\circ}{B}(x, r)\},$$

where $\overset{\circ}{B}(x, r) := \{y \in Z : d(x, y) < r\}$.

For compact metric spaces X_1, X_2, their Hausdorff distance is

$$\begin{aligned} d_H(X_1, X_2) := \inf_Z \ &\{d_H^Z(i(X_1), j(X_2)), \\ &\text{where } i : X_1 \to Z, j : X_2 \to Z \text{ are isometries into a metric space } Z\}. \end{aligned}$$

This distance then defines the notion of Hausdorff convergence of compact metric spaces. Let M_0 be a compact differentiable manifold of dimension n. We say that M_0 admits a collapse to a compact metric space X of lower (Hausdorff) dimension than M_0 if there exists a sequence $(g_j)_{j \in \mathbb{N}}$ of Riemannian metrics with uniformly bounded curvature on M_0 such that the Riemannian manifolds (M_0, g_j) as metric spaces converge to X. This phenomenon has been introduced and studied by Cheeger, Gromov, and Fukaya [47, 48], [89].

It is easy to see that any torus can collapse to a point; for this purpose, one just rescales a given flat metric by a factor ε and lets $\varepsilon \to 0$. The diameter then shrinks to 0, while the curvature always remains 0. Berger showed that S^3 admits a collapse onto S^2. The construction is based on the Hopf fibration $\pi : S^3 \to S^2 = \mathbb{CP}^1$ (see §5.1), and one lets the fibers shrink to zero in length.

In this terminology the above theorem (as refined by Ruh) says that those manifolds that can collapse to a point are precisely the infranilmanifolds. More recently, it was shown by Tuschmann[256] that any manifold that admits a collapse onto some flat orbifold is homeomorphic to an infrasolvmanifold and conversely, that any infrasolvmanifold also admits a sequence of Riemannian metrics for which it collapses to a compact flat orbifold. Here, an infrasolvmanifold is a certain type of quotient of a solvable Lie group.

We next mention the following result of Cheeger[44], with the improvements by Peters[209].

Finiteness Theorem. *For any $n \in \mathbb{N}$, $\Lambda < \infty$, $D < \infty, v > 0$, the class of compact differentiable manifolds of dimension n admitting a Riemannian metric with*

$$|K| \le \Lambda,\ \textit{diam} \le D,\ \textit{Volume} \ge v$$

consists of at most finitely many diffeomorphism types.

The lower positive uniform bound on volume prevents collapsing and is necessary for this result to hold.

Diffeomorphism finiteness can however actually also be obtained if no volume bounds are present and collapsing may take place.

This is demonstrated by the following recent finiteness theorem by Petrunin and Tuschmann[213]. Instead of volume bounds this result only uses a merely topological condition:

π_2-Finiteness Theorem. *For any $n \in \mathbb{N}$, $\Lambda < \infty$, and $D < \infty$, the class of compact simply connected differentiable manifolds of dimension n with finite second homotopy group admitting a Riemannian metric with*

$$|K| \geq \Lambda, \; \mathit{diam} \; \leq D$$

consists of at most finitely many diffeomorphism types.

Cheeger's finiteness theorem was refined in the so-called Gromov convergence theorem, which we are going to present in the form proved by Peters[210] and Greene and Wu[100].

Convergence Theorem. *Let $(M_j, g_j)_{j\in\mathbb{N}}$ be a sequence of Riemannian manifolds of dimension n satisfying the assumptions of the finiteness theorem with Λ, D, v independent of j. Then a subsequence converges in the Hausdorff distance and (after applying suitable diffeomorphisms) also in the (much stronger) $C^{1,\alpha}$ topology (for any $0 < \alpha < 1$) to a differentiable manifold with a $C^{1,\alpha}$-metric.*

Such a family of manifolds is known to have a uniform lower bound on their injectivity radius. The crucial ingredient in the proof then are the a-priori estimates of Jost-Karcher for harmonic coordinates described in the Perspectives on §7.7. Namely, these estimates imply convergence of subsequences of local coordinates on balls of fixed size, and the limits of these coordinates then are coordinates for the limiting manifold.

Nikolaev[203] showed that the Hausdorff limits of sequences of compact n-dimensional Riemannian manifolds of uniformly bounded curvature and diameter and with volume bounded away from 0 uniformly are precisely the smooth compact n-manifolds with metrics of bounded curvature in the sense of Alexandrov.

Let us conclude this short survey by listing some other textbooks on Riemannian geometry that treat various selected topics of global differential geometry and which complement the present book, Chavel[43], Cheeger and Ebin[45], do Carmo [65], Gallot, Hulin and Lafontaine[90], Gromoll, Klingenberg and Meyer[103], Klingenberg [168], Petersen[211], Sakai[220]. Finally, we wish to mention the stimulating survey Berger[17].

Chapter 5

Symmetric Spaces and Kähler Manifolds

5.1 Complex Projective Space

We consider the complex vector space $\mathbb{C}^{n+1}$. A complex linear subspace of $\mathbb{C}^{n+1}$ of complex dimension one is called a line. We define the complex projective space $\mathbb{CP}^n$ as the space of all lines in $\mathbb{C}^{n+1}$. Thus, $\mathbb{CP}^n$ is the quotient of $\mathbb{C}^{n+1}\backslash\{0\}$ by the equivalence relation

$$Z \sim W :\Longleftrightarrow \exists \lambda \in \mathbb{C}\backslash\{0\} : W = \lambda Z.$$

Namely, two points of $\mathbb{C}^{n+1}\backslash\{0\}$ are equivalent iff they are complex linearly dependent, i.e. lie on the same line. The equivalence class of Z is denoted by [Z] .

We also write

$$Z = (Z^0, \ldots, Z^n) \in \mathbb{C}^{n+1}$$

and define

$$U_i := \{[Z] : Z^i \neq 0\} \subset \mathbb{CP}^n,$$

i.e. the space of all lines not contained in the complex hyperplane $\{Z^i = 0\}$. We then obtain a bijection

$$\varphi_i : U_i \to \mathbb{C}^n$$

via

$$\varphi_i([Z^0, \ldots, Z^n]) := \left(\frac{Z^0}{Z^i}, \ldots, \frac{Z^{i-1}}{Z^i}, \frac{Z^{i+1}}{Z^i}, \ldots, \frac{Z^n}{Z^i} \right).$$

$\mathbb{CP}^n$ thus becomes a differentiable manifold, because the transition maps

$$\begin{aligned}\varphi_j \circ \varphi_i^{-1} : \varphi_i(U_i \cap U_j) &= \{z = (z^1, \dots, z^n) \in \mathbb{C}^n : z^j \neq 0\} \to \varphi_j(U_i \cap U_j)\\ \varphi_j \circ \varphi_i^{-1}(z^1, \dots, z^n) &= \varphi_j([z^1, \dots, z^i, 1, z^{i+1}, \dots, z^n])\\ &= \left(\frac{z^1}{z^j}, \dots, \frac{z^i}{z^j}, \frac{z^{i+1}}{z^j}, \dots, \frac{z^{j-1}}{z^j}, \frac{z^{j+1}}{z^j}, \dots, \frac{z^n}{z^j}\right),\end{aligned}$$

(w.l.o.g. $i < j$) are diffeomorphisms. They are even holomorphic; namely, with $z^k = x^k + iy^k$ ($i = \sqrt{-1}$) and

$$\begin{aligned}\frac{\partial}{\partial z^k} &:= \frac{1}{2}\left(\frac{\partial}{\partial x^k} - i\frac{\partial}{\partial y^k}\right),\\ \frac{\partial}{\partial z^{\bar k}} &:= \frac{1}{2}\left(\frac{\partial}{\partial x^k} + i\frac{\partial}{\partial y^k}\right),\end{aligned}$$

we have

$$\frac{\partial}{\partial z^{\bar k}}\varphi_j \circ \varphi_i^{-1}(z^1, \dots, z^n) = 0 \quad \text{for } k = 1, \dots, n.$$

Thus, $\mathbb{CP}^n$ is a complex manifold in the sense of Definition 1.8.5.

We consider the $(n+1)$-tuple

$$(Z^0, \dots, Z^n),$$

which satisfies the restriction that not all Z^j vanish identically, as homogeneous coordinates $[Z] = [Z^0, \dots, Z^n]$. These are not coordinates in the usual sense, because a point in a manifold of dimension n here is described by $(n+1)$ complex numbers. The coordinates are defined only up to multiplication with an arbitrary nonvanishing complex number λ

$$[Z^0, \dots, Z^n] = [\lambda Z^0, \dots, \lambda Z^n];$$

this fact is expressed by the adjective "homogeneous". The coordinates $(z^1, \dots, z^n)$ defined by the charts φ_i are called *Euclidean coordinates*. The vector space structure of $\mathbb{C}^{n+1}$ induces an analogous structure on $\mathbb{CP}^n$ by homogenization: Each linear inclusion $\mathbb{C}^{m+1} \subset \mathbb{C}^{n+1}$ induces an inclusion $\mathbb{CP}^m \subset \mathbb{CP}^n$. The image of such an inclusion is called a *linear subspace*. The image of a hyperplane in $\mathbb{C}^{n+1}$ is again called a *hyperplane*, and the image of a twodimensional space $\mathbb{C}^2$ is called a *line*.

Instead of considering $\mathbb{CP}^n$ as a quotient of $\mathbb{C}^{n+1}\backslash\{0\}$, we may also view it as a compactification of $\mathbb{C}^n$. One says that the hyperplane H at infinity is added to $\mathbb{C}^n$; this means the following: The inclusion

$$\mathbb{C}^n \to \mathbb{CP}^n$$

is given by

$$(z^1, \dots, z^n) \mapsto [1, z^1, \dots, z^n].$$

Then

$$\mathbb{CP}^n\backslash\mathbb{C}^n = \{[Z] = [0, Z^1, \dots, Z^n]\} =: H,$$

and H is a hyperplane $\mathbb{CP}^{n-1}$.

It follows that

$$\mathbb{CP}^n = \mathbb{C}^n \cup \mathbb{CP}^{n-1} = \mathbb{C}^n \cup \mathbb{C}^{n-1} \cup \ldots \cup \mathbb{C}^0, \tag{5.1.1}$$

(disjoint union). Topologically, $\mathbb{CP}^n$ thus is the union of $(n+1)$ cells of real dimension $0, 2, \ldots, 2n$. With the help of the Mayer-Vietoris sequence of cohomology theory,[1] we may easily compute the cohomology of $\mathbb{CP}^n$ from (5.1.1). In order to represent $\mathbb{CP}^n$ as the union of two open sets as required for the application of this sequence, we put

$$U := \mathbb{C}^n, V := \{z \in \mathbb{C}^n : \|z\| = z^j z^{\bar{j}} > 1\} \cup \mathbb{CP}^{n-1} \text{ (as in (5.1.1).}$$

Then V has $\mathbb{CP}^{n-1}$ as a deformation retract (consider

$$r_t : V \to V, r_t(z) = tz \text{ for } z \in \mathbb{C}^n, r_t(w) = w \text{ for } w \in \mathbb{CP}^{n-1}$$

and let t run from 1 to ∞), and $U \cap V$ is homotopically equivalent to the unit sphere S^{2n-1} of $\mathbb{C}^n$.

We now observe first that $\mathbb{CP}^1$ is diffeomorphic to S^2. It actually follows already from (5.1.1) that the two spaces are homeomorphic. In order to see that they are diffeomorphic, we recall that S^2 may be described via stereographic projection from the north and south pole by two charts with image $\mathbb{C}$ and transition map

$$z \mapsto \frac{1}{z}$$

(cf. §1.1). This, however, is nothing but the transition map

$$[1, z] \to [\frac{1}{z}, 1]$$

of $\mathbb{CP}^1$.

In particular, $H^0(\mathbb{CP}^1) = H^2(\mathbb{CP}^1) = \mathbb{R}, H^1(\mathbb{CP}^1) = 0$. For the general case, the relevant portion of the Mayer-Vietoris sequence is

$$H^{q-1}(S^{2n-1}) \to H^q(\mathbb{CP}^n) \to H^q(\mathbb{C}^n) \oplus H^q(\mathbb{CP}^{n-1}) \to H^q(S^{2n-1}). \tag{5.1.2}$$

We now want to show by induction w.r.t. n that

$$H^q(\mathbb{CP}^n) = \begin{cases} \mathbb{R} & \text{for } q = 0, 2, \ldots, 2n, \\ 0 & \text{otherwise.} \end{cases}$$

This is obvious for $q = 0$. For $2 \le q \le 2n-1$ we have $H^{q-1}(S^{2n-1}) = 0, H^q(\mathbb{C}^n) = 0$, and for $q = 2, \ldots, 2n-2$ we obtain from (5.1.2) that $H^q(\mathbb{CP}^n) = \mathbb{R}$ since by inductive assumption $H^q(\mathbb{CP}^{n-1}) = \mathbb{R}$, while for $q = 1, 3, \ldots, 2n-1$, again by inductive

[1] This sequence has been derived in the previous editions of this textbook, but for the present edition, we are not including an introduction to cohomology theory anymore as that can be readily found in standard textbooks on algebraic topology.

assumption, $H^q(\mathbb{CP}^{n-1}) = 0$, hence also $H^q(\mathbb{CP}^n) = 0$. The case $q = 1$ is similar. $H^{2n}(\mathbb{CP}^n) = \mathbb{R}$ again follows from (5.1.2) or even more easily from Corollary 2.2.2.

Let us also show that $\mathbb{CP}^n$ can be considered as a quotient of the unit sphere S^{2n+1} in $\mathbb{C}^{n+1}$. Namely, each line in $\mathbb{C}^{n+1}$ intersects S^{2n+1} in a circle S^1, and we obtain the point of $\mathbb{CP}^n$ defined by this line by identifying all points on that circle.

The projection

$$\pi : S^{2n+1} \to \mathbb{CP}^n$$

is called the *Hopf map.* In particular, since $\mathbb{CP}^1 = S^2$, we obtain a map

$$\pi : S^3 \to S^2$$

with fiber S^1.

The unitary group $\mathrm{U}(n+1)$ operates on $\mathbb{C}^{n+1}$ and transforms complex subspaces into complex subspaces, in particular lines into lines. Therefore, $\mathrm{U}(n+1)$ also operates on $\mathbb{CP}^n$.

We now want to introduce a metric on $\mathbb{CP}^n$. For this purpose, let

$$\pi : \mathbb{C}^{n+1}\backslash\{0\} \to \mathbb{CP}^n$$

be the standard projection, $U \subset \mathbb{CP}^n$, $Z : U \to \mathbb{C}^{n+1}\backslash\{0\}$ a lift of π, i.e. a holomorphic map with $\pi \circ Z = \mathrm{id}$. We put

$$\omega := \frac{i}{2}\partial\bar{\partial} \log \|Z\|^2, \tag{5.1.3}$$

putting for abbreviation

$$\begin{aligned} \partial &:= \frac{\partial}{\partial Z^j} dZ^j, \\ \bar{\partial} &:= \frac{\partial}{\partial Z^{\bar{k}}} dZ^{\bar{k}}. \end{aligned}$$

If $Z' : U \to \mathbb{C}^{n+1}\backslash\{0\}$ is another lift, we have

$$Z' = \varphi Z,$$

where φ is a nowhere vanishing holomorphic function.

Hence

$$\begin{aligned} \frac{i}{2}\partial\bar{\partial}\log\|Z'\|^2 &= \frac{i}{2}\partial\bar{\partial}(\log\|Z\|^2 + \log\varphi + \log\bar{\varphi}) \\ &= \omega + \frac{i}{2}(\partial\bar{\partial}\log\varphi - \bar{\partial}\partial\log\bar{\varphi}) \quad \text{(cf. (5.2.3) below)} \\ &= \omega, \end{aligned}$$

since $\bar{\partial}\log\varphi = 0 = \partial\log\bar{\varphi}$, because φ is holomorphic and nowhere vanishing. Therefore, ω does not depend on the choice of chart and thus defines a 2-form on $\mathbb{CP}^n$.

We want to represent ω in local coordinates; for this purpose, let as above

$$U_0 = \{[Z^0, \dots, Z^n] : Z^0 \neq 0\},$$

since $z^i = \frac{Z^i}{Z^0}$ on U_0, $Z = (1, z^1, \dots, z^n)$ is a lift of π over U_0. Then

$$\begin{aligned}\omega &= \frac{i}{2}\partial\bar{\partial}\log(1 + z^j z^{\bar{j}}) \\ &= \frac{i}{2}\partial\left(\frac{z^j dz^{\bar{j}}}{1 + z^k z^{\bar{k}}}\right),\end{aligned}$$

hence

$$\omega = \frac{i}{2}\left\{\frac{dz^j \wedge dz^{\bar{j}}}{1 + z^k z^{\bar{k}}} - \frac{z^{\bar{j}} z^k dz^j \wedge dz^{\bar{k}}}{(1 + z^\ell z^{\bar{\ell}})^2}\right\}. \tag{5.1.4}$$

At $[1, 0, \dots, 0]$ again

$$\omega = \frac{i}{2} dz^j \wedge dz^{\bar{j}} = dx^j \wedge dy^j. \tag{5.1.5}$$

Thus, ω is positive definite (in a sense to be made precise in Definition 5.1.1) at the point $[1, 0, \dots, 0]$. Since ω is invariant under the operation of $\mathrm{U}(n+1)$ on $\mathbb{CP}^n$, it is therefore positive definite everywhere.

We want to generalize the object ω just introduced in the following

Definition 5.1.1. Let M be a complex manifold with local coordinates $z=(z^1, \dots, z^n)$. A *Hermitian metric* on M is given by an expression of the form

$$h_{j\bar{k}}(z) dz^j \otimes dz^{\bar{k}}$$

where $h_{j\bar{k}}(z)$ depends smoothly (i.e. C^∞) on z and is positive definite and Hermitian for every z.

The expression

$$\frac{i}{2} h_{j\bar{k}}(z) dz^j \wedge dz^{\bar{k}}$$

is called the *Kähler form* of the Hermitian metric.

That $h_{j\bar{k}}$ is Hermitian means

$$h_{k\bar{j}} = \overline{h_{j\bar{k}}}. \tag{5.1.6}$$

We also put

$$\begin{aligned} h_{\bar{k}j} &= h_{j\bar{k}}, \\ h_{jk} &= 0 = h_{\bar{j}\bar{k}}. \end{aligned} \tag{5.1.7}$$

Let now

$$\begin{aligned} v &= v^j \frac{\partial}{\partial z^j} + v^{\bar{j}} \frac{\partial}{\partial z^{\bar{j}}}, \\ w &= w^j \frac{\partial}{\partial z^j} + w^{\bar{j}} \frac{\partial}{\partial z^{\bar{j}}} \end{aligned}$$

be tangent vectors (with complex coefficients) in $z \in M$.

We put

$$\langle v, w\rangle := h_{j\bar{k}}(z)v^j w^{\bar{k}} + h_{k\bar{j}}(z)v^{\bar{j}} w^k. \tag{5.1.8}$$

If v and w are tangent vectors with real coefficients, i.e.

$$v = v^j \frac{\partial}{\partial x^j} + v^{j+n}\frac{\partial}{\partial y^j},$$
$$w = w^j \frac{\partial}{\partial x^j} + w^{j+n}\frac{\partial}{\partial y^j},$$

with $v^\alpha, w^\alpha \in \mathbb{R}, \alpha = 1, \ldots, 2n$, then because of

$$\frac{\partial}{\partial x^j} = \frac{\partial}{\partial z^j} + \frac{\partial}{\partial z^{\bar{j}}},$$
$$\frac{\partial}{\partial y^j} = i\left(\frac{\partial}{\partial z^j} - \frac{\partial}{\partial z^{\bar{j}}}\right),$$

we have

$$\begin{aligned}\langle v, w\rangle &= h_{j\bar{k}}(v^j + iv^{j+n})(w^k - iw^{k+n}) + \overline{h_{j\bar{k}}}(v^j - iv^{j+n})(w^k + iw^{k+n})\\ &= 2\operatorname{Re} h_{j\bar{k}}(v^j w^k + v^{j+n}w^{k+n}) + 2\operatorname{Im} h_{j\bar{k}}(v^j w^{k+n} - w^k v^{j+n}).\end{aligned}$$

Consequently, each Hermitian metric induces a Riemannian one. This justifies the name "Hermitian metric".

Definition 5.1.2. A Hermitian metric $h_{j\bar{k}}dz^j \otimes dz^{\bar{k}}$ is called a *Kähler metric*, if for every z there exists a neighborhood U of z and a function $F : U \to \mathbb{R}$ with $\frac{i}{2}h_{j\bar{k}}dz^j \wedge dz^{\bar{k}} = \partial\bar{\partial}F$. $\partial\bar{\partial}F$ then is called the *Kähler form.*

The 2-form ω from (5.1.3) defines a Kähler metric on $\mathbb{CP}^n$, called the *Fubini-Study metric.*

This metric has many special properties. In particular, the operation of $\mathrm{U}(n+1)$ on $\mathbb{C}^{n+1}$ induces an isometric operation of $\mathrm{U}(n+1)$ on $\mathbb{CP}^n$ equipped with this metric. This follows from (5.1.5) and the fact that $\|\cdot\|$ is invariant under the operation of $\mathrm{SO}(2n+2)$, hence in particular invariant under the one of $\mathrm{U}(n+1)$.

For a line L in $\mathbb{C}^{n+1}$ we may also consider the reflection at L, i.e.

$$s_{|L} = \mathrm{id},$$
$$s_{|L^\perp} = -\mathrm{id}.$$

s then induces an isometry σ of $\mathbb{CP}^n$ (equipped with the Fubini-Study metric) with fixed point $\pi(L)$ and

$$d\sigma = -\mathrm{id} : T_{\pi(L)}\mathbb{CP}^n \to T_{\pi(L)}\mathbb{CP}^n.$$

In particular

$$\sigma^2 = \mathrm{id}.$$

Definition 5.1.3. A Riemannian manifold is called *symmetric* if for every $p \in M$ there exists an isometry $\sigma_p : M \to M$ with

$$\begin{aligned} \sigma_p(p) &= p, \\ D\sigma_p(p) &= -\mathrm{id} \qquad \text{(as a self map of } T_pM). \end{aligned}$$

Such an isometry is also called an *involution*.

Thus, $\mathbb{CP}^n$, equipped with the Fubini-Study metric, is a symmetric space.

Thus, complex projective space carries two different structures: it is both a Kähler manifold and a symmetric space. The rest of this chapter is devoted to an investigation of those structures.

5.2 Kähler Manifolds

In the preceding section, we have introduced complex projective space as an example of a Kähler manifold. There exist simpler examples. Namely, $\mathbb{C}^d$ with its standard Euclidean metric is a Kähler manifold with Kähler form

$$\omega = \frac{i}{2} dz^j \wedge dz^{\bar{j}}.$$

Also, any complex 1-dimensional manifold Σ, that is, any Riemann surface (see §8.1) is automatically a Kähler manifold since $d\omega$ is a 3-form and therefore vanishes on the real 2-dimensional manifold Σ.

Moreover, any complex submanifold N of a Kähler manifold M is automatically a Kähler manifold itself; we simply need to restrict the local Kähler potential F of M to N. Therefore, in particular, all complex projective manifolds, that is, those that admit a holomorphic embedding into some complex projective space, are Kähler manifolds. This makes Kähler geometry a useful tool in algebraic geometry.

In this section, we want to give a systematic introduction to Kähler geometry. We start by recalling the rules from Lemma 1.8.4 for the calculus of the operators ∂ and $\bar{\partial}$:

$$d = \partial + \bar{\partial}, \tag{5.2.1}$$

$$\partial\partial = \bar{\partial}\bar{\partial} = 0, \tag{5.2.2}$$

$$\partial\bar{\partial} = -\bar{\partial}\partial\,. \tag{5.2.3}$$

We can now state various equivalent versions of the Kähler condition

$$\omega := \frac{i}{2} h_{j\bar{k}} dz^j \wedge dz^{\bar{k}} = \partial\bar{\partial}F, \tag{5.2.4}$$

that is, that for every z, there exist some neighborhood U and some function F defined on U with this property.

Theorem 5.2.1. *The following conditions are equivalent to a Hermitian manifold M being Kähler.*

(i) The Kähler form ω is closed, i.e.

$$d\omega = 0. \tag{5.2.5}$$

(ii) In local (holomorphic) coordinates

$$\frac{\partial h_{i\bar{j}}}{\partial z^k} = \frac{\partial h_{k\bar{j}}}{\partial z^i}, \qquad \textit{for all } i, j, k, \tag{5.2.6}$$

or equivalently,

$$\frac{\partial h_{i\bar{j}}}{\partial z^{\bar{\ell}}} = \frac{\partial h_{i\bar{\ell}}}{\partial z^{\bar{j}}}, \qquad \textit{for all } i, j, \ell. \tag{5.2.7}$$

(iii) At each $z_0 \in M$, **holomorphic** *normal coordinates can be introduced, i.e.*

$$h_{i\bar{j}}(z_0) = \delta_{ij}, \quad \frac{\partial h_{i\bar{j}}}{\partial z^k}(z_0) = 0 = \frac{\partial h_{i\bar{j}}}{\partial z^{\bar{\ell}}}(z_0), \quad \textit{for all } i, j, k, \ell. \tag{5.2.8}$$

In other words, we can find holomorphic coordinates near any z_0, which we then take the liberty to identify with 0, so that for z near 0,

$$h_{i\bar{j}}(z) = \delta_{ij} + \mathrm{O}(|z|^2). \tag{5.2.9}$$

The last condition expresses the essential content of the Kähler condition, namely the compatibility of the Riemannian and the complex structure. Condition (i) has the advantage of expressing the Kähler condition in a global, coordinate invariant manner. This will make it particularly useful.

Proof. We first show that the Kähler condition implies (i).

$$d(\partial\bar{\partial}F) = (\partial + \bar{\partial})(\partial\bar{\partial}F) = \partial\partial\bar{\partial}F - \partial\bar{\partial}\bar{\partial}F = 0 \quad \text{by (5.2.2), (5.2.3)}.$$

This yields (i). (ii) is the local coordinate version of (i). In turn, (i) implies the Kähler condition by the Frobenius Theorem. Namely, since ω is closed, $d\omega = 0$, on each sufficiently small open set U, we can find a 1-form η with $d\eta = \omega$. ω is a $(1,1)$-form, and so, when we decompose the 1-form η into a $(1,0)$- and a $(0,1)$-form, $\eta = \eta^{1,0} + \eta^{0,1}$, we have

$$\omega = d\eta = (\partial + \bar{\partial})\eta = \partial\eta^{0,1} + \bar{\partial}\eta^{1,0}$$

with

$$\bar{\partial}\eta^{0,1} = 0 = \partial\eta^{1,0}.$$

From the last condition, on our sufficiently small U, we can then find s functions α and β with

$$\eta^{0,1} = \bar{\partial}\alpha, \; \eta^{1,0} = -\partial\beta,$$

and so, keeping (5.2.3) in mind,

$$\omega = \partial\bar{\partial}(\alpha + \beta).$$

Since ω is real ($\bar{\omega} = \omega$), we may then also assume that the function $F := \alpha + \beta$ is real, and we have deduced the Kähler condition from (i). It thus only remains to show that (iii) is equivalent to the other conditions. It is clear that (5.2.9) implies $d\omega(z_0) = 0$, that is, (i). For the converse, we first achieve by a linear change of coordinates that $h_{i\bar{j}}(z_0) = \delta_{ij}$. Thus,

$$\omega = \frac{i}{2} h_{j\bar{k}} dz^j \wedge dz^{\bar{k}} = \frac{i}{2}(\delta_{jk} + a_{jkl}z^l + a_{jk\bar{l}}z^{\bar{l}})dz^j \wedge dz^{\bar{k}}.$$

Here, (5.1.6) implies that

$$a_{kj\bar{l}} = \bar{a}_{jkl}, \tag{5.2.10}$$

and (i) yields

$$a_{jkl} = a_{lkj}. \tag{5.2.11}$$

We shall now make the linear terms disappear by the following change of coordinates

$$z^j = \zeta^j - \frac{1}{2} a_{ljk}\zeta^k\zeta^l. \tag{5.2.12}$$

Using (5.2.10), (5.2.11), this yields

$$\begin{aligned}
\omega &= \frac{i}{2}(d\zeta^j - a_{ljk}\zeta^k d\zeta^l) \wedge (d\zeta^{\bar{j}} - \bar{a}_{njm}\zeta^{\bar{m}} d\zeta^{\bar{n}}) \\
&\quad + \frac{i}{2}(a_{jkl}\zeta^l + a_{jk\bar{l}}\zeta^{\bar{l}})d\zeta^j \wedge d\zeta^{\bar{k}} + \mathrm{O}(|z|^2) \\
&= \frac{i}{2}\delta_{jk} d\zeta^j \wedge d\zeta^{\bar{k}} + \mathrm{O}(|z|^2),
\end{aligned}$$

This is (5.2.9). □

In particular, the Kähler form ω, being closed, represents a (complex) cohomology class, i.e. an element of $H^2(M) \otimes \mathbb{C}$.

Lemma 5.2.1. *The Kähler form μ of a Kähler metric on a complex manifold M with $\dim_{\mathbb{C}} M = n$ satisfies*

$$\mu^n = n! * (1). \tag{5.2.13}$$

Proof. (5.2.13) is a pointwise identity. Let $p \in M$. Since a Hermitian form can be diagonalized by a unitary transformation, we may assume that local coordinates are chosen such that at p

$$\mu = \frac{i}{2} dz^j \wedge dz^{\bar{j}} = dx^j \wedge dy^j.$$

Therefore,

$$\begin{aligned} \mu^n &= n!\, dx^1 \wedge dy^1 \wedge dx^2 \wedge dy^2 \wedge \ldots \wedge dx^n \wedge dy^n \\ &= n! * (1), \end{aligned}$$

since $dx^1, dy^1, dx^2, dy^2, \ldots, dx^n, dy^n$ constitute a positive orthonormal basis of $T_p^* M$. □

Corollary 5.2.1. *The Kähler form of a Kähler metric on a compact manifold represents a nontrivial cohomology class, and so does every $\mu^j, j = 1, \ldots, n$. Therefore, the cohomology groups $H^2(M), H^4(M), \ldots, H^{2n}(M)$ of a compact Kähler manifold are nontrivial.*

Proof. By Lemma 5.2.1

$$\int_M \mu^n = n! \int_M *(1) = n! \operatorname{Vol}(M) > 0.$$

If we now had $\mu^j = d\psi$ for some $j \in \{1, \ldots, n\}$, then we would also have

$$\begin{aligned} \int_M \mu^n &= \int_M \mu^j \wedge \mu^{n-j} \\ &= \int_M d\psi \wedge \mu^{n-j} \\ &= \int_M d(\psi \wedge \mu^{n-j}) \quad \text{since } \mu \text{ is closed by Theorem 5.2.1} \\ &= 0 \quad \text{by Stokes Theorem.} \end{aligned}$$

This is a contradiction. □

Corollary 5.2.1 expresses an instance of the important fact that the existence of a Kähler metric yields nontrivial topological restrictions for a manifold. We shall soon derive some deeper such results. Before doing that, however, we state some useful local formulae in Kähler geometry.

For the inverse of the Hermitian metric $(h_{i\bar{j}})$, we use the convention

$$h^{i\bar{j}} h_{k\bar{j}} = \delta_{ik} \tag{5.2.14}$$

(note the switch of indices). With $h := \det(h_{i\bar{j}})$, the Laplace-Beltrami operator (2.1.13) becomes

$$\Delta = -\frac{1}{h}\frac{\partial}{\partial z^i}\left(h\, h^{i\bar{j}}\frac{\partial}{\partial z^{\bar{j}}}\right) = -h^{i\bar{j}}\frac{\partial^2}{\partial z^i \partial z^{\bar{j}}}. \tag{5.2.15}$$

This is most easily seen by using the coordinates given in (iii) of Theorem 5.2.1 and then observing that both expressions transform in the right manner under coordinate transformations.

Similarly, we have for the Christoffel symbols of a Kähler manifold

$$\begin{aligned} \Gamma^k_{ij} &= h^{k\bar{\ell}} h_{i\bar{\ell},j}, \\ \Gamma^{\bar{k}}_{\bar{i}\bar{j}} &= h^{m\bar{k}} h_{m\bar{i},\bar{j}}, \end{aligned} \tag{5.2.16}$$

because of (5.1.6), (5.2.6), (5.2.7). All other Christoffel symbols, that is, all those that contain both bared and unbared indices, vanish. Using this, the formulae (3.1.31), (3.3.6) for the Riemannian curvature tensor also simplifies to become

$$R_{i\bar{j}k\bar{\ell}} = \frac{\partial^2}{\partial z^k \partial z^{\bar{\ell}}} h_{i\bar{j}} - h^{m\bar{n}}\left(\frac{\partial}{\partial z^k} h_{i\bar{n}}\right)\left(\frac{\partial}{\partial z^{\bar{\ell}}} h_{m\bar{j}}\right). \tag{5.2.17}$$

Also,

$$R_{ij\bar{k}\bar{\ell}} = R_{\bar{i}\bar{j}k\ell} = 0. \tag{5.2.18}$$

With the first Bianchi identity (3.3.8), and

$$R_{i\bar{\ell}\bar{j}k} = -R_{i\bar{\ell}k\bar{j}}, \tag{5.2.19}$$

we then obtain

$$R_{i\bar{j}k\bar{\ell}} = R_{i\bar{\ell}k\bar{j}}, \tag{5.2.20}$$

and analogously,

$$R_{k\bar{j}i\bar{\ell}} = R_{i\bar{j}k\bar{\ell}}. \tag{5.2.21}$$

The Ricci tensor (3.3.18) of a Kähler metric is given by

$$R_{k\bar{\ell}} = h^{i\bar{j}} R_{i\bar{j}k\bar{\ell}}. \tag{5.2.22}$$

From (5.2.17), we then have a simple formula for the so-called Ricci form

$$R_{k\bar{\ell}} dz^k \wedge dz^{\bar{\ell}} = -\partial\bar{\partial} \log\det(h_{i\bar{j}}). \tag{5.2.23}$$

Finally, the scalar curvature of a Kähler metric is

$$R = \Delta \log\det(h_{i\bar{j}}). \tag{5.2.24}$$

The Ricci form is closed by (5.2.1) – (5.2.3) and therefore defines a cohomology class, the so-called first Chern class

$$c_1(M) := \frac{i}{2\pi} R_{k\bar{\ell}} dz^k \wedge dz^{\bar{\ell}}, \tag{5.2.25}$$

which is independent of the choice of Kähler metric. Namely, if $h'_{i\bar{j}}$ is another Kähler metric on M with Ricci class

$$R'_{k\bar{\ell}} dz^k \wedge dz^{\bar{\ell}} = -\partial\bar{\partial} \log \det(h'_{i\bar{j}}),$$

then

$$(R_{k\bar{\ell}} - R'_{k\bar{\ell}}) dz^k \wedge dz^{\bar{\ell}} = -\partial\bar{\partial} \log \frac{\det(h_{i\bar{j}})}{\det(h'_{i\bar{j}})}, \tag{5.2.26}$$

and this is exact since $\frac{\det(h_{i\bar{j}})}{\det(h'_{i\bar{j}})}$ is a globally defined function independent of the choice of coordinates (this follows from the transformation formula (1.4.3).

We recall from the end of §1.8 that on a complex manifold, the space of (complex-valued) k-forms $\Omega^k(M)$ admits a decomposition

$$\Omega^k(M) = \sum_{p+q=k} \Omega^{p,q}(M). \tag{5.2.27}$$

The elements of $\Omega^{p,q}$ are called (p,q)-forms. $\Omega^{p,q}$ is generated by forms of the type

$$\varphi(z) dz^{i_1} \wedge \ldots \wedge dz^{i_p} \wedge dz^{\bar{j}_1} \wedge \ldots \wedge dz^{\bar{j}_q}. \tag{5.2.28}$$

We now use the Kähler form ω to define

$$L : \Omega^{p,q} \to \Omega^{p+1,q+1}, \quad L(\eta) := \eta \wedge \omega \tag{5.2.29}$$

and its adjoint w.r.t. the L^2-product

$$(\eta, \sigma) = \int_M \eta \wedge *\bar{\sigma}, \tag{5.2.30}$$

(where the star-operator $*$ introduced in §2.1 has been linearly extended from the real to the complex case),

$$\Lambda := L^* : \Omega^{p,q} \to \Omega^{p-1,q-1}.$$

For example, for $\eta = \eta_{j\bar{k}} dz^j \wedge dz^{\bar{k}}$, recalling

$$\omega = \frac{i}{2} h_{j\bar{k}} dz^j \wedge dz^{\bar{k}},$$

we have,

$$\Lambda(\eta) = -2i h^{j\bar{k}} \eta_{j\bar{k}}. \tag{5.2.31}$$

Theorem 5.2.2. *On a Kähler manifold, we have the identities*

$$[\Lambda, \bar{\partial}] = -i\partial^*, \tag{5.2.32}$$

$$[\Lambda, \partial] = \quad i\bar{\partial}^*. \tag{5.2.33}$$

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

Proof. Since Λ is a real operator because ω is real, each of these two identities implies the other by conjugation. We shall now verify (5.2.33). For this, we shall use the Kähler condition in an essential way. Namely, Λ being the adjoint of the multiplication with the Kähler form ω, its operation involves the Hermitian metric $h_{i\bar{j}}$, but no derivatives of it, see e.g. (5.2.31). Thus, the commutator of Λ with the first derivative operator ∂ involves at most first derivatives of the Hermitian metric. By (iii) of Theorem 5.2.1, we may assume that these first derivatives vanish at the point under consideration. Therefore, we can neglect them and compute as on Euclidean space. Thus, we only need to verify (5.2.33) on $\mathbb{C}^d$, and we proceed to do so. In fact, most of the relevant formalism has been developed already in §1.11 and §2.1; we briefly recall it here. We have the L^2-product of k-forms

$$(\alpha, \beta) = \int_{\mathbb{C}^d} \alpha \wedge *\bar{\beta}. \tag{5.2.34}$$

To see the pattern, we check that in the case $d = 1$,

$$*dz = *(dx + idy) = dy - idx = -idz,$$

and

$$*d\bar{z} = *(dx - idy) = dy + idx = id\bar{z},$$

as well as

$$*(1) = dx \wedge dy = \frac{i}{2} dz \wedge d\bar{z}. \tag{5.2.35}$$

We let ϵ_j be the exterior product with $\frac{1}{\sqrt{2}} dz^j$,

$$\epsilon_j \alpha := \frac{1}{\sqrt{2}} dz^j \wedge \alpha,$$

and similarly,

$$\epsilon_{\bar{j}} \alpha := \frac{1}{\sqrt{2}} dz^{\bar{j}} \wedge \alpha.$$

The factor $\frac{1}{\sqrt{2}}$ here is inserted because the Euclidean norm of $dz^j = dx^j + idy^j$ is $\frac{1}{\sqrt{2}}$. Thus, the L^2-adjoint ι_j of ϵ_j is given by contraction with $\frac{1}{\sqrt{2}} dz^j$, that is

$$\begin{aligned} &\iota_j (dz^{j_1} \wedge \ldots \wedge dz^{j_p} \wedge dz^{\bar{\ell}_1} \wedge \ldots \wedge dz^{\bar{\ell}_q}) \\ &= \begin{cases} 0 & \text{if } j \notin \{j_1, \ldots, j_p\}, \\ (-1)^{\mu-1} \sqrt{2}\, dz^{j_1} \wedge \ldots \wedge \widehat{dz^{j_\mu}} \wedge & \\ \qquad \ldots \wedge dz^{j_p} \wedge dz^{\bar{\ell}_1} \wedge \ldots \wedge dz^{\bar{\ell}_q} & \text{if } j = j_\mu. \end{cases} \end{aligned}$$

We check this in a simple case – the general pattern will then be clear:

$$\begin{aligned}
(\epsilon_1 dz^{\bar 1}, dz^1 \wedge dz^{\bar 1}) &= \left(\frac{1}{\sqrt{2}} dz^1 \wedge dz^{\bar 1}, dz^1 \wedge dz^{\bar 1}\right) \\
&= \sqrt{2}(dz^{\bar 1}, dz^{\bar 1}) \\
&= (dz^{\bar 1}, \iota_1(dz^1 \wedge dz^{\bar 1})).
\end{aligned}$$

Next, we either recall (1.11.22), (1.11.23) (where, however, a somewhat different notation had been employed) or check directly that

$$\epsilon_j \iota_j + \iota_j \epsilon_j = 1, \tag{5.2.36}$$

$$\epsilon_j \iota_\ell + \iota_\ell \epsilon_j = 0 \quad \text{for } j \neq \ell, \tag{5.2.37}$$

$$\epsilon_j \iota_{\bar\ell} + \iota_{\bar\ell} \epsilon_j = 0 \quad \text{for all } j, \ell. \tag{5.2.38}$$

Putting $\partial_j := \frac{\partial}{\partial z^j}$ and $\partial_{\bar j} := \frac{\partial}{\partial z^{\bar j}}$, we then have

$$\begin{gathered}
\partial = \sqrt{2} \sum_j \partial_j \epsilon_j = \sqrt{2} \sum_j \epsilon_j \partial_j, \\
\bar\partial = \sqrt{2} \sum_j \bar\partial_j \epsilon_{\bar j} = \sqrt{2} \sum_j \epsilon_{\bar j} \bar\partial_j, \\
\partial^* = -\sqrt{2} \sum_j \bar\partial_j \iota_j, \quad \bar\partial^* = -\sqrt{2} \sum_j \partial_j \iota_{\bar j}, \\
L = i \sum_j \epsilon_j \epsilon_{\bar j}, \qquad \Lambda = -i \sum_j \iota_{\bar j} \iota_j.
\end{gathered}$$

Equipped with these formulae, it is now straightforward to complete the proof:

$$\begin{aligned}
\Lambda\partial &= -i\sqrt{2} \sum_{j,\ell} \iota_{\bar\ell} \iota_\ell \partial_j \epsilon_j \\
&= -i\sqrt{2} \sum_{j,\ell} \partial_j \iota_{\bar\ell} \iota_\ell \epsilon_j \\
&= -i\sqrt{2}\Big(\sum_j \partial_j \iota_{\bar j} \iota_j \epsilon_j + \sum_{j\neq\ell} \partial_j \iota_{\bar\ell} \iota_\ell \epsilon_j\Big) \\
&= -i\sqrt{2}\Big(-\sum_j \partial_j \iota_{\bar j} \epsilon_j \iota_j + \sum_j \partial_j \iota_{\bar j} - \sum_{j\neq\ell} \partial_j \iota_{\bar\ell} \epsilon_j \iota_\ell\Big) \\
&= -i\sqrt{2}\Big(\sum_j \partial_j \epsilon_j \iota_{\bar j} \iota_j + \sum_j \partial_j \iota_{\bar j} + \sum_{j\neq\ell} \partial_j \epsilon_j \iota_{\bar\ell} \iota_\ell\Big) \\
&= -i\sqrt{2} \sum_{j,\ell} \partial_j \epsilon_j \iota_{\bar\ell} \iota_\ell - i\sqrt{2} \sum_j \partial_j \iota_{\bar j} \\
&= \partial\Lambda + i\bar\partial^*.
\end{aligned}$$

Thus, we have shown the identity on $\mathbb{C}^d$, and the Kähler condition then makes this also valid on a general Kähler manifold, as explained. □

In addition to the Laplacian

$$\Delta = dd^* + d^*d, \tag{5.2.39}$$

we can also build the operators

$$\Delta_\partial := \partial\partial^* + \partial^*\partial, \tag{5.2.40}$$

and

$$\Delta_{\bar\partial} := \bar\partial\bar\partial^* + \bar\partial^*\bar\partial. \tag{5.2.41}$$

Theorem 5.2.3. *On a Kähler manifold,*

$$\Delta = 2\Delta_\partial = 2\Delta_{\bar\partial}. \tag{5.2.42}$$

Proof. From Theorem 5.2.2,

$$\begin{aligned}\Delta_\partial &= i(\partial[\Lambda,\bar\partial] + [\Lambda,\bar\partial]\partial)\\ &= i(\partial\Lambda\bar\partial - \partial\bar\partial\Lambda + \Lambda\bar\partial\partial - \bar\partial\Lambda\partial)\\ &= i(\partial\Lambda\bar\partial + \bar\partial\partial\Lambda - \Lambda\partial\bar\partial - \bar\partial\Lambda\partial) \qquad \text{by (5.2.3)}\\ &= -i(\bar\partial[\Lambda,\partial] + [\Lambda,\partial]\bar\partial)\\ &= \Delta_{\bar\partial}.\end{aligned} \tag{5.2.43}$$

Next,

$$\partial\bar\partial^* + \bar\partial^*\partial = -i(\partial(\Lambda\partial - \partial\Lambda) + (\Lambda\partial - \partial\Lambda)\partial) = 0, \tag{5.2.44}$$

by Theorem 5.2.2 and (5.2.2).

Finally, from (5.2.44) and (5.2.1), we easily get

$$\Delta = \Delta_\partial + \Delta_{\bar\partial}. \tag{5.2.45}$$

The relations (5.2.43) and (5.2.45) yield (5.2.42). □

In §2.2, we had defined the cohomology groups $H^k(M)$ and identified them with spaces of harmonic forms, that is, solutions of

$$\Delta\eta = 0, \tag{5.2.46}$$

see Theorem 2.2.1. From Theorem 5.2.3, we infer that the operator Δ preserves the decomposition (1.8.7) which in fact is orthogonal w.r.t. the L^2-product,

$$\Omega^k(M) = \bigoplus_{p+q=k} \Omega^{p,q}(M), \tag{5.2.47}$$

that is,

$$\Delta : \Omega^{p,q}(M) \to \Omega^{p,q}(M). \tag{5.2.48}$$

If we then define $H^{p,q}(M) := H^k(M) \cap \Omega^{p,q}(M)$ $(p+q=k)$ as the space of harmonic forms of bidegree (p,q), we obtain the first part of the **Hodge decomposition theorem**, while the second part follows from the fact that Δ is a real operator and therefore, complex conjugation maps harmonic forms to harmonic forms:

Corollary 5.2.2. *For a compact Kähler manifold M,*

$$H^k(M,\mathbb{C}) = \bigoplus_{p+q=k} H^{p,q}(M,\mathbb{C}), \tag{5.2.49}$$

$$H^{p,q}(M,\mathbb{C}) = \overline{H^{q,p}(M,\mathbb{C})} \quad \textit{(complex conjugate)}. \tag{5.2.50}$$

The k-th Betti number of the compact manifold M (see Definition 2.2.1) is given by

$$b_k(M) = \dim_{\mathbb{C}} H^k(M,\mathbb{C}), \tag{5.2.51}$$

and if we put

$$h^{p,q}(M) = \dim_{\mathbb{C}} H^{p,q}(M,\mathbb{C}), \tag{5.2.52}$$

we obtain

Corollary 5.2.3. *For a compact Kähler manifold M,*

$$b_k(M) = \sum_{p+q=k} h^{p,q}(M) \tag{5.2.53}$$

and

$$h^{q,p}(M) = h^{p,q}(M), \tag{5.2.54}$$

and consequently

$$b_k(M) \text{ is even for odd } k. \tag{5.2.55}$$

We have already seen a restriction on the topology of a compact Kähler manifold in Corollary 5.2.1. (5.2.55) is a deeper such restriction.

Perspectives. Kähler geometry started with the remarkable paper of Kähler[159] that introduced the Kähler condition and derived all the basic formulae and the perspectives for the subsequent development of the subject. A thorough discussion of Kähler's paper can be found in [27] and [161].

Some references that we have used in the present section are[259], [101] and [145].

Let us briefly mention some further aspects of Kähler geometry. Metrics on Kähler manifolds satisfying

$$R_{i\bar{j}} = \mu h_{i\bar{j}} \quad \text{for some constant } \mu$$

are called *Kähler-Einstein metrics.*

Since the Ricci form represents a cohomology class $c_1(M)$, there are necessary conditions for the existence of a Kähler-Einstein metric with positive, negative or vanishing μ.

Namely, $c_1(M)$ has to be representable by a positive or negative cohomology class, or has to be cohomologous to 0, resp. For nonpositive μ, these conditions were also shown to be sufficient for the existence of a Kähler-Einstein metric on a compact M in famous work of S.T. Yau[271] (the case of negative μ was also independently solved by Aubin, see the account in [9]).

The case of positive μ is not yet completely solved. In that case, there exist obstructions for the existence of Kähler-Einstein metrics. Existence results in cases where these obstructions vanish were obtained by Tian[253], Tian and Yau[255], Siu[235], Nadel[201]. Yau, Problem 65 in [272], conjecturally related the existence of a Kähler-Einstein metric to stability properties in the sense of algebraic geometry of the underlying manifold. Tian[254] developed the appropriate stability notion and showed its necessity for the existence of a Kähler-Einstein metric. He thus disproved the conjecture that a compact Kähler manifold with positive Chern class always admits a Kähler-Einstein metric if it has no nontrivial holomorphic vector field (another condition that is known to be necessary).

As noted, every complex manifold with $\dim_{\mathbb{C}} M = 1$, i.e. every Riemann surface (see Definition 8.1.1), is Kähler since condition (i) above is trivially satisfied for any Hermitian metric. Moreover, in that case, the Kähler-Einstein metrics are simply the ones of constant curvature, and by the uniformization theorem, every Riemann surface admits such a metric since its universal cover ($\mathbb{C}, S^2$ or the hyperbolic upper half-plane $H = \{z = x + iy, y > 0\}$) does; in the latter case, the metric is $\frac{1}{y^2}(dx^2 + dy^2)$, see also §4.4. Moreover, the metric is unique up to isometries.

If one studies the space of all compact Riemann surfaces of a given topological type (Teichmüller theory), it is then convenient to investigate the space of all metrics of constant curvature on a given differentiable surface, because one can exploit additional geometric information. In a similar vein, the aforementioned results of S.T. Yau have found important applications in the classification of Kähler manifolds and algebraic varieties.

A certain class of Kähler manifolds, the so-called special Kähler manifolds (see [85]), has become important in string theory.

5.3 The Geometry of Symmetric Spaces

Besides $\mathbb{CP}^n$, we have already seen other examples of symmetric spaces:

1. $\mathbb{R}^d$, equipped with the Euclidean metric, i.e. d-dimensional Euclidean space E^d. The involution at $p \in E^d$ is the map $\sigma_p(x) = 2p - x$.

2. The sphere S^d : Since its isometry group operates transitively on S^d, it suffices to display an involution σ at the north pole $(1, 0, \ldots, 0)$; such an involution is given by
$$\sigma(x^1, \ldots, x^{d+1}) = (x^1, -x^2, \ldots, -x^{d+1})$$
in the usual coordinates.

3. Hyperbolic space H^d from §4.4. Again, the isometry group operates transitively, and it suffices to consider the point $(1, 0, \ldots, 0)$ (in the notations from §4.4), the isometry here is

$$\sigma(x^0, \ldots, x^d) = (x^0, -x^1, \ldots, -x^d).$$

In the sequel, ∇ will always denote the Levi-Civita connection.

Lemma 5.3.1. *An involution $\sigma_p : M \to M$ of a symmetric space reverses the geodesics through p. Thus, if $c : (-\varepsilon, \varepsilon) \to M$ is geodesic with $c(0) = p$ (as always parametrized proportionally to arc length), then $\sigma_p c(t) = c(-t)$.*

Proof. As an isometry, σ_p maps geodesics to geodesics. If c is a geodesic through p (with $c(0) = p$), then

$$D\sigma_p \dot{c}(0) = -\dot{c}(0).$$

The claim follows since a geodesic is uniquely determined by its initial point and initial direction (cf. Theorem 1.4.2). $\square$

Lemma 5.3.2. *Let c be a geodesic in the symmetric space M, $c(0) = p$, $c(\tau) = q$. Then*

$$\sigma_q \sigma_p(c(t)) = c(t + 2\tau) \tag{5.3.1}$$

(for all t, for which $c(t)$ and $c(t + 2\tau)$ are defined). For $v \in T_{c(t)}M$, $D\sigma_q D\sigma_p(v) \in T_{c(t+2\tau)}M$ is the vector at $c(t + 2\tau)$ obtained by parallel transport of v along c.

Proof. Let $\tilde{c}(t) := c(t + \tau)$. $\tilde{c}$ then is geodesic with $\tilde{c}(0) = q$. It follows that

$$\begin{aligned}
\sigma_q \sigma_p(c(t)) &= \sigma_q(c(-t)) \qquad \text{by Lemma 5.3.1} \\
&= \sigma_q(\tilde{c}(-t - \tau)) \\
&= \tilde{c}(t + \tau) \\
&= c(t + 2\tau).
\end{aligned}$$

Let $v \in T_pM$ and let V be the parallel vector field along c with $V(p) = v$. Since σ_p is an isometry, $D\sigma_p V$ is likewise parallel. Moreover, $D\sigma_p V(p) = -V(p)$. Hence

$$\begin{aligned}
D\sigma_p V(c(t)) &= -V(c(-t)), \\
D\sigma_q \circ D\sigma_p V(c(t)) &= \quad V(c(t + 2\tau)) \qquad \text{as before.}
\end{aligned}$$

$\square$

Corollary 5.3.1. *A symmetric space is geodesically complete, i.e. each geodesic can be indefinitely extended in both directions, i.e. may be defined on all of $\mathbb{R}$.*

Proof. (5.3.1) implies that geodesics can be indefinitely extended. One simply uses the left hand side of (5.3.1) to define the right hand side. □

The Hopf-Rinow Theorem 1.7.1 implies

Corollary 5.3.2. *In a symmetric space, any two points can be connected by a geodesic.* □

By Lemma 5.3.1, the operation of σ_p on geodesics through p is given by a reversal of the direction. Since by Corollary 5.3.2, any point can be connected with p by a geodesic, we conclude

Corollary 5.3.3. *σ_p is uniquely determined.* □

Definition 5.3.1. Let M be a symmetric space, $c : \mathbb{R} \to M$ a geodesic. The *translation along c by the amount $t \in \mathbb{R}$* is

$$\tau_t := \sigma_{c(t/2)} \circ \sigma_{c(0)}.$$

By Lemma 5.3.2, τ_t thus maps $c(s)$ onto $c(s+t)$, and $D\tau_t$ is parallel transport along c from $c(s)$ to $c(s+t)$.

Remark. τ_t is an isometry defined on all of M. $\tau = \tau_t$ maps the geodesic c onto itself. The operation of τ on geodesics other than c in general is quite different, and in fact τ need not map any other geodesic onto itself. One may see this for $M = S^n$.

Convention: For the rest of this paragraph, M will be a symmetric space. G denotes the isometry group of M. G_0 is the following subset of G :

$$G_0 := \{g_t \text{ for } t \in \mathbb{R}, \text{where } s \mapsto g_s \text{ is a group homomorphism from } \mathbb{R} \text{ to } G\},$$

i.e. the union of all one-parameter subgroups of G. (It may be shown that G_0 is a subgroup of G.)

Examples of such one-parameter subgroups are given by the families of translations $(\tau_t)_{t\in\mathbb{R}}$ along geodesic lines.

Theorem 5.3.1. *G_0 operates transitively on M.*

Proof. By Corollary 5.3.2, any two points $p, q \in M$ can be connected by a geodesic c; let $p = c(0), q = c(s)$. If $(\tau_t)_{t\in\mathbb{R}}$ is the family of translations along c, then

$$q = \tau_s(p).$$

We thus have found an isometry from G_0 that maps p to q. □

Definition 5.3.2. A Riemannian manifold with a transitive group of isometries is called *homogeneous*.

Theorem 5.3.2. *The curvature tensor R of M is parallel, $\nabla R \equiv 0$.*

Proof. Let c be a geodesic, and let X, Y, Z, W be parallel vector fields along c, $p = c(t_0), q = c(t_0 + t)$. Then $q = \tau_t(p)$ and by Lemma 5.3.2

$$\begin{aligned}\langle R(X(q), Y(q))Z(q), W(q)\rangle &= \langle R(d\tau_t X(p), d\tau_t Y(p))d\tau_t Z(p), d\tau_t W(p)\rangle \\ &= \langle R(X(p), Y(p))Z(p), W(p)\rangle \quad \text{since } \tau_t \text{ is an isometry.}\end{aligned}$$

Let now $v := \dot{c}(t_0)$. The preceding relation gives

$$v\langle R(X, Y)Z, W\rangle = 0,$$

and since X, Y, Z, W are parallel,

$$\langle(\nabla_v R)(X, Y)Z, W\rangle = 0.$$

Since $\nabla_v R$ like R is a tensor, $(\nabla_v R)(X, Y)Z$ depends only on the values of X, Y, Z at p. Since this holds for all c, X, Y, Z, W we get $\nabla R \equiv 0$. □

Definition 5.3.3. A complete Riemannian manifold with $\nabla R \equiv 0$ is called *locally symmetric*.

Remark. One can show that for each locally symmetric space N there exist a simply connected symmetric space M and a group Γ operating on M discretely, without fixed points, and isometrically, such that

$$N = M/\Gamma. \tag{5.3.2}$$

Conversely, it is clear that such a space is locally symmetric. Examples are given by compact Riemann surfaces of genus $g \geq 2$ which may be realized as quotients of the hyperbolic plane H^2.
Let us also introduce different examples, the so called *lens spaces*:
We consider S^3 as unit sphere in $\mathbb{C}^2$:

$$S^3 = \{(z^1, z^2) \in \mathbb{C}^2 : |z^1|^2 + |z^2|^2 = 1\}.$$

On S^3, we then have an isometric action of the torus $S^1 \times S^1$, namely

$$(z^1, z^2) \mapsto (e^{i\varphi^1} z^1, e^{i\varphi^2} z^2) \quad \text{for } 0 \leq \varphi^1, \varphi^2 \leq 2\pi.$$

Let now $p, q \in \mathbb{N}$ be relatively prime with $1 \leq p < q$.

Let $\mathbb{Z}_q$ be the cyclic group of order q. We then obtain a homomorphism

$$\begin{aligned}\mathbb{Z}_q &\to S^1 \times S^1, \\ r &\mapsto (e^{2\pi i r/q}, e^{2\pi i p r/q}).\end{aligned}$$

Thus, $\mathbb{Z}_q$ operates isometrically on S^3. Since p and q are relatively prime, this operation has no fixed points, and the *lens space*

$$L(q,p) := S^3/\mathbb{Z}_q$$

is a manifold.

Actually, $L(2,1)$ is not only locally symmetric, but symmetric. More precisely, $L(2,1)$ is the three dimensional real projective space.

For $q > 2$, however, the lens spaces are not symmetric. For example, the involution at $p = (1,0) \in S^3$ is given (in our complex notation) by

$$\sigma_p(z^1, z^2) = (z^{\bar{1}}, -z^2)$$

(recall the definition of S^d at the beginning of this paragraph). σ_p therefore does not commute with the $\mathbb{Z}_q$ action. Therefore, the involution σ_p does not carry over to $L(q,p)$. Since on the other hand each involution is already determined by its operation on the tangent space and since an involution would have to operate in the same way as σ_p on the tangent space of the point corresponding to p in the lens space, the lens space cannot possess any such involution and hence cannot be symmetric.

We now want to determine the Jacobi fields on (locally) symmetric spaces. For a Riemannian manifold N, $p \in N, v \in T_pN$ we define an operator

$$R_v : T_pN \to T_pN$$

by

$$R_v(w) = R(w,v)v. \tag{5.3.3}$$

For a geodesic c, $R_{\dot{c}(t)}$ maps the orthogonal complement of $\dot{c}(t)$ in $T_{c(t)}N$ onto itself.

The operator $R_{\dot{c}(t)}$ is self-adjoint. This follows from (3.3.10) and (3.3.9) or (3.3.7):

$$\langle R_v(w), w'\rangle = \langle R(w,v)v, w'\rangle = \langle R(w',v)v, w\rangle = \langle R_v(w'), w\rangle.$$

Since R is parallel for a locally symmetric space, $R_{\dot{c}(t)}$ commutes with parallel transport along c.

Let v be an eigenvector of $R_{\dot{c}(0)}$ with eigenvalue ρ with $\|v\| = 1, \|\dot{c}(0)\| = 1$ (this can be achieved by reparametrization), and

$$\langle v, \dot{c}(0)\rangle = 0,$$

i.e.

$$R(v, \dot{c}(0))\dot{c}(0) = R_{\dot{c}(0)}(v) = \rho v.$$

Let $v(t)$ be the vector field obtained by parallel transport of v along c. Then $v(t)$ is an eigenvector of $R_{\dot{c}(t)}$ with eigenvalue ρ, since R is parallel. Thus

$$R(v(t), \dot{c}(t))\dot{c}(t) = \rho v(t). \tag{5.3.4}$$

(5.3.4) implies that the vector fields

$$\begin{aligned} J_1(t) &:= c_\rho(t)v(t), \\ J_2(t) &:= s_\rho(t)v(t), \end{aligned} \tag{5.3.5}$$

(c_ρ and s_ρ defined as in §4.5) satisfy the Jacobi equation:

$$\ddot{J}_i(t) + R\left(J_i(t), \dot{c}(t)\right)\dot{c}(t) = 0, \quad \text{for } i = 1, 2. \tag{5.3.6}$$

Thus

Theorem 5.3.3. *Let N be a locally symmetric space, c geodesic in N, $c(0) =: p$, $v_1, \ldots, v_{n-1}$ an orthonormal basis of eigenvectors of $R_{\dot{c}(0)}$ orthogonal to $\dot{c}(0)$ with eigenvalues $\rho_1, \ldots, \rho_{n-1}$, $v_1(t), \ldots, v_{n-1}(t)$ the parallel vector fields along c with $v_j(0) = v_j \quad (j = 1, \ldots, n-1)$. The Jacobi fields along c (orthogonal to $\dot{c}$) then are linear combinations of Jacobi fields of the form*

$$c_{\rho_j}(t)v_j(t) \quad \text{and} \quad s_{\rho_j}(t)v_j(t). \tag{5.3.7}$$

□

Definition 5.3.4. Let $\mathfrak{g}$ be the Lie algebra of Killing fields (cf. Lemma 1.9.8) on the symmetric space M, and let $p \in M$. We put

$$\begin{aligned} \mathfrak{k} &:= \{X \in \mathfrak{g} : X(p) = 0\}, \\ \mathfrak{p} &:= \{X \in \mathfrak{g} : \nabla X(p) = 0\}. \end{aligned}$$

Theorem 5.3.4.

$$\begin{aligned} \mathfrak{k} \oplus \mathfrak{p} &= \mathfrak{g}, \\ \mathfrak{k} \cap \mathfrak{p} &= \{0\}. \end{aligned}$$

Proof. $\mathfrak{k} \cap \mathfrak{p} = \{0\}$ follows from the facts that each Killing field is a Jacobi field (Corollary 4.2.1) (along any geodesic) and that Jacobi fields that vanish at some point together with their derivative vanish identically (by Lemma 4.2.3) and finally that by Corollary 5.3.2 any two points can be connected by a geodesic. Let now $X \in \mathfrak{g}$ with $X(p) \neq 0$. Let $c(t) := \exp_p tX(p)$ be the geodesic with $\dot{c}(0) = X(p)$, and let τ_t be the group of translations along c (Definition 5.3.1). Then

$$Y(q) := \frac{d}{dt}\tau_t(q)_{|t=0} \tag{5.3.8}$$

is a Killing field, since the τ_t are isometries (Lemma 1.9.7).

We have

$$Y(p) = X(p). \tag{5.3.9}$$

For $v \in T_pM$, let $\gamma(s)$ be a curve with $\gamma'(0) = v$. Then

$$\begin{aligned} \nabla_v Y(p) &= \nabla_{\frac{\partial}{\partial s}} \frac{\partial}{\partial t} \tau_t(\gamma(s))_{|s=t=0} \\ &= \nabla_{\frac{\partial}{\partial t}} \frac{\partial}{\partial s} \tau_t(\gamma(s))_{|s=t=0} \\ &= \nabla_{\frac{\partial}{\partial t}} D\tau_t(v)_{|t=0} \\ &= 0, \end{aligned} \tag{5.3.10}$$

since by Lemma 5.3.2 $D\tau_t$ is parallel transport along c, and hence $D\tau_t(v)$ is a parallel vector field along c.

We conclude

$$X = (X - Y) + Y,$$

with $(X - Y) \in \mathfrak{k}$ by (5.3.9), and with $Y \in \mathfrak{p}$ by (5.3.10). □

Theorem 5.3.5. *As a vector space, $\mathfrak{p}$ is isomorphic to T_pM. The one-parameter subgroup of isometries generated by $Y \in \mathfrak{p}$ is the group of translations along the geodesic $\exp_p tY(p)$.*

Proof. Let $w \in T_pM$. Let $c(t) := \exp_p tw$ be the geodesic with $\dot{c}(0) = w$. Let τ_t be the group of translations along c. As in (5.3.8), we put

$$Y(q) := \frac{d}{dt} \tau_t(q)_{|t=0} \quad \text{for all } q \in M. \tag{5.3.11}$$

As in the proof of Theorem 5.3.4, we obtain

$$Y(p) = w \text{ and } Y \in \mathfrak{p}.$$

This induces a linear map from T_pM to $\mathfrak{p}$. The inverse of this map is simply the restriction mapping $Y \in \mathfrak{p}$ to $Y(p)$. Thus, we have found a bijective linear map between T_pM and $\mathfrak{p}$. By (5.3.11), Y also generates the one-parameter-subgroup τ_t (surjectivity follows from the proof of Theorem 5.3.4). □

Let us introduce the following notation: For a Killing field we denote the (at this point only local) 1-parameter group of isometries generated by X by e^{tX} (instead of the previous notation ψ_t or φ_t).

Lemma 5.3.3. *Let X be a Killing field on the symmetric space M. Then e^{tX} is defined for all $t \in \mathbb{R}$. Thus, $(e^{tX})_{t \in \mathbb{R}}$ is a 1-parameter-group of isometries.*

Proof. Let $q \in M$. We want to show that $e^{tX}(q)$ is defined for all $t \in \mathbb{R}$. We shall show that this is true for $t > 0$, since the case $t < 0$ is analogous. Let now

$$T := \sup\{t \in \mathbb{R} : e^{\tau X}(q) \text{ is defined for all } \tau \leq t\}.$$

We assume $T < \infty$ and want to reach a contradiction.

We put
$$m := \sup\{d(q, e^{tX}(q)) : t \le T/2\}.$$
Since each $g \in G$ is an isometry, we have for all $x, y \in M$
$$d(gx, gy) = d(x, y),$$
hence also
$$d(g^2 q, gq) = d(gq, q),$$
and thus
$$d(g^2 q, q) \le 2d(gq, q).$$
Therefore for $0 \le t < T$,
$$d(e^{tX}(q), q) \le 2d(e^{t/2\,X}(q), q) \le 2m.$$
Therefore, for all $0 \le t < T$, $e^{tX}(q)$ is contained in
$$B(q, 2m),$$
which is a compact set.

As in the proof of Corollary 1.4.3, we see that there exists $\varepsilon > 0$ with the property that for all $x \in B(q, 2m)$ $e^{tX}(x)$ is defined for $|t| \le \varepsilon$. Thus, for $\tau := T - \varepsilon/2$,
$$e^{\varepsilon X}(e^{\tau X}(q)) = e^{(T+\varepsilon/2)X}(q)$$
is defined.

This contradicts the assumption on T and proves the claim. □

For $Y \in \mathfrak{p}$, we thus obtain from Theorem 5.3.5
$$e^{tY} = \tau_t, \tag{5.3.12}$$
where (τ_t) is the family of translations along the geodesic $\exp_p tY(p)$.

We now define a group homomorphism
$$s_p : G \to G$$
by
$$s_p(g) = \sigma_p \circ g \circ \sigma_p, \tag{5.3.13}$$
where $\sigma_p : M \to M$ is the involution at p. Since $\sigma_p^2 = \mathrm{id}$, we have
$$s_p(g) = \sigma_p \circ g \circ \sigma_p^{-1}. \tag{5.3.14}$$

We obtain a map

$$\theta_p : \mathfrak{g} \to \mathfrak{g}$$

by

$$\theta_p(X) := \frac{d}{dt} s_p(e^{tX})_{|t=0}. \tag{5.3.15}$$

Theorem 5.3.6.

$$\begin{aligned} \theta_{p|\mathfrak{k}} &= \ \ \mathrm{id}, \\ \theta_{p|\mathfrak{p}} &= -\mathrm{id}. \end{aligned}$$

Proof. Let $X \in \mathfrak{k}$, i.e. $X(p) = 0$. Then for all t,

$$e^{tX}(p) = p. \tag{5.3.16}$$

Let c_1 be a geodesic with $c_1(0) = p$. Then for all t,

$$c_2(s) := e^{tX} c_1(s)$$

likewise defines a geodesic through p, i.e. $c_2(0) = p$. It follows that

$$\begin{aligned} s_p(e^{tX})c_1(s) &= \sigma_p \circ e^{tX} \circ \sigma_p c_1(s) \\ &= \sigma_p \circ e^{tX} c_1(-s) \qquad \text{by Lemma 5.3.1} \\ &= \sigma_p c_2(-s) \\ &= c_2(s), \end{aligned}$$

i.e.

$$s_p(e^{tX})c_1(s) = e^{tX} c_1(s).$$

Since each $q \in M$ can be connected with p by a geodesic (Corollary 5.3.2), we obtain

$$s_p(e^{tX})(q) = e^{tX}(q)$$

for all $q \in M$, i.e. $s_p(e^{tX}) = e^{tX}$, and hence also

$$\theta_p(X) = X,$$

i.e.

$$\theta_{p|\mathfrak{k}} = \mathrm{id}.^2$$

Let now $Y \in \mathfrak{p}$. From (5.3.12) (cf. Theorem 5.3.5),

$$e^{tY} = \tau_t = \sigma_{c(t/2)} \circ \sigma_p \quad \text{by Definition 5.3.1},$$

[2] One may easily modify the proof at this place so as to avoid using the completeness of M.

where $c(t) = \exp_p tY(p)$. Hence

$$\begin{aligned} s_p(e^{tY}) &= \sigma_p \circ \sigma_{c(t/2)} \circ \sigma_p \circ \sigma_p \\ &= \sigma_p \circ \sigma_{c(t/2)} \qquad \text{because of } \sigma_p^2 = \mathrm{id} \\ &= \tau_{-t}, \end{aligned}$$

which may be seen e.g. as follows: Let $q = c(t/2), \tilde{c}(s) = c(t/2 - s)$. Then $p = \tilde{c}(t/2), \tilde{c}(0) = q$, hence

$$\sigma_p \circ \sigma_c(t/2) = \sigma_{\tilde{c}}(t/2) \circ \sigma_{\tilde{c}(0)}.$$

Therefore, this is the translation along $\tilde{c}$ by the amount t. Since $\tilde{c}$ is traversed in opposite direction as c, this is the same as translation along c by the amount $-t$.

Since

$$\tau_{-t} = e^{-tY},$$

it follows that,

$$s_p(e^{tY}) = e^{-tY},$$

hence

$$\theta_p(Y) = -Y,$$

i.e.

$$\theta_{p|\mathfrak{p}} = -\mathrm{id}.$$

□

Lemma 5.3.4. *$\theta_p[X, Y] = [\theta_p X, \theta_p Y]$ for all $X, Y \in \mathfrak{g}$. Thus, θ_p is a Lie algebra homomorphism.*

Proof. By definition of θ_p (5.3.15), $\theta_p(X)$ generates the 1-parameter group $e^{t\theta_p(X)}$, i.e.

$$s_p(e^{tX}) = e^{t\theta_p(X)}. \tag{5.3.17}$$

Now

$$\begin{aligned} [X, Y] &= \frac{d}{dt} De^{-tX} \circ Y \circ e^{tX}|_{t=0} \quad \text{cf. Theorem 1.9.4 (ii)} \\ &= \frac{\partial^2}{\partial t \partial s} e^{-tX} e^{sY} e^{tX}|_{t=s=0}. \end{aligned} \tag{5.3.18}$$

Hence

$$\begin{aligned} \theta_p[X, Y] &= \frac{\partial^2}{\partial t \partial s} \sigma_p e^{-tX} e^{sY} e^{tX} \sigma_{p|t=s=0} \\ &= \frac{\partial^2}{\partial t \partial s} \sigma_p e^{-tX} \sigma_p^{-1} \sigma_p e^{sY} \sigma_p^{-1} \sigma_p e^{tX} \sigma_{p|t=s=0} \\ &= \frac{\partial^2}{\partial t \partial s} s_p(e^{-tX}) s_p(e^{sY}) s_p(e^{tX})_{|t=s=0} && \text{cf. (5.3.14)} \\ &= \frac{\partial^2}{\partial t \partial s} e^{-t\theta_p(X)} e^{s\theta_p(Y)} e^{t\theta_p(X)}|_{t=s=0} && \text{by (5.3.17)} \\ &= [\theta_p(X), \theta_p(Y)] && \text{by (5.3.18).} \end{aligned}$$

□

Theorem 5.3.7.

$$[\mathfrak{k}, \mathfrak{k}] \subset \mathfrak{k},$$
$$[\mathfrak{p}, \mathfrak{p}] \subset \mathfrak{k},$$
$$[\mathfrak{k}, \mathfrak{p}] \subset \mathfrak{p}.$$

Proof. Because of $\theta_p^2 = \mathrm{id}$, θ_p has eigenvalues -1 and 1. By Theorem 5.3.6, $\mathfrak{k}$ is the eigenspace with eigenvalue 1, $\mathfrak{p}$ the eigenspace with eigenvalue -1 (note that by Theorem 5.3.5, $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$). If X is an eigenvector with eigenvalue λ, Y one with eigenvalue μ, then, since θ_p is a Lie algebra homomorphism (Lemma 5.3.3), $[X, Y]$ is an eigenvector with eigenvalue $\lambda\mu$. This easily gives the claim. □

Corollary 5.3.4. *$\mathfrak{k}$ is a Lie subalgebra of $\mathfrak{g}$.*

Proof. $\mathfrak{k}$ is a subspace of $\mathfrak{g}$ and closed w.r.t. the Lie bracket by Theorem 5.3.7. □

Corollary 5.3.5. *With the identification*

$$T_pM \simeq \mathfrak{p}$$

from Theorem 5.3.5, the curvature tensor of M satisfies

$$R(X, Y)Z(p) = -[[X, Y], Z](p) \tag{5.3.19}$$

for $X, Y, Z \in \mathfrak{p}$.

Proof. Let $X \in \mathfrak{g}, Y \in \mathfrak{p}$. The geodesic $\exp_p tY(p)$ satisfies

$$Y(c(t)) = \dot{c}(t) \quad \text{for all } t \in \mathbb{R}.$$

This follows e.g. from Theorem 5.3.5.

Since by Corollary 4.2.1, X is a Jacobi field along c, we obtain

$$\nabla_Y \nabla_Y X + R(X, Y)Y = 0 \tag{5.3.20}$$

along c, hence in particular at p.

This implies that we have also for $Y, Z \in \mathfrak{p}$, since then also $Y + Z \in \mathfrak{p}$, that

$$\nabla_Y \nabla_Z X + \nabla_Z \nabla_Y X + R(X, Y)Z + R(X, Z)Y = 0 \tag{5.3.21}$$

at p.

Now by (3.3.7),

$$R(X,Z)Y = -R(Z,X)Y, \tag{5.3.22}$$

by (3.3.8),

$$R(X,Y)Z + R(Y,Z)X + R(Z,X)Y = 0, \tag{5.3.23}$$

and by (3.3.3)

$$R(Y,Z)X = \nabla_Y \nabla_Z X - \nabla_Z \nabla_Y X - \nabla_{[Y,Z]} X. \tag{5.3.24}$$

By Theorem 5.3.7, for $Y, Z \in \mathfrak{p}, [Y,Z] \in \mathfrak{k}$, hence

$$[Y,Z](p) = 0. \tag{5.3.25}$$

(5.3.21) – (5.3.25) imply

$$\nabla_Y \nabla_Z X + R(X,Y)Z = 0 \tag{5.3.26}$$

at p.

By (5.3.23) and (5.3.22) for $X, Y, Z \in \mathfrak{p}$,

$$\begin{aligned} R(X,Y)Z(p) &= -R(Y,Z)X(p) + R(X,Z)Y(p) \\ &= \nabla_Z \nabla_X Y(p) - \nabla_Z \nabla_Y X(p) \qquad \text{by (5.3.26)} \\ &= \nabla_Z [X,Y](p) \\ &= \nabla_{[X,Y]} Z(p) - [[X,Y],Z](p) \\ &= -[[X,Y],Z](p), \end{aligned}$$

because of $[X,Y](p) = 0$ (Theorem 5.3.7). □

Corollary 5.3.6. *The sectional curvature of the plane in T_pM spanned by the orthonormal vectors $Y_1(p), Y_2(p)$ $(Y_1, Y_2 \in \mathfrak{p})$ satisfies*

$$K(Y_1(p) \wedge Y_2(p)) = -\langle [[Y_1, Y_2], Y_2], Y_1 \rangle (p).$$

Proof. From (5.3.19). □

5.4 Some Results about the Structure of Symmetric Spaces

In this paragraph, we shall employ the conventions established in the previous one.

Let us first quote the following special case of a theorem of Myers and Steenrod:

Theorem 5.4.1. *The isometry group of a symmetric space M is a Lie group, and so is the group G_0 defined in §5.3. Moreover $\mathfrak{g}$ is the Lie algebra of both G and G_0.*

A proof may May be found, e.g., in [123]. Technically, this result will not be indispensable for the sequel, but it is useful in order to gain a deeper understanding of symmetric spaces.

We now start with some constructions that are valid not only for the isometry group of a symmetric space but more generally for an arbitrary Lie group G with Lie algebra denoted by $\mathfrak{g}$.

Each $h \in G$ defines an inner automorphism of G by conjugation:

$$\begin{aligned} \mathrm{Int}\,(h) : G &\to G, \\ g &\mapsto hgh^{-1}. \end{aligned}$$

Putting $h = \sigma_p$, here we obtain s_p from §5.3.

$\mathfrak{g}$ as a Lie algebra in particular is a vector space, and we denote the group of vector space automorphisms of $\mathfrak{g}$ by $\mathrm{Gl}(\mathfrak{g})$.

Definition 5.4.1. The *adjoint representation* of G is given by

$$\begin{aligned} \mathrm{Ad}\, : G &\to \mathrm{Gl}(\mathfrak{g}), \\ h &\mapsto D_e \,\mathrm{Int}\,(h) \end{aligned}$$

where $e \in G$ is the identity element.

In the notations of §5.3 we thus have

$$\theta_p = \mathrm{Ad}\,(\sigma_p). \tag{5.4.1}$$

Lemma 5.4.1. Ad *is a group homomorphism, and for each $h \in G$, $\mathrm{Ad}\ h \in \mathrm{Gl}(\mathfrak{g})$ is a Lie algebra homomorphism, i.e.*

$$\mathrm{Ad}\ h[X,Y] = [\mathrm{Ad}\ hX,\ \mathrm{Ad}\,, hY] \quad \textit{for all } X, Y \in \mathfrak{g}. \tag{5.4.2}$$

This result generalizes Lemma 5.3.4. *Proof.* That Ad is a group homomorphism follows from

$$\mathrm{Int}\,(h_1 h_2) = \mathrm{Int}\,(h_1)\mathrm{Int}\,(h_2).$$

That $\mathrm{Ad}\ h$ is a Lie algebra homomorphism follows as in the proof of Lemma 5.3.4. □

Definition 5.4.2. The *adjoint representation* of $\mathfrak{g}$ is given by

$$\begin{aligned} \mathrm{ad}\, : \mathfrak{g} &\to \mathfrak{gl}(\mathfrak{g}), \\ X &\mapsto (D_e \,\mathrm{Ad}\,)(X), \end{aligned}$$

where $\mathfrak{gl}(\mathfrak{g})$ is the space of linear self maps of $\mathfrak{g}$.

Lemma 5.4.2.
$$(\mathrm{ad}\ X)Y = -[X, Y]. \tag{5.4.3}$$

Proof.
$$\begin{aligned}(\mathrm{ad}\ X)Y &= \frac{d}{dt} D_e \operatorname{Int}(e^{tX}) Y_{|t=0} \\ &= \frac{\partial^2}{\partial t \partial s} \operatorname{Int}(e^{tX}) e^{sY}|_{t=s=0} \\ &= [-X, Y] \qquad \text{by Theorem 1.9.4 (ii).}\end{aligned}$$
□

Corollary 5.4.1.
$$(\mathrm{ad}\ X)[Y, Z] = [(\mathrm{ad}\ X)Y, Z] + [Y, (\mathrm{ad}\ X)Z].$$

Proof. From Lemma 5.4.2 and the Jacobi identity (Lemma 1.9.5). □

Corollary 5.4.2.
$$e^{\mathrm{ad}\ X} = \mathrm{Ad}\ e^X \ \textit{for all } X \in \mathfrak{g}.$$

Proof.
$$\begin{aligned}\frac{d}{dt} e^{\mathrm{ad}\ tX}|_{t=0} &= \mathrm{ad}\ X \\ &= (D_e \operatorname{Ad})X \\ &= \frac{d}{dt} \operatorname{Ad}\ e^{tX}|_{t=0},\end{aligned}$$
which easily implies the claim. □

Definition 5.4.3. The *Killing form* of $\mathfrak{g}$ is the bilinear form
$$\begin{aligned} B : \mathfrak{g} \times \mathfrak{g} &\to \mathbb{R}, \\ (X, Y) &\mapsto \operatorname{tr}(\mathrm{ad}\ X \circ \mathrm{ad}\ Y). \end{aligned}$$
$\mathfrak{g}$ (and likewise G) is called *semisimple* if the Killing form of $\mathfrak{g}$ is nondegenerate.

Lemma 5.4.3. *The Killing form B of $\mathfrak{g}$ is symmetric. B is invariant under automorphisms of $\mathfrak{g}$. In particular*
$$B((\mathrm{Ad}\, g)X, (\mathrm{Ad}\ g)Y) = B(X, Y) \quad \textit{for all } X, Y \in \mathfrak{g}, g \in G. \tag{5.4.4}$$
Moreover
$$B((\mathrm{ad}\ X)Y, Z) + B(Y, (\mathrm{ad}\ X)Z) = 0 \quad \textit{for all } X, Y, Z \in \mathfrak{g}. \tag{5.4.5}$$

Proof. The symmetry of B is a direct consequence of the formula

$$\mathrm{tr}\,(AC) = \mathrm{tr}\,(CA) \tag{5.4.6}$$

for linear self maps of a vector space.

Let now σ be an automorphism of $\mathfrak{g}$. Then

$$\begin{aligned}(\mathrm{ad}\ \sigma X)(Y) &= [\sigma(-X), Y] \qquad \text{by (5.4.3)}\\ &= [\sigma(-X), \sigma\sigma^{-1}Y]\\ &= \sigma[-X, \sigma^{-1}Y]\\ &= (\sigma \circ \mathrm{ad}\ X \circ \sigma^{-1})(Y).\end{aligned}$$

Therefore

$$\mathrm{tr}\,(\mathrm{ad}\ \sigma X\ \mathrm{ad}\ \sigma Y) = \mathrm{tr}\,(\sigma\,\mathrm{ad}\ X\,\mathrm{ad}\ Y\sigma^{-1}) = \mathrm{tr}\,(\mathrm{ad}\ X\,\mathrm{ad}\ Y)$$

with (5.4.6), i.e.

$$B(\sigma X, \sigma Y) = B(X, Y). \tag{5.4.7}$$

Therefore, B is invariant under automorphisms of $\mathfrak{g}$. We now choose

$$\sigma = \mathrm{Ad}\,(e^{tX}).$$

Differentiating (5.4.7) w.r.t. t at $t = 0$ yields (5.4.5). □

We also define

$$K := \{g \in G : g(p) = p\}.$$

K then is a subgroup of G. For $X \in \mathfrak{k}$, we have $e^{tX} \in K$.

We now have two scalar valued products on $\mathfrak{p}$. Namely, for $Y, Z \in \mathfrak{p}$, we may form $\langle Y(p), Z(p)\rangle$, where $\langle ., .\rangle$ denotes the Riemannian metric of M, as well as

$$B(Y, Z).$$

We now want to compare these two products.

Lemma 5.4.4. *$\mathrm{Ad}\,K$ leaves $\mathfrak{p}$ and the product $\langle\cdot,\cdot\rangle$ on $\mathfrak{p}$ invariant.*

Proof. Since for $k \in K, k(p) = p$, for $Y \in \mathfrak{p}$, $\mathrm{Int}\,(k)$ maps the geodesic $\exp_p tY(p)$ through p onto another geodesic through p, and this geodesic is generated by $Dk \circ Y(k^{-1}(p)) = DkY(p)$.

Therefore, $(\mathrm{Ad}\ k)(Y) = Dk \circ Y(k^{-1})$ is in $\mathfrak{p}$ as well (cf. the proof of Theorem 5.3.6). Moreover, for $Y, Z \in \mathfrak{p}$,

$$\begin{aligned}\langle Y(p), Z(p)\rangle &= \langle Dk \circ Y(p), Dk \circ Z(p)\rangle && \text{since } k \text{ is an isometry}\\ &= \langle Dk \circ Y(k^{-1}(p)), Dk \circ Z(k^{-1}(p))\rangle && \text{since } k^{-1}(p) = p\\ &= \langle \mathrm{Ad}\ kY(p), \mathrm{Ad}\ kZ(p)\rangle.\end{aligned}$$

□

Corollary 5.4.3. *The Killing form B is negative definite on $\mathfrak{k}$.*

Proof. Let $X \in \mathfrak{k}, Y, Z \in \mathfrak{p}$. By Lemma 5.4.4

$$\langle \mathrm{Ad}\,(e^{tX})Y(p),\, \mathrm{Ad}\,(e^{tX})Z(p)\rangle = \langle Y(p), Z(p)\rangle. \tag{5.4.8}$$

We differentiate (5.4.8) at $t = 0$ w.r.t. t and obtain

$$\langle \mathrm{ad}\,(X)Y(p), Z(p)\rangle + \langle Y(p),\, \mathrm{ad}\,(X)Z(p)\rangle = 0. \tag{5.4.9}$$

By Theorem 5.3.7 or Lemma 5.4.4, $\mathrm{ad}\, X$ yields a linear self map of $\mathfrak{p}$, and by (5.4.9), this map is skew symmetric w.r.t. the scalar products $\langle \cdot, \cdot \rangle(p)$ on $\mathfrak{p}$. We choose an orthonormal basis of $\mathfrak{p}$ w.r.t. $\langle \cdot, \cdot \rangle(p)$ and write ad $X = (a_{ij})_{i,j=1,\dots,n}$ as a matrix w.r.t. this basis. Since ad X is skew symmetric, we have

$$a_{ij} = -a_{ji} \quad \text{for}\, i, j = 1, \dots, n.$$

Therefore

$$B(X, X) = \mathrm{tr}\ \mathrm{ad}\ X \circ \mathrm{ad}\ X = -\sum_{i,j=1}^{n} a_{ij}^2,$$

and negative definiteness follows, since for $X \in \mathfrak{k}, X \neq 0$, also ad $X \neq 0$ because otherwise $\mathrm{Ad}\, e^{tX} = \mathrm{id}$, hence by $e^{tX} \in K, De^{tX}$ would be the identity of T_pM, i.e. e^{tX}, i.e. $X = 0$. □

We now define the following scalar product on $\mathfrak{g}$:

for $Y, Z \in \mathfrak{p}$, $\quad \langle Y, Z\rangle_{\mathfrak{g}} := \langle Y(p), Z(p)\rangle$,

where the scalar product on the right hand side is the Riemannian metric on T_pM;

for $X, W \in \mathfrak{k}$, $\quad \langle X, W\rangle_{\mathfrak{g}} := -B(X, W)$,
for $X \in \mathfrak{k}, Y \in \mathfrak{p}$, $\quad \langle X, Y\rangle_{\mathfrak{g}} := 0$.

Lemma 5.4.5. $\langle \cdot, \cdot \rangle_{\mathfrak{g}}$ *is positive definite and* $\mathrm{Ad}\, K$*-invariant.*

Proof. Positive definiteness follows from positive definiteness of the Riemannian metric on T_pM and Corollary 5.4.3. $\mathrm{Ad}\, K$-invariance follows from Lemmas 5.4.3 and 5.4.4. □

The infinitesimal version of the $\mathrm{Ad}\, K$-invariance of $\langle \cdot, \cdot \rangle_{\mathfrak{g}}$ is

$$\langle (\mathrm{ad}\ X)Y, Z\rangle_{\mathfrak{g}} + \langle Y, (\mathrm{ad}\ X)Z\rangle_{\mathfrak{g}} = 0 \quad \text{for}\, Y, Z \in \mathfrak{g}, X \in \mathfrak{k}. \tag{5.4.10}$$

For $Y \in \mathfrak{p}$, we now consider the linear functional

$$\begin{aligned} &\mathfrak{p} \to \mathbb{R}, \\ &X \mapsto B(X, Y), \end{aligned}$$

where B again denotes the Killing form of $\mathfrak{g}$. Then there exists $Y^* \in \mathfrak{p}$ with

$$B(X,Y) = \langle X, Y^* \rangle_{\mathfrak{g}}.$$

Since B is symmetric (Lemma 5.4.3), the map

$$\begin{aligned} \mathfrak{p} &\to \mathfrak{p}, \\ Y &\mapsto Y^*, \end{aligned}$$

is self adjoint w.r.t. $\langle \cdot, \cdot \rangle$. Therefore, there exists an orthonormal basis $Y_1, \dots, Y_n$ of eigenvectors:

$$Y_j^* = \lambda_j Y_j \quad (j = 1, \dots, n).$$

Then

$$\begin{aligned} B(Y_i, Y_j) &= \langle Y_i, Y_j^* \rangle = \lambda_j \langle Y_i, Y_j \rangle \\ &= \langle Y_j, Y_i^* \rangle = \lambda_i \langle Y_i, Y_j \rangle. \end{aligned}$$

Thus, eigenspaces of different eigenvalues are orthogonal not only w.r.t. $\langle \cdot, \cdot \rangle_{\mathfrak{g}}$, but also w.r.t. B. We write the decomposition of $\mathfrak{p}$ into eigenspaces as

$$\mathfrak{p} = \mathfrak{p}_1 \oplus \ldots \oplus \mathfrak{p}_m.$$

The eigenvalue of $\mathfrak{p}_j$ is denoted by μ_j $(j = 1, \dots m)$.

Lemma 5.4.6.

$$[\mathfrak{p}_i, \mathfrak{p}_j] = 0 \quad \textit{for } i \neq j. \tag{5.4.11}$$

If $\mathfrak{g}$ is semisimple, i.e. B nondegenerate, then

$$\langle \cdot, \cdot \rangle_{\mathfrak{g}} = -B_{|\mathfrak{k}} + \frac{1}{\mu_1} B_{|\mathfrak{p}_1} + \ldots + \frac{1}{\mu_m} B_{|\mathfrak{p}_m}. \tag{5.4.12}$$

Proof. Let $Y_i \in \mathfrak{p}_i, Y_j \in \mathfrak{p}_j$. Then

$$\begin{aligned} B([Y_i, Y_j], [Y_i, Y_j]) &= -B(Y_j, [Y_i, [Y_i, Y_j]]) \qquad \text{by (5.4.3), (5.4.5)} \\ &= -\mu_j \langle Y_j, [Y_i, [Y_i, Y_j]] \rangle \\ &= -\mu_j \langle Y_i, [Y_j, [Y_j, Y_i]] \rangle, \end{aligned}$$

for example by Corollary 5.3.6 and by the symmetries of the curvature tensor.

In the same manner, however, we also obtain

$$B([Y_i, Y_j], [Y_i, Y_j]) = -\mu_i \langle Y_i, [Y_j, [Y_j, Y_i]] \rangle, \tag{5.4.13}$$

and hence, since B is nondegenerate, we must have

$$[Y_i, Y_j] = 0.$$

Namely, by Theorem 5.3.7, $[Y_i, Y_j] \in \mathfrak{k}$, and by Corollary 5.4.3 B is negative definite on $\mathfrak{k}$. That the restriction of $\langle \cdot, \cdot \rangle_{\mathfrak{g}}$ onto $\mathfrak{k}$ coincides with $-B_{|\mathfrak{k}}$ is a consequence of the definition of $\langle \cdot, \cdot \rangle_{\mathfrak{g}}$. Moreover, for $Y, Z \in \mathfrak{p}_j$

$$B(Y, Z) = \mu_j \langle Y, Z \rangle,$$

and $\mathfrak{p}_i$ and $\mathfrak{p}_j$ for $i \neq j$ are orthogonal w.r.t. $\langle \cdot, \cdot \rangle$ and B. This implies (5.4.12), because, since B is nondegenerate, all μ_j must be $\neq 0$. □

Definition 5.4.4. Let $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ be the usual decomposition of the space of Killing fields of the symmetric space M.

M is called of *Euclidean type*, if

$$[\mathfrak{p}, \mathfrak{p}] = 0,$$

i.e. if the restriction of the Killing form vanishes identically on $\mathfrak{p}$.

M is called *semisimple*, if $\mathfrak{g}$ is semisimple.

M is called *of compact (noncompact)* type, if it is semisimple and of nonnegative (nonpositive) sectional curvature.

Corollary 5.4.4. *A semisimple symmetric space is of (non)compact type if and only if B is negative (positive) definite on $\mathfrak{p}$.*

Proof. Since B is negative definite on $\mathfrak{p}$, all μ_i are < 0, and Corollary 5.4.3 and 5.4.13 imply

$$-\langle Y_i, [Y_j, [Y_j, Y_k]] \rangle \geq 0,$$

hence $K \geq 0$ by Corollary 5.3.6. If conversely $K \geq 0$, B must be negative definite on $\mathfrak{p}$, because otherwise we would contradict (5.4.13), since by Corollary 5.4.3 $B([Y_i, Y_j], [Y_i, Y_j]) \leq 0$. The case $K \leq 0$ is analogous. □

Perspectives. Symmetric spaces were introduced and investigated by E. Cartan. They form a central class of examples in Riemannian geometry, combining the advantage of a rich variety of geometric phenomena with the possibility of explicit computations. Moreover, symmetric spaces can be completely classified in a finite number of series (like $S^n = \mathrm{SO}(n+1)/\mathrm{SO}(n)$, hyperbolic space $H^n = \mathrm{SO}_0(n,1)/\mathrm{SO}(n)$, $\mathbb{CP}^n = \mathrm{SU}(n+1)/\mathrm{S}(\mathrm{U}(n)\times\mathrm{U}(1))$, $\mathrm{Sl}(n,\mathbb{R})/\mathrm{SO}(n)$, $\mathrm{Sp}(p+q)/\mathrm{Sp}(p) \times \mathrm{Sp}(q)$, etc.) plus a finite list of exceptional spaces. Moreover, there exists a duality between the ones of compact and of noncompact type. For example, the dual companion of the sphere $S^n = \mathrm{SO}(n+1)/\mathrm{SO}(n)$ is hyperbolic space $H^n = \mathrm{SO}_0(n,1)/\mathrm{SO}(n)$. A reference for the theory of symmetric spaces is Helgason[123].

5.5 The Space $\mathrm{Sl}(n,\mathbb{R})/\mathrm{SO}(n,\mathbb{R})$

We now want to consider examples: In fact, we shall specialize the examples of §1.10 to the case where we identify the vector space with $\mathbb{R}^n$.

Let M^n be the space of $(n\times n)$-matrices over $\mathbb{R}$ ($M^n \simeq \mathbb{R}^{n^2}$),

$$\begin{aligned}
\mathrm{Gl}(n,\mathbb{R}) &:= \{A\in M^n : \det A \neq 0\} && \text{(linear group)},\\
\mathrm{Sl}(n,\mathbb{R}) &:= \{A\in M^n : \det A = 1\} && \text{(special linear group)},\\
\mathrm{SO}(n) := \mathrm{SO}(n,\mathbb{R}) &:= \{A\in M^n : A^t = A^{-1}, \det A = 1\} && \text{(special orthogonal group)}.
\end{aligned}$$

(Note that A^t is the adjoint A^* of A w.r.t. the Euclidean scalar product).

Obviously, these are Lie groups.

$$\mathfrak{gl}(n,\mathbb{R}) := M^n$$

when equipped with the Lie bracket

$$[X,Y] := XY - YX$$

becomes a Lie algebra, and so do

$$\begin{aligned}
\mathfrak{sl}(n,\mathbb{R}) &:= \{X\in M^n : \mathrm{tr}\ X = 0\},\\
\mathfrak{so}(n) := \mathfrak{so}(n,\mathbb{R}) &:= \{X\in M^n : X^t = -X\};
\end{aligned}$$

these are the Lie algebras of $\mathrm{Gl}(n,\mathbb{R}), \mathrm{Sl}(n,\mathbb{R}), \mathrm{SO}(n,\mathbb{R})$. As in §1.10, one verifies this by considering for $X\in\mathfrak{gl}(n,\mathbb{R})$ the exponential series

$$e^{tX} := \mathrm{Id} + tX + \frac{t^2}{2}X^2 + \dots\ .$$

We have,

$$\det(e^X) = e^{\mathrm{tr}\,X}, \tag{5.5.1}$$

as is easily seen with the help of the Jordan normal form.

In particular, for all $t\in\mathbb{R}$

$$e^{tX}\in \mathrm{Gl}(n,\mathbb{R}).$$

By (5.5.1), if $X\in\mathfrak{sl}(n,\mathbb{R})$, then $e^X\in\mathrm{Sl}(n,\mathbb{R})$. Moreover, for $X\in\mathfrak{so}(n,\mathbb{R})$

$$(e^X)^t = \mathrm{Id} + X^t + \frac{1}{2}(X^t)^2 + \ldots = \mathrm{Id} - X + \frac{1}{2}(X)^2 - \ldots = e^{-X} = (e^X)^{-1},$$

i.e. $e^X\in\mathrm{SO}(n,\mathbb{R})$.

The series representation of e^{tX} also easily implies that the derivative of

$$\begin{aligned}
\mathfrak{gl}(n,\mathbb{R}) &\to \mathrm{Gl}(n,\mathbb{R}),\\
X &\mapsto e^X
\end{aligned}$$

at $X = 0$ is the identity; note in particular that $\mathfrak{gl}(n,\mathbb{R})$ and $\mathrm{Gl}(n,\mathbb{R})$ are of the same dimension. Therefore, the exponential map $X \mapsto e^X$ is a diffeomorphism in the vicinity of $X = 0$.

The exponential map then also yields a diffeomorphism between neighborhoods of 0 in $\mathfrak{sl}(n,\mathbb{R})$ and $\mathfrak{so}(n)$, resp., and neighborhoods of Id in $\mathrm{Sl}(n,\mathbb{R})$ and $\mathrm{SO}(n)$, resp., because the corresponding spaces again have the same dimension.

From §1.10, we recall that for $A, B \in \mathrm{Gl}(n,\mathbb{R})$,

$$\mathrm{Int}\,(A)B = ABA^{-1}.$$

Therefore, for $X \in \mathfrak{gl}(n,\mathbb{R})$

$$(\mathrm{Ad}\,A)X = \frac{d}{dt} A e^{tX} A^{-1}|_{t=0} = AXA^{-1} \tag{5.5.2}$$

and for $Y \in \mathfrak{gl}(n,\mathbb{R})$ then

$$(\mathrm{ad}\,Y)X = \frac{\partial^2}{\partial t \partial s} e^{sY} e^{tX} e^{-sY}|_{s=t=0} = YX - XY = [Y,X]. \tag{5.5.3}$$

We now let $E^{ij} \in m^n$ be the matrix with entry 1 at the intersection of the i^{th} row and the j^{th} column and entries 0 otherwise, $E^{ij} = (e^{ij}_{k\ell})_{k,\ell=1,\ldots,n}$.

Then with $X = (x_{k\ell}), Y = (y_{k\ell})$,

$$\begin{aligned}&\mathrm{ad}\,X\,\mathrm{ad}\,Y E^{ij} = \\ &\Big(x_{k\ell} y_{\ell m} e^{ij}_{mh} - x_{k\ell} e^{ij}_{\ell m} y_{mh} - y_{k\ell} e^{ij}_{\ell m} x_{mh} + e^{ij}_{k\ell} x_{\ell m} y_{mh}\Big)_{k,h=1,\ldots,n}\end{aligned}$$

and hence

$$\begin{aligned}\mathrm{tr}\,\mathrm{ad}\,X\,\mathrm{ad}\,Y &= \langle E^{ij},\, \mathrm{ad}\,X\,\mathrm{ad}\,Y E^{ij}\rangle \\ &= n x_{ij} y_{ji} - x_{ii} y_{jj} - y_{ii} x_{jj} + n x_{ji} y_{ij} \\ &= 2n\,\mathrm{tr}\,XY - 2\,\mathrm{tr}\,X\,\mathrm{tr}\,Y.\end{aligned} \tag{5.5.4}$$

If $X = \lambda\,\mathrm{Id}$ (Id = identity matrix), then

$$\mathrm{ad}\,X = 0.$$

Therefore, $\mathfrak{gl}(n,\mathbb{R})$ is not semisimple. On $\mathfrak{sl}(n,\mathbb{R})$, however, the Killing form satisfies by (5.5.4)

$$B(X,Y) = 2n\,\mathrm{tr}\,XY. \tag{5.5.5}$$

Therefore, for $X \neq 0$

$$B(X, X^t) > 0, \tag{5.5.6}$$

and the Killing form is nondegenerate.

A similar computation applies to $\mathfrak{so}(n)$: $\mathfrak{so}(2) = \mathbb{R}$ is not semisimple. For $n > 2$, we choose $\{\frac{1}{\sqrt{2}}(E^{ij} - E^{ji}) : i < j\}$ as a basis for $\mathfrak{so}(n)$. Then

$$\Big\langle \tfrac{1}{\sqrt{2}}(E^{ij} - E^{ji}),\ \tfrac{1}{\sqrt{2}}(E^{k\ell} - E^{\ell k})\Big\rangle = \delta_{ik}\delta_{j\ell} \quad \text{for } i < j, k < \ell$$

and

$$\begin{aligned}\operatorname{tr}\operatorname{ad}X\operatorname{ad}Y &= \left\langle \tfrac{1}{\sqrt{2}}(E^{ij}-E^{ji}),\ \operatorname{ad}X\operatorname{ad}Y\tfrac{1}{\sqrt{2}}(E^{ij}-E^{ji})\right\rangle \\ &= (n-1)x_{k\ell}y_{\ell k} + x_{ij}y_{ij}.\end{aligned}$$

using $\operatorname{tr}X = \operatorname{tr}Y = 0$ for $X, Y \in \mathfrak{so}(n)$.

Since $X = -X^t$ for $X \in \mathfrak{so}(n)$, we obtain

$$\operatorname{tr}\operatorname{ad}X\operatorname{ad}Y = (n-2)\operatorname{tr}X\cdot Y.$$

In particular, for $n > 2$ let B now denote the Killing form of $\mathfrak{so}(n)$, then for $X \neq 0$

$$B(X, X^t) < 0,$$

and also

$$B(X, X) < 0.$$

Thus, the Killing form of $\mathfrak{so}(n)$ is negative definite for $n > 2$. Note that the Killing form of $\mathfrak{so}(n)$ does not coincide with the restriction of the Killing form of $\mathfrak{sl}(n,\mathbb{R})$ onto $\mathfrak{so}(n)$. In the sequel, we shall employ the latter one.

(5.5.5) directly implies that B is $\operatorname{Ad}(\mathrm{Sl}(n,\mathbb{R}))$ invariant. We now put

$$G = \mathrm{Sl}(n,\mathbb{R}), \qquad K = \mathrm{SO}(n),$$

$$\mathfrak{g} = \mathfrak{sl}(n,\mathbb{R}), \quad \mathfrak{k} = \mathfrak{so}(n), \quad \mathfrak{p} = \{X \in \mathfrak{sl}(n,\mathbb{R}) : X^t = X\}.$$

Then because of $X = \frac{1}{2}(X - X^t) + \frac{1}{2}(X + X^t)$,

$$\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}. \tag{5.5.7}$$

Moreover, because of $(XY - YX)^t = Y^tX^t - X^tY^t$

$$[\mathfrak{k},\mathfrak{k}] \subset \mathfrak{k}, [\mathfrak{p},\mathfrak{p}] \subset \mathfrak{k}, [\mathfrak{k},\mathfrak{p}] \subset \mathfrak{p}. \tag{5.5.8}$$

Next, let

$$M := G/K;$$

more precisely, M is the space of equivalence classes w.r.t. the following equivalence relation on G :

$$g_1 \sim g_2 :\Longleftrightarrow \exists k \in K : g_2 = g_1 k.$$

Thus, M is the space of left cosets of K in G. As K is not a normal subgroup of G, M is not a group. We want to equip M with a symmetric space structure. G operates transitively on M by

$$g'K \mapsto gg'K \quad \text{for } g \in G.$$

Let

$$\pi : G \to M$$

be the projection.

A subset Ω of M is called open, if $\pi^{-1}(\Omega)$ is open in G. Then π becomes an open map.

We want to show that M is a Hausdorff space. The preimage of K under the continuous map

$$G \times G \to G,$$
$$(g_1, g_2) \mapsto g_1^{-1} g_2$$

is closed since K is closed in G. Thus, if $g_1^{-1} g_2 \notin K$, in $G \times G$ there exists a neighborhood of (g_1, g_2) of the form $\tilde{\Omega}_1 \times \tilde{\Omega}_2$ which is disjoint from the preimage of K. If now $g_1 K \neq g_2 K$, then $g_1^{-1} g_2 \notin K$, and $\Omega_i := \pi(\tilde{\Omega}_i), i = 1, 2$, are disjoint neighborhoods of $g_1 K$ and $g_2 K$. Namely, if $gK \in \Omega_i$, there exists $k_i \in K$ with $gk_i \in \tilde{\Omega}_i$, and if we had $gK \in \Omega_1 \cap \Omega_2$, (gk_1, gk_2) would be mapped to $k_1^{-1} k_2 \in K$, and $\tilde{\Omega}_1 \times \tilde{\Omega}_2$ would not be disjoint to the preimage of K. This shows the Hausdorff property.

In order to construct coordinate charts, we first have to recall the Cauchy polar decomposition of an invertible matrix.

Lemma 5.5.1. *For $A \in \mathrm{Gl}(n, \mathbb{R})$, there exist an orthogonal matrix R and a symmetric positive definite matrix V with*

$$A = VR,$$

and this decomposition is unique.

Proof. Since A is invertible,

$$H := AA^t$$

is symmetric and positive definite. We are going to show that there exists a unique symmetric, positive definite matrix V with $V^2 = H$. For this purpose, we first observe that H may be diagonalized by an orthogonal matrix S :

$$H = S^t \Lambda S \text{ with } \Lambda = \mathrm{diag}\,(\lambda_i), \lambda_i > 0 \text{ by positive definiteness.}$$

We put

$$V := S^t \,\mathrm{diag}\,(\sqrt{\lambda_i}) S.$$

V then is symmetric, positive definite, and because of $S^t = S^{-1}$, it satisfies

$$V^2 = H.$$

This shows existence. For uniqueness, we first show that for a symmetric, positive definite matrix V, each eigenvector of V^2 with eigenvalue λ is an eigenvector of V with eigenvalue $\lambda^{\frac{1}{2}}$. Namely, from $V^2 x = \lambda x$ it follows that

$$(V + \sqrt{\lambda}\,\mathrm{Id})(V - \sqrt{\lambda}\,\mathrm{Id})x = 0,$$

and therefore we must have $y := (V - \sqrt{\lambda}\,\mathrm{Id})x = 0$, because otherwise y would be an eigenvector of V with eigenvalue $-\sqrt{\lambda} < 0$, contradicting the positive definiteness of V.

This implies that the relation $V^2 = H$ uniquely determines V, because all eigenvalues and eigenvectors of V are determined by those of H.

We now put

$$R = V^{-1}A.$$

Then

$$RR^t = V^{-1}AA^tV^{-1} = V^{-1}V^2V^{-1} = \mathrm{Id},$$

and R is orthogonal.

This shows the existence of the decomposition. Uniqueness is likewise easy:

If $A = VR$, with orthogonal R and with symmetric, positive definite V, then

$$AA^t = VRR^tV^t = V^2,$$

and by the preceding, this uniquely determines V. R then is unique as well. $\square$

Let

$$P := \{A \in \mathrm{Sl}(n,\mathbb{R}) : A^t = A, A \text{ pos. def.}\}$$

(Note that P is not a group.)

For $X \in \mathfrak{p}$, then

$$e^X \in P$$

and the exponential map again yields a diffeomorphism between a neighborhood of O in $\mathfrak{p}$ and a neighborhood of Id in $\mathfrak{p}$. We now decompose $A \in \mathrm{Sl}(n,\mathbb{R})$ according to Lemma 5.5.1

$$A = VR$$

with $R \in \mathrm{O}(n), V \in P$.

Let A be contained in a sufficiently small neighborhood of Id.

There then exist unique

$$X \in \mathfrak{so}(n), Y \in \mathfrak{p}$$

with

$$e^X = R, e^Y = V.$$

This implies the existence of neighborhoods Ω_1 of 0 in $\mathfrak{p}$, Ω_2 of 0 in $\mathfrak{so}(n)$ for which

$$\begin{aligned} \Omega_1 \times \Omega_2 &\to G, \\ (Y,X) &\mapsto e^Y e^X \end{aligned}$$

is a diffeomorphism onto its image.

Lemma 5.5.2. *G/K is homeomorphic to P. If G/K is equipped with the differentiable structure of P,*

$$\exp : \mathfrak{p} \to G/K \simeq P, V \mapsto e^V$$

becomes a local diffeomorphism between a neighborhood of 0 in $\mathfrak{p}$ and a neighborhood of $\mathrm{Id} \cdot K$ in G/K.

Proof. We first construct a homeomorphism Φ between G/K and P. For gK we write by Lemma 5.5.1

$$g = VR \text{ with } V \in P, R \in \mathrm{SO}(n)$$

and put

$$\Phi(g) = V.$$

This does not depend on the choice of representative of gK. Namely, if $gK = g'K$, there exists $S \in \mathrm{SO}(n) = K$ with $gS = g'$, hence $g' = VRS = VR'$ with $R' := RS \in \mathrm{SO}(n)$, and $\Phi(g') = V = \Phi(g)$. If conversely $\Phi(g) = \Phi(g') =: V$, then $g = VR, g' = VS$ with $R, S \in \mathrm{SO}(n)$, hence $g' = g(R^{-1}S)$ with $R^{-1}S \in \mathrm{SO}(n)$, hence $gK = g'K$. Therefore, Φ is bijective. Φ is continuous in both directions, because

$$\pi : G \to G/K$$

and

$$\begin{aligned} \pi : G &\to P, \\ A &\mapsto V, \end{aligned}$$

with $A = VF$ (the unique decomposition of Lemma 5.5.1), both are continuous and open.

Moreover $\exp(\mathfrak{p}) \subset P$, and since $\exp : \mathfrak{gl}(n, \mathbb{R}) \to \mathrm{Gl}(n, \mathbb{R})$ is a local diffeomorphism, and $\mathfrak{p}$ and P have the same dimension, $\exp_{|\mathfrak{p}}$ is a local diffeomorphism, too, between a neighborhood of 0 in $\mathfrak{p}$ and a neighborhood of Id in P. $\square$

By Lemma 5.5.2, G/K becomes a differentiable manifold. We have already displayed a chart near $\mathrm{Id} \cdot K$. In order to obtain a chart at gK, we simply map a suitable neighborhood U of gK via g^{-1} onto a neighborhood $g^{-1}U$ of $\mathrm{Id} \cdot K$ and use the preceding chart.

G then operates transitively on G/K by diffeomorphisms,

$$\begin{aligned} G \times G/K &\to G/K, \\ (h, gK) &\mapsto hgK. \end{aligned}$$

The isotropy group of $\mathrm{Id} \cdot K$ is K itself. The isotropy group of gK is gKg^{-1}, and this group is conjugate to K.

We want to construct Riemannian metrics on G on G/K w.r.t. which G operates isometrically on G/K.

For this purpose, we use the Killing form B of $\mathfrak{sl}(n,\mathbb{R})$ and the decomposition $\mathfrak{g} = \mathfrak{sl}(n,\mathbb{R}) = \mathfrak{k} \oplus \mathfrak{p}$ (with $\mathfrak{k} = \mathfrak{so}(n)$). We put

$$\langle X, Y\rangle_{\mathfrak{g}} = \begin{cases} B(X,Y) & \text{for } X, Y \in \mathfrak{p}, \\ -B(X,Y) & \text{for } X, Y \in \mathfrak{k}, \\ 0 & \text{for } X \in \mathfrak{p}, Y \in \mathfrak{k} \text{ or vice versa.} \end{cases}$$

By (5.5.5), $\langle \cdot, \cdot \rangle_{\mathfrak{g}}$ is positive definite.

For abbreviation, we put

$$e := \mathrm{Id} \qquad \text{(identity matrix)}$$

and we identify $\mathfrak{g}$ with T_eG. For each $g \in G$, we then also obtain a metric on T_gG by requesting that the left translation

$$\begin{aligned} L_g : G &\to G, \\ h &\mapsto gh \end{aligned}$$

is an isometry between T_eG and T_gG ($dL_g : T_eG \to T_gG$). We also obtain a metric on G/K : restricting $\langle \cdot, \cdot \rangle_{\mathfrak{g}}$ to $\mathfrak{p}$, we get a metric on $T_{eK}G/K \simeq \mathfrak{p}$; the metric on $T_{gK}G/K$ then is produced by

$$\begin{aligned} \tilde{L}_g : G/K &\to G/K, \\ hK &\mapsto ghK \end{aligned}$$

by requesting again that those maps are isometries.

The metric is well defined; namely, if

$$gK = g'K,$$

then

$$g' = gk \quad \text{with} \quad k \in K,$$

hence $\tilde{L}_{g'} = \tilde{L}_g \circ \tilde{L}_k$. $\tilde{L}_k$ now maps eK onto itself, and $d\tilde{L}_k : T_{eK}G/K \to T_{eK}G/K$ is an isometry, since for $V \in P$, $L_kV = kV = (kVk^{-1})k = ((\mathrm{Int}\, k)V)k$, hence $d\tilde{L}_k(X) = (\mathrm{Ad}\, , k)X$ for $X \in \mathfrak{p} \simeq T_{eK}G/K$, and $\mathrm{Ad}\, k$ is an isometry of $\mathfrak{p}$ because it leaves the Killing form invariant. Therefore, the metric on G/K is indeed well defined. By definition, G then operates isometrically on G/K.

We want to define involutions on G/K so as to turn G/K into a symmetric space.

We first have an involution

$$\begin{aligned} \sigma_e : G &\to G, \\ h &\mapsto (h^{-1})^t \end{aligned}$$

with

$$\begin{aligned} d\sigma_e : \mathfrak{g} &\to \mathfrak{g}, \\ X &\mapsto -X^t, \end{aligned}$$

hence

$$d\sigma_{e|\mathfrak{k}} = \mathrm{id} d\sigma_{e|\mathfrak{p}} = -\mathrm{id}, \sigma_{e|K} = \mathrm{id}.$$

For $g \in G$, we then obtain an involution

$$\sigma_g : G \to G$$

by

$$\sigma_g h = L_g \sigma_e (L_{g^{-1}} h) = g((g^{-1}h)^{-1})^t = gg^t (h^{-1})^t.$$

We have

$$\sigma_g^2(h) = gg^t(((gg^t(h^{-1})^t)^{-1})^t) = h,$$

hence

$$\sigma_g^2 = \mathrm{id}$$

and

$$\sigma_g(g) = g.$$

Since $\sigma_{e|K} = \mathrm{id}$, σ_e induces an involution

$$\sigma_{eK} : G/K \to G/K$$

with $\sigma_{eK}(eK) = eK, d\sigma_{eK} : T_{eK}G/K \to T_{eK}G/K$, $d\sigma_{eK} = -\mathrm{id}$. Since G operates transitively on G/K, at each $gK \in G/K$, we then also obtain an involution $\sigma_{gK} = \tilde{L}_g \circ \sigma_{eK} \circ \tilde{L}_{g^{-1}}$.

We have thus shown

Theorem 5.5.1. *G/K carries a symmetric space structure.* □

The group of orientation preserving isometries of G/K is G itself. Namely, that group cannot be larger than G, because any such isometry is already determined by its value and its derivative at one point, and G operates transitively on $M = G/K$, and so does K on $T_{eK}M$, and hence G already generates all such isometries. We want to establish the connection with the theory developed in §5.3 and 5.4. We first want to compare the exponential map on $\mathfrak{sl}(n, \mathbb{R})$ and the induced map on G/K with the Riemannian exponential map. Let a one parameter subgroup of G be given, i.e. a Lie group homomorphism

$$\varphi : \mathbb{R} \to G.$$

Thus $\varphi(s+t) = \varphi(s) \circ \varphi(t)$, hence

$$\frac{\varphi(t+h) - \varphi(t)}{h} = \varphi(t) \frac{\varphi(h) - 1}{h},$$

hence

$$\frac{d\varphi}{dt}(t) = \frac{d\varphi}{dt}(0)\varphi(t).$$

As usual, this implies

$$\varphi(t) = e^{t\frac{d\varphi}{dt}(0)}.$$

Thus, the exponential map generates all one parameter subgroups of G.

If c is a geodesic in G/K with $c(0) = eK =: p$, the translations τ_t along c yield a one parameter subgroup of G, hence

$$\exp_p t\dot{c}(0) = c(t) = \tau_t(p) = e^{tX}(p) \qquad \text{for some } X\, T_{eK}G/K \cong \mathfrak{p}. \tag{5.5.9}$$

Here, on the left, we have the Riemannian exponential map, whereas on the right, we have the one of G. Since the derivative of the Lie group exponential map at 0 is the identity, we obtain $X = \dot{c}(0)$, and the two exponential maps coincide. In particular, the Lie group exponential map, when applied to the straight lines through the origin in $\mathfrak{p}$, generates the geodesics of G/K.

We also obtain a map ψ from the Lie algebra $\mathfrak{sl}(n,\mathbb{R})$ of $\mathrm{Sl}(n,\mathbb{R})$ into the Lie algebra of Killing fields of G/K. For $X \in \mathfrak{sl}(n,\mathbb{R})$ we put

$$\begin{aligned}\psi(X)(q) &= \frac{d}{dt} g e^{tX}(p)_{|t=0} \qquad \text{for } q = g(p)\\ &= \frac{d}{dt} L_g e^{tX}(p)_{|t=0}.\end{aligned}$$

Now

$$\begin{aligned}\psi(XY)(q) &= dgXY(p)\\ &= \frac{\partial^2}{\partial t \partial s} g e^{tX} e^{sY}(p)_{|t=s=0}\\ &= \frac{d}{dt}\psi(Y)(ge^{tX}(p))_{|t=0}\\ &= \psi(Y)\psi(X)(q),\end{aligned}$$

hence

$$\psi([X,Y]) = [\psi(Y),\psi(X)] = -[\psi(X),\psi(Y)].$$

We thus obtain an antihomomorphism of Lie algebras. This explains the difference in sign between (5.4.3) and (5.5.3).

Corollary 5.5.1. $\mathrm{Sl}(n,\mathbb{R})/\mathrm{SO}(n)$ *is a symmetric space of noncompact type. The sectional curvature of the plane spanned by the orthonormal vectors* $Y_1, Y_2 \in \mathfrak{p}$ *is given by*

$$K = B([Y_2,Y_1],[Y_2,Y_1]) = -\|[Y_1,Y_2]\|_{\mathfrak{g}}^2 \leq 0.$$

Proof. As observed above (5.5.5), the Killing form is nondegenerate, and the symmetric space is semisimple. By Corollary 5.3.6 the sectional curvature of the plane spanned by $Y_1, Y_2 \in \mathfrak{p}$ satisfies

$$\begin{aligned} K &= -\langle [[Y_1, Y_2], Y_2], Y_1 \rangle \\ &= -B([[Y_1, Y_2], Y_2], Y_1) \\ &= -B([Y_2, [Y_2, Y_1]], Y_1) \\ &= \quad B([Y_2, Y_1], [Y_2, Y_1]), \end{aligned} \tag{5.5.10}$$

because the Killing form is $\operatorname{Ad} G$ invariant.

This expression is ≤ 0, because by $[\mathfrak{p}, \mathfrak{p}] \subset \mathfrak{k}$, $[Y_2, Y_1] \in \mathfrak{k}$ and B is negative definite on $\mathfrak{k}$. □

Definition 5.5.1. A subalgebra $\mathfrak{a}$ of $\mathfrak{g}$ is called *abelian* if $[A_1, A_2] = 0$ for all $A_1, A_2 \in \mathfrak{a}$.

We want to find the maximal abelian subspaces of $\mathfrak{p}$. Let $\mathfrak{a}$ be an abelian subspace of $\mathfrak{p}$, i.e. an abelian subalgebra of $\mathfrak{g}$ that is contained in $\mathfrak{p}$. Thus

$$A_1 A_2 - A_2 A_1 = 0 \quad \text{for all } A_1, A_2 \in \mathfrak{a}.$$

The elements of $\mathfrak{a}$ therefore constitute a commuting family of symmetric $(n \times n)$ matrices. Hence, they can be diagonalized simultaneously. Thus, there exists an orthonormal basis $v_1, \ldots, v_n$ of $\mathbb{R}^n$ consisting of common eigenvectors of the elements of $\mathfrak{a}$. We write our matrices w.r.t. an orthonormal basis $e_1, \ldots, e_n$ of $\mathbb{R}^n$, and we choose $S \in \mathrm{SO}(n)$ with

$$S(v_i) = \pm e_i \quad \text{for } i = 1, \ldots, n.$$

$S\mathfrak{a}S^{-1}$ then is an abelian subspace of $\mathfrak{p}$ with eigenvectors $e_1, \ldots, e_n$. Thus, all elements of $S\mathfrak{a}S^{-1}$ are diagonal matrices (with trace 0 since they are contained in $\mathfrak{p}$). This implies that the space of diagonal matrices of trace 0 is a maximal abelian subspace of $\mathfrak{p}$. Furthermore, it follows that each maximal abelian subspace is conjugate to this one, w.r.t. an element from $K = \mathrm{SO}(n)$. Therefore, any two maximal abelian subspaces of $\mathfrak{p}$ are conjugate to each other.

Let now $\mathfrak{a}$ be an abelian subspace of $\mathfrak{p}$. We put

$$A := \exp \mathfrak{a},$$

where exp, as usual, is the exponential map $\mathfrak{g} \to G$. A then is a Lie subgroup of G. For $g_1, g_2 \in A$, we have

$$g_1 g_2 = g_2 g_1,$$

because for any two commuting elements $X, Y \in \mathfrak{g}$

$$e^{X+Y} = e^X e^Y = e^Y e^X,$$

as is easily seen from the exponential series. Thus, A is an abelian Lie group.

On the other hand, because of $\mathfrak{a} \subset \mathfrak{p}$, A also is a subspace of $M = G/K$.

Lemma 5.5.3. *A is totally geodesic in M and flat, i.e. its curvature vanishes.*

Proof. Let $Y \in \mathfrak{a}$. By definition of A, the geodesic e^{tY} is contained in A. A is thus totally geodesic at the point $eK := P$ in the sense that any geodesic of M through p and tangential to A at p is entirely contained in A. A operates transitively and isometrically on itself by left translations. Let now $q \in A$. There then exists $a \in A$ with $ap = q$. Since a as element of G is an isometry, it maps the geodesics of A and those of M through p onto geodesics through q. This implies that A is totally geodesic Al at q as well, hence everywhere. The curvature formula (5.5.10) implies that A is flat. $\square$

Let conversely N be a flat subspace of M. Since the Killing form of $\mathfrak{k}$ is negative definite, the curvature formula (5.5.10) implies $[Y_1, Y_2] = 0$ for all $Y_1, Y_2 \in T_pN$. Thus, T_pN is an abelian subspace of $\mathfrak{p}$.

We conclude

Corollary 5.5.2. *The maximal flat subspaces of M through $p = eK$, i.e. those not contained in any larger flat subspace of M, bijectively correspond to the maximal abelian subspaces of $\mathfrak{p}$.* $\square$

The assertions of Lemma 5.5.3 and Corollary 5.5.2 are valid for all symmetric spaces.

Definition 5.5.2. The *rank* of a symmetric space M is the dimension of a maximal flat subspace.

Thus, the rank is the dimension of a maximal abelian subalgebra of $\mathfrak{g}$ contained in $\mathfrak{p}$. As remarked above, any two such subalgebras are conjugate to each other. Likewise, because G operates transitively on M, the dimension of a maximal flat subspace through any given point of M is the same.

Corollary 5.5.3.

$$\mathrm{Rank}\,(\mathrm{Sl}(n,\mathbb{R})/\mathrm{SO}(n)) = n - 1.$$

Proof. As observed above, a maximal abelian subalgebra of $\mathfrak{g}$ contained in $\mathfrak{p}$ consists of the space of diagonal matrices with vanishing trace, and the latter space has dimension $n - 1$. $\square$

Corollary 5.5.4. *A symmetric space M of noncompact type has rank 1 if and only if its sectional curvature is negative.*

Proof. The rank is 1 if for two linearly independent $Y_1, Y_2 \in T_pM$, we have $[Y_1, Y_2] \neq 0$. Since B is negative definite on $\mathfrak{k}$ and $[Y_1, Y_2] \in \mathfrak{k}$ for Y_1, $Y_2 \in T_pM$ (identified with $\mathfrak{p}$), (5.5.10) yields the claim. $\square$

Lemma 5.5.4. *For $X \in \mathfrak{k}$, $\operatorname{ad} X : \mathfrak{g} \to \mathfrak{g}$ is skew symmetric w.r.t. $\langle \cdot, \cdot \rangle_{\mathfrak{g}}$, and for $X \in \mathfrak{p}$, it is symmetric.*

Proof. Let $X \in \mathfrak{k}, Y, Z \in \mathfrak{k}$. Then $(\operatorname{ad} X)Y = [X, Y] \in \mathfrak{k}$, hence

$$\begin{aligned} \langle [X,Y], Z\rangle_{\mathfrak{g}} &= -B([X,Y],Z) \\ &= \quad B(Y,[X,Z]) \\ &= -\langle Y,[X,Z]\rangle_{\mathfrak{g}} \qquad \text{by (5.4.5)}. \end{aligned}$$

For $X \in \mathfrak{k}, Y \in \mathfrak{p}, Z \in \mathfrak{k}$, we have $[X,Y] \in \mathfrak{p}, [X,Z] \in \mathfrak{k}$, hence

$$\langle [X,Y], Z\rangle_{\mathfrak{g}} = 0 = \langle Y,[X,Z]\rangle_{\mathfrak{g}}.$$

For $X \in \mathfrak{k}, Y, Z \in \mathfrak{p}$, we have $[X,Y] \in \mathfrak{p}, [X,Z] \in \mathfrak{p}$ and

$$\begin{aligned} \langle [X,Y], Z\rangle_{\mathfrak{g}} &= \quad B([X,Y],Z) \\ &= -B(Y,[X,Z]) \\ &= -\langle Y,[X,Z]\rangle_{\mathfrak{g}} \qquad \text{by (5.4.5)}. \end{aligned}$$

Altogether, this implies that $\operatorname{ad} X$ is skew symmetric for $X \in \mathfrak{k}$. Let now $X \in \mathfrak{p}$, $Y, Z \in \mathfrak{k}$. Then $[X,Y] \in \mathfrak{p}, [X,Z] \in \mathfrak{p}$, hence

$$\langle [X,Y], Z\rangle_{\mathfrak{g}} = 0 = \langle Y,[X,Z]\rangle_{\mathfrak{g}}.$$

For $X \in \mathfrak{p}, Y \in \mathfrak{k}, Z \in \mathfrak{p}$, we have $[X,Y] \in \mathfrak{p}, [X,Z] \in \mathfrak{k}$, hence

$$\begin{aligned} \langle [X,Y], Z\rangle_{\mathfrak{g}} &= \quad B([X,Y],Z) \\ &= -B(Y,[X,Z]) \qquad \text{by (5.4.5)} \\ &= \quad \langle Y,[X,Z]\rangle_{\mathfrak{g}}. \end{aligned}$$

Finally for $X \in \mathfrak{p}, Y, Z \in \mathfrak{p}$, we have $[X,Y] \in \mathfrak{k}, [X,Z] \in \mathfrak{k}$, hence

$$\langle [X,Y], Z\rangle_{\mathfrak{g}} = 0 = \langle Y,[X,Z]\rangle_{\mathfrak{g}}.$$

Altogether, this implies that ad X is skew symmetric for $X \in \mathfrak{p}$. $\square$

Lemma 5.5.5. *If $X, Y \in \mathfrak{g}$ commute, i.e. $[X,Y] = 0$, then so do $\operatorname{ad} X$ and $\operatorname{ad} Y$.*

Proof.

$$\operatorname{ad} X \operatorname{ad} Y Z = \quad [X,[Y,Z]]$$

by the Jacobi identity,

$$= -[Y,[Z,X]] - [Z,[X,Y]]$$

because $[X, Y] = 0$,

$$\begin{aligned} &= [Y, [X, Z]] \\ &= \operatorname{ad} Y \operatorname{ad} X Z. \end{aligned}$$

$\square$

Let now $\mathfrak{a}$ be a fixed maximal abelian subspace of $\mathfrak{p}$. By Lemmas 5.5.4, 5.5.5, for $X \in \mathfrak{a}$, the maps $\operatorname{ad} X : \mathfrak{g} \to \mathfrak{g}$ are symmetric w.r.t. $\langle \cdot, \cdot \rangle_{\mathfrak{g}}$ and commute with each other. Therefore, $\mathfrak{g}$ can be decomposed as a sum orthogonal w.r.t. $\langle \cdot, \cdot \rangle_{\mathfrak{g}}$ of common eigenvectors of the $\operatorname{ad} X, X \in \mathfrak{a}$:

$$\mathfrak{g} = \mathfrak{g}_0 \oplus \sum_{\alpha \in \Lambda} \mathfrak{g}_\alpha.$$

Definition 5.5.3. Λ is called the *set of roots*, and the $\alpha \in \Lambda$ are called the *roots* of $\mathfrak{g}$ w.r.t. $\mathfrak{a}$.

We have

$$[X, Y] = (\operatorname{ad} X) Y = \alpha(X) Y \qquad \text{for } X \in \mathfrak{a}, Y \in \mathfrak{g}_\alpha. \tag{5.5.11}$$

Thus $\alpha(X)$ is the eigenvalue of $\operatorname{ad} X$ on $\mathfrak{g}_\alpha$, with $0(X) := 0$ for all X. Since $\mathfrak{a}$ is abelian, of course

$$\mathfrak{a} \subset \mathfrak{g}_0.$$

Moreover, $\alpha : \mathfrak{a} \to \mathbb{R}$ is linear for all $\alpha \in \Lambda$, since

$$\begin{aligned} \operatorname{ad}(X + Y) &= \operatorname{ad} X + \operatorname{ad} Y, \\ \operatorname{ad}(\mu X) &= \mu \operatorname{ad} X, \end{aligned}$$

for $X, Y \in \mathfrak{a}, \mu \in \mathbb{R}$.

We now recall the involution

$$\sigma_e : G \to G, \qquad \sigma_e(h) = (h^{-1})^t,$$

and

$$\theta := d\sigma_e : \mathfrak{g} \to \mathfrak{g}, \qquad \theta(X) = -X^t,$$

which is also called *Cartan involution*, and the decomposition

$$\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p},$$

$\mathfrak{k}$ being the eigenspace of θ with eigenvalue 1, $\mathfrak{p}$ the one with eigenvalue -1, is called *Cartan decomposition.* We thus may write

$$\langle X, Y \rangle_{\mathfrak{g}} = -B(X, \theta Y). \tag{5.5.12}$$

In the same manner as e does, any element g of G, hence also any element gK of G/K induces a Cartan decomposition $\mathfrak{g} = \mathfrak{k}' \oplus \mathfrak{p}'$ with $\mathfrak{k}' = \operatorname{Ad}(g)\mathfrak{k}$ etc. (cf. also §5.3).

Lemma 5.5.6.

(i) $[\mathfrak{g}_\alpha, \mathfrak{g}_\beta] \subset \mathfrak{g}_{\alpha+\beta}$ *for* $\alpha + \beta \in \Lambda$, $[\mathfrak{g}_\alpha, \mathfrak{g}_\beta] = 0$ *for* $\alpha + \beta \notin \Lambda$.

(ii) $\alpha \in \Lambda \iff -\alpha \in \Lambda$, *and for each* $\alpha \in \Lambda, \theta : \mathfrak{g}_\alpha \to \mathfrak{g}_{-\alpha}$ *is an isomorphism.*

(iii) θ *leaves* $\mathfrak{g}_0$ *invariant,* $\mathfrak{g}_0 = \mathfrak{g}_0 \cap \mathfrak{k} + \mathfrak{a}$.

(iv) *For* $X \in \mathfrak{a}, Y \in \mathfrak{g}_\alpha$, $\mathrm{Ad}\,(e^{tX})Y = e^{t\alpha(X)}Y$.

(v) *For* $\alpha \neq -\beta$, $B(\mathfrak{g}_\alpha, \mathfrak{g}_\beta) = 0$.

Proof. Let $Y \in \mathfrak{g}_\alpha, Z \in \mathfrak{g}_\beta, X \in \mathfrak{a}$. Then

$$(\mathrm{ad}\,X)[Y,Z] = [X,[Y,Z]]$$

because of the Jacobi identity,

$$\begin{aligned} &= -[Y,[Z,X]] - [Z,[X,Y]] \\ &= \beta(X)[Y,Z] + \alpha(X)[Y,Z] \\ &= (\alpha+\beta)(X)[Y,Z]. \end{aligned}$$

This implies (i).

Next

$$[X, \theta Y] = \quad [X, -Y^t]$$

by $X = X^t$, since $X \in \mathfrak{a} \subset \mathfrak{p}$,

$$\begin{aligned} &= -[X^t, Y^t] \\ &= \quad [X,Y]^t \\ &= \quad \alpha(X) Y^t \\ &= -\alpha(X)\theta Y, \end{aligned}$$

hence $\theta Y \in \mathfrak{g}_{-\alpha}$. This proves (ii), and the first part of (iii), too, hence also $\mathfrak{g}_0 = (\mathfrak{g}_0 \cap \mathfrak{k}) + (\mathfrak{g}_0 \cap \mathfrak{p})$.

Since $\mathfrak{a}$ is maximal abelian in $\mathfrak{p}$ and commutes with all elements of $\mathfrak{g}_0$, it follows that $\mathfrak{g}_0 \cap \mathfrak{p} = \mathfrak{a}$ which is the remaining part of (iii).

Next

$$\mathrm{Ad}\,(e^{tX}) = e^{t\,\mathrm{ad}\,X} = \mathrm{Id} + \sum_{n=1}^{\infty} \frac{t^n}{n!} (\mathrm{ad}\,X)^n$$

which implies (iv).

Finally, (v) follows from

$$\begin{aligned} 0 = \quad & \langle \mathfrak{g}_\alpha, \mathfrak{g}_\beta \rangle_{\mathfrak{g}} && \text{for } \alpha \neq \beta \\ = & -B(\mathfrak{g}_\alpha, \theta(\mathfrak{g}_\beta)) && \text{by (5.5.11)} \\ = & -B(\mathfrak{g}_\alpha, \mathfrak{g}_{-\beta}) && \text{by (ii).} \end{aligned}$$

□

We now want to determine the root space decomposition of $\mathfrak{g} = \mathfrak{sl}(n,\mathbb{R})$. For that purpose, let E^{ij} be as above, and

$$H^i := E^{ii} - E^{i+1,i+1}, \quad i = 1, \ldots, n-1.$$

$\{E^{ij} (i \neq j)$ and $H^k (k = 1, \ldots, n-1)\}$ then constitute a basis of $\mathfrak{g}$. Let $\mathfrak{a}$ be the space of diagonal matrices with vanishing trace, i.e. a maximal abelian subspace of $\mathfrak{p}$.

For $X = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n) = \sum_{i=1}^n \lambda_i E^{ii}$, we have

$$\begin{aligned} (\mathrm{ad}\, X) E^{ij} &= (\lambda_i - \lambda_j) E^{ij}, && \text{for } i \neq j, \\ (\mathrm{ad}\, X) H^i &= 0, && \text{for } i = 1, \ldots, n-1, \text{ since } H^i \in \mathfrak{a}. \end{aligned}$$

We thus obtain $n(n-1)$ nonzero roots $\alpha_{ij} (i \neq j)$ with

$$\alpha_{ij}(X) = \lambda_i - \lambda_j \quad (X = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)).$$

The corresponding root spaces $\mathfrak{g}_{\alpha_{ij}}$ are spanned by the E^{ij}. $\mathfrak{g}_0$ is spanned by H^1, ..., H^{n-1}; in particular

$$\mathfrak{g}_0 = \mathfrak{a}.$$

Definition 5.5.4. A maximal flat abelian subspace of G/K is called a *flat.* A geodesic in G/K is called *regular* if contained in one flat only; otherwise it is called *singular.* Tangent vectors of regular (singular) geodesics are called *regular (singular).*

Lemma 5.5.7. *$X \in \mathfrak{a}$ is singular iff there exists $Y \in \mathfrak{g}\backslash\mathfrak{g}_0$ with $[X,Y] = 0$, i.e. if there exists $\alpha \in \Lambda$ with $\alpha(X) = 0$.*

Proof. Let X be singular. Then X is contained in another maximal abelian subspace $\mathfrak{a}'$ of $\mathfrak{p}$ besides $\mathfrak{a}$. Therefore, there exists $Y \in \mathfrak{a}', Y \notin \mathfrak{a}$. Because of $X, Y \in \mathfrak{a}'$,

$$[X,Y] = 0.$$

Since $\mathfrak{g}_0 \cap \mathfrak{p} = \mathfrak{a}$ (Lemma 5.5.6 (ii)), $Y \notin \mathfrak{g}_0$. (5.5.11) implies $\alpha(X) = 0$ for at least one $\alpha \in \Lambda$.

Let now $\alpha(X) = 0$ for such a $\alpha \in \Lambda$. Let $Y \in \mathfrak{g}_\alpha, Y \neq 0$. Then

$$[X,Y] = \alpha(X) Y = 0. \tag{5.5.13}$$

We decompose

$$Y = Y_{\mathfrak{k}} + Y_{\mathfrak{p}} \quad \text{with } Y_{\mathfrak{k}} \in \mathfrak{k}, Y_{\mathfrak{p}} \in \mathfrak{p}. \tag{5.5.14}$$

For $A \in \mathfrak{a}$, we have because of $Y \in \mathfrak{g}_\alpha$

$$[A, Y] = \alpha(A)Y, \tag{5.5.15}$$

and because of $[\mathfrak{k}, \mathfrak{p}] \subset \mathfrak{p}, [\mathfrak{p}, \mathfrak{p}] \subset \mathfrak{k}$, (5.5.14), (5.5.15) imply

$$[A, Y_\mathfrak{k}] = \alpha(A)Y_\mathfrak{p}, \tag{5.5.16}$$
$$[A, Y_\mathfrak{p}] = \alpha(A)Y_\mathfrak{k}. \tag{5.5.17}$$

If we had $Y_\mathfrak{p} = 0$, then by (5.5.16) also $Y_\mathfrak{k} = 0$, since α does not vanish on $\mathfrak{a}$, hence $Y = 0$. Likewise, $Y_\mathfrak{k}$ cannot vanish. By (5.5.17), $Y_\mathfrak{p}$ thus is contained in $\mathfrak{p}\backslash\mathfrak{a}$. Since (5.5.13) – (5.5.17) imply

$$[X, Y_\mathfrak{p}] = 0,$$

X and $Y_\mathfrak{p}$ are contained in some abelian, hence also in some maximal abelian subspace of $\mathfrak{p}$ different from $\mathfrak{a}$. Thus, X is singular. □

By Lemma 5.5.7, the singular elements of $\mathfrak{a}$ constitute the set

$$\mathfrak{a}_{\text{sing}} = \{X \in \mathfrak{a} : \exists \alpha \in \Lambda : \alpha(X) = 0\}.$$

$\mathfrak{a}_{\text{sing}}$ thus is the union of finitely many so called singular hyperplanes

$$\{X \in \mathfrak{a} : \alpha(X) = 0\} \quad \text{for } \alpha \in \Lambda.$$

Likewise, the set of regular elements of $\mathfrak{a}$ is

$$\mathfrak{a}_{\text{reg}} = \{X \in \mathfrak{a} : \forall \alpha \in \Lambda : \alpha(X) \neq 0\}.$$

The singular hyperplanes partition $\mathfrak{a}_{\text{reg}}$ into finitely many components which are called *Weyl chambers*.

For $\mathfrak{g} = \mathfrak{sl}(n, \mathbb{R}), \mathfrak{a} = \{\text{diag}\,(\lambda_1, \ldots, \lambda_n), \sum_{i=1}^n \lambda_i = 0\}$, we have

$$\mathfrak{a}_{\text{sing}} = \{\text{diag}\,(\lambda_1, \ldots, \lambda_n) : \exists i \neq j : \lambda_i = \lambda_j, \sum_{i=1}^n \lambda_i = 0\},$$

the space of those diagonal matrices whose entries are not all distinct. This follows from the fact that the roots are given by

$$\alpha_{ij}(\text{diag}\,(\lambda_1, \ldots, \lambda_n)) = \lambda_i - \lambda_j$$

as computed above.

One of the Weyl chambers then is

$$\mathfrak{a}^+ := \{\text{diag}\,(\lambda_1, \ldots, \lambda_n) : \lambda_1 > \lambda_2 > \ldots > \lambda_n, \Sigma\lambda_j = 0\}.$$

We call

$$\Lambda^+ := \{\alpha \in \Lambda : \forall A \in \mathfrak{a}^+ : \alpha(A) > 0\}$$

the space of positive roots (this obviously depends on the choice of $\mathfrak{a}^+$). In our case,

$$\Lambda^+ = \{\alpha_{ij} : i < j\}.$$

$\Lambda_b^+ := \{\alpha_{12}, \alpha_{23}, \ldots, \alpha_{n-1,n}\} \subset \Lambda^+$ then is a fundamental system of positive roots, meaning that each $\alpha \in \Lambda^+$ can be written as

$$\alpha = \sum_{i=1}^{n-1} s_i \alpha_{i,i+1}$$

with some $s_i \in \mathbb{N}$. For abbreviation, we put $\alpha_i := \alpha_{i,i+1}, i = 1, \ldots, n-1$.

The sets

$$\{A \in \mathfrak{a} : \alpha_{i_\nu}(A) > 0 \text{ for } \nu = 1, \ldots, r, \alpha_{i_\nu}(A) = 0 \text{ for } \nu = r+1, \ldots, n-1\},$$

where $\{i_1, \ldots, i_{n-1}\} = \{1, \ldots, n-1\}$, then are the r'-dimensional "walls" of the Weyl chamber $\mathfrak{a}^+$. The relation "is contained in the closure of" then defines an incidence relation on the space of all Weyl chambers and all Weyl chamber walls of all maximal abelian subspaces of $\mathfrak{p}$. This set with this incidence relation is an example of a so-called Tits building. Via the exponential map, we obtain a corresponding incidence structure on the set of all flats and all images of Weyl chamber walls through each given point of G/K.

We next introduce the Iwasawa decomposition of an element of $\mathrm{Sl}(n,\mathbb{R}) = G$. Let, as before,

$$K = \mathrm{SO}(n),$$

and moreover

$$A := \Big\{\mathrm{diag}\,(\lambda_1, \ldots, \lambda_n) : \lambda_i > 0 \text{ for } i = 1, \ldots, n, \prod_{i=1}^{n} \lambda_i = 1\Big\},$$

$$N := \Big\{\text{upper triangular matrices with entries 1 on the diagonal}\Big\}.$$

Theorem 5.5.2 (Iwasawa Decomposition). *We have*

$$G = KAN.$$

More precisely, for each $g \in G$ there exist unique $k \in K, a \in A, n \in N$ with

$$g = kan.$$

We first prove

Lemma 5.5.8. *For each $g \in \mathrm{Gl}(n,\mathbb{R})$, there exists a unique $h \in \mathrm{O}(n)$ with*

$$(hg)_{ij} = 0 \qquad \textit{for } i < j,$$
$$(hg)_{ii} > 0.$$

Proof. We denote the columns of g by $v_1, \dots, v_n$. The rows $r_1, \dots, r_n$ of $h \in \mathrm{O}(n)$ satisfying the assertions of the lemma must satisfy

(i) $r_1, \dots, r_n$ is an orthonormal basis of $\mathbb{R}^n$ (since $h \in \mathrm{O}(n)$).

(ii) $r_j \cdot v_i = 0$ for $i < j$. ("$\cdot$" here denotes the Euclidean scalar product).

(iii) $r_j \cdot v_j > 0$ for all j.

Conversely, if these three relations are satisfied, h has the desired properties.
We first determine r_n by the conditions

$$r_n \cdot r_n = 1, r_n \cdot v_n > 0, r_n \cdot v_i = 0 \quad \text{for } i = 1, \dots, n-1.$$

Since the columns of $\mathfrak{g}$, i.e. the v_i, are linearly independent, there indeed exists such an r_n. Assume now that we have iteratively determined $r_j, r_{j+1}, \dots, r_n$. Let W_j be the subspace of $\mathbb{R}^n$ spanned by $v_1, \dots, v_{j-2}, r_j, \dots, r_n$. W_j then has codimension 1 because of the properties of the vectors $r_j, \dots, r_n$. Then r_{j-1} has to be orthogonal to W_j and satisfies $r_{j-1} \cdot v_{j-1} > 0$ and $r_{j-1} \cdot r_{j-1} = 1$. There exists a unique such r_{j-1}. Iteratively, we obtain $r_1, \dots, r_n$, hence h. □

Proof of Theorem 5.5.2. By Lemma 5.5.8, there exist $k \in \mathrm{SO}(n)$, namely $k = h^{-1}$ from Lemma 5.5.8 (for $g \in \mathrm{Sl}(n, \mathbb{R})$, we get $h \in \mathrm{SO}(n)$) and an upper triangular matrix $m = (m_{ij})$ with positive diagonal entries with

$$g = km.$$

We put $\lambda_i := m_{ii}, n_{ii} = 1, n_{ij} = \frac{1}{\lambda_i} m_{ij}$ for $i \neq j$, $a = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, $n = (n_{ij})$ and obtain

$$g = km = kan.$$

The uniqueness of this decomposition is implied by the uniqueness statement of Lemma 5.5.8. □

5.6 Symmetric Spaces of Noncompact Type as Examples of Nonpositively Curved Riemannian Manifolds

We continue to study the symmetric space $M = \mathrm{Sl}(n, \mathbb{R})/\mathrm{SO}(n)$. It is complete (Corollary 5.3.1), nonpositively curved (Corollary 5.5.1), and simply connected (this follows from Lemma 5.5.2 since P is simply connected). Thus, the constructions at the end of §4.8 may be applied to M. (Actually, what follows will be valid for any symmetric space of noncompact type.) We continue to use the notations of §5.5, e.g. $G = \mathrm{Sl}(n, \mathbb{R})$, $K = \mathrm{SO}(n)$.

For $x \in M(\infty)$, let

$$G_x := \{g \in G : gx = x\}$$

be the isotropy group of x. G_x then is a subgroup of G. Let $\mathfrak{g}_x$ be the corresponding sub Lie algebra of $\mathfrak{g}$.

Theorem 5.6.1. *Let $x \in M(\infty)$, $p \in M$, $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ be the Cartan decomposition w.r.t. p. Let X be the element of $\mathfrak{p} \cong T_pM$ with*

$$c_{px}(t) = e^{tX}(p) \quad (= \exp_p tX).$$

Let $\mathfrak{a}$ be a maximal abelian subspace of $\mathfrak{p}$ with $X \in \mathfrak{a}$, and let

$$\mathfrak{g} = \mathfrak{g}_0 + \sum_{\alpha \in \Lambda} \mathfrak{g}_\alpha$$

be the root space decomposition of $\mathfrak{g}$ determined by $\mathfrak{a}$. Then

$$\mathfrak{g}_x = \mathfrak{g}_0 + \sum_{\alpha(X) \geq 0} \mathfrak{g}_\alpha. \tag{5.6.1}$$

Corollary 5.6.1. *Let B_1, B_2 be Weyl chambers or Weyl chamber walls with $B_1 \subset \bar{B}_2$. Let $X_1 \in B_1, X_2 \in B_2$ $\|X_1\| = \|X_2\| = 1$, $x_1, x_2 \in M(\infty)$ be the classes of asymptotic geodesic rays determined by X_1 and X_2, resp. Then*

$$G_{x_2} \subset G_{x_1}. \tag{5.6.2}$$

Conversely, $G_{x_2} \subset G_{x_1}$ implies $B_1 \subset \bar{B}_2$.

Proof. B_1 and B are contained in a common maximal abelian subspace $\mathfrak{a}$ of $\mathfrak{p}$. Let Λ be the set of roots of the root space decomposition of $\mathfrak{g}$ determined by $\mathfrak{a}$. Each $\alpha \in \Lambda$ which is nonnegative on B_2 then is nonnegative on B_1, too. Theorem 5.6.1 then implies the claim. □

By Corollary 5.6.1, the geometric relation $B_1 \subset \bar{B}_2$ defining the incidence relation for the Tits building may be replaced by the algebraic relation (5.6.2) between subgroups of G.

Proof of Theorem 5.6.1. For abbreviation, we put

$$c(t) := c_{px}(t).$$

Let $Y \in \mathfrak{g}$. We decompose

$$Y = Y_0 + \sum_{\alpha \in \Lambda} Y_\alpha \quad \text{with } Y_0 \in \mathfrak{g}_0, Y_\alpha \in \mathfrak{g}_\alpha$$

and put

$$Y(t) := \mathrm{Ad}\,(e^{-tX})Y = Y_0 + \sum_{\alpha\in\Lambda} e^{-t\alpha(X)} Y_\alpha, \tag{5.6.3}$$

by Lemma 5.5.6 (iv).

Then for all $s, t \in \mathbb{R}$

$$\begin{aligned} d(e^{sY}c(t), c(t)) &= d(e^{sY}e^{tX}(p), e^{tX}(p)) \\ &= d(e^{-tX}e^{sY}e^{tX}(p), p), \quad \text{since } e^{tX} \text{ is an isometry of } M \\ &= d(\mathrm{Ad}\,(e^{-tX})e^{sY}(p), p) \\ &= d(e^{sY(t)}(p), p). \end{aligned} \tag{5.6.4}$$

Let now

$$Y \in \mathfrak{g}_0 + \sum_{\alpha(X)\geq 0} \mathfrak{g}_\alpha.$$

We put

$$Y' := Y_0 + \sum_{\alpha(X)=0} Y_\alpha.$$

(5.6.3), (5.6.4) imply for each s

$$\lim_{t\to\infty} d^2(e^{sY}c(t), c(t)) = d^2(e^{sY'}(p), p).$$

Since by Theorem 4.8.2, $d^2(e^{sY}c(t), c(t))$ is convex in t, it has to be bounded for $t \geq 0$. Hence $e^{sY}c$ is asymptotic to c, hence

$$e^{sY} \in G_x \quad \text{for all } s,$$

hence

$$Y \in \mathfrak{g}_x.$$

Let conversely $Y \in \mathfrak{g}_x$. We write $Y = Y_1 + Y_2$ with $Y_1 := Y_0 + \sum_{\alpha(X)\geq 0} Y_\alpha, Y_2 := \sum_{\alpha(X)<0} Y_\alpha$. By what we have just proved, we obtain

$$Y_1 \in \mathfrak{g}_x,$$

hence also

$$Y_2 = Y - Y_1 \in \mathfrak{g}_x.$$

Therefore, for any fixed s,

$$d^2(e^{sY_2}c(t), c(t))$$

is bounded for $t \geq 0$. On the other hand (5.6.3), (5.6.4) imply

$$\lim_{t\to -\infty} d^2(e^{sY}c(t), c(t)) = 0.$$

Since this function is convex by Theorem 4.8.2, it then vanishes identically. We obtain

$$e^{sY_2}c(t) = c(t)$$

hence in particular

$$e^{sY_2}p = p,$$

hence $Y_2 \in \mathfrak{k}$. Therefore, letting θ_p denote the Cartan involution at p,

$$\begin{aligned} Y_2 = \theta_p(Y_2) \in \theta_p\left(\sum_{\alpha(X)<0} \mathfrak{g}_\alpha\right) & \quad \text{since } \theta_{p|\mathfrak{k}} = \mathrm{id}_{|\mathfrak{k}} \\ = \sum_{\alpha(X)>0} \mathfrak{g}_\alpha & \quad \text{by Lemma 5.5.6 (ii).} \end{aligned}$$

By definition of Y_2, this implies $\dim Y_2 = 0$, hence

$$Y = Y_1 \in \mathfrak{g}_0 + \sum_{\alpha(X)\geq 0} \mathfrak{g}_\alpha.$$

$\square$

Remark. The isotropy groups of any two points $p, q \in M$ are conjugate. If $q = gp$, then

$$G_q = gG_pg^{-1}.$$

(The isotropy group of $p \in M$ is by definition $G_p = \{g \in G : gp = p\}$.)

The isotropy groups of points in $M(\infty)$, however, are not necessarily conjugate as one sees from Theorem 5.6.1. However, there are only finitely many conjugacy classes.

Example. Let

$$X = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)$$

and let x be the element in $M(\infty)$ determined by X.

Then

$$\mathfrak{g}_x = \{A = (a_{ij})_{i,j=1,\ldots,n} \in \mathfrak{sl}(n, \mathbb{R}) \text{ with } a_{ij} = 0 \text{ for } \lambda_i < \lambda_j\}.$$

For example, if

$$\lambda_1 > \ldots > \lambda_n,$$

then $\mathfrak{g}_x$ is the space of upper triangular matrices.

Perspectives. For a differential geometric treatment of symmetric spaces of noncompact type, our sources and references are [68, 69], [12].

Let G/K be a symmetric space of noncompact type. A discrete subgroup Γ of G is called a lattice if the quotient $\Gamma\backslash G/K$ has finite volume in the induced locally symmetric metric. Here, Γ operates on G/K by isometries since the whole group G does. Γ may have fixed points so that the quotient need not be a manifold. Any such Γ, however, always contains a subgroup Γ' of finite index which is torsion free, i.e. operates without fixed points (i.e. there do not exist $\gamma \in \Gamma', \gamma \neq id$, and $z \in G/K$ with $\gamma z = z$), and the quotient $\Gamma\backslash G/K$ then is a manifold and a finite covering of $\Gamma\backslash G/K$. Therefore, one may usually assume w.l.o.g. that Γ itself has no fixed points, and we are hence going to do this for simplicity of discussion. A lattice Γ is called uniform or cocompact if the quotient is compact, nonuniform otherwise.

We now discuss the rigidity of such lattices.

For $G = \mathrm{Sl}(2, \mathbb{R}) \simeq \mathrm{SO}_0(2, 1)$ and $K = \mathrm{SO}(2)$, there exist continuous families of compact quotients, namely Riemann surfaces of a given genus $p \geq 2$. Thus, no rigidity result holds in this case. This, however, is a singular phenomenon.

The first rigidity result was obtained by Calabi and Vesentini[37] who showed that compact quotients of any irreducible Hermitian symmetric space of noncompact type other than $\mathrm{Sl}(2, \mathbb{R})/\mathrm{SO}(2)$ are infinitesimally, hence locally, rigid. They showed that the relevant cohomology group rising from the theory of Kodaira and Spencer vanishes in all these cases. Their result means that there do not exist nontrivial continuous families of uniform lattices in G/K other than $\mathrm{Sl}(2, \mathbb{R})/\mathrm{SO}(2)$.

Mostow[200] showed strong rigidity of compact quotients of irreducible symmetric spaces of noncompact type. This means that any two such lattices Γ, Γ' which are isomorphic as abstract groups are lattices in the same G and isomorphic as subgroups of G. Geometrically this means that the quotients $\Gamma\backslash G/K$ and $\Gamma'\backslash G/K$ are isometric. (Here, as always, they carry the Riemannian metric induced from the symmetric metric on G/K.)

Margulis[183] then showed superrigidity if rank $(G/K) \geq 2$. This essentially means that any homomorphism $\rho : \Gamma \to H$(Γ as above) extends to a homomorphism $G \to H$, if H, like G, is a simple noncompact algebraic group (defined over $\mathbb{R}$) and if $\rho(\Gamma)$ is Zariski dense, or that $\rho(\Gamma)$ is contained in a compact subgroup of H, if H is an algebraic subgroup of some $\mathrm{Sl}(n, \mathbb{Q}_p)$. Here, $\mathbb{Q}_p$ stands for the p-adic numbers. More generally and precisely, if G is a semisimple Lie group without compact factors with maximal compact subgroup K, rank $(G/K) \geq 2$, if Γ is an irreducible lattice in G (irreducibility means that no finite cover of the quotient $\Gamma\backslash G/K$ is a nontrivial product; this condition is nontrivial only in the case where G/K itself is not irreducible, i.e. a nontrivial product), and if H is a reductive algebraic group over $\mathbb{R}, \mathbb{C}$, or some $\mathbb{Q}_p$, then any homomorphism $\rho : \Gamma \to H$ with Zariski dense image (this means that $\rho(\Gamma)$ is not contained in a proper algebraic subgroup of H) factors through a homomorphism,

$$\begin{tikzcd} \Gamma \arrow[r, hook] \arrow[dr, "\rho"'] & G \times L \arrow[d, "\sigma"] \\ & H \end{tikzcd}$$

where L is a compact group. The results of Margulis and their proofs can be found in [274]. Important generalizations are given in [184]. Margulis also showed that superrigidity implies arithmeticity of a lattice Γ. This means that Γ is obtained from the prototype $\mathrm{Sl}(n, \mathbb{R})$ by certain finite algebraic operations, namely taking the intersection of $\mathrm{Sl}(n, \mathbb{Z})$ with Lie subgroups of $\mathrm{Sl}(n, \mathbb{R})$, applying surjective homomorphisms between Lie groups with compact kernels, passing to sublattices of finite index or taking finite extensions of lattices.

In the Perspectives on §7.7, we shall discuss how harmonic maps can be used to prove superrigidity.

Exercises for Chapter 5

1. Show that real projective space $\mathbb{RP}^n$ (cf. Exercise 3 of Chapter 1) can be obtained as the space of all (real) lines in $\mathbb{R}^{n+1}$. Show that $\mathbb{RP}^1$ is diffeomorphic to S^1. Compute the cohomology of $\mathbb{RP}^n$. Show that $\mathbb{RP}^n$ carries the structure of a symmetric space.

2. Similarly, define and discuss quaternionic projective space $\mathbb{HP}^n$ as the space of all quaternionic lines in quaternionic space $\mathbb{H}^{n+1}$. In particular, show that it is a symmetric space.

3. Determine all Killing fields on S^n.

4. Determine the Killing forms of the groups $\mathrm{Sl}(n, \mathbb{C}), \mathrm{Sp}(n, \mathbb{R}), \mathrm{SU}(n), \mathrm{U}(n)$.

5. Discuss the geometry of S^n by viewing it as the symmetric space $\mathrm{SO}(n+1)/\mathrm{SO}(n)$.

6. Show that $\mathbb{CP}^n = \mathrm{SU}(n+1)/\mathrm{S}(\mathrm{U}(1) \times \mathrm{U}(n))$. Compute the rank of $\mathbb{CP}^n$ as a symmetric space.

7. Determine the closed geodesics and compute the injectivity radius of the symmetric space $\mathbb{RP}^n$ (cf. Exercise 1).

Chapter 6

Morse Theory and Floer Homology

6.1 Preliminaries: Aims of Morse Theory

Let X be a complete Riemannian manifold, not necessarily of finite dimension.[1] We shall consider a smooth function f on X, i.e. $f \in C^\infty(X, \mathbb{R})$ (actually $f \in C^3(X, \mathbb{R})$ usually suffices). The essential feature of the theory of Morse and its generalizations is the relationship between the structure of the critical set of f,

$$C(f) := \{x \in X : df(x) = 0\}$$

(and the space of trajectories for the gradient flow of f) and the topology of X.

While some such relations can already be deduced for continuous, not necessarily smooth functions, certain deeper structures and more complete results only emerge if additional conditions are imposed onto f besides smoothness. Morse theory already yields very interesting results for functions on finite dimensional, compact Riemannian manifolds. However, it also applies in many infinite dimensional situations. For example, it can be used to show the existence of closed geodesics on compact Riemannian manifolds M by applying it to the energy functional on the space X of curves of Sobolev class $H^{1,2}$ in M, as we shall see in §6.11 below.

Let us first informally discuss the main features and concepts of the theory at some simple example. We consider a compact Riemannian manifold X diffeomorphic to the 2-sphere S^2, and we study smooth functions on X; more specifically let us look at two functions f_1, f_2 whose level set graphs are exhibited in the following figure,

[1] In this textbook, we do not systematically discuss infinite dimensional Riemannian manifolds. The essential point is that they are modeled on Hilbert instead of Euclidean spaces. At certain places, the constructions require a little more care than in the finite dimensional case, because compactness arguments are no longer available.

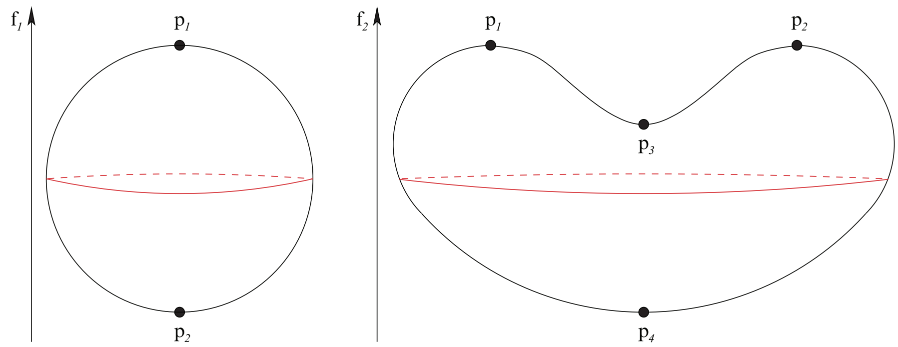

Figure 6.1.1:

with the vertical axis describing the value of the functions. The idea of Morse theory is to extract information about the global topology of X from the critical points of f, i.e. those $p \in X$ with

$$df(p) = 0.$$

Clearly, their number is not invariant; for f_1, we have two critical points, for f_2, four, as indicated in the figure. In order to describe the local geometry of the function more closely in the vicinity of a critical point, we assign a so-called Morse index $\mu(p)$ to each critical point p as the number of linearly independent directions on which the second derivative $d^2 f(p)$ is negative definite (this requires the assumption that that second derivative is nondegenerate, i.e. does not have the eigenvalue 0, at all critical points; if this assumption is satisfied we speak of a Morse function). Equivalently, this is the dimension of the unstable manifold $W^u(p)$. That unstable manifold is defined as follows: We look at the negative gradient flow of f, i.e. we consider the solutions of

$$\begin{aligned} &x : \mathbb{R} \to M, \\ &\dot{x}(t) = -\operatorname{grad} f(x(t)) \qquad \text{for all } t \in \mathbb{R}. \end{aligned}$$

It is at this point that the Riemannian metric of X enters, namely by defining the gradient of f as the vector field dual to the 1-form df. The flow lines $x(t)$ are curves of steepest descent for f. For $t \to \pm\infty$, each flow line $x(t)$ converges to some critical points $p = x(-\infty), q = x(\infty)$ of f, recalling that in our examples we are working on a compact manifold. The unstable manifold $W^u(p)$ of a critical point p then simply consists of all flow lines $x(t)$ with $x(-\infty) = p$, i.e. of those flow lines that emanate from p.

In our examples, we have for the Morse indices of the critical points of f_1

$$\mu_{f_1}(p_1) = 2, \quad \mu_{f_1}(p_2) = 0,$$

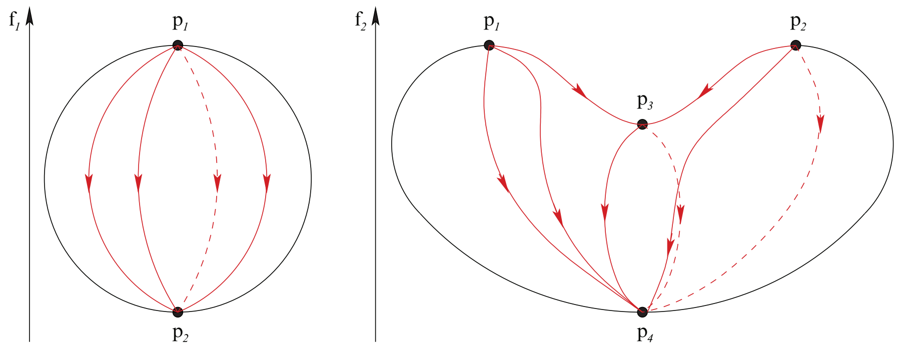

Figure 6.1.2:

and for f_2

$$\mu_{f_2}(p_1) = 2, \quad \mu_{f_2}(p_2) = 2, \quad \mu_{f_2}(p_3) = 1, \quad \mu_{f_2}(p_4) = 0,$$

as f_1 has a maximum point p_1 and a minimum p_2 as its only critical points whereas f_2 has two local maxima p_1, p_2, a saddle point p_3, and a minimum p_4. As we see from the examples, the unstable manifold $W^u(p)$ is topologically a cell (i.e. homeomorphic to an open ball) of dimension $\mu(p)$, and the manifold X is the union of the unstable manifolds of the critical points of the function. Thus, we get a decomposition of X into cells. In order to see the local effects of critical points, we can intersect $W^u(p)$ with a small ball around p and contract the boundary of that intersection to a point. We then obtain a

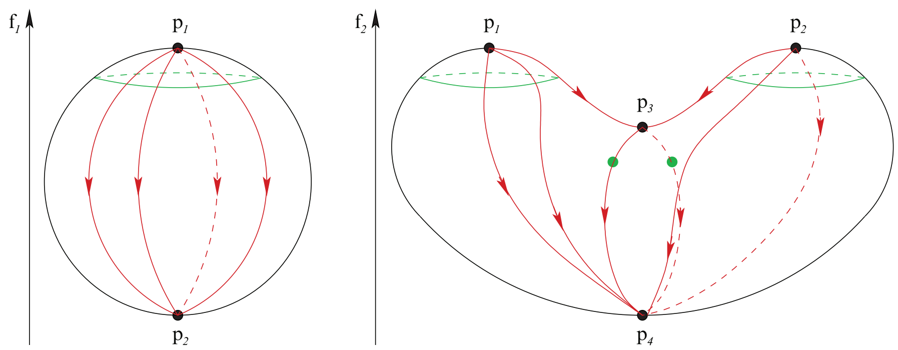

Figure 6.1.3:

pointed sphere $(S^{\mu(p)}, \text{ pt.})$ of dimension $\mu(p)$. These local constructions already yield an important topological invariant, namely the Euler characteristic $\chi(X)$, as the alternating sum of these dimensions,

$$\chi(X) = \sum_{p \text{ critical point of } f} (-1)^{\mu(p)} \mu(p).$$

We are introducing the signs $(-1)^{\mu(p)}$ here in order to get some cancellations between the contributions from the individual critical points. This issue is handled in more generality by the introduction of the boundary operator ∂. From the point of view explored by Floer, we consider pairs (p, q) of critical points with $\mu(q) = \mu(p) - 1$, i.e. of index difference 1. We then count the number of trajectories from p to q modulo 2 (or, more generally, with associated signs as will be discussed later in this chapter):

$$\partial p = \sum_{\substack{q \text{ crit. pt. of } f \\ \mu(q)=\mu(p)-1}} (\#\{\text{flow lines from } p \text{ to } q\} \mod 2)\, q.$$

In this way, we get an operator from $C_*(f, \mathbb{Z}_2)$, the vector space over $\mathbb{Z}_2$ generated by the critical points of f, to itself. The important point then is to show that

$$\partial \circ \partial = 0.$$

On this basis, one can define the homology groups

$$H_k(X, f, \mathbb{Z}_2) := \text{kernel of } \partial \text{ on } C_k(f, \mathbb{Z}_2)/\text{image of } \partial \text{ from } C_{k+1}(f, \mathbb{Z}_2),$$

where $C_k(f, \mathbb{Z}_2)$ is generated by the critical points of Morse index k. (Because of the relation $\partial \circ \partial = 0$, the image of ∂ from $C_{k+1}(f, \mathbb{Z}_2)$ is always contained in the kernel of ∂ on $C_k(f, \mathbb{Z}_2)$.) We return to our examples: In the figure, we now only indicate flow lines between critical points of index difference 1.

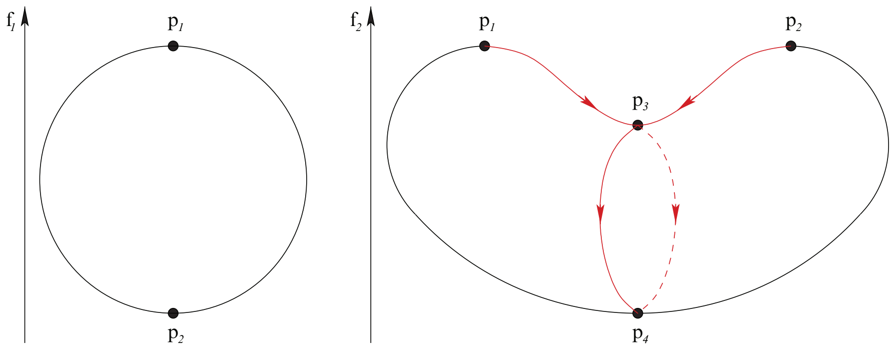

Figure 6.1.4:

For f_1, there are no pairs of critical points of index difference 1 at all. Denoting the restriction of ∂ to $C_k(f, \mathbb{Z}_2)$ by ∂_k, we then have

$$\ker \partial_2 = \{p_1\},$$
$$\ker \partial_0 = \{p_0\},$$

while ∂_1 is the trivial operator as $C_1(f_1, \mathbb{Z}_2)$ is 0. All images are likewise trivial, and so

$$\begin{aligned} H_2(X, f_1, \mathbb{Z}_2) &= \mathbb{Z}_2, \\ H_1(X, f_1, \mathbb{Z}_2) &= 0, \\ H_0(X, f_1, \mathbb{Z}_2) &= \mathbb{Z}_2. \end{aligned}$$

Putting

$$b_k := \dim_{\mathbb{Z}_2} H_k(X, f, \mathbb{Z}_2) \qquad \text{(Betti numbers)},$$

in particular we recover the Euler characteristic as

$$\chi(X) = \sum_j (-1)^j b_j.$$

Let us now look at f_2. Here we have

$$\begin{aligned} &\partial_2 p_1 = \partial_2 p_2 = p_3, \quad \text{hence } \partial_2(p_1 + p_2) = 2p_3 = 0, \\ &\partial_1 p_3 = 2p_4 = 0 \quad \text{(since we are computing mod 2)}, \\ &\partial_0 p_4 = 0. \end{aligned}$$

Thus

$$\begin{aligned} H_2(X, f_2, \mathbb{Z}_2) &= \ker \partial_2 = \mathbb{Z}_2, \\ H_1(X, f_2, \mathbb{Z}_2) &= \ker \partial_1 / \operatorname{image} \partial_2 = 0, \\ H_0(X, f_2, \mathbb{Z}_2) &= \ker \partial_0 / \operatorname{image} \partial_1 = \mathbb{Z}_2. \end{aligned}$$

Thus, the homology groups, and therefore also the Betti numbers are the same for either function. This is the basic fact of Morse theory, and we also see that this equality arises from cancellations between critical points achieved by the boundary operator.

This will be made more rigorous in §§6.3 – 6.10.

As already mentioned, there is one other aspect to Morse theory, namely that it is not restricted to finite dimensional manifolds. While some of the considerations in this chapter will apply in a general setting, here we can only present an application that does not need elaborate features of Morse theory but only an existence result for unstable critical points in an infinite dimensional setting. This will be prepared in §6.2 and carried out in §6.11.

6.2 Compactness: The Palais-Smale Condition and the Existence of Saddle Points

On a compact manifold, any continuous function assumes its minimum. It may have more than one local minimum, however. If a differentiable function on a compact manifold has two local minima, then it also has another critical point which is not a strict local minimum. These rather elementary results, however, in general cease to hold on noncompact spaces, for example infinite dimensional ones. The attempt to isolate conditions that permit an extension of these results to general, not necessarily compact situations is the starting point of the modern calculus of variations. For the existence of a minimum, one usually imposes certain generalized convexity conditions while for the existence of other critical points, one needs the so-called Palais-Smale condition (PS).

Definition 6.2.1. $f \in C^1(X, \mathbb{R})$ satisfies condition (PS) if every sequence $(x_n)_{n\in\mathbb{N}}$ with

(i) $|f(x_n)|$ bounded,

(ii) $\|df(x_n)\| \to 0$ for $n \to \infty$,

contains a convergent subsequence.

Obviously, (PS) is automatically satisfied if X is compact. It is also satisfied if f is proper, i.e. if for every $c \in \mathbb{R}$

$$\{x \in X : |f(x)| \leq c\}$$

is compact. However, (PS) is more general than that and we shall see in the sequel (see §6.11 below) that it holds for example for the energy functional on the space of closed curves of Sobolev class $H^{1,2}$ on a compact Riemannian manifold M.

For the sake of illustration, we shall now demonstrate the following result:

Proposition 6.2.1. *Suppose $f \in C^1(X, \mathbb{R})$ satisfies (PS) and has two strict relative minima $x_1, x_2 \in X$. Then there exists another critical point x_3 of f (i.e. $df(x_3) = 0$) with*

$$f(x_3) = \kappa := \inf_{\gamma\in\Gamma} \max_{x\in\gamma} f(x) > \max\{f(x_1), f(x_2)\} \tag{6.2.1}$$

with $\Gamma := \{\gamma \in C^0([0,1], X) : \gamma(0) = x_1, \gamma(1) = x_2\}$, the set of all paths connecting x_1 and x_2. (x_3 is called a saddle point for f.)

We assume also that solutions of the negative gradient flow of f,

$$\begin{aligned} &\varphi : X \times \mathbb{R} \to X, \\ &\frac{\partial}{\partial t}\varphi(x,t) = -\operatorname{grad} f(\varphi(x,t)), \\ &\varphi(x,0) = x \end{aligned} \tag{6.2.2}$$

exist for all $x \in X$ and $0 \leq t \leq \varepsilon$, for some $\varepsilon > 0$. ($\operatorname{grad} f$ is the gradient of f, see (2.1.14); it is the vector field dual to the 1-form df.)

Proof of Proposition 6.2.1. Since x_1 and x_2 are strict relative minima of f,

$$\begin{gathered}\exists\, \delta_0 > 0 \,\forall\, \delta \text{ with } 0 < \delta \leq \delta_0 \;\exists\, \varepsilon > 0 \,\forall\, x \text{ with } \|x - x_i\| = \delta : \\ f(x) \geq f(x_i) + \varepsilon \quad \text{for } i = 1, 2.\end{gathered}$$

Consequently,

$$\exists\, \varepsilon_0 > 0 \,\forall\, \gamma \in \Gamma \,\exists\, \tau \in (0,1) : f(\gamma(\tau)) \geq \max(f(x_1), f(x_2)) + \varepsilon_0.$$

This implies

$$\kappa > \max(f(x_1), f(x_2)). \tag{6.2.3}$$

We want to show that

$$f^\kappa := \{x \in \mathbb{R}^n : f(x) = \kappa\}$$

contains a point x_3 with

$$df(x_3) = 0. \tag{6.2.4}$$

If this is not the case, by (PS) there exist $\eta > 0$ and $\alpha > 0$ with

$$\|df(x)\| \geq \alpha, \tag{6.2.5}$$

whenever $\kappa - \eta \leq f(x) \leq \kappa + \eta$.
Namely, otherwise, we find a sequence $(x_n)_{n \in \mathbb{N}} \subset X$ with $f(x_n) \to \kappa$ and $df(x_n) \to 0$ as $n \to \infty$, hence by (PS) a limit point x_3 that satisfies $f(x_3) = \kappa$, $df(x_3) = 0$ as f is of class C^1.

In particular,

$$f(x_1), f(x_2) < \kappa - \eta, \tag{6.2.6}$$

since $df(x_1) = 0 = df(x_2)$. Consequently we may find arbitrarily small $\eta > 0$ such that for all $\gamma \in \Gamma$ with $\max f(\gamma(\tau)) \leq \kappa + \eta$:

$$\forall \tau \in [0,1] : \text{ either } f(\gamma(\tau)) \leq \kappa - \eta \text{ or } \|df(\gamma(\tau))\| \geq \alpha. \tag{6.2.7}$$

We let $\varphi(x,t)$ be the solution of (6.2.2) for $0 \leq t \leq \varepsilon$.

We select $\eta > 0$ satisfying (6.2.7) and $\gamma \in \Gamma$ with

$$\max_{\tau \in [0,1]} f(\gamma(\tau)) \leq \kappa + \eta. \tag{6.2.8}$$

Then

$$\begin{aligned}\frac{d}{dt} f(\varphi(\gamma(\tau), t)) &= -\langle (df)(\varphi(\gamma(\tau), t), \operatorname{grad} f(\varphi(\gamma(\tau), t)\rangle \\ &= -\|df(\varphi(\gamma(\tau), t))\|^2 \leq 0.\end{aligned} \tag{6.2.9}$$

Therefore

$$\max f(\varphi(\gamma(\tau),t)) \leq \max f(\gamma(\tau)) \leq \kappa + \eta. \tag{6.2.10}$$

Since $\operatorname{grad} f(x_i) = 0, i = 1, 2$, because x_1, x_2 are critical points of f, also $\varphi(x_i, t) = x_i$ for $i = 1, 2$ and all $t \in \mathbb{R}$, hence

$$\varphi(\gamma(\cdot), t) \in \Gamma.$$

(6.2.9), (6.2.6), (6.2.7) and (6.2.2) then imply

$$\frac{d}{dt} f(\varphi(\gamma(\tau),t)) \leq -\frac{\alpha^2}{4}, \text{whenever } f(\varphi(\gamma(\tau),t) > \kappa - \eta. \tag{6.2.11}$$

We may assume that the above $\eta > 0$ satisfies

$$\frac{8\eta}{\alpha^2} \leq \varepsilon.$$

Then the negative gradient flow exists at least up to $t = \frac{8\eta}{\alpha^2}$. (6.2.10) and (6.2.11), however, imply that for $t_0 = \frac{8\eta}{a^2}$, we have

$$f(\varphi(\gamma(\tau), t_0)) \leq \kappa - \eta \quad \text{for all } \tau \in [0,1].$$

Since $\varphi(\gamma(\cdot), t_0) \in \Gamma$, this contradicts the definition of κ. We conclude that there has to exist some x_3 with $f(x_3) = \kappa$ and $df(x_3) = 0$. $\square$

The issue of the existence of the negative gradient flow for f will be discussed in the next section. Essentially the same argument as in the proof of Proposition 6.2.1 will be presented once more in Theorem 6.11.3 below.

Perspectives. The role of the Palais-Smale condition in the calculus of variations is treated in [149]. A thorough treatment of many further examples can be found in [243] and [42]. A recent work on Morse homology in an infinite dimensional context is Abbondandolo, Majer[1].

6.3 Local Analysis: Nondegeneracy of Critical Points, Morse Lemma, Stable and Unstable Manifolds

The next condition provides a nontrivial restriction already on compact manifolds.

Definition 6.3.1. $f \in C^2(X, \mathbb{R})$ is called a Morse function if for every $x_0 \in C(f)$, the Hessian $d^2 f(x_0)$ is nondegenerate. (This means that the continuous linear operator

$$A : T_{x_0}X \to T^*_{x_0}X$$

defined by

$$(A_u)(v) = d^2 f(x_0)(u, v) \quad \text{for } u, v \in T_{x_0}X$$

is bijective.) Moreover, we let

$$V^- \subset T_{x_0}X$$

be the subspace spanned by eigenvectors of (the bounded, symmetric, bilinear form) $d^2 f(x_0)$ with negative eigenvalues and call

$$\mu(x_0) := \dim V^-$$

the Morse index of $x_0 \in C(f)$. For $k \in \mathbb{N}$, we let

$$C_k(f) := \{x \in C(f) : \mu(x) = k\}$$

be the set of critical points of f of Morse index k.

The Morse index $\mu(x_0)$ may be infinite. In fact, however, for Morse theory in the sense of Floer one only needs finite *relative* Morse indices. Before we can explain what this means we need to define the stable and unstable manifolds of the negative gradient flow of f at x_0.

The first point to observe here is that the preceding notion of nondegeneracy of a critical point does not depend on the choice of coordinates. Indeed, if we change coordinates via

$$x = \xi(y), \quad \text{for some local diffeomorphism } \xi,$$

then, computing derivatives now w.r.t. y, and putting $y_0 = \xi^{-1}(x_0)$,

$$d^2(f \circ \xi)(y_0)(u, v) = (d^2 f)(\xi(y_0))(d\xi(y_0)u, d\xi(y_0)v) \quad \text{for any } u, v,$$

if

$$df(x_0) = 0.$$

Since $d\xi(y_0)$ is an isomorphism by assumption, we see that

$$d^2(f \circ \xi)(y_0)$$

has the same index as

$$d^2 f(x_0).$$

The negative gradient flow for f is defined as the solution of

$$\begin{aligned} &\phi : X \times \mathbb{R} \to X, \\ &\frac{\partial}{\partial t}\phi(x, t) = -\operatorname{grad} f(\phi(x, t)), \\ &\phi(x, 0) = x. \end{aligned} \tag{6.3.1}$$

Here, grad f of course is the gradient of f for all $x \in X$, defined with the help of some Riemannian metric on X, see (2.1.14).

The theorem of Picard-Lindelöf yields the local existence of this flow (see Lemma 1.9.1), i.e. for every $x \in X$, there exists some $\varepsilon > 0$ such that $\phi(x,t)$ exists for $-\varepsilon < t < \varepsilon$. This holds because we assume $f \in C^2(X,\mathbb{R})$ so that grad f satisfies a local Lipschitz condition as required for the Picard-Lindelöf theorem. We shall assume in the sequel that this flow exists globally, i.e. that ϕ is defined on all of $X \times \mathbb{R}$. In order to assure this, we might for example assume that $d^2 f(x)$ has uniformly bounded norm on X.

(6.3.1) is an example of a flow of the type

$$\begin{aligned} \phi : X \times \mathbb{R} &\to X, \\ \frac{\partial}{\partial t}\phi &= V(\phi(x,t)), \\ \phi(x,0) &= x, \end{aligned}$$

for some vector field V on X which we assume bounded for the present exposition as discussed in §1.9. The preceding system is autonomous in the sense that V does not depend explicitly on the "time" parameter t (only implicitly through its dependence on ϕ). Therefore, the flow satisfies the group property

$$\phi(x, t_1 + t_2) = \phi(\phi(x,t_1), t_2) \quad \text{for all } t_1, t_2 \in \mathbb{R} \text{ (see Theorem 1.9.1).}$$

In particular, for every $x \in X$, the flow line or orbit $\gamma_x := \{\phi(x,t) : t \in \mathbb{R}\}$ through x is flow invariant in the sense that for $y \in \gamma_x$, $t \in \mathbb{R}$

$$\phi(y,t) \in \gamma_x.$$

Also, for every $t \in \mathbb{R}$, $\phi(\cdot,t) : X \to X$ is a diffeomorphism of X onto its image (see Theorem 1.9.1).

As a preparation for our treatment of Morse theory, in the present section we shall perform a local analysis of the flow (6.3.1) near a critical point x_0 of f, i.e. grad $f(x_0) = 0$.

Definition 6.3.2. The stable and unstable manifolds at x_0 of the flow ϕ are defined as

$$\begin{aligned} W^s(x_0) &:= \big\{y \in X : \lim_{t \to +\infty} \phi(y,t) = x_0\big\}, \\ W^u(x_0) &:= \big\{y \in X : \lim_{t \to -\infty} \phi(y,t) = x_0\big\}. \end{aligned}$$

Of course, the question arises whether $W^s(x_0)$ and $W^u(x_0)$ are indeed manifolds.

In order to understand the stable and unstable manifolds of a critical point, it is useful to transform f locally near a critical point x_0 into some simpler, so-called "normal" form, by comparing f with a local diffeomorphism. Namely, we want to find a local diffeomorphism

$$x = \xi(y),$$

with
$$x_0 = \xi(0) \quad \text{for simplicity}$$
such that
$$f(\xi(y)) = f(x_0) + \frac{1}{2} d^2 f(x_0)(y, y). \tag{6.3.2}$$
In other words, we want to transform f into a quadratic polynomial. Having achieved this, we may then study the negative gradient flow in those coordinates w.r.t. the Euclidean metric. It turns out that the qualitative behaviour of this flow in the vicinity of 0 is the same as the one of the original flow in the vicinity of $x_0 = \xi(0)$.

That such a local transformation is possible is the content of the Morse-Palais-Lemma:

Lemma 6.3.1. *Let B be a Banach space, U an open neighborhood of $x_0 \in B$, $f \in C^{k+2}(U, \mathbb{R})$ for some $k \geq 1$, with a nondegenerate critical point at x_0. Then there exist a neighborhood V of $0 \in B$ and a diffeomorphism*
$$\xi : V \to \xi(V) \subset U$$
of class C^k with $\xi(0) = x_0$ satisfying (6.3.2) in V. In particular, nondegenerate critical points of a function f of class C^3 are isolated.

Proof. We may assume $x_0 = 0$, $f(0) = 0$ for simplicity of notation.

We want to find a flow
$$\varphi : V \times [0, 1] \to B,$$
with
$$\varphi(y, 0) = y, \tag{6.3.3}$$
$$f(\varphi(y, 1)) = \frac{1}{2} d^2 f(0)(y, y) \quad \text{for all } y \in V. \tag{6.3.4}$$
$\xi(y) := \varphi(y, 1)$ then has the required property. We shall construct $\varphi(y, t)$ so that with
$$\eta(y, t) := t f(y) + \frac{1}{2}(1 - t) d^2 f(0)(y, y),$$
we have
$$\frac{\partial}{\partial t} \eta(\varphi(y, t), t) = 0, \tag{6.3.5}$$
implying
$$\begin{aligned} f(\varphi(y, 1)) &= \eta(\varphi(y, 1), 1) \\ &= \eta(\varphi(y, 0), 0) \\ &= \frac{1}{2} d^2 f(0)(y, y) \end{aligned}$$

as required. (6.3.5) means

$$\begin{aligned} 0 =& f(\varphi(y,t)) + t\, df(\varphi(y,t))\frac{\partial}{\partial t}\varphi(y,t) \\ &- \frac{1}{2}d^2 f(0)(\varphi(y,t),\varphi(y,t)) + (1-t)d^2 f(0)(\varphi(y,t),\frac{\partial}{\partial t}\varphi(y,t)). \end{aligned} \tag{6.3.6}$$

Now by Taylor expansion, using $df(0) = 0$,

$$\begin{aligned} f(x) &= \int_0^1 (1-\tau)d^2 f(\tau x)(x,x)\, d\tau, \\ df(x) &= \int_0^1 d^2 f(\tau x)x\, d\tau. \end{aligned}$$

Inserting this into (6.3.6), with $x = \varphi(y,t)$, we observe that we have a common factor $\varphi(y,t)$ in all terms. Thus, abbreviating

$$\begin{aligned} T_0(x) &:= -\frac{1}{2}d^2 f(0) + \int_0^1 (1-\tau)d^2 f(\tau x)\, d\tau, \\ T_1(x,t) &:= \quad d^2 f(0) + t\int_0^1 (d^2 f(\tau x) - d^2 f(0))\, d\tau, \end{aligned}$$

(6.3.6) would follow from

$$0 = T_0(\varphi(y,t))\varphi(y,t) + T_1(\varphi(y,t),t)\frac{\partial}{\partial t}\varphi(y,t). \tag{6.3.7}$$

Here, we have deleted the common factor $\varphi(y,t)$, meaning that we now consider e.g. $d^2 f(0)$ as a linear operator on B.

Since we assume that $d^2 f(0)$ is nondegenerate, $d^2 f(0)$ is invertible as a linear operator, and so then is $T_1(x,t)$ for x in some neighborhood W of 0 and all $t \in [0,1]$.

Therefore,

$$-T_1(\varphi(y,t),t)^{-1} \circ T_0(\varphi(y,t))\varphi(y,t)$$

exists and is bounded if $\varphi(y,t)$ stays in W. Therefore, a solution of (6.3.7), i.e. of

$$\frac{\partial}{\partial t}\varphi(y,t) = -T_1(\varphi(y,t)^{-1} \circ T_0(\varphi(y,t))\varphi(y,t), \tag{6.3.8}$$

stays in W for all $t \in [0,1]$ if $\varphi(y,0)$ is contained in some possibly smaller neighborhood V of 0. The existence of such a solution then is a consequence of the theorem of Picard-Lindelöf for ODEs in Banach spaces. This completes the proof. $\square$

Remark. The preceding lemma plays a fundamental role in the classical expositions of Morse theory. The reason is that it allows to describe the change of topology in the vicinity of a critical point x_0 of f of the sublevel sets

$$f_\lambda := \{y \in X : f(y) \le \lambda\}$$

as λ decreases from $f(x_0)+\varepsilon$ to $f(x_0)-\varepsilon$, for $\varepsilon > 0$.

The gradient flow w.r.t. the Euclidean metric for f of the form (6.3.2) now is very easy to describe. Assuming w.l.o.g. $f(x_0) = 0$, we are thus in the situation of

$$g(y) = \frac{1}{2}B(y,y),$$

where $B(\cdot,\cdot)$ is a bounded symmetric quadratic form on a Hilbert space H. Denoting the scalar product on H by $\langle\cdot,\cdot\rangle$, B corresponds to a selfadjoint bounded linear operator

$$L : H \to H$$

via

$$\langle L(u), v\rangle = B(u,v)$$

by the Riesz representation theorem, and the negative gradient flow for g then is the solution of

$$\begin{aligned} \frac{\partial}{\partial t}\phi(y,t) &= -L\phi(y,t), \\ \phi(y,0) &= y. \end{aligned}$$

If v is an eigenvector of L with eigenvalue λ, then

$$\phi(v,t) = e^{-\lambda t}v.$$

Thus, the flow exponentially contracts the directions corresponding to positive eigenvalues, and these are thus stable directions, while the ones corresponding to negative eigenvalues are expanded, hence unstable.

Let us describe the possible geometric pictures in two dimensions. If we have one positive and one negative eigenvalue, we have a so-called saddle, and the flow lines in the vicinity of our critical point look like:
If we have two negative eigenvalues, hence two unstable directions, we have a node. If the two eigenvalues are equal, all directions are expanded at the same speed, and the local picture is
If they are different, we may get the following picture, if the one of largest absolute value corresponds to the horizontal direction.
The situations of Figures 6.3.2 and 6.3.3 are topologically conjugate, but not differentiably. However, if we want to preserve conditions involving derivatives like the transversality condition imposed in the next section, we may only perform differentiable transformations of the local picture. It turns out that the situation of Figure 6.3.1 is better behaved in that sense.

Namely, the main point of the remainder of this section is to show that the decomposition into stable and unstable manifolds always has the same qualitative features in the differentiable sense as in our model situation of a linear system of ODEs (although the situation for a general system is conjugate to the one for the

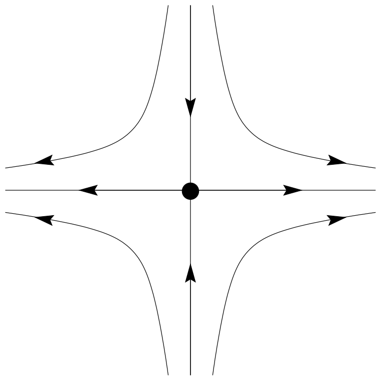

Figure 6.3.1: The horizontal axis is the unstable, the vertical one the stable manifold.

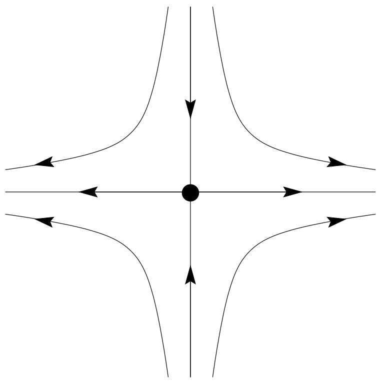

Figure 6.3.2:

linearized one only in the topological sense, as stated by the Hartmann-Grobman-Theorem). All these results will depend crucially on the nondegeneracy condition near a critical point, and the analysis definitely becomes much more complicated without such a condition. In particular, even the qualitative topological features may then cease to be stable against small perturbations. While many aspects can still be successfully addressed in the context of the theory of Conley, we shall confine ourselves to the nondegenerate case.

By Taylor expansion, the general case may locally be considered as a perturbation

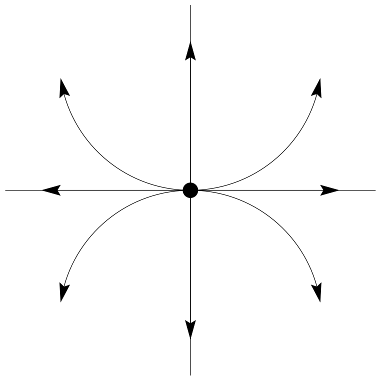

Figure 6.3.3:

of the linear equation just considered. Namely, we study

$$\begin{aligned} \frac{\partial}{\partial t}\phi(y,t) &= -L\phi(y,t) + \eta(\phi(y,t)), \\ \phi(y,0) &= y, \end{aligned} \tag{6.3.9}$$

in some neighborhood U of 0, where $\eta : H \to H$ satisfies

$$\begin{aligned} \eta(0) &= 0, \\ \|\eta(x) - \eta(y)\| &\le \delta(\varepsilon)\|x - y\| \end{aligned} \tag{6.3.10}$$

for $\|x\|, \|y\| < \varepsilon$, with $\delta(\varepsilon)$ a continuous monotonically increasing function of $\varepsilon \in [0,\infty)$ with $\delta(0) = 0$. The local unstable and stable manifolds of 0 then are defined as

$$\begin{aligned} &W^u(0,U) = \\ &\qquad \big\{x \in U : \phi(x,t) \text{ exists and is contained in } U \text{ for all } t \le 0, \lim_{t\to-\infty}\phi(x,t) = 0\big\}, \\ &W^s(0,U) = \\ &\qquad \big\{x \in U : \phi(x,t) \text{ exists and is contained in } U \text{ for all } t \ge 0, \lim_{t\to+\infty}\phi(x,t) = 0\big\}. \end{aligned}$$

We assume that the bounded linear selfadjoint operator L is nondegenerate, i.e. that 0 is not contained in the spectrum of L. As L is selfadjoint, the spectrum is real. H then is the orthogonal sum of subspaces H_+, H_- invariant under L for which $L_{|H_+}$ has positive, $L_{|H_-}$ negative spectrum, and corresponding projections

$$\begin{aligned} P_\pm &: H \to H_\pm, \\ P_+ + P_- &= \mathrm{Id}. \end{aligned}$$

Since L is bounded, we may find constants $c_0, \gamma > 0$ such that

$$\begin{aligned} \|e^{-Lt}P_+\| &\le c_0 e^{-\gamma t} && \text{for } t \ge 0, \\ \|e^{-Lt}P_-\| &\le c_0 e^{\gamma t} && \text{for } t \le 0. \end{aligned} \tag{6.3.11}$$

Let now $y(t) = \phi(x,t)$ be a solution of (6.3.9) for $t \ge 0$. We have for any $\tau \in [0, \infty)$,

$$y(t) = e^{-L(t-\tau)}y(\tau) + \int_\tau^t e^{-L(t-s)}\eta(y(s))\,ds, \tag{6.3.12}$$

hence also

$$P_\pm y(t) = e^{-L(t-\tau)}P_\pm y(\tau) + \int_\tau^t e^{-L(t-s)}P_\pm \eta(y(s))\,ds. \tag{6.3.13}$$

If we assume that $y(t)$ is bounded for $t \ge 0$, then by (6.3.11)

$$\lim_{\tau\to\infty} e^{-L(t-\tau)}P_- y(\tau) = 0, \tag{6.3.14}$$

and hence such a solution $y(t)$ that is bounded for $t \ge 0$ can be represented as

$$\begin{aligned} y(t) &= P_+y(t) + P_-y(t) \\ &= e^{-Lt}P_+x \\ &\quad + \int_0^t e^{-L(t-s)}P_+\eta(y(s))\,ds - \int_t^\infty e^{-L(t-s)}P_-\eta(y(s))\,ds, \ \text{with } x = y(0) \end{aligned} \tag{6.3.15}$$

(putting $\tau = 0$ in $(6.3.13)_+$ and $\tau = \infty$ in $(6.3.13)_-$). Conversely, any solution of (6.3.15), bounded for $t \ge 0$, satisfies (6.3.12), hence (6.3.9). For a solution that is bounded for $t \le 0$, we analogously get the representation

$$y(t) = e^{-Lt}P_-x - \int_t^0 e^{-L(t-s)}P_-\eta(y(s))\,ds + \int_{-\infty}^t e^{-L(t-s)}P_+\eta(y(s))\,ds.$$

Theorem 6.3.1. *Let $\phi(y,t)$ satisfy (6.3.9), with a bounded linear nondegenerate selfadjoint operator L and η satisfying (6.3.10). Then we may find a neighborhood U of 0 such that $W^s(0,U)$ ($W^u(0,U)$) is a Lipschitz graph over $P_+H \cap U$ ($P_-H \cap U$), tangent to P_+H (P_-H) at 0. If η is of class C^k in U, so are $W^s(0,U)$ and $W^u(0,U)$.*

Proof. We consider, for $x \in P_+H$,

$$T(y,x)(t) := e^{-Lt}x + \int_0^t e^{-L(t-s)}P_+\eta(y(s))\,ds - \int_t^\infty e^{-L(t-s))}P_-\eta(y(s))\,ds. \tag{6.3.16}$$

From (6.3.15) we see that we need to find fixed points of T, i.e.

$$y(t) = T(y,x)(t). \tag{6.3.17}$$

In order to apply the Banach fixed point theorem, we first need to identify an appropriate space on which $T(\cdot, x)$ operates as a contraction. For that purpose, we consider, for $0 < \lambda < \gamma$, $\varepsilon > 0$, the space

$$M_\lambda(\varepsilon) := \left\{ y(t) : \|y\|_{\exp,\lambda} := \sup_{t\geq 0} e^{\lambda t}\|y(t)\| \leq \varepsilon \right\}. \tag{6.3.18}$$

$M_\lambda(\varepsilon)$ is a complete normed space. We fix λ, e.g. $\lambda = \frac{\gamma}{2}$, in the sequel. Because of (6.3.10), (6.3.11), we have for $y \in M_\lambda(\varepsilon)$

$$\begin{aligned}
\|T(y,x)(t)\| \leq c_0 e^{-\gamma t}\|x\| & \\
&+ c_0\delta(\varepsilon)\left(\int_0^t e^{-\gamma(t-s)}\|y(s)\|\,ds + \int_t^\infty e^{\gamma(t-s)}\|y(s)\|\,ds\right)\\
&\leq c_0 e^{-\gamma t}\|x\|\\
+ c_0\delta(\varepsilon)\Bigg(\sup_{0\leq s\leq t} e^{\lambda s}\|y(s)\| \int_0^t & e^{-\gamma(t-s)}e^{-\lambda s}\,ds + \sup_{t\leq s\leq\infty} e^{\lambda s}\|y(s)\|\int_t^\infty e^{\gamma(t-s)}e^{-\lambda s}\,ds\Bigg).
\end{aligned} \tag{6.3.19}$$

Now since

$$\begin{aligned}
\int_0^t e^{-\gamma(t-s)}e^{-\lambda s}\,ds &= e^{-\gamma t}\frac{1}{\gamma-\lambda}\left(e^{(\gamma-\lambda)t} - 1\right) \leq \frac{1}{\gamma-\lambda}e^{-\lambda t},\\
\int_t^\infty e^{\gamma(t-s)}e^{-\lambda s}\,ds &= e^{\gamma t}\frac{1}{\gamma+\lambda}e^{-(\gamma+\lambda)t} = \frac{1}{\gamma+\lambda}e^{-\lambda t},
\end{aligned}$$

(6.3.19) implies

$$\|T(y,x)(t)\| \leq c_0 e^{-\gamma t}\|x\| + \frac{2c_0\delta(\varepsilon)}{\gamma-\lambda}e^{-\lambda t}\|y\|_{\exp,\lambda}. \tag{6.3.20}$$

Similarly, for $y_1, y_2 \in M_\lambda(\varepsilon)$

$$\|T(y_1,x)(t) - T(y_2,x)(t)\| \leq \frac{4c_0\delta(\varepsilon)}{\gamma-\lambda}e^{-\lambda t}\|y_1 - y_2\|_{\exp,\lambda}. \tag{6.3.21}$$

Because of our assumptions on $\delta(\varepsilon)$ (see (6.3.10), we may choose ε so small that

$$\frac{4c_0}{\gamma-\lambda}\delta(\varepsilon) \leq \frac{1}{2}. \tag{6.3.22}$$

Then from (6.3.21), for $y_1, y_2 \in M_\lambda(\varepsilon)$

$$\|T(y_1,x) - T(y_2,x)\|_{\exp,\lambda} \leq \frac{1}{2}\|y_1 - y_2\|_{\exp,\lambda}. \tag{6.3.23}$$

If we assume in addition that

$$\|x\| \leq \frac{\varepsilon}{2c_0}, \tag{6.3.24}$$

then for $y \in M_\lambda(\varepsilon)$, by (6.3.20),

$$\|T(y,x)\|_{\exp,\lambda} \leq \varepsilon. \tag{6.3.25}$$

Thus, if ε satisfies (6.3.22), and $\|x\| \leq \frac{\varepsilon}{2c_0}$, then $T(\cdot,x)$ maps $M_\lambda(\varepsilon)$ into itself, with a contraction constant $\frac{1}{2}$. Therefore applying the Banach fixed point theorem, we get a unique solution $y_x \in M_\lambda(\varepsilon)$ of (6.3.17), for any $x \in P_+H$ with $\|x\| \leq \frac{\varepsilon}{2c_0}$.

Obviously, $T(0,0) = 0$, and thus $y_0 = 0$. Also, since $y_x \in M_\lambda(\varepsilon)$ is decaying exponentially, we have for any x (with $\|x\| \leq \frac{\varepsilon}{2c_0}$)

$$\lim_{t\to\infty} y_x(t) = 0,$$

i.e.

$$y_x(0) \in W^s(0).$$

From (6.3.16), we have

$$y_x(t) = e^{-Lt}x + \int_0^t e^{-L(t-s)}P_+\eta(y_x(s))\,ds - \int_t^\infty e^{-L(t-s)}P_-\eta(y_x(s))\,ds.$$

y_x lies in $M(\varepsilon)$ and so in particular is bounded for $t \geq 0$. Thus, it also satisfies (6.3.15), i.e.

$$y_x(t) = e^{-Lt}P_+y_x(0) + \int_0^t e^{-L(t-s)}P_+\eta(y_x(s))\,ds - \int_t^\infty e^{-L(t-s)}P_-\eta(y_x(s))\,ds,$$

and comparing these two representations, we see that

$$x = P_+y_x(0). \tag{6.3.26}$$

Thus, for any $U \subset \{\|x\| \leq \frac{\varepsilon}{2c_0}\}$, we have a map

$$\begin{aligned} H_+ \cap U &\to W^s(0),\\ x &\mapsto y_x(0), \end{aligned}$$

with inverse given by P_+, according to (6.3.26). We claim that this map is a bijection between $H_+ \cap U$ and its image in $W^s(0)$. For that purpose, we observe that as in (6.3.20), we get, assuming (6.3.24),

$$\|y_{x_1}(t) - y_{x_2}(t)\| \leq c_0e^{-\gamma t}\|x_1 - x_2\| + \frac{1}{2}\|y_{x_1} - y_{x_2}\|_{\exp,\lambda},$$

hence

$$\|y_{x_1}(0) - y_{x_2}(0)\| \leq \|y_{x_1} - y_{x_2}\|_{\exp,\lambda} \leq 2c_0\|x_1 - x_2\|. \tag{6.3.27}$$

We insert the second inequality in (6.3.27) into the integrals in (6.3.16) and use (6.3.11) as before to get from (6.3.16)

$$\|y_{x_1}(0) - y_{x_2}(0)\| \geq \|x_1 - x_2\| - \frac{4c_0^2\delta(\varepsilon)}{\gamma - \lambda}\|x_1 - x_2\|.$$

If in addition to the above requirement $\frac{1}{\gamma} c_0 \delta(\varepsilon) < \frac{1}{4}$ we also impose the condition upon ε that

$$\frac{4c_0^2 \delta(\varepsilon)}{\gamma - \lambda} \leq \frac{1}{2},$$

the above inequality yields

$$\|y_{x_1}(0) - y_{x_2}(0)\| \geq \frac{1}{2}\|x_1 - x_2\|. \tag{6.3.28}$$

Thus, the above map indeed is a bijection between $\{x \in P_+H, \|x\| \leq \frac{\varepsilon}{2c_0}\}$ and its image W in $W^s(0)$. (6.3.27) also shows that our map $x \mapsto y_x(0)$ is Lipschitz, whereas its inverse is Lipschitz by (6.3.28).

In particular, since $y_0 = 0$ as used above, W contains an open neighborhood of 0 in $W^s(0)$, hence is of the form $W^s(0, U)$ for some open U.

We now verify that $W^s(0, U)$ is tangent to P_+H at 0. (6.3.10), (6.3.16) and (6.3.27) yield (for $x_1 = x, x_2 = 0$, recalling $y_0 = 0$)

$$\begin{aligned}
\|P_- y_x(0)\| &= \|\int_0^\infty e^{Ls} P_- \eta(y_x(s))\, ds\| \\
&\leq c_0 \int_0^\infty e^{-\gamma s} \delta(\|y_x(s)\|)\|y_x(s)\|\, ds \\
&\leq c_0 \int_0^\infty e^{-\gamma s} \delta\left(2c_0 e^{-\lambda s}\|x\|\right) 2c_0 e^{-\lambda s}\|x\| \\
&\leq \frac{2c_0^2}{\gamma - \lambda} \delta(2c_0\|x\|)\|x\|.
\end{aligned}$$

This implies

$$\frac{\|P_- y_x(0)\|}{\|P_+ y_x(0)\|} = \frac{\|P_- y_x(0)\|}{\|x\|} \to 0,$$

as $y_x(0) \to 0$ in $W^s(0, U)$, or equivalently, $x \to 0$ in P_+H. This shows that $W^s(0, U)$ indeed is tangent to P_+H at 0.

The regularity of $W^s(0, U)$ follows since $T(y, x)$ in (6.3.16) depends smoothly on η. (It is easily seen from the proof of the Banach fixed point theorem that the fact that the contraction factor is < 1 translates smoothness of T as a function of a parameter into the same type of smoothness of the fixed point as a function of that parameter.)

Obviously, the situation for $W^u(0, U)$ is symmetric to the one for $W^s(0, U)$. $\square$

The preceding theorem provides the first step in the local analysis for the gradient flow in the vicinity of a critical point of the function f. It directly implies a global result.

Corollary 6.3.1. *The stable and unstable manifolds $W^s(x), W^u(x)$ of the negative gradient flow ϕ for a smooth function f are injectively immersed smooth manifolds. (If f is of class C^{k+2}, then $W^s(x)$ and $W^u(x)$ are of class C^k.)*

Proof. We have

$$W^s(x) = \bigcup_{t \le 0} \phi(\cdot, t)(W^s(x, U)),$$
$$W^u(x) = \bigcup_{t \ge 0} \phi(\cdot, t)(W^u(x, U)),$$

for any neighborhood U of x. $\square$

Of course, the corollary holds more generally for the flows of the type (6.3.9) (if we consider only those flow lines $\phi(\cdot, t)$ that exist for all $t \le 0$ resp. $t \ge 0$). (The stable and unstable sets then are as smooth as η is.) The point is that the flow $\phi(\cdot, t)$, for any t and any open set U, provides a diffeomorphism between U and $\phi(U, t)$, and the sets $\phi(U, t)$ cover the image of $\phi(\cdot, \cdot)$.

The stable and unstable manifolds $W^s(0), W^u(0)$ for the flow (6.3.9) are invariant under the flow, i.e. if e.g.

$$x = \phi(x, 0) \in W^u(0),$$

then also

$$x(t) = \phi(x, t) \in W^u(0), \quad \text{for all } t \in \mathbb{R} \text{ for which it exists.}$$

In §6.4, we shall easily see that because f is decreasing along flow lines, the stable and unstable manifolds are in fact embedded, see Corollary 6.4.1.

We return to the local situation. The next result says that more generally, in some neighborhood of our nondegenerate critical point 0, we may find a so-called stable foliation with leaves $\Lambda^s(z_u)$ parametrized by $z_u \in W^u(0)$, such that where defined, $\Lambda^s(0)$ coincides with $W^s(0)$ while all leaves are graphs over $W^s(0)$, and if a flow line starts on the leaf $\Lambda^s(z_u)$ at $t = 0$, then at other times t, we find it on $\Lambda^s(\phi(z_u, t))$, the leaf over the flow line on $W^u(0)$ starting at z_u at $t = 0$. Also, as t increases, different flow lines starting on the same leaf approach each other at exponential speed.

The precise result is

Theorem 6.3.2. *Suppose that the assumptions of Theorem 6.3.1 hold. There exist constants $c_1, \lambda > 0$, and neighborhoods U of 0 in H, V of 0 in P_+H with the following properties:*

For each $z_u \in W^u(0, U)$, there is a function

$$\varphi_{z_u} : V \to H.$$

$\varphi_{z_u}(z_+)$ is as smooth in z_u, z_+ as η is, for example of class C^k if η belongs to that class. If

$$z \in \Lambda^s(z_u) = \varphi_{z_u}(V),$$

then

$$\phi(z,t) = \varphi_{\phi(z_u,t)}(P_+\phi(z,t)), \tag{6.3.29}$$

and

$$\|\phi(z,t) - \phi(z_u,t)\| \le c_1 e^{-\lambda t}, \tag{6.3.30}$$

as long as $\phi(z,t)$, $\phi(z_u,t)$ *remain in* U.

We thus have a smooth (of class C^k, *if* $\eta \in C^k$*), so-called stable foliation which is flow invariant in the sense that the flow maps leaves to leaves. In particular,* $\Lambda^s(0)$ *is the stable manifold* $W^s(0) \cap V$, $\phi(z,t)$ *approaches* $W^s(0) \cap V$ *exponentially for negative* t, *as long as it stays in* U.

Of course, there also exists an unstable foliation with analogous properties.

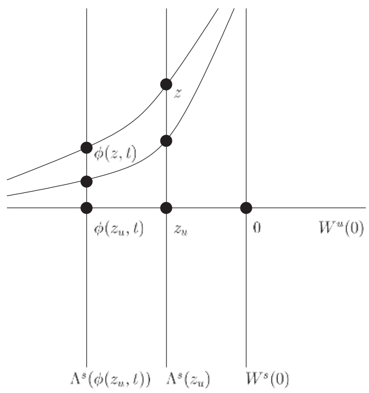

Figure 6.3.4:

Corollary 6.3.2. *Let* $f : X \to \mathbb{R}$ *be of class* C^{k+2}, $k \ge 1$, x *a nondegenerate critical point of* f. *Then in some neighborhood* U *of* x, *there exist two flow-invariant foliations of class* C^k, *the stable and the unstable one. The leaves of these two foliations intersect transversally in single points, and conversely each point of* U *is the intersection of precisely one stable and one unstable leaf.*

The Corollary is a direct consequence of the Theorem, and we thus turn to the

Proof of Theorem 6.3.2. Changing η outside a neighborhood U of 0 will not affect the local structure of the flow lines in that neighborhood. By choosing U sufficiently small and recalling (6.3.10), we may thus assume that the Lipschitz constant of η is as small as we like. We apply (6.3.12) to $\phi(z,t)$ and $\phi(z_u,t)$ and get for $\tau \ge 0$, putting $y(t;z,z_u) := \phi(z,t) - \phi(z_u,t)$,

$$\begin{aligned} y(t;z,z_u) = {} & e^{-L(t-\tau)} y(\tau;z,z_u) \\ & + \int_\tau^t e^{-L(t-s)}(\eta(\phi(z,s)) - \eta(\phi(z_u,s)))\, ds. \end{aligned} \tag{6.3.31}$$

If this is bounded for $t \to \infty$, then (6.3.11) implies, as in (6.3.14),

$$\lim_{\tau\to\infty} e^{-L(t-\tau)} P_- y(t; z, z_u) = 0. \tag{6.3.32}$$

Consequently, as in (6.3.15) we get,

$$\begin{aligned} y(t; z, z_u) = e^{-Lt} P_+ y(0; z, z_u) \\ &+ \int_0^t e^{-L(t-s)} P_+(\eta(\phi(z_u, s) + y(s; z, z_u)) - \eta(\phi(z_u, s)))\, ds \\ &- \int_t^\infty e^{-L(t-s)} P_-(\eta(\phi(z_u, s) + y(s; z, z_u)) - \eta(\phi(z_u, s)))\, ds. \end{aligned} \tag{6.3.33}$$

As in the proof of Theorem 6.3.1, we want to solve this equation by an application of the Banach fixed point theorem, i.e. by finding a fixed point of the iteration of

$$\begin{aligned} T(y, z_u, z_+) := e^{-Lt} z_+ \\ &+ \int_0^t e^{-L(t-s)} P_+(\eta(\phi(z_u, s) + y(s)) - \eta(\phi(z_u, s)))\, ds \\ &- \int_t^\infty e^{-L(t-s)} P_-(\eta(\phi(z_u, s) + y(s)) - \eta(\phi(z_u, s)))\, ds, \end{aligned} \tag{6.3.34}$$

for $z_+ \in P_+ H$. As in the proof of Theorem 6.3.1, we shall use a space $M_\lambda(\varepsilon_0)$ for some fixed $0 < \lambda < \gamma$.

Before we proceed to verify the assumptions required for the application of the fixed point theorem, we wish to describe the meaning of the construction. Namely, given $z_u \in W^u(0)$, and the orbit $\phi(z_u, t)$ starting at z_u and contained in $W^u(0)$, and given $z_+ \in P_+ H$, we wish to find an orbit $\phi(z, t)$ with $P_+\phi(z, 0) = P_+ z = z_+$ that exponentially approaches the orbit $\phi(z_u, t)$ for $t \geq 0$. The fixed point argument will then show that in the vicinity of 0, we may find a unique such orbit. If we keep z_u fixed and let z_+ vary in some neighborhood of 0 in $P_+ H$, we get a corresponding family of orbits $\phi(z, t)$, and the points $z = \phi(z, 0)$ then constitute the leaf through z_u of our foliation. The leaves are disjoint because orbits on the unstable manifold $W^u(0)$ with different starting points for $t = 0$ diverge exponentially for positive t. Thus, any orbit $\phi(z, t)$ can approach at most one orbit $\phi(z_u, t)$ on $W^u(0)$ exponentially. In order to verify the foliation property, however, we also will have to show that the leaves cover some neighborhood of 0, i.e. that any flow line $\phi(z, t)$ starting in that neighborhood for $t = 0$ approaches some flow line $\phi(z_u, t)$ in $W^u(0)$ exponentially. This is equivalent to showing that the leaf through z_u depends continuously on z_u, and this in turn follows from the continuous dependence of the fixed point of $T(\cdot, z_u, z_+)$ on z_u.

Precisely as in the proof of Theorem 6.3.1, we get for $0 < \lambda < \gamma$ (say $\lambda = \frac{\gamma}{2}$), with c_0, γ as in (6.3.11), $\|z_+\| \leq \varepsilon_1$, $y \in M_\lambda(\varepsilon_0)$, i.e. $\|y(t)\| \leq e^{-\lambda t}\varepsilon_0$, and with $[\eta]_{Lip}$ being the Lipschitz constant of η

$$\|T(y, z_u, z_+)(t)\| \leq c_0 \varepsilon_1 e^{-\gamma t} + \frac{2c_0\varepsilon_0}{\gamma - \lambda}[\eta]_{Lip} e^{-\lambda t} \tag{6.3.35}$$

and

$$\|T(y_1, z_u, z_+)(t) - T(y_2, z_u, z_+)(t)\| \leq \frac{4c_0[\eta]_{Lip}}{\gamma - \lambda} e^{-\lambda t} \|y_1 - y_2\|_{\exp,\lambda}. \tag{6.3.36}$$

As remarked at the beginning of this proof, we may assume that $[\eta]_{Lip}$ is as small as we like. Therefore, by choosing $\varepsilon_1 > 0$ sufficiently small, we may assume from (6.3.35) that $T(\cdot, z_u, z_+)$ maps $M_\lambda(\varepsilon_0)$ into itself, and from (6.3.36) that it satisfies

$$\|T(y_1, z_u, z_+) - T(y_2, z_u, z_+)\|_{\exp,\lambda} \leq \frac{1}{2} \|y_1 - y_2\|_{\exp,\lambda}.$$

Thus, the Banach fixed point theorem, applied to $T(\cdot, z_u, z_+)$ on the space $M_\lambda(\varepsilon_0)$, yields a unique fixed point y_{z_u,z_+} on this space. We now put

$$\begin{aligned} \varphi_{z_u}(z_1) :&= y_{z_u,z_+}, \\ z &= y_{z_u,z_+}(0). \end{aligned} \tag{6.3.37}$$

We then have all the required relations:

$$P_+ z = P_+ y_{z_u,z_1}(0) = z_1 \quad \text{from (6.3.34)},$$

and hence y_{z_u,z_+} solves (6.3.33), i.e. is of the form $y(t; z, z_u)$ with z from (6.3.37), and $\phi(z,t) = y(t; z, z_u) + \phi(z_u, t)$ is a flow line. Condition (6.3.29) thus holds at $t = 0$. Since the construction is equivariant w.r.t. time shifts, because of the group property

$$\phi(z, t + \tau) = \phi(\phi(z,t), \tau) \quad \text{for all } t, \tau,$$

(6.3.29) holds for any t, as long as $\phi(z,t)$ stays in our neighborhood U of 0. The exponential decay of $\phi(z,t) - \phi(z_u,t) = y(t; z, z_u)$ follows since we have constructed our fixed point of T in the space of mappings with precisely that decay.

Since T is linear in z_+, we see as before in the proof of Theorem 6.3.1 that a smoothness property of η translates into a smoothness property of y_{z_u} as a function of z_+. It remains to show the smoothness of y_{z_u,z_+} as a function of z_u. This, however is a direct consequence of the fact that y_{z_u,z_+} is a fixed point of $T(\cdot, z_u, z_+)$, an operator with a contraction constant < 1 on the space under consideration ($M_\lambda(\varepsilon_0)$), and so the smooth dependence of T (see (6.3.34) on the parameters z_u and z_+ (which easily follows from estimates of the type used above) translates into the corresponding smoothness of the fixed point as a function of the parameters z_u, z_+.

The foliation property is then clear, because leaves corresponding to different $z'_u, z''_u \in W^u(0, U)$ cannot intersect as we had otherwise $z = y_{z'_u,z_+}(0) = y_{z''_u,z_+}(0)$ for some z with $z_+ = P_+ z$, hence also $z'_u = \phi(z'_u, 0) = \phi(z, 0) - y_{z'_u,z_+}(0) = \phi(z,0) - y_{z''_u,z_+}(0) = z''_u$.

As the leaves depend smoothly on z_u, they approach the stable manifold $W^s(0)$ at the same speed as z_u does. More precisely, any orbit $\phi(z_u, t)$ converges to 0 exponentially for $t \to -\infty$, and the leaf over $\phi(z_u, t)$ then has to converge exponentially to the one over 0 which is $W^s(0)$.

The last statement easily follows by changing signs appropriately, for example by replacing t by $-t$ throughout. $\square$

Perspectives. The theory of stable and unstable manifolds for a dynamical system is classical. Our presentation is based on the one in [54], although we have streamlined it somewhat by consistently working with function spaces with exponential weights.

6.4 Limits of Trajectories of the Gradient Flow

As always in this chapter, X is a complete Riemannian manifold, with metric $\langle\cdot,\cdot\rangle$, associated norm $\|\cdot\|$, and distance function $d(\cdot,\cdot)$. $f : X \to \mathbb{R}$ is a C^2-function. We consider the negative gradient flow

$$\begin{aligned} \dot{x}(t) &= -\operatorname{grad} f(x(t)) \qquad && \text{for } t \in \mathbb{R},\\ x(0) &= x && \text{for } x \in X. \end{aligned} \tag{6.4.1}$$

We assume that the norms of the first and second derivative of f are bounded. Applying the Picard-Lindelöf theorem (see §1.9), we then infer that our flow is indeed defined for all $t \in \mathbb{R}$. Also, differentiating (6.4.1), we get

$$\begin{aligned} \ddot{x}(t)\ (= \nabla_{\frac{d}{dt}}\dot{x}(t)) &= -\left(\nabla_{\frac{\partial}{\partial x}} \operatorname{grad} f(x(t))\right)\dot{x}(t)\\ &= \quad \left(\nabla_{\frac{\partial}{\partial x}} \operatorname{grad} f(x(t))\right)\operatorname{grad} f(x(t)). \end{aligned}$$

In particular, the first and second derivative of any flow line is uniformly bounded. For later use, we quote this fact as:

Lemma 6.4.1. *There exists a constant c_0 with the property that for any solution $x(t)$ of (6.4.1),*

$$\|\dot{x}\|_{C^1(\mathbb{R},TX)} \le c_0.$$

In particular, $\dot{x}(t)$ is uniformly Lipschitz continuous.

(6.4.1) is a system of so-called autonomous ordinary differential equations, meaning that the right hand side does not depend explicitly on the "time" t, but only implicitly through the solution $x(t)$.

In contrast to the previous section, where we considered the local behaviour of this flow near a critical point of f, we shall now analyze the global properties, and the gradient flow structure will now become more important.

In the sequel, $x(t)$ will always denote a solution of (6.4.1), and we shall exploit (6.4.1) in the sequel without quoting it explicitly. We shall call each curve $x(t)$, $t \in \mathbb{R}$, a flow line, or an orbit (of the negative gradient flow). We also put, for simplicity

$$x(\pm\infty) := \lim_{t \to \pm\infty} x(t),$$

assuming that these limits exist.

Lemma 6.4.2. *The flow lines of (6.4.1) are orthogonal to the level hypersurfaces* $f = \text{const}$.

Proof. This means the following: If for some $t \in \mathbb{R}$, $V \in T_{x(t)}X$ is tangent to the level hypersurface $\{y : \ f(y) = f(x(t))\}$, then

$$\langle V, \dot{x}(t) \rangle = 0.$$

Now

$$\begin{aligned} \langle V, \dot{x}(t) \rangle &= -\langle V, \operatorname{grad} f(x(t)) \rangle \\ &= -V(f)(x(t)) \end{aligned}$$

by the definition of $\operatorname{grad} f$, see (2.1.15),

$$= 0,$$

since V is tangent to a hypersurface on which f is constant. □

We compute

$$\begin{aligned} \frac{d}{dt} f(x(t)) &= df(x(t))\dot{x}(t) \\ &= \langle \operatorname{grad} f(x(t)), \dot{x}(t) \rangle \qquad \text{by (2.1.15)} \\ &= -\|\dot{x}(t)\|^2. \end{aligned} \tag{6.4.2}$$

As a consequence, we observe

Lemma 6.4.3. *f is decreasing along flow lines. In particular, there are no nonconstant homoclinic orbits, i.e. nonconstant orbits with*

$$x(-\infty) = x(\infty).$$

□

Thus, we see that there are only two types of flow lines or orbits, the "typical" ones diffeomorphic to the real axis $(-\infty, \infty)$ on which f is strictly decreasing, and the "exceptional" ones, namely those that are reduced to single points, the critical points of f. The issue now is to understand the relationship between the two types.

Another consequence of (6.4.2) is that for $t_1, t_2 \in \mathbb{R}$

$$\begin{aligned} f(x(t_1)) - f(x(t_2)) &= -\int_{t_1}^{t_2} \frac{d}{dt} f(x(t))\, dt \\ &= \int_{t_1}^{t_2} \|\dot{x}(t)\|^2 \\ &= \int_{t_1}^{t_2} \|\operatorname{grad} f(x(t))\|^2\, dt. \end{aligned} \tag{6.4.3}$$

We also have the estimate

$$\begin{aligned} d(x(t_1), x(t_2)) &\le \int_{t_1}^{t_2} \|\dot{x}(t)\|\, dt \\ &\le (t_2 - t_1)^{\frac{1}{2}} \left(\int_{t_1}^{t_2} \|\dot{x}(t)\|^2 \right)^{\frac{1}{2}} \quad \text{by Hölder's inequality} \\ &= (t_2 - t_1)^{\frac{1}{2}} (f(x(t_1)) - f(x(t_2)))^{\frac{1}{2}} \text{ by (6.4.3).} \end{aligned} \tag{6.4.4}$$

Lemma 6.4.4. *For any flow line, we have for $t \to \pm\infty$ that*

$$\operatorname{grad} f(x(t)) \to 0,$$

or,

$$|f(x(t))| \to \infty.$$

Proof. If e.g. $f_\infty = \lim_{t\to\infty} f(x(t)) > -\infty$, then for $0 \le t \le \infty$

$$f_0 := f(x(0)) \ge f(x(t)) \ge f_\infty,$$

and (6.4.3) implies

$$\int_0^\infty \|\dot{x}(t)\|^2 := f_0 - f_\infty < \infty. \tag{6.4.5}$$

Since $\dot{x}(t) = -\operatorname{grad} f(x(t))$ is uniformly Lipschitz continuous by Lemma 6.4.1, (6.4.5) implies that

$$\lim_{t\to\infty} \operatorname{grad} f(x(t)) = \lim_{t\to\infty} \dot{x}(t) = 0.$$

□

We also obtain the following strengthening of Corollary 6.3.1:

Corollary 6.4.1. *The stable and unstable manifolds $W^s(x), W^u(x)$ of the negative gradient flow ϕ for a smooth function f are embedded manifolds.*

Proof. The proof is an easy consequence of what we have already derived, but it may be instructive to see how all those facts are coming together here.

We have already seen in Corollary 6.3.1 that $W^s(x)$ and$W^u(x)$ are injectively immersed. By Corollary 1.9.1, each point in X is contained in a unique flow line, but the typical ones of the form $(-\infty, \infty)$ are not compact, and so, their closures may contain other points. By Lemma 6.4.4, any such point is a critical point of f. The local situation near such a critical point has already been analyzed in Theorem 6.3.1. The only thing that still needs to be excluded to go from Corollary 6.3.1 to the present statement is that a flow line $x(t)$ emanating at one critical point $x(-\infty)$ returns to that same point for $t \to \infty$. This, however, is exluded by Lemma 6.4.3. □

In the sequel, we shall also make use of

Lemma 6.4.5. *Suppose $(x_n)_{n\in\mathbb{N}} \subset X$ converges to x_0. Then for any $T > 0$, the curves $x_n(t)_{|[-T,T]}$ (with $x_n(0) = x_n$) converge in C^1 to the curve $x_0(t)_{|[-T,T]}$.*

Proof. This follows from the continuous dependence of solutions of ODEs on the initial data under the assumption of the Picard-Lindelöf theorem (the proof of that theorem is based on the Banach fixed point theorem, and the fixed point produced in that theorem depends continuously on a parameter, cf. J.Jost, Postmodern Analysis, Springer, 1998, p.129). Thus the curves $x_n(t)$ converge uniformly to $x_0(t)$ on any finite interval $[-T, T]$. By Lemma 6.4.1, $\ddot{x_n}(t)$ are uniformly bounded, and so x_n has to converge in C^1. □

We now assume for the remainder of this section that f satisfies the Palais-Smale condition (PS), and that all critical points of f are nondegenerate.

These assumptions are rather strong as they imply

Lemma 6.4.6. *f has only finitely many critical points in any bounded region of X, or, more generally in any region where f is bounded. In particular, in every bounded interval in $\mathbb{R}$ there are only finitely many critical values of f, i.e. $\gamma \in \mathbb{R}$ for which there exists $p \in X$ with $df(p) = 0$, $f(p) = \gamma$.*

Proof. Let $(p_n)_{n\in\mathbb{N}} \subset X$ be a sequence of critical points of f, i.e. $df(p_n) = 0$. If they are contained in a bounded region of X, or, more generally, if $f(p_n)$ is bounded, the Palais-Smale condition implies that after selection of a subsequence, they converge towards some critical point p_0. By Theorem 6.3.1, we may find some neighborhood U of p_0 in which the flow has the local normal form as described there and which in particular contains no other critical point of f besides p_0. This implies that almost all p_n have to coincide with p_0, and thus there can only be finitely many of them. □

Our assumptions – (PS) and nondegeneracy of all critical points – also yield

Lemma 6.4.7. *Let $x(t)$ be a flow line for which $f(x(t))$ is bounded. Then the limits $x(\pm\infty) := \lim_{t\to\pm\infty} x(t)$ exist and are critical points of f. $x(t)$ converges to $x(\pm\infty)$ exponentially as $t \to \pm\infty$.*

Proof. By Lemma 6.4.4, $\operatorname{grad} f(x(t)) \to 0$ for $t \to \pm\infty$. Analyzing w.l.o.g. the situation $t \to -\infty$, (PS) implies that we can find a sequence $(t_n)_{n\in\mathbb{N}} \subset \mathbb{R}$, $t_n \to -\infty$ for $n \to \infty$, for which $x(t_n)$ converges to some critical point $x_{-\infty}$ of f. We wish to show that $\lim_{t\to-\infty} x(t)$ exists, and it then has to coincide with $x_{-\infty}$.

This, however, directly follows from the nondegeneracy condition, since by Theorem 6.3.1 we may find a neighborhood U of the critical point $x_{-\infty}$ with the property that any flow line in that neighborhood containing $x_{-\infty}$ as an accumulation point of some sequence $x(t_n)$, $t_n \to -\infty$, is contained in the unstable manifold of $x_{-\infty}$. Furthermore, as shown in Theorem 6.3.1, the convergence is exponential. $\square$

Remark. Without assuming that the critical point $x(-\infty)$ is nondegenerate, we still may use (PS) (see Lemma 6.4.8 below) and $\operatorname{grad} f(x(t)) \to 0$ for $t \to -\infty$ to see that there exists $t_0 \in \mathbb{R}$ for which $U := \{x(t) : t \leq t_0\}$ is precompact and in particular bounded. By Taylor expansion, we have in U

$$\|\operatorname{grad} f(x)\| \leq \|\operatorname{grad} f(x_{-\infty})\| + cd(x, x_{-\infty}) = cd(x, x_{-\infty}),$$

for some constant c, as $\operatorname{grad} f(x_{-\infty}) = 0$.

Thus, for $t \leq t_n$

$$d(x(t), x_{-\infty}) \leq \int_{-\infty}^{t} \|\dot{x}(s)\|\, ds \leq c \int_{-\infty}^{t} d(x(s), x_{-\infty})\, ds.$$

The latter integral may be infinite. As soon as it is finite, however, we already get

$$d(x(t), x_{-\infty}) \leq c_1 e^{ct} \quad \text{for some constant } c_1,$$

i.e. exponential convergence of $x(t)$ towards $x_{-\infty}$ as $t \to -\infty$.

We shall also use the following simple estimate

Lemma 6.4.8. *Suppose $\|\operatorname{grad} f(x(t))\| \geq \varepsilon$, for $t_1 \leq t \leq t_2$. Then*

$$d(x(t_1), x(t_2)) \leq \frac{1}{\varepsilon}(f(x(t_1)) - f(x(t_2))).$$

Proof.

$$\begin{aligned} d(x(t_1), x(t_2)) &\leq \int_{t_1}^{t_2} \|\dot{x}(t)\|\, dt \\ &\leq \frac{1}{\varepsilon} \int_{t_1}^{t_2} \|\dot{x}(t)\|^2\, dt \text{ since } \|\dot{x}(t)\| = \|\operatorname{grad} f(x(t))\| \geq \varepsilon \\ &= \frac{1}{\varepsilon}(f(x(t_1)) - f(x(t_2))) \ \text{ by (6.4.3).} \end{aligned}$$

□

We now need an **additional assumption**:

There exists a flow-invariant **compact** *set $X^f \subset X$ containing the critical points p and q.*

What we have in mind here is a certain set of critical points together with all connecting trajectories between them. We shall see in Theorem 6.4.1 below that we need to include here all critical points that can arise as limits of flow lines between any two critical points of the set we wish to consider.

Lemma 6.4.9. *Let $(x_n(t))_{n\in\mathbb{N}}$ be a sequence of flow lines in X^f with*

$$\begin{aligned} x_n(-\infty) &= p, \\ x_n(\infty) &= q. \end{aligned}$$

Then after selection of a subsequence, $x_n(t)$ converges in C^1 on any compact interval in $\mathbb{R}$ towards some flow line $x_0(t)$.

Proof. Let $t_0 \in \mathbb{R}$. If (for some subsequence)

$$\|\operatorname{grad} f(x_n(t_0))\| \to 0,$$

then by (PS) ($\gamma_1 = f(p)$, $\gamma_2 = f(q)$, noting $f(p) \geq f(x(t)) \geq f(q)$ by Lemma 6.4.3), we may assume that $x_n(t_0)$ converges, and the convergence of the flow lines on compact intervals then follows from Lemma 6.4.5. We thus assume

$$\|\operatorname{grad} f(x_n(t_0))\| \geq \varepsilon \quad \text{for all } n \text{ and some } \varepsilon > 0.$$

Since $f(x_n(t))$ is bounded between $f(p)$ and $f(q)$, Lemma 6.4.4 implies that we may find $t_n < t_0$ with

$$\|\operatorname{grad} f(x_n(t_n))\| = \varepsilon$$

and

$$\|\operatorname{grad} f(x_n(t))\| \geq \varepsilon \quad \text{for } t_n \leq t \leq t_0.$$

From (6.4.3), we get

$$|t_n - t_0| \leq \frac{1}{\varepsilon^2}(f(t_n) - f(t_0)) \leq \frac{1}{\varepsilon^2}(f(p) - f(q)).$$

Applying our compactness assumption on X^f, we may assume that $x_n(t_n)$ converges. From Lemma 6.4.5 we then see that $x_n(t)$ converges on any compact interval towards some flow line $x_0(t)$. □

In general, $x_n(t)$ will not converge uniformly on all of $\mathbb{R}$ towards $x_0(t)$. We need an additional assumption as in the next

Lemma 6.4.10. *Under the assumption of Lemma 6.4.9, assume*

$$x_0(-\infty) = p,$$
$$x_0(\infty) = q,$$

i.e. $x_0(t)$ has the same limit points as the $x_n(t)$. Then the $x_n(t)$ converge to $x_0(t)$ in the Sobolev space $H^{1,2}(\mathbb{R}, X)$. In fact, this holds already if we only assume

$$f(x_0(-\infty)) = f(p),$$
$$f(x_0(\infty)) = f(q).$$

Proof. The essential point is to show that

$$\lim_{t\to-\infty} x_n(t) = p, \quad \lim_{t\to\infty} x_n(t) = q, \quad \text{uniformly in } n.$$

Namely in that case, we may apply the local analysis provided by Theorem 6.3.1 uniformly in n to conclude convergence for $t \le t_1$ and $t \ge t_2$ for certain $t_1, t_2 \in \mathbb{R}$, and on the compact interval $[t_1, t_2]$, we get convergence by the preceding lemma.

Because of (PS), we only have to exclude that after selection of a subsequence of $x_n(t)$, we find a sequence $(t_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ converging to ∞ or $-\infty$, say $-\infty$, with

$$\|\operatorname{grad} f(x_n(t_n))\| \ge \varepsilon \quad \text{for some } \varepsilon > 0. \tag{6.4.6}$$

From (6.4.4), we get the uniform estimate

$$\|\operatorname{grad} f(x_n(t_1)) - \operatorname{grad} f(x_n(t_2))\| \le c(t_2 - t_1)^{\frac{1}{2}} \quad \text{for some constant } c. \tag{6.4.7}$$

By (6.4.6), (6.4.7), we may find $\delta > 0$ such that for $t_n - \delta \le t \le t_n$,

$$\|\operatorname{grad} f(x_n(t))\| \ge \frac{\varepsilon}{2},$$

hence

$$f(p) - f(x_n(t_n)) \ge f(x_n(t_n - \delta)) - f(x_n(t_n)) \ge \delta\,\frac{\varepsilon^2}{4} \quad \text{by (6.4.3)}.$$

On the other hand, by our assumption on $x_0(t)$, we may find $t_0 \in \mathbb{R}$ with

$$f(p) - f(x_0(t_0)) = \delta\,\frac{\varepsilon^2}{8}. \tag{6.4.8}$$

If $t_n \le t_0$, we have

$$f(p) - f(x_n(t_0)) \ge f(p) - f(x_n(t_n)) \ge \delta\,\frac{\varepsilon^2}{4},$$

and so $x_n(t_0)$ cannot converge to $x_0(t_0)$, contrary to our assumption. Thus (6.4.6) is impossible, and the proof is complete, except for the last remark, which, however, also directly follows as the only assumption about $x_0(t)$ that we need is (6.4.8). □

We are now ready to demonstrate the following compactness

Theorem 6.4.1. *Let p, q be critical points of f, and let $\mathcal{M}^f_{p,q} \subset X^f$ be a space of flow lines $x(t)(t \in \mathbb{R})$ for f with $x(-\infty) = p$, $x(\infty) = q$. Here we assume that X^f is a flow-invariant compact set. Then for any sequence $(x_n(t))_{n\in\mathbb{N}} \subset \mathcal{M}^f_{p,q}$, after selection of a subsequence, there exist critical points*

$$p = p_1, p_2, \ldots, p_k = q,$$

flow lines $y_i \in \mathcal{M}^f_{p_i,p_{i+1}}$ and $t_{n,i} \in \mathbb{R}$ $(i = 1, \ldots, k-1,\ n \in \mathbb{N})$ such that the flow lines $x_n(t + t_{n,i})$ converge to y_i for $n \to \infty$. In this situation, we say that the sequence $x_n(t)$ converges to the broken trajectory $y_1\#y_2\#\ldots\#y_{k-1}$.

Proof. By Lemma 6.4.9, $x_n(t)$ converges (after selection of a subsequence, as always) towards some flow line $x_0(t)$. $x_0(t)$ need not be in $\mathcal{M}^f_{p,q}$, but the limit points $x_0(-\infty)$, $x_0(\infty)$ (which exist by Lemma 6.4.7) must satisfy

$$f(p) \geq f(x_0(-\infty)) \geq f(x_0(\infty)) \geq f(q).$$

If e.g. $f(p) = f(x_0(-\infty))$ then the proof of Lemma 6.4.10 shows that

$$x_0(-\infty) = p.$$

If $f(p) > f(x_0(-\infty))$, we choose $f(x_0(-\infty)) < a < f(p)$ and $t_{n,i}$ with

$$f(x_n(t_n, i)) = a.$$

We apply Lemma 6.4.9 to $x_n(t + t_{n,i})$ to get a limiting flow line $y_0(t)$. Clearly,

$$f(p) \geq f(y_0(-\infty)),$$

and we must also have

$$f(y_0(\infty)) \geq f(x_0(-\infty)),$$

because otherwise the flow line $y_0(t)$ would contain the critical point $x_0(-\infty)$ in its interior.

If $f(p) > f(y_0(-\infty))$ of $f(y_0(\infty)) > f(x_0(-\infty))$, we repeat the process. The process must stop after a finite number of such steps, because the critical points of f are isolated because of (PS) and the nondegeneracy assumption yielding to the local picture of Theorem 6.3.1 (see Lemma 6.4.6). $\square$

6.5 The Morse-Smale-Floer Condition: Transversality and $\mathbb{Z}_2$-Cohomology

In this section, we shall continue to assume the Palais-Smale condition and the nondegeneracy of all critical points of our function $f : X \to \mathbb{R}$. Here, we assume that f is of class C^3.

The central object of Morse-Floer theory is the space of connecting trajectories between the critical points of a function f. If f is bounded, then by Lemma 6.4.6, any $x \in X$ lies on some such trajectory connecting two critical points of f. In the general case, one may simply restrict the considerations in the sequel to the subspace X^f of X of such connecting trajectories, and one may even consider only some subset of the critical points of f and the connecting trajectories between them, including those limiting configurations that arise by Theorem 6.4.1. As in §6.4, we need to assume that the set of flow-lines under consideration is contained in a compact flow-invariant set. Thus, we shall assume X is such a closed space of connecting trajectories.

X then carries two stratifications S^s and S^u, consisting of the stable resp. unstable manifolds of the critical points of f. Thus, each point lies on precisely one stratum of S^s, and likewise on one stratum of S^u, and each such stratum is a smooth manifold, by Corollary 6.3.1.

Definition 6.5.1. The pair (X, f) satisfies the *Morse-Smale-Floer condition* if all intersections between the strata of S^s and the ones of S^u are finite-dimensional and transversal.

We recall that two submanifolds X_1, X_2 of X intersect transversally if for all $x \in X_1 \cap X_2$, the tangent space T_xX is the linear span of the tangent spaces T_xX_1 and T_xX_2. If the dimension of X is finite, then if X_1 and X_2 intersect transversally at x, we have

$$\dim X_1 + \dim X_2 = \dim(X_1 \cap X_2) + \dim X. \tag{6.5.1}$$

It easily follows from the implicit function theorem that in the case of a transversal intersection of smooth manifolds X_1, X_2, $X_1 \cap X_2$ likewise is a smooth manifold.

In addition to (PS) and the nondegeneracy of all critical points of f, we shall **assume** *for the rest of this section that (X, f) satisfies the* **Morse-Smale-Floer condition**.

Definition 6.5.2. Let p, q be critical points of f. If the unstable manifold $W^u(p)$ and the stable manifold $W^s(q)$ intersect, we say that p is connected to q by the flow, and we define the relative index of p and q as

$$\mu(p, q) := \dim(W^u(p) \cap W^s(q)).$$

$\mu(p, q)$ is finite because of the Morse-Smale-Floer condition.

If X is finite dimensional, then the Morse indices $\mu(p)$ of all critical points p of f themselves are finite, and in the situation of Definition 6.5.2, we then have

$$\mu(p, q) = \mu(p) - \mu(q) \tag{6.5.2}$$

as one easily deduces from (6.5.1). Returning to the general situation, we start with the following simple observation

Lemma 6.5.1. *Any nonempty intersection $W^u(p) \cap W^s(q)$ $(p, q \in C(f), p \neq q)$ is a union of flow lines. In particular, its dimension is at least* 1.

Proof. If $x \in W^u(p)$, then so is the whole flow line $x(t)$ $(x(0) = x)$, and the same holds for $x \in W^s(q)$. □

p is thus connected to q by the flow if and only if there is a flow line $x(t)$ with $x(-\infty) = p$ and $x(\infty) = q$. Expressed in another way, the intersections $W^u(p) \cap W^s(q)$ are flow invariant. In particular, in the case of a nonempty such intersection, p and q are both contained in the closure of $W^u(p) \cap W^s(q)$.

The following lemma is fundamental:

Lemma 6.5.2. *Suppose that p is connected to r and r to q by the flow. Then p is also connected to q by the flow, and*

$$\mu(p, q) = \mu(p, r) + \mu(r, q).$$

Proof. By assumption, $W^u(p)$ intersects $W^s(r)$ transversally in a manifold of dimension $\mu(p, r)$. Since $W^s(r)$ is a leaf of the smooth stable foliation of r in some neighborhood U of r by Theorem 6.3.2, in some possibly smaller neighborhood of r, $W^u(p)$ intersects each leaf of this stable foliation transversally in some manifold of dimension $\mu(p, r)$. Similarly, in the vicinity of r, $W^s(q)$ also intersects each leaf of the unstable foliation of r in some manifold, this time of dimension $\mu(r, q)$. Thus, the following considerations will hold in some suitable neighborhood of r.

The space of leaves of the stable foliation of r is parametrized by $W^u(r)$, and we thus get a family of $\mu(p, r)$-dimensional manifolds parametrized by $W^u(r)$. Likewise, we get a second family of $\mu(r, q)$-dimensional manifolds parametrized by $W^s(r)$. The leaves of the stable and unstable foliations satisfy uniform C^1-estimates (in the vicinity of r) by Theorem 6.3.2, because of our assumption that f is of class C^3. The two finite-dimensional families that we have constructed may also be assumed to satisfy such uniform estimates. The stable and unstable foliations yield a local product structure in the sense that each point near r is the intersection of precisely one stable and one unstable leaf.

If we now have two such foliations with finite-dimensional smooth subfamilies of dimension n_1 and n_2, say, all satisfying uniform estimates, it then easily follows by induction on n_1 and n_2 that the leaves of these two subfamilies need to intersect in a submanifold of dimension $n_1 + n_2$. The case where $n_1 = n_2 = 0$ can be derived from the implicit function theorem. □

We also have the following converse result

Lemma 6.5.3. *In the situation of Theorem 6.4.1, we have*

$$\sum_{i=1}^{k-1} \mu(p_i, p_{i+1}) = \mu(p, q).$$

Proof. It suffices to treat the case $k = 3$ as the general case then will easily follow by induction. This case, however, easily follows from Lemma 6.5.2 with $p = p_1$, $r = p_2$, $q = p_3$. □

We shall now need to make the **assumption** *that the space X^f of connecting trajectories that we are considering is compact. (At this moment, we are considering the space $W^u(p) \cap W^s(q)$.)*

Lemma 6.5.4. *Suppose that p, q ($p \neq q$) are critical points of f, connected by the flow, with*

$$\mu(p, q) = 1.$$

Then there exist only finitely many trajectories from p to q.

Proof. For any point x on such a trajectory, we have

$$f(p) \geq f(x) \geq f(q).$$

We may assume that $\varepsilon > 0$ is so small that on each flow line from p to q, we find some x with $\| \operatorname{grad} f(x)\| = \varepsilon$, because otherwise we would have a sequence of flow lines $(s_i)_{i \in \mathbb{N}}$ from p to q with $\sup_{x \in s_i} \| \operatorname{grad} f(x)\| \to 0$ for $i \to \infty$. By (PS) a subsequence would converge to a flow line s (see Lemma 6.4.5) with $\operatorname{grad} f(x) \equiv 0$ on s. s would thus be constant, in contradiction to Theorem 6.4.1. Thus, if, contrary to our assumption, we have a sequence $(s_i)_{i \in \mathbb{N}}$ of trajectories from p to q, we select $x_i \in s_i$ with $\| \operatorname{grad} f(x_i)\| = \varepsilon$, use the compactness assumption on the flow-invariant set containing the s_i to get a convergent subsequence of the x_i, hence also of the s_i by Theorem 6.4.1. The limit trajectory s also has to connect p to q, because our assumption $\mu(p, q) = 1$ and Lemmas 6.5.1 and 6.5.3 rule out that s is a broken trajectory containing further critical points of f. The Morse-Smale-Floer condition implies that s is isolated in the one-dimensional manifold $W^u(p) \cap W^s(q)$. This is not compatible with the assumption that there exists a sequence (s_i) of different flow lines converging to s. Thus, we conclude finiteness. □

We can now summarize our results about trajectories:

Theorem 6.5.1. *Suppose our general assumptions ($f \in C^3$, (PS), nondegeneracy of critical points, Morse-Smale-Floer condition) continue to hold. Let p, q be critical points of f connected by the flow with*

$$\mu(p, q) = 2.$$

Then each component of the space of flow lines from p to q, $\mathcal{M}^f_{p,q} := W^u(p) \cap W^s(q)$ either is compact after including p, q (and diffeomorphic to the 2-sphere), or its boundary (in the sense of Theorem 6.4.1) consists of two different broken trajectories from p to q.

Conversely each broken trajectory $s = s_1 \# s_2$ from p to q (this means that there exists a critical point p' of f with $\mu(p,p') = 1 = \mu(p',q)$, $s_1(-\infty) = p$, $s_1(\infty) = p' = s_2(-\infty)$, $s_2(\infty) = q$) is contained in the boundary of precisely one component of $\mathcal{M}^f_{p,q}$.

Remark. Let $s_1' \# s_2'$ and $s_1'' \# s_2''$ be broken trajectories contained in the boundary of the same component of $\mathcal{M}^f_{p,q}$. It is then possible that $s_1' = s_1''$ or $s_2' = s_2''$, but the theorem says that we cannot have both equalities simultaneously.

Proof of Theorem 6.5.1. If a component $\mathcal{M}$ of $\mathcal{M}^f_{p,q}$ is compact then it is a $2-$dimensional manifold that is a smooth family of curves, flow lines from p to q with common end points p, q, but disjoint interiors. Thus, such a component is diffeomorphic to S^2.

If $\mathcal{M}$ is not compact, Theorem 6.4.1 implies the existence of broken trajectories from p to q in the boundary of this component.

Let a be a regular value of f with $f(p) > a > f(q)$. By Lemma 6.4.2, $\mathcal{M}$ intersects the level hypersurface $f^{-1}(a)$ transversally, and $\mathcal{M} \cap f^{-1}(a)$ thus is a 1-dimensional manifold. It can thus be compactified by adding one or two points. By Theorem 6.4.1, these points correspond to broken trajectories from p to q. We thus need to exclude that $\mathcal{M}$ can be compactified by a single broken trajectory $s_1 \# s_2$. We have $s_1(-\infty) = p$, $s_2(\infty) = q$, and we put $p' := s_1(\infty) = s_2(-\infty)$. In view of the local normal form provided by Theorem 6.3.2, we have the following situation near p': $\mathcal{M}^f_{p,q}$ is a smooth surface containing s_1 in its interior. $\mathcal{M}^f_{p,q}$ then intersects a smooth $1-$dimensional family of leaves of the stable foliation near p' in a $1-$dimensional manifold. The family of those stable leaves intersected by $\mathcal{M}^f_{p,q}$ then is parametrized by a smooth curve in $W^u(p')$ containing p' in its interior. It thus contains the initial pieces of different flow lines originating from p in opposite directions, and these flow lines are contained in limits of flow lines from $\mathcal{M}^f_{p,q}$. Therefore, in order to compactify $\mathcal{M}^f_{p,q}$ in $W^u(p')$, a single flow line s_2 does not suffice.
Finally, if a broken trajectory through some p' would be a $2-$sided limit of $\mathcal{M}^f_{pq}$, this again would not be compatible with the local flow geometry near p' as just described. □

Definition 6.5.3. Let $C_*(f, \mathbb{Z}_2)$ be the free Abelian group with $\mathbb{Z}_2$-coefficients generated by the set $C_*(f)$ of critical points of f. For $p \in C_*(f)$, we put

$$\partial p := \sum_{\substack{r \in C_*(f) \\ \mu(p,r)=1}} \left(\#_{\mathbb{Z}_2} \mathcal{M}^f_{p,r} \right) r,$$

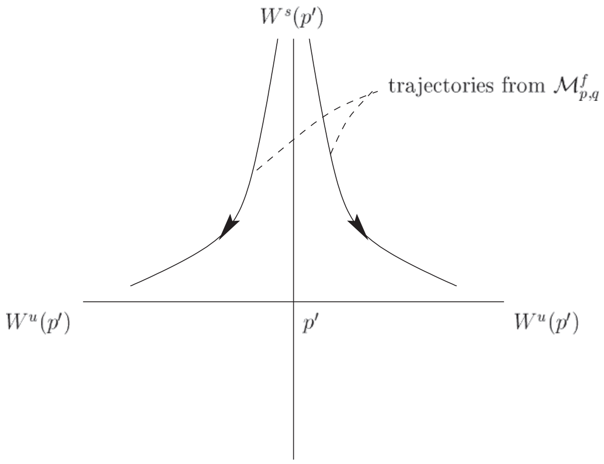

Figure 6.5.1:

where $\#_{\mathbb{Z}_2}\mathcal{M}^f_{p,r}$ is the number mod 2 of trajectories from p to r (by Lemma 6.5.4 there are only finitely many such trajectories), and we extend this to a group homomorphism

$$\partial : C_*(f, \mathbb{Z}_2) \to C_*(f, \mathbb{Z}_2).$$

Theorem 6.5.2. *We have*

$$\partial \circ \partial p = 0,$$

and thus $(C_*(f, \mathbb{Z}_2), \partial)$ *is a chain complex.*

Proof. We have

$$\partial \circ \partial p = \sum_{\substack{r \in C_*(f) \\ \mu(p,r)=1}} \sum_{\substack{q \in C_*(f) \\ \mu(r,q)=1}} \#_{\mathbb{Z}_2}\mathcal{M}^f_{p,r} \, \#_{\mathbb{Z}_2}\mathcal{M}^f_{r,q} q.$$

We are thus connecting the broken trajectories from p to q for $q \in C_*(f)$ with $\mu(p, q) = 2$, by Lemma 6.5.1. By Theorem 6.5.1 this number is always even, and so it vanishes mod 2. This implies $\partial \circ \partial p = 0$ for each $p \in C_*(f)$, and thus the extension to $C_*(f, \mathbb{Z}_2)$ also satisfies $\partial \circ \partial = 0$. □

We are now ready for

Definition 6.5.4. Let f be a C^3 function satisfying the Morse-Smale-Floer and Palais-Smale conditions, and assume that we have a compact space X of trajectories as investigated above. If we are in the situation of an absolute Morse index, we let

$C_k(f, \mathbb{Z}_2)$ be the group with coefficient in $\mathbb{Z}_2$ generated by the critical points of Morse index k. Otherwise, we choose an arbitrary grading in a consistent manner, i.e. we require that if $p \in C_k(f)$, $q \in C_l(f)$, then

$$k - l = \mu(p, q)$$

whenever the relative index is defined. We then obtain boundary operators

$$\partial = \partial_k : C_k(f, \mathbb{Z}_2) \to C_{k-1}(f, \mathbb{Z}_2),$$

and we define the associated homology groups as

$$H_k(X, f, \mathbb{Z}_2) := \frac{\ker \partial_k}{\operatorname{image} \partial_{k+1}},$$

i.e. two elements $\alpha_1, \alpha_2 \in \ker \partial_k$ are identified if there exists some $\beta \in C_{k+1}(f, \mathbb{Z}_2)$ with

$$\alpha_1 - \alpha_2 = \partial \beta.$$

Instead of a homology theory, we can also define a Morse-Floer cohomology theory by dualization. For that purpose, we put

$$C^k(f, \mathbb{Z}_2) := \operatorname{Hom}\left(C_k(f, \mathbb{Z}_2), \mathbb{Z}_2\right)$$

and define coboundary operators

$$\delta^k : C^k(f, \mathbb{Z}_2) \to C^{k+1}(f, \mathbb{Z}_2)$$

by

$$\delta^k \omega^k(p_{k+1}) = \omega^k(\partial_{k+1} p_{k+1})$$

for $\omega^k \in C^k(f, \mathbb{Z}_2)$ and $p_{k+1} \in C_k(f, \mathbb{Z}_2)$.

If there are only finitely many critical points $p_{1,k}, ..., p_{m,k}$ of index k, then we have a canonical isomorphism

$$\begin{aligned} &C_k(f, \mathbb{Z}_2) \to C^k(f, \mathbb{Z}_2), \\ &\qquad p_{j,k} \mapsto p_j^k \text{ with } p_j^k(p_{i,k}) = \delta_{ij} \quad (\delta_{ij} = 1 \text{ for } i = j \text{ and } 0 \text{ otherwise}) \end{aligned}$$

and

$$\delta^k p_j^k = \sum_{\substack{q_{i,k+1} \text{ critical point} \\ \text{of index } k+1}} p_{j,k}(\partial q_{i,k+1}) q_i^{k+1},$$

provided that sum is finite, too. Of course, this cohomology theory and the coboundary operator δ can also be constructed directly from the function f, by looking at the positive instead of the negative gradient flow, i.e. at the solution curves of

$$\begin{aligned} &y : \mathbb{R} \to X, \\ &\dot{y}(t) = \operatorname{grad} f(y(t)) \quad \text{for all } t. \end{aligned}$$

The preceding formalism then goes through in the same manner as before.

Remark. In certain infinite dimensional situations in the calculus of variations, there may be an analytic difference between the positive and negative gradient flow. Often, one faces the task of minimizing a certain function $f : X \to \mathbb{R}$ that is bounded from below, but not from above, and then also of finding other critical points of such a function. In such a situation, flow lines for the negative gradient flow

$$\dot{x}(t) = -\operatorname{grad} f(x(t))$$

might be well controlled, simply because f is decreasing on such a flow line, and therefore bounded, while along the positive gradient flow

$$\dot{y}(t) = \operatorname{grad} f(y(t)),$$

f may not be so well controlled, and one may not be able to derive the asymptotic estimates necessary for the analysis.

6.6 Orientations and $\mathbb{Z}$-homology

In the present section, we wish to consider the group $C_*(f, \mathbb{Z})$ with integer coefficients generated by the set $C_*(f)$ of critical points of f and define a boundary operator

$$\partial : C_*(f, \mathbb{Z}) \to C_*(f, \mathbb{Z})$$

satisfying

$$\partial \circ \partial = 0$$

as in the $\mathbb{Z}_2$-case, in order that $(C_*(f, \mathbb{Z}), \partial)$ be a chain complex. We assume that the general assumptions of §6.5 ($f \in C^3$, (PS), nondegeneracy of critical points, Morse-Smale-Floer condition) continue to hold.

We shall attempt to define ∂ as in Definition 6.5.3, by counting the number of connecting trajectories between critical points of relative index 1, but now we cannot simply take that number $\mod 2$, but we need to introduce a sign for each such trajectory and add the corresponding signs ± 1. In order to define these signs, we shall introduce orientations.

In order to motivate our subsequent construction, we shall first consider the classical case where X is a finite dimensional, compact, oriented, differentiable manifold. Let $f : X \to \mathbb{R}$ thus be a Morse function. The index $\mu(p)$ of a critical point p is the number of negative eigenvalues of $d^2 f(p)$, counted with multiplicity. The corresponding eigenvectors span the tangent space $V_p^u \subset T_p X$ of the unstable manifold $W^u(p)$ at p. We choose an arbitrary orientation of V_p^u, i.e. we select some basis

$e^1, \dots, e^{\mu(p)}$ of V_p^u as being positive. Alternatively, we may represent this orientation by $dx^1 \wedge \dots \wedge dx^{\mu(p)}$, where $dx^1, \dots, dx^{\mu(p)}$ are the cotangent vectors dual to $e^1, \dots, e^{\mu(p)}$.

As X is assumed to be oriented, we get an induced orientation of the tangent space $V_p^s \subset T_pX$ of the stable manifold $W^s(p)$ by defining a basis $e^{\mu(p)+1}, \dots, e^n$ ($n = \dim X$) as positive if $e^1, \dots, e^{\mu(p)}, e^{\mu(p)+1}, \dots, e^n$ is a positive basis of T_pX. In the alternative description, with $dx^{\mu(p)+1}, \dots, \dots dx^n$ dual to $e^{\mu(p)+1}, \dots, e^n$, the orientation is defined by $dx^{\mu(p)+1} \wedge \dots \wedge dx^n$ precisely if $dx^1 \wedge \dots dx^{\mu(p)} \wedge dx^{\mu(p)+1} \wedge \dots \wedge dx^n$ yields the orientation of T_pX.

Now if q is another critical point of f, of index $\mu(q) = \mu(p) - 1$, we choose any regular value a of f with $f(q) < a < f(p)$ and consider the intersection

$$W^u(p) \cap W^s(q) \cap f^{-1}(a).$$

The orientation of X also induces an orientation of $f^{-1}(a)$, because $f^{-1}(a)$ is always transversal to $\operatorname{grad} f$, and so we can consider a basis $\eta^2, \dots, \eta^n$ of $T_y f^{-1}(a)$ as positive if $\operatorname{grad} f(y), \eta^2, \dots, \eta^n$ is a positive basis of T_yX.

As we are assuming the Morse-Smale-Floer condition,

$$W^u(p) \cap W^s(q) \cap f^{-1}(a)$$

is a finite number of points by Lemma 6.5.4, and since $W^u(p), W^s(p)$ and $f^{-1}(a)$ all are equipped with an orientation, we can assign the sign $+1$ or -1 to any such intersection point depending on whether this intersection is positive or negative.

These intersection points correspond to the trajectories s of f from p to q, and we thus obtain a sign

$$n(s) = \pm 1$$

for any such trajectory, and we put

$$\partial p := \sum_{\substack{r \in C_*(f) \\ \mu(r) = \mu(p) - 1 \\ s \in \mathcal{M}^f_{p,r}}} n(s) r.$$

It thus remains to show that with this definition of the boundary operator ∂, we get the relation

$$\partial \circ \partial = 0.$$

In order to verify this, and also to free ourselves from the assumptions that X is finite dimensional and oriented and to thus preserve the generality achieved in the previous section, we shall now consider a relative version.

We let p, q be critical points of f connected by the flow with

$$\mu(p, q) = 2,$$

and we let $\mathcal{M}$ be a component of $\mathcal{M}^f_{p,q} = W^u(p) \cap W^s(q)$. For our subsequent analysis, only the second case of Theorem 6.5.1 will be relevant, i.e. where $\mathcal{M}$ has a boundary which then consists of two different broken trajectories from p to q. It is clear from the analysis of the proof of Theorem 6.5.1 that $\mathcal{M}$ is orientable. In fact, $\mathcal{M}$ is homeomorphic to the open disk, and it contains two transversal one-dimensional foliations, one consisting of the flow lines of f and the other one of the intersections of $\mathcal{M}$ with the level hypersurfaces $f^{-1}(a)$, $f(q) < a < f(p)$ (as $\mathcal{M}$ does not contain any critical points in its interior, all intersections with level hypersurfaces of f are transversal). We may thus choose an orientation of $\mathcal{M}$.

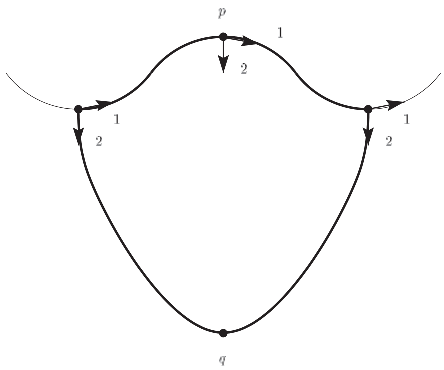

Figure 6.6.1:

This orientation then also induces orientations of the corner points of the broken trajectories in the boundary of $\mathcal{M}$ in the following sense: Let $s = s_1 \# s_2$ be such a broken trajectory, with intermediate critical point $r = s_1(\infty) = s_2(-\infty)$. The plane in T_rX spanned by $\dot{s_1}(\infty) := \lim_{t\to\infty} \dot{s_1}(t)$ and $\dot{s_2}(-\infty) := \lim_{t\to-\infty} \dot{s_2}(t)$ then is a limit of tangent planes of $\mathcal{M}$ and thus gets an induced orientation from $\mathcal{M}$.

This now implies that if we choose an orientation of s_1, we get an induced orientation of s_2, by requiring that if v_1, v_2 are positive tangent vectors of s_1 and s_2, resp. at r, then v_1, v_2 induces the orientation of the above plane in T_rX. Likewise, $\mathcal{M} \cap f^{-1}(a)$, for $f(q) < a < f(p)$ gets an induced orientation from the one of $\mathcal{M}$ and the one of the flow lines inside $\mathcal{M}$ which we always orient by $-\operatorname{grad} f$. Then the signs $n(s_1)$, $n(s_2)$ of s_1 and s_2, resp. are defined by checking whether s_1 resp. s_2 intersects these level hypersurfaces $f^{-1}(a)$ positively or negatively. Alternatively, what amounts to the same is simply checking whether s_1, s_2 have the orientation defined by $-\operatorname{grad} f$, or the opposite one, and thus, we do not even need the level hypersurfaces $f^{-1}(a)$.

Obviously, the problem now is that the choice of orientation of many trajectories connecting two critical points p, r of relative index $\mu(p, r) = 1$ depends on the choice of orientation of some such $\mathcal{M}$ containing s in its boundary, and the question is whether conversely, the orientations of these $\mathcal{M}$ can be chosen consistently in the sense that they all induce the same orientation of a given s. In the case of a finite dimensional, oriented manifold, this is no problem, because we get induced orientations on all such $\mathcal{M}$ from the orientation of the manifold and choices of orientations on all unstable manifolds, and these orientations fit together properly. In the general case, we need to make the global assumption that this is possible:

Definition 6.6.1. The Morse-Smale-Floer flow f is called orientable if we may define orientations on all trajectories $\mathcal{M}^f_{p,q}$ for critical points p, q with relative index $\mu(p,q)=2$ in such a manner that the induced orientations on trajectories s between critical points of relative index 1 are consistent.

With these preparations, we are ready to prove

Theorem 6.6.1. *Assume that the general assumptions ($f \in C^3$, (PS)), nondegeneracy of critical points, Morse-Smale-Floer conditions continue to hold, and that the flow is orientable in the sense of Definition 6.6.1. For the group $C_*(f,\mathbb{Z})$ generated by the set $C_*(f)$ of critical points of f, with integer coefficients, the operator*

$$\partial : C_*(f,\mathbb{Z}) \to C_*(f,\mathbb{Z})$$

defined by

$$\partial p := \sum_{\substack{r\in C_*(f)\\ \mu(p,r)=1\\ s\in\mathcal{M}^f_{p,r}}} n(s) r$$

for $p \in C_*(f)$ and linearly extended to $C_*(f,\mathbb{Z})$, satisfies

$$\partial \circ \partial = 0.$$

Thus, $C_*((f,\mathbb{Z}),\partial)$ becomes a chain complex, and we may define homology groups $H_k(X,f,\mathbb{Z})$ in the same manner as in Definition 6.5.4.

Proof. We have

$$\begin{aligned}
\partial \circ \partial p &= \sum_{\substack{q\in C_*(f)\\ \mu(r,q)=1\\ s_2\in\mathcal{M}^f_{r,q}}} \sum_{\substack{r\in C_*(f)\\ \mu(p,r)=1\\ s_1\in\mathcal{M}^f_{r,p}}} n(s_2)n(s_1)q \\
&= \sum_{\substack{q\in C_*(f)\\ \mu(p,q)=2\\ (s_1,s_2)\text{ broken trajectory}\\ \text{from } p \text{ to } q}} n(s_2)n(s_1)q.
\end{aligned}$$

By Theorem 6.5.1, these broken trajectories always occur in pairs $({s'}_1, {s'}_2)$, $({s''}_1, {s''}_2)$ bounding some component $\mathcal{M}$ of $\mathcal{M}^f_{p,q}$.

It is then geometrically obvious, see Figure 6.6.1, that

$$n(s_1')n(s_2') = -n(s_1'')n(s_2'').$$

Thus, the contributions of the two members of each such pair cancel each other, and the preceding sum vanishes. $\square$

In the situation of Theorem 6.6.1, we put

$$b_k(X, f) := \dim_{\mathbb{Z}} H_k(X, f, \mathbb{Z}).$$

We shall see in §§6.7, 6.9 that these numbers in fact do not depend on f. As explained at the end of the preceding section, one may also construct a dual cohomology theory, with

$$C^k(f, \mathbb{Z}) := \mathrm{Hom}\,(C_k(f, \mathbb{Z}), \mathbb{Z})$$

and coboundary operators

$$\delta^k : C^k(f, \mathbb{Z}) \to C^{k+1}(f, \mathbb{Z})$$

with

$$\delta^k \omega^k (p_{k+1}) = \omega^k (\partial_{k+1} p_{k+1})$$

for $\omega^k \in C^k(f, \mathbb{Z}), p_{k+1} \in C_{k+1}(f, \mathbb{Z})$.

6.7 Homotopies

We have constructed a homology theory for a Morse-Smale-Floer function f on a manifold X, under the preceding assumptions. In order to have a theory that captures invariants of X, we now ask to what extent the resulting homology depends on the choice of f. To formulate the question differently, given two such functions f^1, f^2, can one construct an isomorphism between the corresponding homologies? If so, is this isomorphism canonical?

A first geometric approach might be based on the following idea, considering again the case of a finite dimensional, compact manifold:

Given a critical point p of f^1 of Morse index μ, and a critical point q of f^2 of the same Morse index, the unstable manifold of p has dimension μ, and the stable one of q dimension $n - \mu$ if $n = \dim X$. Thus, we expect that generally, these two manifolds intersect in finitely many points $x_1, \ldots, x_k$ with signs $n(x_j)$ given by the sign of the intersection number, and we might put

$$\phi^{21}(p) = \sum_{\substack{q \in C_*(f^2) \\ \mu_{f^2}(q) = \mu_{f^1}(p)}} \sum_{x \in W^u_{f^1}(p) \cap W^s_{f^2}(q)} n(x) q \tag{6.7.1}$$

(we introduce additional indices f^1, f^2 in order to indicate the source of the objects) to get a map

$$\phi^{21} : C_*(f^1, G) \to C_*(f^2, G)$$

extended to coefficients $G = \mathbb{Z}_2$ or $\mathbb{Z}$ that hopefully commutes with the boundary operators ∂^{f^1}, ∂^{f^2} in the sense that

$$\phi^{21} \circ \partial^{f^1} = \partial^{f^2} \circ \phi^{21}. \tag{6.7.2}$$

One difficulty is that for such a construction, we need the additional assumption that the unstable manifolds for f^1 intersect the stable ones for f^2 transversally. Even if f^1 and f^2 are Morse-Smale-Floer functions, this need not hold, however. For example, one may consider $f^2 = -f^1$; then for any critical point p,

$$W^u_{f^1}(p) = W^s_{f^2}(p)$$

which is not compatible with transversality.

Of course, one may simply assume that all such intersections are transversal but that would not be compatible with our aim to relate the homology theories for any pair of Morse-Smale-Floer functions in a canonical manner. We note, however, that the construction would work in the trivial case where $f^2 = f^1$, because then $W^u_{f^1}(p)$ and $W^s_{f^2}(p) = W^s_{f^1}(p)$ intersect precisely at the critical point p itself.

In order to solve this problem, we consider homotopies

$$F : X \times \mathbb{R} \to \mathbb{R}$$

with

$$\lim_{t \to -\infty} F(x,t) = f^1(x), \quad \lim_{t \to \infty} F(x,t) = f^2(x), \quad \text{for all } x \in X.$$

In fact, for technical reasons it will be convenient to impose the stronger requirement that

$$\begin{aligned} F(x,t) &= f^1(x) \quad \text{for } t \le -R, \\ F(x,t) &= f^2(x) \quad \text{for } t \ge R, \end{aligned} \tag{6.7.3}$$

for some $R > 0$.

Given such a function F, we consider the flow

$$\begin{aligned} \dot{x}(t) &= -\operatorname{grad} F(x(t),t) \quad \text{for } t \in \mathbb{R}, \\ x(0) &= x, \end{aligned} \tag{6.7.4}$$

where grad denotes the gradient w.r.t. the x-variables. In order to avoid trouble with cases where this gradient is unbounded, one may instead consider the flow

$$\dot{x}(t) = \frac{-1}{\sqrt{1 + \left|\frac{\partial F}{\partial t}\right| |\operatorname{grad} F|^2}} \operatorname{grad} F(x(t),t), \tag{6.7.5}$$

but for the moment, we ignore this point and consider (6.7.4) for simplicity.

If p and q are critical points of f^1 and f^2, resp., with index μ the strategy then is to consider the number of flow lines $s(t)$ of (6.7.5) with

$$\begin{aligned} s(-\infty) &= p, \\ s(\infty) &= q, \end{aligned}$$

equipped with appropriate signs $n(s)$, denote the space of these flow lines by $\mathcal{M}^F_{p,q}$, and put

$$\phi^{21}(p) = \sum_{\substack{q \in C_*(f^2) \\ \mu(q)=\mu(p)}} \sum_{s \in \mathcal{M}^F_{p,q}} n(s) q. \tag{6.7.6}$$

Let us again discuss some trivial examples:
If $f^1 = f^2$ and F is the constant homotopy, then clearly

$$\phi^{21}(p) = p,$$

for every critical point p. If $f^2 = -f^1$ and we construct F by

$$F(x,t) := \begin{cases} f^1(x) & \text{for } -\infty < t \le -1, \\ -tf^1(t) & \text{for } -1 \le t \le 1, \\ -f^1(x) & \text{for } \quad 1 \le t < \infty. \end{cases} \tag{6.7.7}$$

we have

$$s(t) = s(-t) \tag{6.7.8}$$

for any flow line. Thus, also

$$s(\infty) = s(-\infty),$$

and a flow line cannot connect a critical p of f^1 of index μ^{f^1} with a critical point q of f^2 of index $\mu^{f^2} = n - \mu^{f^1}$, unless $p = q$ and $\mu^{f^2} = \frac{n}{2}$. Consequently, we seem to have the same difficulty as before. This is not quite so, however, because we now have the possibility to perturb the homotopy if we wish to try to avoid such a peculiar behavior. In other words, we try to employ only generic homotopies.

In order to formulate what we mean by a generic homotopy we recall the concept of a Morse function. There, we required that the Hessian $d^2 f(x_0)$ at a critical point is nondegenerate. At least in the finite dimensional case that we consider at this moment, this condition is generic in the sense that the Morse functions constitute an open and dense subset of the set of all C^2 functions on X. The Morse condition means that at a critical point x_0, the linearization of the equation

$$\dot{x}(t) = -\operatorname{grad} f(x(t))$$

has maximal rank. A version of the implicit function theorem then implies that the linearization of the equation locally already describes the qualitative features of the original equation. In this sense, we formulate

Definition 6.7.1. The homotopy F satisfying (6.7.6) is called regular if whenever

$$\operatorname{grad} F(x_0, t) = 0 \quad \text{for all } t \in \mathbb{R},$$

the operator

$$\frac{\partial}{\partial t} + d^2 F(x_0, t) : H^{1,2}(x_0^* TX) \to L^2(x_0^* TX)$$

is surjective.

This is satisfied for a constant homotopy, if f^1 is a Morse function, but not for the homotopy (6.7.7) because in that case only sections satisfying (6.7.8) are contained in the range of $\frac{\partial}{\partial t} + d^2F(x_0, t)$.

Let us continue with our heuristic considerations:

If f^1 is a Morse function as before, $\varphi : (-\infty, 0] \to \mathbb{R}^+$ satisfies $\varphi(t) = 1$ for $t \leq -1$, $\varphi(0) = 0$, we consider the flow

$$\begin{aligned} \dot{x}(t) &= -\varphi(t) \operatorname{grad} f^1(x(t)) \quad \text{for } -\infty < t \leq 0, \\ x(0) &= x. \end{aligned}$$

We obtain a solution for every $x \in X$, and as before $x(-\infty)$ always is a critical point of f^1. Thus, while all the flow lines emanate at a critical point for $t = -\infty$, they cover the whole manifold at $t = 0$. If we now extend φ to $(0, \infty)$ by putting

$$\varphi(t) := \varphi(-t) \quad \text{for } t \geq 0,$$

and if we have another Morse function f^2 and put

$$\dot{x}(t) = -\varphi(t) \operatorname{grad} f^2(x(t)) \quad \text{for } t \geq 0,$$

in the same manner, the flow lines will converge to critical points of f^2 at $t = \infty$. We thus relate the flow asymptotic regimes governed by f^1 and f^2 through the whole manifold X at an intermediate step. Of course, this only works under generic conditions, and we may have to deform the flow slightly to achieve that, but here we rather record the following observation: The points $x(0)$ for flow lines with $x(-\infty) = p$ cover the unstable manifolds of the critical point p of f^1, and likewise the points $x(0)$ for the flow lines with $x(\infty) = q$ for the critical point q of f^2 cover the stable manifold of q. Thus the flow lines with $x(-\infty) = p$, $x(\infty) = q$ correspond to the intersection of the unstable manifold of p (w.r.t. f^1) with the stable manifold of q (w.r.t. f^2), and we now have the flexibility to deform the flow if problems arise from nontransversal intersections.

Let us return once more to the trivial example $f^1 = f^2$, and a constant homotopy F. We count the flow lines not in X, but in $X \times \mathbb{R}$. This simply means that in contrast to the situation in previous sections, we now consider the flow lines $x(\cdot)$ and $x(\cdot + t_0)$, for some fixed $t_0 \in \mathbb{R}$, as different. Of course, if the homotopy F is not constant in t, the time shift invariance is broken anyway, and in a certain sense this is the main reason for looking at the nonautonomous equation (6.7.4) as opposed to the autonomous one $\dot{x}(t) = -\operatorname{grad} f(x(t))$ considered previously. Returning for a moment to our constant homotopy, if p and q are critical points of indices $\mu(p)$ and $\mu(q) = \mu(p) - 1$, resp. , connected by the flow of f^1, the flow lines for F cover a two-dimensional region in $X \times \mathbb{R}$. This region is noncompact, and it can be compactified by adding broken trajectories of the type

$$s_1 \# s_2$$

where s_1 is a flow for f^1 from p to q and s_2 is the constant flow line for $f^1 = f^2$ from p to q. This looks analogous to the situation considered in §6.5, and in fact with the same methods one shows the appropriate analogue of Theorem 6.5.1. When it come to orientations, however, there is an important difference. Namely, in the situation of Figure 6.7.1 (where we have compactified $\mathbb{R}$ to a bounded interval), the two broken trajectories from p to q in the boundary of the square should now be given the same orientation if we wish to maintain the aim that the homotopy given through (6.7.6) commutes with the boundary operator even in the case of coefficients in $\mathbb{Z}$.

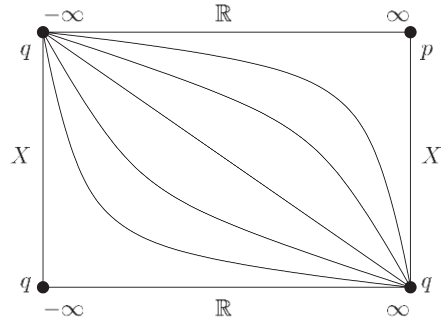

Figure 6.7.1:

The considerations presented here only in heuristic terms will be taken up with somewhat more rigour in §6.9 below.

6.8 Graph flows

In this section, we shall assume that X is a compact, oriented Riemannian manifold. A slight variant of the construction of the preceding section would be the following:

Let f_1, f_2 be two Morse-Smale-Floer functions, as before. In the preceding section, we have treated the general situation where the unstable manifolds of f_1 need not intersect the stable ones of f_2 transversally. The result was that there was enough flexibility in the choice of homotopy between f_1 and f_2 so that that did not matter. In fact, a consequence of that analysis is that we may always find a sufficiently small perturbation of either one of the two functions so that such a transversality property holds, without affecting the resulting algebraic invariants.

Therefore from now on, we shall assume that for all Morse-Smale-Floer functions $f_1, f_2, \ldots$ occuring in any construction in the sequel, all unstable manifolds of any one of them intersect all the stable manifolds of all the other functions transversally. We call this the **generalized Morse-Smale-Floer condition**.

Thus, assuming that property, we consider continuous paths

$$x : \mathbb{R} \to X$$

with

$$\dot{x}(t) = -\operatorname{grad} f_i(x(t)), \quad \text{with } i = \begin{cases} 1 & \text{for } t < 0, \\ 2 & \text{for } t > 2. \end{cases}$$

The continuity requirement then means that we are switching at $t = 0$ in a continuous manner from the flow for f_1 to the one for f_2. As we are assuming the generalized Morse-Smale-Floer condition, this can be utilized in the manner described in the previous section to equate the homology groups generated by the critical points of f_1 and f_2 resp.

This construction admits an important generalization:

Let Γ be a finite oriented graph with n edges, n_1 of them parametrized by $(-\infty, 0]$, n_2 parametrized by $[0, \infty)$, and the remaining ones by $[0, 1]$. We also assume that to each edge e_i of Γ, there is associated a Morse-Smale-Floer function f_i and that the generalized Morse-Smale-Floer condition holds for this collection $f_1, \ldots, f_n$.

Definition 6.8.1. A continuous map $x : \Gamma \to X$ is called a solution of the graph flow for the collection $(f_1, \ldots, f_n)$ if

$$\dot{x}(t) = -\operatorname{grad} f_i(x(t)) \quad \text{for } t \in e_i. \tag{6.8.1}$$

Again, the continuity requirement is relevant only at the vertices of Γ as the flow is automatically smooth in the interior of each edge. If $p_1, \ldots, p_{n_1}$ are critical points for the functions $f_1, \ldots, f_{n_1}$ resp. corresponding to the edges $e_1, \ldots, e_{n_1}$ parametrized on $(-\infty, 0]$, $p_{n_1+1}, \ldots, p_{n_1+n_2}$ critical points corresponding to the edges $e_{n_1+1}, \ldots, e_{n_1+n_2}$ resp. parametrized on $[0, \infty)$, we let $\mathcal{M}^{\Gamma}_{p_1,\ldots,p_{n_1+n_2}}$ be the space of all solutions of (6.8.1) with

$$\lim_{\substack{t \to -\infty \\ t \in e_i}} x(t) = p_i \quad \text{for } i = 1, \ldots, n_1,$$

$$\lim_{\substack{t \to \infty \\ t \in e_i}} x(t) = p_i \quad \text{for } i = n_1 + 1, \ldots, n_1 + n_2,$$

i.e. we assume that on each edge e_i, $i = 1, \ldots, n_1 + n_2$, $x(t)$ asymptotically approaches the critical p_i of the function f_i.

If X is a compact Riemannian manifold of dimension d, we have

Theorem 6.8.1. *Assume, as always in this section, the generalized Morse-Smale-Floer condition. Then* $\mathcal{M}^{\Gamma}_{p_1,\ldots,p_{n_1+n_2}}$ *is a smooth manifold, for all tuples* $(p_1, \ldots, p_{n_1+n_2})$*, where* p_i *is a critical point of* f_i*, with*

$$\dim \mathcal{M}^{\Gamma}_{p_1,\ldots,p_{n_1+n_2}} = \sum_{i=1}^{n_1} \mu(p_i) - \sum_{j=n_1+1}^{n_1+n_2} \mu(p_j) - d(n_1 - 1) - d \dim H_1(\Gamma, \mathbb{R}), \tag{6.8.2}$$

where $\mu(p_k)$ *is the Morse index of the critical point* p_k *for the function* f_k*.*

Proof. We simply need to count the dimensions of intersections of the relevant stable and unstable manifolds for the edges modeled on $[0, \infty)$ and $(-\infty, 0]$ and the contribution of internal loops. Each unstable manifold corresponding to a point $p_i, i = n_1 + 1, \ldots, n_1 + n_2$ has dimension $d - \mu(p_i)$. If a submanifold X_1 of X is intersected transversally by another submanifold X_2, then the intersection has dimension $d - (d - \dim X_1) - (d - \dim X_2)$, and this accounts for the first three terms in (6.8.2). If we have an internal loop in Γ, this reduces the dimension by d, as the following argument shows:

Let Γ be constituted by two e_1, e_2 with common end points, and let the associated Morse functions be f_1, f_2, resp. For f_i, $i = 1, 2$, we consider the graph of the flow induced by that function, i.e. we associate to each $x \in X$ the point $x_i(1)$, where x_i is the solution of $\dot{x}_i(t) = -\operatorname{grad} f_i(x_i(t))$, $x_i(0) = x$. These two graphs for f_1 and f_2 are then submanifolds of dimension d of $X \times X$, and if they intersect transversally, they do so in isolated points, as $\dim(X \times X) = 2d$. Thus, if we start with a d-dimensional family of initial points, we get a finite number of common end points. $\square$

Again $\mathcal{M}^{\Gamma}_{p_1,\ldots,p_{n_1+n_2}}$ is not compact, but can be compactified by flows with broken trajectories on the noncompact edges of Γ.

The most useful case of Theorem 6.8.1 is the one where the dimension of $\mathcal{M}^{\Gamma}_{p_1,\ldots,p_{n_1+n_2}}$ is 0. In that case, $\mathcal{M}^{\Gamma}_{p_1,\ldots,p_{n_1+n_2}}$ consists of a finite number of continuous maps $x : \Gamma \to X$ solving (6.8.1) that can again be given appropriate signs. The corresponding sum is denoted by

$$n(\Gamma; p_1, \ldots p_{n_1+n_2}).$$

We then define a map

$$\begin{aligned} q(\Gamma) : \bigotimes_{i=1}^{n_1} C_*(f_i, \mathbb{Z}) &\to \bigotimes_{j=n_1+1}^{n_1+n_2} C_*(f_j, \mathbb{Z}), \\ (p_1 \otimes \cdots \otimes p_{n_1}) &\mapsto n(\Gamma; p_1, \ldots, p_{n_1+n_2})(p_{n_1+1} \otimes \cdots \otimes p_{n_1+n_2}). \end{aligned}$$

With

$$C^*(f_i, \mathbb{Z}) := \operatorname{Hom}(C_*(f_i, \mathbb{Z}), \mathbb{Z}),$$

we may consider $q(\Gamma)$ as an element of

$$\bigotimes_{i=1}^{n_1} C^*(f_i, \mathbb{Z}) \bigotimes_{j=n_1+1}^{n_1+n_2} C_*(f_j, \mathbb{Z}).$$

With the methods of the previous section, one verifies

Lemma 6.8.1. $\partial q = 0$.

Consequently, we consider $q(\Gamma)$ also as an element of

$$\bigotimes_{i=1}^{n_1} H^*(f_i, \mathbb{Z}) \bigotimes_{j=n_1+1}^{n_1+n_2} H_*(f_j; \mathbb{Z}).$$

Besides the above example where Γ had the edges $(-\infty, 0]$ and $[0, \infty)$, there are other examples of topological significance:

1) $\Gamma = [0, \infty)$. Thus, $n_1 = 0$, $n_2 = 1$, and with $p = p_{n_1} = p_1$,

$$\dim \mathcal{M}_p^\Gamma = d - \mu(p).$$

This is 0 precisely if $\mu(p) = d$, i.e. if p is a local maximum. In that case $q(\Gamma) \in H_d(X; \mathbb{Z})$ is the so-called fundamental class of X.

2) Γ consisting of two edges modeled on $(-\infty, 0]$, and joined by identifying the two right end points 0. Thus $n_1 = 2$, $n_2 = 0$, and

$$\dim \mathcal{M}_{p_1,p_2}^\Gamma = \mu(p_1) + \mu(p_2) - d,$$

and this is 0 if $\mu(p_2) = d - \mu(p_1)$. With $k := \mu(p_1)$, thus

$$\begin{aligned} q(\Gamma) &\in H^k(X, \mathbb{Z}) \otimes H^{d-k}(X, \mathbb{Z}) \\ &\cong \operatorname{Hom}(H_k(X, \mathbb{Z}), H^{d-k}(X, \mathbb{Z})) \end{aligned}$$

is the so-called Poincaré duality isomorphism.

3) Γ consisting of one edge modeled on $(-\infty, 0]$, and two ones modeled on $[0, \infty)$, all three identified at the common point 0. Thus $n_1 = 1, n_2 = 2$, and

$$\dim \mathcal{M}_{p_1,p_2,p_3}^\Gamma = \mu(p_1) - \mu(p_2) - \mu(p_3).$$

Hence, if this is 0,

$$\begin{aligned} q(\Gamma) &\in \bigotimes_{j \le k} H^k(K, \mathbb{Z}) \otimes H_j(X, \mathbb{Z}) \otimes H_{k-j}(X, \mathbb{Z}) \\ &\cong \bigotimes_{j \le k} \operatorname{Hom}(H^j(X, \mathbb{Z}) \otimes H^{k-j}(X, \mathbb{Z}), H^k(X, \mathbb{Z})). \end{aligned}$$

We thus obtain a product

$$\cup : H^j(X, \mathbb{Z}) \otimes H^{k-j}(X, \mathbb{Z}) \to H^k(X, \mathbb{Z}),$$

the so-called cup product.

4) Γ consisting of one edge $(-\infty, 0]$ together with a closed loop based at 0. In that case

$$\dim \mathcal{M}_p^\Gamma = \mu(p) - d,$$

which vanishes for $\mu(p) = d$, i.e.

$$q(\Gamma) \in H^d(X, \mathbb{Z}).$$

This cohomology class is called the Euler class.

6.9 Orientations

We are considering solution curves of

$$\dot{x}(t) + \operatorname{grad} f(x(t)) = 0, \tag{6.9.1}$$

or more generally of

$$\dot{x}(t) + \operatorname{grad} F(x(t), t) = 0, \tag{6.9.2}$$

and we wish to assign a sign to each such solution in a consistent manner.

For that purpose, we linearize those equations. We consider a curve $x(t)$ of class $H^{1,2}(\mathbb{R}, X)$ and a section $\varphi(t)$ of class $H^{1,2}$ of the tangent bundle of X along x, i.e. $\varphi \in H^{1,2}(\mathbb{R}, x^*TX)$. Then, in the case of (6.9.1), the linearization is

$$\nabla_{\frac{d}{ds}} \Big((\exp_x(t) s\varphi(t))^{\bullet} + \operatorname{grad} f(\exp_{x(t)} s\varphi(t)) \Big)_{|s=0} = \nabla_{\frac{d}{dt}} \varphi(t) + D_{\varphi(t)} \operatorname{grad} f(x(t)) \,,$$

with $\nabla_{\frac{d}{dt}} := \nabla_{\dot{x}(t)}$, ∇ the Levi-Civita connection of X, and likewise, for (6.9.2), we get

$$\nabla_{\frac{d}{dt}} \varphi(t) + D_{\varphi(t)} \operatorname{grad} F(x(t), t).$$

We shall thus consider the operator

$$\begin{aligned} \nabla_{\dot{x}} + D \operatorname{grad} F : H^{1,2}(x^*TX) &\to L^2(x^*TX), \\ \varphi &\mapsto \nabla_{\dot{x}}\varphi + D_{\varphi} \operatorname{grad} F. \end{aligned} \tag{6.9.3}$$

This is an operator of the form

$$\nabla + A : H^{1,2}(x^*TX) \to L^2(x^*TX),$$

where A is a smooth section of $x^*\operatorname{End} TX$ which is selfadjoint, i.e. for each $t \in \mathbb{R}$, $A(t)$ is a selfadjoint linear operator on $T_{x(t)}X$.

We are thus given a vector bundle E on $\mathbb{R}$ and an operator

$$\nabla + A : H^{1,2}(E) \to L^2(E),$$

with A a selfadjoint endomorphism of E. $H^{1,2}(E)$ and $L^2(E)$ are Hilbert spaces, and $\nabla + A$ will turn out to be a Fredholm operator if we assume that A has boundary values $A(\pm\infty)$ at $\pm\infty$.

Let $L : V \to W$ be a continuous linear operator between Hilbert spaces V, W, with associated norms $\|\cdot\|_V$, $\|\cdot\|_W$ resp. (we shall often omit the subscripts ${}_{V,W}$ and simply write $\|\cdot\|$ in place of $\|\cdot\|_V$ or $\|\cdot\|_W$). L is called a Fredholm operator iff

(i) $V_0 := \ker L$ is finite dimensional,

(ii) $W_1 := L(V)$, the range of L, is closed and has finite dimensional complement $W_0 =: \operatorname{coker} L$, i.e.
$$W = W_1 \oplus W_0.$$

From (i), we infer that there exists a closed subspace V_1 of V with

$$V = V_0 \oplus V_1,$$

and the restriction of L to V_1 is a bijective continuous linear operator $L^{-1} : V_1 \to W_1$.

By the inverse operator theorem,

$$L^{-1} : W_1 \to V_1$$

then is also a bijective continuous linear operator. We put

$$\begin{aligned}\operatorname{ind} L : &= \dim V_0 - \dim W_0 \\ &= \dim \ker L - \dim \operatorname{coker} L.\end{aligned}$$

The set of all Fredholm operators from V to W is denoted by $F(V, W)$.

Lemma 6.9.1. *$F(V, W)$ is open in the space of all continuous linear operators from V to W, and*

$$\operatorname{ind} : F(V, W) \to \mathbb{Z}$$

is continuous, and therefore constant on each component of $F(V, W)$.

For a proof, see e.g. [149].

By trivializing E along $\mathbb{R}$, we may simply assume $E = \mathbb{R}^n$, and we thus consider the operator

$$\frac{d}{dt} + A(t) : H^{1,2}(\mathbb{R}, \mathbb{R}^n) \to L^2(\mathbb{R}, \mathbb{R}^n), \tag{6.9.4}$$

and we assume that $A(t)$ is continuous in t with boundary values

$$A(\pm\infty) = \lim_{t \to \pm\infty} A(t),$$

and that $A(-\infty)$ and $A(\infty)$ are nondegenerate. In particular, since these limits exists, we may assume that

$$\|A(t)\| \leq \text{const.},$$

independently of t. For a selfadjoint $B \in Gl(n, \mathbb{R})$, we denote by

$$\mu(B)$$

the number of negative eigenvalues, counted with multiplicity.

Lemma 6.9.2. *$L_A := \frac{d}{dt} + A(t) : H^{1,2}(\mathbb{R}, \mathbb{R}^n) \to L^2(\mathbb{R}, \mathbb{R}^n)$ is a Fredholm operator with*

$$\operatorname{ind} L_A = \mu(A(-\infty)) - \mu(A(\infty)).$$

Proof. We may find a continuous map $C : \mathbb{R} \to \mathrm{Gl}(n, \mathbb{R})$ and continuous functions $\lambda_1(t), \ldots, \lambda_n(t))$ such that

$$C(t)^{-1} A(t) C(t) = \mathrm{diag}\,(\lambda_1(t), \ldots, \lambda_n(t)), \quad \lambda_1(t) \leq \lambda_2(t) \leq \cdots \leq \lambda_n(t),$$

i.e. we may diagonalize the selfadjoint linear operators $A(t)$ in a continuous manner. By continuously deforming $A(t)$ (using Lemma 6.9.1), we may also assume that $A(t)$ is asymptotically constant, i.e. there exists $T > 0$ with

$$\begin{aligned} A(t) &= A(-\infty) \quad \text{for } t \leq -T, \\ A(t) &= A(\infty) \quad\ \ \text{for } t \geq T. \end{aligned}$$

Thus, $C(t)$, $\lambda_1(t), \ldots, \lambda_n(t)$ are also asymptotically constant. If $s(t)$ is in $H^{1,2}$, then it is also continuous, and hence if it solves

$$\frac{d}{dt} s(t) + A(t) s(t) = 0,$$

then it is also of class C^1, since $\frac{d}{dt} s(t) = -A(t) s(t)$ is continuous. On $(-\infty, -T]$, it has to be a linear combination of the functions

$$e^{-\lambda_i(-\infty) t},$$

and on $[T, \infty)$, it is a linear combination of

$$e^{-\lambda_i(\infty) t}, \quad i = 1, \ldots, n.$$

Since a solution on $[-T, T]$ is uniquely determined by its values at the boundary points $\pm T$, we conclude that the space of solutions is finite dimensional. In fact, the requirement that s be in $H^{1,2}$ only allows linear combinations of those exponential functions of the above type with $\lambda_i(-\infty) < 0$, on $(-\infty, -T)$, and likewise we get the condition $\lambda_i(\infty) > 0$. Thus

$$\dim \ker L_A = \max(\mu(A(-\infty)) - \mu(A(\infty)), 0)$$

is finite.

Now let $\sigma \in L^2(\mathbb{R}, \mathbb{R}^n)$ be in the orthogonal complement of the image of L_A, i.e.

$$\int \left(\frac{d}{dt} s(t) + A(t) s(t) \right) \cdot \sigma(t)\, dt = 0 \quad \text{for all } s \in H^{1,2}(\mathbb{R}, \mathbb{R}^n),$$

where the "$\cdot$" denotes the Euclidean scalar product in $\mathbb{R}^n$. In particular, this relation implies that the weak derivative $\frac{d}{dt} \sigma(t)$ equals $-A(t)\sigma(t)$, hence is in L^2. Thus $\sigma \in H^{1,2}(\mathbb{R}, \mathbb{R}^n)$ is a solution of

$$\frac{d}{dt} \sigma(t) - A(t) \sigma(t) = 0.$$

In other words, L_A has $-L_{-A}$ as its adjoint operator, which then by the above argument satisfies

$$\begin{aligned}\dim\ker L_{-A} &= \max(\mu(-A(-\infty)) - \mu(-A(\infty)), 0)\\ &= \max(\mu(A(\infty)) - \mu(A(-\infty)), 0).\end{aligned}$$

L_A then has as its range the orthogonal complement of the finite dimensional space $\ker L_{-A}$, which then is closed, and

$$\begin{aligned}\operatorname{ind} L_A &= \dim\ker L_A - \dim\operatorname{coker} L_A\\ &= \dim\ker L_A - \dim\ker L_{-A}\\ &= \mu(A(-\infty)) - \mu(A(\infty)).\end{aligned}$$

□

Corollary 6.9.1. *Let x_1, x_2 be $H^{1,2}$ curves in X, E_i vector bundles along x_i, A_i continuous selfadjoint sections of* $\operatorname{End} E_i$, *$i = 1, 2$, with $x_1(\infty) = x_2(-\infty)$, $E_1(\infty) = E_2(-\infty)$, $A_1(\infty) = A_2(-\infty)$. We assume again that $A_1(-\infty), A_1(\infty) = A_2(-\infty)$, $A_2(\infty)$ are nondegenerate. We consider diffeomorphisms*

$$\sigma_1 : (-\infty, 0) \to \mathbb{R}, \quad \sigma_2 : (0, \infty) \to \mathbb{R},$$

with $\sigma_t(t) = t$ for $|t| \geq T$ for some $T > 0$, $i = 1, 2$, and consider the curve

$$x(t) := \begin{cases} x_1(\sigma_1(t)) & \text{for } t < 0,\\ x_1(\infty) = x_2(-\infty) & \text{for } t = 0,\\ x_2(\sigma_2(t)) & \text{for } t > 0,\end{cases}$$

with the corresponding bundle $E(t)$ and $A(t)$ glued together from E_1, E_2, A_1, A_2, resp. in the same manner. Then

$$\operatorname{ind} L_A = \operatorname{ind} L_{A_1} + \operatorname{ind} L_{A_2}.$$

Proof.

$$\begin{aligned}\operatorname{ind} L_{A_1} + \operatorname{ind} L_{A_2} &= \mu(A_1(-\infty)) - \mu(A_1(\infty)) + \mu(A_2(-\infty)) - \mu(A_2(\infty))\\ &= \mu(A(-\infty)) - \mu(A(\infty))\\ &= \operatorname{ind} L_A, \qquad \text{by Lemma 6.9.2 and construction.}\end{aligned}$$

□

We now need to introduce the notion of the determinant of a Fredholm operator. In order to prepare that definition, we first let V, W be finite dimensional vector spaces of dimension m, equipped with inner products, and put

$$\operatorname{Det} V := \Lambda^m(V), \quad \text{with } \Lambda^0 V := \mathbb{R}.$$

Then $(\mathrm{Det}\, V)^* \otimes \mathrm{Det}\, V$ is canonically isomorphic to $\mathbb{R}$ via $v^* \otimes w \mapsto v^*(w)$. A linear map

$$l : V \to W$$

then induces

$$\det l : \mathrm{Det}\, V \to \mathrm{Det}\, W,$$

i.e.

$$\det l \in (\mathrm{Det}\, V)^* \otimes \mathrm{Det}\, W.$$

The transformation behavior w.r.t. bases $e_1, \dots, e_m$ of V, $f_1, \dots, f_m$ of W is given by

$$\det l(e_1 \wedge \dots \wedge e_m) = le_1 \wedge \dots \wedge le_m =: \Delta_l f_1 \wedge \dots \wedge f_m.$$

We may e.g. use the inner product on W to identify the orthogonal complement of $l(V)$ with $\mathrm{coker}\, l$. The exact sequence

$$0 \to \ker l \to V \xrightarrow{l} W \to \mathrm{coker}\, l \to 0$$

and the multiplicative properties of det allow the identification

$$(\mathrm{Det}\, V)^* \otimes \mathrm{Det}\, W \cong (\mathrm{Det}\, \ker l)^* \otimes \mathrm{Det}\,(\mathrm{coker}\, l) =: \mathrm{Det}\, l.$$

This works as follows:

Put $V_0 = \ker l$, $W_0 = \mathrm{coker}\, L$ $(= l(V)^\perp)$, and write $V = V_0 \otimes V_1$, $W = W_0 \otimes W_1$. Then

$$l_1 := l_{|V_1} : V_1 \to W_1$$

is an isomorphism, and if $e_1, \dots, e_k$ is a basis of V_0, $e_{k+1}, \dots e_m$ one of V_1, $f_1, \dots f_k$ one of W_0, and if we take the basis $le_{k+1}, \dots, le_m$ of W_1, then

$$(e_1 \wedge \dots e_k \wedge e_{k+1} \wedge \dots \wedge e_m)^* \otimes (f_1 \wedge \dots \wedge f_k \wedge le_{k+1} \wedge \dots \wedge le_m)$$

is identified with

$$(e_1 \wedge \dots \wedge e_m)^* \otimes (f_1 \wedge \dots \wedge f_m).$$

According to the rules of linear algebra, this identification does not depend on the choices of the basis. In this manner, we obtain a trivial line bundle over $V^* \otimes W$, with fiber $(\mathrm{Det}\, V)^* \otimes \mathrm{Det}\, W \cong (\mathrm{Det}\, \ker l)^* \otimes \mathrm{Det}\, \mathrm{coker}\, l$ over l. $\det l$ then is a section of this line bundle, vanishing precisely at those l that are not of maximal rank m. On the other hand, if l is of maximal rank, then $(\mathrm{Det}\, \ker l)^* \otimes \mathrm{Det}\, \mathrm{coker}\, l$ can be canonically identified with $\mathbb{R}$, and $\det l$ with $1 \in \mathbb{R}$, by choosing basis $e_1, \dots, e_m$ of V and the basis $le_1, \dots le_m$ of W, as above.

In a more abstract manner, this may also be derived from the above exact sequence

$$0 \to \ker l \to V \xrightarrow{l} W \to \mathrm{coker}\, l \to 0$$

on the basis of the following easy algebraic

Lemma 6.9.3. *Let* $0 \to V_1 \xrightarrow{l_1} V_2 \xrightarrow{l_2} \dots \xrightarrow{l_{k-1}} V_k \to 0$ *be an exact sequence of linear maps between finite dimensional vector spaces. Then there exists a canonical isomorphism*

$$\bigotimes_{i\ odd} \Lambda^{\max} V_i \xrightarrow{\sim} \bigotimes_{i\ even} \Lambda^{\max} V_i.$$

One simply uses this Lemma plus the above canonical identification $(\operatorname{Det} V)^* \otimes \operatorname{Det} V \cong \mathbb{R}$.

Suppose now that V, W are Hilbert spaces, that Y is a connected topological space and that $l_y \in F(V, W)$ is a family of Fredholm operators depending continuously on $y \in Y$. Again, we form the determinant line

$$\operatorname{Det} l_y := (\operatorname{Det} \ker l_y)^* \otimes (\operatorname{Det} \operatorname{coker} l_y)$$

for each y. We intend to show that these lines $(\operatorname{Det} l_y)_{y \in Y}$ constitute a line bundle over Y.

$$l_y : (\ker l_y)^\perp \to (\operatorname{coker} l_y)^\perp,$$
$$v \mapsto l_y v$$

is an isomorphism, and

$$\operatorname{ind} l_y = \dim \ker l_y - \dim \operatorname{coker} l_y$$

is independent of $y \in Y$, as Y is connected. For y in a neighborhood of some $y_0 \in Y$, let $V'_y \subset V$ be a continuous family of finite dimensional subspaces with $\ker l_y \subset V'_y$ for each y, and put

$$W'_y := l_y(V'_y) \oplus \operatorname{coker} l_y.$$

Then as above

$$(\operatorname{Det} V'_y)^* \otimes \operatorname{Det} W'_y \cong (\operatorname{Det} \ker l_y)^* \otimes \operatorname{Det} \operatorname{coker} l_y.$$

The point now is that this construction is independent of the choice of V'_y in the sense that if V''_y is another such family, we get a **canonical** identification

$$(\operatorname{Det} V''_y)^* \otimes \operatorname{Det} W''_y \cong (\operatorname{Det} V'_y)^* \otimes \operatorname{Det} W'_y.$$

Once we have verified that property, we can piece the local models $(\operatorname{Det} V'_y)^* \otimes \operatorname{Det} W'_y$ for $\operatorname{Det} l_y$ unambiguously together to get a line bundle with fiber $\operatorname{Det} l_y$ over y on Y.

It suffices to treat the case

$$V'_y \subset V''_y,$$

and we write

$$V''_y = V'_y \oplus \bar{V}_y,$$

and

$$W''_y = W'_y \oplus \bar{W}_y.$$

$l_y : \bar{V}_y \to \bar{W}_y$ is an isomorphism, and

$$\det l_y : \operatorname{Det} \bar{V}_y \to \operatorname{Det} \bar{W}_y$$

yields a nonvanishing section Δ_{l_y} of $(\operatorname{Det} \bar{V}_y)^* \otimes \operatorname{Det} \bar{W}_y$. We then get the isomorphism

$$\begin{aligned}(\operatorname{Det} V'_y)^* \otimes \operatorname{Det} W'_y \to \\ (\operatorname{Det} V'_y)^* \otimes \operatorname{Det} W'_y \otimes (\operatorname{Det} \bar{V}_y)^* \otimes \operatorname{Det} \bar{W}_y \cong (\operatorname{Det} V''_y)^* \otimes (\operatorname{Det} W''_y) \\ s_y \mapsto s_y \otimes \Delta_{l_y},\end{aligned}$$

and this isomorphism is canonically determined by l_y.

We have thus shown

Theorem 6.9.1. *Let $(l_y)_{y \in Y} \subset F(V, W)$ be a family of Fredholm operators between Hilbert spaces V, W depending continuously on y in some connected topological space Y. Then we may construct a line bundle over Y with fiber*

$$\operatorname{Det} l_y = (\operatorname{Det} \ker l_y)^* \otimes (\operatorname{Det} \operatorname{coker} l_y)$$

over y, and with a continuous section $\det l_y$ vanishing precisely at those $y \in Y$ where $\ker l_y \neq 0$.

Definition 6.9.1. Let $l = (l_y)_{(y \in Y)} \subset F(V, W)$ be a family of Fredholm operators between Hilbert spaces V, W depending continuously on y in some connected topological space Y. An orientation of this family is given by a nowhere vanishing section of the line bundle $\operatorname{Det} l$ of the preceding theorem.

If $\ker l_y = 0$ for all $y \in Y$, then of course $\det l_y$ yields such a section. If this property does not hold, then such a section may or may not exist.

We now wish to extend Corollary 6.9.1 to the determinant lines of the operators involved, i.e. we wish to show that

$$\det L_A \cong \det L_{A_1} \otimes \det L_{A_2}.$$

In order to achieve this, we need to refine the glueing somewhat. We again trivialize a vector bundle E over $\mathbb{R}$, so that E becomes $\mathbb{R} \times \mathbb{R}^n$. Of course, one has to check that the subsequent constructions do not depend on the choice of trivialization.

We again consider the situation of Corollary 6.9.1, and we assume that A_1, A_2 are asymptotically constant in the sense that they do not depend on t for $|t| \geq T$, for some $T > 0$. For $\tau \in \mathbb{R}$, we define the shifted operator $L^\tau_{A_1}$ via

$$L^\tau_{A_1} s(t) = \frac{ds}{dt} + A_1(t - \tau) s(t).$$

As we assume A_1 asymptotically constant, $A_1^\tau(t) := A_1(t+\tau)$ does not depend on t over $[-1,\infty)$ for τ sufficiently large. Likewise, $A_2^{-\tau}(t)$ does not depend on t over $(-\infty, 1]$ for τ sufficiently large. We then put

$$A(t) := A_1 \#_\tau A_2(t) := \begin{cases} A_1(t+\tau) & \text{for } t \in (-\infty, 0], \\ A_2(t-\tau) & \text{for } t \in [0,\infty), \end{cases}$$

and obtain a corresponding Fredholm operator

$$L_{A_1 \#_\tau A_2}.$$

Lemma 6.9.4. *For τ sufficiently large,*

$$\operatorname{Det} L_{A_1 \#_\tau A_2} \cong \operatorname{Det} L_{A_1} \otimes \operatorname{Det} L_{A_2}.$$

Sketch of Proof. We first consider the case where L_{A_1} and L_{A_2} are surjective. We shall show

$$\dim\ker L_A \leq \dim\ker L_{A_1} + \dim\ker L_{A_2}, \tag{6.9.5}$$

which in the surjective case, by Corollary 6.9.1 equals

$$\operatorname{ind} L_{A_1} + \operatorname{ind} L_{A_2} = \operatorname{ind} L_A \leq \dim\ker L_A,$$

hence equality throughout.

Now if $s_\tau(t) \in \ker L_{A_1 \#_\tau A_2}$, we have

$$\frac{d}{dt} s_\tau(t) + A(t) s_\tau(t) = 0, \tag{6.9.6}$$

and we have

$$A(t) = A_1(\infty) (= A_2(-\infty))$$

for $|t| \leq \tau$, for arbitrarily large T, provided τ is sufficiently large. Since $A_1(\infty)$ is assumed to be nondegenerate, the operator

$$\frac{d}{dt} + A_1(\infty)$$

is an isomorphism, and thus, if we have a sequence

$$(s_{\tau_n})_{n\in\mathbb{N}}$$

of solutions of (6.9.6) for $\tau = \tau_n$, with $\|s_{\tau_n}\|_{H^{1,2}} \leq 1$, $\tau_n \to \infty$, then

$$s_{\tau_n} \to 0 \text{on } [-T,T], \quad \text{for any } T > 0.$$

On the other hand, for t very negative, we get a solution of

$$\frac{d}{dt} s_\tau(t) + A_1(-\infty) s_\tau(t) = 0,$$

or more precisely, $s_\tau(t-\tau)$ will converge to a solution of

$$\frac{d}{dt}s(t) + A_1(t)s(t) = 0,$$

i.e. an element of $\ker L_{A_1}$. Likewise $s_\tau(t+\tau)$ will yield an element of $\ker L_{A_2}$. This shows (6.9.5).

If L_{A_1}, L_{A_2} are not necessarily surjective, one finds a linear map $\Lambda : \mathbb{R}^k \to L^2(\mathbb{R}, \mathbb{R}^n)$ such that

$$\begin{aligned} L_{A_i} + \Lambda : H^{1,2}(\mathbb{R}, \mathbb{R}^n) \times \mathbb{R}^k &\to L^2(\mathbb{R}, \mathbb{R}^n), \\ (s, v) &\mapsto L_{A_i}s + \Lambda v \end{aligned}$$

are surjective for $i = 1, 2$. One then performs the above argument for these perturbed operators, and observes that the corresponding determinants of the original and the perturbed operators are isomorphic. $\square$

We now let Y be the space of all pairs (x, A), where $x : \mathbb{R} \to X$ is a smooth curve with limits $x(\pm\infty) = \lim_{t\to\pm\infty} x(t) \in X$, and A is a smooth section of $x^*\mathrm{End}\, TX$ for which $A(t)$ is a selfadjoint linear operator on $T_{x(t)}X$, for each $t \in \mathbb{R}$, with limits $A(\pm\infty) = \lim_{t\to\pm\infty} A(t)$ that are nondegenerate, and for each $y \in (x, A) \in Y$, we consider the Fredholm operator

$$L_{(x,A)} := \nabla + A : H^{1,2}(x^*TX) \to L^2(x^*TX),$$

Lemma 6.9.5. *Suppose X is a finite dimensional orientable Riemannian manifold. Let (x_1, A_1), $(x_1, A_2) \in Y$ satisfy*

$$\begin{aligned} x_1(\pm\infty) &= x_2(\pm\infty), \\ A_1(\pm\infty) &= A_2(\pm\infty). \end{aligned}$$

Then the determinant lines $\mathrm{Det}\, L_{(x_1,A_1)}$ and $\mathrm{Det}\, L_{(x_2,A_2)}$ can be identified through a homotopy.

Proof. We choose trivializations $\sigma_i : x_i^*TX \to \mathbb{R} \times \mathbb{R}^n$ $(n = \dim X)$ extending continuously to $\pm\infty$, for $i = 1, 2$. Thus, $L_{(x_i,A_i)}$ is transformed into an operator

$$L_{A_i} = \frac{d}{dt} + A_i(t) : H^{1,2}(\mathbb{R}, \mathbb{R}^n) \to L^2(\mathbb{R}, \mathbb{R}^n)$$

(with an abuse of notation, namely using the same symbol $A_i(t)$ for an endomorphism of $T_{x(t)}X$ and of $\mathbb{R}^n = \sigma_i(t)(T_{x(t)}X)$). Since X is orientable, we may assume that

$$\sigma_1(\pm\infty) = \sigma_2(\pm\infty)$$

(for a nonorientable X, we might have $\sigma_1(-\infty) = \sigma_2(-\infty)$, but $\sigma_1(\infty) = -\sigma_2(\infty)$, or vice versa, because $\mathrm{Gl}(n, \mathbb{R})$ has two connected components, but in the orientable case, we can consistently distinguish these two components acting on the tangent spaces T_xX with the help of the orientations of the spaces T_xX). Thus, the relations $A_1(\pm\infty) = A_2(\pm\infty)$ are preserved under these trivializations.

From the proof of Lemma 6.9.2, $\operatorname{ind} L_{A_1} = \operatorname{ind} L_{A_2}$, and $\operatorname{coker} L_{A_i} = 0$ or $\ker L_{A_i} = 0$, depending on whether $\pm\mu(A_i(-\infty)) \geq \pm\mu(A_i(\infty))$. It then suffices to consider the first case. Since the space of all adjoint endomorphisms of $\mathbb{R}^n$ can be identified with $\mathbb{R}^{\frac{n(n+1)}{2}}$ (the space of symmetric $(n \times n)$ matrices), we may find a homotopy between A_1 and A_2 in this space with fixed endpoints $A_1(\pm\infty) = A_2(\pm\infty)$. As a technical matter, we may always assume that everything is asymptotically constant as in the proof of Lemma 6.9.2, and that proof then shows that such a homotopy yields an isomorphism between the kernels of L_{A_1} and L_{A_2}. □

Thus, Fredholm operators with coinciding ends at $\pm\infty$ as in Lemma 6.9.5 can be consistently oriented. Expressed differently, we call such operators equivalent, and we may define an orientation on an equivalence class by choosing an orientation of one representative and then defining the orientations of the other elements of the class through a homotopic deformation as in that lemma.

Definition 6.9.2. An assignment of an orientation $\sigma(x, A)$ to each equivalence class (x, A) is called coherent if it is compatible with glueing, i.e.

$$\sigma((x_1, A_1)\#(x_2, A_2)) = \sigma(x_1, A_1) \otimes \sigma(x_2, A_2)$$

(assuming, as always, the conditions required for glueing, i.e. $x_1(\infty) = x_2(-\infty)$, $A_1(\infty) = A_2(-\infty)$).

Theorem 6.9.2. *Suppose X is a finite dimensional orientable Riemannian manifold. Then a coherent orientation exists.*

Proof. We first consider an arbitrary constant curve

$$x(t) \equiv x_0 \in X, \quad A(t) = A_0.$$

The corresponding Fredholm operator

$$L_{A_0} = \frac{d}{dt} + A_0 : H^{1,2}(\mathbb{R}, T_{x_0}X) \to L^2(T_{x_0}X)$$

then is an isomorphism by the proof of Lemma 6.9.2, or an easy direct argument. Thus, $\operatorname{Det} L_{A_0}$ is identified with $\mathbb{R}\otimes\mathbb{R}^*$, and we choose the orientation $1\otimes 1^* \in \mathbb{R}\otimes\mathbb{R}^*$. We next choose an arbitrary orientation for each class of operators $L_{(x,A)}$ different from $L_{(x_0,A_0)}$ with

$$x(-\infty) = x_0, \quad A(-\infty) = A_0$$

(note that the above definition does not require any continuity e.g. in $A(\infty)$). This then determines orientations for classes of operators $L_{(x,A)}$ with

$$x(\infty) = x_0, \quad A(\infty) = A_0,$$

because the operator $L_{(x^{-1},A^{-1})}$, with $x^{-1}(t) := x(-t)$, $A^{-1}(t) := A(-t)$, then is in the first class, and

$$L_{(x^{-1},A^{-1})}\#L_{(x,A)} \text{ is equivalent to } L_{(x_0,A_0)},$$

and by Lemmas 6.9.4 and 6.9.5,

$$\operatorname{Det} L_{(x^{-1},A^{-1})} \otimes \operatorname{Det} L_{(x,A)} \equiv \operatorname{Det} L_{(x_0,A_0)}.$$

Finally, for an arbitrary class $L_{(x,A)}$, we find (x_1, A_1) and (x_2, A_2) with

$$\begin{array}{cccc} x_1(-\infty) = x_0, & A_1(-\infty) = x_0, & x_1(\infty) = x(-\infty), & A_1(\infty) = A(-\infty), \\ x_2(\infty) = x_0, & A_2(\infty) = x_0, & x_2(-\infty) = x(\infty), & A_2(-\infty) = A(\infty). \end{array}$$

and the glueing relation

$$L_{(x_1,A_1)}\#L_{(x,A)}\#L_{(x_2,A_2)} \text{ equivalent to } L_{(x_0,A_0)}.$$

The relation of Lemma 6.9.4, i.e.

$$\operatorname{Det} L_{(x_1,A_1)} \otimes \operatorname{Det} L_{(x,A)} \otimes \operatorname{Det} L_{(x_2,A_2)} \cong \operatorname{Det} L_{(x_0,A_0)}$$

then fixes the orientation of $L_{(x,A)}$. □

We shall now always assume that X is a compact finite dimensional, orientable Riemannian manifold. According to Theorem 6.9.2, we may assume from now on that a coherent orientation on the class of all operators $L_{(x,A)}$ as above has been chosen.

We now consider a Morse-Smale-Floer function

$$f : X \to \mathbb{R}$$

as before, and we let $p, q \in X$ be critical points of f with

$$\mu(p) - \mu(q) = 1.$$

Then for each gradient flow line $x(t)$ with $x(-\infty) = p$, $x(\infty) = q$, i.e.

$$\dot{x}(t) + \operatorname{grad} f(x(t)) = 0,$$

the linearization of that operator, i.e.

$$L := \nabla_{\dot{x}(t)} + d^2 f(x(t)) : H^{1,2}(x^*TX) \to L^2(x^*TX)$$

is a surjective Fredholm operator with one-dimensional kernel, according to Lemma 6.9.2 and its proof. However, we can easily find a generator of the kernel: as the equation satisfied by $x(t)$ is autonomous, for any $\tau_0 \in \mathbb{R}$, $x(t+\tau)$ likewise is a solution, and therefore $\dot{x}(t)$ must lie in the kernel of the linearization. Altogether, $\dot{x}(t)$ defines an orientation of $\operatorname{Det} L$, called the canonical orientation.

Definition 6.9.3. We assign a sign $n(x(t)) = \pm 1$ to each such trajectory of the negative gradient flow of f with $\mu(x(-\infty)) - \mu(x(\infty)) = 1$ by putting $n = 1$ precisely if the coherent and the canonical orientation for the corresponding linearized operator $\nabla + d^2 f$ coincide.

This choice of sign enables us to take up the discussion of §6.6 and define the boundary operator as

$$\partial p = \sum_{\substack{r \in C_*(f) \\ \mu(r) = \mu(p) - 1 \\ s \in \mathcal{M}^f_{p,r}}} n(s) r,$$

now with our present choice of sign. Again, the crucial point is to verify the relation

$$\partial^2 = 0.$$

As in Thmeorem 6.5.1, based on Theorem 6.3.1, we may again consider a component $\mathcal{M}$ of $\mathcal{M}^f_{p,q}$ (p, q critical points of f with $\mu(p) - \mu(q) = 2$), homeomorphic to the open disk. We get a figure similar to Figure 6.6.1

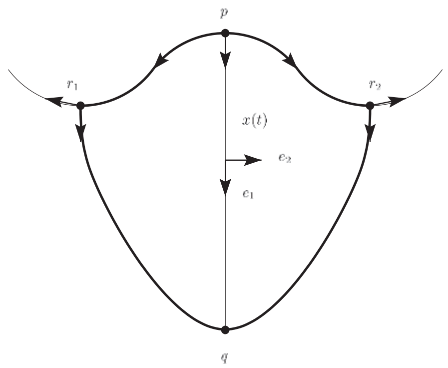

Figure 6.9.1:

On the flow line $x(t)$ from p to q, we have indicated a coherent orientation, chosen such that e_1 corresponds to the negative flow line direction, and e_2 corresponds to an arbitrarily chosen orientation of the one-dimensional manifold $f^{-1}(a) \cap \mathcal{M}$, where $f(q) < a < f(p)$, as in §6.6. The kernel of the associated Fredholm operators L_x is two-dimensional, and $e_1 \wedge e_2$ then induces an orientation of $\operatorname{Det} L_x$. The coherence condition then induces corresponding orientations on the two broken trajectories from p to q, passing through the critical points r_1, r_2 resp. In the figure, we have indicated the canonical orientations of the trajectories from p to r_1 and r_2 and from r_1 and r_2 to q. Now if for example the coherent orientations of the two trajectories from p to r_1 and r_2, resp. both coincide with those canonical orientations, then this will take place for precisely one of the two trajectories from r_1 and r_2 resp. to q. Namely, it is clear now from the figure that the combination of the canonical orientations on the broken trajectories leads to opposite orientations at q, which however is not compatible with the coherence condition. From this simple geometric observation, we infer the relation $\partial \circ \partial = 0$ as in §6.6.

We may also take up the discussion of §6.7 and consider a regular homotopy (as in Definition 6.7.1) F between two Morse functions f^1, f^2, and the induced map

$$\phi^{21} : C_*(f^1, \mathbb{Z}) \to C_*(f^2, \mathbb{Z}).$$

In order to verify the relationship

$$\phi^{21} \circ \partial f^1 = \partial f^2 \circ \phi^{21} \tag{6.9.7}$$

with the present choice of signs, we proceed as follows. If p_1 is a critical point of f^1, p_2 one of f^2, with

$$\mu(p_1) = \mu(p_2),$$

and if $s : \mathbb{R} \to X$ with $s(-\infty) = p_1$, $s(\infty) = p_2$ satisfies (6.7.4), i.e.

$$\dot{s}(t) = -\operatorname{grad} F(s(t), t), \tag{6.9.8}$$

we consider again the linearized Fredholm operator

$$L_s := \nabla + d^2 F : H^{1,2}(s^*TX) \to L^2(s^*TX).$$

Since $\mu(p_1) = \mu(p_2)$, Lemma 6.9.2 implies

$$\operatorname{ind} L_s = 0.$$

Since by definition of a regular homotopy, L_s is surjective, we consequently get

$$\ker L_s = 0.$$

Thus, $\operatorname{Det} L_s$ is the trivial line bundle $\mathbb{R} \otimes \mathbb{R}^*$, and we may orient it by $1 \otimes 1^*$, and we call that orientation again canonical. Thus, we may assign a sign $n(s)$ to each trajectory from p_1 to p_2 solving (6.9.8) as before by comparing the coherent and the canonical orientations.

Now in order to verify (6.9.7), we look at Figure 6.9.2. Here, we have indicated a flow line w.r.t. f^1 from p_1 to another critical point r_1 of f^1 with $\mu(p_1) - \mu(r_1) = 1$, and likewise one w.r.t. f^2 from p_2 to r_2 with $\mu(p_2) - \mu(r_2) = 1$, both of them equipped with the canonical orientations as defined above for the relative index 1. Since now the solution curves of (6.9.8) from p_1 to p_2, and likewise from r_1 to r_2 carry the orientation of a trivial line bundle, we may choose the coherent orientations so as to coincide with the canonical ones.

We now compute for a critical point p_1 of f^1 with $\mu(p_1) = \beta$, and with $\mathcal{M}^F_{p_1,q_1}$ the

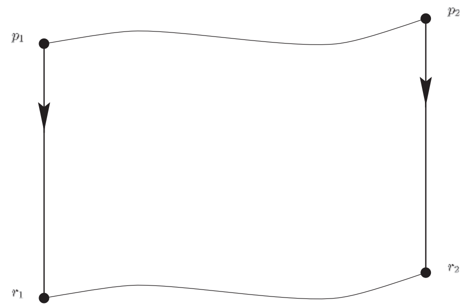

Figure 6.9.2:

space of solutions of (6.9.8) from p_1 to p_2,

$$\begin{aligned}
&(\partial f^2 \circ \phi^{21} - \phi^{21} \circ \partial f^1)(p_1) \\
&\quad = \partial f^2 \left(\sum_{\mu(p_2)=\beta} \sum_{s \in \mathcal{M}^F_{p_1,p_2}} n(s) p_2 \right) - \phi^{21} \left(\sum_{\mu(r_1)=\beta-1} \sum_{s_1 \in \mathcal{M}^{f^1}_{p_1,r_1}} n(s_1) r_1 \right) \\
&\quad = \sum_{\mu(r_2)=\beta-1} \left(\sum_{\mu(p_2)=\beta} \sum_{s_1 \in \mathcal{M}^F_{p_1,p_2}} \sum_{s_2 \in \mathcal{M}^{f^2}_{p_2,r_2}} n(s) n(s_2) \right. \\
&\qquad\qquad \left. - \sum_{\mu(r_1)=\beta-1} \sum_{s_1 \in \mathcal{M}^{f^1}_{p_1,r_1}} \sum_{s' \in \mathcal{M}^F_{r_1,r_2}} n(s_1) n(s') \right) r_2 \ .
\end{aligned}$$

Again, as in Theorem 6.5.1, trajectories occur in pairs, but the pairs may be of two different types: within each triple sum, we may have a pair $(s^{(1)}, s_2^{(1)})$ and $(s^{(2)}, s_2^{(2)})$, and the two members will carry opposite signs as we are then in the situation of Figure 6.9.1. The other type of pair is of the form (s, s_2) and (s_1, s'), i.e. one member each from the two triple sums. Here, the two members carry the same sign, according to the analysis accompanying Figure 6.9.2, but since there are opposite signs in front of the two triple sums, we again get a cancellation.

In conclusion, all contributions in the preceding expression cancel in pairs, and we obtain

$$\partial f^2 \circ \phi^{21} - \phi^{21} \circ \partial f^1 = 0,$$

as desired. We thus obtain

Theorem 6.9.3. *Let X be a compact, finite dimensional, orientable Riemannian manifold. Let f^1, f^2 be Morse-Smale-Floer functions, and let F be a regular homotopy between them. Then F induces a map*

$$\phi^{21} : C_*(f^1, \mathbb{Z}) \to C_*(f^2, \mathbb{Z})$$

satisfying

$$\partial \circ \phi^{21} = \phi^{21} \circ \partial,$$

and hence an isomorphism of the corresponding homology groups defined by f^1 *and* f^2*, resp.* □

Corollary 6.9.2. *Under the assumptions of Theorem 6.9.3, the numbers* $b_k(X, f)$ *defined at the end of §6.6 do not depend on the choice of a Morse-Smale-Floer function* f *and thus define invariants* $b_k(X)$ *of* X*.* □

Definition 6.9.4. The numbers $b_k(X)$ are called the Betti numbers of X.

Remark. The Betti numbers have been defined through the choice of a Riemannian metric. In fact, however, they turn out not to depend on that choice. See the Perspectives for some further discussion.

Perspectives. The relative approach to Morse theory presented in this chapter was first introduced by Floer in [81]. It was developed in detail by Schwarz[230], and starting with §6.4 we have followed here essentially the approach of Schwarz although in certain places some details are different (in particular, we make a more systematic use of the constructions of §6.3), and we cannot penetrate here into all the aspects worked out in that monograph. An approach to Floer homology from the theory of hyperbolic dynamical system has been developed in [258]. We also refer the reader to the bibliography of [230] for an account of earlier contributions by Thom, Milnor, Smale, and Witten. (Some references can also be found in the Perspectives on §6.10.)

In particular, Witten[264], inspired by constructions from supersymmetry, established an isomorphism between the cohomology groups derived from a Morse function and the ones coming from the Hodge theory of harmonic forms as developed in Chapter 2 of the present work.

In some places, we have attempted to exhibit geometric ideas even if considerations of space did not allow the presentation of all necessary details. This applies for example to the §6.8 on graph flows which is based on [22]. As in Schwarz' monograph, the construction of coherent orientations in §6.9 is partly adapted from Floer, Hofer[82]. This in turn is based on the original work of Quillen[214] on determinants.

The theory as presented here is somewhat incomplete because we did not develop certain important aspects, among which we particularly wish to mention the following three:

1) Questions of genericity:
   A subset of a Baire topological space is called generic if it contains a countable intersection of open and dense sets. In the present context, one equips the space of (sufficiently smooth) functions on a differentiable manifold X as well as the space of Riemannian metrics on X with some C^k topology, for sufficiently large k. Then at least if X is finite dimensional and compact, the set of all functions satisfying the Morse condition as well as the set of all Riemannian metrics for which a given Morse function satisfies the Morse-Smale-Floer condition are generic.
2) We have shown (see §6.7 and this section) that a regular homotopy between two Morse functions induces an isomorphism between the corresponding homology theory. It remains to verify that this isomorphism does not depend on the choice of homotopy and is flow canonical.

3) Independence of the choice of Riemannian metric on X: We recall that by Lemma 1.8.1, a Riemannian metric on X is given by a symmetric, positive definite covariant 2-tensor. Therefore, for any two such metrics g_0, g_1 and $0 \leq t \leq 1, g_t := tg_0 + (1-t)g_1$ is a metric as well, and so the space of all Riemannian metrics on a given differentiable manifold is a convex space, in particular connected. If we now have a Morse function f, then the gradient flows w.r.t. two metrics g_0, g_1 can be connected by a homotopy of metrics. The above linear interpolation g_t may encounter the problem that for some t, the Morse-Smale-Floer transversality condition may not hold, and so one needs to consider more general homotopies. Again, for a generic homotopy, all required transversality conditions are satisfied, and one then conclude that the homology groups do not depend on the choice of Riemannian metric. Thus, they define invariants of the underlying differentiable manifold. In fact, they are even invariants of the topological structure of the manifold, because they satisfy the abstract Eilenberg-Steenrood axioms of homology theory, and therefore yield the same groups as the singular homology theory that is defined in purely topological terms.

These points are treated in detail in [230] to which we consequently refer.

As explained in this chapter, we can also use a Morse function to develop a cohomology theory. The question then arises how this cohomology theory is related to the de Rham-Hodge cohomology theory developed in Chapter 2. One difference is that the theory in Chapter 2 is constructed with coefficients $\mathbb{R}$, whereas the theory in this chapter uses $\mathbb{Z}_2$ and $\mathbb{Z}$ as coefficients. One may, however, extend those coefficients to $\mathbb{R}$ as well. Then, in fact, the two theories become isomorphic on a compact differentiable manifolds, as are all cohomology theories satisfying the Eilenberg-Stennrod axioms. These axioms are verified for Morse-Floer cohomology in [230]. The background in algebraic topology can be found in [238]. Witten[264] derived that isomorphism in a direct manner. For that purpose, Witten considered the operators

$$d_t := e^{-tf} d e^{tf},$$

their formal adjoints

$$d_t^* = e^{tf} d^* e^{-tf},$$

and the corresponding Laplacian

$$\Delta_t := d_t d_t^* + d_t^* d_t.$$

For $t = 0$, Δ_0 is the usual Laplacian that was used in Chapter 2 in order to develop Hodge theory and de Rham cohomology, whereas for $t \to \infty$, one has the following expansion

$$\Delta_t = dd^* + d^*d + t^2 \|df\|^2 + t\Sigma_{k,j} \frac{\partial^2 h}{\partial x^k \partial x^j} \left[i \left(\frac{\partial}{\partial x^k} \right), dx^j \right]$$

where $\left(\frac{\partial}{\partial x^j}\right)_{j=1,\dots,n}$ is an orthonormal frame at the point under consideration. This becomes very large for $t \to \infty$, except at the critical points of f, i.e. where $df = 0$. Therefore, the eigenfunctions of Δ_t will concentrate near the critical points of f for $t \to \infty$, and we obtain an interpolation between de Rham cohomology and Morse cohomology.

An elementary discussion of Morse theory, together with applications to closed geodesics, can be found in [191].

Finally, as already mentioned, Conley developed a very general critical point theory that encompasses Morse theory but applies to arbitrary smooth functions without the requirement of nondegenerate critical points. This theory has found many important applications, but here we have to limit ourselves to quoting the references Conley[55], Conley and Zehnder[56]. In another direction, different approaches to Morse theory on singular (stratified) spaces have been developed by Goresky and MacPherson[97] and Ludwig[182].

6.10 The Morse Inequalities

The Morse inequalities express relationships between the Morse numbers μ_i, defined as the numbers of critical points of a Morse function f of index i, and the Betti numbers b_i of the underlying manifold X. In order to simplify our exposition, in this section, we assume that X is a **compact** Riemannian manifold, and we only consider homology with $\mathbb{Z}_2$-coefficients (the reader is invited to extend the considerations to a more general setting). As before, we also assume that $f : X \to \mathbb{R}$ is of class C^3 and that all critical points of f are nondegenerate, and that (X, f) satisfies the Morse-Smale-Floer condition.

As a preparation, we need to consider relative homology groups. Let A be a compact subset of X, with the property that flow lines can enter, but not leave A. This means that if

$$\dot{x}(t) = -\operatorname{grad} f(x(t)) \quad \text{for } t \in \mathbb{R}$$

and

$$x(t_0) \in A \quad \text{for some } t_0 \in \mathbb{R} \cup \{-\infty\},$$

then also

$$x(t) \in A \quad \text{for all } t \geq t_0.$$

We obtain a new boundary operator ∂^A in place of ∂ by taking only those critical points of f into account that lie in $X \backslash A$. Thus, for a critical point $p \in X \backslash A$, we put

$$\partial^A p := \sum_{\substack{r \in C_*(f) \cap X \backslash A \\ \mu(p,r)=1}} (\#_{\mathbb{Z}_2} \mathcal{M}^f_{p,r})\, r. \tag{6.10.1}$$

By the above condition that flow lines cannot leave A once they hit it, all flow lines between critical points $p, r \in X \backslash A$ are entirely contained in $X \backslash A$ as well. In particular, as in Theorem 6.5.2, we have

$$\partial^A \cdot \partial^A p = 0 \quad \text{for all critical points of } f \text{ in } X \backslash A. \tag{6.10.2}$$

Defining $C_*^A(f,\mathbb{Z}_2)$ as the free Abelian group with $\mathbb{Z}_2$-coefficients generated by the critical points of f in $X \backslash A$, we conclude that

$$(C_*^A(f,\mathbb{Z}_2), \partial^A)$$

is a chain complex. We then obtain associated homology groups

$$H_k(X, A, f, \mathbb{Z}_2) := \frac{\ker \partial_k^A}{\text{image}\, \partial_{k+1}^A}\,, \tag{6.10.3}$$

as in §6.5.

We shall actually need a further generalization: Let $A \subset Y \subset X$ be compact, and let $f : X \to \mathbb{R}$ satisfy:

(i) If the flow line $x(t)$, i.e.

$$\dot{x}(t) = -\text{grad}\, f(x(t)) \quad \text{for all } t,$$

satisfies

$$x(t_0) \in A \quad \text{for some } t_0 \in \mathbb{R} \cup \{-\infty\},$$

then there is no $t > t_0$ with $x(t) \in Y \backslash A$.

(ii) If the flow line $x(t)$ satisfies

$$x(t_1) \in Y, x(t_2) \in X \backslash \overset{\circ}{Y}, \text{ with } -\infty \le t_1 < t_2 \le \infty,$$

then there exists $t_1 \le t_0 \le t_2$ with

$$x(t_0) \in A.$$

Thus, by (i), flow lines cannot reenter the rest of Y from A, whereas by (ii), they can leave the interior of Y only through A. If $p \in Y \backslash A$ is a critical point of f, we put

$$\partial^{Y,A} p := \sum_{\substack{r \in C_*(f) \cap Y \backslash A \\ \mu(p,r)=1}} (\#_{\mathbb{Z}_2} \mathcal{M}_{p,r}^f)\, r. \tag{6.10.4}$$

Again, if p and r are critical points in $Y \backslash A$, then any flow line between them also has to stay entirely in $Y \backslash A$, and so as before

$$\partial^{Y,A} \circ \partial^{Y,A} = 0, \tag{6.10.5}$$

and we may define the homology groups

$$H_k(Y, A, f, \mathbb{Z}_2) := \frac{\ker \partial_k^{Y,A}}{\text{image}\, \partial_{k+1}^{Y,A}}\,. \tag{6.10.6}$$

We now apply these constructions in three steps:

1) Let p be a critical point of f with Morse index

$$\mu(p) = k.$$

We consider the unstable manifold

$$W^u(p) = \{x(\cdot) \text{ flow line with } x(-\infty) = p\}. \tag{6.10.7}$$

As the parametrization of a flow line is only defined up to an additive constant, we use the following simple device to normalize that constant. It is easy to see, for example by Theorem 6.3.1, that for sufficiently small $\varepsilon > 0, W^u(p)$ intersects the sphere $\partial B(p, \varepsilon)$ transversally, and each flow line in $W^u(p)$ intersects that sphere exactly once. We then choose the parametrization of the flow lines $x(\cdot)$ in $W^u(p)$ such that $x(0)$ always is that intersection point with the sphere $\partial B(p, \varepsilon)$. Having thus fixed the parametrization, for any $T \in \mathbb{R}$, we cut all the flow lines off at time T:

$$Y_p^T := \{x(t) : -\infty \leq t \leq T, x(\cdot) \text{ flow line in } W^u(p)\} \tag{6.10.8}$$

and

$$A_p^T := \{x(T) : x(\cdot) \text{ flow line in } W^u(p)\}. \tag{6.10.9}$$

It is easy to compute the homology $H_*(Y_p^T, A_p^T, f, \mathbb{Z}_2)$: p is the only critical point of f in $Y_p^T \backslash A_p^T$, and so

$$\partial^{Y_p^T, A_p^T} p = 0. \tag{6.10.10}$$

Thus, the kernel of $\partial_k^{Y_p, A_p}$ is generated by p. All the other kernels and images of the $\partial_j^{Y_p^T, A_p^T}$ are trivial and therefore

$$H_j(Y_p^T, A_p^T, f, \mathbb{Z}_2) = \begin{cases} \mathbb{Z}_2 & \text{if } j = k, \\ 0 & \text{otherwise,} \end{cases} \tag{6.10.11}$$

for all $T \in \mathbb{R}$.

Thus, the groups $H_j(Y_p^T, A_p^T, f, \mathbb{Z}_2)$ encode the local information expressed by the critical points and their indices. No relations between different critical points are present at this stage. Thus, for this step, we do not yet need the Morse-Smale-Floer condition.

2) We now wish to let T tend to ∞, i.e. to consider the entire unstable manifold $W^u(p)$. $W^u(p)$, however, is not compact, and so we need to compactify it. This can be done on the basis of the results of §§6.4, 6.5. Clearly, we need to include all critical points r of f that are end points of flow lines in $W^u(p)$, i.e.

$$r = x(\infty) \quad \text{for some flow line} \quad x(\cdot) \text{ in } W^u(p).$$

In other words, we consider all critical points r to which p is connected by the flow in the sense of Definition 6.5.2. In particular, for any such r

$$\mu(r) < \mu(p),$$

because of the Morse-Smale-Floer condition, see (6.5.2). Adding those critical points, however, is not yet enough for compactifying $W^u(p)$. Namely, we also need to add the unstable manifolds $W^u(r)$ of all those r. If the critical point q is the asymptotic limit $y(\infty)$ of some flow line $y(\cdot)$ in $W^u(r)$, then, by Lemma 6.5.2, we may also find a flow line $x(\cdot)$ in $W^u(p)$ with $x(\infty) = q$, and furthermore, as the proof of Lemma 6.5.2 shows, the flow line $y(\cdot)$ is the limit of flow lines $x(\cdot)$ from $W^u(p)$. Conversely, by Theorem 6.4.1, any limit of flow lines $x_n(\cdot)$ from $W^u(p)$, $n \in \mathbb{N}$, is a union of flow lines in the unstable manifolds of critical points to which p is connected by the flow, using also Lemma 6.5.2 once more. As these results are of independent interest, we summarize them as

Theorem 6.10.1. *Let $f \in C^3(X, \mathbb{R}), X$ a compact Riemann manifold, be a function with only nondegenerate critical points, satisfying the Morse-Smale-Floer condition. Let p be a critical point of f with unstable manifold $W^u(p)$. Then $W^u(p)$ can be compactified by adding all the unstable manifolds $W^u(r)$ of critical points r for which there exists some flow line from p to r, and conversely, this is the smallest compactification of $W^u(p)$.* □

We now let Y be that compactification of $W^u(p)$, and $A := Y \backslash W^u(p)$, i.e. the union of the unstable manifolds $W^u(r)$ of critical points r to which p is connected by the flow. Again, the only critical point of f in $Y \backslash A$ is p, and so we have as in 1)

$$H_j(Y, A, f, \mathbb{Z}_2) = \begin{cases} \mathbb{Z}_2 & \text{if } j = \mu(p), \\ 0 & \text{otherwise.} \end{cases} \tag{6.10.12}$$

The present construction, however, also allows a new geometric interpretation of the boundary operator ∂. For that purpose, we let $C'_*(f, \mathbb{Z}_2)$ be the free Abelian group with $\mathbb{Z}_2$-coefficients generated by the set $C'_*(f)$ of unstable manifolds $W^u(p)$ of critical points p of f, and

$$\partial' W^u(p) := \sum_{\substack{r \in C_*(f) \\ \mu(r) = \mu(p) - 1}} (\#_{\mathbb{Z}_2} \mathcal{M}^f_{p,r})\, W^u(r). \tag{6.10.13}$$

Thus, if $\mu(p) = k$, the boundary of the k-dimensional manifold $W^u(p)$ is a union of $(k-1)$-dimensional manifolds $W^u(r)$. Clearly, $\partial' \circ \partial' = 0$ by Theorem 6.5.2, as we have simply replaced all critical points by their unstable manifolds. This brings us into the realm of classical or standard homology theories on differentiable manifolds. From that point of view, the idea of Floer then was to encode all information about certain submanifolds of X that generate the homology, namely the unstable manifolds $W^u(p)$ in the critical points p themselves and the flow lines between them. The advantage is that this allows a formulation of homology in purely relative terms, and thus greater generality and enhanced conceptual clarity, as already explained in this chapter.

3) We now generalize the preceding construction by taking unions of unstable manifolds. For a critical point p of f, we now denote the above compactification of $W^u(p)$ by $Y(p)$. We consider a space Y that is the union of some such $Y(p)$, and a subspace A that is the union of some $Y(q)$ for critical points $q \in Y$. As before, we get induced homology groups $H_k(A), H_k(Y), H_k(Y, A)$, omitting f and $\mathbb{Z}_2$ from the notation from now on for simplicity. As explained in 2), we may consider the elements of these groups as equivalence classes (up to boundaries) either of collections of critical points of f or of their unstable manifolds.

We now need to derive some standard facts in homology theory in our setting. A reader who knows the basics of homology theory may skip the following until the end of the proof of Lemma 6.10.4.

We recall the notation from algebraic topology that a sequence of linear maps f_j between vector spaces A_j

$$\cdots\ A_{i+1} \xrightarrow{f_{i+1}} A_i \xrightarrow{f_i} A_{i-1} \xrightarrow{f_{i-1}} \cdots$$

is called exact if always

$$\ker(f_i) = \text{image}\,(f_{i+1}).$$

We consider the maps

$$\begin{aligned} i_k &: H_k(A) \to H_k(Y), \\ j_k &: H_k(Y) \to H_k(Y, A), \\ \partial_k &: H_k(Y, A) \to H_{k-1}(A) \end{aligned}$$

defined as follows:

If $\pi \in C_k(A)$, the free Abelian group with $\mathbb{Z}_2$-coefficients generated by the critical points of f in A, we can consider π also as an element of $C_k(Y)$, from the inclusion $A \hookrightarrow Y$. If π is a boundary in $C_k(A)$, i.e. $\pi = \partial_{k+1}\gamma$ for some $\gamma \in C_{k+1}(A)$, then by the same token, γ can be considered as an element of $C_{k+1}(Y)$, and so π is a boundary in $C_k(Y)$ as well.

Therefore, this procedure defines a map i_k from $H_k(A)$ to $H_k(Y)$.

Next, if $\pi \in C_k(Y)$, we can also consider it as an element of $C_k(Y, A)$, by forgetting about the part supported on A, and again this defines a map j_k in homology.

Finally, if $\pi \in C_k(Y)$ with $\partial\pi \in C_{k-1}(A)$ and thus represents an element of $H_k(Y, A)$, then we may consider $\partial\pi$ as an element of $H_{k-1}(A)$, because $\partial \circ \partial\pi = 0$. $\partial\pi$ is not necessarily trivial in $H_{k-1}(A)$, because π need not be supported on A, but $\partial\pi$ as an element of $H_{k-1}(A)$ does not change if we replace π by $\pi + \gamma$ for some

$\gamma \in C_k(A)$. Thus, $\partial\pi$ as an element of $H_{k-1}(A)$ depends on the homology class of π in $H_k(Y, A)$, and so we obtain the map $\partial_k : H_k(Y, A) \to H_{k-1}(A)$.

The proof of the following result is a standard routine in algebraic topology:

Lemma 6.10.1.

$$\cdots\ H_k(A) \xrightarrow{i_k} H_k(Y) \xrightarrow{j_k} H_k(Y, A) \xrightarrow{\partial_k} H_{k-1}(A) \longrightarrow\ \cdots$$

is exact.

Proof. We denote the homology classes of an element γ by $[\gamma]$.

1) Exactness at $H_k(A)$:

   Suppose $[\gamma] \in \ker i_k$, i.e.
   $$i_k[\gamma] = 0.$$
   This means that there exists $\pi \in C_{k+1}(Y)$ with
   $$\partial\pi = i_k(\gamma).$$
   Since $i_k(\gamma)$ is supported on A, π represents an element of $H_{k+1}(Y, A)$, and so $[\gamma] \in \text{image}\,(\partial_{k+1})$. Conversely, for any such π, $\partial\pi$ represents the trivial element in $H_k(Y)$, and so $i_k[\partial\pi] = 0$, hence $[\partial\pi] \in \ker i_k$. Thus $i_k \circ \partial_{k+1} = 0$.

2) Exactness at $H_k(Y)$:

   Suppose $[\pi] \in \ker j_k$. This means that π is supported on A, and so $[\pi]$ is in the image of i_k. Conversely, obviously $j_k \circ i_k = 0$.

3) Exactness at $H_k(Y, A)$:

   Let $[\pi] \in \ker \partial_k$. Then $\partial\pi = 0$, and so π represents an element in $H_k(Y)$. Conversely, for any $[\pi] \in H_k(Y), \partial\pi = 0$, and therefore $\partial_k \circ j_k = 0$.

□

In the terminology of algebraic topology, a diagram

$$\begin{array}{ccc} A_2 & \xrightarrow{a} & A_1 \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle g} \\ B_2 & \xrightarrow{b} & B_1 \end{array}$$

of linear maps between vector spaces is called commutative if

$$g \circ a = b \circ f.$$

Let now (Y_1, Y_2) and (Y_2, Y_3) be pairs of the type (Y, A) just considered. We then have the following simple result

Lemma 6.10.2. *The diagram*

$$\begin{tikzcd}
\cdots \arrow[r] & H_k(Y_2,Y_3) \arrow[r, "\partial_k^{2,3}"] \arrow[d, "i_k^{1,2,3}"] & H_{k-1}(Y_3) \arrow[r, "i_k^{2,3}"] \arrow[d, "i_k^{2,3}"] & H_{k-1}(Y_2) \arrow[r, "j_k^{2,3}"] \arrow[d, "i_k^{1,2}"] & H_{k-1}(Y_2,Y_3) \arrow[r] \arrow[d, "i_{k-1}^{1,2,3}"] & \cdots \\
\cdots \arrow[r] & H_k(Y_1,Y_2) \arrow[r, "\partial_k^{1,2}"] & H_{k-1}(Y_2) \arrow[r, "i_k^{1,2}"] & H_{k-1}(Y_1) \arrow[r, "j_k^{1,2}"] & H_{k-1}(Y_1,Y_2) \arrow[r] & \cdots
\end{tikzcd}$$

where the vertical arrows come from the inclusions $Y_3 \hookrightarrow Y_2 \hookrightarrow Y_1$, and where superscripts indicate the spaces involved, is commutative.

Proof. Easy; for example, when we compute $i_k^{2,3} \circ \partial_k^{2,3}[\pi]$, we have an element π of $C_k(Y_2)$, whose boundary $\partial\pi$ is supported on Y_3, and we consider that as an element of $C_{k-1}(Y_2)$. If we apply $i_k^{1,2,3}$ to $[\pi]$, we consider π as an element of $C_k(Y_1)$ with boundary supported on $C_{k-1}(Y_2)$, and $\partial_k^{1,2}[\pi]$ is that boundary. Thus $i_k^{2,3} \circ \partial_k^{2,3} = \partial_k^{1,2} \circ i_k^{i,2,3}$. □

Lemma 6.10.3. *Let $Y_3 \subset Y_2 \subset Y_1$ be as above. Then the sequence*

$$\cdots \longrightarrow H_{k+1}(Y_1,Y_2) \xrightarrow{j_{k+1}^{2,3} \circ \partial_{k+1}^{1,2}} H_k(Y_2,Y_3) \xrightarrow{i_k^{1,2}} H_k(Y_1,Y_3) \xrightarrow{j_k^{1,2}} H_k(Y_1,Y_2) \longrightarrow \cdots$$

is exact. (Here, the map $i_k^{1,2}$ comes from the inclusion $Y_2 \hookrightarrow Y_1$, whereas $j_k^{1,2}$ arises from considering an element of $C_{k-1}(Y_1,Y_3)$ also as an element of $C_{k-1}(Y_1,Y_2)$ (since $Y_3 \subset Y_2$), in the same way as above).

Proof. Again a simple routine:

1) Exactness at $H_k(Y_2,Y_3)$:

$$i_k^{1,2}[\pi] = 0 \Leftrightarrow \exists \gamma \in C_{k+1}(Y_1,Y_3) : \partial\gamma = \pi,$$

and in fact, we may consider γ as an element of $C_{k+1}(Y_1,Y_2)$ as the class of π in $H_k(Y_2,Y_3)$ is not influenced by adding $\partial\omega$ for some $\omega \in C_{k+1}(Y_2)$. Thus π is in the image of $j_{k+1}^{2,3} \circ \partial_{k+1}^{1,2}$.

2) Exactness at $H_k(Y_1,Y_3)$:

$$j_k^{1,2}[\pi] = 0 \Leftrightarrow \exists \gamma \in C_{k+1}(Y_1,Y_2) : \partial\gamma = \pi,$$

and so π is trivial in homology up to an element of $C_k(Y_2,Y_3)$, and so it is in the image of $i_k^{i,2}$.

3) Exactness at $H_k(Y_1,Y_2)$:

$$\begin{aligned} j_k^{2,3} \circ \partial_k^{1,2}[\pi] = 0 &\Leftrightarrow \partial_k\pi \text{ vanishes up to an element of } C_{k-1}(Y_3) \\ &\Leftrightarrow \pi \text{ is in the image of } j_k^{1,2}. \end{aligned}$$

□

Finally, we need the following algebraic result:

Lemma 6.10.4. *Let*

$$\cdots \longrightarrow A_3 \xrightarrow{a_3} A_2 \xrightarrow{a_2} A_1 \xrightarrow{a_1} 0$$

be an exact sequence of linear maps between vector spaces. Then for all $k \in \mathbb{N}$,

$$\dim A_1 - \dim A_2 + \dim A_3 - \cdots - (-1)^k \dim A_k + (-1)^k \dim(\ker a_k) = 0. \quad (6.10.14)$$

Proof. For any linear map $\ell = V \to W$ between vector spaces,

$$\dim V = \dim(\ker \ell) + \dim(\operatorname{image} \ell).$$

Since by exactness

$$\dim(\operatorname{image} a_j) = \dim(\ker a_{j-1}),$$

we obtain

$$\dim(A_j) = \dim(\ker a_j) + \dim(\ker a_{j-1}).$$

Since $\dim A_1 = \dim \ker a_1$, we obtain

$$\dim A_1 - \dim A_2 + \dim A_3 - \cdots + (-1)^k \dim(\ker a_k) = 0.$$

□

We now apply Lemma 6.10.4 to the exact sequence of Lemma 6.10.3. With

$$b_k(X,Y) := \dim(H_k(X,Y)),$$
$$\nu_k(Y_1,Y_2,Y_3) = \dim(\ker j^{2,3}_{k+1} \circ \partial^{1,2}_k),$$

we obtain

$$\sum_{i=0}^{k} (-1)^i (b_i(Y_1,Y_2) - b_i(Y_1,Y_3) + b_i(Y_2,Y_3)) - (-1)^k \nu_k(Y_1,Y_2,Y_3) = 0.$$

Hence

$$\begin{aligned}(-1)^{k-1}\nu_{k-1}(Y_1,Y_2,Y_3) = \quad & (-1)^k \nu_k(Y_1,Y_2,Y_3) - (-1)^k b_k(Y_1,Y_2) \\ & + (-1)^k b_k(Y_1,Y_3) \qquad - (-1)^k b_k(Y_2,Y_3).\end{aligned} \quad (6.10.15)$$

We define the following polynomials in t:

$$P(t,X,Y) := \sum_{k\ge 0} b_k(X,Y)\, t^k,$$
$$Q(t,Y_1,Y_2,Y_3) := \sum_{k\ge 0} \nu_k(Y_1,Y_2,Y_3)\, t^k.$$

Multiplying the preceding equation by $(-1)^k t^k$ and summing over k, we obtain

$$\begin{aligned} Q(t, Y_1, Y_2, Y_3) = -tQ(t, Y_1, Y_2, Y_3) + P(t, Y_1, Y_2) \\ - P(t, Y_1, Y_3) + P(t, Y_2, Y_3). \end{aligned} \tag{6.10.16}$$

We now order the critical points $p_1, \dots, p_m$ of the function f in such a manner that

$$\mu(p_i) \geq \mu(p_j) \quad \text{whenever } i \leq j.$$

For any i, we put

$$\begin{aligned} Y_1 := Y_1(i) := &\bigcup_{k \geq i} Y(p_k), \\ Y_2 := Y_2(i) := &\bigcup_{k \geq i+1} Y(p_k), \\ Y_3 := &\ \emptyset. \end{aligned}$$

Thus $Y_2 = Y_1 \backslash W^k(p_i)$. The pair (Y_1, Y_2) may differ from the pair $(Y, A) = (Y(P_i), Y(P_i) \backslash W^k(p_i))$ in so far as both Y_1 and Y_2 may contain in addition the same unstable manifolds of some other critical points. Thus, they are of the form $(Y \cup B, A \cup B)$ for a certain set B. It is, however, obvious that the previous constructions are not influenced by adding a set B to both pairs, i.e. we have

$$H_k(Y \cup B, A \cup B) = H_k(Y, A) \quad \text{for all } k,$$

because all contributions in B cancel. Therefore, we have

$$H_k(Y_1(i), Y_2(i)) = H_k(Y(p_i), Y(p_i) \backslash W^k(p_i)) = \begin{cases} \mathbb{Z}_2 & \text{for } k = \mu(p_i), \\ 0 & \text{otherwise.} \end{cases} \tag{6.10.17}$$

Consequently,

$$P(t, Y_1, Y_2) = t^{\mu(p_i)}. \tag{6.10.18}$$

We now let μ_ℓ be the number of critical points of f of Morse index ℓ. Since the dimension of any unstable manifold is bounded by the dimension of X, we have $\mu_\ell = 0$ for $\ell > \dim X$. (6.10.18) implies

$$\sum_{i=0}^{\dim X} P(t, Y_1(i), Y_2(i)) = \sum_\ell t^\ell \mu_\ell. \tag{6.10.19}$$

From (6.10.16), we obtain for our present choice of the triple (Y_1, Y_2, Y_3)

$$P(t, Y_1(i), Y_2(i)) = P(t, Y_1(i), \emptyset) - P(t, Y_2(i), \emptyset) + (1+t)(X(t, Y_1(i), Y_2(i), \emptyset),$$

and summing w.r.t. i and using $Y_1(1) = X$, we obtain

$$\sum_{i=0}^{\dim X} P(t, Y_1(i), Y_2(i)) = P(t, X, \emptyset) + (1+t)\, Q(t) \tag{6.10.20}$$

for a polynomial $Q(t)$ with nonnegative coefficients. Inserting (6.10.19) in (6.10.20) and using the relation

$$\begin{aligned} P(t, X, \phi) &= \sum t^j \dim H_j(X) \quad &&(\text{since } H_j(X, \emptyset) = H_j(X)) \\ &= \sum t^j b_j(X) &&(\text{see Corollary 6.9.1}). \end{aligned}$$

we conclude

Theorem 6.10.2. *Let f be a Morse-Smale-Floer function on the compact, finite dimensional orientable Riemannian manifold X. Let μ_ℓ be the number of critical points of f of Morse index ℓ, and let $b_k(X)$ be the k-th Betti number of X. Then*

$$\sum_{\ell=0}^{\dim X} t^\ell \mu_\ell = \sum_j t^j b_j(X) + (1+t)\, Q(t) \tag{6.10.21}$$

for some polynomial $Q(t)$ in t with nonnegative integer coefficients.

We can now deduce the **Morse inequalities**

Corollary 6.10.1. *Let f be a Morse-Smale-Floer function on the compact, finite dimensional, orientable Riemannian manifold X. Then, with the notations of Theorem 6.10.2,*

(i) $\mu_k \geq b_k(X)$ *for all k.*

(ii) $\mu_k - \mu_{k-1} + \mu_{k-2} - ... \pm \mu_0 \geq b_k(X) - b_{k-1}(X) ... \pm b_0(X)$.

(iii) $\sum_j (-1)^j \mu_j = \sum_j (-1)^j b_j(X)$ *(this expression is called the* **Euler characteristic** *of X).*

Proof.

(i) The coefficients of t^k on both sides of (6.10.21) have to coincide, and $Q(t)$ has nonnegative coefficients.

(ii) Let $Q(t) = \sum t^i q_i$. From (6.10.21), we get the relation

$$\sum_{j=0}^{k} t^j \mu_j = \sum_{j=0}^{k} t^j b_j(X) + (1+t) \sum_{j=0}^{k-1} t^i q_i + t^k q_k$$

for the summands of order no larger than k. We put $t = -1$. Since $q_k \geq 0$, we obtain

$$\sum_{j=0}^{k} (-1)^{j-k} \mu_j \geq \sum_{j=0}^{k} (-1)^{j-k} b_j.$$

(iii) We put $t = -1$ in (6.10.21).

□

Let us briefly return to the example discussed in §6.1 in the light of the present constructions. We obtain interesting aspects only for the function f_2 of §6.1. The essential feature behind the Morse inequality (i) is that for a triple (Y_1, Y_2, Y_3) satisfying $Y_3 \subset Y_2 \subset Y_1$ as in our above constructions, we always have

$$b_k(Y_1, Y_3) \le b_k(Y_1, Y_2) + b_k(Y_2, Y_3). \tag{6.10.22}$$

In other words, by inserting the intermediate space Y_2 between Y_1 and Y_3, we may increase certain topological quantities, by inhibiting cancellations caused by the boundary operator ∂. If, in our example from §6.1, we take $Y_1 = X, Y_3 = \emptyset$, we may take any intermediate Y_2. If we take $Y_2 = Y(p_2)$ (p_2 being one of two maximum points), then $Y_1 \backslash Y_2 = W^k(p_1)$ (p_1 the other maximum), and so

$$b_k(Y_1, Y_2) = \begin{cases} 1 & \text{for } k = 2, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$b_k(Y_2, Y_3) = \begin{cases} 1 & \text{for } k = 0, \\ 0 & \text{otherwise,} \end{cases}$$

(we have $\partial p_2 = p_3, \partial p_3 = 2p_4 = 0$ in Y_2), and so, since

$$b_k(X) = \begin{cases} 1 & \text{for } k = 0, 2, \\ 0 & \text{for } k = 1, \end{cases}$$

we have equality in (6.10.22).
If we take $Y_2 = Y(p_3)$ (p_3 the saddle point), however, we get

$$b_k(Y_1, Y_2) = \begin{cases} 2 & \text{for } k = 2, \\ 0 & \text{otherwise,} \end{cases}$$

(since $\partial p_1 = 0 = \partial p_2$ in (Y_1, Y_2)) and

$$b_k(Y_2, Y_3) = \begin{cases} 1 & \text{for } k = 1, \\ 0 & \text{otherwise,} \end{cases}$$

(since $\partial p_3 = 0$, but there are no critical points of index 2 in Y_2). Thus, in the first case, the boundary operator ∂ still achieved a cancellation between the second maximum and the saddle point while in the second case, this was prevented by placing p_2 and p_3 into different sets. Generalizing this insight, we conclude that the Morse numbers μ_ℓ arise from placing all critical points in different sets and thus gathering only strictly local information while the Betti numbers b_ℓ incorporate all the cancellations induced by the boundary operator ∂. Thus, the μ_ℓ and the b_ℓ only coincide if no cancellations at all take place, as in the example of the function f_1 in §6.1.

Perspectives. In this section, we have interpreted the insights of Morse theory, as developed by Thom[250], Smale[237], Milnor[192], Franks[83] 199-215, in the light of Floer's approach. Schwarz[231] used these constructions to construct an explicit isomorphism between Morse homology and singular homology.

6.11 The Palais-Smale Condition and the Existence of Closed Geodesics

Let M be a compact Riemannian manifold of dimension n, with metric $\langle \cdot, \cdot \rangle$ and associated norm $||\cdot|| = \langle \cdot, \cdot \rangle^{\frac{1}{2}}$. We wish to define the Sobolev space $\Lambda_0 = H^1(S^1, M)$ of closed curves on M with finite energy, parametrized on the unit circle S^1. We first consider $H^1(I, \mathbb{R}^n) := H^{1,2}(I, \mathbb{R}^n)$, where I is some compact interval $[a, b]$, as the closure of $C^\infty(I, \mathbb{R}^n)$ w.r.t. the Sobolev $H^{1,2}$-norm. This norm is induced by the scalar product

$$(c_1, c_2) := \int_a^b c_1(t) \cdot c_2(t)\, dt + \int_a^b \frac{dc_1(t)}{dt} \cdot \frac{dc_2(t)}{dt}\, dt\,, \tag{6.11.1}$$

where the dot $\cdot$ denotes the Euclidean scalar product on $\mathbb{R}^n$. $H^1(I, \mathbb{R}^n)$ then is a Hilbert space.

Since I is 1-dimensional, by Sobolev's embedding theorem (Theorem A.1.7), all elements in $H^1(I, \mathbb{R}^n)$ are continuous curves. Therefore, we can now define the Sobolev space $H^1(S^1, M)$ of Sobolev curves in M via localization with the help of local coordinates:

Definition 6.11.1. The Sobolev space $\Lambda_0 = H^1(S^1, M)$ is the space of all those curves $c : S^1 \to M$ for which for every chart $x : U \to \mathbb{R}^n$ (U open in M), (the restriction to any compact interval of)

$$x \circ c : c^{-1}(U) \to \mathbb{R}^n$$

is contained in the Sobolev space $H^{1,2}(c^{-1}(U), \mathbb{R}^n)$.

Remark. The space Λ_0 can be given the structure of an infinite dimensional Riemannian manifold, with charts modeled on the Hilbert space $H^{1,2}(I, \mathbb{R}^n)$. Tangent vectors at $c \in \Lambda_0$ then are given by curves $\gamma \in H^1(S^1, TM)$, i.e. Sobolev curves in the tangent bundle of M, with $\gamma(t) \in T_{c(t)}M$ for all $t \in S^1$. For $\gamma_1\gamma_2 \in T_c\Lambda_0$, i.e. tangent vectors at c, their product is defined as

$$(\gamma_1, \gamma_2) := \int\limits_{t \in S^1} \langle D\gamma_1(t), D\gamma_2(t)\rangle \, dt \,,$$

where $D\gamma_i(t)$ is the weak first derivative of γ_i at t, as defined in §A.1. This then defines the Riemannian metric of Λ_0. While this becomes conceptually very satisfactory, one needs to verify a couple of technical points to make this completely rigorous. For that reason, we rather continue to work with ad hoc constructions in local coordinates. In any case, Λ_0 assumes the role of the space X in the general context described in the preceding sections.

The Sobolev space Λ_0 is the natural space on which to define the energy functional

$$E(c) = \frac{1}{2} \int\limits_{S^1} ||Dc(t)||^2 \, dt$$

for curves $c : S^1 \to M$, with Dc denoting the weak first derivative of c.

Definition 6.11.2. $(u_n)_{n\in\mathbb{N}} \subset \Lambda_0$ converges to $u \in \Lambda_0$ in $H^{1,2}$ iff

(i) u_n converges uniformly to u $(u_n \rightrightarrows u)$.

(ii) $E(u_n) \to E(u)$ as $n \to \infty$.

Uniform convergence $u_n \rightrightarrows u$ implies that there exist coordinate charts $f_\mu : U_\mu \to \mathbb{R}^n$ $(\mu = 1, \ldots, m)$ and a covering of $S^1 = \cup_{\mu=1}^m V_\mu$ by open sets such that for sufficiently large n,

$$u_n(V_\mu), u(V_\mu) \subset U_\mu \quad \text{for } \mu = 1, \ldots, m.$$

If now $\varphi \in C_0^\infty(V_\mu, \mathbb{R}^n)$ for some μ, then for sufficiently small $|\varepsilon|$,

$$f_\mu(u(t) + \varepsilon\varphi(t)) \subset f_\mu(U_\mu) \quad \text{for all } t \in V_\mu,$$

i.e. we can perform local variations without leaving the coordinate chart. In this sense we write

$$u + \varepsilon\varphi$$

instead of $f_\mu \circ u + \varepsilon\varphi$. For such φ then

$$\frac{d}{d\varepsilon}E(u + \varepsilon\varphi)_{|\varepsilon=0} = \frac{1}{2}\frac{d}{d\varepsilon}\int g_{ij}(u + \varepsilon\varphi)(\dot{u}^i + \varepsilon\dot{\varphi}^i)(\dot{u}^j + \varepsilon\dot{\varphi}^j)\, dt_{|\varepsilon=0} \,,$$

where everything is written w.r.t. the local coordinate $f_\mu : U_\mu \to \mathbb{R}^n$ (the dot $\dot{}$ of course denotes a derivative w.r.t. $t \in S^1$),
using $g_{ij} = g_{ji}$,

$$= \int \left(g_{ij}(u)\dot{u}^i\dot{\varphi}^j + \frac{1}{2}g_{ij,k}(u)\dot{u}^i\dot{u}^j\varphi^k \right) dt \qquad (6.11.2)$$

if $u \in H^{2,2}(S^1, M)$, this is

$$= -\int \left(g_{ij}(u)\ddot{u}^i\varphi^j + g_{ij,\ell}\dot{u}^\ell\dot{u}^i\varphi^j - \frac{1}{2}g_{ij,k}\dot{u}^i\dot{u}^j\varphi^k \right) dt \qquad (6.11.3)$$
$$= -\int \left(\ddot{u}^i + \Gamma^i_{k\ell}(u)\dot{u}^k\dot{u}^\ell\right) g_{ij}(u)\varphi^j \, dt\,,$$

as in §1.4. We observe that $\varphi \in H^{1,2}$ is bounded by Sobolev's embedding theorem (Theorem A.1.7) (see also the argument leading to (6.11.6) below) so that also the second terms in (6.11.2) and (6.11.3) are integrable.

We may put

$$\|DE(u)\| = \sup\left\{ \frac{d}{d\varepsilon}E(u+\varepsilon\varphi)_{|\varepsilon=0} : \varphi \in H_0^{1,2}(V_\mu, \mathbb{R}^n) \text{ for some } \mu, \int g_{ij}(u)\dot{\varphi}^i\dot{\varphi}^j \, dt \le 1 \right\}. \qquad (6.11.4)$$

For second derivatives of E, we may either quote the formula of Theorem 4.1.1 or compute directly in local coordinates

$$\frac{d^2}{d\varepsilon^2}E(u+\varepsilon\varphi)_{|\varepsilon=0} = \frac{1}{2}\frac{d^2}{d\varepsilon^2}\int g_{ij}(u+\varepsilon\varphi)(\dot{u}^i+\varepsilon\dot{\varphi}^i)(\dot{u}^j+\varepsilon\dot{\varphi}^j)\, dt$$
$$= \int \left(g_{ij}(u)\dot{\varphi}^i\dot{\varphi}^j + 2g_{ij,k}\dot{u}^i\dot{\varphi}^j\varphi^k + g_{ij,k\ell}\dot{u}^i\dot{u}^j\varphi^k\varphi^\ell \right) dt,$$

which is also bounded for u and φ of Sobolev class $H^{1,2}$.

Suppose now that $u \in \Lambda_0$ satisfies

$$DE(u) = 0.$$

This means

$$\int \left(g_{ij}(u)\dot{u}^i\dot{\varphi}^j + \frac{1}{2}g_{ij,k}(u)\dot{u}^i\dot{u}^j\varphi^k \right) dt = 0 \quad \text{for all } \varphi \in H^{1,2}. \qquad (6.11.5)$$

Lemma 6.11.1. *Any $u \in \Lambda_0$ with $DE(u) = 0$ is a closed geodesic (of class C^∞).*

Proof. We have to show that u is smooth. Then (6.11.3) is valid, and Theorem A.1.5 gives

$$\ddot{u}^i + \Gamma^i_{k\ell}(u)\dot{u}^k\dot{u}^\ell = 0 \quad \text{for } i = 1, \ldots, \dim M,$$

thus u is geodesic.

We note that u is continuous so that we can localize in the image. More precisely, we can always find sufficiently small subsets of S^1 whose image is contained in one coordinate chart. Therefore, we may always write our formulae in local coordinates. We first want to show

$$u \in H^{2,1}.$$

For this, we have to find $v \in L^1$ with

$$\int u^i \ddot{\eta}_i = \int v^i \eta_i,$$

where we always assume that the support of $\eta \in C_0^\infty(S^1, M)$ is contained in a small enough subset of S^1 so that we may write things in local coordinates as explained before.

We put

$$\varphi^j(t) := g^{ij}(u(t))\eta_i(t).$$

Then

$$\int u^i \ddot{\eta}_i \, dt = -\int \dot{u}^i \dot{\eta}_i \, dt$$

which is valid since $u \in H^{1,2}$,

$$\begin{aligned}
&= -\int \left(g_{ij}(u(t))\dot{u}^i \dot{\varphi}^j + g_{ij,k}\dot{u}^k \dot{u}^i \varphi^j \right) dt \\
&= \int \left(\frac{1}{2} g_{ij,k}(u)\dot{u}^i \dot{u}^j \varphi^k - g_{ij,k}(u)\dot{u}^k \dot{u}^i \varphi^j \right) dt \quad \text{by (6.11.5)} \\
&= \int \left(\frac{1}{2} g_{ij,k} g^{k\ell} \dot{u}^i \dot{u}^j - g_{ij,k}\dot{u}^k \dot{u}^i g^{j\ell} \right) \eta_\ell \, dt \\
&= \int \left(\frac{1}{2} g^{i\ell}(g_{jk,\ell} - g_{j\ell,k} - g_{k\ell,j} \right) \dot{u}^j \dot{u}^k \eta_i \, dt, \quad \text{renaming indices} \\
&= -\int \Gamma^i_{jk} \dot{u}^j \dot{u}^k \eta_i \, dt \, .
\end{aligned} \tag{6.11.6}$$

With $v^i = -\Gamma^i_{jk}\dot{u}^j \dot{u}^k \in L^1$, the desired formula

$$\int u^i \ddot{\eta}_i = \int v^i \eta_i \quad \text{for } \eta \in C_0^\infty(S^1, M) \text{ with sufficiently small support}$$

then holds, and

$$u \in H^{2,1}.$$

By the Sobolev embedding theorem (Theorem A.1.7) we conclude

$$u \in H^{1,q} \quad \text{for all } q < \infty.$$

(We note that since S^1 has no boundary, the embedding theorem holds for the $H^{k,p}$ spaces and not just for $H_0^{k,p}$. For the norm estimates, however, one needs

$\|f\|_{H^{k,p}(\Omega)}$ on the right hand sides in Theorem A.1.7 and Corollary A.1.2, instead of just $\|D^k f\|_{L^p}$.)

In particular, $u \in H^{1,4}(\Omega)$, hence

$$\Gamma^i_{jk}(u)\dot{u}^j\dot{u}^k \in L^2.$$

(6.11.6) then implies

$$u \in H^{2,2},$$

hence $\dot{u} \in C^0$ by Theorem A.1.2 again.

Now

$$\begin{aligned}\frac{d}{dt}(\Gamma^i_{jk}(u)\dot{u}^j\dot{u}^k) &= 2\Gamma^i_{jk}\dot{u}^j\ddot{u}^k + \Gamma^i_{jk,\ell}\dot{u}^\ell\dot{u}^j\dot{u}^k \qquad \text{using } \Gamma^i_{jk} = \Gamma^i_{kj}\\ &\in L^2,\end{aligned}$$

since $\ddot{u} \in L^2$, $\dot{u} \in L^\infty$. Thus

$$\Gamma^i_{jk}(u)\dot{u}^j\dot{u}^k \in H^{1,2},$$

and then

$$u \in H^{3,2},$$

by (6.11.6) again.

Iterating this argument, we conclude

$$u \in H^{k,2} \quad \text{for all } k \in \mathbb{N},$$

hence

$$u \in C^\infty$$

by Corollary A.1.2. □

We now verify a version of the Palais-Smale condition:

Theorem 6.11.1. *Any sequence* $(u_n)_{n\in\mathbb{N}} \subset \Lambda_0$ *with*

$$E(u_n) \le \text{const},$$
$$\|DE(u_n)\| \to 0 \quad \text{as } n \to 0$$

contains a strongly convergent subsequence with a closed geodesic as limit.

Proof. First, by Hölder's inequality, for every $v \in \Lambda_0, t_1, t_2 \in S^1$,

$$\begin{aligned} d(v(t_1), v(t_2)) &\le \int_{t_1}^{t_2} \left(g_{ij}(v)\dot{v}^i\dot{v}^j\right)^{\frac{1}{2}}\, dt \\ &\le \left((t_2 - t_1)\int_{t_1}^{t_2} g_{ij}(v)\dot{v}^i\dot{v}^j\, dt\right)^{\frac{1}{2}} \\ &\le \sqrt{2}\,|t_2 - t_1|^{\frac{1}{2}} E(v)^{\frac{1}{2}}. \end{aligned} \tag{6.11.7}$$

Thus

$$\Lambda_0 \subset C^{\frac{1}{2}}(S^1, M),$$

i.e. every H^1-curve is Hölder continuous with exponent $\frac{1}{2}$, and the Hölder $\frac{1}{2}$-norm is controlled by $\sqrt{2E(v)}$.

The Arzela-Ascoli theorem therefore implies that a sequence with $E(u_n) \le \text{const}$ contains a uniformly convergent subsequence. We call the limit u. u also has finite energy, actually

$$E(u) \le \liminf_{n\to\infty} E(u_n).$$

We could just quote Theorem 7.3.2 below. Alternatively, by uniform convergence everything can be localized in coordinate charts, and lower semicontinuity may then be verified directly. For our purposes it actually suffices at this point that u has finite energy, and this follows because the $H^{1,2}$-norm (defined w.r.t. local coordinates) is lower semicontinuous under L^2-convergence.

We now let $(\eta_\mu)_{\mu=1,\dots,m}$ be a partition of unity subordinate to $(V_\mu)_{\mu=1,\dots,m}$, our covering of S^1 as above. Then

$$E(u_n) - E(u) = \int \sum_{\mu=1}^{m} \eta_\mu \left(g_{ij}^{\mu}(u_n)\dot{u}_n^i\dot{u}_n^j - g_{ij}^{\mu}(u)\dot{u}^i\dot{u}^j\right) dt\,, \tag{6.11.8}$$

where the superscript μ now refers to the coordinate chart $f_\mu : U_\mu \to \mathbb{R}^n$.

In the sequel, we shall omit this superscript, however.

By assumption (cf. (6.11.2)),

$$\int \left(g_{ij}(u_n)\dot{u}_n^i\dot{\varphi}^j + \frac{1}{2}g_{ij,k}(u_n)\dot{u}_n^i\dot{u}_n^j\varphi^k\right) dt \to 0 \quad \text{as } n \to \infty,$$

for all $\varphi \in H^{1,2}$.

We use

$$\varphi^j = \eta_\mu(u_n^j - u^j)$$

(where, of course, the difference is computed in local coordinates f_μ). Then,

$$\int g_{ij,k}(u_n)\dot{u}_n^i\dot{u}_n^j\eta_\mu(u_n^k - u^k)\, dt \le \text{const}\cdot\max_t d(u_n(t), u(t))\, E(u_n) \to 0,$$

as $n \to \infty$ since $E(u_n) \le \text{const}$ and $u_n \rightrightarrows u$ (after selecting a subsequence).

Consequently from (6.11.2), since $\|DE(u_n)\| \to 0$,

$$\int \left(g_{ij}(u_n)\dot{u}_n^i(\dot{u}_n^j - \dot{u}^j)\eta_\mu + g_{ij}(u^n)\dot{u}_n^i\dot{\eta}_\mu(u_n^j - u^j)\right) dt \to 0.$$

The second term again goes to zero by uniform convergence.

We conclude

$$\int g_{ij}(u_n)\dot{u}_n^i(\dot{u}_n^j - \dot{u}^j)\eta_\mu \to 0 \quad \text{as } n \to \infty. \tag{6.11.9}$$

Now

$$\int \left(g_{ij}(u_n)\dot{u}_n^i\dot{u}_n^j - g_{ij}(u)\dot{u}^i\dot{u}^j\right)\eta_\mu =$$
$$\int \left\{g_{ij}(u_n)\dot{u}_n^i(\dot{u}_n^j - \dot{u}^j) + (g_{ij}(u_n) - g_{ij}(u))\,\dot{u}_n^i\dot{u}^j + g_{ij}(u)(\dot{u}_n^i - \dot{u}^i)\dot{u}^j\right\}\eta_\mu\,. \tag{6.11.10}$$

The first term goes to zero by (6.11.9). The second one goes to zero by uniform convergence and Hölder's inequality.

For the third one, we exploit that (as observed above, after selection of a subsequence) $\dot{u}_n$ converges weakly in L^2 to $\dot{u}$ on V_μ. This implies that the third term goes to zero as well.

(6.11.10) now implies

$$E(u_n) \to E(u) \text{ as } n \to \infty$$

(cf. (6.11.8)).

u then satisfies

$$DE(u) = 0$$

and is thus geodesic by Lemma 6.11.1. □

As a technical tool, we shall have to consider the negative gradient flow of E.

Remark. In principle, this is covered by the general scheme of §6.3, but since we are working with local coordinates here and not intrinsically, we shall present the construction in detail. For those readers who are familiar with ODEs in Hilbert manifolds, the essential point is that the Picard-Lindelöf theorem applies because the second derivative of E is uniformly bounded on sets of curves with uniformly bounded energy E. Therefore, the negative gradient flow for E exists for all positive times, and by the Palais-Smale condition always converges to a critical point of E, i.e. a closed geodesic.

The gradient of E, ∇E, is defined by the requirement that for any $c \in \Lambda_0$, $\nabla E(c)$ is the H^1-vector field along c satisfying for all H^1-vector fields along c

$$(\nabla E(c), V)_{H^1} = DE(c)(V) = \int_{S^1} \langle \dot{c}, \dot{V}\rangle\, dt\,. \tag{6.11.11}$$

Since the space of H^1-vector fields along c is a Hilbert space, $\nabla E(c)$ exists by the Riesz representation theorem. (The space of H^1-vector fields along an H^1-curve can be defined with the help of local coordinates).

We now want to solve the following differential equation in Λ_0:

$$\begin{aligned} \frac{d}{dt}\Phi(t) &= -\nabla E(\Phi(t)), \\ \Phi(0) &= c_0, \end{aligned} \tag{6.11.12}$$

where $c_0 \in \Lambda_0$ is given and $\Phi : \mathbb{R}^+ \to \Lambda_0$ is to be found.

We first observe

Lemma 6.11.2. *Let $\Phi(t)$ be a solution of (6.11.12). Then*

$$\frac{d}{dt}E(\Phi(t)) \le 0.$$

Proof. By the chain rule,

$$\begin{aligned} \frac{d}{dt}E(\Phi(t)) &= DE(\Phi(t))\left(\frac{d}{dt}\Phi(t)\right) \\ &= -\|\nabla E(\Phi(t))\|_{H^1}^2 \le 0. \end{aligned} \tag{6.11.13}$$

□

Theorem 6.11.2. *For any $c_0 \in \Lambda_0$, there exists a solution $\Phi : \mathbb{R}^+ \to \Lambda_0$ of*

$$\begin{aligned} \frac{d}{dt}\Phi(t) &= -\nabla E(\Phi(t)), \\ \Phi(0) &= c_0. \end{aligned} \tag{6.11.14}$$

Proof. Let

$$A := \{T > 0 : \text{ there exists } \Phi : [0,T] \to \Lambda_0 \text{ solving (6.11.14) with } \Phi(0) = c_0\}.$$

(That Φ is a solution on $[0,T]$ means that there exists some $\varepsilon > 0$ for which Φ is a solution on $[0, T + \varepsilon)$.)

We are going to show that A is open and nonempty on the one hand and closed on the other hand. Then $A = \mathbb{R}^+$, and the result will follow. To show that A is open and nonempty, we are going to use the theory of ODEs in Banach spaces. For $c \in \Lambda_0$, we have the following bijection between a neighborhood U of c in Λ_0 and a neighborhood V of 0 in the Hilbert space of H^1-vector fields along c:

$$\xi(\tau) \mapsto \exp_{c(\tau)} \xi(\tau) \quad \text{for } \xi \in V. \tag{6.11.15}$$

(By Theorem 1.4.3 and compactness of c, there exists $\rho_0 > 0$ with the property that for all $\tau \in S^1$ $\exp_{c(\tau)}$ maps the ball $B(0, \rho_0)$ in $T_{c(\tau)}M$ diffeomorphically onto its image in M.)

If Φ solves (6.11.14) on $[s, s+\varepsilon]$ we may assume that $\varepsilon > 0$ is so small that for all t with $s \leq t \leq s + \varepsilon$, $\Phi(t)$ stays in a neighborhood U of $c = \Phi(s)$ with the above property. This follows because Φ, since differentiable, in particular is continuous in t. Therefore, (6.11.15) transforms our differential equation (with its solution $\Phi(t)$ having values in U for $s \leq t < s + \varepsilon$) into a differential equation in V, an open subset of a Hilbert space. Since DE, hence ∇E is continuously differentiable, hence Lipschitz continuous, the standard existence result for ODE (theorem of Cauchy or Picard-Lindelöf) may be applied to show that given any $c \in \Lambda_0$, there exists $\varepsilon > 0$ and a unique solution $\Psi : [0, \varepsilon] \to \Lambda_0$ of $\frac{d}{dt}\Psi(t) = -\nabla E(\Psi(t))$ with $\Psi(0) = c$. If Φ solves (6.11.14) on $[0, t_0]$, then putting $c = \Phi(t_0)$, we get a solution on $[0, t_0 + \varepsilon]$, putting $\Phi(t) = \Psi(t - t_0)$.

This shows openness, and also nonemptyness, putting $t_0 = 0$. To show closedness, suppose $\Phi : [0, t) \to \Lambda_0$ solves (6.11.14), and $0 < t_n < T, t_n \to T$ for $n \to \infty$.

Lemma 6.11.2 implies

$$E(\Phi(t_n)) \leq \text{const}\,. \tag{6.11.16}$$

Therefore, the curves $\Phi(t_n)$ are uniformly Hölder continuous (cf. (6.11.7)), and hence, by the theorem of Arzela-Ascoli, after selection of a subsequence, they converge uniformly to some $c_T \in \Lambda_0$; c_T indeed has finite energy because we may assume that $(\Phi(t_n))_{n\in\mathbb{N}}$ also converges weakly in $H^{1,2}$ to c_T, as in the proof of Theorem 6.11.1. By the openness argument, consequently we can solve

$$\begin{aligned}\frac{d}{dt}\Phi(t) &= -\nabla E(\Phi(t)),\\ \Phi(t) &= c_T\end{aligned}$$

for $T \leq t \leq T+\varepsilon$ and some $\varepsilon > 0$. Thus, we have found $\Phi : [0, T+\varepsilon)$ solving (6.11.14), and closedness follows. $\square$

We shall now display some applications of the Palais-Smale condition for closed geodesics. The next result holds with the same proof for any C^2-functional on a Hilbert space satisfying (PS) with two strict local minima.

While this result is simply a variant of Proposition 6.2.1 above, we shall present the proof once more as it will serve as an introduction to the proof of the theorem of Lyusternik and Fet below.

Theorem 6.11.3. *Let c_1, c_2 be two homotopic closed geodesics on the compact Riemannian manifold M which are strict local minima for E (or, equivalently, for the length functional L). Then there exists another closed geodesic c_3 homotopic to c_1, c_2 with*

$$E(c_3) = \kappa := \inf_{\lambda\in\Lambda} \max_{\tau\in[0,1]} E(\lambda(\tau)) > \max\{E(c_1), E(c_2)\}\,, \tag{6.11.17}$$

where $\Lambda := \Lambda(c_1, c_2) := \{\lambda \in C^0([0,1], \Lambda_0) : \lambda(0) = c_1, \lambda(1) = c_2\}$, the set of all homotopies between c_1 and c_2.

Proof. We first claim

$$\begin{aligned} &\exists \delta_0 > 0 \ \forall \delta \text{ with } 0 < \delta \le \delta_0 \ \exists \varepsilon > 0 \ \forall c \text{ with } d_1(c, c_i) = \delta : \\ &\qquad E(c) \ge E(c_i) + \varepsilon \quad \text{for } i = 1, 2. \end{aligned} \tag{6.11.18}$$

Indeed, otherwise, for $i = 1$ or 2,

$$\begin{aligned} &\forall \delta_0 \ \exists 0 < \delta \le \delta_0 \ \forall n \ \exists \gamma_n \text{ with } d_1(\gamma_n, c_i) = \delta : \\ &\qquad E(\gamma_n) < E(c_i) + \frac{1}{n}. \end{aligned}$$

If $\|DE(\gamma_n)\| \to 0$, then (γ_n) is a Palais-Smale sequence and by Theorem 6.11.1 converges (after selection of a subsequence) to some γ_0 with

$$\begin{aligned} d_1(\gamma_0, c_i) &= \delta, \\ E(\gamma_0) &= E(c_i), \end{aligned}$$

contradicting the strict local minimizing property of c_i.

If $\|DE(\gamma_n)\| \ge \eta > 0$ for all n, then there exists $\rho > 0$ with

$$\|DE(\gamma)\| \ge \frac{\eta}{2} \quad \text{whenever } d_1(\gamma_n, \gamma) \le \rho. \tag{6.11.19}$$

This follows, because $\|D^2 E\|$ is uniformly bounded on E-bounded sets.

(6.11.19) can then be used to derive a contradiction to the local minimizing property of c_i by a gradient flow construction. Such a construction will be described in detail below. We may thus assume that (6.11.18) is correct.

(6.11.18) implies

$$\kappa > \max(E(c_1), E(c_2)). \tag{6.11.20}$$

We let now K^κ be the set of all closed geodesics, i.e. curves c in Λ^0 with $DE(c) = 0$, $E(c) = \kappa$, homotopic to c_1 and c_2.

We have to show

$$K^\kappa \ne \emptyset.$$

We assume on the contrary

$$K^\kappa = \emptyset. \tag{6.11.21}$$

We claim that there exist $\eta > 0$, $\alpha > 0$ with

$$\|DE(c)\| \ge \alpha, \tag{6.11.22}$$

whenever c is homotopic to c_1, c_2 and satisfies

$$\kappa - \eta \le E(c) \le \kappa + \eta. \tag{6.11.23}$$

Namely, otherwise, there exists a sequence $(\gamma_n)_{n\in\mathbb{N}}$ of H^1-curves homotopic to c_1, c_2, with

$$\begin{aligned} \lim_{n\to\infty} E(\gamma_n) &= \kappa, \\ \lim_{n\to\infty} DE(\gamma_n) &= 0. \end{aligned}$$

$(\gamma_n)_{n\in\mathbb{N}}$ then is a Palais-Smale sequence and converges to a closed geodesic c_3 with $E(c_3) = \kappa$, contradicting our assumption $K^\kappa = \emptyset$.

Thus (6.11.22) has to hold if $\kappa - \eta \le E(c) \le \kappa + \eta$.

From Theorem 6.11.2, we know that for any $t > 0$, there is a map

$$\begin{aligned} \Lambda_0 &\to \Lambda_0, \\ c &\mapsto \Phi_t(c), \end{aligned}$$

where $\Phi_t(c) = \Phi(t)$ solves

$$\begin{aligned} \frac{d}{dt}\Phi(t) &= -\nabla E(\Phi(t)), \\ \Phi(0) &= c. \end{aligned}$$

With the help of this gradient flow, we may now decrease the energy below the level κ, contradicting (6.11.21). For that purpose, let $\lambda \in \Lambda$ satisfy

$$\max_{\tau\in[0,1]} E(\lambda(\tau)) \le \kappa + \eta. \tag{6.11.24}$$

Then, as in the proof of Lemma 6.11.2,

$$\frac{d}{dt}E(\Phi_t(\lambda(\tau))) = -\|\nabla E(\Phi_t(\lambda(\tau)))\|^2 \le 0. \tag{6.11.25}$$

In particular, for $t > 0$,

$$\max E(\Phi_t(\lambda(\tau))) \le \max E(\lambda(\tau)) \le \kappa + \eta. \tag{6.11.26}$$

Since c_1 and c_2 are closed geodesics, i.e. critical points of E, $\nabla E(c_i) = 0$ for $i = 1, 2$, hence

$$\Phi_t(c_i) = c_i \quad \text{for all } t \ge 0.$$

Therefore

$$\Phi_t \circ \lambda \in \Lambda \quad \text{for } t \ge 0.$$

(6.11.22), (6.11.25) imply

$$\frac{d}{dt}E(\Phi_t(\lambda(\tau))) \le -\alpha^2 \quad \text{whenever } E(\Phi_t(\lambda(\tau))) > \kappa - \eta. \tag{6.11.27}$$

(6.11.24), (6.11.27) imply

$$E(\Phi_s(\lambda(\tau))) \le \kappa - \eta,$$

for $s \ge \frac{2\eta}{\alpha^2}$ and all $\tau \in [0, 1]$, contradicting the definition of κ. Therefore, (6.11.21) cannot hold, and the theorem is proved. $\square$

As the culmination of this section, we now prove the **theorem of Lyusternik and Fet**

Theorem 6.11.4. *Each compact Riemannian manifold contains a nontrivial closed geodesic.*

For the proof, we shall need the following result from algebraic topology which, however, we do not prove here. (A proof may be found e.g. in E. Spanier, Algebraic topology, McGraw Hill, 1966.)

Lemma 6.11.3. *Let M be a compact manifold of dimension n. Then there exist some $i, 1 \leq i \leq n$, and a continuous map*

$$h : S^i \to M,$$

which is not homotopic to a constant map.

In case M is a differentiable manifold, then h can also be chosen to be differentiable. □

Proof of Theorem 6.11.4. We start with a very simple construction that a reader with a little experience in topology may skip.

Let i be as in Lemma 6.11.3. If $i = 1$, the result is a consequence of Theorem 1.5.1. We therefore only consider the case $i \geq 2$. h from Lemma 6.11.3 then induces a continuous map H of the $(i-1)$-cell D^{i-1} into the space of differentiable curves in M, mapping ∂D^{i-1} to point curves. In order to see this, we first identify D^{i-1} with the half equator $\{x^1 \geq 0, x^2 = 0\}$ of the unit sphere S^i in $\mathbb{R}^{i+1}$ with coordinates $(x^1, \ldots, x^{i+1})$. To $p \in D^{i-1} \subset S^i$, we assign that circle $c_p(t), t \in [0,1]$, parametrized proportionally to arc length that starts at p orthogonally to the hyperplane $\{x^2 = 0\}$ into the half sphere $\{x^2 \geq 0\}$ with constant values of $x^3, \ldots, x^{i+1}$. For $p \in \partial D^{i-1}$, c_p then is the trivial (i.e. constant) circle $c_p(t) = p$. The map H is then given by

$$H(p)(t) := h \circ c_p(t).$$

Each $q \in S^i$ then has a representation of the form $q = c_p(t)$ with $p \in D^{i-1}$. p is uniquely determined, and t as well, unless $q \in \partial D^{i-1}$. A homotopy of H, i.e. a continuous map

$$\tilde{H} : D^{i-1} \times [0,1] \to \{\text{closed curves in } M\}$$

that maps $\partial D^{i-1} \times [0,1]$ to point curves and satisfies $\tilde{H}_{|D^{i-1} \times \{0\}} = H$, then induces a homotopy $\tilde{h} : S^i \times [0,1] \to M$ of h by

$$\tilde{h}(q,s) = \tilde{h}(c_p(t), s) = \tilde{H}(p,s)(t),$$

($q = c_p(t)$, as just described).

We now come to the core of the proof and consider the space

$$\begin{aligned}\Lambda := \{\lambda : D^{i-1} \to \Lambda_0, \lambda &\text{ homotopic to } H \text{ as described above,}\\ &\text{in particular mapping } \partial D^{i-1} \text{to point curves } \},\end{aligned}$$

and put

$$\kappa := \inf_{\lambda \in \Lambda} \max_{z \in D^{i-1}} E(\lambda(z)).$$

As in the proof of Theorem 6.11.3, we see that there exists a closed geodesic γ with

$$E(\gamma) = \kappa.$$

It only remains to show that $\kappa > 0$, in order to exclude that γ is a point curve and trivial. Should $\kappa = 0$ hold, however, then for every $\varepsilon > 0$, we would find some $\lambda_\varepsilon \in \Lambda$ with

$$\max_{z \in D^{i-1}} E(\lambda_\varepsilon(z)) < \varepsilon.$$

All curves $\lambda_\varepsilon(z)$ would then have energy less than ε. We choose $\varepsilon < \frac{\rho_0^2}{2}$. Then, for every curve $c_z := \lambda_\varepsilon(z)$ and each $t \in [0,1]$,

$$d(c_z(0), c_z(t))^2 \leq 2E(c_z) < \rho_0^2.$$

The shortest connection from $c_z(0)$ to $c_z(t)$ is uniquely determined; denote it by $q_{z,t}(s), s \in [0,1]$. Because of its uniqueness, $q_{z,t}$ depends continuously on z and t. $\bar{H}(z,s)(t) := q_{z,t}(1-s)$ then defines a homotopy between λ_ε and a map that maps D^{i-1} into the space of point curves in M, i.e. into M.

Such a map, however, is homotopic to a constant map, for example since D^{i-1} is homotopically equivalent to a point. (The more general maps from D^{i-1} considered here into the space of closed curves on M are not necessarily homotopic to constant maps since we have imposed the additional condition that $\partial D^{i-1} = S^{i-2}$ is mapped into the space of point curves which is a proper subspace of the space of all closed curves.) This implies that λ_ε is homotopic to a constant map, hence so are H and h, contradicting the choice of h. Therefore, κ cannot be zero. □

Perspectives. It has been conjectured that every compact manifold admits infinitely many geometrically distinct closed geodesics. "Geometrically distinct" means that geodesics which are multiple coverings of another closed geodesic are not counted. The loop space, i.e. the space of closed curves on a manifold has a rich topology, and Morse theoretic constructions yield infinitely many critical points of the energy function. The difficulty, however, is to show that those correspond to geometrically distinct geodesics. Besides many advances, most notably by Klingenberg[167], the conjecture is not verified in many cases. Among the hardest cases are Riemannian manifolds diffeomorphic to a sphere S^n. For $n = 2$, however, in that case, the existence of infinitely many closed geodesics was shown in work of Franks[84] and Bangert[13]. For an explicit estimate for the growth of the number of closed geodesics of length $\leq \ell$, see Hingston[126] where also the proof of Franks' result is simplified.

We would also like to mention the beautiful theorem of Lyusternik and Schnirelman that any surface with a Riemannian metric diffeomorphic to S^2 contains at least three *embedded* closed geodesics (the number 3 is optimal as certain ellipsoids show). See e.g. Ballmann[10], Grayson[98], Jost[135], as well as Klingenberg[167].

Exercises for Chapter 6

1. Show that if f is a Morse function on the compact manifold X, $a < b$, and if f has no critical point p with $a \leq f(p) \leq b$, then the sublevel set $\{x \in X : f(x) \leq a\}$ is diffeomorphic to $\{x \in X : f(x) \leq b\}$.

2. Compute the Euler characteristic of a torus by constructing a suitable Morse function.

3. Show that the Euler characteristic of any compact odd-dimensional differentiable manifold is zero.

4. Show that any smooth function $f : S^n \to \mathbb{R}$ always has an even number of critical points, provided all of them are nondegenerate.

5. Prove the following

   Theorem (Reeb). *Let M be a compact differentiable manifold, and let $f \in C^3(M, \mathbb{R})$ have precisely two critical points, both of them nondegenerate. Then M is homeomorphic to the sphere S^n ($n = \dim M$).*

6. Is it possible, for any compact differentiable manifold M, to find a smooth function $f : M \to \mathbb{R}$ with only nondegenerate critical points, and with $\mu_j = b_j$ for all j (notations of Theorem 5.3.1)?

   Hint: Consider $\mathbb{RP}^3$ (cf. Chapter 1, Exercise 3 and Chapter 4, Exercise 5) and use Bochner's theorem 3.5.1, Poincaré duality (Corollary 2.2.2), and Reeb's theorem (Exercise 5).

7. State conditions for a complete, but noncompact Riemannian manifold to contain a nontrivial closed geodesic. (Note that such conditions will depend not only on the topology, but also on the metric as is already seen for surfaces of revolution in $\mathbb{R}^3$.)

8. Let M be a compact Riemannian manifold, $p, q \in M$, $p \neq q$. Show that there exist at least two geodesic arcs with endpoints p and q.

9. In (6.2.1), assume that f has two relative minima, not necessarily strict anymore. Show that again there exists another critical point x_3 of f with $f(x_3) \geq \max\{f(x_1), f(x_2)\}$. Furthermore, if $\kappa = \inf_{\gamma \in \Gamma} \max_{x \in \gamma} f(x) = f(x_1) = f(x_2)$, show that f has infinitely many critical points.

10. Prove the following statement:

    Let γ be a smooth convex closed Jordan curve in the plane $\mathbb{R}^2$. Show that there exists a straight line ℓ in $\mathbb{R}^2$ (not necessarily through the origin, i.e. $\ell = \{ax^1 + bx^2 + c = 0\}$ with fixed coefficients a, b, c) intersecting γ orthogonally in two points.

Hint: γ bounds a compact set A in $\mathbb{R}^2$ by the Jordan curve theorem. For every line ℓ in $\mathbb{R}^2$, put

$$L_A(\ell) := \text{length}\,(A \cap \ell).$$

Find a nontrivial critical point ℓ_0 for L_A (i.e. $L_A(\ell_0) > 0$) on the set of all lines by a saddle point construction. See also J. Jost, X. Li-Jost, Calculus of variations, Cambridge Univ. Press, 1998, Chapter I.3.

11. Generalize the result of Exercise 10 as follows:

    Let M be diffeomorphic to S^2, γ a smooth closed Jordan curve in M. Show that there exists a nontrivial geodesic arc in M meeting γ orthogonally at both endpoints.

    Hint: For the boundary condition, see Exercise 1 of Chapter 4.

12. If you know some algebraic topology (relative homotopy groups and a suitable extension of Lemma 6.11.3, see E. Spanier, Algebraic topology, McGraw Hill (1966)), you should be able to show the following generalization of 11:

    Let M_0 be a compact (differentiable) submanifold of the compact Riemannian manifold M. Show that there exists a nontrivial geodesic arc in M meeting M_0 orthogonally at both end points.

13. For $p > 1$ and a smooth curve $c(t)$ in M, define

$$E_p(c) := \frac{1}{p} \int \|\dot{c}\|^p dt.$$

    Define more generally a space $H^{1,p}(M)$ of curves with finite value of E_p. What are the critical points of E_p (derive the Euler-Lagrange equations)? If M is compact, does E_p satisfy the Palais-Smale condition?

Chapter 7

Harmonic Maps between Riemannian Manifolds

7.1 Definitions

We let M and N be Riemannian manifolds of dimension m and n, resp.

If we use local coordinates, the metric tensor of M will be written as

$$(\gamma_{\alpha\beta})_{\alpha,\beta=1,\dots,m},$$

and the one of N as

$$(g_{ij})_{i,j=1,\dots,n}.$$

We shall also use the following notations

$$\begin{aligned}
&(\gamma^{\alpha\beta})_{\alpha,\beta=1,\dots,m} = (\gamma_{\alpha\beta})^{-1}_{\alpha,\beta}, && \text{(inverse metric tensor)}\\
&\gamma := \det(\gamma_{\alpha\beta}),\\
&\Gamma^{\alpha}_{\beta\eta} := \frac{1}{2}\gamma^{\alpha\delta}(\gamma_{\beta\delta,\eta} + \gamma_{\eta\delta,\beta} - \gamma_{\beta\eta,\delta}), && \text{(Christoffel symbols of } M\text{)}
\end{aligned}$$

and similarly

$$g^{ij}, \Gamma^{i}_{jk}.$$

If $f : M \to N$ is a map of class C^1, we define its *energy density* as

$$e(f)(x) := \frac{1}{2}\gamma^{\alpha\beta}(x) g_{ij}(f(x)) \frac{\partial f^i(x)}{\partial x^\alpha} \frac{\partial f^j(x)}{\partial x^\beta} \tag{7.1.1}$$

in local coordinates $(x^1, \dots, x^m)$ on M, $(f^1, \dots, f^n)$ on N.

The value of $e(f)(x)$ seems to depend on the choices of local coordinates; we are now going to interpret $e(f)$ intrinsically and see that this is not so. For this purpose, we consider the differential of f,

$$df = \frac{\partial f^i}{\partial x^\alpha} dx^\alpha \otimes \frac{\partial}{\partial f^i},$$

a section of the bundle $T^*M \otimes f^{-1}TN$.

$f^{-1}TN$ is a bundle over M with metric $(g_{ij}(f(x)))$, while T^*M of course has metric $(\gamma^{\alpha\beta}(x))$, cf. (1.8.5). Likewise, we have for the Levi-Civita connections:

$$\begin{aligned} \nabla_{\frac{\partial}{\partial x^\alpha}} \frac{\partial}{\partial f^i} &= \nabla_{\frac{\partial f^j}{\partial x^\alpha}\frac{\partial}{\partial f^j}} \frac{\partial}{\partial f^i} \quad \text{by the chain rule} \\ &= \frac{\partial f^j}{\partial x^\alpha} \Gamma^k_{ij} \frac{\partial}{\partial f^k}, \end{aligned} \tag{7.1.2}$$

$\nabla_{\frac{\partial}{\partial x^\alpha}} dx^\beta = -\Gamma^\beta_{\alpha\gamma} dx^\gamma$, cf. (3.1.20), which follows from

$$dx^\beta\Big(\frac{\partial}{\partial x^\gamma}\Big) = \delta_{\beta\gamma},$$

hence,

$$\begin{aligned} 0 &= \frac{\partial}{\partial x^\alpha}\Big(dx^\beta\Big(\frac{\partial}{\partial x^\gamma}\Big)\Big) \\ &= (\nabla_{\frac{\partial}{\partial x^\alpha}} dx^\beta)\Big(\frac{\partial}{\partial x^\gamma}\Big) + dx^\beta\Big(\nabla_{\frac{\partial}{\partial x^\alpha}} \frac{\partial}{\partial x^\gamma}\Big) \\ &= (\nabla_{\frac{\partial}{\partial x^\alpha}} dx^\beta)\Big(\frac{\partial}{\partial x^\gamma}\Big) + \Gamma^\beta_{\alpha\gamma}. \end{aligned} \tag{7.1.3}$$

We shall also employ the convention that the metric of a vector bundle E over M will be denoted as

$$\langle\cdot,\cdot\rangle_E.$$

Then, with $\frac{\partial f}{\partial x^\alpha} = \frac{\partial f^i}{\partial x^\alpha}\frac{\partial}{\partial f^i}$,

$$\begin{aligned} e(f) &= \frac{1}{2}\gamma^{\alpha\beta}\Big\langle \frac{\partial f}{\partial x^\alpha}, \frac{\partial f}{\partial x^\beta}\Big\rangle_{f^{-1}TN} \\ &= \frac{1}{2}\langle df, df\rangle_{T^*M\otimes f^{-1}TN}. \end{aligned} \tag{7.1.4}$$

$\langle \frac{\partial f}{\partial x^\alpha}, \frac{\partial f}{\partial x^\beta}\rangle_{f^{-1}TN}$ is the pullback by f of the metric tensor of N, and consequently $e(f)$ is its trace (up to the factor $\frac{1}{2}$) w.r.t. the metric on T^*M. We may also express (7.1.4) as

$$e(f) = \frac{1}{2}\|df\|^2, \tag{7.1.5}$$

where the norm $\|\cdot\|$ involves the metrics on T^*M and $f^{-1}TN$.

Definition 7.1.1. The energy of a C^1-map $f : M \to N$ is

$$E(f) := \int_M e(f) dM \tag{7.1.6}$$

(with $dM = \sqrt{\gamma} dx^1 \wedge \ldots \wedge dx^m$ in local coordinates, being the volume form of M).

Of course, E generalizes the energy of a curve in N, i.e. a map from, say, S^1 to N as considered in Chapter 9 and earlier.

Another, even simpler special case is where $N = \mathbb{R}$. We then have the Dirichlet integral of a function $f : M \to \mathbb{R}$,

$$E(f) = \frac{1}{2} \int_M \gamma^{\alpha\beta}(x) \frac{\partial f}{\partial x^\alpha} \frac{\partial f}{\partial x^\beta} \sqrt{\gamma} dx^1 \ldots dx^m.$$

Our aim in this chapter is to find critical points of E. These will then be higher dimensional generalizations of closed geodesics on N. One can also consider them as nonlinear analogues of harmonic functions on M.

Lemma 7.1.1. *The Euler-Lagrange equations for E are*

$$\frac{1}{\sqrt{\gamma}} \frac{\partial}{\partial x^\alpha} (\sqrt{\gamma} \gamma^{\alpha\beta} \frac{\partial}{\partial x^\beta} f^i) + \gamma^{\alpha\beta}(x) \Gamma^i_{jk}(f(x)) \frac{\partial}{\partial x^\alpha} f^j \frac{\partial}{\partial x^\beta} f^k = 0. \tag{7.1.7}$$

Definition 7.1.2. Solutions of (7.1.7) are called *harmonic maps.*

Remark. If $M = S^1$ with its metric in standard coordinates, (7.1.7) reduces to the familiar equation for geodesics.

Proof. Let f be a smooth critical point of E. Then f is in particular continuous, and we may localize our computations in local coordinates in both domain and image. In this sense, let a smooth φ be given in such local coordinates, with compact support, and consider the variation $f + t\varphi$ for sufficiently small $|t|$, the sum being taken again in local coordinates. As f is a critical point of E,

$$\frac{d}{dt} E(f + t\varphi)|_{t=0} = 0. \tag{7.1.8}$$

So far, in fact, it sufficed to suppose f to be of class C^1. We now assume f to be of class C^2 so that the equations (7.1.7) are meaningful. (7.1.8) gives

$$\begin{aligned}
0 =& \frac{d}{dt} \frac{1}{2} \int_M \gamma^{\alpha\beta}(x) g_{ij}(f(x) + t\varphi(x)) \\
& \Big(\frac{\partial f^i}{\partial x^\alpha} + t \frac{\partial \varphi^i}{\partial x^\alpha} \Big) \Big(\frac{\partial f^j}{\partial x^\beta} + t \frac{\partial \varphi^j}{\partial x^\beta} \Big) \sqrt{\gamma} dx^1 \ldots dx^m |_{t=0} \\
=& \int_M (\gamma^{\alpha\beta}(x) g_{ij}(f(x)) \frac{\partial f^i}{\partial x^\alpha} \frac{\partial \varphi^j}{\partial x^\beta} \\
& + \frac{1}{2} \gamma^{\alpha\beta}(x) g_{ij,k}(f(x)) \frac{\partial f^i}{\partial x^\alpha} \frac{\partial f^j}{\partial x^\beta} \varphi^k \sqrt{\gamma} dx^1 \ldots dx^m
\end{aligned}$$

making use of the symmetry $g_{ij} = g_{ji}$,

$$\begin{aligned}
&= -\int_M \frac{\partial}{\partial x^\beta}\Big(\sqrt{\gamma}\gamma^{\alpha\beta}\frac{\partial f^i}{\partial x^\alpha}\Big) g_{ij}(f(x))\varphi^j dx^1 \dots dx^m \\
&\quad - \int_M \gamma^{\alpha\beta}(x)\frac{\partial f^i}{\partial x^\alpha}\frac{\partial f^k}{\partial x^\beta} g_{ij,k}(f(x))\varphi^j \sqrt{\gamma} dx^1 \dots dx^m \\
&\quad + \int_M \frac{1}{2}\gamma^{\alpha\beta}(x) g_{ij,k}(f(x))\frac{\partial f^i}{\partial x^\beta}\frac{\partial f^j}{\partial x^\alpha}\varphi^k \sqrt{\gamma} dx^1 \dots dx^m,
\end{aligned}$$

where we may integrate by parts since φ has compact support in M. We put $\eta_i = g_{ij}\varphi^j$, and thus $\varphi^j = g^{ij}\eta_i$. We then obtain

$$\begin{aligned}
0 = &-\int_M \frac{1}{\sqrt{\gamma}}\frac{\partial}{\partial x^\beta}\Big(\sqrt{\gamma}\gamma^{\alpha\beta}\frac{\partial f^i}{\partial x^\alpha}\Big)\eta_i \sqrt{\gamma} dx^1 \dots dx^m \\
&- \int_M \frac{1}{2}\gamma^{\alpha\beta} g^{\ell j}(g_{ij,k} + g_{kj,i} - g_{ik,j})\frac{\partial f^i}{\partial x^\alpha}\frac{\partial f^k}{\partial x^\beta}\eta_\ell \sqrt{\gamma} dx^1 \dots dx^m,
\end{aligned} \tag{7.1.9}$$

using the symmetry $\gamma^{\alpha\beta} = \gamma^{\beta\alpha}$ in the second integral above.

The claim then follows from Theorem A.1.5. □

Later on, the smoothness of critical points of E will be an important and often difficult issue. For the moment, however, rather than discussing this question further, we want to interpret (7.1.7) from an intrinsic point of view.

We let ψ be a vector field along f; this just means that ψ is a section of $f^{-1}TN$. In local coordinates

$$\psi = \psi^i(x)\frac{\partial}{\partial f^i}.$$

ψ induces a variation of f by

$$f_t(x) := \exp_{f(x)}(t\psi(x)). \tag{7.1.10}$$

We want to compute

$$\frac{d}{dt}E(f_t)|_{t=0}.$$

As an auxiliary computation

$$\begin{aligned}
d\psi &= \nabla^{f^{-1}TN}_{\frac{\partial}{\partial x^\alpha}}\Big(\psi^i \frac{\partial}{\partial f^i}\Big) \otimes dx^\alpha \\
&= \frac{\partial \psi^i}{\partial x^\alpha}\frac{\partial}{\partial f^i} \otimes dx^\alpha + \psi^i \Gamma^k_{ij}\frac{\partial f^j}{\partial x^\alpha}\frac{\partial}{\partial f^k} \otimes dx^\alpha,
\end{aligned} \tag{7.1.11}$$

writing $\frac{\partial}{\partial x^\alpha} = \frac{\partial f^j}{\partial x^\alpha}\frac{\partial}{\partial f^j}$ as above.

We now also have to take derivatives w.r.t. t. Here $\frac{\partial}{\partial t}$ is a vector tangent to $M \times \mathbb{R}$. The Levi-Civita connection on M and the trivial connection on $\mathbb{R}$ yield the

Levi-Civita connection on $T^*(M \times \mathbb{R}) \otimes f^{-1}TN$. Moreover, instead of $\nabla^N_{df(\frac{\partial}{\partial x^\alpha})} = \left(f^*\nabla^N\right)_{\frac{\partial}{\partial x^\alpha}}$, we shall simply write $\nabla_{\frac{\partial}{\partial x^\alpha}}$.

$$\nabla_{\frac{\partial}{\partial t}} df_t = \nabla_{\frac{\partial}{\partial t}} \frac{\partial f_t^i}{\partial x^\alpha} \frac{\partial}{\partial f^i} \otimes dx^\alpha \qquad \text{(cf. (7.1.2))} \tag{7.1.12}$$

$$= \nabla_{\frac{\partial}{\partial x^\alpha}} (\frac{\partial f_t^i}{\partial t} \frac{\partial}{\partial f^i}) \otimes dx^\alpha$$

since $\frac{\partial}{\partial t}$ and $\frac{\partial}{\partial x^\alpha}$ commute and ∇ is torsion free

$$= d\psi. \qquad \text{(cf. (7.1.11))}$$

Since the derivative of $\exp_p$ at $0 \in T_pM$ is the identity (1.4.10). Then

$$\begin{aligned}
\frac{d}{dt} E(f_t)|_{t=0} &= \frac{1}{2} \int_M \frac{d}{dt} \langle df_t, df_t \rangle dM_{|t=0} \\
&= \int_M \langle df, \nabla_{\frac{\partial}{\partial t}} df_t \rangle dM_{|t=0} \\
&= \int_M \langle df, d\psi \rangle dM \quad \text{by (7.1.12)} \\
&= \int_M \langle df, \nabla_{\frac{\partial}{\partial x^\alpha}} (\psi) \otimes dx^\alpha \rangle dM \quad \text{by (7.1.11)} \\
&= -\int_M \langle \nabla_{\frac{\partial}{\partial x^\alpha}} df, \psi \otimes dx^\alpha \rangle dM \quad \text{since } \nabla \text{ is metric} \\
&= -\int_M \langle \operatorname{trace} \nabla df, \psi \rangle dM.
\end{aligned} \tag{7.1.13}$$

Thus, intrinsically, the Euler-Lagrange equations for E are

$$\tau(f) := \operatorname{trace} \nabla df = 0. \tag{7.1.14}$$

τ is called the *tension field* of f.

Later on, we shall also be concerned with weak solutions, that is, critical points of E that are not necessarily, or not yet known to be, smooth, but are only in a Sobolev space $H^{1,2}(M,N)$. That Sobolev space will only be defined in Section 7.3, but for the moment, it suffices that f have finite energy. We can then formulate

Definition 7.1.3. f is a *critical point* of the energy integral E if

$$\frac{d}{dt} E(\exp_f t\psi)_{|t=0} = 0 \tag{7.1.15}$$

whenever ψ is a compactly supported bounded section of $f^{-1}TN$ of class $H^{1,2}$, i.e.

$$\int_M \langle d\psi, d\psi \rangle dM < \infty$$

(cf. (7.1.11) for the definition of $d\psi$; all partial derivatives are to be understood as weak derivatives).

Lemma 7.1.2. *$f \in H^{1,2}_{\mathrm{loc}}(M,N)$ is a critical point of E iff*

$$\int_M \langle df, d\psi \rangle \, dM = 0 \tag{7.1.16}$$

for all ψ as in Definition 7.1.3.

Proof. This follows from the computation of $\frac{d}{dt}E(\exp_f t\psi)$ leading to (7.1.13). $\square$

Definition 7.1.4. A solution of (7.1.16) is called *weakly harmonic.*

Corollary 7.1.1. *The weakly harmonic maps are the critical points of E.* $\square$

Lemma 7.1.3. *$f \in H^{1,2}_{\mathrm{loc}}(M,N)$ is weakly harmonic if in local coordinates*

$$\int_M \gamma^{\alpha\beta} \frac{\partial f^i}{\partial x^\alpha} \frac{\partial \eta_i}{\partial x^\beta} \sqrt{\gamma}\, dx^1 \dots dx^m = - \int_M \gamma^{\alpha\beta} \Gamma^i_{jk}(f(x)) \frac{\partial f^j}{\partial x^\alpha} \frac{\partial f^k}{\partial x^\beta} \eta_i \sqrt{\gamma}\, dx^1 \dots dx^m \tag{7.1.17}$$

for all $\eta \in H^{1,2}_0 \cap L^\infty$ (w.r.t. local coordinates).

Proof. This follows from the proof of Lemma 7.1.1 and the derivation of (7.1.13). $\square$

Remarks.

1. Under coordinate changes $g = g(f)$ in the image, η transforms into $\tilde{\eta}$ with

$$\tilde{\eta}_j = \frac{\partial f^i}{\partial g^j} \eta_i.$$

   With this transformation behaviour, (7.1.17) is invariantly defined.

2. The only variations that we shall need in the sequel are of the form

$$\psi(x) = s(f(x))\varphi(x) \tag{7.1.18}$$

   where s is a compactly supported smooth section of TN and φ is a compactly supported Lipschitz continuous real valued function. For such ψ, $f \in H^{1,2}_{\mathrm{loc}}(M,N)$ implies $\psi \in H^{1,2}$ by the chain rule.

In particular, for such variations, (7.1.16) and (7.1.17) are meaningful even if f should not be localizable in the sense of Section 7.3.

We return to the smooth case and, as an alternative to the above treatment, we now check directly that (7.1.7) and (7.1.14) are equivalent:

We let ∇ denote the Levi-Civita connection in $T^*M \otimes f^{-1}TN$ as before.

$$\begin{aligned}
\nabla_{\frac{\partial}{\partial x^\beta}}(df) &= \nabla_{\frac{\partial}{\partial x^\beta}}\Big(\frac{\partial f^i}{\partial x^\alpha} dx^\alpha \frac{\partial}{\partial f^i}\Big) \\
&= \frac{\partial}{\partial x^\beta}\Big(\frac{\partial f^i}{\partial x^\alpha}\Big) dx^\alpha \frac{\partial}{\partial f^i} + (\nabla^{T^*M}_{\frac{\partial}{\partial x^\beta}} dx^\alpha)\frac{\partial f^i}{\partial x^\alpha}\frac{\partial}{\partial f^i} + \Big(\nabla^{f^{-1}TN}_{\frac{\partial}{\partial x^\beta}}\frac{\partial}{\partial f^i}\Big)\frac{\partial f^i}{\partial x^\alpha} dx^\alpha \\
&= \frac{\partial^2 f^i}{\partial x^\alpha \partial x^\beta} dx^\alpha \frac{\partial}{\partial f^i} - \Gamma^\alpha_{\beta\gamma} dx^\gamma \frac{\partial f^i}{\partial x^\alpha}\frac{\partial}{\partial f^i} + \Gamma^k_{ij}\frac{\partial}{\partial f^k}\frac{\partial f^j}{\partial x^\beta}\frac{\partial f^i}{\partial x^\alpha} dx^\alpha . \qquad (7.1.19)
\end{aligned}$$

We then obtain for the components of $\tau(f) = \operatorname{trace} \nabla df$,

$$\tau^i(f) = \gamma^{\alpha\beta}\frac{\partial^2 f^i}{\partial x^\alpha \partial x^\beta} - \gamma^{\alpha\beta}\Gamma^\gamma_{\alpha\beta}\frac{\partial f^i}{\partial x^\gamma} + \gamma^{\alpha\beta}\Gamma^i_{jk}\frac{\partial f^j}{\partial x^\alpha}\frac{\partial f^k}{\partial x^\beta} . \qquad (7.1.20)$$

This shows that (7.1.7) and (7.1.14) are indeed equivalent, since one easily computes

$$\frac{1}{\sqrt{\gamma}}\frac{\partial}{\partial x^\alpha}\Big(\sqrt{\gamma}\gamma^{\alpha\beta}\frac{\partial}{\partial x^\beta}\Big) = \gamma^{\alpha\beta}\frac{\partial^2}{\partial x^\alpha \partial x^\beta} - \gamma^{\alpha\beta}\Gamma^\gamma_{\alpha\beta}\frac{\partial}{\partial x^\gamma} . \qquad (7.1.21)$$

The operator

$$-\Delta_M = \frac{1}{\sqrt{\gamma}}\frac{\partial}{\partial x^\alpha}\Big(\sqrt{\gamma}\gamma^{\alpha\beta}\frac{\partial}{\partial x^\beta}\Big)$$

is the negative of the Laplace-Beltrami operator of the Riemannian manifold M, cf. (2.1.13). We recall

$$\Delta_M f = -\operatorname{div} \operatorname{grad} f,$$

with

$$\operatorname{grad} f = \gamma^{\alpha\beta}\frac{\partial f}{\partial x^\alpha}\frac{\partial}{\partial x^\beta}, \qquad \text{cf. (2.1.14)}$$

$$\operatorname{div}\Big(Z^\alpha\frac{\partial}{\partial x^\alpha}\Big) = \frac{1}{\sqrt{\gamma}}\frac{\partial}{\partial x^\alpha}(\sqrt{\gamma}Z^\alpha) \quad \text{cf. (2.1.16)}$$

$f : M \to \mathbb{R}$ is a harmonic function iff

$$\Delta_M f = 0.$$

Besides closed geodesics and harmonic functions, there is another easy example of a harmonic map.

The identity map $\mathrm{id} : M \to M$ of any Riemannian manifold is harmonic. This follows for example from (7.1.20):

if $f(x) = x$, then

$$\frac{\partial f^i}{\partial x^\gamma} = \delta_{i\gamma},$$

$$\frac{\partial f^j}{\partial x^\alpha} = \delta_{j\alpha},$$

$$\frac{\partial f^k}{\partial x^\beta} = \delta_{k\beta},$$

and thus

$$\tau(f) = 0.$$

Also,

Corollary 7.1.2. *An isometric immersion $f : M \to N$ is harmonic if and only if it represents a minimal submanifold of N.*

Proof. From (3.6.24). □

Perspectives. An intrinsic calculus for operators on vector bundles and harmonic maps is developed in [73]. Some older survey articles on harmonic maps are [72, 74] and [134], the latter also containing a list of open problems with detailed references. Some more recent references will be given in the Perspectives on the subsequent sections.

7.2 Formulae for Harmonic Maps. The Bochner Technique

A.

We first want to derive the formula for the second variation of energy. For this purpose, let

$$f_{st}(x) = f(x, s, t),$$
$$f : M \times (-\varepsilon, \varepsilon) \times (-\varepsilon, \varepsilon) \to N$$

be a smooth family of maps between Riemannian manifolds of finite energy. M (but not N) may have nonempty boundary, in which case we require $f(x, s, t) = f(x, 0, 0)$ for all $x \in \partial M$ and all s, t.

We put

$$V := \frac{\partial f_{st}}{\partial s}\Big|_{s=t=0},$$
$$W := \frac{\partial f_{st}}{\partial t}\Big|_{s=t=0}.$$

We want to compute

$$\frac{\partial^2 E(f_{st})}{\partial s \partial t}\Big|_{s=t=0}.$$

To simplify notation, we usually write f instead of f_{st}, and also

$$df = \frac{\partial f}{\partial x^\alpha} dx^\alpha = \frac{\partial f^i}{\partial x^\alpha} dx^\alpha \otimes \frac{\partial}{\partial f^i},$$

a section of $T^*M \otimes f^{-1}TN$.

Then

$$\frac{\partial^2}{\partial s \partial t} E(f_{st}) = \frac{1}{2} \int_M \frac{\partial}{\partial t} \frac{\partial}{\partial s} \langle df, df \rangle \, d\mathrm{Vol}\,(M).$$

We compute the integrand: ∇ will denote the Levi-Civita connection in $f^{-1}TN$, and everything will be evaluated at $s = t = 0$:

$$\begin{aligned}
&\frac{\partial}{\partial t} \frac{\partial}{\partial s} \frac{1}{2} \Big\langle \frac{\partial f}{\partial x^\alpha} dx^\alpha, \frac{\partial f}{\partial x^\beta} dx^\beta \Big\rangle_{T^*M \otimes f^{-1}TN} \\
&\qquad = \frac{\partial}{\partial t} \Big\langle \nabla_{\frac{\partial}{\partial s}} \frac{\partial f}{\partial x^\alpha} dx^\alpha, \frac{\partial f}{\partial x^\beta} dx^\beta \Big\rangle_{T^*M \otimes f^{-1}TN} \quad \text{since } \nabla \text{ is metric} \\
&\qquad = \frac{\partial}{\partial t} \Big\langle \nabla_{\frac{\partial}{\partial x^\alpha}} \Big(\frac{\partial f}{\partial s} \Big) dx^\alpha, \frac{\partial f}{\partial x^\beta} dx^\beta \Big\rangle_{T^*M \otimes f^{-1}TN} \quad \text{since } \nabla \text{ is torsion free} \\
&\qquad = \Big\langle \nabla_{\frac{\partial}{\partial t}} \nabla_{\frac{\partial}{\partial x^\alpha}} \Big(\frac{\partial f}{\partial s} \Big) dx^\alpha, \frac{\partial f}{\partial x^\beta} dx^\beta \Big\rangle_{T^*M \otimes f^{-1}TN} \\
&\qquad\quad + \Big\langle \nabla_{\frac{\partial}{\partial x^\alpha}} \Big(\frac{\partial f}{\partial s} \Big) dx^\alpha, \nabla_{\frac{\partial}{\partial x^\beta}} \Big(\frac{\partial f}{\partial t} \Big) dx^\beta \Big\rangle_{T^*M \otimes f^{-1}TN} \\
&\qquad = \Big\langle \nabla_{\frac{\partial}{\partial x^\alpha}} \nabla_{\frac{\partial}{\partial t}} \Big(\frac{\partial f}{\partial s} \Big) dx^\alpha, \frac{\partial f}{\partial x^\beta} dx^\beta \Big\rangle_{T^*M \otimes f^{-1}TN} \\
&\qquad\quad + \Big\langle R^N \Big(\frac{\partial f}{\partial t}, \frac{\partial f}{\partial x^\alpha} \Big) \frac{\partial f}{\partial s} dx^\alpha, \frac{\partial f}{\partial x^\beta} dx^\beta \Big\rangle_{T^*M \otimes f^{-1}TN} \\
&\qquad\quad + \Big\langle \nabla_{\frac{\partial}{\partial x^\alpha}} V dx^\alpha, \nabla_{\frac{\partial}{\partial x^\beta}} W dx^\beta \Big\rangle_{T^*M \otimes f^{-1}TN}
\end{aligned}$$

by definition of the curvature tensor R^N of N

$$\begin{aligned}
&= \Big\langle \nabla \nabla_{\frac{\partial}{\partial t}} \Big(\frac{\partial f}{\partial s} \Big), df \Big\rangle_{T^*M \otimes f^{-1}TN} \\
&\quad - \operatorname{trace}_M \langle R^N(df, V)W, df \rangle_{f^{-1}TN} \\
&\quad + \operatorname{trace}_M \langle \nabla V, \nabla W \rangle_{f^{-1}TN}.
\end{aligned}$$

Thus

$$\begin{aligned}
\frac{\partial^2 E(f_{st})}{\partial s \partial t}\Big|_{s=t=0} &= \int_M \langle \nabla V, \nabla W \rangle_{f^{-1}TN} - \int_M \operatorname{trace}_M \langle R^N(df, V)W, df \rangle_{f^{-1}TN} \\
&\quad + \int_M \Big\langle \nabla \nabla_{\frac{\partial}{\partial t}} \Big(\frac{\partial f}{\partial s} \Big), df \Big\rangle_{T^*M \otimes f^{-1}TN}.
\end{aligned} \tag{7.2.1}$$

We want to examine the third term in (7.2.1) more closely.

Since ∇ is metric, integrating by parts we have

$$
\begin{aligned}
&\int_M \left\langle \nabla_{\frac{\partial}{\partial x^\alpha}} \nabla_{\frac{\partial}{\partial t}} \frac{\partial f}{\partial s} dx^\alpha, \frac{\partial f}{\partial x^\beta} dx^\beta \right\rangle_{T^*M\otimes f^{-1}TN} \\
&\qquad = -\int_M \left\langle \nabla_{\frac{\partial}{\partial t}} \frac{\partial f}{\partial s} dx^\alpha, \nabla_{\frac{\partial}{\partial x^\alpha}} \frac{\partial f}{\partial x^\beta} dx^\beta \right\rangle_{T^*M\otimes f^{-1}TN} \\
&\qquad = -\int_M \left\langle \nabla_{\frac{\partial}{\partial t}} \frac{\partial f}{\partial s}, \operatorname{trace}_M \nabla df \right\rangle_{f^{-1}TN} .
\end{aligned}
\tag{7.2.2}
$$

Theorem 7.2.1. *For a smooth family $f_{st} : M \to N$ of finite energy maps between Riemannian manifolds, with $f_{st}(x) = f_{00}(x)$ for all $x \in \partial M$ (in case $\partial M \neq \emptyset$) and all s, t, we have for the second variation of energy, with $V = \frac{\partial f}{\partial s}_{|s=0}$, $W = \frac{\partial f}{\partial t}_{|t=0}$,*

$$
\begin{aligned}
&\frac{\partial^2 E(f_{st})}{\partial s \partial t}\Big|_{s=t=0} = \int_M \langle \nabla V, \nabla W \rangle_{f^{-1}TN} \\
&- \int_M \operatorname{trace}_M \langle R^N(df, V)W, df \rangle_{f^{-1}TN} + \int_M \left\langle \nabla_{\frac{\partial}{\partial t}} \frac{\partial f}{\partial s}, \operatorname{trace}_M \nabla df \right\rangle_{f^{-1}TN} .
\end{aligned}
\tag{7.2.3}
$$

If f_{00} is harmonic, or if $\nabla_{\frac{\partial}{\partial t}} \frac{\partial f}{\partial s} \equiv 0$ for $s = t = 0$, then the second variation depends only on V and W, but not on higher derivatives of f w.r.t. s, t, and

$$
\begin{aligned}
I_f(V, W) &:= \frac{\partial^2 E(f_{st})}{\partial s \partial t} \\
&= \int_M \langle \nabla V, \nabla W \rangle_{f^{-1}TN} - \int_M \operatorname{trace}_M \langle R^N(df, V)W, df \rangle_{f^{-1}TN} .
\end{aligned}
\tag{7.2.4}
$$

Proof. (7.2.3) follows from (7.2.1), (7.2.2). (7.2.4) holds if either $\nabla_{\frac{\partial}{\partial t}} \frac{\partial f}{\partial s} \equiv 0$ or $\operatorname{trace}_M \nabla df \equiv 0$, and the latter is the harmonic map equation (cf. (7.1.14)). □

We look at the special case where we only have one parameter:

$$
\begin{aligned}
f(x, t) &= f_t(x), \\
f : M \times (-\varepsilon, \varepsilon) &\to N, \\
W &:= \frac{\partial f}{\partial t}\Big|_{t=0} .
\end{aligned}
$$

Then

Corollary 7.2.1. *Under the assumptions of Theorem 7.2.1,*

$$
\begin{aligned}
I_f(W, W) &= \frac{\partial^2}{\partial t^2} E(f_t)_{|t=0} \\
&= \int_M \|\nabla W\|^2_{f^{-1}TN} - \int_M \operatorname{trace}_M \langle R^N(df, W)W, df \rangle_{f^{-1}TN} ,
\end{aligned}
\tag{7.2.5}
$$

if f is harmonic or if $f(x, \cdot)$ is geodesic for every x.

Proof. If $f(x, \cdot)$ is geodesic for every x,

$$\nabla_{\frac{\partial}{\partial t}} \frac{\partial f}{\partial t} \equiv 0.$$

All assertions follow from Theorem 7.2.1. □

Remark. For geodesics, the second variation of energy was already derived in Theorem 4.1.1.

Corollary 7.2.2. *Under the assumptions of Theorem 7.2.1, if N has nonpositive sectional curvature, then a harmonic map is a stable critical point of the energy functional in the sense that the second variation of energy is nonnegative.*

Proof. If N has nonpositive sectional curvature,

$$\langle R^N(df(\Phi), W(x))W(x), df(\Phi)\rangle \le 0$$

for every $x \in M$, $\Phi \in T_xM$, and every section of W of $f^{-1}TN$, and the claim follows from (7.2.5). □

B.

We next want to calculate

$$\Delta e(f) = \Delta \frac{1}{2}\gamma^{\alpha\beta}(x) g_{ij}(f(x)) f^i_{x^\alpha} f^j_{x^\beta}$$

for a harmonic map $f : M \to N$. (Here $f^i_{x^\alpha} := \frac{\partial f^i}{\partial x^\alpha}$.) The computation may be carried out in the same manner as at the end of §3.3 and in §3.5. It is somewhat easier, however, to perform it in local coordinates.

In order to simplify the computation, we introduce normal coordinates at x and at $f(x)$. Thus

$$\gamma_{\alpha\beta} = \delta_{\alpha\beta}, \quad g_{ij}(f(x)) = \delta_{ij}, \tag{7.2.6}$$

and

$$\gamma_{\alpha\beta,\delta}(x) = 0, \quad g_{ij,k}(f(x)) = 0 \tag{7.2.7}$$

for all indices. Therefore, in our computations, we only have to take second derivatives of the metric into account; these will yield curvature terms.

We rewrite the harmonic map equation (7.1.7) as

$$0 = \gamma^{\alpha\beta}(x) f^i_{x^\alpha x^\beta} - \gamma^{\alpha\beta}(x)\Gamma^\eta_{\alpha\beta}(x) f^i_{x^\eta} + \gamma^{\alpha\beta}(x)\Gamma^i_{jk}(f(x)) f^j_{x^\alpha} f^k_{x^\beta}\,. \tag{7.2.8}$$

(Here, the Christoffel symbols of M have Greek indices, those of N Latin ones.)

We differentiate (7.2.8) at x w.r.t. x^ε and obtain, recalling (7.2.6), (7.2.7),

$$\begin{aligned} f^i_{x^\alpha x^\alpha x^\varepsilon} &= \frac{1}{2}(\gamma_{\alpha\eta,\alpha\varepsilon} + \gamma_{\alpha\eta,\alpha\varepsilon} - \gamma_{\alpha\alpha,\eta\varepsilon}) f^i_{x^\eta} \\ &\quad - \frac{1}{2}(g_{ki,\ell m} + g_{\ell i,km} - g_{k\ell,im}) f^m_{x^\varepsilon} f^k_{x^\alpha} f^\ell_{x^\alpha} \,. \end{aligned} \tag{7.2.9}$$

Moreover, by (7.2.6), (7.2.7)

$$\gamma^{\alpha\beta}{}_{,\varepsilon\varepsilon} = -\gamma_{\alpha\beta,\varepsilon\varepsilon} \tag{7.2.10}$$

and from the chain rule

$$-\Delta g_{ij}(f(x)) = g_{ij,k\ell} f^k_{x^\varepsilon} f^\ell_{x^\varepsilon} \,. \tag{7.2.11}$$

(7.2.9) – (7.2.11) yield

$$\begin{aligned} &-\Delta\Big(\frac{1}{2}\gamma^{\alpha\beta}(x) g_{ij}(f(x)) f^i_{x^\alpha} f^j_{x^\beta}\Big) \\ &\qquad = f^i_{x^\alpha x^\varepsilon} f^i_{x^\alpha x^\varepsilon} - \frac{1}{2}(\gamma_{\alpha\beta,\varepsilon\varepsilon} + \gamma_{\varepsilon\varepsilon,\alpha\beta} - \gamma_{\varepsilon\alpha,\varepsilon\beta} - \gamma_{\varepsilon\beta,\varepsilon\alpha}) f^i_{x^\alpha} f^i_{x^\beta} \\ &\qquad\qquad + \frac{1}{2}(g_{ij,k\ell} + g_{k\ell,ij} - g_{ik,j\ell} - g_{j\ell,ik}) f^i_{x^\alpha} f^j_{x^\alpha} f^k_{x^\varepsilon} f^\ell_{x^\varepsilon} \\ &\qquad = f^i_{x^\alpha x^\varepsilon} f^i_{x^\alpha x^\varepsilon} + R^M_{\alpha\beta} f^i_{x^\alpha} f^i_{x^\beta} - R^N_{i\ell jk} f^i_{x^\alpha} f^j_{x^\alpha} f^k_{x^\varepsilon} f^\ell_{x^\varepsilon} \end{aligned} \tag{7.2.12}$$

(cf. (3.3.15)), where

$$R^M_{\alpha\beta} = \gamma^{\delta\varepsilon} R^M_{\alpha\delta\beta\varepsilon} = R^M_{\alpha\varepsilon\beta\varepsilon}$$

is the Ricci tensor of M, and $R^N_{i\ell jk}$ is the curvature tensor of N.

In invariant notation, if $e_1, \dots, e_m$ is an orthonormal basis of T_xM, (7.2.12) becomes

$$\begin{aligned} -\Delta e(f)(x) &= \|\nabla df\|^2 + \langle df(\operatorname{Ric}^M(e_\alpha)), df(e_\alpha)\rangle_{f^{-1}TN} \\ &\quad - \langle R^N(df(e_\alpha), df(e_\beta)) df(e_\beta), df(e_\alpha)\rangle_{f^{-1}TN} \,. \end{aligned} \tag{7.2.13}$$

Corollary 7.2.3. *Let M be a compact Riemannian manifold with nonnegative Ricci curvature, N a Riemannian manifold with nonpositive sectional curvature.*

Let $f : M \to N$ be harmonic.

Then f is totally geodesic[1] *(i.e. $\nabla df \equiv 0$) and $e(f) \equiv \text{const}$. If the Ricci curvature of M is (nonnegative, but) not identically zero, then f is constant.*

If the sectional curvature of N is negative, then f is either constant or maps M onto a closed geodesic.

Proof. By Stokes' theorem,

$$\int_M \Delta e(f) = 0.$$

[1] See Lemma 7.2.1 below.

Therefore, the integral of the right hand side of (7.2.13) also vanishes. Since the integrand is the sum of three terms which are all everywhere nonnegative by assumption, all three terms have to vanish identically.

We first conclude

$$\|\nabla df\| \equiv 0, \tag{7.2.14}$$

hence $\nabla df \equiv 0$ so that f is totally geodesic.

Secondly,

$$\Delta e(f) \equiv 0,$$

and since harmonic functions on compact Riemannian manifolds are constant (cf. Corollary 2.1.2),

$$e(f) \equiv \text{const.} \tag{7.2.15}$$

If for some $x \in M$,

$$R^M_{\alpha\beta}(x) \text{ is positive definite,}$$

then

$$R^M_{\alpha\beta}(x) f^i_{x^\alpha} f^i_{x^\beta} = 0$$

implies

$$df(x) = 0,$$

hence $e(f)(x) = 0$, hence

$$e(f) \equiv 0$$

by (7.2.15), and f is constant.

If N has negative sectional curvature, then

$$\langle R^N(df(e_\alpha), df(e_\beta))df(e_\beta), df(e_\alpha)\rangle \equiv 0$$

implies that $df(e_\alpha)$ and $df(e_\beta)$ are linearly dependent everywhere. Therefore, $f(M)$ is at most one-dimensional. If the dimension is zero, f is constant, and if the dimension is one, $f(M)$ is a closed geodesic because f is totally geodesic and M is compact. (See Lemma 7.2.1 below.) □

Remark. The method of proof of Corollary 7.2.3 is another instance of the so-called Bochner method which is very important in Riemannian and complex geometry. The prototype of the technique was already given in §3.5.

Lemma 7.2.1. *A smooth map $f : M \to N$ between Riemannian manifolds is totally geodesic iff f maps every geodesic of M onto a geodesic of N.*

Proof. Let $\gamma(t)$ be a geodesic in M. Then

$$\begin{aligned}
\nabla^{TN}_{\frac{\partial}{\partial t}} \frac{\partial}{\partial t}(f \circ \gamma(t)) &= \nabla^{TN}_{\frac{\partial}{\partial t}} \Big(df\Big(\frac{\partial \gamma}{\partial t}\Big)\Big) \\
&= (\nabla^{TN}_{\frac{\partial}{\partial t}} df \circ \gamma)\Big(\frac{\partial \gamma}{\partial t}\Big) + df\Big(\nabla^{TM}_{\frac{\partial}{\partial t}} \frac{\partial \gamma}{\partial t}\Big) \\
&= \nabla df\Big(\frac{\partial \gamma}{\partial t}, \frac{\partial \gamma}{\partial t}\Big),
\end{aligned}$$

since γ is geodesic. Thus $(f \circ \gamma)(t)$ is geodesic iff

$$\nabla df\Big(\frac{\partial \gamma}{\partial t}, \frac{\partial \gamma}{\partial t}\Big) = 0.$$

□

C.

We finally want to derive and exploit a chain rule. If $f : M \to N$ and $h : N \to Q$ are smooth maps between Riemannian manifolds,

$$\begin{aligned}
\tau(h \circ f) &= \operatorname{trace} \nabla d(h \circ f) \\
&= \gamma^{\alpha\beta} \nabla_{\frac{\partial}{\partial x^\alpha}} \frac{\partial}{\partial x^\beta}(h \circ f) \\
&= \gamma^{\alpha\beta} \nabla_{\frac{\partial}{\partial x^\alpha}} \Big(\frac{\partial h}{\partial f^i} \frac{\partial f^i}{\partial x^\beta}\Big) \\
&= \gamma^{\alpha\beta} \Big(\nabla_{\frac{\partial}{\partial f^j}} \frac{\partial h}{\partial f^i}\Big) \frac{\partial f^j}{\partial x^\alpha} \frac{\partial f^i}{\partial x^\beta} + \gamma^{\alpha\beta} \frac{\partial h}{\partial f^i} \nabla_{\frac{\partial}{\partial x^\alpha}} \frac{\partial f^i}{\partial x^\beta} \\
&= \gamma^{\alpha\beta} \nabla dh\Big(\frac{\partial f}{\partial x^\alpha}, \frac{\partial f}{\partial x^\beta}\Big) + (dh)(\tau(f)),
\end{aligned}$$

where ∇dh is the Hessian of h (see Definition 3.3.5), and $\tau(f)$ is the tension field of f.

Thus

Lemma 7.2.2. *For smooth maps $f : M \to N, h : N \to Q$ between Riemannian manifolds, the following chain rule holds*

$$\tau(h \circ f) = \gamma^{\alpha\beta} \nabla dh\Big(\frac{\partial f}{\partial x^\alpha}, \frac{\partial f}{\partial x^\beta}\Big) + (dh) \circ (\tau(f)). \tag{7.2.16}$$

In particular, if f is harmonic

$$\tau(h \circ f) = \gamma^{\alpha\beta} \nabla dh\Big(\frac{\partial f}{\partial x^\alpha}, \frac{\partial f}{\partial x^\beta}\Big). \tag{7.2.17}$$

□

Remark. If $Q = \mathbb{R}$, of course $\tau = -\Delta$, where Δ is the Laplace-Beltrami operator. This, in fact, is the case that we shall use in the sequel. Therefore, it might be useful to observe that in this case we can also use the Euclidean chain rule. Using Riemann normal coordinates on M and arbitrary local coordinates on N, we then have

$$-\Delta(h \circ f) = \frac{\partial^2 h}{\partial f^i \partial f^j} \frac{\partial f^i}{\partial x^\alpha} \frac{\partial f^j}{\partial x^\alpha} + \frac{\partial h}{\partial f^k} \frac{\partial^2 f^k}{(\partial x^\alpha)^2} \tag{7.2.18}$$

which, of course, is equivalent to the Riemannian chain rule of Lemma 7.2.2 which here becomes, using (3.3.47),

$$-\Delta(h \circ f) = \left(\frac{\partial^2 h}{\partial f^i \partial f^j} - \frac{\partial h}{\partial f^k} \Gamma^k_{ij} \right) \frac{\partial f^i}{\partial x^\alpha} \frac{\partial f^j}{\partial x^\alpha} + \frac{\partial h}{\partial f^k} \left(\frac{\partial^2 f^k}{(\partial x^\alpha)^2} + \Gamma^k_{ij} \frac{\partial f^i}{\partial x^\alpha} \frac{\partial f^j}{\partial x^\alpha} \right). \tag{7.2.19}$$

Definition 7.2.1. $g : N \to \mathbb{R}$ (N a Riemannian manifold) is called subharmonic if

$$-\Delta g \geq 0.$$

Corollary 7.2.4. *If $f : M \to N$ is harmonic, and $h : N \to \mathbb{R}$ is convex, then $h \circ f$ is subharmonic, i.e.*

$$-\Delta(h \circ f) \geq 0.$$

Conversely, if $f : M \to N$ is a smooth map such that for all open $V \subset N$ and convex $h : V \to \mathbb{R}$, with $U := f^{-1}(V)$,

$$h \circ f \quad \textit{is subharmonic},$$

then f is harmonic.

Proof. (7.2.17) implies the first part. For the second part, if f is not harmonic, we may find some $x_0 \in M$ with

$$\tau(f)(x_0) \neq 0.$$

We then need to find a convex function h on some neighborhood V of $f(x_0)$ for which

$$-\Delta(h \circ f)(x_0) < 0.$$

If N were Euclidean, we could simply take a linear function h, i.e. $\nabla dh \equiv 0$, with $(\operatorname{grad} h)(f(x_0)) = -\tau(f)(x_0)$.

We then have

$$\begin{aligned} -\Delta(h \circ f)(x_0) &= dh \circ \tau(f)(x_0) \\ &= \langle (\operatorname{grad} h)(f(x_0)), \tau(f)(x_0) \rangle \\ &= -\|\tau(f)(x_0)\|^2 < 0. \end{aligned}$$

In the Riemannian case, in general, we may not find local functions with $\nabla dh \equiv 0$, but if we consider sufficiently small neighborhoods V, we may find such functions h_0 for which $\|\nabla dh_0\|$ is arbitrarily small while we still have a prescribed gradient $(\operatorname{grad} h_0)(f(x_0)) = -\tau(f)(x_0)$. This follows from the definition of the Hessian ∇dh_0, see (3.3.47), together with the fact that in Riemannian coordinates centered at $f(x_0)$, $\Gamma^i_{jk}(f(x_0)) = 0$, see (1.4.12), and so Γ^i_{jk} can be made arbitrarily small in a sufficiently small neighborhood V of $f(x_0)$.

Still, h_0 is not convex, but since $\|\nabla dh_0\|$ is small, it can be made convex in a small neighborhood V of x_0 by adding a small multiple of $d^2(f(x_0), \cdot)$, the squared distance function from $f(x_0)$, using (4.6.6). Since that multiple is small, say ε, the new function

$$h = h_0 + \varepsilon d^2(f(x_0), \cdot)$$

can still be assumed to satisfy

$$dh \circ \tau(f)(x_0) < \gamma^{\alpha\beta} \nabla dh\Big(\frac{\partial f}{\partial x^\alpha}, \frac{\partial f}{\partial x^\beta}\Big)$$

i.e.

$$-\Delta h \circ f(x_0) < 0.$$

This completes the proof. □

Corollary 7.2.5. *If $f : M \to N$ is harmonic and if N has nonpositive sectional curvature, and is simply connected and complete, then for any $p \in N$,*

$$-\Delta d^2(f(x), p) \geq 2\|df(x)\|^2.$$

Proof. (7.2.17) and Lemma 4.8.2. □

Corollary 7.2.6. *Let M be a compact Riemannian manifold, N a Riemannian manifold, $f : M \to N$ harmonic.*

If there exists a strictly convex function h on $f(M)$, then f is constant.

Proof. By Corollary 7.2.3, $h \circ f$ is subharmonic. The following Lemma shows that $h \circ f$ then is constant. Since h is strictly convex, (7.2.17) implies $f \equiv$ const. □

Lemma 7.2.3. *Let M be a compact Riemannian manifold. Then any subharmonic function φ is constant.*

Proof. By Stokes' theorem

$$\int_M \Delta\varphi = 0,$$

so that a subharmonic function is harmonic, hence constant by Corollary 2.1.2. □

Corollary 7.2.7. *If $f : M \to N$ is harmonic and $h : N \to Q$ is totally geodesic, then $h \circ f$ is harmonic.*

Proof. (7.2.17). $\square$

Perspectives. For more special domains, other Bochner type formulae for harmonic maps have been found. Here, we only want to quote two such formulae.

Siu[234] derived the following formula that is actually valid for any smooth, not necessarily harmonic map between Kähler manifolds $f : M \to N$

$$\partial\bar{\partial}(g_{i\bar{j}}\bar{\partial}f^i \wedge \partial f^{\bar{j}}) = R_{i\bar{j}k\bar{\ell}}\bar{\partial}f^i \wedge \partial f^{\bar{j}} \wedge \partial f^k \wedge \overline{\partial f^\ell} - g_{i\bar{j}}D'\bar{\partial}f^i \wedge D''\partial f^{\bar{j}}.$$

Here, $(g_{i\bar{j}})$ is the Kähler metric of N in local holomorphic coordinates $(f^1, \dots, f^n)$, $R_{i\bar{j}k\bar{\ell}}$ its curvature tensor, Γ^i_{jk} its Christoffel symbols,

$$\begin{aligned} D'\bar{\partial}f^i &= \partial\bar{\partial}f^i + \Gamma^i_{jk}\partial f^j \wedge \bar{\partial}f^k, \\ D''\partial f^{\bar{j}} &= \bar{\partial}\partial f^{\bar{j}} + \Gamma^{\bar{j}}_{\bar{\ell}\bar{k}}\bar{\partial}f^{\bar{\ell}} \wedge \partial f^{\bar{k}} \end{aligned}$$

the covariant derivatives.

The assumption that f is harmonic is needed if one wants to know the sign of the second term on the right hand side. Namely, in that case

$$g_{i\bar{j}}D'\bar{\partial}f^i \wedge D''\partial f^{\bar{j}} \wedge \omega^{n-2} = q\omega^n$$

for some nonpositive function q on M, where ω is the Kähler form of M. Furthermore, if the curvature tensor is "strongly seminegative", then the first term on the right hand side is a nonnegative multiple of ω^n, and integration by parts then gives as in the proof of Corollary 7.2.3 that under these conditions, a harmonic map f satisfies

$$D'\bar{\partial}f = D''\partial\bar{f} = 0.$$

This means that f is pluriharmonic.

If the curvature of N is even "strongly negative" and if the real rank of df is at least 3 at some point, then Siu showed that f has to be holomorphic or antiholomorphic. If N is a Riemann surface of negative curvature then the real dimension of the image is 2, hence $\mathrm{Rank}_{\mathbb{R}} df \le 2$ and Siu's result does not apply. Nevertheless, in that case, Jost and Yau[154] showed that the level sets of f still define a holomorphic foliation of M although f itself need not be holomorphic.

We now want to derive a Bochner type identity for harmonic maps from Einstein manifolds, due to Jost and Yau[156]. In order to simplify the formula and its derivation, we always use normal coordinates at the point under consideration and denote (covariant) derivatives by subscripts, e.g.

$$\begin{aligned} f_\alpha &:= \frac{\partial}{\partial x^\alpha} f, \\ \nabla_\beta &:= \nabla_{\frac{\partial}{\partial x^\beta}}. \end{aligned}$$

The formula then is

Theorem. *Let $f : M \to N$ be a harmonic map between Riemannian manifolds, where M is* *compact and Einstein. Then for any* $\lambda \in \mathbb{R}$,

$$\begin{aligned}\lambda \int_M \langle f_{\alpha\beta}, f_{\alpha\beta}\rangle + 2\int_M R^M_{\alpha\beta\gamma\delta}\langle f_{\alpha\delta}, f_{\beta\gamma}\rangle = \\ -\lambda \int_M R^M_{\alpha\beta}\langle f_\alpha, f_\beta\rangle - \int_M R^M_{\alpha\beta\gamma\delta}R^M_{\eta\beta\gamma\delta}\langle f_\alpha, f_\eta\rangle \\ +\lambda\int_M \langle R^N(f_\alpha, f_\beta)f_\beta, f_\alpha\rangle + \int_M R^M_{\alpha\beta\gamma\delta}\langle R^N(f_\gamma, f_\delta)f_\beta, f_\alpha\rangle\,.\end{aligned}$$

Let us give the

Proof. We start with (7.2.13), i.e.

$$\frac{-1}{2}\Delta\langle f_\alpha, f_\alpha\rangle = \{\langle f_{\alpha\beta}, f_{\alpha\beta}\rangle + R^M_{\alpha\beta}\langle f_\alpha, f_\beta\rangle - \langle R^N(f_\alpha, f_\beta)f_\beta, f_\alpha\rangle\}\,. \tag{7.2P.1}$$

We compute

$$(\nabla_\gamma\nabla_\delta - \nabla_\delta\nabla_\gamma)(f_\alpha dx^\alpha) = -R^M_{\beta\alpha\gamma\delta}f_\beta dx^\alpha + R^N(f_\gamma, f_\delta)f_\alpha dx^\alpha\,. \tag{7.2P.2}$$

From (7.2P.2),

$$\begin{aligned}&\langle(\nabla_\gamma\nabla_\delta - \nabla_\delta\nabla_\gamma)f_\alpha dx^\alpha, (\nabla_\gamma\nabla_\delta - \nabla_\delta\nabla_\gamma)f_\beta dx^\beta\rangle = \\ &\quad R^M_{\beta\alpha\gamma\delta}R^M_{\eta\alpha\gamma\delta}\langle f_\beta, f_\eta\rangle + \langle R^N(f_\gamma, f_\delta)f_\alpha, R^N(f_\gamma, f_\delta)f_\alpha\rangle \\ &\quad - 2R^M_{\alpha\beta\gamma\delta}\langle R^N(f_\gamma, f_\delta)f_\beta, f_\alpha\rangle\,.\end{aligned} \tag{7.2P.3}$$

Denoting the L^2-product on $T^*M \otimes f^*TN$ by $(\cdot,\cdot)$, we get

$$\begin{aligned}&((\nabla_\gamma\nabla_\delta - \nabla_\delta\nabla_\gamma)f_\alpha dx^\alpha, (\nabla_\gamma\nabla_\delta - \nabla_\delta\nabla_\gamma)f_\beta dx^\beta) \\ &\quad = (-R^M_{\beta\alpha\gamma\delta}f_\beta dx^\alpha, (\nabla_\gamma\nabla_\delta - \nabla_\delta\nabla_\gamma)f_\eta dx^\eta) \\ &\quad\quad + (R^N(f_\gamma, f_\delta)f_\alpha dx^\alpha, (\nabla_\gamma\nabla_\delta - \nabla_\delta\nabla_\gamma)f_\beta dx^\beta) \\ &\quad = 2(-R^M_{\beta\alpha\gamma\delta}f^\beta dx^\alpha, \nabla_\gamma\nabla_\delta f_\eta dx^\eta) \\ &\quad\quad + \int_M \langle R^N(f_\gamma, f_\delta)f_\alpha, R^N(f_\gamma, f_\delta)f_\alpha\rangle - \int_M R^M_{\alpha\beta\gamma\delta}\langle R^N(f_\gamma, f_\delta)f_\beta, f_\alpha\rangle\end{aligned}$$

integrating the first term by parts, we get

$$\begin{aligned}&= 2\int_M \left\langle \frac{\partial}{\partial\gamma}(R^M_{\beta\alpha\gamma\delta}f_\beta), f_{\alpha\delta}\right\rangle \\ &\quad + \int_M \langle R^N(f_\gamma, f_\delta)f_\alpha, R^N(f_\gamma, f_\delta)f_\alpha\rangle - \int_M R^M_{\alpha\beta\gamma\delta}\langle R^N(f_\gamma, f_\delta)f_\beta, f_\alpha\rangle\,.\end{aligned}$$

Now

$$\left\langle \frac{\partial}{\partial\gamma}(R^M_{\beta\alpha\gamma\delta}f_\beta), f_{\alpha\delta}\right\rangle = R^M_{\beta\alpha\gamma\delta}\langle f_{\beta\gamma}, f_{\alpha\delta}\rangle + R^M_{\beta\alpha\gamma\delta,\gamma}\langle f_\beta, f_{\alpha\delta}\rangle\,.$$

The second term vanishes for any Einstein metric since

$$R^M_{\beta\alpha\gamma\delta,\gamma} = (R^M_{\delta\gamma\alpha\gamma,\beta} - R^M_{\delta\gamma\beta\gamma,\alpha})$$

by the Bianchi identity, and this vanishes if the Ricci tensor is parallel.

We obtain for an Einstein metric on M,

$$((\nabla_\gamma\nabla_\delta - \nabla_\delta\nabla_\gamma)f_\alpha dx^\alpha, (\nabla_\gamma\nabla_\delta - \nabla_\delta\nabla_\gamma)f_\beta dx^\beta) = -2\int_M R^M_{\alpha\beta\gamma\delta}\langle f_{\alpha\delta}, f_{\beta\gamma}\rangle$$
$$+\int_M \langle R^N(f_\gamma, f_\delta)f_\alpha, R^N(f_\gamma, f_\delta)f_\alpha\rangle - \int_M R^M_{\alpha\beta\gamma\delta}\langle R^N(f_\gamma, f_\delta)f_\beta, f_\alpha\rangle\,. \quad (7.2\text{P}.4)$$

From (7.2P.1), (7.2P.3), (7.2P.4) we get the desired formula. For the special case $N = \mathbb{R}$, the formula is simpler and due to Matsushima[185].

For an application of the formula, see the Perspectives on §7.7.

A general discussion of identities for harmonic maps and applications can be found in Xin[268].

The characterization of harmonic mappings given in Corollary 7.2.4, i.e. that a (smooth) map between Riemannian manifolds is harmonic if and only if locally the composition with all convex functions is subharmonic has been observed by Ishihara[130].

It might be tempting (and it has been proposed) to use that characterization for an axiomatic approach to harmonic maps. That, however, would loose the deeper aspects of harmonic maps based on their variational properties. It will be explained later in this chapter that a rather general and satisfactory theory can be developed for harmonic mappings with values in Riemannian manifolds of nonpositive sectional curvature based on an abstract variational approach. By way of contrast, a characterization analogous to Corollary 7.2.4 can also be obtained for solutions of other nonlinear elliptic systems for maps between manifolds that need not have a variational origin. For example, Jost and Yau[155] considered the system

$$\gamma^{\alpha\bar\beta}\Big(\frac{\partial^2 f^i}{\partial z^\alpha \partial \bar z^\beta} + \Gamma^i_{jk}\frac{\partial f^j}{\partial z^\alpha}\frac{\partial f^k}{\partial \bar z^\beta}\Big) = 0$$

for maps $f : X \to N$, where N is a Riemannian manifold as before, but X is a Hermitian manifold with metric $(\gamma_{\alpha\bar\beta})_{\alpha,\beta=1,\dots,\dim_{\mathbb{C}} M}$. The preceding system is equivalent to the harmonic map system if the metric $(\gamma_{\alpha\bar\beta})$ is a Kähler metric, but not for a general Hermitian metric, and in fact, in the general case, it need not arise from a variational integral. Analogous to Corollary 7.2.4, solutions can be characterized by the property that local compositions with convex functions

$$h : V(\subset N) \to \mathbb{R}$$

satisfy

$$-\gamma^{\alpha\bar\beta}\frac{\partial^2(h\circ f)}{\partial z^\alpha \partial z^{\bar\beta}} \ge 0.$$

However, as examples show, one does not always get the existence of solutions of the new system in a prescribed homotopy class of maps $f : X \to N$, N of nonpositive sectional curvature, as in the existence theory for harmonic maps. The reason for the failure of the existence theory is the lack of variational structure. We refer to [155] for details.

7.3 Definition and Lower Semicontinuity of the Energy Integral

For the analysis of harmonic maps, it is necessary to consider classes of maps more general than C^1. A natural space of maps is $L^2(M,N)$. One then needs to define the energy integral and derive conditions for a map to be a critical point of that integral.

The idea of defining the energy functional is quite simple and may be described as follows:

We let, for $h > 0$,

$$\sigma_h : \mathbb{R}^+ \to \mathbb{R}$$

be some nonnegative function with $\sigma_h(s) = 0$ for $s \geq h$ and

$$\int\limits_{B(0,h)} \sigma_h(|x|)\, dx = 1,$$

where $B(0,h)$ is a ball of radius h in $\mathbb{R}^m$ (m will be the dimension of our domain M in the sequel). For $x, y \in M$, we put

$$\eta_h(x,y) := \sigma_h(d(x,y)). \tag{7.3.1}$$

The typical example we have in mind is

$$\sigma_h(s) = \begin{cases} \frac{1}{\omega_m h^m} & \text{for } 0 \leq s < h \quad (\omega_m = \text{volume of the unit ball in } \mathbb{R}^m), \\ 0 & \text{for } s < h, \end{cases} \tag{7.3.2}$$

and so, $\eta_h(x,\cdot)$ is a multiple of the characteristic function of the ball $B(x,h)$, for every x. That multiple is chosen so that the integral of $\eta_h(x,\cdot)$ w.r.t. the Euclidean volume form dy on $B(x,h)$ is 1, i.e. the one induced from the Euclidean volume form on T_xM via the exponential map $\exp_x : T_xM \to M$. We note that by Theorem 1.4.4, the difference between the Euclidean and Riemannian volume forms is of order $\mathrm{O}(h^{m+2})$. The advantage of the Euclidean volume form is that the normalization does not depend on x so that η_h becomes symmetric in x and y.

For a map $f \in L^2(M,N)$ between Riemannian manifolds M and N, we then define

$$E_h(f) := \int\limits_M \int\limits_{B(x,h)} \eta_h(x,y) \frac{d^2(f(x),f(y))}{h^2}\, d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x), \tag{7.3.3}$$

where $d\mathrm{Vol}$ is the Riemannian volume form on M.

In order to understand the geometric meaning of the functionals E_h, we observe

Lemma 7.3.1. *$f : M \to N$ minimizes E_h iff $f(x)$ is a center of mass for the measure $f_\#(\eta_h(x,y)\, d\mathrm{Vol}\,(y))$ for almost all $x \in M$, i.e. if $f(x)$ minimizes*

$$F(p) = \int\limits_{B(x,h)} \eta_h(x,y) d^2(p, f(y))\, d\mathrm{Vol}\,(y).$$

Proof. If $f(x)$ did not minimize $F(p)$, then

$$\int\limits_{B(x,h)} \eta_h(x,y) d^2(f(x), f(y)) d\mathrm{Vol}\,(y)$$

could be decreased by replacing $f(x)$ by some minimizer p. Since $\eta_h(x,y)$ is symmetric, that would also decrease $E_h(f)$ if happening on a set of positive measure. $\square$

It is also instructive to consider the following computation that leads to a proof of Lemma 7.3.1 in the smooth case. We consider variations

$$f_t(x) = f(x) + t\varphi(x)$$

of f. If f minimizes E_h, then

$$\begin{aligned}
0 &= \frac{d}{dt} E_h(f_t)_{|t=0} \\
&= \frac{1}{h^2} \frac{d}{dt} \iint \eta_h(x,y) d^2(f_t(x), f_t(y))\, d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x) \\
&= \frac{1}{h^2} \iint \eta_h(x,y) \big\{ \nabla_1 d^2(f(x), f(y))(\varphi(x)) + \\
&\qquad\qquad \nabla_2 d^2(f(x), f(y))(\varphi(y)) \big\}\, d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x) \\
&= \frac{2}{h^2} \iint \eta_h(x,y) \nabla_1 d^2(f(x), f(y)) \varphi(x)\, d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x)
\end{aligned}$$

because of the symmetry of η_h

$$= \frac{2}{h^2} \iint \eta_h(x,y) \exp^{-1}_{\varphi(x)} f(y) \varphi(x)\, d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x)\,.$$

Since this has to hold for all smooth φ with compact support,

$$\int \eta_h(x,y) \exp^{-1}_{\varphi(x)} f(y)\, d\mathrm{Vol}\,(y) = 0$$

for all x. Thus $f(x)$ is the center of mass of $f_{\#}(\eta_h(x,y)\, d\mathrm{Vol}\,(y))$.

We now consider the functionals E_ε for $h = \varepsilon$ with the kernel η_ε defined by (7.3.1), (7.3.2), and we let $\varepsilon \to 0$ and define the energy E as the limit of the functionals E_ε. The functionals E_ε increase towards E, and it is not excluded that $E(f)$ takes the value ∞ for some $f \in L^2(M,N)$. We shall see that E coincides with the usual energy functional for those mappings for which the latter is defined. Also, the functionals E_ε are continuous w.r.t. L^2-convergence, and the limit of an increasing sequence of continuous functions is lower semicontinuous. We shall thus obtain the lower semicontinuity of the energy w.r.t. L^2-convergence.

Actually, the described monotonicity of the sequence E_ε as $\varepsilon \to 0$ only holds up to an error term that comes from the geometry of M. It is not hard to control this error term sufficiently well so that the desired conclusion about E can still be reached.

Lemma 7.3.2. *$E_\varepsilon(f)$ is continuous on $L^2(M,N)$, i.e. if $(f_\nu)_{\nu\in\mathbb{N}}$ converges to f in $L^2(M,N)$, then*

$$E_\varepsilon(f) = \lim_{\nu\to\infty} E_\varepsilon(f_\nu).$$

Proof. Elementary. □

We estimate for $0<\lambda<1$,

$$\begin{aligned}
E_\varepsilon(f) &= \frac{1}{\omega_m\varepsilon^{m+2}}\int_M\int_{B(x,\varepsilon)} d^2(f(x),f(y))\,d\mathrm{Vol}\,(y)d\mathrm{Vol}\,(x)\\
&\le \frac{1}{\omega_m\varepsilon^{m+2}}\int_M\int_{B(x,\varepsilon)} \big\{d(f(x),f(x+\lambda(y-x)))\\
&\qquad + d(f(x+\lambda(y-x)),f(y))\big\}^2\,d\mathrm{Vol}\,(y)d\mathrm{Vol}\,(x)
\end{aligned}$$

(by the triangle inequality)

$$\begin{aligned}
&\le \frac{1}{\omega_m\varepsilon^{m+2}}\int_M\int_{B(x,\varepsilon)} \Big\{\frac{1}{\lambda}d^2(f(x),f(x+\lambda(y-x)))\\
&\qquad + \frac{1}{1-\lambda}d^2(f(x+\lambda(y-x)),f(y))\Big\}\,d\mathrm{Vol}\,(y)d\mathrm{Vol}\,(x)
\end{aligned}$$

(using the inequality $(a+b)^2\le\frac{1}{\lambda}a^2+\frac{1}{1-\lambda}b^2$, valid for any real numbers a, b).

In local coordinates, with metric tensor (g_{ij}), we have

$$d\mathrm{Vol}\,(y) = \det(g_{ij})^{\frac12}\,dy^1\dots dy^m.$$

By Corollary 1.4.3, we may assume that ε is so small that Riemannian normal coordinates may be introduced on $B(x,\varepsilon)$. In those coordinates, we have from Theorem 1.4.4 that

$$\det(g_{ij}(y))^{\frac12} = 1+\mathrm{O}(\varepsilon^2)\quad\text{for } y\in B(x,\varepsilon).$$

Therefore

$$\frac{d\mathrm{Vol}\,(y)}{d\mathrm{Vol}\,(\lambda y)} = \frac{1}{\lambda^{m-1}}(1+0(\varepsilon^2)).$$

We then substitute $z=\lambda y$ and obtain (noting that x has the coordinate representation 0)

$$\int_{B(x,\varepsilon)}\frac{1}{\lambda}d^2(f(0),f(\lambda y))\,d\mathrm{Vol}\,(y) = \frac{1}{\lambda^m}(1+\mathrm{O}(\varepsilon^2))\int_{B(x,\lambda\varepsilon)} d^2(f(0),f(z))\,d\mathrm{Vol}\,(z).$$

In that manner, we obtain

$$\begin{aligned}
E_\varepsilon(f) &\le \frac{1}{\omega_m\varepsilon^{m+2}}(1+\mathrm{O}(\varepsilon^2))\Big\{\int_M\frac{1}{\lambda^m}\int_{B(x,\lambda\varepsilon)} d^2(f(x),f(z))\,d\mathrm{Vol}\,(z)d\mathrm{Vol}\,(x)\\
&\qquad + \int_M\frac{1}{(1-\lambda)^m}\int_{B(z,(1-\lambda)\varepsilon)} d^2(f(z),f(y))\,d\mathrm{Vol}\,(y)d\mathrm{Vol}\,(z)\Big\}\\
&= (1+\mathrm{O}(\varepsilon^2))(\lambda E_{\lambda\varepsilon}(f)+(1-\lambda)E_{(1-\lambda)\varepsilon}(f)).
\end{aligned} \tag{7.3.4}$$

We put

$$E^n(f) := E_{2^{-n}}(f).$$

Definition 7.3.1. The *energy* of a map $f \in L^2(M, N)$ is defined as

$$E(f) = \lim_{n\to\infty} E^n(f) = \lim_{\varepsilon\to 0} E_\varepsilon(f) \in \mathbb{R} \cup \{+\infty\}. \tag{7.3.5}$$

We also say that $f \in L^2(M, N)$ belongs to the Sobolev space $H^{1,2}(M, N)$ if $E(f) < \infty$.

In order to make contact with more classical definitions of Sobolev spaces, we start with the following

Definition 7.3.2. A map $f : M \to N$ between manifolds is *localizable* if for every $x_0 \in M$ there exists a neighborhood U of x_0 in M and a domain V of a coordinate chart in N with the property that

$$f(U) \subset V.$$

In the sequel, we shall look at maps which are localizable in the sense of Definition 7.3.2. For such maps, all relevant regularity properties can be studied in local coordinates. In particular, it can be defined with the help of local coordinates whether such a map between Riemannian manifolds is of Sobolev class $H^{1,2}(M, N)$.

We now want to establish the result that for such localizable maps, our general definition of the energy coincides with the one obtained by local coordinate representations.

Theorem 7.3.1. *For a localizable map $f \in L^2(M, N)$,*

$$E(f) = d(m) \int_M \langle df, df \rangle \, d\mathrm{Vol}\,(x)$$

whenever the latter expression is defined and finite (where the weak derivative df is defined with the help of local coordinates), and

$$E(f) = \infty$$

otherwise.

Here, $d(m)$ is some factor depending on the dimension of M that can be safely ignored in the sequel.

In the proof of Theorem 7.3.1, we shall employ the following auxiliary result:

Lemma 7.3.3. *For a localizable f, $f \in H^{1,2}(M, N)$, $(M, N$ compact) iff for all Lipschitz functions $\ell : N \to \mathbb{R}$, $l \circ f \in H^{1,2}(M, \mathbb{R})$.*

Proof. We have assumed f to be localizable, and so the $H^{1,2}$-property may be tested in local coordinates. Therefore, if the $H^{1,2}$-property holds for composition with Lipschitz functions it holds for coordinate functions. Conversely, if f is in $H^{1,2}$, then

$\ell \circ f$ is also in $H^{1,2}$ for all Lipschitz functions ℓ by Lemma A.1.3. $\square$

Proof of Theorem 7.3.1. For $f \in C^1$, it is an elementary consequence of Taylor's formula that

$$E(f) = \lim_{\varepsilon \to 0} E_\varepsilon(f). \tag{7.3.6}$$

For $f \in H^{1,2}$ (defined with the help of local coordinates), we choose a sequence $(f_\nu)_{\nu \in \mathbb{N}} \subset C^1$ converging to f in $H^{1,2}$. Given $\delta > 0$, we find ν_0 such that for all $\nu, \mu \geq \nu_0$,

$$|E(f_\nu) - E(f)| < \frac{\delta}{3}. \tag{7.3.7}$$

We write

$$\begin{aligned}
E_\varepsilon(f_\nu) - E_\varepsilon(f_\mu) &= \frac{1}{\omega_m \varepsilon^{m+2}} \int_M \int_{B(x,\varepsilon)} \big(d^2(f_\nu(x), f_\nu(y)) \\
&\qquad - d^2(f_\mu(x), f_\mu(y))\big)\, d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x) \\
&= \frac{1}{\omega_m \varepsilon^{m+2}} \int_M \int_{B(x,\varepsilon)} \big(d(f_\nu(x), f_\nu(y)) \\
&\qquad - d(f_\mu(x), f_\mu(y))\big) d(f_\nu(x), f_\nu(y))\, d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x) \\
&\quad + \frac{1}{\omega_m \varepsilon^{m+2}} \int_M \int_{B(x,\varepsilon)} \big(d(f_\nu(x), f_\nu(y)) \\
&\qquad - d(f_\mu(x), f_\mu(y))\big) d(f_\mu(x), f_\mu(y))\, d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x)\,.
\end{aligned}$$

Now

$$d(f_\nu(x), f_\nu(y)) = \int_0^1 D_2 d\left(f_\nu(x), f_\nu(x + t(y-x))\right)(y-x)\, dt \tag{7.3.8}$$

(for almost all y), where D_2 denotes the derivative w.r.t. the second variable, and we use local coordinates on $B(x,\varepsilon)$. This derivative exists a.e. by Lemma A.1.3 since d is Lipschitz.

Consequently

$$
\begin{aligned}
&\frac{1}{\omega_m \varepsilon^{m+2}} \left| \int_M \int_{B(x,\varepsilon)} \big(d(f_\nu(x), f_\nu(y)) - d(f_\mu(x), f_\mu(y))\big) d(f_\nu(x), f_\nu(y))\, d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x) \right| \\
&\quad \le \frac{1}{\omega_m \varepsilon^{m+2}} \Bigg\{ \int_M \int_{B(x,\varepsilon)} \int_0^1 \big| D_2 d\big(f_\nu(x), f_\nu(x + t(y-x))\big) \\
&\qquad\qquad - D_2 d\big(f_\mu(x), f_\mu(x + t(y-x))\big)\big|^2 \big|y - x\big|^2 \, dt\, d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x) \Bigg\}^{\frac{1}{2}} \\
&\qquad \cdot \Bigg\{ \int_M \int_{B(x,\varepsilon)} d^2(f_\nu(x), f_\nu(y))\, d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x) \Bigg\}^{\frac{1}{2}} \quad \text{by Hölder's inequality} \\
&\quad \le E_\varepsilon(f_\nu)^{\frac{1}{2}} \frac{1}{\omega_m^{\frac{1}{2}} \varepsilon^{\frac{m}{2}}} \Bigg\{ \int_M \int_{B(x,\varepsilon)} \int_0^1 \big| D_2 d\big(f_\nu(x), f_\nu(x + t(y-x))\big) \\
&\qquad\qquad - D_2 d\big(f_\mu(x), f_\mu(x + t(y-x))\big)\big|^2 \, dt\, d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x) \Bigg\}^{\frac{1}{2}} . \qquad (7.3.9)
\end{aligned}
$$

Since (f_ν) converges in $H^{1,2}$, by Lemma 7.3.3 then $(D_2 d(f_\nu(x), f_\nu(\cdot)))_{\nu \in \mathbb{N}}$ converges in L^2 for every x. Therefore, given $\eta > 0$, there exists $\nu_1 \ge \nu_0$ so that for all $\nu, \mu \ge \nu_1$, the preceding expression is bounded by

$$\eta E_\varepsilon(f_\nu)^{\frac{1}{2}}.$$

(For M a compact Riemannian manifold, and an integrable function $\varphi : M \to \mathbb{R}$, $\int_M \int_{B(x,\varepsilon)} \int_0^1 \varphi(x + t(y-x))\, dt d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x)$ behaves like $\omega_m \varepsilon^m \int_M \varphi(z)\, d\mathrm{Vol}\,(z)$ as $\varepsilon \to 0$.)

We thus obtain

$$|E_\varepsilon(f_\nu) - E_\varepsilon(f_\mu)| \le \eta (E_\varepsilon(f_\nu)^{\frac{1}{2}} + E_\varepsilon(f_\mu)^{\frac{1}{2}}). \qquad (7.3.10)$$

From (7.3.8), we see that $E_\varepsilon(f_\nu)$ is controlled by the energy $E(f_\nu)$ and since the latter is bounded since it converges to $E(f)$, we may assume

$$E_\varepsilon(f_\nu) \le K \quad \text{for some constant } K \text{ and all } \nu.$$

Hence by a suitable choice of η in (7.3.10), we have for all $\nu, \mu \ge \nu_1$

$$|E_\varepsilon(f_\nu) - E_\varepsilon(f_\mu)| < \frac{\delta}{3}. \qquad (7.3.11)$$

We then choose $\varepsilon > 0$ so small that

$$|E_\varepsilon(f_{\nu_1}) - E(f_{\nu_1})| < \frac{\delta}{3} \qquad (7.3.12)$$

which is possible by (7.3.6). From (7.3.7), (7.3.11), (7.3.12), we conclude

$$|E_\varepsilon(f) - E(f)| < \delta$$

for all sufficiently small ε. This is the claim for $f \in H^{1,2}$.

In order to establish the result for general (localizable) $f \in L^2(M, N)$, we show that if $E_\varepsilon(f)$ stays bounded for $\varepsilon \to 0$, then $f \in H^{1,2}(M, N)$. For that purpose, we use the characterization of Lemma 7.3.3.

Let $\ell : N \to \mathbb{R}$ be Lipschitz. If $E_\varepsilon(f)$ is bounded, so then is

$$E_\varepsilon(\ell \circ f) = \frac{1}{\omega_m \varepsilon^{m+2}} \int_M \int_{B(x,\varepsilon)} |\ell \circ f(x) - \ell \circ f(y)|^2 \, d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x).$$

Introducing Riemannian polar coordinates (r, φ) on $B(x, \varepsilon)$ (ε sufficiently small, cf. Corollary 1.4.3), we compute

$$E_\varepsilon(\ell \circ f) = \frac{1}{\omega_m} \int_M \int_{B(0,1)} \frac{|\ell \circ f(x + \varepsilon y) - \ell \circ f(x)|^2}{\varepsilon^2} \frac{\varepsilon^m}{\varepsilon^m} \, dy \, d\mathrm{Vol}\,(x),$$

up to an error term that goes to 0 for $\varepsilon \to 0$. Since this is assumed to be bounded as $\varepsilon \to 0$, for almost all $y \in B(0, 1)$, the difference quotients

$$\Delta_y^\varepsilon(\ell \circ f)(x) = \frac{\ell \circ f(x + \varepsilon y) - \ell \circ f(x)}{\varepsilon}$$

are uniformly bounded in L^2. By Lemma A.2.2, we conclude that $\ell \circ f \in H^{1,2}$. Since this holds for every Lipschitz function ℓ, by Lemma 7.3.3, $f \in H^{1,2}$. This completes the proof. $\square$

We now want to show the lower semicontinuity of the energy E w.r.t. L^2-convergence.

Theorem 7.3.2. *If $(f_\nu)_{\nu \in \mathbb{N}}$ converges to f in $L^2(M, N)$, then*

$$E(f) \le \liminf_{\nu \to \infty} E(f_\nu).$$

Proof. We may assume

$$\liminf_{\nu \to \infty} E(f_\nu) < \infty,$$

hence also

$$E(f_\nu) \le K \tag{7.3.13}$$

for some constant K and all ν. By definition

$$E(f) = \lim_{n \to \infty} E^n(f).$$

Given $\delta > 0$, there then exists n_0 such that for all $n \ge n_0$

$$E(f) \le E^n(f) + \delta.$$

By Lemma 7.3.2, E^n is continuous on L^2. Hence there exists ν_0 such that for all $\nu \geq \nu_0$ with ν_0 depending on δ and n_0,

$$E(f) \leq E^{n_0}(f_\nu) + 2\delta. \tag{7.3.14}$$

Applying (7.3.4) with $\lambda = \frac{1}{2}$, we obtain

$$E^n(f_\nu) \leq (1 + 0(2^{-2n}))E^{n+1}(f_\nu).$$

Possibly choosing n_0 larger, we obtain for all $n \geq n_0$

$$E^n(f_\nu) \leq E(f_\nu) + \delta \tag{7.3.15}$$

using (7.3.13).

(7.3.14) and (7.3.5) imply

$$E(f) \leq E(f_\nu) + 3\delta \quad \text{for all } \nu \geq \nu_0.$$

Since $\delta > 0$ was arbitrary, the claim follows. □

We now wish to relate the above results to a general concept of variational convergence, the Γ-convergence in the sense of de Giorgi. In order to introduce that concept, let Z be a topological space satisfying the first axiom of countability;[2] that means that for every $x \in Z$, we may find a sequence $(U_\nu)_{\nu \in \mathbb{N}}$ of open subsets of Z such that every open set containing x also contains some U_ν. In our applications, Z of course will be $L^2(M, N)$ or some subspace of that space.

Let

$$F_n : Z \to \mathbb{R} \cup \{\pm\infty\},\ n \in \mathbb{N},$$

be a sequence of functionals.

Definition 7.3.3. The functional

$$F : Z \to \mathbb{R} \cup \{\pm\infty\}$$

is the Γ-limit of $(F_n)_{n \in \mathbb{N}}$, written as

$$F = \Gamma - \lim_{n \to \mathbb{N}} F_n$$

if

(i) whenever $(x_n)_{n \in \mathbb{N}} \subset Z$ converges to $x \in Z$,

$$F(x) \leq \liminf_{n \in \infty} F_n(x_n),$$

[2]This is assumed only for the simplicity of presentation; the concept is meaningful also for spaces that do not satisfy the first axiom of countability; one has to replace sequences by filters in that case.

(ii) for every $x \in Z$, we can find a sequence $(x_n)_{n\in\mathbb{N}} \subset Z$ that converges to x and satisfies

$$F(x) = \lim_{n\to\infty} F_n(x_n).$$

Lemma 7.3.4. *$E = \Gamma - \lim E_\varepsilon$ w.r.t. L^2-convergence.*

Proof. By monotonicity (see (7.3.4)), it suffices to show the result for E^n instead of E_ε.

(i) For every $f \in L^2(M,N)$, there exists a sequence $(f_\nu)_{\nu\in\mathbb{N}} \subset L^2(M,N)$

$$E(f) = \lim_{\nu\to\infty} E^\nu(f_\nu).$$

According to the definition of E, we may simply take $f_\nu = f$ for all ν.

(ii) For every sequence $(f_\nu)_{\nu\in\mathbb{N}} \subset L^2(M,N)$ converging to f we have

$$E(f) \le \liminf_{\nu\to\infty} E^\nu(f_\nu).$$

From the definition of E, for any $\delta > 0$ there exists $n_0 \in \mathbb{N}$ such that for $\nu \ge n_0$

$$E(f) \le E^\nu(f) + \delta.$$

Using this estimate and that E^ν is continuous on L^2 by Lemma 7.3.1, we may find ν_0 (depending on δ and n_0) such that for $\nu \ge \nu_0$,

$$E(f) \le E^{n_0}(f_\nu) + 2\delta.$$

From (7.3.2) with $\lambda = \frac{1}{2}$, we get

$$E^n(f_\nu) \le (1 + c\,2^{-2n})E^{n+1}(f_\nu),$$

for some constant c, depending on the geometry of M.

We may have chosen n_0 in the preceding also satisfying

$$\prod_{n\ge n_0} (1 + c\,2^{-2n}) \le 1 + \delta.$$

Then from the preceding estimate

$$E^{n_0}(f_\nu) \le (1+\delta)E^\nu(f_\nu) \quad \text{for } \nu \ge n_0.$$

Putting the estimates together,

$$E(f) \le (1+\delta)E^\nu(f_\nu) + 2\delta \quad \text{for } \nu \ge n_0, \nu_0.$$

As this holds for any $\delta > 0$,

$$E(f) \le \liminf_{\nu\to\infty} E^\nu(f_\nu).$$

□

This result is quite useful, because, in view of the next lemma, it tells us that if for some sequence $\varepsilon_n \to 0$, we can find a minimizer f_n for every E_{ε_n} and if this sequence converges to some f, then f automatically minimizes E. In other words, we can find a minimizer for E by minimizing the simpler approximating functionals E_ε.

Lemma 7.3.5. *Let*

$$F = \Gamma - \lim_{n\to\infty} F_n$$

in the above setting. Assume that every F_n is bounded from below, and that x_n minimizes F_n. If x_n converges to $x \in Z$, then x minimizes F, and

$$F(x) = \lim_{n\to\infty} F_n(x_n). \tag{7.3.16}$$

Proof. Let $z \in Z$.

Since F is the Γ-limit of the F_n, we can find some sequence $(z_n)_{n\in\mathbb{N}}$ converging to z with

$$\lim_{n\to\infty} F_n(z_n) = F(z).$$

Given $\varepsilon > 0$, we choose $n \in \mathbb{N}$ so large that

$$F_n(z_n) < F(z) + \frac{\varepsilon}{2}$$

and also

$$F_n(x_n) > F(x) - \frac{\varepsilon}{2} \quad \text{(property (i) of } \Gamma\text{-convergence).}$$

Since x_n minimizes F_n,

$$F_n(x_n) \le F_n(z_n).$$

Altogether

$$F(x) < F(z) + \varepsilon.$$

Since this holds for every $z \in Z$ and every $\varepsilon > 0$, x minimizes F. By Γ-convergence

$$F(x) \le \liminf_{n\to\infty} F_n(x_n),$$

and we may find a sequence $(z_n)_{n\in\mathbb{N}}$ converging to x with

$$F(x) = \lim_{n\to\infty} F_n(z_n).$$

Since

$$F_n(x_n) \le F_n(z_n)$$

because of the minimizing property of x_n, (7.3.16) follows. □

Γ-limits are automatically lower semicontinuous, and so, we could have deduced Theorem 7.3.2 from that general result about Γ-convergence.

Perspectives. The definition and treatment of the energy functional presented here are taken from Jost[138]. (See also [139].) A similar theory is developed by Korevaar and Schoen[171]. For the usual definition of the Sobolev space $H^{1,2}(M,N)$, see Exercise 8. The concept of Γ-convergence is treated in dal Maso[58] and Jost, Li-Jost[149].

7.4 Higher Regularity

In this section, we study continuous solutions $f \in H^{1,2}(\Omega, \mathbb{R}^n)$, Ω open in $\mathbb{R}^m$, of a system

$$\int_\Omega a^{\alpha\beta}(x) D_\alpha f^i(x) D_\beta \varphi^i(x)\, dx = \int_\Omega G^i(x, f(x), Df(x))\varphi^i(x)\, dx \tag{7.4.1}$$

for all $\varphi \in H_0^{1,2} \cap L^\infty(\Omega, \mathbb{R}^n)$.

We shall assume the following structure conditions:

$$\begin{aligned} &(a^{\alpha\beta}(x))_{\alpha,\beta=1,\dots,m} \text{is symmetric for almost all } x, \text{the coefficients } \alpha^{\alpha\beta}(x) \\ &\qquad\qquad \text{are measurable;} \\ &a^{\alpha\beta}(x)\xi_\alpha\xi_\beta \geq \lambda|\xi|^2 \text{ for all } \xi = (\xi_1, \dots, \xi_m) \in \mathbb{R}^m \text{ and almost all } x \in \Omega \end{aligned} \tag{A1}$$

with a constant $\lambda > 0$, and

$$|a^{\alpha\beta}(x)| \leq K \quad \text{for almost all } x \in \Omega \tag{A2}$$

with a constant K.

$G(x, f, p) = (G^1, \dots, G^n)$ is measurable in x and continuous in f and p. This implies that $G(x, f(x), Df(x))$ is measurable in x for $f \in H^{1,1}_{\text{loc}}$.

$$|G(x, f, p)| \leq c_0 + c_1|p|^2 \quad \text{for all } (x, f, p) \in \Omega \times \mathbb{R}^n \times \mathbb{R}^{mn} \tag{G1}$$

with constants c_0, c_1.

Later on, $a^{\alpha\beta}$ and G^i will be assumed even differentiable, and so we may as well assume here that they are continuous instead of just measurable.

If f is a continuous weakly harmonic map, then continuity allows us to localize the situation not only on the domain, but also in the image, i.e. to write everything down in fixed local coordinates. The preceding structural conditions then are satisfied, cf. Lemma 7.4.2.

Some notational conventions:
We usually omit the indices in the image; thus e.g.

$$D_\alpha f \cdot D_\beta \varphi := D_\alpha f^i D_\beta \varphi^i \quad \text{with the standard summation convention.}$$

(Usually, also the dot "$\cdot$" will be omitted.) Also, we shall always integrate w.r.t. to the Euclidean volume element dx on Ω, and this will often be omitted.

We start with the following auxiliary result

Lemma 7.4.1. *Suppose $f \in C^0 \cap H^{1,2}(\Omega, \mathbb{R}^n)$ solves (7.4.1), where the coefficients satisfy (A1), (A2), (G1).*

Then for every $\varepsilon > 0$, there exists $\rho > 0$, depending on ε, m, the structural constants λ, K, c_0, c_1, and on the modulus of continuity of f, with

$$\int_{B(x_1,\rho)} |Df|^2 \eta^2(x)\, dx \le \varepsilon \int_{B(x_1,\rho)} |D\eta|^2\, dx \tag{7.4.2}$$

whenever $B(x_1,\rho) \subset \Omega$ and $\eta \in H_0^{1,2}(B(x_1,\rho), \mathbb{R})$.

Proof. We choose

$$\varphi(x) := (f(x) - f(x_1))\eta^2(x)$$

in (7.4.1). We obtain

$$\begin{aligned}
\int_{B(x_1,\rho)} a^{\alpha\beta}(x) D_\alpha f D_\beta f\, \eta^2 &\le c_2 \sup_{x\in B(x_1,\rho)} |f(x) - f(x_1)| \int_{B(x_1,\rho)} |Df|^2 \eta^2 \\
&\quad + c_3 \sup_{x \in B(x_1,\rho)} |f(x) - f(x_1)| \int \eta^2 \\
&\quad + 2 \int_{B(x_1,\rho)} a^{\alpha\beta}(x) D_\alpha f D_\beta \eta\, (f(x) - f(x_1))\eta \quad \text{because of (G1)} \\
&\le c_2 \sup |f(x) - f(x_1)| \int_{B(x_1,\rho)} |Df|^2 \eta^2 \\
&\quad + c_4 \sup |f(x) - f(x_1)| \rho^2 \int_{B(x_1,\rho)} |D\eta|^2 \\
&\quad + \frac{1}{2} \int_{B(x_1,\rho)} a^{\alpha\beta}(x) D_\alpha f D_\beta f \eta^2 \\
&\quad + 8 \sup |f(x) - f(x_1)|^2 \int a^{\alpha\beta}(x) D_\alpha \eta D_\beta \eta\,,
\end{aligned}$$

where we have used the Poincaré inequality (Corollary A.1.1) for the second term. The claim follows with (A1), (A2) because we can make $\sup_{B(x_1,\rho)} |f(x) - f(x_1)|$ arbitrarily small by choosing ρ sufficiently small, since f is continuous. $\square$

In order to proceed, we have to make additional structural assumptions about the system (7.4.1):

The coefficients $a^{\alpha\beta}(x)$ are differentiable and

$$|D_\gamma a^{\alpha\beta}(x)| \le K_1 \quad \text{for all } \alpha, \beta, \gamma = 1, \dots, m, x \in \Omega \tag{A3}$$

with a constant K_1.

$G = (G^1, \dots, G^n)$ is differentiable with

$$\begin{aligned} |D_x G(x, f, p)| &\le \gamma_0 + \gamma_1 |p|^3, \\ |D_f G(x, f, p)| &\le \gamma_2 + \gamma_3 |p|^2, \\ |D_p G(x, f, p)| &\le \gamma_4 + \gamma_r |p| . \end{aligned} \tag{G2}$$

In order to show the main idea of the subsequent regularity argument, we shall first derive a so-called *a priori estimate*. This means that assuming that we already have a *regular* solution, we can estimate its norms.

Lemma 7.4.2. *Suppose*

$$f \in C^0 \cap H^{1,4} \cap H^{3,2}(B(x_0, R), \mathbb{R}^n)$$

is a solution of (7.4.1) with $\Omega = B(x_0, R)$*, where the structural conditions (A1), (A2), (A3), (G1), (G2) are satisfied. Then*

$$\|D^2 f\|_{L^2(B(x_0, \frac{R}{2}))} + \|Df\|^2_{L^4(B(x_0, \frac{R}{2}))} \le C_0 R^{\frac{m}{2}} + C_1 \|Df\|_{L^2(B(x_0, R))} , \tag{7.4.3}$$

where C_0 *and* C_1 *depend on the structural constants in (A1) – (G2), on* m*, and the modulus of continuity of* f*.*

Proof. Since $f \in H^{2,2}$, for $\varphi \in H^{1,2}_0$,

$$\int a^{\alpha\beta} D_\alpha f D_\beta \varphi = -\int D_\beta(a^{\alpha\beta} D_\alpha f)\varphi . \tag{7.4.4}$$

We now put

$$\varphi = D_\gamma(\xi^2 D_\gamma f)$$

with $\xi \in L^\infty \cap h^{1,2}_0(B(x_0, R), \mathbb{R})$ to be determined later on.

From (7.4.1), (7.4.4)

$$\begin{aligned} \int_{B(x_0,R)} D_\gamma(a^{\alpha\beta} D_\alpha f) D_\beta(\xi^2 D_\gamma f) &= -\int_{B(x_0,R)} a^{\alpha\beta} D_\alpha f \cdot D_\beta(D_\gamma(\xi^2 D_\gamma f)) \\ &= -\int_{B(x_0,R)} G(x, f, Df) D_\gamma(\xi^2 D_\gamma f) \\ &= \int_{B(x_0,R)} D_\gamma(G(x, f, Df)) D_\gamma f \cdot \xi^2 . \end{aligned} \tag{7.4.5}$$

Now

$$
\begin{aligned}
D_\gamma(a^{\alpha\beta}D_\alpha f)D_\beta(\xi^2 D_\gamma f) = a^{\alpha\beta}D_\gamma D_\alpha f \cdot D_\beta D_\gamma f \cdot \xi^2 \\
+ a^{\alpha\beta}D_\gamma D_\alpha f \cdot D_\gamma f \cdot D_\beta \xi^2 \\
+ D_\gamma a^{\alpha\beta} \cdot D_\alpha f \cdot D_\beta D_\gamma f \cdot \xi^2 \\
+ D_\gamma a^{\alpha\beta} \cdot D_\alpha f \cdot D_\gamma f \cdot D_\beta \xi^2 \,,
\end{aligned}
\tag{7.4.6}
$$

and from (G2),

$$
|D_\gamma G(x,f,Df)||D_\gamma f| \le c_5|Df| + c_6|Df|^4 + c_7|Df|\cdot|D^2 f| + c_8|Df|^2|D^2 f|\,, \tag{7.4.7}
$$

and from (A1),

$$
|D^2 f|^2 \le \frac{1}{\lambda} a^{\alpha\beta} D_\gamma D_\alpha f \cdot D_\gamma D_\beta f. \tag{7.4.8}
$$

From (7.4.5) – (7.4.8) we conclude, using also (A2), (A3),

$$
\begin{aligned}
\int_{B(x_0,R)} |D^2 f|^2 \cdot \xi^2 &\le c_9 \int_{B(x_0,R)} |D^2 f||Df||\xi D\xi| + c_{10}\int_{B(x_0,R)} |D^2 f||Df|\xi^2 \\
&\quad + c_{11}\int_{B(x_0,R)} |Df|^2|\xi D\xi| + c_5 \int_{B(x_0,R)} \xi^2 \\
&\quad + c_6 \int_{B(x_0,R)} |Df|^4 \xi^2 + c_8 \int_{B(x_0,R)} |D^2 f||Df|^2\xi^2 \\
&\le \varepsilon_1 c_9 \int_{B(x_0,R)} |D^2 f|^2 \xi^2 + \frac{c_9}{4\varepsilon_1}\int_{B(x_0,R)} |Df|^2|D\xi|^2 \\
&\quad + \varepsilon_2 c_{10}\int_{B(x_0,R)} |D^2 f|^2\xi^2 + \frac{c_{10}}{4\varepsilon_2}\int_{B(x_0,R)} |Df|^2\xi^2 \\
&\quad + \frac{c_{11}}{2}\int_{B(x_0,R)} |Df|^2|D\xi|^2 + \frac{c_{11}}{2}\int_{B(x_0,R)} |Df|^2\xi^2 \\
&\quad + c_6\int_{B(x_0,R)} |Df|^4\xi^2 + c_5\int_{B(x_0,R)} \xi^2 \\
&\quad + \varepsilon_3 c_8 \int_{B(x_0,R)} |D^2 f|^2\xi^2 + \frac{c_8}{4\varepsilon_3}\int_{B(x_0,R)} |Df|^4\xi^2
\end{aligned}
\tag{7.4.9}
$$

with arbitrary positive ε_1, ε_2, ε_3, where we have used the inequality

$$
ab \le \varepsilon a^2 + \frac{1}{4\varepsilon} b^2 \quad \text{for arbitrary } \varepsilon > 0, a, b \in \mathbb{R}. \tag{7.4.10}
$$

We may choose $\varepsilon_1, \varepsilon_2, \varepsilon_3 > 0$ so small that

$$
\varepsilon_1 c_9 + \varepsilon_2 c_{10} + \varepsilon_3 c_8 \le \frac{1}{2}
$$

and obtain

$$
\int_{B(x_0,R)} |D^2 f|^2\xi^2 \le c_{12}\int_{B(x_0,R)} |Df|^2|D\xi|^2 + c_{13}\int_{B(x_0,R)} \xi^2 + c_{14}\int_{B(x_0,R)} |Df|^4\xi^2. \tag{7.4.11}
$$

For $\varepsilon > 0$, we now choose $\rho > 0$ as in Lemma 7.4.1.

We assume

$$B(x_1, \rho) \subset B(x_0, R)$$

and choose $\xi \in C_0^\infty(B(x_1, \rho))$ with $0 \le \xi \le 1$,

$$\xi \equiv 1 \quad \text{on } B\Big(x_1, \frac{\rho}{2}\Big),$$
$$|D\xi| \le \frac{4}{\rho}\,.$$

Thus, all preceding integrals need to be evaluated only on $B(x_1, \rho)$. We now write

$$\int_{B(x_1,\rho)} |Df|^4 \xi^2 = \int_{B(x_1,\rho)} |Df|^2 (|Df|^2 \cdot \xi^2)$$

and apply Lemma 7.4.1 with $\eta = |Df| \cdot \xi$ and obtain

$$\begin{aligned} \int_{B(x_1,\rho)} |Df|^4 \xi^2 &\le \varepsilon \int_{B(x_1,\rho)} |D(|Df|\xi^2)|^2 \\ &\le \varepsilon \int_{B(x_1,\rho)} |D^2 f|^2 \xi^2 + \varepsilon \int_{B(x_1,\rho)} |Df|^2 |D\xi|^2. \end{aligned} \tag{7.4.12}$$

We may choose $\varepsilon > 0$ so small that

$$\varepsilon c_{14} \le \frac{1}{2}.$$

(7.4.11) and (7.4.12) give

$$\begin{aligned} \int_{B(x_1,\frac{\rho}{2})} |D^2 f|^2 &\le c_{15} \int_{B(x_1,\rho)} \xi^2 + c_{16} \int_{B(x_1,\rho)} |Df|^2 |D\xi|^2 \\ &\le c_{17} \rho^m + \frac{c_{18}}{\rho^2} \int_{B(x_1,\rho)} |Df|^2\,. \end{aligned} \tag{7.4.13}$$

Covering $B(x_0, \frac{R}{2})$ by balls $B(x_1, \frac{\rho}{2})$ with $B(x_1, \rho) \subset B(x_0, R)$, we obtain the desired estimate for

$$\int_{B(x_0,\frac{R}{2})} |D^2 f|^2.$$

(7.4.12) and (7.4.13) and the same covering argument then also yield the estimate for

$$\int_{B(x_0,\frac{R}{2})} |Df|^4.$$

□

However, we cannot apply Lemma 7.4.2 because we do not know yet that $f \in H^{3,2}$. The point, however, is that the conclusion does not depend on the $H^{3,2}$-norm, and a slight modification will give us the desired regularity result:

Lemma 7.4.3. *Suppose that*

$$f \in C^0 \cap H^{1,2}(B(x_0,R),\mathbb{R}^n)$$

is a solution of (7.4.1) with $\Omega = B(x_0,R)$ *and the structural conditions (A1), (A2), (A3), (G1), (G2). Then*

$$f \in H^{2,2} \cap H^{1,4}\Big(B\Big(x_0,\frac{R}{2}\Big),\mathbb{R}^n\Big),$$

and the same estimate as in Lemma 7.4.2 holds.

Proof. We just replace certain weak derivatives by difference quotients (cf. (A.2.1)) in the proof of Lemma 7.4.2. Namely, we put

$$\varphi := \Delta_\gamma^{-h}(\xi^2 \Delta_\gamma^h f)$$

with ξ as above.

Analogously to (7.4.11), we get with $\Delta^h = (\Delta_1^h,\dots,\Delta_m^h)$,

$$\begin{aligned}\int_{B(x_0,R)} |D(\Delta^h f)|^2\xi^2 \le c_{12}&\int_{B(x_0,R)} |\Delta^h f|^2|D\xi|^2\\ &+ c_{13}\int_{B(x_0,R)}\xi^2 + c_{14}\int_{B(x_0,R)} |Df|^2|\Delta^h f|^2\xi^2\,.\end{aligned} \tag{7.4.14}$$

But

$$\int |\Delta_\gamma^h f|^2|D\xi|^2 \le \int |Df|^2|D\xi|^2$$

(this is similar to Lemma A.2.1).

Using Lemma 7.4.1, we then obtain analogously to (7.4.13),

$$\int_{B(x_1,\frac{\rho}{2})} |D(\Delta^h f)|^2\xi^2 \le c_{17}\rho^n + \frac{c_{18}}{\rho^2}\int_{B(x_1,\rho)} |Df|^2. \tag{7.4.15}$$

Lemma A.2.2 then shows that the weak derivative $D^2 f$ exists and satisfies the same estimate. Likewise, we get control over the L^4-norm of Df. □

Lemma 7.4.4. *Let* $f \in C^0 \cap H^{1,2}(B(x_0,R),\mathbb{R}^n)$ *be a solution of (7.4.1), with structural conditions (A1), (A2), (A3), (G1), (G2) satisfied. Then*

$$Df \in L^p\Big(B\Big(x_0,\frac{R}{4}\Big)\Big) \quad \textit{for every } p<\infty,$$

and

$$\int_{B(x_0,\frac{R}{4})} |Df|^p|D^2 f|^2 < \infty. \tag{7.4.16}$$

Remark. As in Lemma 7.4.2, one also gets a-priori estimates, with constants depending also on p.

Proof. By Lemma 7.4.3, we know already

$$Df \in H^{1,2} \cap L^4\Big(B\Big(x_0, \frac{R}{2}\Big)\Big).$$

We put

$$w := |Df|^2.$$

We are going to show by induction that for every $s \in \mathbb{N}$, $s \geq 2$ and $R_1 < \frac{R}{2}$,

$$\int_{B(x_0,R_1)} (w^s + w^{s-2}|D^2 f|^2) < \infty. \tag{E_s}$$

By Lemma 7.4.3, (E_s) holds for $s = 2$. We assume (E_s) for s and want to conclude (E_s) for $s+1$, i.e. (E_{s+1}).

We put

$$w_L(x) := \min(w(x), L) \quad \text{for } L > 0.$$

We observe

$$Dw_L(x) = 0 \quad \text{if } w(x) > L, \tag{7.4.17}$$

$$|Dw| \leq 2|D^2 f| w^{\frac{1}{2}}, \tag{7.4.18}$$

and

$$|Dw_L| \leq 2|D^2 f| w_L^{\frac{1}{2}} \quad \text{from (7.4.17), (7.4.18).} \tag{7.4.19}$$

Let $\eta \in L^\infty \cap H_0^{1,2}(B(x_0, R_1))$. We compute, for $x_1 \in B(x_0, R_1)$,

$$\begin{aligned}
&\int_{B(x_0,R_1)} \eta^2 w^s w_L \\
&\qquad = \int_{B(x_0,R_1)} \eta^2 Df \cdot Df w^{s-1} w_L \quad \text{by definition of } w \\
&\qquad = \int_{B(x_0,R_1)} \eta^2 D(f - f(x_1)) \cdot Df w^{s-1} w_L \\
&\qquad \leq (2m+1) \sup_{\substack{x \in B(x_0,R_1) \\ \cap\, \operatorname{supp} \eta}} |f(x) - f(x_1)| \int_{B(x_0,R_1)} \eta^2 |D^2 f| w^{s-1} w_L \\
&\qquad\quad + 2 \sup_{\substack{x \in B(x_0,R_1) \\ \cap\, \operatorname{supp} \eta}} |f(x) - f(x_1)| \int_{B(x_0,R_1)} \eta D\eta w^{\frac{2s-1}{2}} w_L \,,
\end{aligned} \tag{7.4.20}$$

integrating by parts and using (7.4.18), (7.4.19).

We now write

$$\eta^2|D^2 f|w^{s-1}w_L = \left(\eta|D^2 f|w^{\frac{s-2}{2}}w_L^{\frac{1}{2}}\right)\left(\eta w^{\frac{s}{2}}w_L^{\frac{1}{2}}\right)$$

and

$$\eta D\eta w^{\frac{2s-1}{2}}w_L = \left(\eta w^{\frac{s}{2}}w_L^{\frac{1}{2}}\right)\left(D\eta w^{\frac{s-2}{2}}w_L^{\frac{1}{2}}\right)$$

and obtain from (7.4.20) (with $2ab \le a^2 + b^2$)

$$\int_{B(x_0,R_1)} \eta^2 w^s w_L \le \sup_{\substack{x\in B(x_0,R_1)\\ \cap\, \operatorname{supp}\eta}} |f(x) - f(x_1)| \cdot$$

$$\left\{c_{19}\int_{B(x_0,R_1)} \eta^2|D^2 f|^2 w^{s-2}w_L + c_{20}\int_{B(x_0,R_1)} \eta^2 w^s w_L + c_{21}\int_{B(x_0,R_1)} |D\eta|^2 w^{s-1}w_L\right\}. \tag{7.4.21}$$

Here, the constants c_{19} and c_{20} also depend on s.

Since f is continuous, given $\varepsilon > 0$ there then exists $\overline{R}(\varepsilon)$ with the property that for $0 < R_2 \le \overline{R}(\varepsilon)$ and $\eta \in H_0^{1,2}(B(x_1,R_2))$ $(B(x_1,R_2) \subset B(x_0,R_1))$,

$$\int_{B(x_1,R_2)} \eta^2 w^s w_L \le \varepsilon \int_{B(x_1,R_2)} \eta^2|D^2 f|^2 w^{s-2}w_L + \varepsilon \int_{B(x_1,R_2)} |D\eta|^2 w^{s-1}w_L\,. \tag{7.4.22}$$

We now require for $\eta \in H_0^{1,2}(B(x_1,R_2))$,

$$\begin{aligned} \eta \equiv& 1 \quad \text{on } B(x_1,\frac{R_2}{2}),\\ 0 \le& \eta \le 1,\\ |D\eta| \le& \frac{2}{R_2}\,. \end{aligned}$$

Since $f \in W^{2,2}$ by Lemma 7.4.3, the equation (7.4.1) yields for $\psi \in H_0^{2,2}$,

$$\begin{aligned} \int D_\gamma(a^{\alpha\beta}D_\alpha f)D_\beta\psi &= -\int a^{\alpha\beta}D_\alpha f D_\gamma D_\beta\psi\\ &= -\int G(x,f,Df)D_\gamma\psi\\ &= \int D_\gamma G(x,f,Df)\psi\,. \end{aligned} \tag{7.4.23}$$

The resulting equation

$$\int D_\gamma(a^{\alpha\beta}D_\alpha f)D_\beta\psi = \int D_\gamma G(x,f,Df)\psi, \tag{7.4.24}$$

then also holds for $\psi \in H_0^{1,2}$ (instead of $H_0^{2,2}$), because we can approximate $\psi \in H_0^{1,2}$ by $H_0^{2,2}$-functions (actually, even by C_0^∞-functions), and an easy application of Lebesgue's theorem on dominated convergence allows the passage to the limit.

We apply (7.4.24) to

$$\psi := \eta^2 w_M^{s-2} w_L D_\gamma f$$

and obtain

$$\begin{aligned}\int_{B(x_1,R_2)} & D_\gamma(a^{\alpha\beta} D_\alpha f)(D_\beta(\eta^2 w_M^{s-2} w_L D_\gamma f)) \\ &= \int_{B(x_1,R_2)} D_\gamma(G(x,f,Df))\eta^2 w_M^{s-2} w_L D_\gamma f\end{aligned} \tag{7.4.25}$$

and from this equation and the structural conditions (as in the derivation of (7.4.9)),

$$\begin{aligned}\int_{B(x_1,R_2)} & a^{\alpha\beta} D_\gamma D_\alpha f \cdot D_\beta(\eta^2 w_M^{s-2} w_L D_\gamma f) \\ &\le c_{22} \int_{B(x_1,R_2)} \eta^2 |D^2 f| w^{\frac{1}{2}} w_M^{s-2} w_L \quad + c_{23} \int_{B(x_1,R_2)} \eta D\eta w w_M^{s-2} w_L \\ &\quad + c_{24} \int_{B(x_1,R_2)} \eta^2 w^{\frac{1}{2}} w_M^{s-2} w_L \qquad + c_{25} \int_{B(x_1,R_2)} \eta^2 w^2 w_M^{s-2} w_L \\ &\quad + c_{26} \int_{B(x_1,R_2)} \eta^2 |D^2 f| w w_M^{s-2} w_L\,,\end{aligned} \tag{7.4.26}$$

where c_{22} again depends on s.

Now

$$\begin{aligned}\int \eta^2 |D^2 f| w w_M^{s-2} w_L &\le \int \eta^2 |D^2 f| w^{s-2} w_L \\ &\le \delta \int \eta^2 |D^2 f|^2 w^{s-2} w_L + \frac{1}{4\delta}\, \eta^2 w^s w_L\,,\end{aligned} \tag{7.4.27}$$

and this is bounded because of (E_s) and since w_L is bounded.

Likewise

$$\int \eta^2 |D^2 f| w^{\frac{1}{2}} w_M^{s-2} w_L \le \delta \int \eta^2 |D^2|^2 w^{s-2} w_L + \frac{1}{4\delta} \int \eta^2 w^{s-1} w_L\,. \tag{7.4.28}$$

Therefore, all terms on the right hand side of (7.4.26) remain bounded as $M \to \infty$. The same then has to happen for the left hand side of (7.4.26). We may hence replace w_M by w in (7.4.26), and conclude that

$$\int_{B(x_1,R_2)} a^{\alpha\beta} D_\alpha D_\gamma f \cdot D_\beta(\eta^2 w^{s-2} w_L D_\gamma f) < \infty. \tag{7.4.29}$$

But this expression equals

$$\begin{aligned}
&\int a^{\alpha\beta} D_\alpha D_\gamma f \cdot D_\beta D_\gamma f w^{s-2} w_L \eta^2 \\
&\quad + (m-2) \int a^{\alpha\beta} D_\alpha D_\gamma f D_\beta w w^{s-3} w_L D_\gamma f \eta^2 \\
&\quad + \int a^{\alpha\beta} D_\alpha D_\gamma f D_\beta w_L w^{s-2} D_\gamma f \eta^2 \\
&\quad + 2 \int a^{\alpha\beta} D_\alpha D_\gamma f \cdot \eta D_\beta \eta w^{s-2} w_L D_\gamma f \,.
\end{aligned} \tag{7.4.30}$$

Since $D_\beta w = D_\beta D_\delta f \cdot D_\delta f$ and $D w_L = 0$ for $w > w_L$, we can rewrite the second and the third integral in (7.4.30) as

$$\frac{m-2}{2} \int a^{\alpha\beta} D_\alpha w D_\beta w w^{s-3} w_L \eta^2 + \frac{1}{2} \int a^{\alpha\beta} D_\alpha w_L D_\beta w_L w^{s-2} \eta^2 \geq 0 \,. \tag{7.4.31}$$

The fourth integral in (7.4.30) is estimated by

$$\delta \int |D^2 f|^2 w^{s-2} w_L \eta^2 + \frac{c_{27}}{\delta} \int w^{s-1} w_L |D\eta|^2 \,, \tag{7.4.32}$$

for $\delta > 0$.

Choosing $\delta > 0$ small enough in (7.4.27), (7.4.28), (7.4.32), we obtain from (7.4.26) – (7.4.32) (recalling $0 \leq \eta \leq 1$, $|D\eta| \leq \frac{2}{R_2}$),

$$\begin{aligned}
\int_{B(x_1,R_2)} |D^2 f|^2 \eta^2 w^{s-2} w_L &\leq \frac{1}{\lambda} \int_{B(x_1,R_2)} a^{\alpha\beta} D_\alpha D_\gamma f D_\beta D_\gamma f w^{s-2} w_L \eta^2 \\
&\leq c_{28} \int_{B(x_1,R_2)} w^s w_L \eta^2 + c_{29} \Big(1 + \frac{1}{R_2^2}\Big) \int_{B(x_1,R_2)} w^s \,.
\end{aligned} \tag{7.4.33}$$

We then choose $\varepsilon > 0$ in (7.4.22) small enough (and thus determine $\overline{R}(\varepsilon)$) to obtain from (7.4.22) and (7.4.33)

$$\int_{B(x_1,\frac{R_2}{2})} (|D^2 f|^2 w^{s-2} w_L + w^s w_L) \leq c_{30} (1 + \frac{1}{R_2^2}) \int_{B(x_1,R_2)} w^s \,. \tag{7.4.34}$$

We may then let $L \to \infty$ in (7.4.34).

A covering argument then gives for every $R_1 < R_0$,

$$\int_{B(x_0,R_1)} (w^{s+1} + w^{s-1} |D^2 f|^2) < \infty. \tag{E_{s+1}}$$

This concludes the induction. □

We obtain

Lemma 7.4.5. *Let $f \in C^0 \cap H_l^{1,2}oc(\Omega, \mathbb{R}^n)$ be a solution of (7.4.1), with structural conditions (A1), (A2), (A3), (G1), (G2) satisfied, and furthermore $a^{\alpha\beta} \in C^2(\Omega)$ for all α, β.*

Then

$$f \in H^{3,2}_{\mathrm{loc}}(\Omega, \mathbb{R}^n).$$

Proof. From (G2),

$$\begin{aligned} \left|\frac{d}{dx} G(x, f(x), Df(x))\right| &= |G_x + G_f Df + G_p D^2 f| \\ &\le k_0 + k_1 |Df|^3 + k_2 |D^2 f| + k_3 |Df||D^2 f|\,, \end{aligned} \tag{7.4.35}$$

and this is in L^2 by Lemmas 7.4.4, 7.4.3.

Consequently, f is a weak solution of an equation

$$D_\beta(a^{\alpha\beta}(x) D_\alpha f) = g(x), \tag{7.4.36}$$

with $g \in H^{1,2}$.

The claim follows from Theorem A.2.1. $\square$

We can now prove

Theorem 7.4.1. *A continuous weakly harmonic map $f : M \to N$ between Riemannian manifolds is smooth.*

Proof. As explained before, by continuity, we may localize in domain and image, and we thus treat a continuous weakly harmonic map as a weak solution of the elliptic system

$$D_\alpha(\gamma^{\alpha\beta}\sqrt{\gamma} D_\beta f^i) = -\sqrt{\gamma}\gamma^{\alpha\beta}\Gamma^i_{jk}(f(x)) D_\alpha f^j D_\beta f^k =: k(x). \tag{7.4.37}$$

The structural conditions (A1) – (G2) then are satisfied.

Lemma 7.4.5 implies

$$f \in H^{3,2}_{\mathrm{loc}}\,.$$

Now

$$\begin{aligned} &|D^2(\sqrt{\gamma}\gamma^{\alpha\beta}\Gamma^i_{jk}(f) D_\alpha f^j D_\beta f^k)| \\ &\qquad \le \kappa_0 |Df|^2 + \kappa_1 |Df|^4 + \kappa_2 |D^2 f||Df|^2 + \kappa_3 |D^2 f|^2 + \kappa_4 |Df||D^3 f|\,. \end{aligned}$$

If $m := \dim M \le 3$ then Sobolev's embedding theorem (Theorem A.1.7) already implies that this is in L^2_{loc}. Hence, the right hand side k of (7.4.37) is in $H^{2,2}_{\mathrm{loc}}$ and by Theorem A.2.1,

$$f \in H^{4,2}_{\mathrm{loc}}\,.$$

In this manner, inductively

$$f \in H^{\nu,2}_{\mathrm{loc}} \Rightarrow k \in H^{\nu-1,2}_{\mathrm{loc}} \Rightarrow f \in H^{\nu+1,2}_{\mathrm{loc}}\,, \tag{7.4.38}$$

and Corollary A.1.2 implies $f \in C^\infty$.

If $m = \dim M$ is arbitrary, one either can apply more refined elliptic regularity results, or alternatively observe that Df satisfies a system with similar (actually, even better) structural conditions, and so the preceding results may be applied to Df instead of f. Iteratively, the same is true for higher derivatives of f, and thus one gets again

$$D^\nu f \in H^{3,2}_{\mathrm{loc}}$$

for all ν, i.e.

$$f \in H^{\ell,2}_{\mathrm{loc}}$$

for all ℓ, hence $f \in C^\infty$ by Corollary A.1.2. □

Perspectives. The regularity results and proofs of this paragraph are due Ladyzhenskaya and Ural'ceva[175] although this is usually not acknowledged in the western literature on harmonic maps. Their proof has been adapted to harmonic maps into spheres in [26].

7.5 Harmonic Maps into Manifolds of Nonpositive Sectional Curvature: Existence

Let M and N be compact Riemannian manifolds, N of nonpositive sectional curvature. In this section, we wish to show that any continuous map $g : M \to N$ is homotopic to some – essentially unique – harmonic map. This result will be deduced from convexity properties of the energy functional E that follow from the assumption that the target manifold N has nonpositive sectional curvature. The relevant geometric results have been collected in §4.8 already.

As an application in §7.7, we shall derive Preissmann's theorem about the fundamental group of compact manifolds of negative sectional curvature. Further applications will be described in the Perspectives.

A continuous map

$$g : M \to N$$

induces a homomorphism

$$\rho = g_\sharp : \pi_1(M, p) \to \pi_1(N, g(p))$$

of fundamental groups (p any point in M). As described in Appendix B, we may then find a lift

$$\tilde{g} : \tilde{M} \to \tilde{N}$$

to universal covers that is ρ-equivariant, i.e.

$$\tilde{g}(\lambda x) = \rho(\lambda)\tilde{g}(x)$$

for all $x \in \tilde{M}$, $\lambda \in \pi_1(M,p)$ where the fundamental groups $\pi_1(M,p)$ and $\pi_1(N, g(p))$ operate by deck transformations on $\tilde{M}$ and $\tilde{N}$, resp.

$Y := \tilde{N}$ is a simply connected complete Riemannian manifold of nonpositive sectional curvature. In particular, all the results derived in §4.8 for such manifolds apply. We let

$$d : Y \times Y \to \mathbb{R}$$

be the distance function induced by the Riemannian metric, as always.

For ρ-equivariant maps

$$h_1, h_2 : X := \tilde{M} \to Y,$$

we can define an L^2-distance by

$$d(h_1, h_2) := \left(\int d^2(h_1(x), h_2(x))\, d\mathrm{Vol}\,(M) \right)^{\frac{1}{2}},$$

where the integration is w.r.t. the volume form of the Riemannian metric on M and over some fundamental domain of M in $X = \tilde{M}$. The ρ-equivariance of h_1 and h_2 implies that this integral does not depend on the choice of fundamental domain.

We then put

$$Z := L^2_\rho(M,N) := \{h : X \to Y \ \rho\text{-equivariant with } d^2(h, \tilde{g}) < \infty\}.$$

$Z = L^2_\rho(M,N)$ then is a complete metric space; the completeness is shown as for the standard spaces of L^2-functions (that result is quoted in Theorem A.1.1), because Y is complete.

Curves in Z are simply given by families

$$(f_t)_{t \in [0,1]}$$

of ρ-equivariant maps $f_t : X \to Y$, depending continuously on t. We say that such a curve is a shortest geodesic if

$$d(f_0, f_t) = t d(f_0, f_1)$$

for all $t \in [0,1]$. (It is not difficult to show that this property characterizes shortest geodesics in Riemannian manifolds, and so it is natural to use this property also in other metric spaces.) It is then easy to describe such geodesics:

Lemma 7.5.1. *For every $x \in \tilde{M}$, let $\gamma_x : [0,1] \to \tilde{N}$ be a shortest geodesic with $\gamma_x(0) = f_0(x)$, $\gamma_x(1) = f_1(x)$, chosen equivariantly, i.e. $\rho(\lambda)\gamma_x = \gamma_{\lambda x}$ for all x, λ. Then the family of maps*

$$f_t(x) := \gamma_x(t), \quad t \in [0,1]$$

defines a shortest geodesic in $L^2_\rho(M,N)$ between f_0 and f_1.

Proof.

$$
\begin{aligned}
d^2(f_0, f_t) &= \int d^2(f_0(x), f_t(x))\, d\mathrm{Vol}\,(x) \\
&= \int t^2 d^2(f_0(x), f_1(x))\, d\mathrm{Vol}\,(x)
\end{aligned}
$$

because γ_x defines a shortest geodesic from $f_0(x)$ to $f_1(x)$

$$= t^2 d^2(f_0, f_1).$$

□

Thus, if f_0 and f_1 are ρ-equivariant maps, the geodesic in $L^2_\rho(M,N)$ from f_0 to f_1 is simply obtained by taking for each $x \in \tilde{M}$ the shortest geodesic from $f_0(x)$ to $f_1(x)$ and defining maps f_t through this family of geodesics.

Corollary 7.5.1. *Let $c_1, c_2 : [0,1] \to L^2_\rho(M,N)$ be shortest geodesics. Then*

$$d^2(c_1(t), c_2(t))$$

is a convex function of t.

Proof. We can use Lemma 7.5.1 to derive this property by integration from the corresponding property of $\tilde{N}$ that has been demonstrated in Theorem 4.8.2. Namely, by Theorem 4.8.2, for each $x \in \tilde{N}$, if $\gamma_{i,x}(t)$ is the shortest geodesic from $c_i(0)(x)$ to $c_i(1)(x)$, $i = 1, 2$, then

$$d^2(\gamma_{1,x}(t), \gamma_{2,x}(t))$$

is a convex function of t. But then also

$$d^2(c_1(t), c_2(t)) = \int d^2(\gamma_{1,x}(t), \gamma_{2,x}(t))\, d\mathrm{Vol}\,(M) \quad \text{by Lemma 7.5.1}$$

is a convex function of t. □

Similarly, if $c : [0,1] \to L^2_\rho(M,N)$ is a shortest geodesic, and $z \in L^2_\rho(M,N)$, we have the analogue of (4.8.7)

$$d^2(c(t), z) \le t d^2(c(1), z) + (1-t) d^2(c(0), z) - t(1-t) d^2(c(0), c(1)) \tag{7.5.1}$$

for all $t \in [0,1]$.

We now consider the functionals E_ε and E defined in §7.3, but this time, we define them on the space $L^2_\rho(M,N)$, carrying out all corresponding integrals on a fundamental domain for M in $\tilde{M}$.

Corollary 7.5.2. *E_ε and E are convex functionals on $L^2_\rho(M,N)$, in the sense that for any shortest geodesic $c:[0,1]\to L^2_\rho(M,N)$,*

$$E_\varepsilon(c(t)) \quad \text{and} \quad E(c(t))$$

are convex functions of t.

Proof. As explained in Lemma 7.5.1, such a shortest geodesic is given by a family of ρ-equivariant maps

$$f_t : \tilde{M} \to \tilde{N}$$

such that for each $x \in \tilde{M}$, $f_t(x)$ is geodesic w.r.t. t.

Applying Theorem 4.8.2 to the geodesics $f_t(x)$ and $f_t(y)$, we obtain

$$d^2(f_t(x), f_t(y)) \le t d^2(f_1(x), f_1(y)) + (1-t) d^2(f_0(x), f_0(y)).$$

Integrating this inequality w.r.t. x and y as in the definition of $E_\varepsilon(f)$ (cf. (7.3.3)) yields the convexity of E_ε, and the convexity of E follows by passing to the limit $\varepsilon \to 0$ as explained in §7.3. □

We are now ready to start our minimization scheme for the functionals E_ε and E on the space $Z = L^2_\rho(M,N)$. In fact, we shall demonstrate a general result about minimizing convex and lower semicontinuous functionals (recall Lemma 7.3.1 and Theorem 7.3.2) on Z; in fact, the constructions will be valid for more general spaces than Z as the only essential property that we shall use about Z is the convexity property of Corollary 7.5.1.

Definition 7.5.1. Let $F : Z \to \mathbb{R} \cup \{\infty\}$ be a function. For $\lambda > 0$, $z \in Z$, the Moreau-Yosida approximation F^λ of F is defined as

$$F^\lambda(z) := \inf_{y \in Z} (\lambda F(y) + d^2(y,z)).$$

Lemma 7.5.2. *Let $F : Z \to \mathbb{R} \cup \{\infty\}$ be convex, lower semicontinuous, $\not\equiv \infty$ and bounded from below. For every $\lambda > 0$, $z \in Z$, there exists a unique $y_\lambda \in Z$ with*

$$F^\lambda(z) = \lambda F(y_\lambda) + d^2(y_\lambda, z). \tag{7.5.2}$$

Proof. We take a minimizing sequence $(y_n)_{n\in\mathbb{N}}$ for $F^\lambda(z)$. This means that

$$\lim_{n\to\infty} (\lambda F(y_n) + d^2(y_n, z)) = F^\lambda = \inf_{y\in Z} (\lambda F(y) + d^2(y,z)). \tag{7.5.3}$$

For $y_m, y_n \in Z$, we take a shortest geodesic $\gamma : [0,1] \to Z$ with

$$\gamma(0) = y_m, \quad \gamma(1) = y_n$$

and define the midpoint as

$$y_{m,n} = \gamma\Big(\frac{1}{2}\Big).$$

The convexity of F then implies

$$\begin{aligned} F^\lambda &\le \lambda F(y_{m,n}) + d^2(y_{m,n}, z) \\ &\le \frac{1}{2}\lambda F(y_m) + \frac{1}{2}\lambda F(y_n) + d^2(y_{m,n}, z) \quad \text{by convexity of } F \\ &\le \frac{1}{2}\lambda F(y_m) + \frac{1}{2}\lambda F(y_n) + \frac{1}{2}d^2(y_m, z) + \frac{1}{2}d^2(y_n, z) - \frac{1}{4}d^2(y_m, y_n) \qquad (7.5.4) \\ &\text{by (7.5.1).} \end{aligned}$$

Since, by (7.5.3) $(\lambda F(y_m) + d^2(y_m, z))$ and $(\lambda F(y_n) + d^2(y_n, z))$ converge to F^λ, we conclude that $d^2(y_m, y_n)$ has to tend to 0 as $m, n \to \infty$. Thus $(y_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in Z, and so it tends towards some limit y_λ. Because F is lower semicontinuous, (7.5.2) then follows from (7.5.3). □

We can now state our abstract existence result:

Theorem 7.5.1. *Let $F : Z \to \mathbb{R} \cup \{\infty\}$ be a convex, lower semicontinuous function that is bounded from below and not identically $+\infty$. Let y_λ be as constructed in Lemma 7.5.2 for $\lambda > 0$.*

If $(y_{\lambda_n})_{n\in\mathbb{N}}$ is bounded for some sequence $\lambda_n \to \infty$, then $(y_\lambda)_{\lambda>0}$ converges to a minimizer of F as $\lambda \to \infty$.

Proof. Take any $z \in Z$. By definition of y_{λ_n}, y_{λ_n} minimizes $F(y) + \frac{1}{\lambda_n}d^2(y, z)$. Since y_{λ_n} is bounded and λ_n tends to ∞, $(y_{\lambda_n})_{n\in\mathbb{N}}$ therefore constitutes a minimizing sequence for F. We claim that $d^2(y_\lambda, z)$ is a nondecreasing function of λ. To see this, let $0 < \mu < \lambda$.

By definition of y_μ,

$$F(y_\lambda) + \frac{1}{\mu}d^2(y_\lambda, z) \ge F(y_\mu) + \frac{1}{\mu}d^2(y_\mu, z).$$

This implies

$$F(y_\lambda) + \frac{1}{\lambda}d^2(y_\lambda, z) \ge F(y_\mu) + \frac{1}{\lambda}d^2(y_\mu, z) + \Big(\frac{1}{\mu} - \frac{1}{\lambda}\Big)(d^2(y_\mu, z) - d^2(y_\lambda, z)).$$

This is compatible with the definition of y_λ only if

$$d^2(y_\mu, z) \le d(y_\lambda, z)$$

showing the claimed monotonicity property of $d^2(y_\lambda, z)$.

Since $d^2(y_\lambda, z)$ is bounded on the sequence $(\lambda_n)_{n\in\mathbb{N}}$ tending to ∞ and monotonic, it has to be a bounded function of $\lambda > 0$. It follows from the definition of y_λ that

$$F(y_\lambda) = \inf\{F(y) : d^2(y, z) \le d^2(y_\lambda, z)\}.$$

Since $d^2(y_\lambda, z)$ is nondecreasing, this implies that $F(y_\lambda)$ is a nonincreasing function of λ, and as noted in the beginning, it tends to $\inf_{y\in Z} F(y)$ for $\lambda \to \infty$. Let now

$\varepsilon > 0$. By the preceding boundedness and monotonicity results, we may find $\Lambda > 0$ such that for $\lambda, \mu > \Lambda$

$$|d^2(y_\lambda, z) - d^2(y_\mu, z)| < \frac{\varepsilon}{2}. \tag{7.5.5}$$

If $\Lambda < \mu \leq \lambda$, we have $F(y_\mu) \geq F(y_\lambda)$ as $F(y_\lambda)$ is nonincreasing. If $y_{\mu,\lambda}$ is the midpoint of y_μ and y_λ as in the proof of Lemma 7.5.2, we obtain from the definition of y_μ

$$\begin{aligned} F(y_\mu) + \frac{1}{\mu} d^2(y_\mu, z) &\leq F(y_{\lambda,\mu}) + \frac{1}{\mu} d^2(y_{\lambda,\mu}, z) \\ &\leq F(y_{\lambda,\mu}) + \frac{1}{\mu}\Big(d^2(y_\mu, z) + \frac{\varepsilon}{4} - \frac{1}{2} d^2(y_\lambda, y_\mu)\Big) \end{aligned}$$

by (7.5.1) and (7.5.5).

Also, by convexity of F, and since $F(y_\mu) \geq F(y_\lambda)$, $F(y_{\lambda,\mu}) \leq F(y_\mu)$. Therefore,

$$d(y_\lambda, y_\mu) < \varepsilon.$$

Thus, $(y_\lambda)_{\lambda > 0}$ is a Cauchy family for $\lambda \to \infty$.

Since Z is complete, there then exists a unique $y_\infty = \lim_{\lambda\to\infty} y_\lambda$. Since we have already seen that

$$\lim_{\lambda\to\infty} F(y_\lambda) = \inf_{y\in Z} F(y),$$

the lower semicontinuity of F implies that

$$F(y_\infty) = \inf_{y\in Z} F(y).$$

□

In order to apply Theorem 7.5.1 to show the existence of a minimizer of E (or, by the same argument, for the functionals E_ε), we need to verify that in our situation the y_λ in the statement of Theorem 7.5.1 remain bounded. This is the content of the proof of

Theorem 7.5.2. *Let M and N be compact Riemannian manifolds, N of nonpositive sectional curvature. Then every continuous map $g : M \to N$ is homotopic to a minimizer f of the energy E, in the sense that E achieves its minimum in the class $L^2_\rho(M,N)$ of ρ-equivariant maps between the universal covers $\tilde{M}$ and $\tilde{N}$, where $\rho : \pi_1(M) \to \pi_1(N)$ is the homeomorphism of fundamental groups induced by g. (We shall verify subsequently that f is smooth, and so in particular continuous.)*

Proof. We first consider the case where $g(M)$ is simply connected. It is not difficult to verify that in that case, g is homotopic to a constant map (and a constant map obviously minimizes the energy). Since that verification is instructive for the general strategy, we proceed to perform it. Let $y_0 \in g(M)$. For each $y \in g(M)$, we choose a curve γ_y from y_0 to y. Let $c_y : [0,1] \to N$ be the geodesic from y_0 to y homotopic to

γ_y. It is unique because N has nonpositive sectional curvature (Theorem 4.8.1), and it does not depend on the choice of γ_y, because any two curves in $g(M)$ from y_0 to y are homotopic to each other as $g(M)$ is simply connected. We put

$$g_t(x) = c_{g(x)}(t).$$

$g_t(x)$ is continuous w.r.t. t, and also w.r.t. x, because

$$d^2(c_{y_1}(t), c_{y_2}(t)) \leq t d^2(c_{y_1}(1), c_{y_2}(1)) \quad \text{by Corollary 4.8.3.}$$

Since $g_0 \equiv y_0$, $g_1 = g$, g_t provides a homotopy between a constant map and g, as desired.

If $g(M)$ is not simply connected, we choose some closed curve γ in $g(M)$ that is not homotopically trivial. Let c be a closed geodesic in N that is homotopic to γ (Theorem 1.5.1). Let $\tilde{g} \in L^2_\rho(M, N)$ be the lift of g to universal covers. In order to apply Theorem 7.5.1, we have to exclude that the $L^2_\rho(M, N)$-maps y_λ constructed for $z = \tilde{g}$ in Lemma 7.5.2 become unbounded, i.e. that the L^2_ρ-distance between $\tilde{g}$ and y_λ becomes unbounded for $\lambda \to \infty$. y_λ projects to a map $g_\lambda : M \to N$ homotopic to g. Let γ_λ be a closed curve in $g_\lambda(M)$ that is homotopic to γ. Let $x \in M$ with $g(x) \in \gamma$, and $y_0 \in c$. Let $c_\lambda : [0,1] \to N$ be the geodesic from y_0 to $g_\lambda(x)$ in the homotopy class determined by a homotopy between g and g_λ. Let b_λ be the geodesic loop (which exists by Theorem 1.5.1) from $g_\lambda(x)$ to itself that is homotopic to $c_\lambda c c_\lambda^{-1}$. Thus, $b_\lambda^{-1} c_\lambda c c_\lambda^{-1}$ is homotopic to a constant curve. Likewise, let $b_{\lambda,t}$ be the geodesic loop based at $c_\lambda(t)$ homotopic to $c_{\lambda|[0,t]} c (c_{\lambda|[0,t]})^{-1}$. By lifting to the universal cover $\tilde{N}$, we see that the energy $E(b_{\lambda,t})$ becomes the squared distance between two different lifts of c_λ, i.e. two geodesics, and so it is convex by Theorem 4.8.2. Since $c = b_{\lambda,0}$ is a shortest geodesic, $E(b_{\lambda,t})$ is minimal at $t = 0$. Thus, assuming $d^2(g, g_\lambda) \to \infty$ for $\lambda \to \infty$, $E(b_{\lambda,t})$ either tends to a constant function, or $E(b_{\lambda,1})$ goes to ∞. In the latter case, however, the lengths of all curves in $g_\lambda(M)$ homotopic in N to c would also go to ∞, and that would let the energy of g_λ tend to ∞ as well, in contradiction to g_λ being a minimizing family for $\lambda \to \infty$ by the proof of Theorem 7.5.1.

If the lengths are constant, i.e. $b_{\lambda,1}$ is asymptotically of the same length as $b_{\lambda,0} = c$, we either find another homotopy class of curves for which the length goes to ∞ – which is impossible as already argued – or the length remains constant for all homotopy classes. In that case, however, the construction of the Moreau-Yosida approximation implies that $d^2(g, g_\lambda)$ cannot tend to ∞, because E is not changed, while $d^2(g, g_\lambda)$ is decreased if we move the image of M closer to c along the curves c_λ ("closer" here refers to the lifts to the universal cover $\tilde{N}$), i.e. replacing x by $c_\lambda(t)$ for $t < 1$.

Thus, in any case, $d^2(g, g_\lambda)$ stays bounded, and Theorem 7.5.1 yields the result after all. □

Perspectives. See the Perspectives on §7.7.

7.6 Harmonic Maps into Manifolds of Nonpositive Sectional Curvature: Regularity

In the preceding section, we have shown the existence of a minimizer of the energy functional E in a given homotopy class, or more precisely, in the class of L^2-maps that induce the same action by deck transformations on the universal covers as some given continuous map g. It is the purpose of this section to show the regularity, i.e. the smoothness of such a minimizer. In fact, we shall present different regularity proofs with the purpose of showing a more representative sample of techniques from geometric analysis.

It is clear that a minimizer f of E is a critical point of E in the sense of Definition 7.2.1. Namely, in §7.1, we have computed that for a compactly supported vector field ψ along f and $f_t(x) = \exp_{f(x)} t\psi(x)$,

$$\frac{d}{dt} E(f_t)_{|t=0} = \int \langle df, d\psi \rangle.$$

Thus, in particular, $E(f_t)$ is a differentiable function of t, and since $f = f_0$ minimizes E, this derivative at $t = 0$ has to vanish, for all such ψ.

If k is some smooth function on the image of f, and if φ is a smooth function on M with compact support, we may consider the test vector

$$(dk) \circ f(x) \cdot \varphi(x).$$

We obtain (referring to §7.1 for the notation)

$$\begin{aligned} 0 = \int \langle df, d\psi \rangle &= \int \langle df, d(dk)\varphi(x) \rangle \qquad \text{with } dk \text{ being evaluated at } f(x) \\ &= \int \varphi(x) \langle df, \nabla_{\frac{\partial}{\partial x^\alpha}} (dk) \otimes dx^\alpha \rangle + \int \langle d\varphi(x), d(k \circ f)(x) \rangle \\ &= \int \varphi(x) \nabla dk(df, df)(x) + \int \langle d\varphi(x), d(k \circ f)(x) \rangle \ , \end{aligned} \tag{7.6.1}$$

recalling (7.1.11) and (3.3.48).
We now take

$$k(z) = \frac{1}{2} d^2(z, p),$$

lifting to universal covers as always.

By Lemma 4.8.2,

$$\nabla dk(df, df) \geq \|df\|^2.$$

Inserting this into (7.6.1) yields

$$\int \langle d\varphi(x), d(k \circ f)(x) \rangle \leq - \int \varphi(x) \|df(x)\|^2. \tag{7.6.2}$$

(7.6.2) means that $k \circ f$ is a weak subsolution of

$$-\Delta(k \circ f) \geq \|df\|^2. \tag{7.6.3}$$

(Cf. Corollary 7.2.5 for the corresponding result in the case where f is a smooth harmonic map.)

We shall now use this differential inequality to derive the Hölder continuity of our minimizer f. Theorem 7.4.1 will then imply that f is smooth.

The same argument actually shows that for any smooth convex function k on the image of f, we have

$$-\Delta(k \circ f) \geq 0. \tag{7.6.4}$$

In the sequel, however, the functions $k(z) = \frac{1}{2}d^2(z,p)$, for various choices of p, will entirely suffice.

We shall need a version of the Poincaré inequality

Lemma 7.6.1. *Let M be a compact Riemannian manifold, Y the universal covering of a Riemannian manifold of nonpositive curvature. Then there exist $r_0 > 0$ and a constant $c_0 < \infty$ such that for any ball $B(x_0, r) \subset M$, $0 < r \leq r_0$, and any L^2-map with finite energy,*

$$f : B(x_0, r) \to Y,$$

the following inequality holds

$$\int_{B(x_0,r)} d^2(f(x), f_B) \leq c_0 r^2 \int_{B(x_0,r)} \|df(x)\|^2, \tag{7.6.5}$$

where $f_B \in Y$ is the center of mass of f, i.e. f_B minimizes

$$\int_{B(x_0,r)} d^2(f(x), p)\, d\mathrm{Vol}\,(x) \quad \textit{w.r.t. } p \in Y.$$

Proof. The factor r^2 on the right hand side of (7.6.5) comes from a simple scaling argument; such a scaling argument is possible because for sufficiently small $r > 0$, the geometry of the ball deviates to an arbitrary little degree from the one of a Euclidean ball of the same radius. Thus, we neglect the factor r^2 in the sequel.

If the inequality (7.6.5) then is not valid, we can find a sequence $(f_n)_{n \in \mathbb{N}}$ of maps from some such ball $B(x_0, r)$ into Y for which

$$\int d^2(f_n(x), f_{n,B}) \geq n \int \|df_n(x)\|^2. \tag{7.6.6}$$

Since Y has a compact quotient, we may compose f_n with deck transformations, i.e. isometries of Y, which leave both sides of (7.6.6) invariant, such that $f_{n,B}$ always stays in some compact region of Y. Thus, we may assume that the $f_{n,B}$ converge to some $p \in Y$. If the left hand side of (7.6.6) happens to be smaller than one, we may rescale Y, i.e. we consider the chart

$$\exp_p : T_pY \to Y$$

and replace the Riemannian metric $g_{ij}(z)$ of Y in this chart by the metric $g_{ij}(\rho z)$ for a suitable $\rho \geq 1$. This multiplies the distance function d and the norm $\|\cdot\|$ by a factor ρ which we can thus adjust to make the left hand side of (7.6.6) equal to 1. The curvature of Y gets multiplied by $\frac{1}{\rho^2}$, and as $\rho \to \infty$, the rescaled Riemannian manifold $(Y, g_{ij}(\rho z))$ becomes Euclidean, and the Poincaré inequality reduces to the Euclidean one.

We now turn to the case where the left hand side of (7.6.6) is bigger than 1.

For any map $g : B(x_0, r) \to Y$, we may perform the following construction:

$$g_t(x) := \exp t(\exp_{g_B}^{-1} g(x)) \quad \text{for } 0 \leq t \leq 1.$$

Thus, for any x,

$$d(g_t(x), g_B) = td(g(x), g_B), \tag{7.6.7}$$

and since g_B is characterized by the property that

$$\int \exp_{g_B}^{-1}(g(x))\, d\mathrm{Vol}\,(x) = 0,$$

we see that

$$g_B = g_{t,B},$$

i.e. g_B remains the center of mass for the maps g_t.

Since Y has nonpositive curvature

$$d(g_t(x), g_t(y)) \leq td(g(x), g(y)) \quad \text{for all } x, y, 0 \leq t \leq 1, \tag{7.6.8}$$

by (4.8.8). Therefore also

$$\|dg_t(x)\|^2 \leq t^2 \|dg(x)\|^2, \tag{7.6.9}$$

whenever this expression is well defined.

For each $n \in \mathbb{N}$ for which the left hand side of (7.6.6) should happen to be bigger than one, we choose $t = t_n$, $0 \leq t \leq 1$, such that

$$\int d^2(f_{n,t}(x), f_{n,t,B}) = 1.$$

Because of (7.6.7) and (7.6.9), we may then replace f_n by $f_{n,t}$ without making (7.6.6) invalid, and so, we may assume w.l.o.g.

$$\int d^2(f_n(x), f_{n,B}) = 1 \quad \text{for all } n \in \mathbb{N}. \tag{7.6.10}$$

Then

$$\int \|df_n(x)\|^2 \to 0 \quad \text{for } n \to \infty,$$

and therefore f_n has to converge to a constant map $f_0 \equiv p$ for some $p \in Y$. By Rellich's theorem (see Theorem A.1.8; the standard proof for functions, see, e.g. J. Jost, Postmodern Analysis, Springer, 1998, p. 265 ff., carries over to maps with values in Y, because we have constructed in §4.8 the mollifiers on which that proof depends)

$$\int d^2(f_n(x), f_{n,B})$$

converges to

$$\int d^2(f_0(x), f_{0,B}) = \int d^2(p,p) = 0.$$

This, however, contradicts (7.6.10). This concludes the proof. □

Let us also present an alternative **proof** of the Poincaré inequality that does not use Rellich's theorem, but rather employs the constructions of §4.8 directly:

Proof. By (4.8.21),

$$d(f(x), f_B) \leq \int_{B(x_0,r)} d(f(x), f(y))\, dy.$$

We may work with the Euclidean volume form on dy on $B(x_0,r)$ induced by the exponential map $\exp_{x_0} : T_{x_0}M \to M$, rather than with the Riemannian one. Since the two are uniformly equivalent, this will only affect the constant c_0 in the estimate. In other words, we assume that $B(x_0,r)$ is a Euclidean ball

$$\{y \in \mathbb{R}^m : d(x_0,y) = |x_0 - y| < r\}.$$

We may also assume that f is differentiable, because a general f may be approximated by the differentiable mollified maps f_h as explained in §4.8.

Then

$$d(f(x), f(y)) \leq \int_0^{|x-y|} \Big\| \frac{\partial}{\partial r} f\Big(x + r\frac{y-x}{|y-x|}\Big)\Big\| dr$$

(the meaning of $\frac{\partial f}{\partial r}$ should be obvious), and so

$$\int d(f(x), f(y))\, dy \leq \frac{1}{m\omega_m} \int \frac{1}{|x-y|^{m-1}} \|df(y)\|\, dy,$$

for $m = \dim M$, $\omega_m =$ volume of the m-dimensional unit sphere.

Therefore,

$$\begin{aligned}\int d^2(f(x), f_B)\,dx &\leq \int\left(\int d(f(x), f(y))\,dy\right)^2 dx\\ &\leq \frac{1}{m^2\omega_m^2}\int\left(\int \frac{1}{|x-y|^{m-1}}\|df(y)\|\,dy\right)^2 dx\\ &\leq \frac{1}{m^2\omega_m^2}\int\left(\int \frac{1}{|x-y|^{m-1}}\|df(y)\|^2 dy\right)\left(\int \frac{1}{|x-y|^{m-1}}\,dy\right)dx\end{aligned}$$

by Hölder's inequality (Theorem A.1.2),

$$= \frac{1}{m^2\omega_m^2}\int \|df(y)\|^2\left(\int \frac{1}{|x-y|^{m-1}}\,dx\right)^2 dy$$

by Fubini's theorem.

Since

$$\int_{B(x_0,r)} \frac{1}{|x-y|^{m-1}}\,dx \leq m\omega_m r \quad \text{for all } y \in B(x_0, r),$$

we obtain

$$\int_{B(x_0,r)} d^2(f(x), f_B)\,dx \leq r^2 \int_{B(x_0,r)} \|df(x)\|^2 dx$$

and the constant c_0 arises from estimating the Euclidean volume dx against the Riemannian volume $d\mathrm{Vol}\,(x)$. In fact, employing Riemannian normal coordinates at x_0, we see that this yields a factor of magnitude $(1 + c_1 r^2)$. □

In the sequel, we shall assume that the radii R of all balls $B(x_0, R), x_0 \in M$, are smaller than the injectivity radius of M. We then do not need to distinguish between such a ball and its lift to the universal cover $\tilde{M}$. Also, on such a ball, the negative Laplace-Beltrami operator in local coordinates,

$$-\Delta = \frac{1}{\sqrt{\gamma}}\frac{\partial}{\partial x^\alpha}\left(\sqrt{\gamma}\gamma^{\alpha\beta}\frac{\partial}{\partial x^\beta}\right) \quad \text{(notations as in Section 7.1)}$$

is of the type considered in §A.2, and therefore on such a ball, the Harnack inequalities stated in Theorem A.2.2 hold.

By the Harnack inequality (Theorem A.2.2 (i)), we have for $x_0 \in M$, $p \in \tilde{N}$, $m = \dim M$,

$$\sup_{B(x_0,r)} d^2(f(x), p) \leq c_2\left(\frac{1}{r^m}\int_{B(x_0,2r)} d^{2q}(f(x), p)d\mathrm{Vol}\,(x)\right)^{\frac{1}{q}} \quad \text{for } q > 1, \qquad (7.6.11)$$

because of the inequality

$$-\Delta d^2(f(x),p) \geq 0 \tag{7.6.12}$$

that follows from (7.6.3).

In order to control the right hand side of (7.6.11), we observe that we can control

$$d(p, f_B)$$

where f_B is the center of mass of f on $B(x_0, 2r)$, because f is in L^2. We therefore need to estimate

$$\int_{B(x_0,2r)} d^{2q}(f(x), f_B) d\mathrm{Vol}\,(x).$$

As in the second proof of the Poincaré inequality, we have (replacing again $d\mathrm{Vol}\,(x)$ by the Euclidean volume element dx)

$$\begin{aligned}
\int d^{2q}(f(x), f_B)\, dx &\leq \int \Big(\int d(f(x), f(y))\, dy \Big)^{2q} dx \\
&\leq \frac{1}{m^{2q}\omega_m^{2q}} \int \Big(\int \frac{1}{|x-y|^{m-1}} \|df(y)\|\, dy \Big)^{2q} dx \\
&\leq \frac{1}{m^{2q}\omega_m^{2q}} \cdot
\end{aligned}$$

$$\int \Big(\int \frac{1}{|x-y|^{(m-1)\frac{2q}{1+q}}} \|df(y)\|^2 dy \Big) \Big(\int \frac{1}{|x-y|^{(m-1)\frac{2q}{1+q}}}\, dy \Big)^{q} \Big(\int \|df(y)\|^2 dy \Big)^{q-1} dx$$

by Hölder's inequality (Theorem A.1.2) with exponents $p_1 = 2$, $p_2 = 2q$, $p_3 = \frac{2q}{q-1}$ and writing

$$\begin{aligned}
&\frac{1}{|x-y|^{m-1}} \|df\| \\
&\quad = \Big\{ \Big(\frac{1}{|x-y|^{m-1}} \Big)^{\frac{q}{1+q}} \Big\} \Big\{ \Big(\frac{1}{|x-y|^{m-1}} \Big)^{\frac{1}{1+q}} \|df\|^{\frac{1}{q}} \Big\} \Big\{ \Big\| df \Big\|^{1-\frac{1}{q}} \Big\} \\
&\quad = \frac{1}{m^{2q}\omega_m^{2q}} \Big(\int \frac{1}{|x-y|^{(m-1)\frac{2q}{1+q}}}\, dy \Big)^{q+1} \Big(\int \|df(y)\|^2 dy \Big)^{q}
\end{aligned}$$

by Fubini's theorem as in the second proof the Poincaré inequality.

Now

$$\int \frac{1}{|x-y|^{(m-1)\frac{2q}{1+q}}}\, dy < \infty \quad \text{if} \quad \frac{2q}{1+q} < \frac{m}{m-1}\,,$$

and if we choose $q > 1$ satisfying that condition, we can bound $d^2(f(x), p)$ by $d^2(p, f_B)$ and

$$\int_{B(x_0,2r)} \|df(y)\|^2 dy.$$

(The first proof of the Poincaré inequality given above can also be strengthened to yield the present stronger conclusion, by making use of Kondrachov's extension of Rellich's theorem, see Theorem A.1.8.)

In particular, $d^2(f(x),p)$ is bounded on $B(x_0,r)$, since f has finite energy. We record this as

Lemma 7.6.2. *Let $f : B(x_0,4r) \to Y$ (complete, simply connected, nonpositive sectional curvature) be a map of finite energy, satisfying*

$$-\Delta d^2(f(x),p) \geq 0 \quad \textit{weakly for all } p \in Y.$$

Then f is bounded on $B(x_0,r)$. □

Lemma 7.6.3. *Let $f : B(x_0,4r) \to Y$ satisfy $-\Delta d^2(f(x),p) \geq 0$ weakly for every $p \in Y$, where $B(x_0,4r)$ is a ball in some Riemannian manifold M, $0 < 2r < i(M)$ and Y is a manifold of nonpositive sectional curvature, the universal cover of a compact manifold N. Let $0 < \kappa_1 \leq \kappa \leq \kappa_0$, and suppose that*

$$\operatorname{diam} f(B(x_0,2r)) := \sup_{\substack{x_1,x_2\in \\ B(x_0,2r)}} d(f(x_1),f(x_2)) = \kappa.$$

There exists $\varepsilon > 0$ depending on the geometry of M and N and on κ_0 and κ_1 with the property that if $0 < \varepsilon \leq \varepsilon_0$ and

$$f(B(x_0,2r))$$

is covered by k balls $B_1,\ldots,B_k$ of radius ε, then

$$f(B(x_0,r))$$

can be covered already by $k-1$ of those balls.

Proof. Since we may obviously assume that each ball B_i contains some point $f(x_i)$ we have

$$B_i \subset B(p_i,2\varepsilon), \quad \text{with } p_i = f(x_i), i = 1,\ldots,k. \tag{7.6.13}$$

If we assume $\varepsilon \leq \varepsilon_0 \leq \frac{\kappa}{16}$, the balls

$$B\Big(p_i, \frac{\kappa}{8}\Big), \quad i = 1,\ldots,\kappa,$$

cover $f(B(x_0,2r))$. Since its diameter is κ, $f(B(x_0,2r))$ is contained in some ball of radius at most 2κ. Because the geometry of Y is uniformly controlled as Y admits a compact quotient,[3] there is some integer k_1 such that any such ball of radius $\leq 2\kappa \leq$

[3]Actually, what is needed at this point is solely a lower bound on the Ricci curvature of Y, combined with the assumption that Y has nonpositive sectional curvature, but we do not pursue this issue here.

$2\kappa_0$ contains at most k_1 points whose mutual distance is always at least $\frac{\kappa}{8}$. Therefore, already k_1 of the balls $B(p_i, \frac{\kappa}{4})$ cover $f(B(x_0, 2r))$, say for $i = 1, \ldots, k_1$.

Therefore, for at least one of those p_i, say for p_1,

$$\begin{aligned} \operatorname{meas}\Big(f^{-1}\Big(B\Big(p_1, \frac{\kappa}{4}\Big)\Big) \cap B(x_0, r)\Big) &\geq \frac{1}{k_1} \operatorname{meas}(B(x_0, r)) \\ &\geq \frac{\eta}{k_1} r^m \end{aligned} \tag{7.6.14}$$

for some constant $\eta > 0$ depending on the geometry of M.[4]

We consider the auxiliary function

$$g(x) := \frac{1}{\kappa^2} d^2(p_1, f(x)).$$

We put

$$\mu := \sup_{x \in B(x_0, 2r)} g(x) \leq \frac{1}{\kappa^2} (\operatorname{diam}(f(B(x_0, r)))) \leq 1. \tag{7.6.15}$$

By the triangle inequality, and since $\operatorname{diam}(f(B(x_0, 2r))) = \kappa$, there also has to exist some $y \in B(x_0, 2r)$ with

$$d(f(y), p_1) \geq \frac{\kappa}{2},$$

hence

$$\mu \geq \frac{1}{4}.$$

On $f^{-1}(B(p_1, \frac{\kappa}{4}))$, we have

$$g(x) \leq \frac{1}{16}.$$

We consider the auxiliary function

$$h(x) := \mu - g(x) \geq 0 \quad \text{on } B(x_0, 2r), \tag{7.6.16}$$

and

$$h(x) \geq \frac{1}{8} \qquad\qquad \text{on } f^{-1}\Big(B\Big(p_1, \frac{\kappa}{4}\Big)\Big). \tag{7.6.17}$$

By (7.6.12) and the definition of g and h, we also have

$$-\Delta h(x) \leq 0 \quad \text{weakly in } B(x_0, 2r).$$

Because of (7.6.17), we may apply the Harnack inequality Theorem A.2.2 (ii) to obtain

$$\begin{aligned} \inf_{B(x_0, r)} h(x) &\geq \delta_0 \frac{1}{r^m} \int_{B(x_0, r)} h(x)\, dx \quad \text{for some } \delta_0 > 0 \\ &\geq \delta \qquad \text{for some } \delta > 0, \quad \text{by (7.6.17), (7.6.14).} \end{aligned} \tag{7.6.18}$$

[4] η is controlled from below by an upper bound for the sectional curvature of M, but again this is not pursued here.

This inequality now implies that for sufficiently small ε, we cannot have

$$f(B(x_0, r)) \cap B(p_i, 2\varepsilon) \neq \emptyset \quad \text{for all } i = 1, \ldots, k. \tag{7.6.19}$$

Namely, the balls $B(p_i, 2\varepsilon)$ cover $f(B(x_0, 2r))$, and thus, if the supremum is realized in (7.6.15) for $y \in B(x_0, 2r)$, i.e.

$$\frac{1}{\kappa^2} d^2(p_i, f(y)) = \mu,$$

we can find some p_i with

$$d(p_1, f(y)) \leq 2\varepsilon.$$

So, if (7.6.19) held, we would have $d(f(x_1), f(y)) \leq 4\varepsilon$ for some $x_1 \in B(x_0, r)$, and thus

$$\begin{aligned}\inf_{B(x_0,r)} h(x) \leq h(x_1) &= \mu - \frac{1}{\kappa^2} d^2(p_1, f(x_1)) \\ &\leq 16 \frac{\varepsilon\sqrt{\mu}}{\kappa},\end{aligned}$$

which contradicts (7.6.18) for

$$\varepsilon < \frac{\delta\kappa_1}{16}.$$

Thus, for such an ε, $f(B(x_0, r))$ is disjoint to one of the balls $B(p_i, 2\varepsilon)$, hence also to one of the balls B_i, because of (7.6.13). Thus, it can be covered by the remaining ones. $\square$

Equipped with the preceding Lemma, we may now prove

Theorem 7.6.1. *Let $B(x_1, 12r)$ be a ball in some Riemannian manifold, $0 < 12r < i(M)$, Y the universal cover of a compact Riemannian manifold of nonpositive sectional curvature (and thus complete, simply connected, and nonpositively curved itself), and let*

$$f : B(x_1, 12r) \to Y$$

satisfy

$$E(f) < \infty$$

and

$$-\Delta d^2(f(x), p) \geq 0$$

weakly for every $p \in Y$.

Then f is continuous on $B(x_1, r)$.

Here, with the notation of §7.1 for the metric on the domain M, Δ is the Laplace-Beltrami operator

$$-\frac{1}{\sqrt{\gamma}}\frac{\partial}{\partial x^\alpha}\Big(\sqrt{\gamma}\gamma^{\alpha\beta}\frac{\partial}{\partial x^\beta}\Big).$$

Proof. By Lemma 7.6.2, f is bounded on $B(x_1, 3r)$, hence on $B(x_0, 2r)$ for every $x_0 \in B(x_1, r)$.

Thus,

$$\operatorname{diam} f(B(x_0, 2r)) \leq \kappa_0 \tag{7.6.20}$$

for some $\kappa_0 < \infty$. Let now $0 < \kappa_1 < \kappa_0$. We want to find some $\rho > 0$ with

$$\operatorname{diam} f(B(x_0, \rho)) < \kappa_1. \tag{7.6.21}$$

Let $\varepsilon_0 = \varepsilon_0(\kappa_0, \kappa_1)$ be as in Lemma 7.6.3.

Because of (7.6.20), we can bound the number k_0 of balls $B_1, \dots, B_{k_0}$ of radius ε_0 in Y that are needed to cover $f(B(x_0, 2r))$. By Lemma 7.6.3, $f(B(x_0, r))$ can be covered by at most $k_0 - 1$ of them. If

$$\operatorname{diam} f(B(x_0, r)) \geq \kappa_1,$$

we may apply Lemma 7.6.3 again with $\frac{r}{2}$ in place of r and $k = k_0 - 1$ and cover $f(B(x_0, \frac{r}{2}))$ by at most $k_0 - 2$ balls. We can repeat this construction until, for some $\nu \in \mathbb{R}$,

$$f(B(x_0, 2^{-\nu} r))$$

is covered by so few balls of radius ε_0 that we must have

$$\operatorname{diam} f(B(x_0, 2^{-\nu} r)) < \kappa_1.$$

Since this holds for every $x_0 \in B(x_1, r)$ and every $\kappa_1 > 0$, we see that f is continuous on $B(x_1, r)$. □

We shall now present an alternative (and more general) derivation of Theorem 7.6.1 not based on Lemma 7.6.3. Of course, the Harnack inequality will again be used in a crucial manner. The geometry of the domain M will only enter through the Poincaré inequality (Lemma 7.6.1) (which implies the Harnack inequality) and the following ball doubling property for the volume form:

$$\operatorname{Vol}(B(x, 2r)) \leq c_0 \operatorname{Vol}(B(x, r)) \tag{7.6.22}$$

for some constant c_0, all $x \in M$, and all sufficiently small radii $r > 0$.

We shall make use of the following abbreviations:

For $v \in L^\infty(B(x_0, R))$:

$$v_{+,R} := \sup_{B(x_0,R)} v,$$
$$v_{-,R} := \inf_{B(x_0,R)} v,$$
$$v_R := \fint_{B(x_0,R)} v\mu(dx)$$

(as always, sup and inf are the essential supremum and infimum).

Lemma 7.6.4. *Let v be a bounded weak subsolution* $(-\Delta v \geq 0)$ *on* $B(x_0, 4R)$. *There exists a constant δ_0, independent of v and R, with*

$$v_{+,R} \leq (1-\delta_0)v_{+,4R} + \delta_0 v_R.$$

Proof.

$$\begin{aligned} v_{+,4R} - v_R &= \fint_{B(x_0,R)} (v_{+,4R} - v) \\ &\leq \left(\fint_{B(x_0,R)} |v_{+,4R} - v|^p \right)^{\frac{1}{p}} \quad \text{since } p \geq 1 \\ &\leq c_2 \left(\fint_{B(x_0,2R)} |v_{+,4R} - v|^p \right)^{\frac{1}{p}} \quad \text{by (7.6.22)} \\ &\leq c_3 (v_{+,4R} - v_{+,R}) \end{aligned}$$

by Theorem A.2.2 (ii), since $v_{+,4R} - v$ is a nonnegative supersolution on $B(x_0, 4R)$. Consequently,

$$v_{+,R} \leq \frac{c_3 - 1}{c_3} v_{+,4R} + \frac{1}{c_3} v_R.$$

□

From Lemma 7.6.4, we derive

Lemma 7.6.5. *Let v satisfy the assumptions of Lemma 7.6.4, and suppose* $0 < \varepsilon < \frac{1}{4}$. *There exists $m \in \mathbb{N}$ (independent of v and ε) such that*

$$v_{+,\varepsilon^m R} \leq \varepsilon^2 v_{+,R} + (1-\varepsilon^2) v_{R'},$$

for some R' with $\varepsilon^m R \leq R' \leq \frac{R}{4}$ (R' may depend on v and ε).

Proof. Iterating the estimate of Lemma 7.6.4, we get for $\nu \in \mathbb{N}$

$$v_{+,4^{-\nu}R} \leq (1-\delta_0)^\nu v_{+,R} + (1-(1-\delta_0)^\nu) \sum_{i=1}^{\nu} \tau_i v_{4^{-i}R},$$

with

$$\tau_i = \frac{\delta_0 (1-\delta_0)^{\nu - i}}{1-(1-\delta_0)^\nu}.$$

We choose ν so large that

$$(1-\delta_0)^\nu \leq \varepsilon^2,$$

and choose $R' = 4^{-j}R$ with $j \in \{1, \dots, \nu\}$ such that $v_{4^{-j}R}$ is largest.

Noting that

$$4^{-\nu} \geq \varepsilon^m$$

if

$$m \geq \frac{-(\log 4)}{\log(1-\delta_0)},$$

the result follows. □

Lemma 7.6.6. *Under the assumptions of Lemma 7.6.4,*

$$\lim_{R\to 0} v_R = \lim_{R\to 0} v_{+,R}\,.$$

Proof. This follows directly from Lemma 7.6.4. □

Lemma 7.6.7. *Let $f : B(x_0, 4R) \to Y$ satisfy (7.6.3). Let $p \in Y$. Then, with*

$$v(x) := d^2(f(x), p),$$
$$|B(x_0, R)| := \mathrm{Vol}\,(B(x_0, R)),$$

$$\frac{R^2}{|B(x_0,R)|}\int_{B(x_0,R)} \|df(x)\|^2 d\mathrm{Vol}\,(x) \leq c_5(v_{+,4R} - v_{+,R}). \tag{7.6.23}$$

In particular,

$$\lim_{R\to 0}\frac{R^2}{|B(x_0,R)|}\int_{B(x_0,R)} \|df(x)\|^2 d\mathrm{Vol}\,(x) = 0. \tag{7.6.24}$$

Proof. $v(x) = d^2(f(x), p)$ satisfies (7.6.3), that is,

$$-\Delta v \geq 2\|df\|^2.$$

Let $G^R(x, y)$ be the mollified Green function on $B(x_0, R)$ relative to $B(x_0, 2R)$, i.e. $G^R(x_0, \cdot) \in H^{1,2} \cap C_0^0(B(x_0, 2R))$,

$$\int_{B(x_0,2R)} \langle d\varphi(x), dG^R(x_0, x)\rangle\, d\mathrm{Vol}\,(x) = \fint_{B(x_0,R)} \varphi(x)\, d\mathrm{Vol}\,(x)$$

for all $\varphi \in H^{1,2}$ with $\mathrm{supp}\,\varphi \Subset B(x_0, 2R)$.

We put

$$w^R(x) := \frac{|B(x_0,2R)|}{R^2} G^R(x_0, x).$$

We then have

$$\int_{B(x_0,2R)} \langle d\varphi(x), dw^R(x)\rangle = \frac{1}{R^2}\int_{B(x_0,R)} \varphi(x) \tag{7.6.25}$$

for all $\varphi \in H^{1,2}$ with $\operatorname{supp}\varphi \Subset B(x_0, 2R)$.

Furthermore, from the estimates for G^R of Corollary A.2.1, we have

$$0 \le w^R \le \gamma_1 \quad \text{in } B(x_0, 2R), \tag{7.6.26}$$

$$w^R \ge \gamma_2 > 0 \quad \text{in } B(x_0, R) \tag{7.6.27}$$

for constants γ_1, γ_2 that do not depend on R.

We then have with $z := v - v_{+,4R}$

$$\begin{aligned}
\lambda \int_{B(x_0,2R)} \langle df, df\rangle (w^R)^2 &\le \int_{B(x_0,2R)} (w^R)^2(-\Delta)z \\
&= -\int_{B(x_0,2R)} \langle d(w^R)^2, dz\rangle \quad \text{since } w^R \in H^{1,2}(B(x_0,2R)) \\
&= -2\int \langle dw^R, d(w^R z)\rangle + 2\int z\langle dw^R, dw^R\rangle \\
&\le -2\int \langle dw^R, d(w^R z)\rangle \quad \text{since } z \le 0.
\end{aligned}$$

From (7.6.25), (7.6.26), (7.6.27), we get

$$\begin{aligned}
\int_{B(x_0,R)} \langle df, df\rangle &\le \frac{c_4}{R^2}\int_{B(x_0,R)} (v_{+,4R} - v) \\
&\le c_4 \frac{|B(x_0,R)|}{R^2}(v_{+,4R} - v_R) \\
&\le c_5 \frac{|B(x_0,R)|}{R^2}(v_{+,4R} - v_{+,R}) \quad \text{by Lemma 7.6.4}
\end{aligned}$$

□

We are now ready to prove the Hölder continuity of f. For a point x_0 in the domain and a radius $R > 0$, let

$$\bar{f}_R := \text{mean value of } f \text{ on } B(x_0, R),$$

(that is, as in Lemma 7.6.1, the minimizer of $\int_{B(x_0,R)} d^2(f(x), p) d\mathrm{Vol}\,(x)$ w.r.t. p), and

$$v_p(x) := d(f(x), p), \quad \text{with } p \in Y \text{ chosen subsequently.}$$

We apply Lemma 7.6.5 to

$$v_{\bar{f}_{\frac{R}{4}}} = d^2(f(x), \bar{f}_{\frac{R}{4}})$$

and choose $\varepsilon = \frac{1}{8}$. $\varepsilon^m \le \frac{1}{8}$ and $\varepsilon^m R \le R' \le \frac{R}{4}$, where $m \in \mathbb{N}$ does not depend on ε or $R \le R_0$. We therefore obtain

$$\begin{aligned}
v_{R'} &= \fint_{B(x_0,R')} d^2(f(x), \bar{f}_{\frac{R}{4}}) d\mathrm{Vol}\,(x) \\
&\le C_0 \fint_{B(x_0,\frac{R}{4})} d^2(f(x), \bar{f}_{\frac{R}{4}}) d\mathrm{Vol}\,(x) \quad \text{for some } C_0 \text{ independent of } R
\end{aligned}$$

using the ball doubling property (7.6.22),

$$\leq \frac{C_1 R^2}{|B(x_0,R)|} \int_{B(x_0,\frac{R}{4})} \|df\|^2 d\mathrm{Vol}\,(x)$$

by the Poincaré inequality Lemma 7.6.1,

$$\leq C_2(v_{p,+,R} - v_{p,+,\frac{R}{4}})$$

by Lemma 7.6.7, also using the ball doubling property (7.6.22) once more.

Combining this estimate with Lemma 7.6.5, we get for p in the convex hull of $f(B(x_0,\varepsilon^m R))$

$$\begin{aligned}\sup_{B(x_0,\varepsilon^m R)} d^2(f,p) &\leq 4 \sup_{B(x_0,\varepsilon^m R)} d^2(f,\bar{f}_{\frac{R}{4}})\\ &\leq 4\varepsilon^2 \sup_{B(x_0,R)} d^2(f,\bar{f}_{\frac{R}{4}}) + C_3(v_{p,+,R} - v_{p,+,\frac{R}{4}})\\ &\leq 16\varepsilon^2 \sup_{B(x_0,R)} d^2(f,p) + C_3(v_{p,+,R} - v_{p,+,\varepsilon^m R}) \;\text{ since } \varepsilon^m \leq \frac{1}{4}.\end{aligned}$$

We put, for $0 < \rho$,

$$\omega(\rho) := \sup_{B(x_0,\rho)} d^2(f(x),p) = v_{p,+,\rho}$$

and obtain

$$(1 + C_3)\omega(\varepsilon^m R) \leq \Big(\frac{1}{64} + C_3\Big)\omega(R).$$

Here, ε^m is considered as a constant. By iteration, we obtain

$$\omega(\rho) \leq c\Big(\frac{\rho}{R_0}\Big)^{\alpha}\omega(R_0)$$

for some $c > 0$ and some $0 < \alpha < 1$.

This holds for any p in the convex hull of $f(B(x_0,\rho))$. In particular, we may choose

$$p = \bar{f}_\rho.$$

Since

$$\omega(\rho)^{\frac{1}{2}} \leq \operatorname*{osc}_{B(x_0,\rho)} f \leq 2\omega(\rho)^{\frac{1}{2}},$$

this implies the Hölder continuity of f. Thus, we have obtained another proof of Theorem 7.6.1.

Corollary 7.6.1. *Let $f : M \to N$ be a weakly harmonic map between compact Riemannian manifolds M and N, with N of nonpositive sectional curvature.*

Then f is smooth.

Proof. Let $B(x_1, 6r)$ be a ball in M with $0 < 6r < i(M)$. Since such a ball is simply connected (being the diffeomorphic image of a ball in $T_{x_1}M$ under the exponential map $\exp_{x_1}$), we may lift f to a map

$$f : B(x_1, 6r) \to Y$$

into the universal cover Y of N. Therefore, we may apply Theorem 7.6.1 to get the continuity of f. The smoothness then follows from Theorem 7.4.1. $\square$

In the preceding, we have seen how to use the weak version of the differential inequality

$$-\Delta d^2(f(x), p) \geq 2\|df(x)\|^2 \quad \text{(see (7.6.3)}$$

to derive the continuity of a weakly harmonic map f with values in a manifold of nonpositive sectional curvature.

There is another differential inequality for such a harmonic map that can be used to obtain estimates, namely

$$-\Delta\|df(x)\|^2 \geq -\sigma\|df(x)\|^2, \tag{7.6.28}$$

where $-\sigma$ is a lower bound for the Ricci curvature of M. This inequality follows from (7.2.13).

We shall now display an alternative approach to the regularity result of Corollary 7.6.1 that is based on some weak analogue of (7.6.28). Our construction will exploit the center of mass properties of the approximating functionals E_ε (cf. Lemma 7.3.1) and constructions from §4.8.

Let $f = f_\varepsilon$ be a minimizer of E_ε. (Of course, the existence of a minimizer for E_ε follows by the same method as the one for E, see the proofs of Theorems 7.5.1 and 7.5.2.) By Lemma 7.3.1, for almost every $x \in M$, $f(x)$ is the center of mass of f on the ball $B(x, \varepsilon)$. (As before, we lift f to a map $f : B(x, \varepsilon) \to Y$ into the universal cover of Y where the center of mass then exists by Theorem 4.8.4.)

Let now $x_1, x_2 \in M$ with $d(x_1, x_2) < i(M)$. We define a diffeomorphism

$$\varphi : B(x_1, \varepsilon) \to B(x_2, \varepsilon)$$

as follows: Let

$$\psi : T_{x_1}M \to T_{x_2}M$$

be the linear map that maps an orthonormal frame at x_1 into that orthonormal frame at x_2 that is obtained by parallel transport along the shortest geodesic from x_1 to x_2. ψ then is a Euclidean isometry.

We put

$$\varphi := \exp_{x_2} \circ \psi \circ \exp_{x_1}^{-1},$$

and φ is almost an isometry in the following sense:

If $d\nu_1$ and $d\nu_2$ are the volume forms on $B(x_1,\varepsilon)$ and $B(x_2,\varepsilon)$, resp., then

$$|d\nu_2 - \varphi_* d\nu_1| \leq c\varepsilon^2 \cdot \text{Euclidean volume form},$$

for some constant c. This is easily seen by writing the volume forms in normal coordinates and using Theorem 1.4.4.

Also, if V_i is the volume of $B(x_i,\varepsilon)$, then[5]

$$|V_i - \omega_m \varepsilon^m| \leq c\varepsilon^2.$$

We then apply Corollary 4.8.7 to get

$$\begin{aligned} d(f(x_1), f(x_2)) \leq \frac{1}{V_1} \int_{B(x_1,\varepsilon)} & d(f(y), f(\varphi(y)) d\mathrm{Vol}\,(y) \\ & + \int_{B(x_2,\varepsilon)} d(f(y), f(x_2)) \Big| \frac{d\mathrm{Vol}\,(y)}{V_2} - \frac{\varphi_* d\mathrm{Vol}\,(y)}{V_1} \Big|. \end{aligned} \tag{7.6.29}$$

We note that

$$d(y,\varphi(y)) \leq d(x_1,x_2)\cosh(\sqrt{-\lambda}\,\varepsilon), \quad \text{for all } y \in B(x_1,\varepsilon)$$

if $\lambda \leq 0$ is a lower curvature bound for M; this follows e.g. from Theorem 4.5.2. Again, at this point, it is only needed that

$$d(y,\varphi(y)) \leq d(x_1,x_2)(1 + c\varepsilon^2)$$

for some constant c.

We now iterate (7.6.29), i.e. we estimate the quantities $d(f(y), f(\varphi(y))$ and $d(f(y), f(x_2))$ in the integrals on the right hand side by applying (7.6.29) again. Repeating this a finite number of times, depending on ε, and using the fact that all errors, i.e. deviations from the Euclidean situation, are quadratic in ε, we obtain for $d(x_1,x_2) \leq \varepsilon$

$$\frac{d(f(x_1), f(x_2))}{\varepsilon} \leq c \int_M \int_{B(x,\varepsilon)} \frac{d(f(y), f(x))}{\varepsilon} d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x)$$

for some constant c depending on the geometry of M,

$$\leq c' E_\varepsilon(f)^{\frac{1}{2}} \quad \text{by Hölder's inequality,}$$

for some other constant c'.

This was for $d(x_1,x_2) \leq \varepsilon$. If $d(x_1,x_2) \leq \nu\varepsilon$ for some $\nu \in \mathbb{N}$, we use the triangle inequality to obtain

$$d(f(x_1), f(x_2)) \leq c' E_\varepsilon(f) \nu\varepsilon. \tag{7.6.30}$$

[5] c can be controlled by a lower bound on the Ricci curvature of M and an upper bound for its sectional curvature, but we do not verify this here.

This was for a minimizer $f = f_\varepsilon$ of E_ε. As for example in the proof of the Arzela-Ascoli theorem (see e.g. J. Jost, Postmodern Analysis, Springer, 3rd ed., 2005, p. 55 – 56), one uses (7.6.30) to find a sequence $\varepsilon_n \to 0$ for which the maps f_{ε_n} converge uniformly and hence also in L^2 towards some f. By Lemma 7.3.5, f minimizes E, and it satisfies the limit of the above estimates, i.e.

$$d(f(x_1), f(x_2)) \le c' E(f) d(x_1, x_2)$$

for all $x_1, x_2 \in M$. By the uniqueness theorem proved below (see Theorem 7.7.2), this estimate then holds for any minimizer of E. We have thus shown

Theorem 7.6.2. *Let M and N be compact Riemannian manifolds N of nonpositive sectional curvature, and let $f : M \to N$ minimize the energy in its homotopy class. Then f is Lipschitz continuous.* □

Corollary 7.6.2. *Under the assumption of Theorem 7.6.2, any minimizer f of the energy is smooth.*

Proof. By Lemma 7.2.2, f is a weak solution of

$$\frac{1}{\sqrt{\gamma}} \frac{\partial}{\partial x^\alpha}\Big(\sqrt{\gamma}\gamma^{\alpha\beta}\frac{\partial f^i}{\partial x^\beta}\Big) = -\Gamma^i_{jk}(f)\frac{\partial f^j}{\partial x^\alpha}\frac{\partial f^k}{\partial x^\beta}\,. \tag{7.6.31}$$

By Theorem 7.6.2, the right hand side of (7.6.31) is bounded. By Theorem A.2.3 of Appendix A, therefore $f \in C^{1,\alpha}$, for some $0 < \alpha < 1$. But then the right hand side of (7.6.31) is of class C^α. Applying Theorem A.2.3 once more, yields $f \in C^{2,\alpha}$. Iterating this argument shows that f is smooth. □

Before concluding this section, we want to show how to use (7.6.28) directly to get a-priori estimates for harmonic maps. Since we have not been very precise about the geometric quantities on which the previous estimates derived in our regularity proof depend, we can also use those a-priori estimates to remedy that point. These estimates will use the assumption that f is a **smooth** harmonic map, and so, they cannot be used to show regularity. Such estimates, however, can be employed in various existence schemes (as for example in previous editions of this book).

Theorem 7.6.3. *Let $f : M \to N$ be a harmonic mapping between Riemannian manifolds, where N is complete, simply connected, and of nonpositive sectional curvature. If $x \in M$, $\rho > 0$, and $B(x,\rho) \subset M$, then*

$$e(f)(x) := \frac{1}{2}\|df(x)\|^2 \le c_0\Big(1 + \frac{1}{\rho^2}\Big) \max_{y \in B(x,\rho)} d^2(f(x), f(y)), \tag{7.6.32}$$

where c_0 depends only on $m = \dim M$, on $\Lambda\rho^2$, where Λ is a bound for the absolute value of the sectional curvature of M, and on a lower bound for the Ricci curvature of M.

Proof. We put

$$r(y) := d(x,y),$$
$$q := f(x),$$

and assume for simplicity $m = \dim M \geq 3$. (The proof for $m = 2$ is similar.) We use Lemma 4.7.2 with m instead of n, x instead of p, and $h(y) = d^2(f(y), q)$. We obtain

$$\begin{aligned}
&\int_{B(x,\rho)} \Big(\frac{1}{r(y)^{m-2}} - \frac{1}{\rho^{m-2}}\Big)(-\Delta) d^2(f(y), \rho) \\
&\qquad \leq -(m-2)\omega_m d^2(f(x), q) \\
&\qquad\qquad + \frac{m-2}{2}\Lambda \int_{B(x,\rho)} \frac{d^2(f(y),q)}{r(y)^{m-2}} + \frac{m-2}{\rho^{m-1}} \int_{\partial B(x,\rho)} d^2(f(y), q) \\
&\qquad \leq c_1 \max_{y \in B(x,\rho)} d^2(f(y), q),
\end{aligned} \tag{7.6.33}$$

with c_1 depending on m and $\Lambda\rho^2$.

We next let $\eta \in C_0^\infty(B(x, \frac{\rho}{2}))$ be a cut-off function,

$$0 \leq \eta \leq 1 \quad \text{on } B\Big(x, \frac{\rho}{2}\Big),$$
$$\eta(x) = 1,$$
$$|\nabla \eta| \leq \frac{c_2}{\rho}, \quad |\Delta \eta| \leq \frac{c_3}{\rho^2}.$$

We then apply Lemma 4.7.2 to $h(y) = \eta^2(y) e(f)(y)$ and obtain

$$\begin{aligned}
(m-2)\omega_m e(f)(x) \leq &\int_{B(x,\rho)} \Big(\frac{1}{r(y)^{m-2}} - \frac{1}{\rho^{m-2}}\Big)\Delta(\eta^2 e(f))(y) \\
&+ 2\Lambda \int_{B(x,\rho)} \frac{(\eta^2 e(f))(y)}{r(y)^{m-2}}.
\end{aligned} \tag{7.6.34}$$

Now

$$\begin{aligned}
\Delta(\eta^2 e(f)) &\leq |\Delta \eta^2| e(f) + 4\eta |\nabla \eta| \|\nabla df\| \cdot \|df\| + \eta^2 \Delta e(f) \\
&\leq \frac{c_3}{\rho^2} e(f) + \eta^2 \|\nabla df\|^2 + \frac{c_4}{\rho^2} e(f) - \eta^2 \|\nabla df\|^2 + c_5 e(f)
\end{aligned}$$

by (7.2.13), since N has nonpositive curvature, where $-c_5$ is a lower bound for the Ricci curvature of M,

$$\leq c_6 \Big(1 + \frac{1}{\rho^2}\Big) e(f). \tag{7.6.35}$$

From (7.6.34), (7.6.35),

$$e(f)(x) \leq c_7 \Big(1 + \frac{1}{\rho^2}\Big) \int_{B(x,\rho)} \Big(\frac{1}{r(y)^{m-2}} - \frac{1}{\rho^{m-2}}\Big) e(f)(y) \tag{7.6.36}$$

noting $\frac{1}{r^{m-2}} \le \text{const} \cdot (\frac{1}{r^{m-2}} - \frac{1}{\rho^{m-2}})$, if $r \le \frac{1}{2}\rho$, where the constant depends only on m, as well as $\eta(y) = 0$ if $r(y) \ge \frac{\rho}{2}$.

We now recall (7.6.3), i.e.

$$-\Delta d^2(f(y), p) \ge 4e(f)(y). \tag{7.6.37}$$

Then (7.6.32) follows from (7.6.36), (7.6.37), (7.6.33). □

From the proof, we also obtain

Theorem 7.6.4. *Under the assumptions of Theorem 7.6.2,*

$$e(f)(x) \le \gamma_0 \Big(1 + \frac{1}{\rho^m}\Big) \int_{B(x,\rho)} e(f)(y) = \gamma_0 \Big(1 + \frac{1}{\rho^m}\Big) E(f_{|B(x,\rho)})$$

where γ_0 depends on the same quantities as c_0 in Theorem 7.6.2.

Proof. From (7.6.36), with ρ' instead of ρ and also assuming $B(x, \rho') \subset M$, we obtain

$$e(f)(x) \le c_7 \Big(1 + \frac{1}{\rho'^2}\Big) \int_{B(x,\rho')} d^{2-m}(x, y) e(f)(y)\, dy. \tag{7.6.38}$$

We put

$$\begin{aligned} g_1(y, z) &:= d^{2-m}(y, z), \\ g_k(y, z) &:= \int_{B(z,\rho')} g_{k-1}(y, w) g_1(z, w)\, dw. \end{aligned}$$

We observe that

$$g_k(y, z) \le c_m d^{2k-m}(y, z).$$

For example, for $k = 2$,

$$g_2(y, z) = \int_{B(z,\rho')} d^{2-m}(y, w) d^{2-m}(z, w)\, dw.$$

We split this integral into integrals over the regions

$$\begin{aligned} I &:= \{w : d(y, w) \le \tfrac{1}{2} d(y, z)\}, \\ II &:= \{w : d(z, w) \le \tfrac{1}{2} d(y, z)\}, \\ III &:= B(z, \rho') \backslash I \cup II. \end{aligned}$$

Then

$$\begin{aligned} g_2(y, z) &\le \int_I + \int_{II} + \int_{III} \\ &\le c_m d^{4-m}(y, z), \end{aligned}$$

as desired. In particular, $g_k(y,z)$ is bounded for $k \geq \frac{m}{2}$. We now iterate (7.6.38); this means that we estimate $e(f)(y)$ in the integral on the right hand side of (7.6.38) by using (7.6.38) for x instead of y. We need then $B(y,\rho') \subset M$, and so we choose $\rho' = \frac{\rho}{k}$ to guarantee this condition also for the subsequent steps.

After at most $\frac{m}{2}$ steps, we obtain the desired estimate. $\square$

Perspectives. The literature on the regularity of harmonic maps has become too numerous and extensive to be reviewed here. (See, however, the Perspectives on §7.7 for some references.)

Therefore, in this section, I have rather tried to present a representative sample of techniques from geometric analysis. The proof of Theorem 7.6.1 given here is due to Lin[179]. I have selected that proof because it employs a fundamental tool, namely Moser's Harnack inequalities, in a particularly elegant and geometrically instructive manner. The alternative proof is taken from Jost[141]; it is the most general and powerful regularity proof presently known; in particular, in contrast to the preceding proof, it does not utilize a compactness argument in the target. (The telescoping argument Lemma 7.6.7 is originally due to [94]; the Harnack inequalities for the mollified Green function can be replaced by a more elementary geometric argument, see [132].) The proof of Theorem 7.6.2 given here is taken from Jost[140]. I have selected that proof because it elucidates the interplay between the geometric meaning of the energy functional and its approximations and the geometric features of nonpositively curved manifolds. Finally, the proofs of Theorems 7.6.3 and 7.6.4 (variants of results of Eells and Sampson[75]) have been developed here because of their elementary nature, depending only on the geometry of the distance function as described in §4.7.

7.7 Harmonic Maps into Manifolds of Nonpositive Curvature: Uniqueness and Other Properties

The results of §§7.5, 7.6 can be summarized as

Theorem 7.7.1. *Let M and N be compact Riemannian manifolds, N of nonpositive sectional curvature.*

Let $g : M \to N$ be a continuous map. Then g is homotopic to a smooth harmonic map f, and f can be obtained by minimizing the energy among maps homotopic to g.

The existence result was deduced from a convexity property of the energy functional E. That convexity also suggests a uniqueness result for minimizers of E. Here, we shall present a variant of such a reasoning that applies to all harmonic maps and shows that they are in fact all minimizers of E.

Theorem 7.7.2. *Let M be a compact, N a complete Riemannian manifold. We assume that N has nonpositive curvature. Let $f_0, f_1 : M \to N$ be homotopic harmonic* *maps. Then there exists a family $f_t : M \to N$, $t \in [0,1]$, of harmonic maps connecting them, for which the energy $E(f_t)$ is independent of t, and for which every curve $\gamma_x(t) := f_t(x)$ is geodesic, and $\|\frac{\partial}{\partial t}\gamma_x(t)\|$ is independent of x and t. If N has negative curvature, then f_0 and f_1 either are both constant maps, or they both map M onto the same closed geodesic, or they coincide.*

If M is a compact manifold with boundary, and if $f_{0|\partial M} = f_{1|\partial M}$. then again $f_0 = f_1$.

Proof. We let
$$H : M \times [0,1] \to N$$
be a homotopy between f_0 and f_1, with fixed boundary values if $\partial M \neq \emptyset$. In particular $H(x,0) = f_0(x)$, $H(x,1) = f_1(x)$. We let $\gamma_x(t)$ be the geodesic arc homotopic to the arc $H(x,t)$. By Lemma 4.8.1, $\gamma_x(t)$ is unique. Again, $t \in [0,1]$, and of course $\gamma_x(t)$ is parametrized proportionally to arc length, and we put $f_t(x) := \gamma_x(t)$.

By Corollary 7.2.1, since N has nonpositive curvature,

$$\begin{aligned}\frac{d^2}{dt^2}E(f_t) &= \int_M \left(\left\| \nabla \frac{\partial}{\partial t}\gamma_x(t) \right\|^2 - \operatorname{trace}_M \left\langle R^N \left(df_t, \frac{\partial}{\partial t}\gamma_x(t) \right) \frac{\partial}{\partial t}\gamma_x(t), df_t \right\rangle \right) \\ &\geq 0.\end{aligned} \tag{7.7.1}$$

Since $\frac{d}{dt}E(f_t)_{|t=0} = 0 = \frac{d}{dt}E(f_t)_{|t=1}$, we obtain
$$E(f_t) \equiv \text{const.} \tag{7.7.2}$$
From (7.7.1) then $\nabla \frac{\partial}{\partial t}\gamma_x(t) \equiv 0$; hence $\frac{\partial}{\partial t}\gamma_x(t)$ is also constant in x. If $\partial M \neq \emptyset$, hence $\frac{\partial}{\partial t}\gamma_x(t) = 0$ for all x, since this is true for $x \in \partial M$; hence $f_0 = f_1$ in this case.

One also sees that f_0 and f_1 and hence by (7.7.2) all maps f_t are energy minimizing in their homotopy class, hence all harmonic. We also get from (7.7.1), 7.7.2, by the nonpositivity of the curvature of N
$$\left\langle R^N \left(df_t, \frac{\partial}{\partial t}\gamma_x(t) \right) \frac{\partial}{\partial t}\gamma_x(t), df_t \right\rangle \equiv 0.$$
If N has negative sectional curvature, then either $\frac{\partial}{\partial t}\gamma_x(t) \equiv 0$ in which case again $f_0 = f_1$, or $\operatorname{Rank}_{\mathbb{R}} df_t(x) \leq 1$ for every x, so that f_t is constant or maps M onto the geodesic $\gamma_x(t)$. If $\partial M = \emptyset$, the image of M under f_0 and f_1 in this case has to be a closed geodesic. □

From Theorem 7.7.1 and Corollary 7.2.3, we obtain

Corollary 7.7.1. *Let N be a compact manifold of nonpositive sectional curvature.*

Then every map from a compact manifold with positive Ricci curvature into N is homotopic to a constant map. Every map from a compact manifold with nonnegative Ricci curvature, in particular from a flat manifold into N is homotopic to a totally geodesic map. If the sectional curvature of N is even negative, then any such map is homotopic to a constant map or a map onto a closed geodesic. □

An implication of Corollary 7.7.1 is that manifolds of positive Ricci curvature are topologically very different from those of nonpositive sectional curvature.

We shall now prove **Preissmann's Theorem.**

Corollary 7.7.2. *Let N be a compact Riemannian manifold of negative sectional curvature. Then every abelian subgroup of the fundamental group is infinite cyclic, i.e. isomorphic to $\mathbb{Z}$.*

Proof. Let $\alpha, \beta \in \pi_1(N, x_0)$. Thus α and β are represented by closed loops with base point x_0. If α and β commute, the homotopy between $\alpha\beta$ and $\beta\alpha$ induces a map $g : T^2 \to N$, where T^2 is a two-dimensional torus, i.e.

$$\begin{array}{ccc} & \xrightarrow{\ \alpha\ } & \\ \beta\Big\uparrow & & \Big\uparrow\beta \\ & \xrightarrow[\ \alpha\ ]{} & \end{array} \qquad \text{with} \quad \begin{array}{rl} g : [0,1] \times [0,1] \to & N, \\ g(s,0) = g(s,1) = & \alpha(s), \\ g(0,t) = g(1,t) = & \beta(t), \end{array} \qquad \text{for all } s, t,$$

and in particular $g(0,0) = g(1,0) = g(0,1) = g(1,1) = x_0$, since α, β have base point x_0.

By Theorem 7.7.1, g is homotopic to a harmonic map

$$f : T^2 \to N.$$

During the homotopy between g and f, the base point may change, but of course the two loops corresponding to α and β will always have the same base point at each step of the homotopy.

Since N has negative sectional curvature, by Corollary 7.2.1, $f(T^2)$ is contained in a closed geodesic γ, with base point $x_1 = f(0,0)$. Therefore, our two loops in $\pi_1(N, x_1)$ (the ones obtained from α and β through the homotopy from g to f, i.e. the curves $f(0,\cdot)$ and $f(\cdot,0)$) are both multiples of γ. Thus they are contained in a cyclic subgroup of $\pi_1(N, x_1)$. This cyclic subgroup has to be infinite as otherwise γ^k for some $k \in \mathbb{N}$ would be homotopic to a constant loop (representing the trivial element of $\pi_1(N, x_1)$), contradicting uniqueness of geodesics (Lemma 4.8.1), as γ^k is a geodesic since γ is, and of course a constant loop is also geodesic.

Thus, the subgroup of $\pi_1(N, x_0)$ generated by α and β is isomorphic to an infinite cyclic subgroup. This is true for any two commuting elements in $\pi_1(N, x_0)$, and the conclusion follows. $\square$

Perspectives. While the concept of harmonic maps had been introduced earlier by Bochner [24], the insight that led to the existence theorem 7.7.1, namely that nonpositive target curvature leads to a useful differential inequality via a Bochner formula was obtained by Al'ber[4, 5] and Eells and Sampson[75]. Once this had been noted, one could essentially apply the linear argument of [188] to obtain regularity and existence of harmonic maps with values in manifolds of nonpositive curvature. In fact, Al'ber[5] also showed uniqueness (Theorem 7.7.2) and already conceived the general scheme of applying harmonic maps to the investigation of the topology of manifolds of nonpositive curvature; in particular, he was the first to derive Preismann's theorem from a harmonic map identity. Thus, his work is one of the several instances encountered in this book when mathematicians in the former Soviet Union obtained results that were not given credit in the Western countries, sometimes from ignorance, but sometimes also deliberately. Hartman[119] also obtained the uniqueness result for harmonic maps into manifolds of nonpositive curvature. For the case of manifolds with boundary, such results were obtained by Hamilton[115]. These authors used a parabolic method. They considered the so-called heat flow, i.e. the problem

$$\begin{aligned} &f : M \times [0, \infty) \to N, \\ &\frac{\partial f}{\partial t}(x,t) = \tau(f(x,t)) \text{ where the tension field } \tau \text{ is taken w.r.t. the } x\text{-variable on } M, \\ &f(x,0) = g(x). \end{aligned}$$

They showed that a solution exists for all $t > 0$, and as $t \to \infty$, $f(x,t)$ converges to a harmonic map homotopic to g. This needs parabolic analogues of the estimates of Theorem 7.6.3, Theorem 7.6.4. A detailed and simplified presentation of this approach is given in [132]. Elliptic methods were first introduced into harmonic map theory by Hildebrandt, Kaul, and Widman[124, 125].

Hildebrandt-Kaul-Widman[125] were also able to handle positive image curvature. They solved the Dirichlet boundary problem for harmonic maps with values in a ball $B(p,\rho)$ in some Riemannian manifold N, with $\rho < \min\,(i(p), \frac{\pi}{2\sqrt{\kappa}})$, where $\kappa \geq 0$ is an upper bound on the sectional curvature of N.

The proof allows an important simplification by a result of Kendall[163]. He constructed suitable convex functions on such a ball. Such geometric constructions adapted to positive curvature had earlier allowed Jäger, Kaul[131] to show that the solution for the harmonic Dirichlet problem in such a ball is unique. See also [132] for a presentation of these results.

Hildebrandt-Kaul-Widman[125] also discovered that without that convexity condition on the target ball, critical points of the energy can be discontinuous, and they found the basic example of a singularity, namely the map

$$\begin{aligned} f : B(0,1)(\subset \mathbb{R}^n) &\to S^{n-1}, \\ x &\mapsto \frac{x}{|x|}. \end{aligned}$$

For $n \geq 3$, it has finite energy and is a critical point for the energy.

Schoen and Uhlenbeck[224, 225] and in a somewhat different context also Giaquinta and Giusti[94, 95] then developed a regularity theory for energy minimizing maps. They discovered that the above example is the prototype of a singularity, that energy minimizing maps are regular except possibly on set of Hausdorff dimension at most dimension $M-3$ and that singularities can be precluded if there are no nontrivial energy minimizing harmonic maps from a sphere $S^k (k \geq 2)$ into the target. Note that in the above example, for $r \geq 1$, $f(\frac{x}{r})$ defines a harmonic map from the sphere S^{n-1} into S^{n-1}. In the general case of a singularity, the same has to happen at least in the limit $r \to \infty$. For a detailed account of the theory and its subsequent developments, we recommend Steffen[239].

Returning to nonpositive image curvature, as mentioned above, Al'ber[5] was the first to observe that harmonic maps can be used to prove Preissmann's theorem. Extensions of Preissmann's theorem, i.e. further restrictions on fundamental groups of compact manifolds of nonpositive curvature, were found by Yau[270], Gromoll and Wolf[104], Lawson and Yau[177]. The harmonic map approach to these results is presented in [145]. Recently, a general theory of harmonic maps between metric spaces has been developed. A systematic description, together with the appropriate references, can be found in Jost[142].

We now want to discuss some further results about harmonic maps and their applications.

The first topic are so-called harmonic coordinates. Let M be an n-dimensional Riemannian manifold. Local coordinates are diffeomorphisms from an open subset U of M onto an open subset of $\mathbb{R}^d$. They are called harmonic if the coordinate functions are harmonic. Harmonic coordinates have been employed in general relativity. They were introduced into Riemannian geometry by Sabitov and Shefel'[217] and by de Turck and Kazdan[59] by showing that the metric tensor when written in harmonic coordinates has the best possible regularity properties. (In particular, the regularity properties are better than those of normal coordinates.) Explicit estimates were developed parallely and independently by Jost and Karcher[148] and Nikolaev[202]. The precise result of Jost-Karcher is

Theorem. *Let $p \in M$. There exists $R_0 > 0$, depending only on the injectivity radius of p, the dimension n of M, and a bound Λ for the absolute value of the sectional curvature on $B(p, R_0)$ with the property that for any $R \le R_0$, there exist harmonic coordinates on $B(p, R)$ the metric tensor $g = (g_{ij})$ of which satisfies on each ball $B(p, (1-\delta)R)$ for every $0 < \alpha < 1$*

$$|g|_{C^{1,\alpha}} \le \frac{c(\Lambda R_0, n, \alpha)}{\delta^2} \Lambda^2 R^2.$$

(Here, the norm is the usual one of the Hölder space $C^{1,\alpha}$.) In particular the α-Hölder norms of the Christoffel-symbols are bounded in terms of ΛR_0 and n.

(See also the presentation in [132].) It is easy to construct harmonic functions on balls $B(p, R_0)$. A difficult point, however, is to construct n harmonic functions that furnish an injective map of maximal rank into $\mathbb{R}^n$. This is the main achievement of the preceding result. As an application, one obtains $C^{2,\alpha}$ estimates for harmonic maps between Riemannian manifolds, depending only on the dimensions, injectivity radii and curvature bounds of the manifolds involved provided one knows an estimate for the modulus of continuity of the maps already. (Otherwise, no estimate can hold, see Theorem 8.1.2). These estimates were also the crucial tool for the proofs of the Gromov compactness theorem (see Short survey on curvature and topology, above).

We already described in the Perspectives on §4.8 how to define a notion of a metric space of nonpositive curvature. Now, by an extension of the construction presented in §7.3, one may define an energy integral for maps between metric spaces as a generalization of the energy integral in the Riemannian case considered here. Again, it turns out to be expedient not to work with maps between compact spaces as we did in this section, but rather to lift to their universal covers and consider equivariant maps. Thus, let X and Y be metric spaces with isometry groups $I(X)$ and $I(Y)$, resp., Γ a (typically discrete) subgroup of $I(X), \rho : \Gamma \to I(Y)$ a homomorphism. We then call $f : X \to Y$ ρ-equivariant if

$$f(\gamma x) = \rho(\gamma) f(x) \quad \text{for all } x \in X, \gamma \in \Gamma.$$

Of course, if M and N are compact Riemannian manifolds with fundamental groups $\pi_1(M)$ and $\pi_1(N)$, resp., then these groups operate by deck transformations on the universal covers $X := \tilde{M}$, $Y := \tilde{N}$, and a homotopy class of maps from M to N defines a homoporphism

$$\rho : \pi_1(M) \to \pi_1(N) \subset I(Y),$$

and the lift of any map in that homotopy class to the universal covers then has to be ρ-equivariant. In fact, if N is a so-called $\kappa(\pi,1)$-space, meaning that all higher homotopy groups $\pi_k(N), k \geq 2$, are trivial (such an N is also called aspherical, because that means that every continuous map $\varphi : S^k \to N$, $k \geq 2$, is homotopic to a constant map), then conversely the push down of any ρ-equivariant map lies in the homotopy class defining ρ. This device, namely to work with ρ-equivariant maps, among other things, has the important advantage that it also naturally applies in situations where some of the elements γ and $\rho(\gamma)$ have nontrivial fixed points, i.e. where the spaces X/Γ and/or $Y/\rho(\Gamma)$ may have singularities. The energy of a ρ-equivariant map then is simply defined by integration over a fundamental region of Γ in X. Minimizers are called generalized harmonic maps. The key feature of the assumption of nonpositive curvature then is that it makes the energy integral a convex functional on spaces of ρ-equivarant, square integrable maps as in §7.5.

As already indicated, this works in considerable generality, and in fact, such generality is useful for example in the context of superrigidity discussed below where certain metric spaces of nonpositive curvature that are quite far from being manifolds naturally occur. Some of those spaces even are not locally compact anymore.

A theory of such generalized harmonic mappings has been developed by J. Jost[138, 139, 140, 142] and independently (but under more restrictive assumptions, like local compactness) by Korevaar and Schoen[171]. In fact, a key point of the approach of Jost is that the convexity of the functional can compensate the lack of local compactness of the target in existence proofs. (Subsequently, Korevaar and Schoen[172] reproved a special case of those existence results by a variant of the method of Jost.) Actually, still more generality can be achieved, and new light can be shed on why nonpositive curvature is the fundamental assumption for harmonic maps. Namely, a space of ρ-equivariant, square integrable maps into a space of nonpositive curvature is itself a space of nonpositive curvature (of course, not locally compact anymore even if the original target had been locally compact), and the existence of generalized harmonic maps can then be deduced from an existence theorem for minima of convex functionals on spaces of nonpositive curvature. In fact, we have displayed this existence method in §7.5 in the setting of a Riemannian target. For a comprehensive treatment, we refer to [142].

We finally want to discuss the applications of harmonic maps to superrigidity results (see the Perspectives on §5.5).

As explained in the Perspectives on §7.2, Siu derived a Bochner type identity for harmonic maps between Kähler manifolds. If the image has nonpositive curvature in a suitable sense, it implies that the product of the Hessian of the map with the Kähler form of the domain vanishes, or in other words, that the map is pluriharmonic. A detailed study of the curvature tensors of Hermitian symmetric spaces (i.e. those that are Kähler) of noncompact type then allowed him to conclude that a harmonic homotopy equivalence between compact quotients of such spaces is holomorphic or antiholomorphic. It then also is a diffeomorphism. If the domain is also a quotient of a Hermitian symmetric space, one can then show that the map is an isometry, proving Mostow's theorem in the Hermitian case. It is interesting to note that the curvature terms to be investigated here come from the image and not from the domain. Sampson[222] found a different formula that applies to harmonic maps from Kählerian to Riemannian manifolds. Corlette[57] showed that the product of the Hessian of a harmonic map with any parallel form on the domain vanishes if the image has nonpositive curvature. For quotients of quaternionic hyperbolic space and the hyperbolic Cayley plane this allowed him to conclude that the Hessian itself vanishes, i.e. that a harmonic map from such a quotient into a nonpositively curved manifold is totally geodesic. This again implies a rigidity theorem.

If one wants to derive so-called nonarchimedean superrigidity and arithmeticity of lattices (see Perspectives on §5.5), one has to study homomorphisms of lattices into $\mathrm{Sl}(n, \mathbb{Q}_p)$ ($\mathbb{Q}_p$ = p-adic numbers). It turns out that this group operates on a so-called Tits building, a certain simplicial metric space with nonpositive curvature in the sense of Alexandrov. Gromov and Schoen[111] then developed a theory of harmonic maps from Riemannian manifolds into such spaces. In particular, they could extend Corlette's results to the p-adic case and obtain arithmeticity of the corresponding lattices.

The most general superrigidity results for harmonic maps were obtained by Jost and Yau[156] and Mok, Siu and Yeung[195]. Since the image of a lattice need not be a lattice anymore, once more, one has to work with ρ-equivariant maps.

The result then is that any such harmonic map is totally geodesic, i.e. we have

Theorem. *Let $\tilde{M} = G/K$ be an irreducible symmetric space of noncompact type, other than* $\mathrm{SO}_0(p,1)/\mathrm{SO}(p) \times \mathrm{SO}(1), \mathrm{SU}(p,1)/\mathrm{S}(U(p) \times \mathrm{U}(1))$.

Let Γ be a discrete cocompact subgroup of G (i.e. a cocompact lattice). Let $\tilde{N}$ be a complete simply connected Riemannian manifold of nonpositive curvature operator with isometry group $I(\tilde{N})$. Let $\rho : \Gamma \to I(\tilde{N})$ be a homomorphism for which $\rho(\Gamma)$ either does not have a fixpoint on the sphere at ∞ of $\tilde{N}$ or if it does, it centralizes a totally geodesic flat subspace. Then there exists a totally geodesic ρ-equivariant map,

$$f : \tilde{M} \to \tilde{N} \,.$$

(With the method of Mok-Siu-Yeung, the curvature assumption on $\tilde{N}$ can be weakened; if $\tilde{M} = G/K$ is of rank ≥ 2 then it suffices that $\tilde{N}$ has nonpositive sectional curvature.)

The proof follows from a careful choice of the parameter λ in the Bochner formula of the Perspectives on §7.2 and a detailed study of the curvature tensors of symmetric spaces.

The corresponding result for $\mathrm{SO}_0(p,1)/\mathrm{SO}(p) \times \mathrm{SO}(1)$ and $\mathrm{SU}(p,1)/\mathrm{S}(\mathrm{U}(p) \times \mathrm{U}(1))$ is false, because it would imply that compact quotients have vanishing first Betti number and there are examples of compact quotients of these spaces for which this is not the case. In the case of $\mathrm{SU}(p,1)/\mathrm{S}(\mathrm{U}(p) \times \mathrm{U}(1))$, one gets, however the existence of a pluriharmonic ρ-equivariant map, essentially a special case of the result of Siu quoted above.

For $\mathrm{Sp}(p,1)/\mathrm{Sp}(p) \times \mathrm{Sp}(1)$ and the hyperbolic Cayley plane, the result is Corlette's theorem quoted above. For Hermitian symmetric spaces, the result is due to Mok[194, 193].

Corollary. *Let $\tilde{M} = G/K$ and Γ be as above. Let H be a semisimple noncompact Lie group with trivial center, $\rho : \Gamma \to H$ a homomorphism with Zariski dense image. Then ρ extends to a homomorphism from G onto H.*

As explained above, this result is due to Margulis for rank $(G/K) \geq 2$ and to Corlette for $\mathrm{Sp}(p,1)/\mathrm{Sp}(p) \times \mathrm{Sp}(1)$ and the hyperbolic Cayley plane.

Using the constructions of Gromov and Schoen, the result extends to the nonarchimedean case to show

Theorem. *Let $\tilde{M} = G/K$ and Γ be as above.*

Let $\rho : \Gamma \to \mathrm{Sl}(n, \mathbb{Q}_p)$ be a homomorphism, for some $n \in N$ and some prime p. Then $\rho(\Gamma)$ is contained in a compact subgroup of $\mathrm{Sl}(n, \mathbb{Q}_p)$.

As explained above, the result is again due to Margulis for rank $(G/K) \geq 2$, and to Gromov-Schoen for quaternionic hyperbolic space and the hyperbolic Cayley plane.

The harmonic map approach to rigidity is still not complete:

First of all, so far it has been unable to derive Mostow's rigidity theorem for quotients of real hyperbolic space. Secondly, the results for spaces that are of finite volume but not compact (i.e. for nonuniform lattices) are still not complete. Margulis' results, for example, also hold in the noncompact case. (For rank 1, rigidity results were shown earlier by G. Prasad.) In the Hermitian symmetric case, however, this problem was solved by Jost and Zuo[157].

A new and very interesting approach to rigidity that applies particularly well in the case of real hyperbolic spaces has been developed by Besson, Courtois and Gallot[20, 21].

One open problem that is quite easy to formulate but as yet unsolved is the following one of H. Hopf: Let M^{2m} be a compact manifold of even dimension $2m$ that admits a Riemannian metric of nonpositive sectional curvature. Is it then true that the Euler characteristic of M satisfies

$$(-1)^m \chi(M^{2m}) \geq 0$$

(with strict inequality in the case of negative sectional curvature)? So far, this has only been demonstrated under additional conditions, e.g. that the curvature is pinched between two negative constants, see for example Donnelly, Xavier[67], Bourguignon, Karcher[28], Jost, Xin[153]. If the manifold carries a Kähler metric, then this conjecture has been verified by Gromov[107], in the case of negative sectional curvature, and by Jost, Zuo[158] and Cao, Xavier[40] in the nonpositive case.

Exercises for Chapter 7

1. Determine all harmonic maps between tori.

   (Hint: Use the uniqueness theorem and the fact that affine linear maps between Euclidean spaces are harmonic.)

2. a: We call a closed subset A of a Riemannian manifold N convex if any two points in A can be connected by a geodesic arc in A. We call A strictly convex if this geodesic arc is contained in the interior of A with the possible exception of its endpoints. We call A strongly convex, if its

boundary ∂A is a smooth submanifold (of codimension 1) in N and if all its principal curvatures w.r.t. the normal vector pointing to the interior of A are positive. Show that a strongly convex set is strictly convex.

b: Show that a strongly convex subset A of a complete Riemannian manifold N has a neighborhood whose closure B_1 and $B_0 := A$ satisfy the conclusions of Lemma 8.2.2.

c: Show that Theorem 8.2.1 continues to hold if N is only complete, but not necessarily compact, again with $\pi_2(N) = 0$, provided $\varphi(\Sigma)$ is contained in a compact, strongly convex subset A of N. In that case, the harmonic $f : \Sigma \to N$ also satisfies $f(\Sigma) \subset A$.

3. In this exercise, still another definition of the Sobolev space $H^{1,2}(M,N)$ will be given. The embedding theorem of Nash (see the Perspectives on §1.4) implies that there exists an isometric embedding

$$i : N \to \mathbb{R}^k$$

into some Euclidean space.

We then define

$$H_i^{1,2}(M,N) := \{\in H^{1,2}(M,\mathbb{R}^k) : f(x) \in i(N) \quad \text{for almost all } x \in M\}.$$

Show that

$$H^{1,2}(M,N) = H_i^{1,2}(M,N).$$

(Hint: Theorem 7.2.1 implies that $H^{1,2}(M,\mathbb{R}^k) = H_i^{1,2}(M,\mathbb{R}^k)$ since every map into $\mathbb{R}^k$ is localizable.)

4. a: For $1 < p < \infty$ and $f \in L^p(M,N)$, we define

$$E_{p,\varepsilon}(f) := \frac{1}{\omega_m \varepsilon^{m+p}} \int_M \int_{B(x,\varepsilon)} d^p(f(x), f(y))\, d\mathrm{Vol}\,(y) d\mathrm{Vol}\,(x)$$

(with the same notation as in (7.2.1)), and

$$E_p(f) := \lim_{\varepsilon \to 0} E_{p,\varepsilon}(f) \in \mathbb{R} \cup \{\infty\}$$

(show that this limit exists). We say that $f \in L^p(M,N)$ belongs to the Sobolev space $H^{1,p}(M,N)$ if $E_p(f) < \infty$. Characterize the localizable maps belonging to $H^{1,p}(M,N)$.

b: Show lower semicontinuity of E_p w.r.t. L^p-convergence, i.e. if $(f_\nu)_{\nu\in\mathbb{N}}$ converges to f in $L^p(M,N)$, then

$$E_p(f) \leq \liminf_{\nu \to \infty} E_p(f_\nu).$$

c: Derive the Euler-Lagrange equations for critical points of E_p. (The smooth critical points are called p-harmonic maps. The regularity theory for p-harmonic maps, however, is not as good as the one for harmonic maps. In general, one only obtains weakly p-harmonic maps of regularity class $C^{1,\alpha}$ for some $\alpha > 0$.)

d: Show the existence of a continuous weakly p-harmonic map (minimizing E_p) under the assumptions of Theorem 8.2.1.

e: Extend the existence theory of §7.5 to E_p.

5. Derive formula (7.2.13) in an invariant fashion, i.e. without using local coordinates.

6. Prove the following result that is analogous to Corollary 7.2.4. A smooth map $f : M \to N$ between Riemannian manifolds is totally geodesic if and only if whenever V is open in N, $U = f^{-1}(V)$, $h : V \to \mathbb{R}$ is convex, then $h \circ f : U \to \mathbb{R}$ is convex.

7. Let M be a compact Riemannian manifold with boundary, N a Riemannian manifold, $f : M \to N$ harmonic with $f(\partial M) = p$ for some point p in N. Show that if there exists a strictly convex function h on $f(M)$ with a minimum at p, then f is constant itself.

8. State and prove a version of the uniqueness theorem 7.7.2 for minimizers of the functionals E_ε. Show that, as for the energy functional E, any critical point of E_ε (with values in a space of non-positive sectional curvature, as always) is a minimizer.

Chapter 8

Harmonic maps from Riemann surfaces

8.1 Twodimensional Harmonic Mappings and Holomorphic Quadratic Differentials

Definition 8.1.1. A *Riemann surface* is a complex manifold (cf. Definition 1.1.5) of complex dimension 1.

Thus, coordinate charts on a Riemann surface Σ are given by maps

$$\varphi_i : U_i \to \mathbb{C},$$

U_i open in Σ, for which the transition functions

$$\varphi_j \circ \varphi_i^{-1} : \varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j)$$

are holomorphic maps between open subsets of $\mathbb{C}$.

We write coordinates in $\mathbb{C}$ as

$$z = x + iy\,.$$

For a coordinate transformation $w = w(z)$, $w = u + iv$, we thus have the Cauchy-Riemann equations

$$\begin{aligned} u_x &= v_y, \\ u_y &= -v_x, \end{aligned}$$

and in particular

$$\begin{aligned} &u_x u_x + v_x v_x = u_y u_y + v_y v_y, \\ &u_x u_y + v_x v_y = 0, \end{aligned}$$

and we see that a Riemann surface has a conformal structure in the sense of Definition 3.6.6.

We call $z = \varphi(p)$ for a local chart φ a *local conformal parameter* at $p \in \Sigma$ and define operators (cf. §6.1)

$$\begin{aligned}\frac{\partial}{\partial z} &:= \frac{1}{2}\Big(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\Big),\\ \frac{\partial}{\partial \overline{z}} &:= \frac{1}{2}\Big(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\Big),\end{aligned}$$

and 1-forms

$$dz = dx + idy, \quad d\overline{z} = dx - idy.$$

These satisfy

$$\begin{aligned}dz\Big(\frac{\partial}{\partial z}\Big) &= 1 = d\overline{z}\Big(\frac{\partial}{\partial \overline{z}}\Big),\\ dz\Big(\frac{\partial}{\partial \overline{z}}\Big) &= 0 = d\overline{z}\Big(\frac{\partial}{\partial z}\Big).\end{aligned}$$

A map between Riemann surfaces is called holomorphic or antiholomorphic if it has this property in local coordinates. This does not depend on the choice of local coordinates because all coordinate changes are holomorphic.

Definition 8.1.2. A Riemannian metric $\langle \cdot, \cdot \rangle$ on a Riemann surface Σ is called *conformal* if in local coordinates it can be written as

$$\rho^2(z)\, dz \otimes d\overline{z} \tag{8.1.1}$$

($\rho(z)$ a positive, real valued function).

This means

$$\Big\langle \frac{\partial}{\partial z}, \frac{\partial}{\partial z} \Big\rangle = 0 = \Big\langle \frac{\partial}{\partial \overline{z}}, \frac{\partial}{\partial \overline{z}} \Big\rangle, \tag{8.1.2}$$

$$\Big\langle \frac{\partial}{\partial z}, \frac{\partial}{\partial \overline{z}} \Big\rangle = \rho^2(z). \tag{8.1.3}$$

If we want to express this in real coordinates, we compute

$$dz \otimes d\overline{z} = dx \otimes dx + dy \otimes dy, \tag{8.1.4}$$

hence

$$\begin{aligned}\Big\langle \frac{\partial}{\partial x}, \frac{\partial}{\partial x} \Big\rangle &= \rho^2(z) = \Big\langle \frac{\partial}{\partial y}, \frac{\partial}{\partial y} \Big\rangle,\\ \Big\langle \frac{\partial}{\partial x}, \frac{\partial}{\partial y} \Big\rangle &= 0.\end{aligned} \tag{8.1.5}$$

In the same manner as Theorem 1.4.1 is proved, a partition of unity argument gives

Lemma 8.1.1. *Every Riemann surface admits a conformal metric.* □

Of course, every conformal metric is Hermitian in the sense of Definition 5.1.2, and conversely.

Definition 8.1.3. Let Σ be a Riemann surface, N a Riemannian manifold with metric $\langle \cdot, \cdot \rangle_N$, or $g_{ij} df^i \otimes df^j$ in local coordinates. A C^1-map $f : \Sigma \to N$ is called *conformal*, if

$$\left\langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial x} \right\rangle_N = \left\langle \frac{\partial f}{\partial y}, \frac{\partial f}{\partial y} \right\rangle_N, \qquad \left\langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right\rangle_N = 0. \tag{8.1.6}$$

In local coordinates this is of course expressed as

$$\begin{aligned} g_{ij}(f(z)) \frac{\partial f^i}{\partial x} \frac{\partial f^j}{\partial x} &= g_{ij}(f(z)) \frac{\partial f^i}{\partial y} \frac{\partial f^j}{\partial y}, \\ g_{ij}(f(z)) \frac{\partial f^i}{\partial x} \frac{\partial f^j}{\partial y} &= 0. \end{aligned} \tag{8.1.7}$$

For the sequel, it will also be instructive to write this condition in complex notation, namely

$$\begin{aligned} 0 &= \left\langle \frac{\partial f}{\partial z}, \frac{\partial f}{\partial z} \right\rangle_N \\ &= g_{jk}(f(z)) \Big(\frac{\partial f^j}{\partial x} \frac{\partial f^k}{\partial x} - \frac{\partial f^j}{\partial y} \frac{\partial f^k}{\partial y} - 2i \frac{\partial f^j}{\partial x} \frac{\partial f^k}{\partial y} \Big). \end{aligned} \tag{8.1.8}$$

Lemma 8.1.2. *Holomorphic or antiholomorphic maps between Riemann surfaces are conformal, if the image is equipped with a conformal metric.*

Proof. Obvious. □

Lemma 8.1.3. *Let Σ be a Riemann surface with a conformal metric $\lambda^2(z)$. Then the Laplace-Beltrami operator is*

$$\Delta = -\frac{4}{\lambda^2(z)} \frac{\partial^2}{\partial z \partial \bar{z}}. \tag{8.1.9}$$

Proof. Direct computation. □

Lemma 8.1.4. *Let Σ be a Riemann surface with conformal metric $\lambda^2(z)$, N a Riemannian manifold with metric tensor (g_{ij}). Then a map $f : \Sigma \to N$ of class C^2 is harmonic iff*

$$\frac{\partial^2 f^i}{\partial z \partial \bar{z}} + \Gamma^i_{jk}(f(z)) \frac{\partial f^j}{\partial z} \frac{\partial f^k}{\partial \bar{z}} = 0 \quad \text{for } i = 1, \ldots, \dim N. \tag{8.1.10}$$

It is a parametric minimal surface iff it is harmonic and conformal.

Proof. One checks directly that (8.1.10) is equivalent to (7.1.7). The second claim directly follows from the Definition 3.6.7 of a parametric minimal surface. □

Corollary 8.1.1. *If Σ is a Riemann surface, N a Riemannian manifold, the harmonic map equation for maps $f : \Sigma \to N$ is independent of the choice of conformal metric on Σ. Thus, whether a map is harmonic depends only on the Riemann surface structure of Σ, but does not need any conformal metric.*

Proof. The metric of Σ does not appear in (8.1.10). □

Corollary 8.1.2. *Holomorphic or antiholomorphic maps between Riemann surfaces are harmonic.*

Proof. Such maps obviously satisfy (8.1.10). □

More generally

Corollary 8.1.3. *If $k : \Sigma_1 \to \Sigma_2$ is a holomorphic or antiholomorphic map between Riemann surfaces, and $f : \Sigma_2 \to N$ is harmonic, then so is $f \circ k$.*

Proof. Let w be a local conformal parameter on Σ_1. Then, if for example k is holomorphic, and in local coordinates $k = z(w)$, we have

$$\frac{\partial z}{\partial \overline{w}} = 0,$$

hence

$$\frac{\partial f \circ k}{\partial w} = \frac{\partial f}{\partial z}\frac{\partial z}{\partial w}, \qquad \frac{\partial f \circ k}{\partial \overline{w}} = \frac{\partial f}{\partial \overline{z}}\frac{\partial \overline{z}}{\partial \overline{w}}$$

and

$$\frac{\partial^2 f^i \circ k}{\partial w \partial \overline{w}} + \Gamma^i_{j\ell}\frac{\partial f^j \circ k}{\partial w}\frac{\partial f^\ell \circ k}{\partial \overline{w}} = \Big(\frac{\partial^2 f^i}{\partial z \partial \overline{z}} + \Gamma^i_{j\ell}\frac{\partial f^j}{\partial z}\frac{\partial f^\ell}{\partial \overline{z}}\Big)\frac{\partial z}{\partial w}\frac{\partial \overline{z}}{\partial \overline{w}},$$

and this vanishes if f is harmonic. □

Let Σ, N be as before, $\lambda^2(z)\, dz \otimes d\overline{z}$ a conformal metric on Σ.
The energy of a map $f : \Sigma \to N$ is written as

$$\begin{aligned} E(f) &= \frac{1}{2}\int_\Sigma \frac{4}{\lambda^2(z)} g_{ij}\frac{\partial f^i}{\partial z}\frac{\partial f^j}{\partial \overline{z}}\frac{\sqrt{-1}}{2}\lambda^2(z)\, dz \wedge d\overline{z} \quad \text{since } dx \wedge dy = \frac{1}{2} dz \wedge d\overline{z} \\ &= \int_\Sigma g_{ij}\frac{\partial f^i}{\partial z}\frac{\partial f^j}{\partial \overline{z}}\sqrt{-1}\, dz \wedge d\overline{z}\,. \end{aligned} \tag{8.1.11}$$

Corollary 8.1.4. *The energy of a map from a Riemann surface Σ into a Riemannian manifold is conformally invariant in the sense that it does not depend on the choice of a metric on Σ, but only on the Riemann surface structure. Also, if $k : \Sigma_1 \to \Sigma_2$ is a bijective holomorphic or antiholomorphic map between Riemann surfaces then for any $f : \Sigma_2 \to N$ (of class C^1)*

$$E(f \circ k) = E(f).$$

□

Remark. Even if the image is also a Riemann surface, the energy of f does depend on the image metric.

Theorem 8.1.1. *Let Σ be a Riemann surface, N a Riemannian manifold with metric $\langle \cdot, \cdot \rangle_N$, or $(g_{ij})_{i,j=1,\dots,\dim N}$ in local coordinates. If $f : \Sigma \to N$ is harmonic, then*

$$\varphi(z)\, dz^2 = \left\langle \frac{\partial f}{\partial z}, \frac{\partial f}{\partial z} \right\rangle_N dz^2 \tag{8.1.12}$$

is a holomorphic quadratic differential. (Here, we use the abbreviation

$$dz^2 := dz \otimes dz$$

, and $\varphi(z) dz^2$ is a holomorphic quadratic differential, if $\varphi(z)$ is a holomorphic function. dz^2 just expresses the transformation behavior. Thus

$$\varphi(z)\, dz^2$$

is a section of $T^*_{\mathbb{C}}\Sigma \otimes T^*_{\mathbb{C}}\Sigma$, with $T^*_{\mathbb{C}}\Sigma := T^*\Sigma \otimes \mathbb{C}$.)

Furthermore,

$$\varphi(z)\, dz^2 \equiv 0 \iff f \text{ conformal.}$$

Proof. In local coordinates

$$\varphi(z)\, dz^2 = \left\langle \frac{\partial f}{\partial z}, \frac{\partial f}{\partial z} \right\rangle_N dz^2 = g_{ij}(f(z)) \frac{\partial f^i}{\partial z} \frac{\partial f^j}{\partial z} dz^2,$$

and we have to show for a harmonic f,

$$\frac{\partial}{\partial \bar{z}} \Big(g_{ij}(f(z)) \frac{\partial f^i}{\partial z} \frac{\partial f^j}{\partial z} \Big) = 0.$$

Now

$$\begin{aligned} \frac{\partial}{\partial \bar{z}} \Big(g_{ij}(f(z)) \frac{\partial f^i}{\partial z} \frac{\partial f^j}{\partial z} \Big) &= 2 g_{ij} \frac{\partial^2 f^i}{\partial z \partial \bar{z}} \frac{\partial f^j}{\partial z} + g_{ij,k} \frac{\partial f^k}{\partial \bar{z}} \frac{\partial f^i}{\partial z} \frac{\partial f^j}{\partial z} \\ &= 2 g_{ij} \frac{\partial^2 f^i}{\partial z \partial \bar{z}} \frac{\partial f^j}{\partial z} + (g_{\ell j,k} + g_{\ell k,j} - g_{jk,\ell}) \frac{\partial f^k}{\partial \bar{z}} \frac{\partial f^\ell}{\partial z} \frac{\partial f^j}{\partial z} \\ &= 2 g_{ij} \frac{\partial f^j}{\partial z} \Big(\frac{\partial^2 f^i}{\partial z \partial \bar{z}} + \Gamma^i_{k\ell} \frac{\partial f^k}{\partial \bar{z}} \frac{\partial f^\ell}{\partial z} \Big) \\ &= 0, \text{ if } f \text{ is harmonic.} \end{aligned}$$

Finally, $\varphi(z)\,dz^2 \equiv 0$ is equivalent to the conformality of f, see (8.1.8). $\square$

In intrinsic notation, the proof of Theorem 8.1.1 goes as follows

$$\begin{aligned}\frac{\partial}{\partial \bar z}\Big\langle \frac{\partial f}{\partial z}, \frac{\partial f}{\partial z}\Big\rangle_N &= 2\Big\langle \nabla_{\frac{\partial}{\partial \bar z}} \frac{\partial f}{\partial z}, \frac{\partial f}{\partial z}\Big\rangle_N \\ &= 2\Big\langle \nabla_{\frac{\partial}{\partial f^j}\frac{\partial f^j}{\partial \bar z}} \frac{\partial f^i}{\partial z}\frac{\partial}{\partial f^i}, \frac{\partial f}{\partial z}\Big\rangle_N \\ &= 2\Big\langle \Big(\frac{\partial^2 f^i}{\partial z \partial \bar z} + \Gamma^i_{jk}\frac{\partial f^j}{\partial \bar z}\frac{\partial f^k}{\partial z}\Big)\frac{\partial}{\partial f^i}, \frac{\partial f}{\partial z}\Big\rangle_N \\ &= 0,\end{aligned}$$

since f is harmonic.

We also note from this computation

$$\tau(f) = 4\nabla_{\frac{\partial}{\partial \bar z}} \frac{\partial f}{\partial z}\,. \tag{8.1.13}$$

In real notation, we have of course

$$\begin{aligned}\varphi(z)\,dz^2 &= \Big(\Big\langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial x}\Big\rangle - \Big\langle \frac{\partial f}{\partial y}, \frac{\partial f}{\partial y}\Big\rangle - 2i\Big\langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\Big\rangle\Big)(dx^2 - dy^2 + 2idxdy) \\ &= g_{jk}(f(z))\Big(\frac{\partial f^j}{\partial x}\frac{\partial f^k}{\partial x} - \frac{\partial f^j}{\partial y}\frac{\partial f^k}{\partial y} - 2i\frac{\partial f^j}{\partial x}\frac{\partial f^k}{\partial y}\Big)(dx^2 - dy^2 + 2idxdy).\end{aligned} \tag{8.1.14}$$

The easiest example of a compact Riemann surface is $S^2 = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_1^2 + x_2^2 + x_3^2 = 1\}$ with the following two coordinate charts:

$$\begin{aligned} f_1 : S^2\backslash\{(0,0,1)\} \to \mathbb{C}, \quad f_1(x_1, x_2, x_3) &= \frac{1}{1 - x_3}(x_1 + ix_2), \\ f_2 : S^2\backslash\{(0,0,-1)\} \to \mathbb{C}, \quad f_2(x_1, x_2, x_3) &= \frac{1}{1 + x_3}(x_1 - ix_2). \end{aligned}$$

We compute

$$\frac{1}{f_1(x_1, x_2, x_3)} = f_2(x_1, x_2, x_3)$$

so that $f_2 \circ f_1^{-1}(z) = \frac{1}{z}$ and the coordinate transformation $f_2 \circ f_1^{-1} : \mathbb{C}\backslash\{0\} \to \mathbb{C}\backslash\{0\}$ is holomorphic as required.

Lemma 8.1.5. *Every holomorphic quadratic differential on S^2 vanishes identically.*

Proof. We put $z = f_1(x)$ and write a holomorphic quadratic differential in the chart f_1 as

$$\varphi(z)\,dz^2, \text{ with } \varphi : \mathbb{C}(= f_1(S^2\backslash\{(0,0)\})) \to \mathbb{C} \text{ holomorphic.}$$

Then with $f_2(x) = w = \frac{1}{z}$ for $z \neq 0$,

$$\varphi(z)\,dz^2 = \varphi(z(w))\Big(\frac{\partial z}{\partial w}\Big)^2 dw^2 = \varphi(z(w))\frac{1}{w^4}\,dw^2.$$

Since we have a holomorphic quadratic differential on S^2, this has to be bounded as $w \to 0$. We conclude that φ is a holomorphic function on $\mathbb{C}$ with

$$\varphi(z) \to 0 \text{ as } z \to \infty,$$

hence $\varphi \equiv 0$ by Liouville's theorem. (One may also apply Lemma 8.2.7 below) □

Another Proof. In the preceding notations, for $\lambda \in \mathbb{C}\backslash\{0\}$,

$$z \mapsto \lambda z$$

induces a holomorphic map $h_\lambda : S^2 \to S^2$, fixing $(0,0,1)$ and $(0,0,-1)$. Since h_λ also depends holomorphically on $\lambda \in \mathbb{C}\backslash\{0\}$,

$$\frac{\partial h_\lambda(z)}{\partial \lambda}\Big|_{\lambda=1} = z$$

represents a holomorphic vector field $V(z)$ on S^2.

Now if q is a holomorphic quadratic differential and V_1, V_2 are holomorphic vector fields on a Riemann surface Σ, then

$$q(V_1, V_2)$$

is a holomorphic function on Σ. Thus

$$\eta(z) := \varphi(z)\,dz^2(V(z), V(z)) = \varphi(z)z^2$$

represents a holomorphic function on the compact Riemann surface S^2 and therefore is constant (for example by Corollary 2.1.2 and Corollary 8.1.3 or by an easy application of the maximum principle), hence $\eta \equiv 0$, since $\eta(0) = 0$, hence $\varphi \equiv 0$. □

Corollary 8.1.5. *For any Riemannian manifold N, every harmonic map*

$$h : S^2 \to N$$

is conformal, i.e. a parametric minimal surface.

Proof. From Theorem 8.1.1 and Lemma 8.1.4. □

We look again at the family $h_\lambda = S^2 \to S^2$ of holomorphic selfmaps of S^2, given in the chart f_1 by

$$z \mapsto \lambda z.$$

We equip the image S^2 with any conformal metric and compute the energy E w.r.t. this metric. We observe for $\lambda \in \mathbb{C}\backslash\{0\}$

$$E(h_\lambda) \equiv \text{const} \neq 0.$$

Namely, we write $h_\lambda = \text{id} \circ h_\lambda$ and apply Corollary 8.1.4 with $f = \text{id}\ (= h_1)$, $k = h_\lambda$, hence

$$E(h_1) = E(h_\lambda) \quad \text{for all } \lambda \in \mathbb{C}\backslash\{0\},$$

and since $h_\lambda \neq \text{const}$ for $\lambda \in \mathbb{C}\backslash\{0\}$, this energy cannot vanish. Now if $\lambda \to 0$, h_λ converges pointwise on $S^2\backslash\{(0,0,-1)\}$ to the constant map $h_0(z) = 0$ (again in the chart f_1), and

$$E(h_0) = 0.$$

We thus have found a sequence of holomorphic, hence harmonic (Corollary 8.1.2) maps, hence critical points of E, i.e.

$$DE(h_\lambda) = 0 \quad \text{for all } \lambda \in \mathbb{C}\backslash\{0\}$$

with

$$E(h_\lambda) \equiv \text{const} \neq 0$$

with the property that this sequence converges for $\lambda \to 0$ pointwise almost everywhere to a map h_0 with

$$E(h_0) \neq \lim_{\lambda \to 0} E(h_\lambda). \tag{8.1.15}$$

We conclude

Theorem 8.1.2. *The energy functional for maps from S^2 to S^2 (the image equipped with any conformal metric) cannot satisfy any kind of Palais-Smale condition.* □

The statement is somewhat vague because we have not yet given a precise definition of the Palais-Smale condition in the present context. Any meaningful definition, however, should require that a sequence of critical points $(f_n)_{n\in\mathbb{N}}$ of E contains a subsequence converging in some sense to be specified towards a map f with

$$E(f) = \lim_{n\to\infty} E(f_n).$$

Definition 8.1.4. A *Riemann surface* Σ *with (smooth) boundary* $\partial\Sigma$ is a differentiable manifold with boundary and charts with values in $\mathbb{C}$ and $\mathbb{C}_+ := \{z = x + iy \in \mathbb{C}, y \geq 0\}$, resp., and holomorphic coordinate changes.

Again, in this case $\overset{\circ}{\Sigma} = \Sigma\backslash\partial\Sigma$ is a Riemann surface in the sense of Definition 8.1.1. Also, $\partial\Sigma$ is a differentiable manifold of real dimension 1.

Example. $D := \{z = x + iy \in \mathbb{C} : |z| \leq 1\}$, with $\partial D = \{|z| = 1\}$.

Definition 8.1.5. A holomorphic quadratic differential q on a Riemann surface Σ with boundary $\partial\Sigma$ is called *real on* $\partial\Sigma$ if for all $z_0 \in \partial\Sigma$ and $v_1, v_2 \in T_{z_0}\partial\Sigma$, i.e. vectors tangent to the boundary

$$q(v_1, v_2) \in \mathbb{R}.$$

Let $z_0 \in \partial\Sigma$, $f : U \to \mathbb{C}_+$ a chart defined on a neighborhood of z_0, $z = x + iy \in \mathbb{C}_+$. In this chart, we write a holomorphic quadratic differential as

$$\begin{aligned}\varphi(z)(dx + idy)^2 &= (u + iv)(dx^2 - dy^2 + 2idxdy)\\ &= u(dx^2 - dy^2) - 2vdxdy + i(v(dx^2 - dy^2) + 2iudxdy),\end{aligned} \tag{8.1.16}$$

with $u = \operatorname{Re}\varphi$, $v = \operatorname{Im}\varphi$.

When applied to a vector tangent to $\partial\mathbb{C}_+ = \{y = 0\}$, dy vanishes. Thus, the holomorphic quadratic differential is real on $\partial\Sigma$ if

$$v = \operatorname{Im}\varphi = 0$$

for all such boundary charts.

Lemma 8.1.6. *Any holomorphic quadratic differential on D which is real on ∂D vanishes identically.*

Proof. A holomorphic function h on an open subset Ω of $\mathbb{C}_+$ which takes real values on $\partial\mathbb{C}_+$ can be reflected as a holomorphic function to $\overline{\Omega} := \{x + iy : x - iy \in \Omega\}$ via $h(x + iy) := \overline{h}(x - iy)$. This is the Schwarz reflection principle. In the same manner, a holomorphic quadratic differential on an open subset of $\mathbb{C}_+$ which is real on $\partial\mathbb{C}_+$ can be reflected across $\partial\mathbb{C}_+$. Thus, a holomorphic quadratic differential on D which is real on ∂D can be reflected to a holomorphic quadratic differential on S^2. Namely, since $f_1(S^2\backslash\{(0,0,1)\}) = \mathbb{C}$ in our above notation, we may consider D as a subset of S^2, and we reflect $\varphi(z)dz^2$ across ∂D as

$$\begin{aligned}\varphi(w)\, dw^2 &= \overline{\varphi}(z)\, dz^2 \quad \text{for } w = \frac{1}{z}\\ &= \overline{\varphi}\Big(\frac{1}{w}\Big)\frac{1}{w^4}\, dw^2.\end{aligned}$$

The result now follows from Lemma 8.1.4. ☐

Theorem 8.1.3. *Let $h : D \to N$ be a harmonic map into a Riemannian manifold with*

$$h|_{\partial D} = \text{const.}$$

Then

$$h = \text{const.}$$

Proof. We denote the metric of N by (g_{jk}). In local coordinates defined on an open subset of $\mathbb{C}_+$, the holomorphic quadratic differential associated to h (Theorem 8.1.1) is

$$\varphi\, dz^2 = g_{jk}(h(z))\Big(\frac{\partial h^j}{\partial x}\frac{\partial h^k}{\partial x} - \frac{\partial h^j}{\partial y}\frac{\partial h^k}{\partial y} - 2i\frac{\partial h^j}{\partial x}\frac{\partial h^k}{\partial y}\Big)(dx + i dy)^2,$$

since $h|_{\partial D} = \text{const}$, $\frac{\partial h}{\partial x} = 0$ on $\partial\mathbb{C}_+$. Thus

$$\operatorname{Im}\varphi = 2g_{jk}\frac{\partial h^j}{\partial x}\frac{\partial h^k}{\partial y} = 0 \quad \text{on } \partial\mathbb{C}_+,$$

and $\varphi\, dz^2$ is real on the boundary. Lemma 8.1.6 implies $\varphi\, dz^2 \equiv 0$. Therefore h is conformal. Since $\frac{\partial h}{\partial x} = 0$ on $\partial\mathbb{C}_+$, then also $\frac{\partial h}{\partial y} = 0$ on $\partial\mathbb{C}_+$. Since h is harmonic and $\frac{\partial^2 h}{\partial x^2} = 0$ on $\partial\mathbb{C}_+$, the harmonic map equation gives also $\frac{\partial^2 h}{\partial y^2} = 0$ on $\partial\mathbb{C}_+$. Iteratively, all derivatives of h vanish on $\partial\mathbb{C}_+$. Hence we can reflect h smoothly as a harmonic and conformal map across $\partial\mathbb{C}_+$ via $h(z) = h(\overline{z})$ for $z = x+iy$ with $y < 0$. This means that we can reflect h to a harmonic and conformal map

$$h : S^2 \to N$$

mapping $\partial D = \{|z| = 1\}$ (considering D as a subset of S^2 as above) onto a single point.

In the sequel, we shall use the abbreviation

$$u_z := \frac{1}{2}\Big(\frac{\partial u}{\partial x} - i\frac{\partial u}{\partial y}\Big),$$
$$u_{\overline{z}} := \frac{1}{2}\Big(\frac{\partial u}{\partial x} + i\frac{\partial u}{\partial y}\Big),$$

even for functions $u : \mathbb{C} \to \mathbb{R}^d$, i.e. with real values, with componentwise differentiation. Thus, for every z_0,

$$u_z(z_0) \in \mathbb{C}^d.$$

$\square$

We now need

Lemma 8.1.7 (Hartman-Wintner). *Suppose Ω is a neighborhood of 0 in $\mathbb{C}$, $u \in C^2(\Omega, \mathbb{R}^d)$ satisfies*

$$|u_{z\overline{z}}| \le K|u_z| \tag{8.1.17}$$

for some constant K in Ω.

If

$$\lim_{z\to 0} u(z)z^{-n+1} = 0 \quad \textit{(assume the limit exists)} \tag{8.1.18}$$

for some $n \in \mathbb{N}$, then

$$\lim_{z\to 0} u_z(z)z^{-n}$$

exists. If (8.1.18) holds for all $n \in \mathbb{N}$, then

$$u \equiv 0. \tag{8.1.19}$$

Proof. For a compact subregion B of Ω with smooth boundary and $g \in C^1(B, \mathbb{C})$, we have the integration by parts formula

$$\oint_{\partial B} g u_z d\overrightarrow{n} = \int_B (u_z g_{\overline{z}} + u_{z\overline{z}} g) \frac{dz \wedge d\overline{z}}{2i}\,, \tag{8.1.20}$$

where $\overrightarrow{n}$ is the exterior normal of B.

We assume now

$$\lim_{z \to 0} u_z z^{1-k} = 0 \quad \text{for some } k \in \mathbb{N}. \tag{8.1.21}$$

We choose

$$B := \{z \in \mathbb{C} : \varepsilon \le |z| \le R, |z - w| \ge \varepsilon\}$$

with

$$0 < 3\varepsilon < R < \min\Big(\operatorname{dist}(0, \partial\Omega), \frac{1}{4k}\Big),$$
$$w \in \Omega, 2\varepsilon < |w| < R - \varepsilon,$$

and

$$g(z) = z^{-k}(z - w)^{-1}.$$

Then

$$g_{\overline{z}} \equiv 0 \quad \text{in } B.$$

(8.1.20) yields

$$\begin{aligned} \oint_{|z|=R} u_z z^{-k}(z-w)^{-1}|dz| - \oint_{|z|=\varepsilon} & u_z z^{-k}(z-w)^{-1}|dz| \\ & - \oint_{|z-w|=\varepsilon} u_z z^{-k}(z-w)^{-1}|dz| \\ & = \int_B u_{z\overline{z}} z^{-k}(z-w)^{-1} \frac{dz \wedge d\overline{z}}{2i}. \end{aligned} \tag{8.1.22}$$

We now let $\varepsilon \to 0$. Because of (8.1.21), the second integral on the left hand side of (8.1.22) then tends to 0. The third one tends to

$$2\pi u_z(w) w^{-k}$$

by Cauchy's integral formula. Consequently for $0 < |w| < R$

$$2\pi u_z(w) w^{-k} = \oint_{|z|=R} u_z z^{-k}(z-w)^{-1}|dz| - \int_{|z|\le R} u_{z\overline{z}} z^{-k}(z-w)^{-1} \frac{dz \wedge d\overline{z}}{2i}$$

and (8.1.17) implies for $0 < |w| < R$

$$\begin{aligned} 2\pi|u_z(w)w^{-k}| \le \oint_{|z|=R} |u_z z^{-k}(z-w)^{-1}||dz| \\ + K \int_{|z|\le R} |u_z||z|^{-k}|z-w|^{-1} \frac{dz \wedge d\overline{z}}{2i}. \end{aligned} \tag{8.1.23}$$

Two auxiliary points:

$$\begin{aligned} \int_{|w|\le R} |z-w|^{-1} \frac{dw \wedge d\overline{w}}{2i} &\le \int_{|z-w|\le 2R} |z-w|^{-1} \frac{dw \wedge d\overline{w}}{2i} \\ &\le 4\pi R \frac{1}{z-w} \frac{1}{w-z_0} \\ &= \frac{1}{z-z_0}\Big(\frac{1}{z-w} + \frac{1}{w-z_0}\Big). \end{aligned}$$

We then multiply (8.1.23) by $|w-z_0|^{-1}$ ($|z_0| < R$) and integrate w.r.t. w:

$$\begin{aligned} 2\pi \int_{|w|\le R} |u_z||w^{-k}||w-z_0|^{-1} \frac{dw \wedge d\overline{w}}{2i} \le 8\pi R \oint_{|z|=R} |u_z z^{-k}(z-z_0)^{-1}||dz| \\ +8\pi RK \int_{|z|\le R} |u_z||z|^{-k}|z-z_0|^{-1} \frac{dz \wedge d\overline{z}}{2i}. \end{aligned} \tag{8.1.24}$$

Hence, renaming some of the variables

$$(1-4RK) \int_{|z|\le R} |u_z||z|^{-k}|z-w|^{-1} \frac{dz \wedge d\overline{z}}{2i} \le 4R \oint_{|z|=R} |u_z z^{-k}(z-w)^{-1}||dz|\,. \tag{8.1.25}$$

The right hand side of (8.1.25) remains bounded as $w \to 0$ and consequently so does the left hand side. Then the right hand side of (8.1.23) remains bounded as $w \to 0$, and consequently also the left hand side. Therefore

$$\lim_{z\to 0} u_z(z) z^{-k} \tag{8.1.26}$$

exists.

If $k < n$, this limit then has to vanish because of (8.1.18), and hence (8.1.21) holds for $k+1$ instead of k.

The first assertion now follows by induction on k:

It is trivial for $n = 0$. For $n \ge 1$, (8.1.18) implies (8.1.21) for $k = 1$. By induction, we get (8.1.21) for $k = n$, and hence the limit in (8.1.26) exists which is the first assertion of the lemma.

For the second assertion, $k = n-1$ and $w \to 0$ in (8.1.25) gives

$$(1-4RK) \int_{|z|\le R} |u_z||z|^{-n} \frac{dz \wedge d\overline{z}}{2i} \le 4R \oint_{|z|=R} |u_z||z|^{-n}|dz| \tag{8.1.27}$$

for all n.

If $u \not\equiv 0$, there exists z_0 with $|z_0| < R$ and

$$|u_z(z_0)| = c \neq 0.$$

Then the left hand side of (8.1.27) would grow in u at least like $c|z_0|^{-n}$, the right hand side at most like $c' R^{-n}$, with $c' = 4R \sup_{|z|=R} |u_z|$. Since $|z_0| < R$, (8.1.27) then could not hold for all n This contradiction proves the second assertion. □

We can now easily conclude the

Proof of Theorem 8.1.3. We may assume of course that in local coordinates

$$h(\partial D) = 0.$$

In the same local coordinates as in the beginning of the proof, we have noted above that all derivatives of h vanish on $\partial\mathbb{C}_+$Thus, if e.g. 0 is in the image of our coordinate chart,

$$\lim_{z\to 0} h(z) z^{-n} = 0 \quad \text{for all } n \in \mathbb{N}.$$

Since h is harmonic

$$\begin{aligned} |h_{z\bar z}| &\le c_0 |h_{\bar z}||h_z| \\ &\le K|h_z|, \end{aligned}$$

in a neighborhood of 0 since h is smooth.

Lemma 8.1.7 then yields $h \equiv 0$ $(= h(\partial D))$. □

More generally, Lemma 8.1.7 implies

Corollary 8.1.6. *Let Σ be a Riemann surface, N a Riemannian manifold of dimension d, $h : \Sigma \to N$ harmonic.*

Then for each $z_0 \in \Sigma$ there exists $m \in \mathbb{N}$ with the property that in any local coordinates around $h(z_0)$, there exists $a \in \mathbb{C}^d$ with

$$h_z(z) = a(z - z_0)^m + 0(|z - z_0|^m) \tag{8.1.28}$$

for z near z_0.

If $h_z(z_0) = 0$, $m \ge 1$. In particular, the zeroes of h_z are isolated, unless h is constant.

If h is conformal, i.e.

$$g_{jk} h_z^j h_z^k = 0,$$

then

$$g_{jk}(h(z_0)) a^j a^k = 0.$$

Proof. We apply Lemma 8.1.7 with $u = h - h(z_0)$. As above, since h is harmonic and smooth

$$\begin{aligned} |h_{z\bar{z}}| &\leq c_0 |h_{\bar{z}}||h_z| \\ &\leq K|h_z|, \end{aligned}$$

so that (8.1.17) holds. All claims follow easily. □

We want to discuss a consequence of Theorem 8.1.3.

We look at (continuous) maps

$$f : D \to S^2$$

with

$$f(\partial D) \quad \text{a point, say the north pole.}$$

It is an elementary topological result that the homotopy classes of such maps can be parametrized by their degree, namely up to a constant factor, with $\omega := d\mathrm{Vol}\,(S^2)$, the volume form of S^2 for some Riemannian metric, by

$$\int_D f^*(\omega), \quad \text{in case } f \text{ is smooth.}$$

That $\int_D f^*(\omega)$, for smooth f, depends only on the homotopy class of f is a consequence of Stokes' theorem. Also, one easily constructs $f : D \to S^2$ for which this invariant is not zero. Consequently, not every map $f : D \to S^2$ with $f(\partial D)$ a point is homotopic to a constant map.

Corollary 8.1.7. *There exist smooth maps $f : D \to S^2$ mapping ∂D onto a point which are not homotopic to a harmonic map.*

Proof. By Theorem 8.1.3, any such harmonic map is constant, while not every smooth map as in the statement is homotopic to a constant map. □

Perspectives. In quantum field theory, harmonic maps occur as solutions to the nonlinear σ-problem. The supersymmetric version of this problem recently inspired an extension of the concept of harmonic maps, the so-called Dirac-harmonic maps[49, 50] that couple the map with a nonlinear spinor field while preserving the essential structural properties of harmonic maps. This will be presented in the next chapter.

The method of holomorphic quadratic differentials associated to two-dimensional geometric variational problems was introduced by H. Hopf. He considered the case of closed surfaces of constant mean curvature in $\mathbb{R}^3$ (cf. Exercise 4 of this chapter).

The applicability of the Hartman-Wintner Lemma to two-dimensional geometric variational problems was first discovered by E. Heinz.

8.2 The Existence of Harmonic Maps in Two Dimensions

We start with some simple topological preliminaries.

Let N be a manifold.

Definition 8.2.1. $\pi_2(N) = 0$ means that every continuous map

$$\varphi : S^2 \to N$$

is homotopic to a constant map.

Lemma 8.2.1. $\pi_2(N) = 0$

$$\iff \textit{Any } h_0, h_1 \in C^0(D, N) \textit{ with } h_{0|\partial D} = h_{1|\partial D} \textit{ are homotopic.}$$

Proof.

"$\Leftarrow$": Take $\eta : D \to S^2$ bijective on D with $\eta(\partial D) = p_0$. For $\varphi : S^2 \to N$ define $h_0 = \varphi \circ \eta, h_1 \equiv \varphi(p_0)$.

"$\Rightarrow$": Given h_0, h_1 we define $\varphi : S^2 \to N$ by

$$\begin{aligned} \varphi(p) &:= h_0(f_1(p)) \qquad \text{if } |f_1(p)| \le 1, \\ \varphi(p) &:= h_1(f_2(p)) \qquad \text{if } |f_2(p)| \le 1, \end{aligned}$$

where f_1, f_2 are the coordinate charts of §1.1.

φ is continuous since $h_{0_{|\{|z|=1\}}} = h_{1_{|\{|z|=1\}}}$. If $\pi_2(N) = 0$, there exists a continuous map

$$L : S^2 \times [0, 1] \to N$$

with

$$\begin{aligned} L_{|S^2 \times \{0\}} &= \varphi, \\ L_{|S^2 \times \{1\}} &= \text{const.} \end{aligned}$$

We now define a homotopy

$$H : D \times I \to N$$

by

$$\begin{aligned} H(z,t) &:= L(f_1^{-1}(2z), 2t) && \text{for } |z| \le \frac{1}{2}, 0 \le t \le \frac{1}{2}, \\ H(z,t) &:= L(f_2^{-1}(2z), 2(1-t)) && \text{for } |z| \le \frac{1}{2}, \frac{1}{2} \le t \le 1, \\ H(z,t) &:= L\Big(f_1^{-1}\Big(\frac{z}{|z|}\Big), 4t(1-|z|)\Big) && \text{for } \frac{1}{2} \le |z| \le 1,\ 0 \le t \le \frac{1}{2}, \\ H(z,t) &:= L\Big(f_2^{-1}\Big(\frac{z}{|z|}\Big), 4(1-t)(1-|z|)\Big) && \text{for } \frac{1}{2} \le |z| \le 1, \frac{1}{2} \le t \le 1. \end{aligned}$$

Then H is continuous, $H_{|\{|z|=1\}\times\{t\}} = h_{0_{|\{|z|=1\}}} = h_{1_{|\{|z|=1\}}}$ for all t, and

$$H_{|D\times\{0\}} \text{ is homotopic to } h_0,\ H_{|D\times\{1\}} \text{ to } h_1.$$

□

Remark. While the proof is formal, the claim of Lemma 8.2.1 should be geometrically obvious.

The first aim of this section is the proof of

Theorem 8.2.1. *Let Σ be a compact Riemann surface, N a compact Riemannian manifold with*

$$\pi_2(N) = 0.$$

Then any smooth $\varphi : \Sigma \to N$ is homotopic to a harmonic map $f : \Sigma \to N$. f can be constructed as a map which minimizes energy in its homotopy class.

We need to establish some auxiliary results before we can start the proof of Theorem 8.2.1.

We say that a *continuous* map

$$h : M \to N$$

between differentiable manifolds is of Sobolev class $H^{k,p}_{\mathrm{loc}}$ if it is of this class w.r.t. any coordinate charts on M and N. If M is compact, we can then also define Sobolev classes $H^{k,p}$ for continuous maps. For a better discussion of Sobolev spaces, see §7.3 below.

Lemma 8.2.2. *Let N be a Riemannian manifold, $B_0 \subset B_1 \subset N$, B_0, B_1 closed. Let $\pi : B_1 \to B_0$ be of class C^1,*

$$\pi_{|B_0} = \mathrm{id}_{|B_0} \tag{8.2.1}$$

and

$$\|D\pi(v)\| < \|v\| \quad \text{for every } x \in B_1\backslash B_0, v \in T_xN, v \neq 0. \tag{8.2.2}$$

Let M be a Riemannian manifold with boundary ∂M, and let

$$\begin{gathered} h \in C^0 \cap H^{1,2}(M, B_1), \\ h(\partial M) \subset B_0 \end{gathered} \tag{8.2.3}$$

be energy minimizing in the class of all maps from M into B_1 with the same boundary values as h.

Then

$$h(M) \subset B_0. \tag{8.2.4}$$

Proof. Let us assume that

$$\Omega := h^{-1}(B_1 \backslash B_0) \neq \emptyset.$$

Since h is continuous, Ω is open, and since $h(\partial M) \subset B_0$, h cannot be constant on Ω.

Thus

$$E(h_{|\Omega}) > 0.$$

But then by (8.2.2), since $\pi \circ h \in H^{1,2}$ as $\pi \in C^1$,

$$E(\pi \circ h) < E(h)$$

contradicting the minimizing property of h. (Note that $(\pi \circ h)_{|\partial M} = h_{|\partial M}$ by (8.2.1) and (8.2.3).)

Therefore Ω is empty. □

Lemma 8.2.3. *Let N be a Riemannian manifold, $B_0 \subset B_1 \subset N$, B_0, B_1 compact. Suppose that every point in $B_1 \backslash B_0$ can be joined inside $B_1 \backslash B_0$ to ∂B_0 by a unique geodesic normal to ∂B_0. Also assume that for any two such geodesics $\gamma_1(t)$, $\gamma_2(t)$, parametrized by arc length ($t \geq 0$) with $\gamma_i(0) \in \partial B_0$, $i = 1, 2$, we have*

$$d(\gamma_1(t), \gamma_2(t)) > d(\gamma_1(0), \gamma_2(0)) \quad \textit{for } t > 0. \tag{8.2.5}$$

Then the conclusion of Lemma 8.2.2 holds.

Proof. We define $\pi : B_1 \to B_0$ as the identity on B_0 and the projection along normal geodesics onto ∂B_0 on $B_1 \backslash B_0$, i.e. if $\gamma(t), t \geq 0$, is a geodesic normal to ∂B_0 inside $B_1 \backslash B_0$, with $\gamma(0) \in \partial B_0$, then $\pi(\gamma(t)) = \gamma(0)$. This map satisfies all the hypotheses of Lemma 8.2.2, except that it is only Lipschitz, but not C^1. It is not difficult, however, to approximate π by maps of class C^1 satisfying the same hypothesis, and the result then easily follows from Lemma 8.2.2. □

Lemma 8.2.4. *Let N be a Riemannian manifold, $p \in N$, $i(p)$ the injectivity radius of p, and suppose that the sectional curvature of N is bounded from above by κ, and let*

$$0 < \rho < \frac{1}{3} \min\Big(i(p), \frac{\pi}{2\sqrt{\kappa}}\Big). \tag{8.2.6}$$

Let M be a Riemannian manifold with boundary ∂M, and let $h \in C^0 \cap H^{1,2}(M, N)$ with

$$h(\partial M) \subset B(p, \rho) = \{q \in N : d(p, q) \leq \rho\}. \tag{8.2.7}$$

If h minimizes the energy among all maps with the same boundary values, then

$$h(M) \subset B(p, \rho). \tag{8.2.8}$$

Proof. By (8.2.6), we can introduce geodesic polar coordinates (r, φ) on $B(p, 3\rho)$ $(0 \le r \le 3\rho)$. We now define a map $\pi : N \to B(p, \rho)$, given in these coordinates by

$$\begin{aligned}
\pi(r, \varphi) &= (r, \varphi) && \text{if } r \le \rho, \\
\pi(r, \varphi) &= (\tfrac{3}{2}\rho - \tfrac{1}{2}r, \varphi) && \text{if } \rho \le r \le 3\rho, \\
\pi(q) &= p && \text{if } q \in N \backslash B(p, 3\rho).
\end{aligned}$$

Thus, π maps concentric spheres of radius $\le 3\rho$ onto concentric spheres of possibly smaller radius. It is clear that on $B(p, 3\rho) \backslash B(p, \rho)$, π is length decreasing in the r-direction. In order to see that π is also length decreasing in the φ-directions, let $\gamma(s)$ be a curve given in our coordinates by $(r, \varphi(s))$, i.e. a curve in the distance sphere $\partial B(p, r)$. For each fixed s, $c_s(t) := (t, \varphi(s))$ is a radial geodesic with $c_s(0) = p$, $c_s(r) = \gamma(s)$. Thus

$$J_s(t) := \frac{\partial}{\partial s} c_s(t)$$

is a Jacobi field, and

$$\dot{\gamma}(s) = J_s(r), 0 = J_s(0) \tag{8.2.9}$$

and

$$D\pi(\dot{\gamma}(s)) = J_s(r'), \quad \text{where } (r', \varphi) = \pi(r, \varphi), \tag{8.2.10}$$

i.e.

$$r' < \rho < r \le 3\rho. \tag{8.2.11}$$

The Rauch comparison theorem (Theorem 4.5.1) implies that (assume $\dot{\gamma}(s) \neq 0$)

$$\frac{|J_s(r)|}{|J_s(r')|} \ge \frac{\sin(\sqrt{\kappa} r)}{\sin(\sqrt{\kappa} r')} > 1, \quad \text{since } r' < r \le 3\rho < \frac{\pi}{2\sqrt{\kappa}}\,. \tag{8.2.12}$$

Consequently by (8.2.9), (8.2.10), (8.2.12)

$$|D\pi(\dot{\gamma}(s))| < |\dot{\gamma}(s)|, \quad \text{if } \dot{\gamma}(s)) \neq 0. \tag{8.2.13}$$

Therefore, π is also length decreasing in the φ-directions.

π is not C^1, but only Lipschitz. It can, however, be approximated by C^1-maps with the same length decreasing properties, and Lemma 8.2.2 then again gives the result. □

We shall also need the **Courant-Lebesgue-Lemma.**

Lemma 8.2.5. *Let N be a Riemannian manifold with distance function $d(\cdot, \cdot)$,*

$$u \in H^{1,2}(D, N)$$

with

$$E(u) \le K. \tag{8.2.14}$$

Then

$$\forall\, x_0 \in D, \delta \in (0,1)\, \exists\, \rho \in (\delta, \sqrt{\delta})\, \forall\, x_1, x_2 \in D \text{ with } |x_i - x_0| = \rho \ (i = 1, 2) :$$

$$d(u(x_1), u(x_2)) \le \frac{(8\pi K)^{\frac{1}{2}}}{(\log \frac{1}{\delta})^{\frac{1}{2}}} \,. \tag{8.2.15}$$

Proof. We first recall the following property of an $H^{1,2}$ function u:
For almost all $r > 0$, $u_{|\partial B(x_0,r)}$ is absolutely continuous. (See Lemma A.1.2)

Then for any such r and $x_1, x_2 \in D$ with $|x_i - x_0| = r$, $i = 1, 2$, we have

$$d(u(x_1), u(x_2)) \le \int_0^{2\pi} \Big\| \frac{\partial u(r,\varphi)}{\partial \varphi} \Big\| d\varphi \tag{8.2.16}$$

in polar coordinates (r, φ) with center x_0, w.l.o.g. $B(x_0, r) \subset D$; otherwise, the integration in (8.2.16) is only over those values of φ which correspond to $\partial B(x_0, r) \cap D$.

By Hölder's inequality

$$\int_0^{2\pi} \Big\| \frac{\partial u}{\partial \varphi} \Big\| d\varphi \le (2\pi)^{\frac{1}{2}} \Big(\int_0^{2\pi} \Big\| \frac{\partial u}{\partial \varphi} \Big\|^2 d\varphi \Big)^{\frac{1}{2}} . \tag{8.2.17}$$

The energy of u on $B(x_0, r)$ is

$$E(u_{|B(x_0,r)}) = \frac{1}{2} \int_0^{2\pi} \int_0^{r} \Big(\Big\| \frac{\partial u}{\partial \rho} \Big\|^2 + \frac{1}{\rho^2} \Big\| \frac{\partial u}{\partial \varphi} \Big\|^2 \Big) \rho d\rho d\varphi.$$

Consequently, there exists $\rho \in (\delta, \sqrt{\delta})$ with

$$\int_0^{2\pi} \Big\| \frac{\partial u(\rho,\varphi)}{\partial \varphi} \Big\|^2 d\varphi \le \frac{2E(u_{|B(x_0,\rho)})}{\int_\delta^{\sqrt{\delta}} \frac{1}{r}\, dr} \le \frac{2K}{-\frac{1}{2}\log \delta} = \frac{4K}{\log \frac{1}{\delta}} . \tag{8.2.18}$$

The claim follows from (8.2.16), (8.2.17), (8.2.18). □

As an intermediate result for the proof of Theorem 8.2.1, we now show

Theorem 8.2.2. *Let N be a complete Riemannian manifold with sectional curvature $\le \kappa$ and injectivity radius $i_0 > 0$, $p \in N$. Let*

$$0 < r < \min\Big(\frac{i_0}{2}, \frac{\pi}{2\sqrt{\kappa}} \Big). \tag{8.2.19}$$

Suppose $g : \partial D \to B(p, r) \subset N$ is continuous and admits an extension $\overline{g} : D \to B(p, r)$ of finite energy.

Then there exists a harmonic map

$$h : D \to B(p, r) \subset N$$

with

$$h_{|\partial D} = g,$$

and h minimizes energy among all such maps.

The modulus of continuity of h is controlled by r, κ, $E(\overline{g})$, and the modulus of continuity of g, i.e. given $\varepsilon > 0$, there exists $\delta > 0$ depending on r, κ, g such that $|x_1 - x_2| < \delta$ implies $d(h(x_1), h(x_2)) < \varepsilon$. Finally, for any $\sigma > 0$, the modulus of continuity of h on $\{z : |z| \le 1 - \sigma\}$ is controlled by σ, r, κ, and $E(\overline{g})$.

Proof. We choose r' with

$$r < r' < \min(\frac{i_0}{2}, \frac{\pi}{2\sqrt{\kappa}}). \tag{8.2.20}$$

Using the Rauch comparison theorem as in the proof of Lemma 8.2.4, one sees that

$$\pi : B(p, r') \to B(p, r),$$

with $\pi_{|B(p,r)} = \mathrm{id}$, and projecting $B(p, r') \backslash B(p, r)$ onto $\partial B(p, r)$ along radial geodesics satisfies the assumptions of Lemma 8.2.3.

As a first and preliminary application we show that any two points $p_1, p_2 \in B(p, r)$ can be joined inside $B(p, r)$ (and not just in N) by a unique shortest geodesic. For this purpose, we minimize

$$E(c)$$

in

$$\{c : [0, 1] \to B(p, r') : c(0) = p_1, c(1) = p_2\}.$$

As in §1.4, the infimum is realized by some curve c_0 with image in $B(p, r')$. Because of the distance decreasing properties of π, Lemma 8.2.3 (with $B_0 = B(p, r)$, $B_1 = (B(p, r'))$ implies that the image of c_0 is actually contained in the smaller ball $B(p, r)$. Therefore, we may perform arbitrarily small variations of c_0 without leaving $B(p, r')$. Therefore, c_0 is a critical point of E, hence geodesic by Lemma 9.2.1. Since $p_1, p_2 \in B(p, r)$, they can be joined inside $B(p, r)$ by a curve of length $\leq 2r < i_0$. Therefore, c_0 is the unique shortest geodesic between p_1 and p_2 by the definition of the injectivity radius i_0. This proves the claim about geodesic arcs. We note that c_0 is free from conjugate points, again by Rauch's comparison theorem (Theorem 4.5.1).

In order to find the harmonic map, we now minimize the energy in

$$V := \{v \in H^{1,2}(D, B(p, r')), v - \overline{g} \in H_0^{1,2}(D, B(p, r'))\}$$

(the latter is the weak formulation of the boundary condition). Since $B(p, r')$ is covered by a single coordinate system, namely normal coordinates, the $H^{1,2}$-property can be defined with the help of these coordinates.

A minimizing sequence has a subsequence converging in L^2 by Theorem A.1.8. We shall see below (Theorem 7.3.2), in order not to interrupt the present reasoning, that E is lower semicontinuous w.r.t. to L^2 convergence. Therefore, the limit h minimizes energy in V. By Lemma 8.2.3 again, $h(D)$ is contained in the smaller ball $B(p, r)$, hence a critical point of E because we may again perform arbitrarily small variations of h without leaving the class V.

We now want to show that h is continuous and control its modulus of continuity.

Let $q \in B(p, r), v_1, v_2 \in T_q N$ with $\|v_i\| = 1$, $i = 1, 2$,

$$c_i(t) = \exp_q(tv_i).$$

By Rauch's comparison theorem (Theorem 4.5.1) again, as in the proof of Lemma 8.2.4,

$$d(c_1(t), c_2(t)) \geq d(c_1(\varepsilon), c_2(\varepsilon)) \tag{8.2.21}$$

for

$$\varepsilon \le t \le \frac{\pi}{\sqrt{\kappa}} - \varepsilon\,.$$

With

$$\varepsilon_0 := \frac{\pi}{\sqrt{\kappa}} - 2r,$$

for any $0 < \varepsilon \le \varepsilon_0$,

$$\begin{aligned} B_0 &:= B(q,\varepsilon) \cap B(p,r), \\ B_1 &:= B(p,r) \end{aligned}$$

satisfy the assumptions of Lemma 8.2.3, as any geodesic

$$c(t) := \exp_q tv, \|v\| = 1 \quad (v \in T_q N, q \in B(p,r))$$

leaves $B(p,r)$ for $t \ge 2r$ (i.e. $c(t) \in B(p,r) \Rightarrow t \le 2r$; this is a consequence of (8.2.19) and the resulting uniqueness of geodesics in $B(p,r)$).

We now apply the Courant-Lebesgue Lemma (Lemma 8.2.5). Since h is energy minimizing,

$$E(h) \le E(\overline{g}).$$

For $0 < \varepsilon \le \varepsilon_0$, we compute $\delta \in (0,1)$ with

$$\left(\frac{8\pi E(\overline{g})}{\log \frac{1}{\delta}}\right)^{\frac{1}{2}} \le \varepsilon. \tag{8.2.22}$$

For any $x_0 \in D$, by Lemma 8.2.5 there exists ρ, $\delta \le \rho \le \sqrt{\delta}$, with the property that for any $x_1, x_2 \in D$ with $|x_i - x_0| = \rho$ $(i = 1,2)$,

$$d(h(x_1), h(x_2)) \le \varepsilon, \tag{8.2.23}$$

hence

$$h(\partial B(x_0,\rho) \cap D) \subset B(q,\varepsilon) \quad \text{for some } q \in N. \tag{8.2.24}$$

Since g is continuous, there also exists $\delta' > 0$ with

$$d(g(y_1), g(y_2)) \le \varepsilon, \tag{8.2.25}$$

whenever $y_1, y_2 \in \partial D$ satisfy $|y_1 - y_2| \le \delta'$.

We now require in addition to (8.2.22) that also

$$\sqrt{\delta} \le \delta'.$$

Since $h_{|\partial D} = g$, with ρ as above we then have

$$\begin{gathered} h(\partial(B(x_0,\rho) \cap D)) \subset B(q,\varepsilon) \quad \text{for some } q \in N, \\ (\partial(B(x_0,\rho) \cap D) = (\partial B(x_0,\rho) \cap D) \cup (\partial D \cap B(x_0,\rho)). \end{gathered} \tag{8.2.26}$$

Lemma 8.2.3 then implies

$$h(B(x_0,\rho)\cap D)\subset B(q,\varepsilon). \tag{8.2.27}$$

Likewise, $|x_0|+\rho<1$, then $\partial(B(x_0,\rho)\cap D)=\partial B(x_0,\rho)\cap D$, and so in this case, we do not need g to control h on $\partial(B(x_0,\rho)\cap D)$.

In particular,

$$h(B(x_0,\delta)\cap D)\subset B(q,\varepsilon) \tag{8.2.28}$$

for any $x_0\in D$ and some $q\in N$ (depending, of course, on x_0). (8.2.28) is the desired estimate of the modulus of continuity. The proof of smoothness of h is postponed until after the proof of Theorem 8.2.2 – see Theorem 8.3.1. □

Remark. We actually shall only need the weaker result that there exists $r_0>0$ with the property that for any $r\in(0,r_0)$, the conclusion of Theorem 8.2.2 holds. As an exercise, the reader should simplify the preceding proof in order to show this weaker statement. On the other hand, the injectivity radius i_0 in (8.2.19) can easily be replaced by $i_0(r):=\min\{i(q):q\in B(p,r)\}$, where $i(q)$ is the injectivity radius of q, without affecting the validity of the above proof. This remark is interesting for complete, but non compact manifolds N. In this case, one may have $i_0=0$, but one always has $i_0(r)>0$ for any $r>0$ as N is complete.

Finally, D may be replaced in Theorem 8.2.2 by any compact Riemann surface Σ with boundary $\partial\Sigma$, with only trivial modifications of the proof.

Proof of Theorem 8.2.1. We put

$$[\varphi]:=\{v\in C^0\cap H^{1,2}(\Sigma,N): v \text{ is homotopic to } \varphi\}.$$

We choose

$$\rho:=\frac{1}{3}\min\Big(i_0(N),\frac{\pi}{2\sqrt{\kappa}}\Big), \tag{8.2.29}$$

where $i_0(N)$ is the injectivity radius of N, and $\kappa\geq 0$ is an upper bound for the sectional curvature of N. We choose $\delta_0<1$ to satisfy

$$\left(\frac{8\pi E(\varphi)}{\log\frac{1}{\delta_0}}\right)^{\frac{1}{2}}\leq\frac{\rho}{2}. \tag{8.2.30}$$

For every $\delta\in(0,\delta_0)$, there exists a finite number of points $x_i\in\Sigma$, $i=1,\dots,m=m(\delta)$, for which the disks $B(x_i,\frac{\delta}{2})$ cover Σ. Here, we may define the disks $B(x_i,\frac{\delta}{2})$ w.r.t. any conformal metric on Σ. We may also arrange things so that around each x_i, there exists a coordinate chart f_i with image containing

$$\{z\in\mathbb{C}:|f(x_i)-z|\leq 1\}$$

and put

$$B(x_i,\delta):=\{z\in\mathbb{C}:|f(x_i)-z|\leq\delta\}.$$

We let $(u_n)_{n\in\mathbb{N}}$ be an energy minimizing sequence in $[\varphi]$. By definition of $[\varphi]$, all u_n then are continuous. Also, w.l.o.g.,

$$E(u_n) \le E(\varphi) \quad \text{for all } n. \tag{8.2.31}$$

Lemma 8.2.5 implies, recalling (8.2.30), that for every $n \in \mathbb{N}$, there exists $r_{n,1} \in (\delta, \sqrt{\delta})$ and $p_{n,1} \in N$ with

$$u_n(\partial B(x_1, r_{n,1})) \subset B(p_{n,1}, \rho). \tag{8.2.32}$$

On the other hand, if $u_n(\partial B(x,r)) \subset B(p,\rho)$ for some $x \in \Sigma$, $r > 0$, $p \in N$, then Theorem 8.2.2 (replacing D by $B(x,r)$) yields a solution of the Dirichlet problem

$$h : B(x,r) \to B(p,\rho) \qquad \text{harmonic and energy minimizing}$$

with

$$h_{|\partial B(x,r)} = u_{n|\partial B(x,r)}\,. \tag{8.2.33}$$

We replace u_n on $B(x_1, r_{n,1})$ by the solution of the Dirichlet problem (8.2.33) for $x = x_1$, $r = r_{n,1}$. Outside $B(x_1, r_{n,1})$, we leave u_n unaltered.

We denote the new map by u_n^1. Since $\pi_2(N) = 0$, by Lemma 8.2.1, u_n^1 is homotopic to u_n, hence to φ. Thus

$$u_n^1 \in [\varphi].$$

After selection of a subsequence, $(r_{n,1})_{n\in\mathbb{N}}$ converges to some $r_1 \in [\delta, \sqrt{\delta}]$. By the interior modulus of continuity estimate of Theorem 8.2.2, the maps (u_n^1) are uniformly continuous on $B(x_1, \delta-\eta)$ for any $\eta \in (0,\delta)$. Moreover, by Lemma 8.2.4, u_n^1 minimizes the energy not only among maps into $B(p,\rho)$, but among all maps into N with the same boundary values.

Thus

$$E(u_n^1) \le E(u_n). \tag{8.2.34}$$

Repeating the above argument, we find radii $r_{n,2} \in (\delta, \sqrt{\delta})$ with

$$u_n^1(\partial B(x_2, r_{n,2})) \subset B(p_{n,2}, \rho)$$

for points $p_{n,2} \in N$. We replace u_n^1 on $B(x_2, r_{n,2})$ by the solution of the Dirichlet problem (8.2.33) for $x = x_2$, $r = r_{n,2}$. Again by selecting a subsequence, $(r_{n,2})_{n\in\mathbb{N}}$ converges to some $r_2 \in [\delta, \sqrt{\delta}]$. The new maps u_n^2 are again homotopic to φ, i.e.

$$u_n^2 \in [\varphi],$$

because $\pi_2(N) = 0$.

Since the maps u_n^1 are equicontinuous on $B(x_1, \delta - \frac{\eta}{2})$ whenever $0 < \eta < \delta$, the boundary values for our second replacement are equicontinuous on

$$\partial B(x_2, r_{n,2}) \cap B(x_1, \delta - \frac{\eta}{2}).$$

Therefore, using the estimates of the modulus of continuity in the proof of Theorem 8.2.2, the maps u_n^2 are equicontinuous on $B(x_1, \delta - \eta) \cup B(x_2, \delta - \eta)$ for any η with $0 < \eta < \delta$.

By Lemma 8.2.4 and (8.2.34)

$$E(u_n^2) \le E(u_n^1) \le E(u_n) \tag{8.2.35}$$

as before.

We repeat the replacement argument on disks centered at $x_3, \dots, x_m$.

We obtain a sequence $v_n := u_n^m \in [\varphi]$ with

$$E(v_n) \le E(u_n) \le E(\varphi), \tag{8.2.36}$$

which is equicontinuous on every disk $B(x_i, \frac{\delta}{2})$, $i = 1, \dots, m$, hence on Σ because these disks cover Σ.

After selection of a subsequence, $(v_n)_{n\in\mathbb{N}}$ converges uniformly to some map u which then also is homotopic to φ. $(v_n)_{n\in\mathbb{N}}$ then also converges in L^2 to u.

By Theorem 7.3.2 we have the lower semicontinuity

$$E(u) \le \liminf_{n\to\infty} E(v_n). \tag{8.2.37}$$

Since $u \in [\varphi]$ and (u_n), hence also (v_n) by (8.2.36) was a minimizing sequence for the energy in $[\varphi]$, (8.2.37) implies that u minimizes energy in $[\varphi]$. In particular, u is energy minimizing when restricted to small balls. Either from this observation and Lemma 8.2.4 and Theorem 8.2.2 or alternatively directly from the construction of u, the modulus of continuity of u is controlled by the geometry of N, more precisely by $i_0(N)$ and κ, and by $E(\varphi)$. Smoothness of u follows from Theorem 8.3.1. $\square$

With the same argument, one also shows:

Theorem 8.2.3. *Let Σ be a compact Riemann surface with boundary $\partial\Sigma$, N a compact Riemannian manifold with $\pi_2(N) = 0$, $\varphi \in C^0 \cap H^{1,2}(\Sigma, N)$. Then there exists a harmonic map*

$$u : \Sigma \to N$$

homotopic to φ with

$$u_{|\partial\Sigma} = \varphi_{|\partial\Sigma},$$

and u can be chosen to minimize energy among all such maps. $\square$

Remark. If one does not assume $\pi_2(N) = 0$, one still obtains a harmonic map $u : \Sigma \to N$ with $u_{|\partial\Sigma} = \varphi_{|\partial\Sigma}$ by our reasoning. In that case, however, u need not be homotopic to φ any more. u can be chosen to minimize the energy among all maps with boundary values given by φ.

In the sequel, we shall need the following covering lemma:

Lemma 8.2.6. *For any compact Riemannian manifold M, there exists $\Lambda \in \mathbb{N}$ with the following property: whenever we have points $x_1, \dots, x_m \in M$ and $\rho > 0$ with*

$$X \subset \bigcup_{i=1}^{m} B(x_i, \rho)$$

and

$$x_i \notin B(x_j, \rho) \quad \text{for } i \neq j,$$

then $\{1, \dots, m\}$ is the disjoint union of Λ sets $I_1, \dots, I_\Lambda$ so that for all $\ell \in \{1, \dots, \Lambda\}$ and $i_1, i_2 \in I_\ell$, $i_1 \neq i_2$,

$$B(x_{i_1}, 2\rho) \cap B(x_{i_2}, 2\rho) = \emptyset.$$

Proof. We construct I_1: We first put $x_1^1 := x_1$ and iteratively seek points $x_j^1 \in \{x_1, \dots x_m\}$ with

$$4\rho < d(x_j^1, x_i^1) \quad \text{for all } i < j,$$

until no such point can be found anymore. I_1 is the set of points selected so far. If $x_k \notin I_1$, there exists $x_j^1 \in I_1$ with

$$d(x_k, x_j^1) \leq 4\rho.$$

We construct I_ℓ iteratively for $\ell \geq 2$: We select any $x_k \notin \bigcup_{\lambda=1}^{\ell-1} I_\lambda$, put $x_1^\ell := x_k$ and iteratively seek points $x_j^\ell \in \{x_1, \dots, x_m\} \setminus \bigcup_{\lambda=1}^{\ell-1} I_\lambda$ with

$$4\rho < d(x_j^\ell, x_i^\ell) \quad \text{for all } i < j$$

until no such point can be found anymore.

If $x_k \notin I_\ell$, then for each $\lambda \leq \ell$, we can find some $x_{j(\lambda)}^\lambda \in I_\lambda$ with

$$d(x_k, x_{j(\lambda)}^\lambda) \leq 4\rho.$$

All these points $x_{j(\lambda)}^\lambda$ are distinct, and their mutual distance is bounded from below by ρ by our assumptions. Therefore, there exists some $\Lambda_0 \in \mathbb{N}$ such that there exists at most Λ_0 points $x_{j(\lambda)}^\lambda$ satisfying the preceding inequality. The reader should by now have acquired enough familiarity with the local geometry of Riemannian manifolds to verify the existence of such a Λ_0 with the required properties. The claim follows with $\Lambda := \Lambda_0 + 1$. □

Remark. It is easy to see that one may always construct coverings satisfying the assumption $x_i \notin B(x_j, \rho)$ for $i \neq j$.

We now come to the important phenomenon of splitting off of minimal 2-spheres. Before giving a general theorem below, we first want to isolate the phenomenon in a simpler situation:

Theorem 8.2.4. *Let Σ be a compact Riemann surface, N a compact Riemannian manifold*
$$u_n : \Sigma \to N$$
a sequence of harmonic maps with
$$E(u_n) \le K \quad \textit{for some constant } K.$$
Then either the maps u_n are equicontinuous, and hence a subsequence converges uniformly to a harmonic map $u : \Sigma \to N$, or there exists a nonconstant conformal harmonic map
$$v : S^2 \to N,$$
i.e. a (parametric) minimal 2-sphere in N.

Proof. Let
$$\lambda_n := \sup_{z\in\Sigma} \|du_n(z)\|.$$
We distinguish two cases.

1) $\sup\limits_{n\in\mathbb{N}} \lambda_n < \infty$.

   Then $(u_n)_{n\in\mathbb{N}}$ is equicontinuous, because the derivatives are uniformly bounded. A priori estimates (see §8.3) imply that also higher derivatives of (u_n) are equibounded. By the Arzela-Ascoli theorem, a subsequence converges uniformly, and by these regularity results the limit is also harmonic. Alternatively, the limit is continuous and weakly harmonic, hence smooth and harmonic by Theorem 8.3.1.

2) $\sup \lambda_n = \infty$.

   After selection of a subsequence, λ_n tends monotonically to ∞, and a sequence $(z_n)_{n\in\mathbb{N}} \subset \Sigma$ with
$$\|du_n(z_n)\| = \sup_{z\in\Sigma} \|du_n(z)\| \quad (= \lambda_n)$$
   has a limit point z_0.

We choose suitable local coordinates for which
$$\{z : |z - z_0| \le 2\}$$
is contained in a coordinate chart. All local expressions will be evaluated in this chart. We put
$$D_n := \{w \in \mathbb{C} : |w| \le \lambda_n\}$$
and define
$$v_n : D_n \to N$$

by

$$v_n(w) := u_n\Big(z_0 + \frac{w}{\lambda_n}\Big).$$

By definition of λ_n,

$$\sup_{w \in D_n} \|dv_n(w)\| = 1.$$

By conformal invariance of E

$$E(v_n) \le K.$$

As $n \to \infty$, D_n exhausts all of $\mathbb{C}$. By regularity results for harmonic maps (see §8.3) after selection of a subsequence, $(v_n)_{n\in\mathbb{N}}$ converges uniformly on compact subsets of $\mathbb{C}$ to a harmonic map

$$v : \mathbb{C} \to N.$$

Actually, the convergence takes place even in C^2, by a priori estimates for harmonic maps, see §8.3 and therefore

$$\|dv(0)\| = 1,$$

and v is not constant. Also, $E(v) \le K$.

The holomorphic quadratic differential defined by v,

$$g_{ij}(v(z))v_z^i v_z^j dz^2,$$

((g_{ij}) being the metric of N in local coordinates) therefore yields a holomorphic function

$$\psi(z) = g_{ij}(v(z))v_z^i v_z^j$$

of class L^1, since

$$\int_{\mathbb{C}} |\psi| \le E(v).$$

By a variant of Liouville's theorem, see Lemma 8.2.7 below,

$$\psi \equiv 0$$

and it follows that v is conformal (see §8.1). It remains to show that v extends as a harmonic and conformal map

$$v : S^2 \to N$$

where we consider S^2 as $\mathbb{C} \cup \{\infty\}$. Thus, one has to show that ∞ is a removable singularity. In §7.3, it will be shown more generally that conformal harmonic maps of finite energy on a Riemann surface cannot have isolated singularities. $\square$

Theorem 8.2.5. *Let Σ be a compact Riemann surface, possibly with boundary $\partial\Sigma$, N a compact Riemannian manifold, $\varphi \in C^0 \cap H^{1,2}(\Sigma, N)$. Then there exists a harmonic map*

$$u : \Sigma \to N$$

homotopic to φ, with $u_{|\partial\Sigma} = \varphi_{|\partial\Sigma}$ in case $\partial\Sigma \neq \emptyset$, or there exists a nontrivial conformal harmonic map

$$v : S^2 \to N.$$

i.e. a (parametric) minimal 2-sphere in N.

Proof. We only treat the case $\partial\Sigma = \emptyset$. The case $\partial\Sigma \neq \emptyset$ is handled with easy modifications of the argument for $\partial\Sigma = \emptyset$.

We let

$$\rho := \frac{1}{3}\min\Big(i(N), \frac{\pi}{2\sqrt{\kappa}}\Big), \tag{8.2.38}$$

where $i(N)$ is the injectivity radius of N, and $\kappa \geq 0$ is an upper curvature bound.

We choose a conformal metric on Σ. All distances on Σ will be computed w.r.t. this metric.

We let

$$r_0 := \sup\{R > 0 : \forall\, x \in \Sigma\, \exists\, p \in N : \varphi(B(x, 2R)) \subset B(p, 3^{-\Lambda}\rho)\}, \tag{8.2.39}$$

where Λ is the integer of Lemma 8.2.6 for $M = \Sigma$.

According to Lemma 8.2.6, there exist finite sets $I_1, \ldots, I_\Lambda$ and points $x_i \in \Sigma$ with

$$\Sigma = \bigcup_{\ell=1}^{\Lambda} \bigcup_{i \in I_\ell} B(x_i, r_0) \tag{8.2.40}$$

and

$$B(x_{i_1}, 2r_0) \cap B(x_{i_2}, 2r_0) = \emptyset, \quad \text{whenever } i_1, i_2 \in I_\ell, i_1 \neq i_2, \text{for some } \ell. \tag{8.2.41}$$

We then replace φ on every disk $B(x_i, 2r_0)$ for $i \in I_1$ by the solution of the Dirichlet problem (8.2.33) for $x = x_i$, $r = 2r_0$. This is possible by Theorem 8.2.1. Since the disks $B(x_i, 2r_0)$ for $i \in I_1$ are disjoint by (8.2.41), we can carry out these replacements simultaneously. We obtain a map

$$u_0^1 : \Sigma \to N$$

with

$$E(u_0^1) \leq E(\varphi) \tag{8.2.42}$$

as in the proof of Theorem 8.2.1.

Since

$$u_0^1(B(x_i, 2r_0)) \subset B(p_i, 3^{-\Lambda}\rho) \tag{8.2.43}$$

for every $i \in I_1$ and some $p_i \in N$ by the maximum principle Lemma 8.2.4, we obtain from the definition of r_0 and the triangle inequality

$$u_0^1(B(x, 2r_0)) \subset B(p, 3^{-\Lambda+1}\rho) \tag{8.2.44}$$

for every $x \in \Sigma$ and some $p \in N$ (depending on x).

Having constructed u_0^ℓ for $1 \le \ell \le \Lambda - 1$, we construct $u_0^{\ell+1}$ by replacing u_0^ℓ on every disk $B(x_i, 2r_0)$, $i \in I_{\ell+1}$, by the solution of (8.2.33) for $x = x_i$, $r = 2r_0$. We obtain

$$E(u_0^{\ell+1}) \le E(u_0^\ell) \tag{8.2.45}$$

and

$$u_0^{\ell+1}(B(x, 2r_0)) \subset B(p, 3^{-\Lambda+\ell+1}\rho) \tag{8.2.46}$$

for every $x \in \Sigma$ and some $p \in N$ (depending on x).

We thus arrive at a map

$$u_1 := u_0^\Lambda : \Sigma \to N$$

with

$$E(u_1) \le E(\varphi) \tag{8.2.47}$$

and

$$u_1(B(x, 2r_0)) \subset B(p, \rho)$$

for every $x \in \Sigma$ and some $p = p(x) \in N$.

Having iteratively constructed $u_n : \Sigma \to N$, we construct u_{n+1} by replacing φ by u_n and r_0 by

$$r_n = \sup\{R > 0 : \forall\, x \in \Sigma\, \exists\, p \in N : u_n(B(x, 2R)) \subset B(p, 3^{-\Lambda}\rho)\}.$$

The maps $(u_n)_{n\in\mathbb{N}}$ satisfy

$$E(u_n) \le E(u_{n-1}) \le E(\varphi). \tag{8.2.48}$$

We now distinguish two cases.

1) $s := \inf_{n\in\mathbb{N}} r_n > 0$.

   We claim that in this case $(u_n)_{n\in\mathbb{N}}$ converges to a harmonic map $u : \Sigma \to N$ homotopic to φ.

   We shall first show that the u_n are equicontinuous. We note that for every n, there exist finite sets $I_1, \dots, I_\Lambda$ and points $x_i \in \Sigma$ (everything depending on n, except for Λ) with

$$\Sigma = \bigcup_{\ell=1}^{\Lambda} \bigcup_{i\in I_\ell} B(x_i, r_n), \tag{8.2.49}$$

$$B(x_{i_1}, 2r_n) \cap B(x_{i_2}, 2r_n) = \emptyset, \tag{8.2.50}$$

   whenever $i_1 \ne i_2$, $i_1, i_2 \in I_\ell$ for some ℓ, by Lemma 8.2.6 again.

   By (8.2.49), for every $x \in \Sigma$, there exists some $i \in \bigcup_{\ell=1}^{\Lambda} I_\ell$ with

$$B(x, s) \subset B(x_i, 2r_n). \tag{8.2.51}$$

   There exists $\ell, 1 \le \ell \le \Lambda$, with $i \in I_\ell$. Therefore

$$u^\ell_{n|B(x,s)}$$

is harmonic, since it is even harmonic on the larger disk $B(x_i, 2r_n)$ (u_n^ℓ is constructed in the manner as u_0^ℓ with u_n instead of φ.)

Given ε with $0 < \varepsilon < \rho$ we consider δ with

$$0 < \delta < \min(1, s) \tag{8.2.52}$$

and

$$\left(\frac{8\pi E(\varphi)}{\log \frac{1}{\delta^2}}\right)^{\frac{1}{2}} \leq 3^{-\Lambda}\varepsilon. \tag{8.2.53}$$

For every $x \in \Sigma$, and $n \in \mathbb{N}$ there exists $R_1(x)$ with

$$\delta^2 < R_1(x) < \delta$$

and some $p_1 \in N$ with

$$u_n^\ell(\partial B(x, R_1(x))) \subset B(p_1, 3^{-\Lambda}\varepsilon) \tag{8.2.54}$$

by Lemma 8.2.5. Here ℓ is chosen as in (8.2.51), i.e. so that $i \in I_\ell$ for the i occuring in (8.2.51).

Since

$$u^\ell_{n|B(x,R_1(x))}$$

is harmonic and energy minimizing from Lemma 8.2.4 and (8.2.54),

$$u_n^\ell(B(x, R_1(x))) \subset B(p_1, 3^{-\Lambda}\varepsilon). \tag{8.2.55}$$

We likewise find $R_2(x)$ with

$$\delta^3 < R_2(x) < \delta^2$$

and

$$u_n^{\ell+1}(\partial B(x, R_2(x))) \subset B(p_2, 3^{-\Lambda}\varepsilon)$$

for some $p_2 \in N$. $u_n^{\ell+1}$ need no longer be harmonic on $B(x, R_2(x))$. It is only piecewise harmonic in case

$$\gamma := B(x, R_2(x)) \cap \bigcup_{i \in I_{\ell+1}} \partial B(x_i, 2r_n) \neq \emptyset.$$

Since

$$u_n^{\ell+1}(\gamma) = u_n^\ell(\gamma) \subset B(p_1, 3^{-\Lambda}\varepsilon)$$

and

$$u_n^{\ell+1}(\gamma \cap \partial B(x, R_2(x))) \subset B(p_2, 3^{-\Lambda}\varepsilon),$$

we obtain

$$u_n^{\ell+1}(\gamma \cup \partial B(x, R_2(x))) \subset B(p_2, 3^{-\Lambda+1}\varepsilon).$$

Therefore, the image of the boundary of every subregion of $B(x, R_2(x))$ on which $u_n^{\ell+1}$ is harmonic is contained in $B(p_2, 3^{-\Lambda+1}\varepsilon)$, and since of course all maps are energy minimizing on these subregions, Lemma 8.2.4 gives as usually

$$u_n^{\ell+1}(B(x, R_2(x))) \subset B(p_2, 3^{-\Lambda+1}\varepsilon). \tag{8.2.56}$$

Iterating, we obtain

$$R(x) > \delta^{\Lambda} \tag{8.2.57}$$

and $p = p(x) \in N$ with

$$u_{n+1}(B(x, R(x))) \subset B(p, \varepsilon) \tag{8.2.58}$$

(note $u_{n+1} = u_n^{\Lambda}$).

This proves equicontinuity, since δ and Λ are independent of u and x.

Therefore, after selection of a subsequence, $(u_n)_{n\in\mathbb{N}}$ converges to some map u homotopic to φ, and by (8.2.48) and lower semicontinuity of E (cf. Theorem 7.3.2),

$$E(u) \le \lim_{n\to\infty} E(u_n) \le E(\varphi). \tag{8.2.59}$$

We want to show that u is harmonic.

Replacing r_n by s, we may assume that the points x_i, $i \in \cup I_\ell$, are independent of n. (One may assume, by selecting a subsequence, that the points $x_i(n)$ converge to points x_i, and also $r_n \to s$ as $n \to \infty$.)

We first claim that with $(u_n)_{n\in\mathbb{N}}$ also $(u_n^1)_{n\in\mathbb{N}}$ converges to u, and that u is harmonic on every disk $B(x_i, s)$ for $i \in I_1$.

Since

$$E(u_{n+1}) = E(u_n^{\Lambda}) \le E(u_n^1) \le E(u_n), \tag{8.2.60}$$

$$\lim_{n\to\infty} (E(u_n) - E(u_n^1)) = 0. \tag{8.2.61}$$

Therefore, on each disk $B(x_i, s), i \in I_1$, for sufficiently large n the energy of u_n deviates only by an arbitrarily small amount from the energy of the energy minimizing map

$$u^1_{n|B(x_i,s)}.$$

Consequently, considering the gradient DE of the energy as in Section 6.11, we obtain

$$DE(u_{n|B(x_i,s)}) \to 0 \quad \text{for } i \in I_1$$

Since the maps u_n converge uniformly, the same argument as in the proof of Theorem 6.11.1 shows that

$$u_{|B(x_i,s)} = \lim_{n\to\infty} u_{n|B(x_i,s)}$$

is harmonic (and energy minimizing), and then also

$$u_{|B(x_i,s)} = \lim_{n\to\infty} u^1_{n|B(x_i,s)} \text{ for } i \in I_1. \tag{8.2.62}$$

Having iteratively shown that $(u_n^\ell)_{n\in\mathbb{N}}$ for some ℓ, $1 \le \ell \le \Lambda - 1$, converges to u and that u is harmonic on every disk $B(x_i, s)$ for $i \in I_\ell$, we show in the same manner that $(u_n^{\ell+1})_{n\in\mathbb{N}}$ likewise converges to u and that u is harmonic on every disk $B(x_i, s)$, $i \in I_{\ell+1}$.

We conclude that u is harmonic on $B(x_i, s)$ for every $i \in I_\ell$ and every $\ell \in \{1, \dots, \Lambda\}$, hence on all of Σ.

2) The second case is

$$\inf_{n\in\mathbb{N}} r_n = 0.$$

By selecting a subsequence, we may assume that $(r_n)_{n\in\mathbb{N}}$ is monotonically decreasing and converges to 0.

By definition of r_n, for every u, there exist points $y_0, y_1 \in \Sigma$ with

$$d(y_0, y_1) = 2r_n, \tag{8.2.63}$$

$$d(u_n(y_0), u_n(y_1)) \ge 3^{-1}\rho =: \rho_0. \tag{8.2.64}$$

We choose local coordinates around y_0 and denote the coordinate representations of y_0 and y_1 again by y_0 and y_1 resp.

For $z \in \mathbb{C}$, we put

$$k_n(z) := y_0 + r_n z$$

whenever this defines a point in our coordinate chart, and

$$\tilde{u}_n(z) := u_n(k_n(z)).$$

We thus have maps

$$\tilde{u}_n : \Omega_n \to N$$

with $\Omega_n \subset \mathbb{C}$ and $\Omega_n \to \mathbb{C}$ as $n \to \infty$ (i.e., in the limit, the domain of definition of k_n becomes the whole complex plane $\mathbb{C}$, since $r_n \to 0$). Since k_n is conformal, the maps $\tilde{u}_n$ are piecewise harmonic in the same manner the maps u_n are (see Corollary 8.1.3).

The maps $\tilde{u}_n$ now are equicontinuous by the same argument as in case 1 for $s = 1$ because for every $w_0 \in \Omega_n$ (with $B(w_0, 2) \subset \Omega_n$) there exists $p \in N$ with

$$\tilde{u}_n(B(w_0, 2)) \subset B(p, 3^{-1}\rho), \tag{8.2.65}$$

by definition of r_n, because $k_n(B(w_0, 2))$ is a ball of radius $2r_n$ (w.l.o.g., we may assume that the chosen metric on Σ coincides with the Euclidean one on our coordinate chart around y_0, as a different metric would only introduce some fixed factor in our estimates for the ball radii on Σ and Ω_n).

Likewise, as in case 1, after selection of a subsequence the maps $(\tilde{u}_n)$ converge uniformly on compact subsets to a harmonic map

$$v : \mathbb{C} \to N.$$

Moreover, by Corollary 8.1.4

$$\begin{aligned} E(\tilde{u}_{n|\Omega_n}) &= E(u_{n|k_n(\Omega_n)}) \\ &\le E(u_n) \\ &\le E(\varphi), \end{aligned}$$

hence by lower semicontinuity of E (Theorem 7.3.2)

$$E(v) \le \liminf_{n\to\infty} E(\tilde{u}_n) \le E(\varphi).$$

The holomorphic quadratic differential associated to v,

$$g_{jk}(v(z))v_z^j v_z^k dz^2$$

((g_{jk}) being the metric of N in local coordinates), therefore defines a holomorphic function

$$\psi(z) := g_{jk}(v(z))v_z^j v_z^k$$

of class L^1, because

$$\int_{\mathbb{C}} |\psi| \le 2E(v).$$

Since every holomorphic function on $\mathbb{C}$ of class L^1 vanishes identically (this follows by applying Lemma 8.2.7 below to the real and imaginary parts of ψ), we get $\psi \equiv 0$, and consequently v is conformal (see the discussion in §8.1).

It remains to show that v extends as a harmonic (and then also conformal) map

$$v : S^2 \to N,$$

i.e. that the singularity at ∞ is removable. This will be achieved in §8.3.

□

Corollary 8.2.1. *Let N be a Riemannian manifold with $\pi_2(N) \neq 0$. Then there exists a nonconstant conformal harmonic $v : S^2 \to N$, i.e. a (parametric) minimal 2-sphere in N.*

Proof. Since $\pi_2(N) \neq 0$, there exists $\varphi : S^2 \to N$ which is not homotopic to a constant map. By Theorem 8.2.5 either φ is homotopic to a harmonic map $v : S^2 \to N$ which then is also conformal by Corollary 8.1.5, or if the second alternative of Theorem 8.2.5 holds, there also exists a conformal harmonic $v : S^2 \to N$. □

Lemma 8.2.7. *Any harmonic function h defined on all $\mathbb{R}^n$ and of class $L^1(\mathbb{R}^n)$ is identically zero.*

Proof. By the mean value property of harmonic functions on $\mathbb{R}^n$,

$$|h(x_0)| = \frac{1}{\mathrm{Vol}\,(B(x_0,R))}\left|\int_{B(x_0,R)} h(x)\,dx\right|, \tag{8.2.66}$$

for any $\mathbb{R} > 0$, $x_0 \in \mathbb{R}^n$.

Since

$$\left|\int_{B(x_0,R)} h(x)\,dx\right| \leq \int_{B(x_0,R)} |h(x)|\,dx \leq \|h\|_{L^1(\mathbb{R}^n)}\,,$$

the r.h.s. of (8.2.66) tends to 0 as $R \to \infty$. Thus $h(x_0) = 0$. This holds for any $x_0 \in \mathbb{R}^n$. □

Perspectives. Theorem 8.2.1 is due to Lemaire[178] and Sacks and Uhlenbeck[219]. Theorem 8.2.5 is again due to Sacks and Uhlenbeck[219]. Other approaches to these results were found by Struwe[242], Chang[41] and Jost, see [136]. A detailed proof of Theorem 8.2.1 is given in [144].

The method of M. Struwe and K.C. Chang consists in studying the associated parabolic problem. Thus, given $\varphi : \Sigma \to N$, one studies solutions of

$$\begin{aligned} f : \Sigma \times [0,\infty) &\to N, \\ f(z,0) &= \varphi(\tau), \\ \frac{\partial f}{\partial t}(z,t) &= \tau(f(z,t)), \end{aligned}$$

where the tension field is computed w.r.t. the z variable. One can then show that a solution can develop at most finitely many singularities. These singularities correspond to the splitting off of minimal 2-spheres. In the limit $t \to \infty$, one obtains a harmonic map f.

The construction presented here is refined in [136]. There, also various existence results for unstable harmonic maps are presented. Any type of critical point theory, e.g. Morse theory, for harmonic maps in two dimensions has to take the splitting off of minimal 2-spheres into account. In certain instances, however, one may show that this phenomenon can be excluded. A prototype of such a result is the following

Theorem. *Let Σ be a compact Riemann surface with boundary, N a Riemannian manifold diffeomorphic to S^2 (thus, the condition $\pi_2(N) = 0$ is* **not** *satisfied). Let $g : \partial\Sigma \to N$ nonconstant. Then there exist at least two harmonic maps $f_1, f_2 : \Sigma \to N$ with $f_{i|\partial\Sigma} = g$.*

This result is due to Brézis and Coron[32] and Jost[133].

In order to prove this theorem, one first minimizes the energy over all maps $f : \Sigma \to N$ with $f_{|\partial\Sigma} = g$ and obtains a harmonic u (see the remark after Theorem 8.2.3). By careful comparison constructions one then exhibits another homotopy class α of maps from Σ to N (not containing u) with

$$\inf\{E(f) : f \in \alpha\} < E(u) + \mathrm{Area}\,(N).$$

One then shows that if minimizing energy in some homotopy class leads to the splitting off of a minimal 2-sphere, the energy would be lowered by an amount of at least the energy of that minimal sphere. Since N is diffeomorphic to S^2, the energy of such a minimal sphere would be at least the area of N. Since, however, u realizes the absolute minimum of energy among all maps with the prescribed boundary values, the above inequality excludes the splitting off of a minimal 2-sphere during the minimization of the energy in the class α.

We have described the preceding argument in some detail because it forms a paradigm for other conformally invariant variational problems (Yang-Mills equations in four dimensions, constant mean curvature surfaces, Yamabe problem, etc.). Some further discussion of such limit cases of the Palais-Smale condition may be found in [243] and in the references given there.

Returning to the critical point theory for two dimensional harmonic maps, we also mention Ding[61] and the survey article [137] where many further references can be found.

In this context, we should also discuss the Plateau problem for minimal surfaces. In its simplest form, we consider a smooth (or, more generally, a rectifiable) closed Jordan curve γ in $\mathbb{R}^3$ and seek a minimal surface with boundary γ. In the parametric version of the problem, we look for a harmonic and conformal $f : D \to \mathbb{R}^3$ (D = unit disk) mapping ∂D monotonically onto γ (a monotonic map between curves is defined to be a uniform limit of homeomorphisms). In this form, the problem was solved by J. Douglas and T. Radó. The problem was then extended by Douglas to configurations of more than one disjoint curves $\gamma_1, \ldots, \gamma_k$ and/or minimal surfaces of other topological type. He found a condition (the so-called Douglas condition) guaranteeing the existence of minimal surfaces of some prescribed topological type. It was also asked whether one may find unstable minimal surfaces with prescribed boundary. The most comprehensive critical point theory for minimal surfaces in $\mathbb{R}^3$ was developed in Jost and Struwe[151] where also references to earlier contributions are given.

The Plateau problem in Riemannian manifolds (instead of just $\mathbb{R}^3$) was solved by C. Morrey[199]. Results pointing into the direction of a general Morse theory for minimal surfaces in Riemannian manifolds may be found in Jost[136].

There also exists the geometric measure theory approach to minimal surfaces. Here, one tries to represent a minimal surface not as the image of a map of a Riemann surface, but directly as a submanifold of the given ambient space. In the parametric approach, one had to generalize the space of smooth maps to a Sobolev space, in order to guarantee the existence of limits of minimizing sequences. For the same reason, in the measure theoretic approach, the space of submanifolds has to be generalized to the one of currents. A submanifold of dimension k yields a linear functional on the space of differential forms of degree k by integration, and so the space of k-currents is defined as a space dual to the one of k-forms. One may then minimize a generalized version of area, the so-called mass, on the space of currents. This approach is valid in any dimension and codimension, in contrast to the parametric one that is restricted to 2 dimensions. If the codimension is 1 and the dimension at most 7, then such a mass minimizing current is regular in the sense that it represents a smooth submanifold. Otherwise, singularities may occur. In particular, any smooth Jordan curve in $\mathbb{R}^3$ bounds an embedded minimal surface, see Hardt and Simon[118]. For a general treatment of the concepts and the approach of geometric measure theory, we recommend Federer[80] and Almgren[6].

Minimal surfaces in Riemannian manifolds have found important geometric applications. Let us mention a few selected ones.

In the proof of the Bonnet-Myers Theorem (Corollary 4.3.1), we have seen how information about geodesics and their stability can be used to reach topological consequences for manifolds of positive Ricci curvature. This suggests that information about the stability of minimal surfaces may likewise be used to obtain restrictions on the topology of positively curved manifolds. The first instance of an important application of minimal 2-spheres in the presence of positive curvature is Siu and Yau[236]. Micallef and Moore[187] showed that minimal 2-spheres can be used to prove that any compact Riemannian manifold with positive curvature operator (i.e. $R(\cdot,\cdot) = \Omega^2(M) \to \Omega^2(M)$ is a positive operator; this in particular implies positive sectional curvature) is diffeomorphic to a sphere. Also, the sphere theorem (see Short survey on curvature and topology, above) was proved under the assumption of pointwise pinching only (i.e. at each point, the maximal ratio between sectional curvatures is less than 4).

There are also important applications of minimal surfaces in three-dimensional topology. The so-called Dehn Lemma, whose first complete proof was given by Papkyriakopoulos, asserts that if S is a differentiably embedded surface in a compact differentiable three-manifold M and if γ is an embedded curve on S that is homotopically trivial in M (i.e. $[\gamma] = 0 \in \pi_1(M)$) then γ bounds an embedded disk. Meeks and Yau[186] showed that in this case, if we equip M with a Riemannian metric in such a way that S is convex, the solution of the parametric Plateau problem with boundary γ is embedded. Thus, one obtains an embedded **minimal** disk bounded by γ. This represents an analytical proof of Dehn's Lemma. The important fact is that we have found a canonical solution of the problem. Assume for example that some compact group G acts on M, leaving γ invariant. One may then average the metric of M under the action of G and obtain a new Riemannian metric on M for which G acts by isometries. Since γ is G-invariant, one may then also find a G-invariant minimal disk bounded by γ. If one chooses this disk to be area minimizing in its class, one may then show again that it is embedded. This equivariant version of Dehn's Lemma of Meeks-Yau then has applications to the classification of discrete group actions on 3-manifolds, see [14].

8.3 Regularity Results

Regularity results are usually local in the domain (but the distinctive feature of geometric analysis in contrast to standard PDE theory is that regularity is a global question in the target). Thus, we consider regularity questions for harmonic maps from Riemann surfaces on the unit disk D. Since we shall see that the regularity question for harmonic maps on Riemann surfaces can essentially be reduced to the consideration of isolated singularities, we shall also use the punctured unit disk

$$D^* := D\backslash\{0\}$$

.

Lemma 8.3.1. *Suppose $f \in H^{1,2}(D^*, \mathbb{R}^n)$ satisfies*

$$\int_{D^*} Df(z)D\varphi(z)\,dz = \int_{D^*} g(z, f(z), Df(z))\varphi(z)\,dz \tag{8.3.1}$$

for all

$$\varphi \in H_0^{1,2} \cap L^\infty(D^*, \mathbb{R}^n)$$

where g *fulfills*

$$|g(z, f, p)| \le c_0 + c_1 |p|^2 \tag{8.3.2}$$

with constants c_0, c_1 *for all* $(z, f, p) \in D^* \times \mathbb{R}^n \times \mathbb{R}^{2m}$. *Then also*

$$\int_D Df(z) D\sigma(z)\, dz = \int_D g(z, f(z), Df(z))\sigma(z)\, dz \tag{8.3.3}$$

for all $\sigma \in H_0^{1,2} \cap L^\infty(D, \mathbb{R}^n)$.

The lemma says that weak solutions of (8.3.1) with finite Dirichlet integral extend as weak solutions through isolated singularities. Easy examples show that the assumption of finite Dirichlet integral is essential.

Proof. For $k \in \mathbb{N}$, $k \ge 2$, we put

$$\lambda_k(r) := \begin{cases} 1 & \text{for } r \le (\frac{1}{k})^2, \\ \log(\frac{1}{kr})/\log k & \text{for } (\frac{1}{k})^2 \le r \le \frac{1}{k}, \\ 0 & \text{for } r \ge \frac{1}{k}, \end{cases}$$

and for $\sigma \in H_0^{1,2} \cap L^\infty(D, \mathbb{R}^n)$,

$$\varphi_k(z) := (1 - \lambda_k(|z|))\sigma(z) \in H_0^{1,2} \cap L^\infty(D^*, \mathbb{R}^n).$$

We now observe that

$$\int_D |D\lambda_k(|z|)|^2\, dz = 2\pi \int_{(\frac{1}{k})^2}^{\frac{1}{k}} \Big(\frac{d\lambda_k}{dr}\Big)^2 r\, dr = \frac{2\pi}{\log k} \to 0 \quad \text{as } k \to \infty. \tag{8.3.4}$$

By (8.3.1),

$$\int_{D^*} Df(z) D\varphi_k(z)\, dz = \int_{D^*} g(z, f(z), Df(z))\varphi_k(z)\, dz. \tag{8.3.5}$$

Because of $f \in H^{1,2}$ and (8.3.2),

$$g(z, f(z), Df(z)) \in L^1.$$

Since $|\varphi_k| \le |\sigma| \in L^\infty$ and since φ_k converges to σ almost everywhere, Lebesgue's theorem on dominated convergence therefore implies that for $k \to \infty$, the right hand of (8.3.5) tends to

$$\int_D g(z, f(z), Df(z))\sigma(z)\, dz.$$

By (8.3.4), $\sigma \in L^\infty$, $f \in H^{1,2}$, and by Hölder's inequality, for $k \to \infty$,

$$\int_D Df(z) D(\lambda_k(z))\sigma(z)\, dz \to 0.$$

Therefore, the left hand side of (8.3.5) tends to

$$\int_D Df(z)D\sigma(z)\,dz$$

for $k \to \infty$, and (8.3.3) follows. □

Corollary 8.3.1. *Suppose that Σ is a Riemann surface, $p \in \Sigma$, N a Riemannian manifold, $f \in H^{1,2}(\Sigma \backslash \{p\}, N)$.*

If f is weakly harmonic on $\Sigma\backslash\{p\}$, then f extends as a weakly harmonic map to Σ.

Proof. A consequence of Lemmas 7.1.3, 8.3.1. □

Remark. Suppose that $f : \Sigma\backslash\{p\} \to N$ is localizable and of finite energy

$$E(f, \Sigma\backslash\{p\}) = \int_{\Sigma\backslash\{p\}} \|df\|^2 < \infty.$$

Then we can define the energy of f on Σ as

$$E(f; \Sigma) = E(f; \Sigma\backslash\{p\}).$$

The proof of Lemma 8.3.1 shows that this is meaningful.

Our first aim is to prove the extension result needed in the proofs of Theorems 8.2.4 and 8.2.5, namely that a conformal harmonic map $\mathbb{C} \to N$ of finite energy extends to a conformal harmonic map on $S^2 = \mathbb{C} \cup \{\infty\}$. While the following results are correct even without the assumption of conformality, that assumption considerably simplifies the proofs. We divide the proof into two steps, first continuity and then smoothness. In order to explain the basic idea of the continuity proof, we first consider an easy special case, namely $N = \mathbb{R}^n$. We are thus investigating weak minimal surfaces in Euclidean space:

Definition 8.3.1. A map $h \in H^{1,2}(\Sigma, \mathbb{R}^n)$ from a Riemann surface Σ is called a weak minimal surface if h is weakly harmonic and conformal, i.e.

(i)
$$\int_\Sigma (h_x\varphi_x + h_y\varphi_y)\,dxdy = 0 \tag{8.3.6}$$

for all $\varphi \in H_0^{1,2} \cap L^\infty(\Sigma, \mathbb{R}^n)$ ($z = x + iy$ being a conformal parameter on Σ), and

(ii)

$$\begin{aligned} h_x \cdot h_x &= h_y \cdot h_y, \\ h_x \cdot h_y &= 0, \end{aligned} \tag{8.3.7}$$

almost everywhere.

We now show

Proposition 8.3.1. *Any weak minimal surface $h \in H^{1,2}_{\mathrm{loc}}(\Sigma, \mathbb{R}^n)$ is continuous.*

Proof. Since the result is local, we may assume $\Sigma = D$, that the point where h has finite Dirichlet integral (energy) on D.

We consider $r \in (0,1)$ and

$$\begin{aligned} z_0 \in D_r &:= \{z \in \mathbb{C} : |z| < r\}, \\ p &:= h(z_0). \end{aligned}$$

We assume that for almost all $z \in \partial D_r = \{|z| = r\}$,

$$|h(z) - p| > \overline{\rho} \tag{8.3.8}$$

(this means that the minimal surface $h(D_r)$ has no boundary inside the ball $B(p, \overline{\rho})$).

The plan is to show that if $r \to 0$ then also $\overline{\rho} \to 0$ for $\overline{\rho}$ satisfying (8.3.8). We shall then apply the Courant-Lebesgue lemma to the extent that for suitable r, if $|h(z) - p|$ is small for one $z \in \partial D_r$ then this is so for all $z \in \partial D_r$. Continuity will then follow from the triangle inequality.

We first consider a general compact Riemann surface S with boundary ∂S and a weak minimal surface $h \in H^{1,2}(S, \mathbb{R}^n)$ with

$$|h(z) - p| > \overline{\rho} \quad \text{for all } z \in \partial S. \tag{8.3.9}$$

We let $\eta \in C^\infty(\mathbb{R})$ satisfy

$$\begin{aligned} \eta(t) &\equiv 1 \quad \text{for} \quad t \leq \frac{1}{2}, \\ \eta(t) &\equiv 0 \quad \text{for} \quad t \geq 1, \\ \eta'(t) &\leq 0 \quad \text{for all } t \end{aligned}$$

and choose as a test vector

$$\varphi(z) := \eta\Big(\frac{|h(z) - p|}{\rho}\Big)(h(z) - p)$$

for $0 < \rho \leq \overline{\rho}$.

Because of (8.3.9), φ has compact support in the interior of S. Therefore, φ is an admissible test vector in (8.3.6), and thus

$$\int_S (h_x \varphi_x + h_y \varphi_y)\, dx dy = 0 \quad (z = x + iy). \tag{8.3.10}$$

We now define

$$A_\eta(\rho) := \frac{1}{2}\int_S |Dh|^2 \eta\Big(\frac{|h-p|}{\rho}\Big).$$

If η is the characteristic function $\chi_{(-\infty,1)}$ of $(-\infty,1)$, $A_\eta(\rho)$ is the area of the minimal surface $h(S)$ inside the ball $B(p,\rho)$. We compute

$$A'_\eta(\rho) = -\frac{1}{2\rho^2}\int_S |Dh|^2 |h-p| \eta'\Big(\frac{|h-p|}{\rho}\Big) \tag{8.3.11}$$

and

$$\begin{aligned} &h_x\varphi_x + h_y\varphi_y = \\ &\eta\Big(\frac{|h-p|}{\rho}\Big)|Dh|^2 + \eta'\Big(\frac{|h-p|}{\rho}\Big)\frac{1}{\rho|h-p|}\{((h-p)\cdot h_x)^2 + ((h-p)\cdot h_y)^2\}\,. \end{aligned} \tag{8.3.12}$$

Since the vectors h_x and h_y are orthogonal and of equal length by the weak conformality of h, we estimate

$$\begin{aligned} ((h-p)\cdot h_x)^2 + ((h-p)\cdot h_y)^2 &\le \frac{1}{2}(h_x^2 + h_y^2)|h-p|^2 \\ &= \frac{1}{2}|Dh|^2|h-p|^2\,. \end{aligned} \tag{8.3.13}$$

The factor $\frac{1}{2}$ will be essential, cf. (8.3.14) below and its consequences. Since $\eta' \le 0$, (8.3.12) and (8.3.13) imply

$$h_x\varphi_x + h_y\varphi_y \ge \eta\Big(\frac{|h-p|}{\rho}\Big)|Dh|^2 + \eta'\Big(\frac{|h-p|}{\rho}\Big)\frac{|h-p|}{2\rho}|Dh|^2\,.$$

(8.3.10) and (8.3.11) then yield

$$2A_\eta(\rho) - \rho A'_\eta(\rho) \le 0,$$

hence

$$\Big(\frac{A_\eta(\rho)}{\rho^2}\Big)' \ge 0, \tag{8.3.14}$$

and thus for $0 < \rho_1 \le \rho_2 \le \overline{\rho}$,

$$\frac{A_\eta(\rho_1)}{2\pi\rho_1^2} \le \frac{A_\eta(\rho_2)}{2\pi\rho_2^2}\,. \tag{8.3.15}$$

□

We choose a sequence $(\eta_n)_{n\in\mathbb{N}}$ of smooth functions with the above properties and tending to $\chi_{(-\infty,1)}$. By Lebesgue's theorem on dominated convergence, we obtain in the limit with

$$A(\rho) := \text{Area}\,(h(S)\cap B(p,\rho))$$

the fundamental monotonicity formula for minimal surfaces which we record as

Theorem 8.3.1. *Let S be a compact Riemann surface with boundary ∂S and let $h \in H^{1,2}(S, \mathbb{R}^n)$ be a weak minimal surface, and suppose*

$$h(\partial S) \cap B(p, \overline{\rho}) = \emptyset. \tag{8.3.16}$$

Then $\frac{A(\rho)}{2\pi\rho^2}$ is a nondecreasing function of ρ for $0 < \rho \leq \overline{\rho}$.

The result also holds for $0 < \rho < \infty$ if S is a (noncompact) Riemann surface and $h \in H^{1,2}_{\mathrm{loc}}(S, \mathbb{R}^n)$ is a proper weak minimal surface. Here, "proper" means that the preimage of each compact set in $\mathbb{R}^n$ is compact in S.

Proof. The compact case has just been described. The claim for noncompact S follows by exhausting S by compact subsets. The properness of h guarantees that (8.3.16) is satisfied for sufficiently large compact subsets. □

We want to determine whether $\frac{A(\rho)}{2\pi\rho^2}$ has a limit as $\rho \to 0$.

Definition 8.3.2. Let T be a surface in a Riemannian manifold N, $p \in N$, $A(T, p, \rho) :=$ Area $(T \cap B(p, \rho))$. If

$$\lim_{\rho \to 0} \frac{A(T, p, \rho)}{2\pi\rho^2} =: d(T, p)$$

exists, then this limit is called the *density* of T at p.

We observe that if T is closed and $p \notin T$, then

$$d(T, p) = 0.$$

If h is a smooth minimal surface, then as a consequence of the Hartman-Wintner-Lemma 8.1.7, we have an asymptotic expansion

$$h_z(z_0) = a(z - z_0)^m$$

with some $a \in \mathbb{C}^n$ ($a^2 = 0$ since h defines a minimal surface) at every z_0 with some non-negative integer m, cf. Corollary 8.1.6, and

$$m = 0$$

for almost all z_0, because h_z has only isolated zeroes.

This easily implies

$$d(h(S), h(z_0)) = m + 1$$

and

$$d(h(S), h(z_0)) = 1 \quad \text{for almost all } z_0.$$

We now return to the case of a weak minimal surface $h : S \to \mathbb{R}^n$.

Lemma 8.3.2. *Let $h : S \to \mathbb{R}^n$ be a weak minimal surface. Then the (lower) density of $h(S)$ at $h(z)$ is at least 1 whenever*

$$z \in S_0 := \{y \in S : \textit{his approximately differentiable at } y, \\ y \textit{ is a Lebesgue point for } |Dh|^2, \textit{and } |Dh(y)|^2 \neq 0\}.$$

Consequently, for $z \in S_0$,

$$\text{Area}\,(h(S) \cap B(h(z), \varrho)) \geq 2\pi\varrho^2$$

whenever

$$h(\partial S) \cap B(h(z), \varrho) = \emptyset.$$

Proof. By the monotonicity formula (Theorem 8.3.1), we need to show that with $K_\varrho := \{x \in S : |h(x) - h(z)| \leq \varrho\}$,

$$\lim_{\varrho \to 0} \frac{1}{2\pi\varrho^2} \int\limits_{K_\varrho} |dh(x)|^2\, dx \geq 1.$$

Now, with $K_\varrho^\varepsilon := \{x \in D_\varrho : |h(x) - h(z) - \nabla h(z)(x - z)| \leq \varepsilon |x - z|\}$

$$\int\limits_{D_\varrho} |dh(x)|^2 \geq \int\limits_{K_\varrho^\varepsilon \cap S_0} |dh(x)|^2 = \int\limits_{K_\varrho^\varepsilon \cap S_0} |\nabla h(x)|^2,$$

where ∇h denotes the approximate derivative (see §A.1), and we shall control the latter quantity from below.

The domain of integration here is controlled by a radius in the image. In order to estimate the integral, however, we shall need to convert that radius into a radius in the domain.

We put

$$r_\varepsilon := \varrho \Big(\frac{1}{\sqrt{2}} |\nabla h(z)| + \varepsilon \Big)^{-1}.$$

Then, for

$$x \in B^\varepsilon(z, r_\varepsilon) := \{y \in B(z, r_\varepsilon) : |h(x) - h(z) - \nabla h(z)(x - z)| \leq \varepsilon |x - z|\},$$

$$|h(x) - h(z)| \leq |\nabla h(z)(x - z)| + \varepsilon |x - z|.$$

The conformality relations (8.3.7)) now imply

$$|\nabla h(z)(x - z)|^2 \leq \frac{1}{2} |\nabla h(z)|^2 |x - z|^2.$$

Thus, we obtain

$$|h(x) - h(z)| \leq \Big(\frac{1}{\sqrt{2}} |\nabla h(z)| + \varepsilon \Big) |x - z| \leq \varrho$$

for $x \in B^\varepsilon(z, r_\varepsilon)$. This implies

$$B^\varepsilon(z, r_\varepsilon) \subset K^\varepsilon_\varrho,$$

and so, since $K^\varepsilon_\varrho \setminus (K^\varepsilon_\varrho \cap S_0)$ is a null set,

$$\frac{1}{2\pi\varrho^2} \int\limits_{K^\varepsilon_\varrho \cap S_0} |\nabla h(z)|^2 \geq \frac{2\pi r_\varepsilon^2}{2\pi\varrho^2} |\nabla h(z)|^2 \,,$$

up to an error term (arising from having $B^\varepsilon(z, r_\varepsilon)$ in place of $B(z, r_\varepsilon)$) which, however, goes to 0 as ϱ, and hence also r_ε tends to 0, because h is approximately differentiable at z.

Inserting the value of r_ε, and letting first ϱ and then ε tend to 0, we obtain

$$\lim_{\varrho \to 0} \frac{1}{2\pi\varrho^2} \int\limits_{K\varrho} |\nabla h(z)|^2 \geq 1.$$

The integrand, here, however, is $|\nabla h(z)|^2$, i.e. the value at the center z, and not $|\nabla h(x)|^2$. Thus, in order to complete the proof, we need to estimate

$$\frac{1}{2\pi\varrho^2} \int\limits_{K^\varepsilon_\varrho \cap S_0} \left| |\nabla h(z)|^2 - |\nabla h(x)|^2 \right| \, dx.$$

Again, we need to translate the radius in the image into one in the domain, but this time with an inequality in the opposite direction.

W.l.o.g. $\varepsilon < |\nabla h(z)|$, and so for $x \in K^\varepsilon_\varrho \cap S_0$,

$$|x - z| \leq \varrho(|\nabla h(z)| - \varepsilon)^{-1} =: R_\varepsilon,$$

i.e.

$$K^\varepsilon_\varrho \cap S_0 \subset B(z, R_\varepsilon).$$

Therefore,

$$\begin{aligned} &\frac{1}{2\pi\varrho^2} \int\limits_{K^\varepsilon_\varrho \cap S_0} \left| |\nabla h(z)|^2 - |\nabla h(x)|^2 \right| dx \\ &\quad \leq \frac{1}{(|\nabla h(z)| - \varepsilon)^2} \frac{1}{2\pi R_\varepsilon^2} \int\limits_{B(z, R_\varepsilon) \cap S_0} \left| |\nabla h(z)|^2 - |\nabla h(x)|^2 \right| dx. \end{aligned}$$

If we then let ϱ, and hence R_ε tend to 0, the last integral also goes to 0 because z is a Lebesgue point for $|dh(z)|^2$. Thus, the proof is complete. $\square$

In order to also include points where h is not approximately differentiable, or that are not Lebesgue points for $|dh(z)|^2$, we now claim that the lower density

$$\liminf_{\rho \to 0} \frac{A(h(S), h(z), \rho)}{2\pi\rho^2}$$

is an upper semicontinuous function of z.

Let $\rho_n \to 0$ for $n \to \infty$.

By the above, we find sequences $(z_n)_{n\in\mathbb{N}} \subset S$, $(\varepsilon_n)_{n\in\mathbb{N}} \subset \mathbb{R}$, $\varepsilon_n \to 0$ as $n \to \infty$,

$$|h(z) - h(z_n)| = \varepsilon_n \rho_n.$$

Then, since $B(h(z_n), (1-\varepsilon_n)\rho_n) \subset B(h(z), \rho_n)$,

$$\begin{aligned}
\frac{A(h(S), h(z), \rho_n)}{2\pi\rho_n^2} &\geq \frac{A(h(S), h(z_n), (1-\varepsilon_n)\rho_n)}{2\pi\rho_n^2} \\
&= \frac{A(h(S), h(z_n), (1-\varepsilon_n)\rho_n)}{2\pi((1-\varepsilon_n)\rho_n)^2}(1-\varepsilon_n)^2 \\
&\geq d(h(S), h(z_n))(1-\varepsilon_n)^2 \quad \text{by monotonicity at } h(z_n),
\end{aligned}$$

and upper semicontinuity follows. □

We now return to the

Proof of Proposition 8.3.1. Put $S = D_r$. The preceding argument, Lemma 8.3.2 and Theorem 8.3.1 say

$$1 \leq \frac{A(\rho)}{2\pi\rho^2} \tag{8.3.17}$$

for $0 \leq \rho \leq \overline{\rho}$, unless $\nabla h \equiv 0$ locally, which, however, represents a trivial case.

Since

$$A(\overline{\rho}) \leq \frac{1}{2}\int_{D_r} |Dh|^2, \tag{8.3.18}$$

and

$$\lim_{r\to 0}\int_{D_r} |Dh|^2 = 0 \quad \text{monotonically}$$

(it follows by applying Lebesgue's theorem on dominated convergence to $f\chi_{D_r}$ that $\lim_{r\to 0}\int_{D_r} f = 0$ for any integrable f), we conclude from (8.3.17) that

$$\overline{\rho} \to 0 \quad \text{as } r \to 0.$$

This means, by definition of $\overline{\rho}$,

$$\inf_{z\in\partial D_r} |h(z) - h(z_0)| \to 0. \tag{8.3.19}$$

On the other hand, the Courant-Lebesgue-Lemma 8.2.5 says that for any $r_0 < 1$, there exists r with $r_0 < r < \sqrt{r_0}$ such that for all $z, z' \in \partial D_r$,

$$|h(z) - h(z')| \leq \frac{2\pi^{\frac{1}{2}}}{(\log\frac{1}{r})^{\frac{1}{2}}}\left(\int_{D_{r_0}} |Dh|^2\right)^{\frac{1}{2}}, \tag{8.3.20}$$

and the right hand side goes to 0 when $r_0 \to 0$, hence $r \to 0$.

Let now $\varepsilon > 0$ be given. We then find sufficiently small $r > 0$ so that first the right hand side of (8.3.20) is smaller than $\frac{\varepsilon}{3}$ and that for every $z_0 \in D_r$, the infimum in (8.3.19) is also smaller than $\frac{\varepsilon}{3}$. For $z_0, z_0' \in D_r$, let then z and z' resp., be points in ∂D_r where the infimum in (8.3.19) is attained. The triangle inequality gives $|h(z_0) - h(z_0')| < \varepsilon$, hence continuity. □

We now want to prove continuity of weak minimal surfaces in Riemannian manifolds.

Definition 8.3.3. A map $h \in H^{1,2}_{\text{loc}}(\Sigma, N)$ from a Riemann surface Σ into a Riemannian manifold N is called a *weak minimal surface* if it is weakly harmonic and conformal, i.e.

(i)
$$\int_\Sigma \langle dh, d\varphi \rangle = 0, \tag{8.3.21}$$

for all compactly supported bounded $H^{1,2}$ sections φ of $h^{-1}TN$ ($\langle \cdot, \cdot \rangle$ here is the scalar product in $T^*\Sigma \otimes h^{-1}TN$),

(ii)
$$\begin{aligned} \langle h_x, h_x \rangle &= \langle h_y, h_y \rangle, \\ \langle h_x, h_y \rangle &= 0 \end{aligned} \tag{8.3.22}$$

almost everywhere ($\langle \cdot, \cdot \rangle$ here is the scalar product in $h^{-1}TN$).

For (i), cf. Definition 7.1.3 and Lemma 7.1.2.

In contrast to the existence theory, for regularity results we do not need the compactness of the ambient manifold N. It suffices to have a uniform control on the geometry of N:

Definition 8.3.4. We say that a Riemannian manifold N is of bounded geometry if

(i)
$$i(N) := \inf_{p \in N} i(p) > 0,$$

where i denotes the injectivity radius,

(ii)
$$\Lambda := \sup_N |K| < \infty,$$

where K denotes the sectional curvature.

Theorem 8.3.2. *A weak minimal surface $H \in H^{1,2}_{\text{loc}}(\Sigma, N)$ (Σ a Riemann surface) in a Riemannian manifold N of bounded geometry is continuous.*

Proof. We shall translate the argument of the above Proposition from the Euclidean case into a Riemannian context. Thus, the strategy of proof will be the same as before.
Again, it suffices to treat the case $\Sigma = D$, $h \in H^{1,2}(D, N)$, and to prove continuity at 0.

We let

$$\begin{aligned} &0 < \rho_0 < \frac{1}{2}\min\Big(\frac{\pi}{2\sqrt{\Lambda}}, i(N)\Big), \\ &0 < r < 1, \\ &z_0 \in D_r = \{|z| < r\}, \quad p := h(z_0). \end{aligned}$$

We assume that for almost all $z \in \partial D_r = \{|z| = r\}$,

$$d(h(z), p) > \overline{\rho} \tag{8.3.23}$$

with

$$0 < \overline{\rho} \le \rho_0,$$

where $d(\cdot, \cdot)$ denotes the distance function of the metric of N. As before, we let $\eta \in C^\infty(\mathbb{R})$ satisfy

$$\begin{aligned} \eta(t) &\equiv 1 \quad \text{for} \quad t \le \frac{1}{2}, \\ \eta(t) &\equiv 0 \quad \text{for} \quad t \ge 1, \\ \eta'(t) &\le 0 \quad \text{for all } t, \end{aligned}$$

and again, we later on let η increase to the characteristic function $\chi_{(-\infty,1)}$.

We now choose as test vector

$$\varphi(z) := \eta\Big(\frac{d(h(z), p)}{\rho}\Big)(-\exp_{h(z)}^{-1} p) \in T_{h(z)}N.$$

φ is bounded, of class $H^{1,2}$, namely

$$\int \langle d\varphi, d\varphi \rangle \le \text{const} \int \langle dh, dh \rangle < \infty,$$

for example by (8.3.26) below, or directly from the chain rule, and by (8.3.23), it has compact support in D_r. Therefore, φ is an admissible test vector, and by (8.3.21)

$$\int_\Sigma \langle dh, d\varphi \rangle = 0. \tag{8.3.24}$$

In order to evaluate (8.3.24), we compute

$$\begin{aligned} \langle d\varphi, dh \rangle &= \langle \nabla_{\frac{\partial}{\partial x}} \varphi dx + \nabla_{\frac{\partial}{\partial y}} \varphi dy, h_x dx + h_y dy \rangle \\ &= \eta\Big(\frac{d(h,p)}{\rho}\Big)\big(\langle \nabla_{\frac{\partial}{\partial x}}(-\exp_h^{-1} p), h_x \rangle + \langle \nabla_{\frac{\partial}{\partial y}}(-\exp_h^{-1} p), h_y \rangle\big) \\ &\quad + \eta'\Big(\frac{d(h,p)}{\rho}\Big)\frac{1}{\rho d(h,p)}\big(\langle(-\exp_h^{-1} p), h_x\rangle^2 + \langle(-\exp_h^{-1} p), h_y\rangle^2\big) \end{aligned} \tag{8.3.25}$$

(cf. (4.6.6)).

We have to estimate the covariant derivatives of $(-\exp_h^{-1} p)$.

For this purpose, let $h(s)$ be a smooth curve in N. In order to control $\nabla_{\frac{\partial}{\partial s}} \exp_{h(s)}^{-1} p$, we consider the family of geodesics

$$c(t,s) := \exp_{h(s)}(t \exp_{h(s)}^{-1} p).$$

Then

$$\frac{\partial}{\partial t} c(t,s)_{|t=0} = \exp_{h(s)}^{-1} p$$

and thus

$$\begin{aligned}\nabla_{\frac{\partial}{\partial s}} \exp_{h(s)}^{-1} p &= \nabla_{\frac{\partial}{\partial s}} \frac{\partial}{\partial t} c(t,s)_{|t=0} \\ &= \nabla_{\frac{\partial}{\partial t}} \frac{\partial}{\partial s} c(t,s)_{|t=0} \,.\end{aligned}$$

For fixed s, $J_s(t) := \frac{\partial}{\partial s} c(t,s)$ is a Jacobi field along the geodesic $c(\cdot, s)$ with

$$\begin{aligned} &J_s(0) = h'(s) && := \frac{\partial h}{\partial s}, \\ &J_s(1) = 0 \in T_p N, && \\ &\dot{J}_s(0) = \nabla_{\frac{\partial}{\partial s}} \exp_{h(s)}^{-1} p && := \frac{\partial}{\partial t} J_s(0). \end{aligned}$$

From Corollary 4.5.1, we have the Jacobi field estimate

$$\|J_s(0) + \dot{J}_s(0)\| \le \frac{1}{2} \Lambda d^2(h(s),p) \|J_s(0)\|,$$

hence

$$\|\nabla_{\frac{\partial}{\partial s}} \exp_{h(s)}^{-1} p + h'(s)\| \le \frac{1}{2} \Lambda d^2(h(s),p) \|h'(s)\|. \tag{8.3.26}$$

We shall use (8.3.26) to compare $\nabla_{\frac{\partial}{\partial x}} \exp_h^{-1} p$ with h_x.

The conformality relations

$$\langle h_x, h_x \rangle = \langle h_y, h_y \rangle, \langle h_x, h_y \rangle = 0 \quad \text{almost everywhere},$$

imply

$$\begin{aligned} \langle \exp_h^{-1} p, h_x \rangle^2 + \langle \exp_h^{-1} p, h_y \rangle^2 &\le \frac{1}{2} (\|h_x\|^2 + \|h_y\|^2) \| \exp_h^{-1} p\|^2 \\ &= \frac{1}{2} \|dh\|^2 \cdot d^2(h,p). \end{aligned} \tag{8.3.27}$$

The factor $\frac{1}{2}$ will be crucial.

We define

$$A_\eta(\rho) := \frac{1}{2} \int_{D_r} \|dh\|^2 \eta\Big(\frac{d(h,p)}{\rho}\Big) \,.$$

Then, because of (8.3.23) and $\rho \le \overline{\rho}$,

$$A'_\eta(\rho) = -\frac{1}{2\rho^2}\int_{D_r} \|dh\|^2 d(h,p)\eta'\Big(\frac{d(h,p)}{\rho}\Big).$$

From (8.3.25), we get, since $\eta' \le 0$, $\eta \ge 0$,

$$\begin{aligned}
2\langle d\varphi, dh\rangle \ge{}& \eta\Big(\frac{d(h,p)}{\rho}\Big)\big(\langle \nabla_{\frac{\partial}{\partial x}}(-\exp_p^{-1} h), h_x\rangle + \langle \nabla_{\frac{\partial}{\partial y}}(-\exp_p^{-1} h), h_y\rangle\big)\\
&+ \eta'\Big(\frac{d(h,p)}{\rho}\Big)\frac{d(h,p)}{2\rho}\|dh\|^2 \qquad \text{by (8.3.27)}\\
\ge{}& \eta\Big(\frac{d(h,p)}{\rho}\Big)\|dh\|^2 + \eta'\Big(\frac{d(h,p)}{\rho}\Big)\frac{d(h,p)}{2\rho}\|dh\|^2\\
&- \frac{\Lambda}{2}\eta\Big(\frac{d(h,p)}{\rho}\Big)d^2(h,p)\|dh\|^2 \qquad \text{by (8.3.26)}
\end{aligned}$$

and then from (8.3.24),

$$2A_\eta(\rho) - \rho A'_\eta(\rho) \le \Lambda\rho^2 A_\eta(\rho). \tag{8.3.28}$$

This implies

$$\Big(\frac{A_\eta(\rho)}{\rho^2}e^{\frac{\Lambda}{2}\rho^2}\Big)' \ge 0,$$

hence

$$\frac{A_\eta(\rho_1)}{2\pi\rho_1^2}e^{\frac{\Lambda}{2}\rho_1^2} \le \frac{A_\eta(\rho_2)}{2\pi\rho_2^2}e^{\frac{\Lambda}{2}\rho_2^2} \tag{8.3.29}$$

whenever

$$0 < \rho_1 \le \rho_2 \le \overline{\rho}.$$

We again let η approach the characteristic function $\chi_{(-\infty,1)}$ and obtain with

$$A(\rho) := \mathrm{Area}\,(h(D_r)\cap B(p,\rho))$$

the following *monotonicity formula*

$$\frac{A(\rho_1)}{2\pi\rho_1^2}e^{\frac{\Lambda}{2}\rho_1^2} \le \frac{A(\rho_2)}{2\pi\rho_2^2}e^{\frac{\Lambda}{2}\rho_2^2} \tag{8.3.30}$$

whenever $0 < \rho_1 \le \rho_2 \le \overline{\rho}$.

Again, if $\rho_1 \to 0$, the left hand side of (8.3.30) tends to the density of the minimal surface $h(D_r)$ at $p = h(z_0)$, and this density again is a positive integer.

Therefore, choosing $\rho_2 = \overline{\rho}$ in (8.3.30),

$$\begin{aligned}
\overline{\rho}^2 &\le \frac{1}{2\pi}e^{\frac{\Lambda}{2}\overline{\rho}^2}\int_{D_r}\|dh\|^2\\
&\le \frac{1}{2\pi}e^{\frac{\Lambda}{2}\rho_0^2}\int_{D_r}\|dh\|^2 \qquad \text{since } \overline{\rho} \le \rho_0.
\end{aligned} \tag{8.3.31}$$

This is impossible, if $r \leq r_0$ and r_0 is chosen so small that

$$\int_{D_{r_0}} \|dh\|^2 \leq 2\pi e^{-\frac{\Lambda}{2}\rho_0^2}\overline{\rho}^2. \tag{8.3.32}$$

Therefore, for such r, (8.3.23) cannot hold. Thus, for $0 < r \leq r_0$,

$$\operatorname*{ess\,inf}_{z \in \partial D_r} d(h(z), h(z_0)) \leq \overline{\rho}. \tag{8.3.33}$$

Also, by the intermediate value theorem, we can find r with $\frac{1}{2}r_0 \leq r \leq r_0$ and

$$d(h(z), h(z')) \leq \frac{2\pi}{(\log 2)^{\frac{1}{2}}} \left(\int_{D_{r_0}} \|dh\|^2 \right)^{\frac{1}{2}} \tag{8.3.34}$$

for all $z, z' \in \partial D_r$ (this is an alternative to the use of the Courant-Lebesgue lemma 8.2.5, the proof is similar).

We then choose r_0 so small that in addition to (8.3.32)

$$\int_{D_{r_0}} \|dh\|^2 < \frac{\log 2}{4\pi^2} \overline{\rho}^2 . \tag{8.3.35}$$

For $z_0, z_0' \in D_r$, $\frac{1}{2}r_0 \leq r \leq r_0$, r satisfying (8.3.24), we find $z, z' \in \partial D_r$ for which the infimum is attained in (8.3.33) for z_0 and z_0', resp. Then from (8.3.33) and (8.3.34) and the triangle inequality

$$d(h(z_0), h(z_0')) \leq 3\overline{\rho}.$$

Since this holds for all $z_0, z_0' \in D_r$, where r is estimated in terms of $\overline{\rho}$, continuity at 0 follows. □

Perspectives. In Theorem 8.3.2, we have shown that weakly harmonic and conformal maps of finite energy from a Riemann surface into a Riemannian manifold (of bounded geometry) are continuous. The conformality of the map is not needed for this regularity result as was shown by Hélein[121]. A systematic treatment is given in [122]. The removability of isolated singularities of weakly harmonic maps was already obtained by Sacks and Uhlenbeck[219]. The proof of the continuity of weak minimal surface given here partly uses some arguments of Grüter[112].

Exercises for Chapter 8

1. Show that every two-dimensional torus carries the structure of a Riemann surface.

2. Determine all holomorphic quadratic differentials on a two-dimensional torus, and all holomorphic quadratic differentials on an annular region $\{z \in \mathbb{C} : r_1 \leq |z| \leq r_2\}$ $(0 < r_1 < r_2)$ that are real on the boundary.

3. Show that the conclusions of the Hartman-Wintner-Lemma 8.1.7 continue to hold if (8.1.17) is replaced by

$$|u_{z\bar{z}}| \leq K(|u_z| + |u|).$$

4. We let Σ be a Riemann surface and $H : \mathbb{R}^3 \to \mathbb{R}$ be a smooth function. For a map $f : \Sigma \to \mathbb{R}^3$ we consider the equation

$$\Big(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\Big) f = 2H(f(z)) f_x \wedge f_y$$

where $z = \kappa + iy$ is a conformal parameter on Σ and $\wedge$ denotes the standard vector product in $\mathbb{R}^3$.

a: Show that, if f is conformal, $H(f(z))$ is the mean curvature of the surface $f(\Sigma)$ at the point $f(z)$.

b: If $\Sigma = S^2$, show that every solution is conformal.

c: If Σ is the unit disk D and f is a solution which is constant on ∂D, show that it is constant on all of D.

d: Show that for a nonconstant solution, f_x and f_y have only isolated zeroes.

e: At those points where f_x and f_y do not vanish, we define

$$\begin{aligned} L &:= \frac{\langle f_{xx}, f_x \wedge f_y \rangle}{|f_x \wedge f_y|}, \\ M &:= \frac{\langle f_{xy}, f_x \wedge f_y \rangle}{|f_x \wedge f_y|}, \\ N &:= \frac{\langle f_{yy}, f_x \wedge f_y \rangle}{|f_x \wedge f_y|}. \end{aligned}$$

(using the Euclidean metric of $\mathbb{R}^3$).

Show that for a solution with $H \equiv$ const. $\varphi dz^2 := (L - N - 2iM)dz^2$ is a holomorphic quadratic differential.

Conclude that φ, since holomorphic and bounded, extends to all of Σ as a holomorphic quadratic differential.

f: If $H \equiv$ const and $\Sigma = S^2$, show that every solution $f(\Sigma)$ has constant and equal principal curvatures at each point. Conclude that it is a standard sphere of radius $\frac{1}{\sqrt{H}}$ i.e. $f(\Sigma) = \{x \in \mathbb{R}^3 : |x - x_0|^2 = \frac{1}{H}\}$ for some x_0.

(Hint: Use a), b), e) and Lemma 8.1.4.)

Remark: By the uniformization theorem, every two dimensional Riemannian manifold M diffeomorphic to S^2 admits the structure of a Riemann surface and a conformal diffeomorphism $K : S^2 \to M$. It thus is conformally equivalent to S^2. The exercise then implies that every surface diffeomorphic to S^2 and immersed into $\mathbb{R}^3$ with constant mean curvature is a standard "round" sphere. This result, as well as the method of proof presented here, were discovered by H. Hopf.

5. Prove Theorem 8.2.3, assuming only that N is complete but not necessarily compact.

Chapter 9

Variational Problems from Quantum Field Theory

9.1 The Ginzburg-Landau Functional

A prototypical situation for the functionals that we are going to consider is the following:

M is a compact Riemannian manifold, E a complex vector bundle over M, i.e. a vector bundle with fiber $\mathbb{C}^n$, equipped with a Hermitian metric $\langle\cdot,\cdot\rangle$. We consider sections φ of E and unitary connections $D_A = d + A$ (locally) on E. Here, "unitary" of course means that A is skew Hermitian w.r.t. $\langle\cdot,\cdot\rangle$. We denote the curvature of $D_A = d + A$ by F_A, and we write $|\varphi|$ for $\langle\varphi,\varphi\rangle^{\frac{1}{2}}$.

We consider Lagrangians of the type

$$\mathcal{L}(\varphi, A) := \int\limits_M \big(\gamma_1|F_A|^2 + \gamma_2|D_A\varphi|^2 + \gamma_3 V(\varphi)\big) *(1)\,. \tag{9.1.1}$$

Here γ_1, γ_2, γ_3 are positive constants, while $V(\cdot)$ is some "potential". If $V(\varphi)$ is quadratic in $|\varphi|$, e.g.

$$V(\varphi) = m^2|\varphi|^2, \tag{9.1.2}$$

the resulting Euler-Lagrange equations are linear in φ,

$$D_A^* D_A\varphi + m^2\varphi = 0. \tag{9.1.3}$$

The Euler-Lagrange equations also contain a equation for variations of A, namely

$$\gamma_1 D_A^* F_A = -\frac{1}{2}\gamma_2(\langle\varphi, D_A\varphi\rangle + \langle D_A\varphi, \varphi\rangle) \tag{9.1.4}$$

(see also the proof of Lemma 9.1.1 below for the derivation of these equations).

It leads to a richer structure, however, if we allow $V(\varphi)$ to be a polynomial of higher than quadratic order in $|\varphi|$. Of particular interest to us will be the case of a fourth order polynomial, for example

$$V(\varphi) = \left(\sigma - |\varphi|^2\right)^2 ,$$

for some $\sigma \in \mathbb{R}$.

We first consider the case where the base manifold is a compact Riemann surface Σ equipped with a conformal metric, and where the vector bundle is a Hermitian line bundle L, i.e. with fiber $\mathbb{C}$, and a Hermitian metric $\langle \cdot, \cdot \rangle$ on the fibers.

Definition 9.1.1. The *Ginzburg-Landau functional* for a section φ of L and a unitary connection $D_A = D + A$ on L is defined as

$$\mathcal{L}(\varphi, A) := \int_\Sigma \Big(|F_A|^2 + |D_A\varphi|^2 + \frac{1}{4}\big(\sigma - |\varphi|^2\big)^2 \Big) {*}(1) \,, \tag{9.1.5}$$

for $\sigma \in \mathbb{R}$.

The reason for the factor $\frac{1}{4}$ will emerge in a moment. A simple calculation yields

Lemma 9.1.1. *The Euler-Lagrange equations for the Ginzburg-Landau functional are*

$$D_A^* D_A \varphi = \frac{1}{2}\left(\sigma - |\varphi|^2\right)\varphi, \tag{9.1.6}$$

$$D_A^* F_A = -\mathrm{Re}\,\langle D_A\varphi, \varphi\rangle. \tag{9.1.7}$$

Proof. The term $\int |F_A|^2$ was handled already in §3.2 when we derived the Yang-Mills equation. Varying

$$\int \langle D_A\varphi, D_A\varphi\rangle \tag{9.1.8}$$

w.r.t. A yields

$$\frac{d}{dt}\int \langle D_{A+tB}\varphi, D_{A+tB}\varphi\rangle|_{t=0} = \int \left(\langle D_A\varphi, B\varphi\rangle + \langle B\varphi, D_A\varphi\rangle\right) .$$

Thus (9.1.7) readily follows (cf. also (9.1.4) above). Varying (9.1.8) w.r.t. φ yields

$$\frac{d}{dt}\int \langle D_A(\varphi + t\psi), D_A(\varphi + t\psi)\rangle|_{t=0} = \int \left(\langle D_A^* D_A\varphi, \psi\rangle + \langle \psi, D_A^* D_A\varphi\rangle\right) .$$

Finally, the right hand side of (9.1.6) obviously arises from varying

$$\int \frac{1}{4}\left(\sigma - |\varphi|^2\right)^2$$

w.r.t. φ. □

Remark. (9.1.7) is linear in A. Namely, as explained in §3.2 (cf. (3.2.24)), for an abelian structure group, $D_A^* F_A$ becomes $d^* F_A$, and so (9.1.7) is

$$d^* \left(\partial A^{0,1} - \overline{\partial} A^{1,0}\right) = -\mathrm{Re}\,\langle (d+A)\varphi, \varphi\rangle$$

(in the notations of (9.1.12) below) which is obviously linear in A (but not in φ).

Since D_A is a unitary connection, A is a 1-form with values in $\mathfrak{u}(1)$, the Lie algebra of U(1). This Lie algebra will sometimes be identified with $i\mathbb{R}$. (U(1) is a subgroup of the Lie group $\mathrm{Gl}(1,\mathbb{C})$, and $\mathfrak{u}(1)$ is a subalgebra of the Lie algebra $\mathfrak{gl}(1,\mathbb{C})$. The latter can be identified with $\mathbb{C}$. Likewise, $\mathrm{Gl}(1,\mathbb{C})$ can be identified with $\mathbb{C}^*$, the nonvanishing complex numbers, and U(1) then corresponds to to the complex numbers of the form $\mathrm{e}^{i\vartheta}$, $\vartheta \in \mathbb{R}$. Taking derivatives, $\mathfrak{u}(1)$ then corresponds to the complex numbers of the form it, $t \in \mathbb{R}$.) Thus, A, $A^{1,0}$ $A^{0,1}$, and the curvature F_A will then be considered as imaginary valued forms. This will explain certain factors i appearing in the sequel.

We should point out that the convention adopted here (which is a consequence of more general conventions used in other places in the present book) is different from the convention employed in the physics literature, where one writes a unitary connection as

$$d - iA$$

with a real valued A. In other words, our A corresponds to $-iA$ in the physics literature.

We decompose Ω^1, the space of 1-forms on Σ, as

$$\Omega^1 = \Omega^{1,0} \oplus \Omega^{0,1}, \tag{9.1.9}$$

with $\Omega^{1,0}$ spanned by 1-forms of the type dz, $\Omega^{0,1}$ by 1-forms of the type $d\overline{z}$. Here, z of course is a local conformal parameter on Σ, and with $z = x+iy$, we have $\overline{z} = x-iy$. From the beginning of §8.1, we recall the conventions

$$dz = dx + i dy, \qquad d\overline{z} = dx - i dy,$$

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right), \qquad \frac{\partial}{\partial \overline{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right).$$

If $\frac{\partial}{\partial x}$, $\frac{\partial}{\partial y}$ are an orthonormal basis of the tangent space of Σ at the point under consideration, we get

$$\begin{aligned}
\langle dz, dz\rangle &= \langle dx + i dy, dx + i dy\rangle \\
&= \langle dx, dx\rangle + i\langle dy, dx\rangle - i\langle dx, dy\rangle + \langle dy, dy\rangle \\
&= 2, \\
\langle d\overline{z}, d\overline{z}\rangle &= 2, \\
\langle dz, d\overline{z}\rangle &= 0.
\end{aligned} \tag{9.1.10}$$

The last relation in (9.1.10) implies that (9.1.9) is an orthogonal decomposition. We may also decompose D_A into its $(1,0)$ and $(0,1)$ parts

$$D_A = \partial_A + \overline{\partial}_A.$$

Thus

$$\partial_A \varphi \in \Omega^{1,0}(L), \quad \overline{\partial}_A \varphi \in \Omega^{0,1}(L), \quad \text{for all sections } \varphi \text{ of } L. \tag{9.1.11}$$

We also write

$$\partial_A = \partial + A^{1,0}, \quad \overline{\partial}_A = \overline{\partial} + A^{0,1}, \tag{9.1.12}$$

with

$$d = \partial + \overline{\partial}$$

being the decomposition of the exterior derivative. Here we have

$$\partial f = \frac{\partial f}{\partial z} dz, \quad \overline{\partial} f = \frac{\partial f}{\partial \overline{z}} d\overline{z}, \quad \text{for functions on } \Sigma.$$

We write the conformal metric g on Σ in our local coordinates as

$$\rho^2(z)\, dz d\overline{z}.$$

Given $z_0 \in \Sigma$, we may assume that

$$\rho^2(z_0) = 1, \tag{9.1.13}$$

simply by replacing our coordinates z by $\frac{1}{\rho(z_0)} z$. We may then describe the action of the $*$ operator of the metric $\rho^2 dz d\overline{z}$ at z_0 as follows

$$\begin{aligned} *dz &= *(dx + idy) \\ &= dy - idx \\ &= -idz, \end{aligned} \tag{9.1.14}$$

$$*d\overline{z} = \quad id\overline{z}. \tag{9.1.15}$$

We also recall

$$dz \wedge d\overline{z} = -2i dx \wedge dy, \tag{9.1.16}$$

hence

$$\begin{aligned} *(dz \wedge d\overline{z}) &= -2i*(dx \wedge dy) \\ &= -2i \end{aligned} \tag{9.1.17}$$

and

$$\begin{aligned} *(1) &= dx \wedge dy \\ &= \frac{i}{2} dz \wedge d\overline{z}. \end{aligned} \tag{9.1.18}$$

We compute

$$\begin{aligned}\partial_A\partial_A\varphi &= \left(\partial + A^{1,0}\right)\circ\left(\partial + A^{1,0}\right)\varphi\\ &= \partial\partial\varphi + A^{1,0}\wedge\partial\varphi + A^{1,0}\wedge A^{1,0}\varphi + \left(\partial A^{1,0}\right)\varphi - A^{1,0}\wedge\partial\varphi\\ &= 0,\end{aligned}\tag{9.1.19}$$

since $\partial\partial = 0$ and $A^{1,0}\wedge A^{1,0} + \partial A^{1,0}$ is a $(2,0)$-form which has to vanish as Σ has complex dimension 1.

Likewise

$$\overline{\partial}\,\overline{\partial} = 0.\tag{9.1.20}$$

Moreover,

$$\begin{aligned}\partial_A\overline{\partial}_A\varphi &= \partial\overline{\partial}\varphi + A^{1,0}\wedge\overline{\partial}\varphi + A^{1,0}\wedge A^{0,1}\varphi + \left(\partial A^{0,1}\right)\varphi - A^{0,1}\wedge\partial\varphi,\\ \overline{\partial}_A\partial_A\varphi &= \overline{\partial}\partial\varphi + A^{0,1}\wedge\partial\varphi + A^{0,1}\wedge A^{1,0}\varphi + \left(\overline{\partial}A^{1,0}\right)\varphi - A^{1,0}\wedge\overline{\partial}\varphi\\ &= -\partial_A\overline{\partial}_A\varphi + \left(\overline{\partial}A^{1,0} - \partial A^{0,1}\right)\varphi\\ &= -\partial_A\overline{\partial}_A\varphi - F_A\varphi,\end{aligned}\tag{9.1.21}$$

i.e.

$$F_A = -\left(\partial_A\overline{\partial}_A + \overline{\partial}_A\partial_A\right).\tag{9.1.22}$$

Theorem 9.1.1. *We have*

$$\mathcal{L}(\varphi, A) = \int_\Sigma \left(2|\overline{\partial}_A\varphi|^2 + \left(*(-iF) - \frac{1}{2}\left(\sigma - |\varphi|^2\right)\right)^2\right)*(1) + 2\pi\sigma\deg L \tag{9.1.23}$$

with

$$\deg L := c_1(L)[\Sigma] \qquad \textit{(the degree of the line bundle } L\textit{).}$$

Proof. We compute (writing F in place of F_A)

$$\int\left(*(-iF) - \frac{1}{2}\left(\sigma - |\varphi|^2\right)\right)^2 *(1) = \int\left(|F|^2 + \frac{1}{4}\left(\sigma - |\varphi|^2\right)^2 - \sigma * iF * iF\langle\varphi,\varphi\rangle\right)*(1).\tag{9.1.24}$$

Now

$$\int *iF*(1) = \int iF = 2\pi c_1(L)[\Sigma] = 2\pi\deg L.\tag{9.1.25}$$

Also, using (9.1.22),

$$\int\langle *iF\varphi,\varphi\rangle *(1) = \int\left\langle -i\left(\partial_A\overline{\partial}_A + \overline{\partial}_A\partial_A\right)\varphi, *\varphi\right\rangle *(1).$$

In order to proceed, let $z_0\in\Sigma$, and choose Riemannian normal coordinates with center z_0. Thus, $\rho^2(z_0) = 1$, and the first derivatives of the metric vanish at z_0. Also, we apply a gauge transformation so that $A(z_0) = 0$ (see Lemma 3.2.3). Since we are not going to commute any derivatives any more, no second derivatives of the metric or first derivatives of A will enter our subsequent computations at z_0, and we may therefore proceed with our computations as in the Euclidean case. Thus, we have to evaluate

$$\begin{aligned}
&\int \Big\langle -i\big((\varphi_{\bar{z}})_z dz \wedge d\bar{z} + (\varphi_z)_{\bar{z}} d\bar{z} \wedge dz\big), \frac{i}{2}\varphi dz \wedge d\bar{z} \Big\rangle *(1) \\
&\quad = -\int 2\big((\varphi_{\bar{z}})_z \cdot \overline{\varphi} - (\varphi_z)_{\bar{z}} \cdot \overline{\varphi}\big) *(1) \\
&\text{(since } \langle -i dz \wedge d\bar{z}, i dz \wedge d\bar{z}\rangle = -|dz \wedge d\bar{z}|^2 = -4, \text{as } \langle \cdot, \cdot \rangle \text{ is Hermitian)} \\
&\quad = 2\int (\varphi_{\bar{z}}\overline{\varphi}_z - \varphi_z \overline{\varphi}_{\bar{z}}) *(1) \\
&\quad = -\int (|\partial_A \varphi|^2 - |\overline{\partial}_A \varphi|^2) *(1)
\end{aligned}$$

(the factor 2 disappears since $\langle dz, dz\rangle = \langle d\bar{z}, d\bar{z}\rangle = 2$, and in our coordinates $\partial\varphi = \varphi_z dz$ etc.). Thus we have shown

$$-\int \langle *iF\varphi, \varphi\rangle *(1) = \int \big(|\partial_A \varphi|^2 - |\overline{\partial}_A \varphi|^2\big) *(1). \tag{9.1.26}$$

Finally, of course

$$|D_A \varphi|^2 = |\partial_A \varphi|^2 + |\overline{\partial}_A \varphi|^2, \tag{9.1.27}$$

since the decomposition

$$\Omega^1 = \Omega^{1,0} \oplus \Omega^{0,1}$$

is orthogonal. The result then follows from (9.1.24) – (9.1.27). □

Theorem 9.1.1 has the following useful consequence

Corollary 9.1.1. *Assume* $\deg L \geq 0$. *Then the lowest possible value permitted by the global topology of the bundle for* $\mathcal{L}(\varphi, A)$ *is realised precisely if* φ *and* A *satisfy the set of first order differential equations*

$$\overline{\partial}_A \varphi = 0, \tag{9.1.28}$$

$$*(iF) = \frac{1}{2}(\sigma - |\varphi|^2). \tag{9.1.29}$$

□

Remark. If $\deg L < 0$, then these equations cannot have any solution, because for any solution, $\mathcal{L}(\varphi, A)$ would be negative by (9.1.23) whereas we see from (9.1.5) that for any φ, A, $\mathcal{L}(\varphi, A) \geq 0$. Thus, in case $\deg L < 0$, one has to consider the selfduality equations arising from the following expression for the Ginzburg-Landau functional:

$$\mathcal{L}(\varphi, A) = \int_\Sigma \Big(2|\partial_A \varphi|^2 + \big(*(-iF) - \frac{1}{2}(\sigma - |\varphi|^2)\big)^2\Big) *(1) - 2\pi \deg L, \tag{9.1.30}$$

which is derived through the same computations. W.l.o.g., we shall assume $\deg L \geq 0$ in the sequel.

Integrating (9.1.29) yields the inequality

$$2\pi \deg L = \int iF = \frac{1}{2}\int (\sigma - |\varphi|^2)*(1) \leq \frac{\sigma}{2}\mathrm{Area}\,(\Sigma),$$

with

$$\mathrm{Area}\,(\Sigma) = \int_{\Sigma} *(1).$$

Thus, a necessary condition for the solvability of (9.1.29) is

$$\sigma \geq \frac{4\pi \deg L}{\mathrm{Area}\,(\Sigma)}, \tag{9.1.31}$$

and in fact, we must have strict inequality in (9.1.31) unless $\varphi \equiv 0$.

Corollary 9.1.1 constitutes another instance of the phenomenon of selfduality that we already encountered in §3.2 when we discussed the Yang-Mills functional on a fourdimensional Riemannian manifold. The equations (9.1.28), (9.1.29) are also called selfduality equations because the solutions of these first order equations are precisely those solutions of (9.1.6), (9.1.7) that realize the lower bound imposed by the topology for the functional and, if they exist, yield the absolute minima for the functional considered. In fact, this remark, namely that these equations hold for the absolute minima, makes it clear that any solution of (9.1.28), (9.1.29) automatically also solves (9.1.6), (9.1.7), as the latter are the Euler-Lagrange equations for the Ginzburg-Landau functional, and as such have to be satisfied in particular by minimizers of that functional. Of course, it may also be checked by a direct computation that solutions of (9.1.28), (9.1.29) also solve (9.1.6), (9.1.7).

The selfduality may be generalized as follows. Instead of $\mathcal{L}(\varphi, A)$, we consider for $\epsilon > 0$,

$$\begin{aligned} \mathcal{L}_\epsilon(\varphi, A) :&= \int \Big\{\epsilon^2 |F_A|^2 + |D_A\varphi|^2 + \frac{1}{4\epsilon^2}(\sigma - |\varphi|^2)^2\Big\}*(1) \\ &= \int \Big\{2|\overline{\partial}_A\varphi|^2 + \big(\epsilon*(iF) - \frac{1}{2\epsilon}(\sigma - |\varphi|^2)\big)^2\Big\}*(1) + 2\pi \deg L, \end{aligned} \tag{9.1.32}$$

which leads to the selfduality equations

$$\overline{\partial}_A\varphi = 0, \tag{9.1.33}$$

$$\epsilon^2 *(iF) = \frac{1}{2}(\sigma - |\varphi|^2). \tag{9.1.34}$$

Still more generally, in place of ϵ, one may consider a function $f(z)$ on Σ, for example,

$\frac{\epsilon}{|\varphi(z)|}$. This leads to the functional

$$\begin{aligned}\mathcal{L}_{\frac{\epsilon}{|\varphi(z)|}}(\varphi, A) &= \int \left\{ \frac{\epsilon^2}{|\varphi(z)|^2}|F_A(z)|^2 + |D_A\varphi(z)|^2 + \frac{1}{4\epsilon^2}(\sigma - |\varphi|^2)^2|\varphi(z)|^2 \right\} *(1) \\ &= \int \left\{ 2|\overline{\partial}_A\varphi|^2 + \Big(\frac{\epsilon}{|\varphi(z)|}*(iF) - \frac{1}{2\epsilon}(\sigma - |\varphi|^2)|\varphi| \Big)^2 \right\} *(1) + 2\pi \deg L\end{aligned} \tag{9.1.35}$$

with the selfduality equations

$$\overline{\partial}_A\varphi = 0, \tag{9.1.36}$$

$$\epsilon^2*(iF) = \frac{1}{2}(\sigma - |\varphi|^2)|\varphi|^2 \tag{9.1.37}$$

The functionals $\mathcal{L}_\epsilon$ and $\mathcal{L}_{\frac{\epsilon}{|\varphi(z)|}}$ are quite important for studying phase transitions in superconductivity.

For studying solutions, the following consequence of the maximum principle is very useful

Lemma 9.1.2. *Let Σ be a compact Riemann surface with a conformal metric, L as before. For any solution of (9.1.6), hence in particular for any solution of (9.1.28), we have*

$$|\varphi| \le \sigma \quad \text{on } \Sigma. \tag{9.1.38}$$

Proof. From (9.1.6), we obtain

$$\begin{aligned}\frac{1}{2}\Delta\langle\varphi, \varphi\rangle &= \langle D^{*_A}D_A\varphi, \varphi\rangle - \langle D_A\varphi, D_A\varphi\rangle \qquad \text{(cf. (3.2.7))} \\ &= \frac{1}{2}(\sigma - |\varphi|^2)|\varphi|^2 - |D_A\varphi|^2.\end{aligned}$$

Let $z_0 \in \Sigma$ be a point where $|\varphi|^2$ achieves its maximum. We may assume $A = 0$ at z_0 (cf. Lemma 3.2.3), hence $D_A\varphi = 0$ at z_0. If we had $|\varphi(z_0)| > \sigma$, then at z_0

$$\Delta|\varphi|^2 < 0,$$

which contradicts the maximum principle. □

Perspectives. It was shown by Taubes[245] that on $\mathbb{R}^2$, one may solve the Ginzburg-Landau equations with any given finite collection prescribed as zero set for φ, with prescribed multiplicities. This result was extended to compact Riemann surfaces by Bradlow and Garcí a-Prada, and these authors also found generalizations on higher dimensional Kähler manifolds. References include [29, 30], [91, 92, 93]. We should also mention Hitchin's penetrating study[127] of the equations

$$\begin{aligned}\overline{\partial}_A\varphi &= 0, \\ F_A + [\varphi, \varphi^*] &= 0\end{aligned}$$

on a compact Riemann surface.

The limit analysis for $\epsilon \to 0$ of the functional $\mathcal{L}_\epsilon(\varphi, A)$ and the solutions of the equations (9.1.33), (9.1.34) on a compact Riemann surface has been carried out by Hong, Jost, Struwe[128]. The result is that away from the prescribed zero set of φ_ϵ (the "vortices"), $|\varphi_\epsilon|$ uniformly converges to 1, and $D_{A_\epsilon}\varphi_\epsilon$ and dA_ϵ uniformly converge to 0, whereas the curvature in the limit becomes a sum of delta distributions concentrated at the vortices. Of course, the number of vortices counted with multiplicity has to equal the degree of the line bundle L, $\deg L$. This result thus yields a method for degenerating a line bundle on a Riemann surface into a flat line bundle with $\deg L$ singular points (counted with multiplicity) and a covariantly constant section.

Results for the φ^6 theory on a compact torus can be found in Caffarelli, Yang[36], Tarantello[244], Ding, Jost, Li, Wang[63]. For the case of S^2, see Ding, Jost, Li, Wang[64]. The general case was solved by Ding, Jost, Li, Peng, Wang[62].

9.2 The Seiberg-Witten Functional

Let M be a compact, oriented, four dimensional Riemannian manifold with a spin^c structure $\widetilde{P}^c$, i.e. a spin^c manifold. (As mentioned in §1.11, in the four dimensional case, there always exists some spin^c structure on a given oriented Riemannian manifold.) As in Definition 1.11.10, the determinant line bundle of this spin^c structure will be denoted by L, and as in Definition 3.4.1 (ii), the Dirac operator determined by a unitary connection A on L will be denoted by $\partial\!\!\!/_A$. Finally, we recall the half spin bundle $\mathcal{S}^\pm$ defined by the spin^c structure, as remarked after Definition 1.11.10 (we omit the subscript for the dimension, as the dimension is fixed to be 4 in the present section). By Lemma 3.4.5, $\partial\!\!\!/_A$ maps sections of $\mathcal{S}^\pm$ to sections of $\mathcal{S}^\mp$.

Definition 9.2.1. The *Seiberg-Witten functional* for a unitary connection A on L and a section φ of $\mathcal{S}^+$ is

$$SW(\varphi, A) := \int_M \Big(|\nabla_A \varphi|^2 + |F_A^+|^2 + \frac{R}{4}|\varphi|^2 + \frac{1}{8}|\varphi|^4\Big) *(1), \tag{9.2.1}$$

where ∇_A is the spin^c connection induced by A and the Levi-Civita connection of M (cf. (3.4.6)), F_A^+ is the selfdual part of the curvature of A, and R is the scalar curvature of M.

The discussion of the Seiberg-Witten functional will parallel our discussion of the Ginzburg-Landau functional in §9.1. In fact, the structure of SW is quite similar to the one of $\mathcal{L}$, containing a square norm of the curvature of the connection A, the square of the norm of the covariant derivation of φ, and a nonlinearity that is a fourth order polynomial in $|\varphi|$.

Lemma 9.2.1. *The Euler-Lagrange equations for the Seiberg-Witten functional are*

$$\nabla_A^* \nabla_A \varphi = -\left(\frac{R}{4} + \frac{1}{4}|\varphi|^2\right)\varphi, \tag{9.2.2}$$

$$d^* F_A^+ = -\mathrm{Re}\,\langle \nabla_A \varphi, \varphi \rangle. \tag{9.2.3}$$

Proof. As the proof of Lemma 9.1.1. $\square$

In order to proceed, we need to associate to $s \in S_4^+$ the 2-form $\tau(s)$ defined by

$$\tau(s)(v,w) := \langle v \cdot w \cdot s, s\rangle + \langle v, w\rangle |s|^2.$$

Lemma 9.2.2.

$$\tau(s) \in \Lambda^{2,+}(i\mathbb{R})$$

(i.e. $\tau(s)$ is a selfdual 2-form that assumes imaginary values), and

$$|\tau(s)|^2 = 2|s|^4.$$

Proof. We first show that $\tau(s)$ takes imaginary values. We start with the skew symmetry.

$$\begin{aligned}
\tau(s)(v,w) &= \langle v \cdot w \cdot s, s\rangle + \langle v, w\rangle |s|^2 \\
&= \langle (-w \cdot v - 2\langle v, w\rangle) s, s\rangle + \langle v, w\rangle |s|^2 \\
&= -\tau(s)(w, v),
\end{aligned}$$

next,

$$\begin{aligned}
\overline{\tau(s)(v,w)} &= \overline{\langle v \cdot w \cdot s, s\rangle} + \langle v, w\rangle |s|^2 \\
&= \langle s, v \cdot w \cdot s\rangle + \langle v, w\rangle |s|^2 \\
&= -\langle v \cdot s, w \cdot s\rangle + \langle v, w\rangle |s|^2 \quad \text{by Corollary 1.11.4} \\
&= \langle w \cdot v \cdot s, s\rangle + \langle v, w\rangle |s|^2 \quad \text{for the same reason} \\
&= \quad \tau(s)(w, v) \\
&= -\tau(s)(v, w) \quad \text{by skew symmetry.}
\end{aligned}$$

This implies that $\tau(s)(v,w)$ is in $i\mathbb{R}$.

For the computation of $|\tau(s)|^2$, we recall that the spin representation $\Gamma : \mathrm{Cl}^c(\mathbb{R}^4) \to \mathbb{C}^{4\times 4}$, and the half spin representation that we shall now denote as Γ^+: $\mathrm{Cl}^{c,ev}(\mathbb{R}^4) \to$

$S_4^+ \cong \mathbb{C}^2$. We write $s = (s^1, s^2) \in \mathbb{C}^2$ and obtain from the formulae for $\Gamma(e_\alpha, e_\beta)$ from §1.11,

$$\begin{aligned}
\tau(s)(e_1, e_2) &= i(s^1\overline{s^2} + s^2\overline{s^1}) = \quad \tau(s)(e_3, e_4),\\
\tau(s)(e_1, e_3) &= \quad s^1\overline{s^2} - s^2\overline{s^1} \ = -\tau(s)(e_2, e_4),\\
\tau(s)(e_1, e_4) &= i(s^1\overline{s^1} - s^2\overline{s^2}) = \quad \tau(s)(e_2, e_3).
\end{aligned}$$

This already implies that $\tau \in \Lambda^{2,+}$.

We may now compute

$$\begin{aligned}
|\tau(s)|^2 &= \sum_{i<j} |\tau(s)(e_i, e_j)|^2\\
&= 2\Big((s^1\overline{s^1} - s^2\overline{s^2})^2 + (s^1\overline{s^2} + s^2\overline{s^1})^2 - (s^1\overline{s^2} - s^2\overline{s^1})^2\Big)\\
&= 2\,|s|^4.
\end{aligned}$$

□

In more explicit terms we may write

$$\tau(s) = \langle e_j \cdot e_k \cdot s, s\rangle e^j \wedge e^k$$

where e^j is a frame in T^*M dual to the frame e_j on TM $(j = 1, \ldots, 4)$.

Theorem 9.2.1. *The Seiberg-Witten functional (9.2.1) can be expressed as*

$$SW(\varphi, A) = \int_M \Big(\left|\partial\!\!\!/_A \varphi\right|^2 + \left|F_A^+ - \frac{1}{4}\langle e_j \cdot e_k \cdot \varphi, \varphi\rangle e^j \wedge e^k\right|^2 \Big) {*}(1), \tag{9.2.4}$$

where e^j, $j = 1, \ldots, 4$, are 1-forms dual to the tangent vectors e_j, $j = 1, \ldots, 4$, i.e. $e^j(e_k) = \delta_{jk}$.

Proof. We have

$$\begin{aligned}
&\left|F_A^+ - \frac{1}{4}\langle e_j \cdot e_k \cdot \varphi, \varphi\rangle e^j \wedge e^k\right|^2 =\\
&\qquad |F_A^+|^2 + \frac{1}{16}\left|\langle e_j \cdot e_k \cdot \varphi, \varphi\rangle e^j \wedge e^k\right|^2 - \frac{1}{2}\langle F_A^+, e^j \wedge e^k\rangle\langle e_j \cdot e_k \cdot \varphi, \varphi\rangle.
\end{aligned} \tag{9.2.5}$$

By Lemma 9.2.2

$$\frac{1}{16}\left|\langle e_j \cdot e_k \cdot \varphi, \varphi\rangle e^j \wedge e^k\right|^2 = \frac{1}{8}|\varphi|^4. \tag{9.2.6}$$

Writing $F_A^+ = F_{il}^+ e^i \wedge e^l$, we get

$$\begin{aligned}
-\frac{1}{2}\langle F_A^+, e^j \wedge e^k\rangle\langle e_j \cdot e_k \cdot \varphi, \varphi\rangle &= -\frac{1}{2}\langle F_{jk}^+ e_j \cdot e_k \cdot \varphi, \varphi\rangle\\
&= -\frac{1}{2}\langle F_A^+ \varphi, \varphi\rangle.
\end{aligned} \tag{9.2.7}$$

On the other hand, the Weizenböck formula of Theorem 3.4.2 yields, (applying (3.4.21) to φ, taking the scalar product with φ, integrating, and using the self adjointness of $\partial\!\!\!/_A$) that

$$\int |\partial\!\!\!/_A \varphi|^2 = \int |\nabla_A \varphi|^2 + \frac{1}{4} R |\varphi|^2 + \frac{1}{2} \langle F_A^+ \varphi, \varphi \rangle. \tag{9.2.8}$$

The result follows from (9.2.5) – (9.2.8). □

Corollary 9.2.1. *The lowest topologically possible value of the Seiberg-Witten functional is achieved precisely if φ and A are solutions of*

$$\partial\!\!\!/_A \varphi = 0, \tag{9.2.9}$$

$$F_A^+ = \frac{1}{4} \langle e_j \cdot e_k \cdot \varphi, \varphi \rangle e^j \wedge e^k. \tag{9.2.10}$$

Definition 9.2.2. The equations (9.2.9) and (9.2.10) are called the *Seiberg-Witten equations.*

Thus, we see the mechanism of selfduality at work once more. The absolute minima of the Seiberg-Witten functional for which the above lower bound is achieved satisfy not only the the second order equations (9.2.2), (9.2.3), but also the first order Seiberg-Witten equations (9.2.9), (9.2.10).

So far our discussion of the Seiberg-Witten functional has been completely analogous to the one of the Ginzburg-Landau functional, except that so far, the parameter σ in the latter has had no analogue in the former. However, this can easily be achieved by choosing a 2-form μ and considering the perturbed functional

$$\begin{aligned} SW_\mu(\varphi, A) &= \int \Big(|\partial\!\!\!/_A \varphi|^2 + \big| F_A^+ - \frac{1}{4} \langle e_j \cdot e_k \cdot \varphi, \varphi \rangle e^j \wedge e^k + \mu \big|^2 \Big) *(1) \\ &= \int \Big(|\nabla_A \varphi|^2 + |F_A^+|^2 + \frac{R}{4} |\varphi|^2 \\ &\qquad + \big| \mu - \frac{1}{4} \langle e_j \cdot e_k \cdot \varphi, \varphi \rangle e^j \wedge e^k \big|^2 + 2 \langle F_A^+, \mu \rangle \Big) *(1). \end{aligned} \tag{9.2.11}$$

If we assume that μ is antiselfdual, then

$$\langle F_A^+, \mu \rangle = 0, \tag{9.2.12}$$

as F_A^+ by definition is selfdual and the decomposition of the 2-forms on a four dimensional manifold into selfdual and antiselfdual ones is orthogonal (see §3.2). Thus, in that case the additional term $\langle F_A^+, \mu \rangle$ in (9.2.11) disappears.

If we assume that μ is a closed selfdual form, then

$$\langle F_A^-, \mu \rangle = 0,$$

again since the antiselfdual form F_A^- is orthogonal to the selfdual forms, and hence

$$\langle F_A^+, \mu \rangle = \langle F_A, \mu \rangle.$$

Further, since F_A represents the first Chern class $c_1(L)$ of the determinant line bundle L (see §3.2), and since μ is assumed to be closed, hence represents a cohomology class $[\mu]$,

$$\int_M \langle F_A, \mu\rangle *(1) \tag{9.2.13}$$

does not depend on the connection A (see the discussion of Chern classes in §3.2), hence represents a topological invariant, denoted by $(c_1(L)\wedge[\mu])[M]$. This expression then plays a role that is completely analogous that one of $2\pi \deg L$ in the discussion of the Ginzburg-Landau functional.

The corresponding first order equations for SW_μ are

$$\partial\!\!\!/_A \varphi = 0, \tag{9.2.14}$$

$$F_A^+ = \frac{1}{4}\langle e_j \cdot e_k \cdot \varphi, \varphi\rangle e^j \wedge e^k - \mu. \tag{9.2.15}$$

Since, by our conventions, both F^+ and $\langle e_j \cdot e_k \cdot \varphi, \varphi\rangle e^j \wedge e^k$ are imaginary valued, (9.2.15) may only admit a solution if we assume that μ is imaginary valued as well. As in the Ginzburg-Landau theory, one may also introduce a scaling factor $\epsilon > 0$ or a scaling function like $\frac{\epsilon}{|\varphi|}$ into the Seiberg-Witten functional. For example, one may define

$$\begin{aligned} SW_{\mu,\epsilon}(\varphi, A) &= \int_M \Big\{ |\nabla_A \varphi|^2 + \epsilon^2 |F_A^+|^2 + \frac{R}{4}|\varphi|^2 \\ &\qquad + \frac{1}{\epsilon^2}\big|\,\mu - \langle e_j \cdot e_k \cdot \varphi, \varphi\rangle e^j \wedge e^k\big|^2 + 2\langle F_A^+, \mu\rangle \Big\} *(1) \\ &= \int_M \Big\{ \big|\partial\!\!\!/_A \varphi\big|^2 + \Big|\epsilon F_A^+ - \frac{1}{\epsilon}\Big(\frac{1}{4}\langle e_j \cdot e_k \cdot \varphi, \varphi\rangle e^j \wedge e^k - \mu\Big)\Big|^2 \Big\} *(1). \end{aligned} \tag{9.2.16}$$

We have a maximum principle similar to Lemma 9.1.2:

Lemma 9.2.3. *For any solution φ of (9.2.2), hence in particular for any solution of (9.2.9), on a compact four-dimensional Riemannian manifold, we have*

$$\max_M |\varphi|^2 \le \max_{x\in M}(-R(x), 0). \tag{9.2.17}$$

Proof. (9.2.2) implies

$$\begin{aligned} \frac{1}{2}\Delta|\varphi|^2 &= \langle \nabla_A^* \nabla_A \varphi, \varphi\rangle - |\nabla_A \varphi|^2 \qquad \text{(cf. (3.2.7))} \\ &= -\Big(\frac{R}{4} + \frac{1}{4}|\varphi|^2\Big)|\varphi|^2 - |\nabla_A \varphi|^2. \end{aligned}$$

Let $x_0 \in M$ be a point where $|\nabla_A \varphi|^2$ achieves its maximum. Then

$$\Delta|\varphi(x_0)|^2 \geq 0.$$

Thus,

$$R(x_0) + |\varphi(x_0)|^2 \leq 0,$$

and (9.2.17) follows. □

Corollary 9.2.2. *If the compact, oriented, Riemannian Spin^c manifold M has non-negative scalar curvature, then the only possible solution of the Seiberg-Witten equations is*

$$\varphi \equiv 0,$$
$$F_A^+ \equiv 0.$$

Proof. By Corollary 9.2.1, solutions of the Seiberg-Witten equations (9.2.9), (9.2.10) also solve (9.2.2), (9.2.3). From Lemma 9.2.3 we conclude that in case $R \geq 0$, the only solution of (9.2.2) is

$$\varphi \equiv 0.$$

(9.2.10) then yields $F_A^+ \equiv 0$. □

In fact, the conclusion of Corollary 9.2.2 may also be obtained directly from Theorem 9.2.1 as follows: From (9.2.4) is clear that for any solution of (9.2.9), (9.2.10), we have $SW(\varphi, A) = 0$. If $R \geq 0$, (9.2.1) on the other hand implies that $SW(\varphi, A) = 0$ can only hold if all terms in the integral in (9.2.1) vanish. Hence $\varphi \equiv 0$, $F_A^+ \equiv 0$.

Perspectives. The Seiberg-Witten equations were introduced by Seiberg and Witten[232, 233]. The mathematical relevance of these equations was first shown by Witten[265], Taubes[246, 247], Kronheimer and Mrowka[173]. Further references can be found in the monographs of Salamon[221] and Morgan[196]. The equations and their applications are also described in several survey articles, among which we mention Friedrich[87] (see also [88]). All these references have been useful in assembling the material presented here.

As in the case of other gauge theories like the Yang-Mills theory discussed in §3.2, the functional and the equations are invariant under the action of a gauge group. Here the structure group is U(1), and so the Gauge group $\mathcal{G}$ consists of maps from M into $\mathrm{U}(1) \cong S^1$, $u \in \mathcal{G}$ acts on a pair (φ, A) via

$$u^*(\varphi, A) = (u^{-1}\varphi, u^{-1}du + A).$$

One has

$$\partial\!\!\!/_{u^*A}(u^{-1}\varphi) = u^{-1}\partial\!\!\!/_A\varphi$$

and

$$F_{u^*A} = F_A,$$

so that the functional and the equations (including the perturbed ones) remain invariant under the action of $\mathcal{G}$. For a given spin^c structure $\widetilde{P}^c$, Riemannian metric g and imaginary valued selfdual 2-form μ as pertubation, one considers the space of solutions of (9.2.14), (9.2.15) modulo the action of $\mathcal{G}$. This space is called moduli space $\mathcal{M}(M, \widetilde{P}^c, g, \mu)$ of solutions. One writes the second Betti number b_2 of M as

$$b_2 = b^+ + b^-,$$

where b^+ (b^-) is the dimension of the subspace of $H^2(M, \mathbb{R})$ represented by (anti)selfdual 2-forms. In Seiberg-Witten theory, it is shown that in case $b^+ > 0$, the moduli spaces $\mathcal{M}(M, \widetilde{P}^c, g, \mu)$ are finite dimensional, smooth, compact, oriented manifolds, at least for "generic" μ. The compactness here comes from the fact that solutions satisfy uniform estimates. (Lemma 9.2.3 and estimates for higher derivatives, see e.g. Jost, Peng, Wang[150] for a general presentation) that imply convergence of subsequences of families of solutions. This is different from the situation in Donaldson's theory of (anti)selfdual connections on SU(2) bundles where no uniform estimates hold. The most useful case seems to be where the moduli space is zerodimensional, i.e. where one has a finite number of solutions. The theorem of Seiberg-Witten says that if $b^+ > 1$ and $b^+ - b^-$ is odd, then the number of solutions counted with orientation is independent of the choice of the Riemannian metric g and the pertubation μ and depends only on the spin^c structure $\widetilde{P}^c$ on M. Also, these moduli spaces are nonempty only for finitely many spin^c structures. If (M, g) in addition has positive scalar curvature, then in fact all Seiberg-Witten invariants vanish (cf. Corollary 9.2.2). On the other hand, such Seiberg-Witten invariants, i.e. numbers of solutions counted with orientation, can often be computed from general index theorems, i.e. from topological data alone, and when these numbers are found to be nonzero, this yields an obstruction for certain compact, oriented, differentiable 4-manifolds to carry metrics with positive scalar curvature. For results based on such ideas, see e.g. Le Brun[33]. The Seiberg-Witten theory can be used to prove, to reprove and to extend many results from Donaldson theory. Kronheimer-Mrowka[173] and Morgan, Szabó, Taubes[197] used Seiberg-Witten theory to prove the Thom conjecture, stating that smooth algebraic curves (i.e. compact complex smooth subvarietes of complex dimension one) in $\mathbb{CP}^2$ minimize the genus in their homology classes.

The Seiberg-Witten equations seem to be particulary useful on symplectic 4-manifolds (M, ω). Using $i\omega$ as a perturbation and using the limit $\epsilon \to 0$ for the parameter ϵ introduced into the equations above (see (9.2.16)), Taubes[248, 249] showed that in the limit the zero set of the solution φ is a collection of pseudoholomorphic curves in the sense of Gromov[106]. Also, the curvature F_A will concentrate along the pseudoholomorphic curves in the limit $\epsilon \to 0$. In this way, one may identify the invariants defined by Gromov that are very useful in symplectic geometry, but hard to compute, with the invariants of Seiberg-Witten that can typically be computed from topological index theorems. For a generalization of the Seiberg-Witten functional with a potential term of sixth order, see Ding, Jost, Li, Peng, Wang[62].

9.3 Dirac-harmonic Maps

Let Σ be a compact oriented Riemann surface, equipped with a conformal Riemannian metric as in Definition 8.1.2, in local coordinates

$$\rho^2(z)\, dz \otimes d\bar{z} \tag{9.3.1}$$

for some positive, real valued function $\rho(z)$. In real coordinates, we write the metric as $\gamma_{\alpha\beta}(x)dx^\alpha dx^\beta$, and put $\gamma = \det(\gamma_{\alpha\beta})$, as usual. For a map $f : \Sigma \to N$ into some Riemannian manifold, we shall use the abbreviations

$$f^i_\alpha := \frac{\partial f^i}{\partial x^\alpha} \tag{9.3.2}$$

in local coordinates on N and real coordinates x^1, x^2 on Σ.

As defined in 1.11, we let $\mathcal{S}\Sigma$ be the spinor bundle of Σ, w.r.t. to some choice of spin structure, equipped with a Hermitian product $\langle \cdot, \cdot \rangle$. We also recall the Clifford multiplication

$$\begin{array}{rcl} T_x\Sigma \times_{\mathbb{C}} \mathcal{S}^{\pm}_x\Sigma & \to & \mathcal{S}^{\mp}_x\Sigma \\ v \otimes s & \mapsto & v \cdot s \end{array} \tag{9.3.3}$$

which satisfies the Clifford relations

$$v \cdot w \cdot s + w \cdot v \cdot s = -2\langle v, w\rangle s \tag{9.3.4}$$

for $v, w \in T_x\Sigma$ and $s \in \mathcal{S}_x\Sigma$, and which is skew-symmetric,

$$\langle v \cdot s, s' \rangle = -\langle \cdot s, v \cdot s' \rangle \tag{9.3.5}$$

for $v \in T_x\Sigma$ and $s, s' \in \mathcal{S}_x\Sigma$.

Let f be a smooth map from Σ to a Riemannian manifold (N, g) of dimension $n \geq 2$. $f^{-1}TN$ is the pull-back of the tangent bundle TN by f. We consider the twisted bundle $\mathcal{S}\Sigma \otimes f^{-1}TN$. On this bundle, there is a metric $\langle \cdot, \cdot \rangle$ induced from the metric on $\mathcal{S}\Sigma$ (induced in turn by the metric on Σ) and the metric of N on $f^{-1}TN$. Also, we have a natural connection $\widetilde{\nabla}$ on $\mathcal{S}\Sigma \otimes f^{-1}TN$ induced from those on $\mathcal{S}\Sigma$ and $f^{-1}TN$.

In local coordinates, a section ψ of $\mathcal{S}\Sigma \otimes f^{-1}TN$ can be expressed as

$$\psi(x) = \sum_{j=1}^{n} \psi^j(x) \otimes \frac{\partial}{\partial y^j}(f(x)), \tag{9.3.6}$$

where ψ^i is a spinor and $\frac{\partial}{\partial y^j}, j = 1, \ldots, n$, is the natural local basis of TN. The connection $\widetilde{\nabla}$ can then be expressed as

$$\widetilde{\nabla}\psi = \nabla\psi^i(x) \otimes \frac{\partial}{\partial y^j}(f(x)) + \Gamma^i_{jk}\nabla f^j(x)\psi^k(x) \otimes \frac{\partial}{\partial y^i}(f(x)), \tag{9.3.7}$$

where, of course, the Γ^i_{jk} are the Christoffel symbols of N. Since the connection $\widetilde{\nabla}$ is induced from the Levi-Civita connections of Σ and N, we have

$$v\langle\psi_1,\psi_2\rangle = \langle\widetilde{\nabla}_v\psi_1,\psi_2\rangle + \langle\psi_1,\widetilde{\nabla}_v\psi_2\rangle, \tag{9.3.8}$$

for any vector field v.

After these preparations, we can define the *Dirac operator along the map* f by

$$\not{D}\psi = \not{\partial}\psi^i(x)\otimes\frac{\partial}{\partial y^i}(f(x)) + \Gamma^i_{jk}\nabla_{e_\alpha}f^j(x)(e_\alpha\cdot\psi^k(x))\otimes\frac{\partial}{\partial y^i}(f(x)), \tag{9.3.9}$$

where e_1, e_2 is the local orthonormal basis of Σ and $\not{\partial} = e_\alpha\cdot\nabla_{e_\alpha}$ is the usual Dirac operator as defined in Definition 3.4.1.

Like $\not{\partial}$, see (3.4.11), also the Dirac operator $\not{D}$ is formally self-adjoint, i.e.,

$$\int_\Sigma\langle\psi,\not{D}\xi\rangle = \int_\Sigma\langle\not{D}\psi,\xi\rangle, \tag{9.3.10}$$

for all $\psi,\xi\in\Gamma(\mathcal{S}\Sigma\otimes f^{-1}TN)$, the space of smooth sections of $\mathcal{S}\Sigma\otimes f^{-1}TN$.

We consider the space

$$\mathcal{X} := \{(f,\psi)\,|\,f\in C^\infty(\Sigma,N) \text{ and } \psi\in\Gamma(\mathcal{S}\Sigma\otimes f^{-1}TN)\}$$

of smooth maps from Σ to N together with smooth sections of the bundle $\mathcal{S}\Sigma\otimes f^{-1}TN$ along those maps. On $\mathcal{X}$, we define the functional

$$\begin{aligned} L(f,\psi) &= \frac{1}{2}\int_\Sigma\left(\|df\|^2 + \langle\psi,\not{D}\psi\rangle\right)\rho^2 dz d\bar{z} \\ &= \int_\Sigma\left(g_{ij}(f)\gamma^{\alpha\beta}\frac{\partial f^i}{\partial x_\alpha}\frac{\partial f^j}{\partial x_\beta} + g_{ij}(f)\langle\psi^i,\not{D}\psi^j\rangle\right)\sqrt{\gamma}d^2x. \end{aligned} \tag{9.3.11}$$

So far, we do not make use of the assumption that the domain Σ is two-dimensional. Thus, the next result is, in fact, valid for Riemannian manifolds of arbitrary dimension as domains.

Theorem 9.3.1. *The Euler-Lagrange equations for L are*

$$\tau(f) = \mathcal{R}(f,\psi), \tag{9.3.12}$$

$$\not{D}\psi = 0, \tag{9.3.13}$$

where $\tau(f)$ is the tension field of the map f and $\mathcal{R}(f,\psi)\in\Gamma(f^{-1}TN)$ is defined by

$$\mathcal{R}(f,\psi)(x) = \frac{1}{2}R^m_{lij}(f(x))\langle\psi^i,\nabla\phi^l\cdot\psi^j\rangle\frac{\partial}{\partial y^m}(f(x)). \tag{9.3.14}$$

Here, the R^m_{lij} are the components of the curvature tensor of N.

Definition 9.3.1. Solutions (f,ψ) of (9.3.12) and (9.3.13) are called *Dirac-harmonic maps.*

Proof of Theorem 9.3.1: We first keep f fixed and vary ψ. We consider a family ψ_t with $d\psi_t/dt = \eta$ at $t = 0$. Since $\not{D}$ is formally self-adjoint (see (9.3.10)), we have for a critical point of L for all such η

$$0 = \frac{dL}{dt}|_{t=0} = \int_M \langle \eta, \not{D}\psi \rangle + \langle \psi, \not{D}\eta \rangle = 2\int_M \langle \eta, \not{D}\psi \rangle, \tag{9.3.15}$$

which yields (9.3.13) by Theorem A.1.5.

Next, we consider a variation $\{f_t\}$ of f with $df_t/dt = \xi$ at $t = 0$ for which the coefficients ψ^j $(j = 1, 2, \cdots, n)$ of the spinor $\psi(x) = \psi^j(x) \otimes \frac{\partial}{\partial y^j}(f(x))$ are independent of t. Then

$$\frac{dL(f_t)}{dt}|_{t=0} = \int_M \frac{\partial}{\partial t}\|df_t\|^2|_{t=0} + \int_M \frac{\partial}{\partial t}\langle \psi, \not{D}\psi \rangle|_{t=0}. \tag{9.3.16}$$

By (7.1.13), we have

$$\int_M \frac{\partial}{\partial t}\|df_t\|^2|_{t=0} = -2\int_M \tau^i(f) g_{im}\xi^m. \tag{9.3.17}$$

For the remaining term in (9.3.16), we first compute the variation of $\not{D}\psi$. As usual, we choose Riemann normal coordinates, that is, $\nabla_{\frac{\partial}{\partial x^\alpha}} \frac{\partial}{\partial x^\beta} = 0$ at the point under consideration. We also put $e_\alpha := \frac{\partial}{\partial x^\alpha}$. Then

$$\begin{aligned}
\frac{d}{dt}\not{D}\psi &= e_\alpha \cdot \nabla_{\frac{\partial}{\partial t}} \nabla_{e_\alpha}\psi \\
&= e_\alpha \cdot \nabla_{e_\alpha}\psi^i \otimes \nabla_{\frac{\partial}{\partial t}} \frac{\partial}{\partial y_i} + e_\alpha \cdot \psi^i \otimes \nabla_{\frac{\partial}{\partial t}} \nabla_{e_\alpha} \frac{\partial}{\partial y_i} \\
&= e_\alpha \cdot \nabla_{e_\alpha}\psi^i \otimes \nabla_{\frac{\partial}{\partial t}} \frac{\partial}{\partial y_i} + e_\alpha \cdot \psi^i \otimes (\nabla_{e_\alpha} \nabla_{\frac{\partial}{\partial t}} \frac{\partial}{\partial y_i} + R(df(\frac{\partial}{\partial t}), df(e_\alpha))\frac{\partial}{\partial y_i}) \\
&= e_\alpha \cdot \nabla_{e_\alpha}(\psi^i \otimes \nabla_{\frac{\partial}{\partial t}} \frac{\partial}{\partial y_i}) + e_\alpha \cdot \psi^i \otimes R(df(\frac{\partial}{\partial t}), df(e_\alpha))\frac{\partial}{\partial y_i}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\int_M \frac{\partial}{\partial t}\langle \psi, \not{D}\psi \rangle|_{t=0} &= \int_M \langle \xi, \not{D}\psi \rangle + \int_M \langle \psi, \frac{d}{dt}\not{D}\psi \rangle|_{t=0} \\
&= \int_M \langle \psi, \not{D}(\psi^i \otimes \nabla_{\frac{\partial}{\partial t}} \frac{\partial}{\partial y_i}) \rangle|_{t=0} + \langle \psi, e_\alpha \cdot \psi^i \otimes R(df(\frac{\partial}{\partial t}), df(e_\alpha))\frac{\partial}{\partial y_i} \rangle|_{t=0} \\
&= \int_M \langle \not{D}\psi, \psi^i \otimes \nabla_{\frac{\partial}{\partial t}} \frac{\partial}{\partial y_i} \rangle|_{t=0} + \langle \psi, e_\alpha \cdot \psi^i \otimes R(df(\frac{\partial}{\partial t}), df(e_\alpha))\frac{\partial}{\partial y_i} \rangle|_{t=0} \\
&= \int_M \langle \psi, e_\alpha \cdot \psi^i \otimes R(df(\frac{\partial}{\partial t}), df(e_\alpha))\frac{\partial}{\partial y_i} \rangle|_{t=0} \qquad \text{by (9.3.13)} \\
&= \int_M \langle \psi, e_\alpha \cdot \psi^i \otimes R(\xi^m \frac{\partial}{\partial y_i}, f_\alpha^l \frac{\partial}{\partial y_i})\frac{\partial}{\partial y_i} \rangle \\
&= \int_M \langle \psi, e_\alpha \cdot \psi^i \otimes \xi^m f_\alpha^l R_{iml}^j \frac{\partial}{\partial y_i} \rangle \\
&= \int_M \langle \psi^i, \nabla f^l \cdot \psi^j \rangle R_{mlij}\xi^m.
\end{aligned}$$

Altogether, when (f,ψ) is a critical point of L for such variations, we obtain

$$\frac{dL(f_t)}{dt}|_{t=0} = \int_M \left(-2g_{mi}\tau^i(f) + R_{mlij}\langle\psi^i, \nabla f^l \cdot \psi^j\rangle\right)\xi^m,$$

and hence (9.3.12). □

There are obvious solutions of the Euler-Lagrange equations (9.3.12), (9.3.13), namely those where either f or ψ is trivial. In the first case, we have a constant map f and a harmonic spinor ψ, that is, $\partial\!\!\!/\psi = 0$. In the second case, we have a harmonic map f, that is, a solution of (9.3.12) with vanishing right hand side, and $\psi \equiv 0$. On S^2, we also have an interesting class of nontrivial solutions. For a map $f : S^2 \to S^2$ and a spinor σ on S^2, that is, a smooth section of $\mathbb{S}S^2$, we define a spinor field ψ along f by

$$\psi_{f,\sigma} := e_\alpha \cdot \sigma \otimes f_*(e_\alpha), \tag{9.3.18}$$

for a local orthonormal basis e_α of the tangent space as before.

Proposition 9.3.1. *Let $\psi_{f,\sigma}$ be defined by (9.3.18) from a nonconstant map $f : S^2 \to S^2$ and a spinor σ. Then $(f,\psi_{f,\sigma})$ is a Dirac-harmonic map if and only if f is a (possibly branched) conformal map and σ is a twistor spinor (see (3.4.14)).*

Proof. Let $(f,\psi_{f,\sigma})$ be a Dirac-harmonic map. The spinor field ψ of (9.3.18) satisfies

$$\langle\psi^k, \nabla f^j \cdot \psi^l\rangle = \langle\nabla f^k \cdot \sigma, \nabla f^j \cdot \nabla f^l \cdot \sigma\rangle = f^k_\alpha f^j_\beta f^l_\gamma \langle e_\alpha \cdot \sigma, e_\beta \cdot e_\gamma \cdot \sigma\rangle.$$

Hence $\langle\psi^k, \nabla f^j \cdot \psi^l\rangle$ is purely imaginary by the skew-symmetry of Clifford multiplication. On the other hand, because of the skew-symmetry of R^i_{jkl} with respect to the indices k and l, $R^i_{jkl}\langle\psi^k, \nabla f^j \cdot \psi^l\rangle$ must be real, and hence

$$\frac{1}{2}R^i_{jkl}\langle\psi^k, \nabla f^j \cdot \psi^l\rangle \equiv 0.$$

Thus, if $(f,\psi_{f,\sigma})$ is a Dirac-harmonic map, then f is harmonic by (9.3.12). By Corollary 8.1.5, f therefore is conformal.

We may, as always, choose Riemann normal coordinates so that $\nabla_{e_\alpha} e_\beta = 0$ at the point $x \in S^2$ under consideration. (9.3.13) then yields at x

$$\begin{aligned}
0 &= D\!\!\!\!/\,\psi_{f,\sigma} \\
&= e_\beta \cdot \widetilde{\nabla}_{e_\beta}(e_\alpha \cdot \sigma \otimes f_*(e_\alpha)) \\
&= e_\beta \cdot e_\alpha \cdot \{\nabla_{e_\beta}\sigma \otimes f_*(e_\alpha) + \sigma \otimes \nabla_{e_\beta} f_*(e_\alpha)\} \\
&= -\left(\nabla_{e_\alpha}\sigma \otimes f_*(e_\alpha) + \sigma \otimes \tau(f)\right) \\
&\quad + e_1 \cdot e_2 \cdot (\nabla_{e_1}\sigma \otimes f_*(e_2) - \nabla_{e_2}\sigma \otimes f_*(e_1) + \sigma \otimes (\nabla_{e_1} f_*(e_2) - \nabla_{e_1} f_*(e_2))) \\
&= -\nabla_{e_\alpha}\sigma \otimes f_*(e_\alpha) - e_1 \cdot e_2 \cdot (\nabla_{e_2}\sigma \otimes f_*(e_1) - \nabla_{e_1}\sigma \otimes f_*(e_2))
\end{aligned}$$

Since ϕ is conformal (and non-constant), the above equation is equivalent to

$$e_1 \cdot \nabla_{e_1}\sigma = e_2 \cdot \nabla_{e_2}\sigma, \tag{9.3.19}$$

which says that σ is a twistor spinor, see (3.4.14).

In the other direction, the above computations also yield that if f is a conformal map and σ is a twistor spinor, then $(f, \psi_{f,\sigma})$ is a Dirac-harmonic map. $\square$

We now use the fact that the domain is two-dimensional in order to detect important structural properties of the functional L and its critical points, the Dirac-harmonic maps. These depend on the analogue of Corollary 8.1.4, that is, conformal invariance.

Theorem 9.3.2. *Let $k : \Sigma \to \Sigma$ be a conformal diffeomorphism, with $\mu(z) := |\frac{\partial k}{\partial z}|$. With*

$$\tilde{f} := f \circ k \quad \text{and} \quad \tilde{\psi} = \mu^{-1/2} \psi \circ K, \tag{9.3.20}$$

we have

$$L(f, \psi) = L(\tilde{f}, \tilde{\psi}). \tag{9.3.21}$$

Proof. The conformal invariance of $\int_\Sigma \|\nabla f\|^2 \rho^2 dz d\bar{z}$ follows from Corollary 8.1.4. From (3.4.7), one may infer that the Dirac operator $\tilde{\partial\!\!\!/}$ for the new metric $\rho^2(k(z)) \frac{\partial k}{\partial z} \overline{\frac{\partial k}{\partial z}} dz \otimes d\bar{z}$ satisfies

$$\widetilde{\partial\!\!\!/} \widetilde{\psi} = \mu^{-\frac{3}{2}} \partial\!\!\!/ \psi, \tag{9.3.22}$$

remembering (9.3.20). Hence also

$$\widetilde{D\!\!\!/} \widetilde{\psi} = \mu^{-\frac{3}{2}} D\!\!\!/ \psi, \tag{9.3.23}$$

whence the conformal invariance of $\int \langle \psi, D\!\!\!/ \psi \rangle \rho^2 dz d\bar{z}$. Thus, both terms in L are conformally invariant. $\square$

The conformal invariance of L will now lead to the analogue of Theorem 8.1.1.

Theorem 9.3.3. *Let Σ be a Riemann surface with local holomorphic coordinates $z = x + iy$, N a Riemannian manifold with metric $\langle \cdot, \cdot \rangle_N$ (with associated norm $\|\dot{|}|$), or $(g_{ij})_{i,j=1,\dots,\dim N}$ in local coordinates. If (f, ψ) is Dirac-harmonic, then*

$$\varphi(z)\, dz^2 = \left((\|f_x\|^2 - \|f_y\|^2 - 2i \langle f_x, f_y \rangle) + (\langle \psi, \frac{\partial}{\partial x} \cdot \widetilde{\nabla}_{\frac{\partial}{\partial x}} \psi \rangle - i \langle \psi, \frac{\partial}{\partial x} \cdot \widetilde{\nabla}_{\frac{\partial}{\partial y}} \psi \rangle) \right) dz^2. \tag{9.3.24}$$

is a holomorphic quadratic differential.

Remark. The expression in (9.3.24) involving ψ does not look symmetric in x and y, but the subsequent computations will clarify this issue.

Theorem 9.3.3 can be proved by direct computation, of course, but it is more insightful to derive it from conservation laws.

Define a two-tensor by

$$\phi_{\alpha\beta} := 2\langle f_\alpha, f_\beta \rangle - \delta_{\alpha\beta} \langle f_\gamma, f_\gamma \rangle + \langle \psi, e_\alpha \cdot \widetilde{\nabla}_{e_\beta} \psi \rangle, \tag{9.3.25}$$

where $f_\alpha := f_*(e_\alpha)$. Here, as before, $\{e_\alpha\}$ is a local orthonormal frame on Σ and $\{\eta^\alpha\}$ is a coframe dual to $\{e_\alpha\}$. The tensor $\phi_{\alpha\beta}\eta^\alpha \otimes \eta^\beta$ is called the energy-momentum tensor. This tensor is symmetric and traceless, as we shall now verify. First the symmetry: The equation $\not{D}\psi = 0$ yields

$$e_1 \cdot \widetilde{\nabla}_{e_1}\psi = -e_2 \cdot \widetilde{\nabla}_{e_2}\psi,$$

then

$$e_2 \cdot e_1 \cdot \widetilde{\nabla}_{e_1}\psi = -e_2^2 \cdot \widetilde{\nabla}_{e_2}\psi = \widetilde{\nabla}_{e_2}\psi,$$

that is:

$$-e_1 \cdot e_2 \cdot \widetilde{\nabla}_{e_1}\psi = \widetilde{\nabla}_{e_2}\psi$$

therefore,

$$e_2 \cdot \widetilde{\nabla}_{e_1}\psi = e_1 \cdot \widetilde{\nabla}_{e_2}\psi,$$

which implies that ϕ is symmetric.
The first term in ϕ is traceless by construction, and that the second one is traceless as well follows directly from the equation $\not{D}\psi = 0$.

Proposition 9.3.2. *When (f,ψ) is a Dirac-harmonic map, the energy-momentum tensor is conserved, i.e.,*

$$\sum_\alpha \nabla_{e_\alpha}\phi_{\alpha\beta} = 0. \tag{9.3.26}$$

Proof.

$$\begin{aligned} \nabla_{e_\alpha}\phi_{\alpha\beta} &= \nabla_{e_\alpha}(2\langle f_\alpha, f_\beta\rangle - \delta_{\alpha\beta}\langle f_\gamma, f_\gamma\rangle) + \nabla_{e_\alpha}\langle \psi, e_\alpha \cdot \widetilde{\nabla}_{e_\beta}\psi\rangle \\ &:= I + II. \end{aligned}$$

As before, we choose a local orthonormal basis $\{e_\alpha\}$ on Σ with $\nabla_{e_\alpha}e_\beta = 0$ at the point under consideration. We compute

$$\begin{aligned} I &= 2\langle \nabla_{e_\alpha} f_*(e_\alpha), f_*(e_\beta)\rangle + 2\langle f_*(e_\alpha), \nabla_{e_\alpha} f_*(e_\beta)\rangle \\ &\quad -2\delta_{\alpha\beta}\langle f_*(e_\gamma), \nabla_{e_\alpha} f_*(e_\gamma)\rangle \\ &= 2\langle \tau(f), f_\beta\rangle + 2\langle f_\alpha, \nabla_{e_\beta} f_*(e_\alpha)\rangle - 2\langle f_\gamma, \nabla_{e_\beta} f_*(e_\gamma)\rangle \\ &= 2\langle \tau(f), f_\beta\rangle, \end{aligned}$$

and

$$\begin{aligned} II &= \langle \psi_\alpha, e_\alpha \cdot \psi_\beta\rangle + \langle \psi, e_\alpha \cdot \widetilde{\nabla}_{e_\alpha}\widetilde{\nabla}_{e_\beta}\psi\rangle \\ &= -\langle e_\alpha \cdot \psi_\alpha, \psi_\beta\rangle + \langle \psi, \not{D}\psi_\beta\rangle \\ &= \langle \psi, \not{D}\psi_\beta\rangle. \end{aligned}$$

Therefore, we have

$$\nabla_{e_\alpha}\phi_{\alpha\beta} = 2\langle \tau(f), f_\beta\rangle + \langle \psi, \not{D}\psi_\beta\rangle. \tag{9.3.27}$$

Now

$$\begin{aligned} 2\langle\tau(f), f_\beta\rangle &= 2\langle\frac{1}{2}R^m_{lij}\langle\psi^i, \nabla f^l\cdot\psi^j\rangle\frac{\partial}{\partial y^m}, f^p_\beta\frac{\partial}{\partial y^p}\rangle \\ &= g_{mp}R^m_{lij}\langle\psi^i, \nabla f^l\cdot\psi^j\rangle f^p_\beta \\ &= R_{mlij}\langle\psi^i, \nabla f^l\cdot\psi^j\rangle f^m_\beta. \end{aligned} \tag{9.3.28}$$

We compute $\not{D}\psi_\beta = e_\alpha\cdot\widetilde{\nabla}_{e_\alpha}\widetilde{\nabla}_{e_\beta}\psi$. By a direct computation,

$$\widetilde{\nabla}_{e_\alpha}\widetilde{\nabla}_{e_\beta}\psi - \widetilde{\nabla}_{e_\alpha}\widetilde{\nabla}_{e_\beta}\psi = R^{\mathcal{S}\Sigma}(e_\alpha, e_\beta)\psi^i\otimes\frac{\partial}{\partial y^i} + R^m_{lij}f^i_\alpha f^j_\beta\psi^l\otimes\frac{\partial}{\partial y^m},$$

where $R^{\mathcal{S}\Sigma}$ is the curvature operator of the connection ∇ on the spinor bundle $\mathcal{S}\Sigma$. By (3.4.19), this curvature operator satisfies for a tangent vector V of Σ

$$e_\alpha\cdot R^{\mathcal{S}\Sigma}(e_\alpha, V)\psi^i = \frac{1}{2}Ric(V)\cdot\psi^i. \tag{9.3.29}$$

It follows that

$$\begin{aligned} \langle\psi, e_\alpha\cdot R^{\mathcal{S}\Sigma}(e_\alpha, e_\beta)\psi^i\otimes\frac{\partial}{\partial y^i}\rangle &= \langle\psi^j\otimes\frac{\partial}{\partial y^j}, e_\alpha\cdot R^{\mathcal{S}\Sigma}(e_\alpha, e_\beta)\psi^i\otimes\frac{\partial}{\partial y^i}\rangle \\ &= g_{ij}\langle\psi^j, e_\alpha\cdot R^{\mathcal{S}\Sigma}(e_\alpha, e_\beta)\psi^i\rangle \\ &= \frac{1}{2}g_{ij}\langle\psi^j, Ric(e_\beta)\cdot\psi^i\rangle \\ &= 0. \end{aligned}$$

Therefore

$$\begin{aligned} \langle\psi, \not{D}\psi_\beta\rangle &= \langle\psi, e_\alpha\cdot\widetilde{\nabla}_{e_\alpha}\widetilde{\nabla}_{e_\beta}\psi\rangle \\ &= \langle\psi, \widetilde{\nabla}_{e_\beta}(e_\alpha\cdot\widetilde{\nabla}_{e_\alpha}\psi)\rangle + R^m_{lij}f^j_\beta\langle\nabla f^i\cdot\psi^l\otimes\tfrac{\partial}{\partial y^m}, \psi^p\otimes\tfrac{\partial}{\partial y^p}\rangle \\ &= R^m_{lij}f^j_\beta\langle\psi^p, \nabla f^i\cdot\psi^l\rangle g_{mp} \\ &= -R_{mlij}\langle\psi^i, \nabla f^l\cdot\psi^j\rangle f^m_\beta. \end{aligned}$$

From (9.3.27),(9.3.28) and (9.3.30) we conclude that $\phi_{\alpha\beta}$ is conserved. □

Proof of Theorem 9.3.3: The proof follows directly from Proposition 9.3.2. □

Perspectives. The variational problem presented in this section is a mathematical version of the nonlinear supersymmetric sigma model of quantum field theory. In that model, the variables and fields are Grassmann instead of real valued. In particular, the ones corresponding to the spinor ψ represent fermionic particles and are anticommuting. It was discovered in [50] that one still obtains a rich mathematical structure when one makes all fields real valued, and therefore commuting, even though one then looses the supersymmetry. The conformal invariance of the functional L, however, is not affected. Here, we have followed that reference. Further analytic results are derived in [49]. It remains to explore the geometric significance of Dirac-harmonic maps further, but since they arise from a deep structure in quantum field theory, one naturally also expects deep geometric applications. The physical aspects including supersymmetry are discussed in [77, 78, 144].

Exercises for Chapter 9

1. Show by a direct computation that (9.1.28), (9.1.29) imply (9.1.6), (9.1.7).
2. Derive the Euler-Lagrange equations for the functional defined in (9.2.16).

Appendix A

Linear Elliptic Partial Differential Equations

A.1 Sobolev Spaces

We are going to use the integration theory of Lebesgue. Therefore, we shall always identify functions which differ only on a set of measure zero. Thus, when we speak about a *function*, we actually always mean an equivalence class of functions under the above identification. In particular, a statement like "the function f is continuous" is to be interpreted as "f differs from a continuous function at most on a set of measure zero" or equivalently "the equivalence class of f contains a continuous function".

Replacing functions by their equivalence classes is necessary in order to make the L^p- and Sobolev spaces Banach spaces.

Definition A.1.1. $\Omega \subset \mathbb{R}^d$ open, $p \in \mathbb{R}$, $p \geq 1$,

$$L^p(\Omega) := \Big\{ f : \Omega \to \mathbb{R} \cup \{\pm\infty\} \text{ measurable} \\ \text{and } \|f\|_{L^p(\Omega)} := \Big(\int_\Omega |f(x)|^p dx \Big)^{\frac{1}{p}} < \infty \Big\},$$

$$L^\infty(\Omega) := \Big\{ f : \Omega \to \mathbb{R} \cup \{\pm\infty\} \text{ measurable} \\ \text{and } \|f\|_{L^\infty(\Omega)} := \operatorname*{ess\,sup}_{x \in \Omega} |f(x)| < \infty \Big\}, \quad \text{with}$$

$$\operatorname*{ess\,sup}_{x \in \Omega} f(x) := \inf \big\{ a \in \mathbb{R} \cup \{\infty\} : f(x) \leq a \text{ for almost all } x \in \Omega \big\}.$$

Theorem A.1.1. *With norm* $\|\cdot\|_{L^p(\Omega)}$, $L^p(\Omega)$ *is a Banach space for* $1 \le p \le \infty$.

Theorem A.1.2 (Hölder's Inequality). *Let* $p, q \ge 1$, $\frac{1}{p} + \frac{1}{q} = 1$ *(*$q = \infty$ *for* $p = 1$ *and vice versa),* $f \in L^p(\Omega), g \in L^q(\Omega)$. *Then* $fg \in L^1(\Omega)$ *and*

$$\int_\Omega |f(x)g(x)|\, dx \le \left(\int_\Omega |f(x)|^p dx\right)^{\frac{1}{p}} \left(\int_\Omega |g(x)|^q dx\right)^{\frac{1}{q}}.$$

More generally, for $p_1, \dots p_m \ge 1$, $\frac{1}{p_1} + \dots + \frac{1}{p_m} = 1$, $f_i \in L^{p_i}(\Omega)$, $i = 1, \dots, m$,

$$\int \Big|\prod_{i=1}^m f_i(x)\Big| dx \le \prod_{i=1}^m \left(\int |f_i(x)|^{p_i}\right)^{\frac{1}{p_i}}$$

Theorem A.1.3. *If* $(f_n)_{n\in\mathbb{N}}$ *converges to* f *in* $L^p(\Omega)$, *then a subsequence converges pointwise almost everywhere to* f.

Theorem A.1.4. $C_0^\infty(\Omega)$ *is dense in* $L^p(\Omega)$ *for* $1 \le p < \infty$ *(but not for* $p = \infty$*).*

Theorem A.1.5. *If* $f \in L^2(\Omega)$ *and*

$$\int_\Omega f(x)\varphi(x)\, dx = 0, \quad \textit{for every } \varphi \in C_0^\infty(\Omega),$$

then

$$f = 0.$$

We let

$$L^p_{\text{loc}}(\Omega) := \big\{ f : \Omega \to \mathbb{R} \cup \{\pm\infty\} : f \in L^p(\Omega') \text{ for } \forall\ \Omega' \Subset \Omega \big\}.$$

Definition A.1.2. Let $f \in L^1_{\text{loc}}(\Omega)$. We call $v \in L^1_{\text{loc}}(\Omega)$ the *weak derivative* of f in the direction of x^i, $v = D_i f$, if

$$\int_\Omega v(x)\varphi(x)\, dx = -\int_\Omega f(x) \frac{\partial \varphi(x)}{\partial x^i}\, dx,$$

for all $\varphi \in C_0^1(\Omega)$. Here $x = (x^1, \dots, x^n) \in \mathbb{R}^n$.

Weak derivatives of higher order are similarly defined (notation $D_{\boldsymbol{\alpha}} f$ for a multiindex $\boldsymbol{\alpha}$).

Definition A.1.3. $k \in \mathbb{N}$, $1 \le p \le \infty$. We define the *Sobolev spaces* and *Sobolev norms* as follows:

$$\begin{aligned}
W^{k,p}(\Omega) &:= \{f \in L^p(\Omega) : \forall \boldsymbol{\alpha} \text{ with } |\boldsymbol{\alpha}| \le k : D_{\boldsymbol{\alpha}} f \in L^p(\Omega)\}, \\
\|f\|_{W^{k,p}(\Omega)} &:= \Bigg(\sum_{|\boldsymbol{\alpha}| \le k} \int_\Omega |D_{\boldsymbol{\alpha}} f|^p \Bigg)^{\frac{1}{p}} \text{ for } 1 \le p < \infty, \\
\|f\|_{W^{k,\infty}(\Omega)} &:= \sum_{|\boldsymbol{\alpha}| \le k} \operatorname{ess\,sup}_{\boldsymbol{\alpha} \in \Omega} |D_{\boldsymbol{\alpha}} f(x)|, \\
H_0^{k,p}(\Omega) &:= \text{closure of } C_0^\infty(\Omega) \text{ w.r.t. } \|\cdot\|_{W^{k,p}(\Omega)}, \\
H^{k,p}(\Omega) &:= \text{closure of } C^\infty(\Omega) \text{ w.r.t. } \|\cdot\|_{W^{k,p}(\Omega)}.
\end{aligned}$$

Theorem A.1.6. *$W^{k,p}(\Omega) = H^{k,p}(\Omega)$ for $1 \le p < \infty, k \in \mathbb{N}$. $W^{k,p}(\Omega)$ is a Banach space for $1 \le p \le \infty$, $k \in \mathbb{N}$.*

Some local properties of Sobolev functions:

Lemma A.1.1. *$\Omega \subset \mathbb{R}^d$ open, $f \in H^{1,1}(\Omega)$, $i \in \{1, \dots, d\}$. Then for almost all $\lambda \in \mathbb{R}, f|_{\{x^i = \lambda\}}$ is absolutely continuous.*

Let $f \in L^1(\Omega)$, Ω open in $\mathbb{R}^d$. Then for almost all $x_0 \in \Omega$,

$$\lim_{r \to 0} \frac{1}{|B(x_0, r)|} \int |f(x) - f(x_0))| \, dx = 0$$

($|B(x_0, r)| = \omega_d r^d$ denotes the Lebesgue measure of the ball $B(x_0, r)$).

An x_0 satisfying this property is called a Lebesgue point. If x_0 is a Lebesgue point, then f is approximately continuous at x_0; this means the following:
For $\varepsilon > 0$, let

$$S_\varepsilon := \{y \in \Omega : |f(y) - f(x_0)| < \varepsilon\}.$$

Then

$$\lim_{r \to 0} \frac{|S_\varepsilon \cap B(x_o, r)|}{|B(x_0, r)|} = 1 \quad \text{for all } \varepsilon > 0.$$

Similarly, $f \in H^{1,1}(\Omega)$ is called approximately differentiable at $x_0 \in \Omega$, with approximate derivative $\nabla f(x_0)$, if for

$$S_\varepsilon^1 := \{y \in \Omega : |f(y) - f(x_0)(y - x_0) - \nabla f(x_0)| \le \varepsilon |y - x_0|\},$$

$$\lim_{r \to 0} \frac{|S_\varepsilon^1 \cap B(x_0, r)|}{|B(x_0, r)|} = 1 \quad \text{for all } \varepsilon > 0.$$

We then have

Lemma A.1.2. *A function $f \in H^{1,1}(\Omega)$, $\Omega \subset \mathbb{R}^d$ open, is approximately differentiable almost everywhere, and the weak derivative coincides with the approximate derivative almost everywhere.*

Lemma A.1.3. *$\Omega \subset \mathbb{R}^d$ open, $\ell : \mathbb{R} \to \mathbb{R}$ Lipschitz, $f \in H^{1,p}(\Omega)$. If $\ell \circ f \in L^p(\Omega)$, then $\ell \circ f \in H^{1,p}(\Omega)$ and for almost all $x \in \Omega$,*

$$D_i(\ell \circ f)(x) = \ell'(f(x))D_i f(x), \quad i = 1, \dots, d.$$

Theorem A.1.7 (Sobolev Embedding Theorem). *$\Omega \subset \mathbb{R}^n$ open, bounded, $f \in H_0^{1,p}(\Omega)$. Then*

$$\begin{aligned} &f \in L^{\frac{np}{n-p}} && \text{for } p < n, \\ &f \in C^0(\overline{\Omega}) && \text{for } p > n. \end{aligned}$$

More precisely, $\exists$ constants $c = c(n,p)$:

$$\begin{aligned} \|f\|_{L^{\frac{np}{n-p}}(\Omega)} &\le c\|Df\|_{L^p(\Omega)} && \text{for } p < n, \\ \sup_{x\in\Omega} |f(x)| &\le c\mathrm{Vol}\,(\Omega)^{\frac{1}{n}-\frac{1}{p}}\|Df\|_{L^p(\Omega)} && \text{for } p > n. \end{aligned}$$

For $n = p$, $f \in L^q(\Omega)$ for all $q < \infty$.

Remark. $H^{1,n}(\Omega)$ is not contained in $C^0(\overline{\Omega})$ or $L^\infty(\Omega)$.

Let us consider the following example:
$d \ge 2$, $\Omega = \overset{\circ}{B}(0, \frac{1}{e}) \subset \mathbb{R}^d$, $f(x) := \log\log\frac{1}{|x|}$ is in $H_0^{1,d}(\Omega)$, but has a singularity at $x = 0$ and is unbounded there. Using this example, we may even produce functions in $H^{1,d}$ with a dense set of singular points. For example, take $\Omega = \overset{\circ}{B}(0, \frac{1}{2e}) \subset \mathbb{R}^d$, let $(p_\nu)_{\nu\in\mathbb{N}}$ be a dense sequence of points in Ω and consider

$$g(x) := \sum_\nu 2^{-\nu} f(x - p_\nu).$$

Corollary A.1.1 (Poincaré Inequality). *$\Omega \subset \mathbb{R}^n$ open, bounded,*

$$f \in H_0^{1,2}(\Omega) \Rightarrow \|f\|_{L^2(\Omega)} \le \text{const Vol}\,(\Omega)^{\frac{1}{n}}\|Df\|_{L^2(\Omega)}.$$

Corollary A.1.2. *$\Omega \subset \mathbb{R}^n$ open, bounded, then,*

$$H_0^{k,p}(\Omega) \subset \begin{cases} L^{\frac{np}{n-kp}}(\Omega) & \text{for } kp < n, \\ C^m(\overline{\Omega}) & \text{for } 0 \le m < k - \frac{n}{p}. \end{cases}$$

In particular, if $f \in H_0^{k,p}(\Omega)$ for all $k \in \mathbb{N}$ and some fixed p, then $f \in C^\infty(\overline{\Omega})$.

Theorem A.1.8 (Rellich-Kondrachov Compactness Theorem). *$\Omega \subset \mathbb{R}^n$ open, bounded. Suppose $1 \le q < \frac{np}{n-p}$ if $p < d$, and $1 \le q < \infty$ if $p \ge d$. Then $H_0^{1,p}(\Omega)$ is compactly embedded in $L^q(\Omega)$, i.e. if $(f_n)_{n\in\mathbb{N}} \subset H_0^{1,p}(\Omega)$ satisfies*

$$\|f_n\|_{W^{1,p}(\Omega)} \le \text{const},$$

then a subsequence converges in $L^q(\Omega)$.

Corollary A.1.3. *Ω as before. Then $H_0^{1,2}(\Omega)$ is compactly embedded in $L^2(\Omega)$.*

$H^{k,2}(\Omega)$ is a Hilbert space, the scalar product is

$$(f,g)_{H^{k,2}(\Omega)} := \sum_{|\boldsymbol{\alpha}|\le k} \int_\Omega D_{\boldsymbol{\alpha}} f(x) D_{\boldsymbol{\alpha}} g(x)\, dx.$$

Finally, we recall the concept of weak convergence:

Let H be a Hilbert space with norm $\|\cdot\|$ and a product $\langle\cdot,\cdot\rangle$. Then $(v_n)_{n\in\mathbb{N}} \subset H$ is called *weakly convergent* to $v \in H$,

$$v_n \rightharpoonup v,$$

iff

$$\langle v_n, w\rangle \to \langle v, w\rangle \quad \text{for all } w \in H.$$

Theorem A.1.9. *Every bounded sequence $(v_n)_{n\in\mathbb{N}}$ in H contains a weakly convergent subsequence, and if the limit is v,*

$$\|v\| \le \liminf_{n\to\infty} \|v_n\|$$

(where (v_n) now is the weakly convergent subsequence).

Example. Let (e_n) be an orthonormal sequence in an infinite dimensional Hilbert space. Then $e_n \rightharpoonup 0$. In particular, the inequality in Theorem A.1.9 may be strict.

A.2 Existence and Regularity Theory for Solutions of Linear Elliptic Equations

Ω will always be an open subset of $\mathbb{R}^m$.

For technical purposes, one often has to approximate weak derivatives if they are not yet known to exist by difference quotients which are supposed to exist. Thus, let

$$f \in L^2(\Omega, \mathbb{R}),$$
$$(e_1, \ldots, e_m) \text{ an orthonormal basis of } \mathbb{R}^m,$$
$$h \in \mathbb{R}, \quad h \neq 0.$$

We put

$$\Delta_i^h f(x) := \frac{f(x + he_i) - f(x)}{h} \quad (\text{ifdist}\,(x, \partial\Omega) > |h|).$$

If $\varphi \in L^2(\Omega)$, $\operatorname{supp} \varphi \Subset \Omega$, $|h| < \operatorname{dist}(\operatorname{supp} \varphi, \partial\Omega)$, we have

$$\int_\Omega (\Delta_i^h f(x))\varphi(x)\, dx = -\int_\Omega f(x) \Delta_i^{-h} \varphi(x)\, dx. \tag{A.2.1}$$

Lemma A.2.1. *If $f \in H^{1,2}(\Omega)$, $\Omega' \Subset \Omega$, $|h| < \operatorname{dist}(\Omega', \partial\Omega)$, then $\Delta_i^h f \in L^2(\Omega')$ and*

$$\|\Delta_i^h f\|_{L^2(\Omega')} \le \|D_i f\|_{L^2(\Omega)} \quad \textit{for } i = 1, \ldots, m.$$

Conversely,

Lemma A.2.2. *If $f \in L^2(\Omega)$ and if for some $K < \infty$*

$$\|\Delta_i^{h_n} f\|_{L^2(\Omega')} \le K$$

for some sequence $h_n \to 0$ and all $\Omega' \Subset \Omega$ with $h_n < \operatorname{dist}(\Omega', \partial\Omega)$, then the weak derivative $D_i f$ exists and

$$\|D_i f\|_{L^2(\Omega)} \le K.$$

The fundamental elliptic regularity theorems for Sobolev norms may be proved by approximating weak derivatives by difference quotients.

We now formulate the general regularity theorem.

We consider an operator

$$Lf(x) := \frac{\partial}{\partial x^\alpha}\big(a^{\alpha\beta}(x) \frac{\partial}{\partial x^\beta} f(x)\big) \tag{A.2.2}$$

for $x \in \Omega$, $f : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^m$.

We assume that there exist constants $0 < \lambda \le \mu$ with

$$\lambda |\xi|^2 \le a^{\alpha\beta}(x) \xi_\alpha \xi_\beta \le \mu |\xi|^2 \tag{A.2.3}$$

for all $x \in \Omega, \xi \in \mathbb{R}^m$. We say that L *is uniformly elliptic.* Let $k \in L^2(\Omega)$. Then $f \in H^{1,2}(\Omega)$ is called *weak solution* of

$$Lf = k$$

if

$$\int_\Omega a^{\alpha\beta}(x) D_\beta f(x) D_\alpha \varphi(x)\, dx = -\int_\Omega k(x)\varphi(x)\, dx \tag{A.2.4}$$

for all $\varphi \in H_0^{1,2}(\Omega)$.

Theorem A.2.1. *Let* $f \in H^{1,2}(\Omega)$ *be a weak solution of (A.2.4). Suppose* $k \in H^{\nu,2}(\Omega)$, $a^{\alpha\beta} \in C^{\nu+1}(\Omega)$ $(\nu \in \mathbb{N})$.

Then

$$f \in H^{\nu+2,2}(\Omega')$$

for every $\Omega' \Subset \Omega$.

If

$$\|a^{\alpha\beta}\|_{C^{\nu+1}(\Omega)} \leq K_\nu,$$

then

$$\|f\|_{H^{\nu+2,2}(\Omega')} \leq c(\|f\|_{L^2(\Omega)} + \|k\|_{H^{\nu,2}(\Omega)}), \tag{A.2.5}$$

where c *depends on* m, λ, ν, K_ν *and* $\operatorname{dist}(\Omega', \partial\Omega)$.

The Harnack inequalities of Moser are of fundamental importance for the theory of elliptic partial differential equations:

Theorem A.2.2. *Let* L *be a uniformly elliptic operator as in (A.2.2), (A.2.3).*

(i) Let u *be a weak subsolution, i.e.*

$$Lu \geq 0 \quad \text{in a ball } B(x_0, 4R) \subset \mathbb{R}^m$$

$(\int a^{\alpha\beta} D_p u\, D_\beta\varphi \leq 0 \quad$ *for all* $\varphi \in H_0^{1,2}(B(x_0,4R)))$. *For* $p > 1$ *then*

$$\sup_{B(x_0,R)} u \leq c_1 \Big(\frac{p}{p-1}\Big)^{\frac{2}{p}} \Big(\frac{1}{\omega_m (2R)^m} \int_{B(x_0,2R)} \max(u(x),0)^p\, dx\Big)^{\frac{1}{p}},$$

where c_1 *depends only on* m *and* $\frac{\mu}{\lambda}$ *in (A.2.3).*

(ii) Let u *be a* **positive** *supersolution, i.e.*

$$Lu \leq 0 \quad \text{in a ball } B(x_0, 4R) \subset \mathbb{R}^m.$$

For $m \geq 3$ *and* $0 < p < \frac{m}{m-2}$ *then*

$$\Big(\frac{1}{\omega_m (2R)^m} \int_{B(x_0,2R)} u^p\Big)^{\frac{1}{p}} \leq \frac{c_2}{(\frac{m}{m-2} - p)^2} \inf_{B(x_0,R)} u,$$

c_2 *again depending only on* m *and* $\frac{\mu}{\lambda}$. *For* $m = 2$ *and* $0 < p < \infty$, *the same estimate holds when* $\frac{c_2}{(\frac{m}{m-2}-p)^2}$ *is replaced by a constant* c_3 *depending on* p *and* $\frac{\mu}{\lambda}$.

The Harnack inequality also translates into estimates for the fundamental solutions of the Laplace-Beltrami operator, and their generalizations, the Green functions. The Green function $G(x_0, x)$ of a ball $B \subset M$ (or another sufficiently regular domain), for x_0 in the interior of B, is symmetric in x and x_0, smooth for $x \neq x_0$, becomes singular like $\frac{1}{(d-2)\omega_d} d(x,x_0)^{2-d}$ in case $d = \dim M \geq 3$ ($\omega_d = \operatorname{Vol} S^{d-1}$) (and like $\frac{1}{\omega_2} \log d(x_0,x)$ for $d = 2$), vanishes for $x \in \partial B$, and satisfies

$$h(x_0) = \int_B \Delta h(x)\, G(x_0,x)\, d\mathrm{Vol}\,(x) \quad \text{for all}\, h \in C_0^2(B).$$

A geometric approximation of the Green function (that is exact in the Euclidean case) has been investigated in §4.7. An analytic alternative that allows to avoid the singularity is the use of the mollified Green function. For simplicity, and because that typically suffices for applications, we only consider the case of a ball. The mollified Green function $G^R(x_0,x)$ on the ball $B(x_0,R)$ relative to the ball $B(x_0,2R)$ of double radius, $G^R(x_0,\cdot) \in H^{1,2} \cap C_0^0(B(x_0,2R))$, satisfies

$$\begin{aligned}\int_{B(x_0,2R)} \Delta\varphi(x)\, G^R(x_0,x)\, d\mathrm{Vol}\,(x) &= \int_{B(x_0,2R)} \langle d\varphi(x), dG^R(x_0,x)\rangle\, d\mathrm{Vol}\,(x)\\ &= \fint_{B(x_0,R)} \varphi(x)\, d\mathrm{Vol}\,(x),\end{aligned}$$

for all $\varphi \in H^{1,2}$ with $\operatorname{supp}\varphi \Subset B(x_0,2R)$.

For purposes of normalization, it is convenient to consider

$$w^R(x) := \frac{|B(x_0,2R)|}{R^2} G^R(x_0,x)$$

with $|B| := \operatorname{Vol} B$.

We then have

$$\int_{B(x_0,2R)} \langle d\varphi(x), dw^R(x)\rangle = \frac{1}{R^2} \int_{B(x_0,R)} \varphi(x),$$

for all $\varphi \in H^{1,2}$ with $\operatorname{supp}\varphi \Subset B(x_0,2R)$.

We then have the estimates

Corollary A.2.1.

$$\begin{aligned} 0 \leq w^R \leq \gamma_1 \quad & \text{in } B(x_0,2R),\\ w^R \geq \gamma_2 > 0 \quad & \text{in } B(x_0,R),\end{aligned}$$

for constants γ_1, γ_2 that do not depend on R.

The estimates of J. Schauder are also very important:

Theorem A.2.3. *Let L be as in (A.2.2), (A.2.3), and suppose that the coefficients $a^{\alpha\beta}(x)$ are Hölder continuous in Ω, i.e. contained in $C^\sigma(\Omega)$ for some $0 < \sigma < 1$.*

(i) If u is a weak solution of

$$Lu = k$$

and if k is in $L^\infty(\Omega)$, then u is in $C^{1,\sigma}(\Omega)$, and on every $\Omega_0 \Subset \Omega$, its $C^{1,\sigma}$-norm can be estimated in terms of its L^2-norm and the L^∞-norm of k, with a structural constant depending on Ω, Ω_0, m, σ, λ, μ and the C^σ-norm of the $a^{\alpha\beta}(x)$.

(ii) If u is a weak solution of

$$Lu = k$$

for some $k \in C^{\nu,\sigma}(\Omega)$, $\nu = 0, 1, 2, \ldots$, $0 < \sigma < 1$, and if the coefficients $a^{\alpha\beta}$ are also in $C^{\nu,\sigma}(\Omega)$, then u is in $C^{\nu+2,\sigma}(\Omega)$, and a similar estimate as in (i) holds, this time involving the $C^{\nu,\sigma}$-norm of k and the $a^{\alpha\beta}$.

Finally, we quote the maximum principle.

Theorem A.2.4. *Let $\Omega \subset \mathbb{R}^m$ (or, more generally, $\Omega \subset M$, M a Riemannian manifold) be open and bounded, $f \in C^2(\Omega) \cap C^0(\bar{\Omega})$ with*

$$Lf \geq 0 \quad \text{in } \Omega,$$

L as in (A.2.2), (A.2.3). Then f assumes its maximum on the boundary $\partial\Omega$.

All the preceding results naturally apply to the Laplace-Beltrami operator on a ball $B(x_0, r)$ in a Riemannian manifold M, putting

$$L = -\Delta = \frac{1}{\sqrt{\gamma}} \frac{\partial}{\partial x^\alpha} \left(\sqrt{\gamma} \gamma^{\alpha\beta} \frac{\partial}{\partial x^\beta} \right),$$

$(\gamma_{\alpha\beta})_{\alpha,\beta=1,\ldots,m}$ the metric tensor of M in local coordinates, $(\gamma^{\alpha\beta}) = (\gamma_{\alpha\beta})^{-1}$, $\gamma = \det(\gamma_{\alpha\beta})$.

References for the material in this appendix are: Gilbarg and Trudinger[96], Jost[146] and, with a more elementary presentation, Jost[143]. The results of Corollary A.2.1 about Green functions are systematically derived in [113], and in a more general context in [23]. Some further points about Sobolev spaces can be found in Ziemer[273].

A.3 Existence and Regularity Theory for Solutions of Linear Parabolic Equations

In this section, we consider differential equations on $\Omega \times [0, \infty)$ where Ω is an open subset of $\mathbb{R}^m$ as in A.2, and we continue to use the notations introduced there.

In particular, as before, let the operator L be a *uniformly elliptic* operator of the form

$$Lf(x) := \frac{\partial}{\partial x^\alpha}\big(a^{\alpha\beta}(x)\frac{\partial}{\partial x^\beta}f(x)\big) \tag{A.3.1}$$

with constants $0 < \lambda \leq \mu$ satisfying

$$\lambda|\xi|^2 \leq a^{\alpha\beta}(x,t)\xi_\alpha\xi_\beta \leq \mu|\xi|^2 \tag{A.3.2}$$

for all $x \in \Omega, 0 \leq t, \xi \in \mathbb{R}^m$. The equation we wish to study then is

$$\frac{\partial}{\partial t}f(x,t) - Lf(x,t) = k(x,t) \text{ for } x \in \Omega,\ t \geq 0 \tag{A.3.3}$$

$$f(x,0) = \phi(x) \tag{A.3.4}$$

for some continuous function $\phi(x)$ and some bounded function $k(x,t)$ (and suitable boundary conditions, but since in the text, we are interested in compact manifolds M in place of the open domain Ω, these will not play an essential role and consequently are not emphasized here). (A.3.3) is a linear parabolic partial differential equation

We first state the parabolic maximum principle.

Theorem A.3.1. *Let $\Omega \subset \mathbb{R}^m$ (or, more generally, $\Omega \subset M$, M a Riemannian manifold) be open and bounded, $f \in C^2(\Omega) \cap C^0(\bar{\Omega})$ with respect to x and in $C^1((0,T)) \cap C^0([0,T])$ with respect to t, with*

$$\frac{\partial}{\partial t}f - Lf \leq 0 \quad \text{in } \Omega \times [0,T]. \tag{A.3.5}$$

Then f assumes its maximum for (x,t) with $x \in \partial\Omega$ or for $t = 0$, that is, either on the spatial boundary or at the initial time. In particular, when M is a compact manifold (without boundary), the supremum of $f(\cdot,t)$ is a decreasing function of t.

We have the following existence and regularity theorem for solutions of (A.3.3), with Schauder type estimates

Theorem A.3.2. *Let L be as in (A.3.1), (A.3.2), and suppose that the coefficients $a^{\alpha\beta}(x,t)$ are Hölder continuous in $\Omega \times [0,\infty)$, i.e. contained in $C^\sigma(\Omega \times [0,\infty))$ for some $0 < \sigma < 1$. If we prescribe some boundary values, say $f(y,t) = g(y)$ for all $y \in \partial\Omega$, for some given, e.g. continuous, function g, the solution of (A.3.3) then exists for all $t \geq 0$.*

Furthermore, we have the following estimates:

(i) If u is a weak solution of

$$Lu = k \tag{A.3.6}$$

and if k is in $L^\infty(\Omega \times [0,\infty))$, then as a function of x, u is in $C^{1,\sigma}(\Omega)$, and for every $\Omega_0 \Subset \Omega$ and $t_0 > 0$, its (spatial) $C^{1,\sigma}(\Omega)$-norm on $\Omega_0 \times [t_0,\infty)$ can be estimated in terms of its L^∞-norm and the L^∞-norm of k, with a structural constant depending on Ω, Ω_0, t_0, m, σ, λ, μ and the C^σ-norm of the $a^{\alpha\beta}(x)$.

(ii) If u is a weak solution of

$$Lu = k$$

for some $k \in C^{\nu,\sigma}(\Omega \times [0,\infty))$, $\nu = 0,1,2,\ldots$, $0 < \sigma < 1$, and if the coefficients $a^{\alpha\beta}$ are also in $C^{\nu,\sigma}(\Omega \times [0,\infty))$, then u is in $C^{\nu+2,\sigma}(\Omega)$ with respect to x and of class $C^{\nu+1,\sigma}$ with respect to t, and the corresponding norms can be estimated analogously to (i), this time involving the $C^{\nu,\sigma}$-norm of k and the $a^{\alpha\beta}$.

The restriction to $t \geq t_0 > 0$ can be avoided if the initial values f_0 satisfy appropriate regularity results. The estimates on $[0,\infty)$ will then naturally also involve the corresponding norms of f_0.

Theorem A.3.2 concerns a linear parabolic equation. In the text, we shall encounter nonlinear parabolic equations and systems. For those, the global existence and regularity cannot be deduced from a general result, but rather needs to invoke the detailed structure of the system. What one can deduce from Theorem A.3.2, however, is the short time existence of solutions when the linearization of the differential operator satisfies the assumptions of that theorem. This follows by linearization and the implicit function theorem. That means that for such nonlinear systems, we can obtain the existence of a solution on some interval $[0,T)$ whose length depends on the regularity properties of the initial values. This also implies that the maximal interval of existence for nonlinear parabolic systems is open. For the closedness of the interval of existence, and consequently the existence of a solution for all "time" $t \geq 0$, one then needs to derive specific apriori estimates that prevent solutions from becoming singular in finite time.

A reference for parabolic differential equations and systems is [174]. For a textbook treatment, we refer to [146].

Appendix B

Fundamental Groups and Covering Spaces

In this appendix, we briefly list some topological results. We assume that M is a connected manifold, although the results hold for more general spaces.

A *path* or *curve* in M is a continuous map

$$c : [0, a] \to M \quad (a \geq 0).$$

A *loop* is a path with $c(0) = c(a)$, and that point then is called the *base point* of the loop. The *inverse* of a path c is

$$\begin{aligned} c^{-1} &: [0, a] \to M, \\ c^{-1}(t) &:= c(a - t). \end{aligned}$$

If $c_i : [0, a_i] \to M$ are paths $(i = 1, 2)$ with $c_2(0) = c_1(a_1)$, we can define the product $c_1 \cdot c_2$ as the path $c : [0, a_1 + a_2] \to M$,

$$c(t) = \begin{cases} c_1(t) & \text{for } 0 \leq t \leq a_1, \\ c_2(t - a_1) & \text{for } a_1 \leq t \leq a_1 + a_2. \end{cases}$$

Two paths $c_i : [0, a_i]$ with $c_1(0) = c_2(0)$ and $c_1(a_1) = c_2(a_2)$ are called *equivalent* or *homotopic* if there exists a continuous function

$$H : [0, 1] \times [0, 1] \to M$$

with

$$\begin{aligned} H(t,0) &= c_1\Big(\frac{t}{a_1}\Big), \\ H(t,1) &= c_2\Big(\frac{t}{a_2}\Big), \qquad \text{for all } t, \\ H(0,s) &= c_1(0) = c_2(0), \\ H(1,s) &= c_1(a_1) = c_2(a_2), \qquad \text{for all } s. \end{aligned}$$

In particular, $c : [0,a] \to M$ is equivalent to $\tilde{c} : [0,1] \to M$ with $\tilde{c}(t) = c(\frac{t}{a})$, and so we may assume that all paths are parametrized on the unit interval.

We obtain an equivalence relation on the space of all paths. The equivalence class of c is denoted $[c]$, and it is not hard to verify that $[c_1c_2]$ and $[c^{-1}]$ are independent of the choice of representations. Thus, we may define

$$\begin{aligned} [c_1 \cdot c_2] &=: [c_1] \cdot [c_2], \\ [c^{-1}] &=: [c]^{-1}. \end{aligned}$$

In particular, the equivalence or homotopy classes of loops with fixed base point $p \in M$ form a group $\pi_1(M,p)$, the *fundamental group* of M with *base point* p.

If p and q are in M and $\gamma : [0,1] \to M$ satisfies $\gamma(0) = p$, $\gamma(1) = q$, then for every loop c with base point q, $\gamma^{-1}c\gamma$ is a loop with base point p, and this induces an isomorphism between $\pi_1(M,q)$ and $\pi_1(M,p)$. We may thus speak of the fundamental group $\pi_1(M)$ of M without reference to a base point. M is called *simply connected* if $\pi_1(M) = 0$. A continuous map $f : M \to N$ induces a map $f_\# : \pi_1(M,p) \to \pi_1(N, f(p))$ of fundamental groups.

A continuous map

$$\pi : X \to M$$

is called a *covering map* if each $p \in M$ has a neighborhood U with the property that each connected component of $\pi^{-1}(U)$ is mapped homeomorphically onto U. If $p \in M$ and H is a subgroup of $\pi_1(M,p)$, there exists a covering $\pi : X \to M$ with the property that for any $x \in X$ with $\pi(x) = p$, we have $\pi_*(\pi_1(X,x)) = H$.

If we choose $H = \{1\}$, we obtain a simply connected manifold $\tilde{M}$ and a covering

$$\pi : \tilde{M} \to M.$$

$\tilde{M}$ is called the *universal covering* of M.

If $\pi : X \to M$ is a covering, $c : [0,1] \to M$ a path, $x_0 \in \pi^{-1}(x(0))$, then there exists a unique path

$$\tilde{c} : [0,1] \to X$$

with $\tilde{c}(0) = x_0$ and $c(t) = \pi(\tilde{c}(t))$. $\tilde{c}$ is called the lift of c through x_0.

More generally, if M' is another manifold, $f : M' \to M$ is continuous, $p_0 \in M$, $y_0 \in f^{-1}(p_0)$, $x_0 \in \pi^{-1}(p_0)$, there exists a continuous

$$\tilde{f} : M' \to X$$

with $\tilde{f}(y_0) = x_0$ and $f = \pi \circ \tilde{f}$ if and only if $f_\#(\pi_1(M', y_0)) \subset \pi_\#(\pi_1(X, x_0))$. $\tilde{f}$ is unique if it exists.

Let $\pi : \tilde{M} \to M$ be the universal covering of M. A *deck transformation* is a homeomorphism $\varphi : \tilde{M} \to \tilde{M}$ with

$$\pi = \pi \circ \varphi.$$

Let $\pi(x_0) = p_0$. $\pi_1(M, p_0)$ then bijectively corresponds to $\pi^{-1}(p_0)$. More precisely, $x_1 \in \pi^{-1}(p_0)$ corresponds to the homotopy class of $\pi(\gamma_{x_1})$, where $\gamma_{x_1} : [0,1] \to \tilde{M}$ is any path with $\gamma_{x_1}(0) = x_0$, $\gamma_{x_1}(1) = x_1$. The deck transformations form a group that acts simply transitively on $\pi^{-1}(p_0)$, and associating to a deck transformation $\varphi(x_0) \in \pi^{-1}(p_0)$ then yields an isomorphism between the group of deck transformations and $\pi_1(M, p_0)$.

If M and N are manifolds with universal coverings $\tilde{M}$ and $\tilde{N}$, resp., and if

$$f : M \to N$$

is a continuous map, we consider the induced homomorphism

$$\rho := f_\sharp : \pi_1(M, p) \to \pi_1(N, f(p))$$

of fundamental groups. If $\pi : \tilde{M} \to M$ is the universal covering, we can lift $f \circ \pi : \tilde{M} \to N$ to a map

$$\tilde{f} : \tilde{M} \to \tilde{N},$$

because the above lifting condition is trivially satisfied as $\pi_1(\tilde{M}) = \{1\}$. $\tilde{f}$ is equivariant w.r.t. the above homomorphism ρ in the sense that for every $\lambda \in \pi_1(M, p)$, acting as a deck transformation on $\tilde{M}$, we have

$$\tilde{f}(\lambda x) = \rho(\lambda)\tilde{f}(x) \quad \text{for every } x \in \tilde{M}, \tag{B.1}$$

where $\rho(\lambda)$ acts as a deck transformation on $\tilde{N}$. We say that $\tilde{f}$ is a ρ-equivariant map between the universal covers $\tilde{M}$ and $\tilde{N}$.

Conversely, given any homomorphism

$$\rho : \pi_1(M, p) \to \pi_1(N, q)$$

and any ρ-equivariant map

$$g : \tilde{M} \to \tilde{N} \quad (\text{with } g(p) = q),$$

not necessarily continuous, then g induces a map

$$g' : M \to N$$

whose lift to universal covers is g. g' is continuous if g is.

Finally, if $\tilde{M}$ is the universal cover of a compact Riemannian manifold M, a so-called fundamental domain $F(M)$ for M in $\tilde{M}$ can be constructed as follows:

For simplicity of notation, we denote the group $\pi_1(M, x_0)$ operating by deck transformations on $\tilde{M}$ by Γ, and its trivial element by e.

Let $d(.,.)$ be the Riemannian distance function on $\tilde{M}$. We select any $z_0 \in \tilde{M}$. We then put

$$F(M) := \{z \in \tilde{M} : d(z, z_0) < d(\gamma z, z_0) \quad \text{for all } \gamma \in \Gamma, \gamma \neq e\}.$$

$F(M)$ is open. Since Γ operates by isometries, i.e.

$$d(\lambda z_1, \lambda z_2) = d(z_1, z_2) \quad \text{for all } \lambda \in \Gamma, z_1, z_2 \in \tilde{M},$$

we may also write

$$F(M) = \{z \in \tilde{M} : d(z, z_0) < d(z, \lambda z_0) \quad \text{for all } \lambda \in \Gamma, \lambda \neq e\}.$$

By its definition, $F(M)$ cannot contain any two points that are equivalent under the operation of Γ. On the other hand, for any $z \in \tilde{M}$, we may find some $\mu \in \Gamma$ such that

$$\mu z \in \overline{F(M)}.$$

Thus, the closure of $F(M)$ contains at least one point from every orbit of Γ in $\tilde{M}$.

If $f : M \to \mathbb{R}$ is an integrable function, and if $\tilde{f} : \tilde{M} \to \mathbb{R}$ is its lift to the universal cover of M, then

$$\int_M f(x)\, d\mathrm{Vol}\,(x) = \int_{F(M)} \tilde{f}(y)\, d\mathrm{Vol}\,(y).$$

Examples of fundamental groups.

1. $\pi_1(\mathbb{R}^n) = \{1\}$ for all n.
2. $\pi_1(S^1) = \mathbb{Z}$.

   A generator is given by

$$c : [0,1] \to S^1 = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\},$$
$$c(t) = (\cos 2\pi t, \sin 2\pi t).$$

   The universal covering of S^1 is $\mathbb{R}^1$, and the covering map is likewise given by

$$\pi(t) = (\cos 2\pi t, \sin 2\pi t).$$

3. $\pi_1(S^n) = \{1\}$ for $n \geq 2$.
4. $\pi_1(\mathrm{SO}(n)) = \mathbb{Z}_2$ for $n \geq 3$.

The preceding results can be found in any reasonable textbook on Algebraic Topology, for example in [99] or [240].

Bibliography

[1] A. Abbondandolo and P. Majer. Morse homology on Hilbert spaces. *Comm. Pure Appl. Math.*, 54:689–760, 2001.

[2] U. Abresch and W. Meyer. Pinching below $\frac{1}{4}$, injectivity radius, and conjugate radius. *J. Diff. Geom.*, 40:643–691, 1994.

[3] U. Abresch and W. Meyer. A sphere theorem with pinching constant below $\frac{1}{4}$. *J. Diff. Geom.*, 44:214–261, 1996.

[4] S.I. Al'ber. On n–dimensional problems in the calculus of variations in the large. *Sov. Math. Dokl.*, 5:700–704, 1964.

[5] S.I. Al'ber. Spaces of mappings into a manifold with negative curvature. *Sov. Math. Dokl.*, 9:6–9, 1967.

[6] F. Almgren. Questions and answers about area–minimizing surfaces and geometric measure theory. *Proc. Symp. Pure Math.*, 54(I):29–53, 1993.

[7] M. Atiyah, R. Bott, and A. Shapiro. Clifford moduls. *Topology*, 3(Suppl. I):3–38, 1964.

[8] M. Atiyah and I. Singer. The index of elliptic operators:III. *Ann. Math.*, 87:546–604, 1963.

[9] T. Aubin. *Nonlinear analysis on manifolds. Monge–Ampère equations.* Springer–Verlag, Berlin, 1982.

[10] W. Ballmann. Der Satz von Ljusternik und Schnirelman. *Bonner Math. Schriften*, 102:1–25, 1978.

[11] W. Ballmann. *Lectures on spaces of nonpositive curvature.* DMV Seminar Vol.25, Birkhäuser, 1995.

[12] W. Ballmann, M. Gromov, and V. Schroeder. *Manifolds of nonpositive curvature.* Birkhäuser, 1985.

[13] V. Bangert. On the existence of closed geodesics on two–spheres. *Internat. J. Math*, 4:1–10, 1993.

[14] H. Bass and J. Morgan (eds.). *The Smith conjecture.* Academic Press, 1984.

[15] V.N. Berestovskij and I.G. Nikolaev. Multidimensional generalized Riemannian spaces. In Y. G. Reshetnyak, editor, *Geometry IV*, pages 165 – 243. Encyclopedia Math. Sciences Vol 70, Springer, 1993.

[16] M. Berger. Les variétés Riemanniennes ($\frac{1}{4}$)–pincées. *Ann. Scuola Norm. Sup. Pisa*, III(14):161–170, 1960.

[17] M. Berger. *A panoramic view of Riemannian geometry.* Springer, 2003.

[18] N. Berline, E. Getzler, and M. Vergne. *Heat kernels and Dirac operators.* Springer, 1992.

[19] A. Besse. *Einstein manifolds.* Springer, 1987.

[20] G. Besson, G. Courtois, and S. Gallot. Entropies et rigidités des espaces localement symétriques de courbure strictement négative. *GAFA*, 5:731–799, 1995.

[21] G. Besson, G. Courtois, and S. Gallot. Minimal entropy and Mostow's rigidity theorems. *Ergod. Th. & Dynam. Sys*, 16:623–649, 1996.

[22] M. Betz and R.Cohen. Graph moduli spaces and cohomology operations. Preprint, 1995.

[23] M. Biroli and U. Mosco. A Saint-Venant principle for Dirichlet forms on discontinuous media. *Annali Mat. Pura Appl.(IV)*, 169:125–181, 1995.

[24] S. Bochner. Harmonic surfaces in Riemannian metric. *Trans.AMS*, 47:146–154, 1940.

[25] S. Bochner and K. Yano. *Curvature and Betti numbers.* Princeton Univ. Press, 1953.

[26] J. Borchers and W. Garber. Analyticity of solutions of the O(N) nonlinear σ–model. *Comm. Math. Phys*, 71:299–309, 1980.

[27] J.P. Bourguignon. *Eugenio Calabi and Kähler metrics*, pages 61–85. In: P. de Bartolomeis(ed.), Manifolds and geometry, Symp. Math. 36, Cambridge Univ. Press, 1996.

[28] J.P. Bourguignon and H. Karcher. Curvature operators: Pinching estimates and geometric examples. *Ann. scient. Éc. Norm. Sup*, 11:71–92, 1978.

[29] S. Bradlow. Vortices in holomorphic line bundles over closed Kähler manifolds. *Comm. Math. Phys*, 135:1–17, 1990.

[30] S. Bradlow. Special metrics and stability for holomorphic bundles with global sections. *J. Diff. Geom*, 33:169–214, 1991.

[31] S. Brendle and R. Schoen. Manifolds with 1/4-pinched curvature are space forms. *arxiv*, 0705.0766, 2007.

[32] H. Brézis and J. Coron. Large solutions for harmonic maps in two dimensions. *Comm. Math. Phys*, 92:203–215, 1983.

[33] C. Le Brun. Einstein metrics and Mostow rigidity. *Math. Res. Lett.*, 2:1 – 8, 1995.

[34] Yu. Burago, M. Gromov, and G. Perel'man. A. D. Alexandrov's spaces with curvatures bounded from below. *Russ. Math. Surveys*, 42:1–58, 1992.

[35] P. Buser and H. Karcher. *Gromov's almost flat manifolds.* Astérisque, 1981.

[36] L. Caffarelli and Y.S. Yang. Vortex condensation in the Chern-Simons-Higgs Model. An existence theorem. *Comm. Math. Phys*, 168:321–336, 1995.

[37] E. Calabi and E. Vesentini. On compact, locally symmetric Kähler manifolds. *Ann. Math*, 71:472–507, 1960.

[38] H. D. Cao and X. P. Zhu. A complete proof of the Poincaré and geometrization conjectures - application of the Hamilton-Perelman theory of the Ricci flow. *Asian J. Math.*, 10(2):195–492, 2006.

[39] H. D. Cao and X. P. Zhu. Hamilton-Perelman's proof of the Poincaré conjecture and the geometrization conjecture. arXiv:math.DG/0612069, 2006.

[40] J. Cao and F. Xavier. Kähler parabolicity and the Euler number of compact of non–positive sectional curvature. *Math. Ann*, 319:483–491, 2001.

[41] K.C. Chang. Heat flow and boundary value problem for harmonic maps. *Anal. Nonlinéaire*, 6:363–396, 1989.

[42] K.C. Chang. *Infinite dimensional Morse theory and multiple solution problems.* Birkhäuser, 1993.

[43] I. Chavel. *Riemannian geometry - A modern introduction.* Cambridge University Press, 1993.

[44] J. Cheeger. Finiteness theorems for Riemannian manifolds. *Amer. J. Math.*, 92:61–74, 1970.

[45] J. Cheeger and D. Ebin. *Comparison theorems in Riemannian geometry.* North Holland, 1975.

[46] J. Cheeger and D. Gromoll. The splitting theorems for manifolds of nonnegative Ricci curvature. *J. Diff. Geom*, 6:119–129, 1971.

[47] J. Cheeger and M. Gromov. Collapsing Riemannian manifolds while keeping their curvature bounded. *I. J. Diff. Geom.*, 23:309–346, 1983.

[48] J. Cheeger and M. Gromov. Collapsing Riemannian manifolds while keeping their curvature bounded. *II, J. Diff. Geom.*, 32:269–298, 1990.

[49] Q. Chen, J. Jost, J.Y. Li, and G.F. Wang. Regularity theorems and energy identities for Dirac-harmonic maps. *Math. Zeitschr.*, 251:61–84, 2005.

[50] Q. Chen, J. Jost, J.Y. Li, and G.F. Wang. Dirac-harmonic maps. *Math. Zeitschr.*, 254:409–432, 2006.

[51] W.Y. Chen and J. Jost. A Riemannian version of Korn's inequality. *Calc. Var.*, 14:517–530, 2001.

[52] B. Chow and D. Knopf. *The Ricci flow: An introduction.* Amer. Math. Soc., 2004.

[53] B. Chow, P. Lu, and L. Ni. *Hamilton s Ricci flow.* Amer. Math. Soc. and Intern. Press, 2007.

[54] S. N. Chow and J. Hale. *Methods of bifurcation theory.* Springer, 1982.

[55] C. Conley. *Isolated invariant sets and the Morse index.* CBMS Reg. Conf. Ser. Math. 38, AMS, Providence, R.I., 1978.

[56] C. Conley and E. Zehnder. Morse-type index theory for flows and periodic solutions for Hamiltonian equations. *Comm. Pure Appl. Math.*, 37:207–253, 1984.

[57] K. Corlette. Archimedean superrigidity and hyperbolic geometry. *Ann. Math.*, 135:165–182, 1992.

[58] G. dal Maso. *An introduction to* Γ*–convergence.* Birkhäuser, 1993.

[59] D. de Turck and J. Kazdan. Some regularity theorems in Riemannian geometry. *Ann. Sci. École Norm. Sup.*, 4(14):249–260, 1981.

[60] U. Dierkes, S. Hildebrandt, A. Küster, and O. Wohlrab. *Minimal surfaces, 2 volumes.* Springer, 1992.

[61] W.Y. Ding. Lusternik-Schnirelman theory for harmonic maps. *Acta Math. Sinica*, 2:105–122, 1986.

[62] W.Y. Ding, J. Jost, J.Y. Li, X.W. Peng, and G.F. Wang. Self duality equations for Ginzburg-Landau and Seiberg-Witten type functionals with 6^{th} order potentials. *Comm. Math. Phys.*, 217(2):383–407, 2001.

[63] W.Y. Ding, J. Jost, J.Y. Li, and G.F. Wang. An analysis of the two–vortex case in the Chern-Simons-Higgs model. *Calc. Var.*, 7:87–97, 1998.

[64] W.Y. Ding, J. Jost, J.Y. Li, and G.F. Wang. Multiplicity results for the two vortex Chern-Simons-Higgs model on the two-sphere. *Commentarii Math. Helv.*, 74:118–142, 1999.

[65] M. do Carmo. *Riemannian geometry.* Birkhäuser, 1992.

[66] S. Donaldson and P. Kronheimer. *The geometry of four-manifolds*. Oxford Univ. Press, 1990.

[67] H. Donnelly and F. Xavier. On the differential form spectrum of negatively curved Riemannian manifolds. *Amer. J. Math.*, 106:169–185, 1984.

[68] P. Eberlein. *Rigidity problems for manifolds of nonpositive curvature*. Springer LNM 1156, 1984.

[69] P. Eberlein. *Geometry of nonpositively curved manifolds*. Univ. Chicago Press, 1996.

[70] P. Eberlein, U. Hamenstädt, and V. Schroeder. Manifolds of nonpositive curvature. *Proc. Symp. Pure Math*, 54:Part 3, 179–227, 1993.

[71] P. Eberlein and B. O'Neill. Visibility manifolds. *Pacific J. Math.*, 46:45–109, 1973.

[72] J. Eells and L. Lemaire. A report on harmonic maps. *Bull. London Math. Soc.*, 10:1–68, 1978.

[73] J. Eells and L. Lemaire. Selected topics in harmonic maps. *CBMS Reg. Conf. Ser.*, 50, 1983.

[74] J. Eells and L. Lemaire. Another report on harmonic maps. *Bull. London Math. Soc.*, 20:385–524, 1988.

[75] J. Eells and J. Sampson. Harmonic mappings of Riemannian manifolds. *Am. J. Math.*, 85:109–160, 1964.

[76] J. Eschenburg and J. Jost. *Differentialgeometrie und Minimalflächen*. Springer, 2007.

[77] P. Deligne et al. *Quantum fields and strings: a course for mathematicians, Vol. I.* Amer.Math.Soc. and Inst. Adv. Study, Princeton, N. J., 1999.

[78] P. Deligne et al. *Quantum fields and strings: a course for mathematicians, Vol. II.* Amer.Math.Soc. and Inst. Adv. Study, Princeton, N. J., 1999.

[79] F. Fang and X. Rong. Positive pinching, volume and second betti number. *GAFA*, 9:641–674, 1999.

[80] H. Federer. *Geometric measure theory*. Springer, 1979.

[81] A. Floer. Witten's complex and infinite dimensional Morse theory. *J. Diff. Geom.*, 30:207–221, 1989.

[82] A. Floer and H. Hofer. Coherent orientations for periodic orbit problems in symplectic geometry. *Math. Z.*, 212:13–38, 1993.

[83] J. Franks. Morse-Smale flows and homotopy theory. *Topology*, 18:199–215, 1979.

[84] J. Franks. Geodesics on S^2 and periodic points of annulus diffeomorphisms. *Inv. math.*, 108:403–418, 1992.

[85] D. Freed. Special Kähler manifolds. *Comm. Math. Phys.*, 203:31–52, 1999.

[86] D. Freed and K. Uhlenbeck. *Instantons and 4-manifolds.* Springer, 1984.

[87] Th. Friedrich. *Neue Invarianten der 4–dimensionalen Mannigfaltigkeiten.* SFB 288, Berlin, 1995.

[88] Th. Friedrich. *Dirac-Operatoren in der Riemannschen Geometrie.* Vieweg, 1997.

[89] K. Fukaya. A boundary of the set of Riemannian manifolds with bounded curvatures and diameters. *J. Diff. Geom.*, 28:1–21, 1988.

[90] S. Gallot, D. Hulin, and J. Lafontaine. *Riemannian geometry.* Springer, 1987.

[91] O. García-Prada. Invariant connections and vortices. *Comm. Math. Phys.*, 156:527–546, 1993.

[92] O. García-Prada. Dimensional reduction of stable bundles, vortices and stable pairs. *Intern. J. Math*, 5:1–52, 1994.

[93] O. García-Prada. Proof for the vortex equations over a compact Riemann surface. *Bull. London Math. Soc.*, 26:88–96, 1994.

[94] M. Giaquinta and E. Giusti. On the regularity of the minima of variational integrals. *Acta Math.*, 148:31–46, 1982.

[95] M. Giaquinta and E. Giusti. The singular set of the minima of certain quadratic functionals. *Ann. Scuola Norm. Sup. Pisa, Cl. Sci.*, (4) 11:45–55, 1984.

[96] D. Gilbarg and N.Trudinger. *Elliptic partial differential equations of second order.* Springer, 1977.

[97] M. Goresky and R. MacPherson. *Stratified Morse theory.* Ergebnisse 14, Springer, 1988.

[98] M. Grayson. Shortening embedded curves. *Ann. Math.*, 120:71–112, 1989.

[99] M. Greenberg. *Lectures on algebraic topology.* Benjamin, Reading, Mass., 1967.

[100] R. Greene and H. Wu. Lipschitz convergence of Riemannian manifolds. *Pacific J. Math.*, 131:119–141, 1988.

[101] P. Griffiths and J. Harris. *Principles of algebraic geometry.* Wiley–Interscience, 1978.

[102] D. Gromoll. Spaces of nonnegative curvature. *Proc. Sym. Pure Math.*, 54(Part 3):337–356, 1993.

[103] D. Gromoll, W. Klingenberg, and W. Meyer. *Riemannsche Geometrie im Großen.* Springer LNM 55, ^{2}, 1975.

[104] D. Gromoll and J. Wolf. Some relations between the metric structure and the algebraic structure of the fundamental group in manifolds of nonpositive curvature. *Bull. AMS*, 77:545–552, 1971.

[105] M. Gromov. *Structures métriques pour les variétés riemanniennes.* Rédigé par J. Lafontaine and P. Pansu. Cedic-Nathan, Paris, 1980.

[106] M. Gromov. Pseudoholomorphic curves in symplectic geometry. *Inv. math.*, 82:307–347, 1985.

[107] M. Gromov. Kähler hyperbolicity and L^2-Hodge theory. *J. Diff. Geom.*, 33:263–292, 1991.

[108] M. Gromov. *Metric structures for Riemannian and non-Riemannian spaces.* Birkhäuser, 1999.

[109] M. Gromov and B. Lawson. The classification of simply connected manifolds of positive scalar curvature. *Ann. Math.*, 111:423–434, 1980.

[110] M. Gromov and B. Lawson. Spin and scalar curvature in the presence of a fundamental group. *Ann. Math.*, 111:209–230, 1980.

[111] M. Gromov and R. Schoen. Harmonic maps into singular spaces and p-adic superrigidity for lattices in groups of rank one. *Publ. Math. IHES*, 76:165–246, 1992.

[112] M. Grüter. Regularity of weak H-surfaces. *J. reine angew. Math.*, 329:1–15, 1981.

[113] M. Grüter and K.-O. Widman. The Green function for uniformly elliptic equations. *man. math.*, 37:303–342, 1982.

[114] M. Günther. Zum Einbettungssatz von J. Nash. *Math. Nachr.*, 144:165–187, 1989.

[115] R. Hamilton. *Harmonic maps of manifolds with boundary.* Springer LNM 471, 1975.

[116] R. Hamilton. Three–manifolds with positive Ricci curvature. *J. Diff. Geom.*, 17:255–306, 1982.

[117] R. Hamilton. The Harnack estimate for the Ricci flow. *J. Diff. Geom.*, 37:225–243, 1993.

[118] R. Hardt and L. Simon. Boundary regularity and embedded solutions for the oriented plateau problem. *Ann. Math.*, 110:439–486, 1979.

[119] P. Hartman. On homotopic harmonic maps. *Can. J. Math.*, 19:673–687, 1967.

[120] S. Hawking and Ellis. *The large scale structure of space–time*. Cambridge University Press, 1973.

[121] F. Hélein. Régularité des applications faiblement harmoniques entre une surface et une varieté riemannienne. *C.R. Acad. Sci. Paris*, 312:591–596, 1991.

[122] F. Hélein. *Harmonic maps, conservation laws and moving frames*. Cambridge Univ. Press, 2002.

[123] S. Helgason. *Differential Geometry, Lie Groups and Symmetric Spaces*. Academic Press, 1978.

[124] S. Hildebrandt, H. Kaul, and K.O. Widman. Harmonic mappings into Riemannian manifolds with non–positive sectional curvature. *Math. Scand.*, 37:257–263, 1975.

[125] S. Hildebrandt, H. Kaul, and K.O. Widman. An existence theorem for harmonic mappings of Riemannian manifolds. *Acta Math.*, 138:1–16, 1977.

[126] N. Hingston. On the growth of the number of closed geodesics on the two–sphere. *Intern. Math. Res.*, 9:253–262, 1993.

[127] N. Hitchin. The self–duality equations on a Riemann surface. *Proc. London Math. Soc.*, 55:19–126, 1987.

[128] M.C. Hong, J. Jost, and M. Struwe. Asymptotic limits of a Ginzburg–Landau type functional. In J. Jost, editor, *Geometric Analysis and the Calculus of Variations for Stefan Hildebrandt*, pages 99–123. International Press, Boston, 1996.

[129] D. Husemoller. *Fibre bundles*. Springer, GTM 20, 1975.

[130] T. Ishihara. A mapping of Riemannian manifolds which preserves harmonic functions. *J. Math. Kyoto Univ.*, 19:215–229, 1979.

[131] W. Jäger and H. Kaul. Uniqueness and stability of harmonic maps and their Jacobi fields. *Man. math.*, 28:269–291, 1979.

[132] J. Jost. *Harmonic mappings between Riemannian manifolds*. ANU–Press, Canberra, 1984.

[133] J. Jost. The Dirichlet problem for harmonic maps from a surface with boundary onto a 2–sphere with non–constant boundary values. *J. Diff. Geom.*, 19:393–401, 1984.

[134] J. Jost. The geometric calculus of variations: A short survey and a list of open problems. *Expo. Math.*, 6:111–143, 1988.

[135] J. Jost. A nonparametric proof of the theorem of Ljusternik and Schnirelman. *Arch. Math.*, 53:497–509, 1989. Correction in *Arch. Math.* 56 (1991), 624.

[136] J. Jost. *Two-dimensional geometric variational problems.* Wiley–Interscience, 1991.

[137] J. Jost. Unstable solutions of two–dimensional geometric variational problems. *Proc. Symp. Pure Math*, 54(I):205–244, 1993.

[138] J. Jost. Equilibrium maps between metric spaces. *Calc. Var.*, 2:173–204, 1994.

[139] J. Jost. Convex functionals and generalized harmonic maps into spaces of non-positive curvature. *Comment. Math. Helv.*, 70:659–673, 1995.

[140] J. Jost. Generalized harmonic maps between metric spaces. In J. Jost, editor, *Geometric analysis and the calculus of variations for Stefan Hildebrandt*, pages 143–174. Intern. Press, Boston, 1996.

[141] J. Jost. Generalized Dirichlet forms and harmonic maps. *Calc. Var.*, 5:1–19, 1997.

[142] J. Jost. *Nonpositive curvature: Geometric and analytic aspects.* Birkhäuser, 1997.

[143] J. Jost. *Postmodern analysis.* Springer, 1998, ^{3}2005.

[144] J. Jost. *Geometry and physics.* Springer, 2008.

[145] J. Jost. *Nonlinear methods in Riemannian and Kählerian geometry.* Birkhäuser, ^{2}1991.

[146] J. Jost. *Partial differential equations.* Springer, ^{2}2007.

[147] J. Jost. *Compact Riemann surfaces.* Springer, ^{3}2006.

[148] J. Jost and H. Karcher. Geometrische Methoden zur Gewinnung von a-priori-Schranken für harmonische Abbildungen. *Man. math.*, 40:27–77, 1982.

[149] J. Jost and X. Li-Jost. *Calculus of variations.* Cambridge Univ. Press, 1998.

[150] J. Jost, X.W. Peng, and G.F. Wang. Variational aspects of the Seiberg-Witten functional. *Calc. Var.*, 4:205–218, 1996.

[151] J. Jost and M. Struwe. Morse-Conley theory for minimal surfaces of varying topological type. *Inv. math.*, 102:465–499, 1990.

[152] J. Jost and L. Todjihounde. Harmonic nets in metric spaces. *Pacific J. Math.*, 2007.

[153] J. Jost and Y. L. Xin. Vanishing theorems for L^2-cohomology groups. *J. reine angew. Math.*, 525, 2000.

[154] J. Jost and S.T. Yau. Harmonic mappings and Kähler manifolds. *Math. Ann.*, 262:145–166, 1983.

[155] J. Jost and S.T. Yau. A nonlinear elliptic system for maps from Hermitian to Riemannian manifolds and rigidity theorems in Hermitian geometry. *Acta Math.*, 170:221–254, 1993. Corr. in *Acta Math.* 173 (1994), 307.

[156] J. Jost and S.T. Yau. Harmonic maps and superrigidity. *Proc. Symp. Pure Math.*, 54(I):245–280, 1993.

[157] J. Jost and K. Zuo. Harmonic maps of infinite energy and rigidity results for Archimedean and non-Archimedean representations of fundamental groups of quasiprojective varieties. *J. Diff. Geom.*, 47:469–503, 1997.

[158] J. Jost and K. Zuo. Vanishing theorems for L^2-cohomology on infinite coverings of compact Kähler manifolds and applications in algebraic geometry. *Comm. Anal. Geom.*, 8:1–30, 2000.

[159] E. Kähler. Über eine bemerkenswerte Hermitesche Metrik. *Abh. Math. Sem. Univ. Hamburg*, 9:173–186, 1933.

[160] E. Kähler. Der innere Differentialkalkül. *Rend. Mat. Appl.(5)*, 21:425–523, 1962.

[161] E. Kähler. *Mathematische Werke – Mathematical Works.* edited by R. Berndt and O. Riemenschneider: ; de Gruyter, 2003.

[162] H. Karcher. Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.*, 30:509–541, 1977.

[163] W. Kendall. Convexity and the hemisphere. *J. London Math. Soc.*, 43(2):567–576, 1991.

[164] Kervaire. A manifold which does not admit any differentiable structure. *Comment. Math. Helv.*, 34:257–270, 1960.

[165] B. Kleiner and J. Lott. Notes on Perelman s papers. *arXiv*, math.DG/0605667. p:47–54, 2006.

[166] W. Klingenberg. Über Riemannsche Mannigfaltigkeiten mit positiver Krümmung. *Comm. Math. Helv.*, 35:47–54, 1961.

[167] W. Klingenberg. *Lectures on closed geodesics.* Springer, 1978.

[168] W. Klingenberg. *Riemannian geometry.* de Gruyter, 1982.

[169] S. Kobayashi and K. Nomizu. *Foundations of differential geometry, I.* Wiley–Interscience, 1963.

[170] S. Kobayashi and K. Nomizu. *Foundations of differential geometry, II.* Wiley–Interscience, 1969.

[171] N. Korevaar and R. Schoen. Sobolev spaces and harmonic maps for metric space targets. *Comm. Anal. Geom.*, 1:561–659, 1993.

[172] N. Korevaar and R. Schoen. Global existence theorems for harmonic maps to non–locally compact spaces. *Comm. Anal. Geom.*, 5:213–266, 1997.

[173] P. Kronheimer and T. Mrowka. The genus of embedded surfaces in the projective plane. *Math. Res. Letters*, 1:797–808, 1994.

[174] O.A. Ladyzhenskaya, V. A. Solonnikov, and N.N. Ural'ceva. *Linear and quasilinear equations of parabolic type.* Translated from the Russian, Amer. Math. Soc., Providence, R. I., 1968.

[175] O.A. Ladyzhenskaya and N.N. Ural'ceva. *Linear and quasilinear elliptic equations.* Translated from the Russian, Academic Press, 1968.

[176] H.B. Lawson and M.L. Michelsohn. *Spin geometry.* Princeton University Press, Princeton, 1989.

[177] H.B. Lawson and S.T. Yau. Compact manifolds of nonpositive curvature. *J. Diff. Geom.*, 7:211–228, 1972.

[178] L. Lemaire. Applications harmoniques de surfaces riemanniennes. *J. Diff. Geom.*, 13:51–78, 1978.

[179] F.H. Lin. Analysis on singular spaces. In Ta-Tsien Li, editor, *Geometry, Analysis and Mathematical Physics, in honor of Prof. Chaohao Gu*, pages 114–126. World Scientific, 1997.

[180] J. Lohkamp. Negatively Ricci curved manifolds. *Bull. AMS*, 27:288–292, 1992.

[181] J. Lohkamp. Metrics of negative Ricci curvature. *Ann. Math.*, 140:655–683, 1994.

[182] U. Ludwig. Stratified Morse theory with tangential conditions.

[183] G.A. Margulis. Discrete groups of motion of manifolds of nonpositive curvature. *AMS Transl.*, 190:33–45, 1977.

[184] G.A. Margulis. *Discrete subgroups of semisimple Lie groups.* Springer, 1991.

[185] Y. Matsushima. On the first Betti number of compact quotient spaces of higher-dimensional symmetric spaces. *Ann. Math.*, 75:312–330, 1962.

[186] W. Meeks and S.T. Yau. The classical Plateau problem and the topology of three-dimensional manifolds. *Top.*, 21:409–440, 1982.

[187] M. Micallef and J.D. Moore. Minimal two–spheres and the topology of manifolds with positive curvature on totally isotropic two–planes. *Ann. Math.*, 127:199–227, 1988.

[188] A. Milgram and P. Rosenbloom. Harmonic forms and heat conduction, I: Closed Riemannian manifolds. *Proc. Nat. Acad. Sci.*, 37:180–184, 1951.

[189] J. Milnor. On manifolds homeomorphic to the 7-sphere. *Ann. Math.*, 64:399–405, 1956.

[190] J. Milnor. Differentiable structures on spheres. *Am. J. Math.*, 81:962–972, 1959.

[191] J. Milnor. *Morse theory, Ann. Math. Studies 51.* Princeton Univ. Press, 1963.

[192] J. Milnor. *Lectures on the h-cobordism theorem.* Princeton Univ. Press, 1965.

[193] N. Mok. Geometric Archimedean superrigidity in the Hermitian case.

[194] N. Mok. Aspects of Kähler geometry on arithmetic varieties. *Proc.Symp. Pure Math.*, 52:335–396, 1991.

[195] N. Mok, Y.T. Siu, and S.K. Yeung. Geometric superrigidity. *Inv. math.*, 113:57–84, 1993.

[196] J. Morgan. *The Seiberg-Witten equations and applications to the topology of smooth four manifolds.* Princeton University Press, 1996.

[197] J. Morgan, Z. Szabó, and C. Taubes. A product formula for the Seiberg-Witten invariants and the generalized Thom conjecture. *J. Diff. Geom.*, 44:706–788, 1996.

[198] J. Morgan and G. Tian. *Ricci Flow and the Poincare Conjecture.* arXiv:math/0607607v2, 2006.

[199] C. Morrey. The problem of Plateau on a Riemannian manifold. *Ann. Math.*, 49:807–851, 1948.

[200] G. Mostow. *Strong rigidity of locally symmetric spaces.* Ann. Math. Studies 78, Princeton Univ. Press, 1973.

[201] A. Nadel. Multiplier ideal sheaves and Kähler-Einstein metrics of positive scalar curvature. *Ann. Math.*, 132:549–596, 1990.

[202] I.G. Nikolaev. Smoothness of the metric of spaces with curvature that is bilaterally bounded in the sense of A.D. Aleksandrov. *Sib. Math. J.*, 24:247–263, 1983.

[203] I.G. Nikolaev. Bounded curvature closure of the set of compact Riemannian manifolds. *Bull. AMS*, 24:171–177, 1991.

[204] I.G. Nikolaev. *Synthetic methods in Riemannian geometry.* Lecture Notes. Univ. Illinois at Urbana–Champaign, 1992.

[205] J. Nitsche. *Lectures on minimal surfaces, Vol. I.* Cambridge Univ. Press, 1989.

[206] G. Perel'man. The entropy formula for the Ricci flow and its geometric application. arxiv:math.DG/0211159, 2002.

[207] G. Perel'man. Finite extinction time for the solutions to the Ricci flow on certain three-manifolds. arxiv:math.DG/0307245, 2003.

[208] G. Perel'man. Ricci flow with surgery on three-manifolds. arxiv:math.DG/0303109, 2003.

[209] S. Peters. Cheeger's finiteness theorem for diffeomorphism classes of Riemannian manifolds. *J. reine angew. Math.*, 394:77–82, 1984.

[210] S. Peters. Convergence of Riemannian manifolds. *Compos. Math.*, 62:3–16, 1987.

[211] P. Petersen. *Riemannian geometry.* Springer, 1998.

[212] A. Petrunin, X. Rong, and W. Tuschmann. Collapsing vs. positive pinching. *GAFA*, 9:699–735, 1999.

[213] A. Petrunin and W. Tuschmann. Diffeomorphism finiteness, positive pinching, and second homotopy. *GAFA*, 9:736–774, 1999.

[214] D. Quillen. Determinants of Cauchy-Riemann operators over a Riemann surface. *Funct. Anal. Appl.*, 19:31–34, 1985.

[215] Y.G. Reshetnyak. Inextensible mappings in a space of curvature no greater than K. *Sib. Math. J.*, 9:683–689, 1968.

[216] E. Ruh. Almost flat manifolds. *J. Diff. Geom.*, 17:1–14, 1982.

[217] I.K. Sabitov and S.Z. Shefel'. Connections between the order of smoothness of a surface and that of its metric. *Sib. Math. J.*, 17:687–694, 1976.

[218] R. Sachs and H. Wu. *General relativity for mathematicians.* Springer GTM, 1977.

[219] J. Sacks and K. Uhlenbeck. The existence of minimal immersions of 2-spheres. *Ann. Math.*, 113:1–24, 1981.

[220] T. Sakai. *Riemannian geometry.* Amer. Math. Soc., 1995.

[221] D. Salamon. *Spin geometry and Seiberg-Witten invariants.* 1995.

[222] J. Sampson. Applications of harmonic maps to Kähler geometry. *Contemp. Math.*, 49:125–134, 1986.

[223] M. Schlicht. Another proof of Bianchi's identity in arbitrary bundles. *Ann. Global Anal. Geom.*, 13:19–22, 1995.

[224] R. Schoen and K. Uhlenbeck. A regularity theory for harmonic maps. *J. Diff. Geom.*, 17:307–335, 1982.

[225] R. Schoen and K. Uhlenbeck. Boundary regularity and miscellaneous results on harmonic maps. *J. Diff. Geom.*, 18:253–268, 1983. Correction in J. Diff. Geom. 18 (1983), 329.

[226] R. Schoen and S.T. Yau. Existence of incompressible minimal surfaces and the topology of three dimensional manifolds with nonnegative scalar curvature. *Ann. Math.*, 110:127–142, 1979.

[227] R. Schoen and S.T. Yau. The structure of manifolds with positive scalar curvature. *Man. math.*, 28:159–183, 1979.

[228] J.A. Schouten. *Der Riccikalkül*. Springer, 1924.

[229] J.A. Schouten. *Ricci-calculus*. Springer, 1954.

[230] M. Schwarz. *Morse homology*. Birkhäuser, 1993.

[231] M. Schwarz. Equivalences for Morse homology. *Contemp. Math.*, 246:197–216, 1999.

[232] N. Seiberg and E. Witten. Electromagnetic duality, monopol condensation, and confinement in $N = 2$ supersymmetric Yang-Mills theory. *Nucl. Phys.*, 431(B):581–640, 1994.

[233] N. Seiberg and E. Witten. Monopoles, duality and chiral symmetry breaking in $n = 2$ supersymmetric QCD. *Nucl. Phys. B*, 431:581–640, 1994.

[234] Y.T. Siu. The complex analyticity of harmonic maps and the strong rigidity of compact Kähler manifolds. *Ann. Math.*, 112:73–111, 1980.

[235] Y.T. Siu. The existence of Kähler-Einstein metrics on manifolds with positive anticanonical line bundle and suitable finite symmetry group. *Ann.Math.*, 127:585–627, 1988.

[236] Y.T. Siu and S.T. Yau. Compact Kähler manifolds of positive bisectional curvature. *Inv. math.*, 59:189–204, 1980.

[237] S. Smale. On gradient dynamical systems. *Ann. Math.*, 74:199–206, 1961.

[238] E. Spanier. *Algebraic topology*. McGraw Hill, 1966.

[239] K. Steffen. *An introduction to harmonic mappings*. SFB 256, Vorlesungsreihe, Vol. 18, Bonn, 1991.

[240] R. Stoecker and H. Zieschang. *Algebraische Topologie*. Teubner, ^{2}1994.

[241] S. Stolz. Simply connected manifolds of positive scalar curvature. *Ann. Math.*, 136:511–540, 1992.

[242] M. Struwe. On the evolution of harmonic mappings. *Comm. Math. Helv.*, 60:558–581, 1985.

[243] M. Struwe. *Variational methods.* Springer, ^{3}1999.

[244] G. Tarantello. Multiple condensate solutions for the Chern-Simons-Higgs theory. *J. Math. Phys.*, 37:3769–3796, 1996.

[245] C. Taubes. Arbitrary n–vortex solutions to the first order Ginzburg-Landau equations. *Comm. Math. Phys.*, 72:227–292, 1980.

[246] C. Taubes. The Seiberg-Witten invariant and symplectic forms. *Math. Res. Letters*, 1:809–822, 1994.

[247] C. Taubes. More constraints on symplectic manifolds from Seiberg-Witten invariants. *Math. Res. Letters*, 2:9–14, 1995.

[248] C. Taubes. The Seiberg-Witten and the Gromov invariants. *Math. Res. Letters*, 9:809–822, 1995.

[249] C. Taubes. Counting pseudoholomorphic submanifolds in dimension 4. *J. Diff. Geom.*, 44:818 – 893, 1996.

[250] R. Thom. Sur une partition cellules associés à une fonction sur une variété. *C. R. Acad. Sci. Paris*, 228:973–975, 1949.

[251] W. Thurston. Hyperbolic structures on 3-manifolds. Preprint, 1980.

[252] W. Thurston. *Three-dimensional geometry and topology, Vol. 1.* Princeton University Press, 1997.

[253] G. Tian. On Kähler-Einstein metrics on certain Kähler manifolds with $c_1(M) > 0$. *Inv. math.*, 89:225–246, 1987.

[254] G. Tian. Kähler-Einstein metrics with positive scalar curvature. *Invent. Math.*, 130:1–39, 1997.

[255] G. Tian and S.T. Yau. Kähler-Einstein metrics on complex surfaces with $c_1(M) > 0$. *Comm. Math. Phys.*, 112:175–203, 1987.

[256] W. Tuschmann. Collapsing, solvmanifolds, and infrahomogenous spaces. *Diff. Geom. Appl.*, 7:251–264, 1997.

[257] W. Tuschmann. Endlichkeitssätze und positive Krümmung. Habilitation thesis, Leipzig, 2000.

[258] J. Weber. The Morse-Witten complex via dynamical systems. arXiv:math.GT/0411465, 2004.

[259] R. Wells. *Differential analysis on complex manifolds.* Springer, ^{2}1980.

[260] H. Weyl. *Die Idee der Riemannschen Fläche.* Teubner, Leipzig, Berlin, 1913.

[261] H. Weyl. *Space, time, matter.* Dover, 1952. Translated from the German.

[262] H. Weyl. *Das Kontinuum und andere Monographien.* Chelsea, New York, 1973. Reprint.

[263] B. Wilking. Manifolds with positive sectional curvature almost everywhere. *Invent.Math.*, 148:117–141, 2002.

[264] E. Witten. Supersymmetry and Morse theory. *J. Diff. Geom*, 17:661–692, 1982.

[265] E. Witten. Monopoles and 4-manifolds. *Math. Res. Letters*, 1:764–796, 1994.

[266] J. Wolf. *Spaces of constant curvature.* Publish or Perish, Boston, 1974.

[267] H. Wu. The Bochner technique in differential geometry. *Math. Rep.*, 3(2):289–538, 1988.

[268] Y.L. Xin. *Geometry of harmonic maps.* Birkhäuser, 1996.

[269] Y.L. Xin. *Minimal submanifolds and related topics.* World Scientific, 2004.

[270] S.T. Yau. On the fundamental group of compact manifolds of non-positive curvature. *Ann. Math.*, 93:579–585, 1971.

[271] S.T. Yau. On the Ricci curvature of a compact Kähler manifold and the complex Monge-Ampère equation, I. *Comm. Pure Appl. Math.*, 31:339–411, 1978.

[272] S.T. Yau. Open problems in geometry. *Proc. Symp. Pure Math.*, 54(I):1–28, 1993.

[273] W. Ziemer. *Weakly differentiable functions.* Springer, 1989.

[274] R. Zimmer. *Ergodic theory and semisimple groups.* Birkhäuser, 1984.

Index

Universitext

Aguilar, M.; Gitler, S.; Prieto, C.: Algebraic Topology from a Homotopical Viewpoint
Aksoy, A.; Khamsi, M. A.: Methods in Fixed Point Theory
Alevras, D.; Padberg M. W.: Linear Optimization and Extensions
Andersson, M.: Topics in Complex Analysis
Aoki, M.: State Space Modeling of Time Series
Arnold, V. I.: Lectures on Partial Differential Equations
Arnold, V. I.; Cooke, R.: Ordinary Differential Equations
Audin, M.: Geometry
Aupetit, B.: A Primer on Spectral Theory
Bachem, A.; Kern, W.: Linear Programming Duality
Bachmann, G.; Narici, L.; Beckenstein, E.: Fourier and Wavelet Analysis
Badescu, L.: Algebraic Surfaces
Balakrishnan, R.; Ranganathan, K.: A Textbook of Graph Theory
Balser, W.: Formal Power Series and Linear Systems of Meromorphic Ordinary Differential Equations
Bapat, R.B.: Linear Algebra and Linear Models
Benedetti, R.; Petronio, C.: Lectures on Hyperbolic Geometry
Benth, F. E.: Option Theory with Stochastic Analysis
Berberian, S. K.: Fundamentals of Real Analysis
Berger, M.: Geometry I, and II
Bhattacharya, R.; Waymire, E. C.: A Basic Course in Probability Theory
Bliedtner, J.; Hansen, W.: Potential Theory
Blowey, J. F.; Coleman, J. P.; Craig, A. W. (Eds.): Theory and Numerics of Differential Equations
Blowey, J. F.; Craig, A.; Shardlow, T. (Eds.): Frontiers in Numerical Analysis, Durham 2002, and Durham 2004
Blyth, T. S.: Lattices and Ordered Algebraic Structures
Börger, E.; Grädel, E.; Gurevich, Y.: The Classical Decision Problem
Böttcher, A; Silbermann, B.: Introduction to Large Truncated Toeplitz Matrices
Boltyanski, V.; Martini, H.; Soltan, P. S.: Excursions into Combinatorial Geometry
Boltyanskii, V. G.; Efremovich, V. A.: Intuitive Combinatorial Topology
Bonnans, J. F.; Gilbert, J. C.; Lemaréchal, C.; Sagastizábal, C. A.: Numerical Optimization
Booss, B.; Bleecker, D. D.: Topology and Analysis
Borkar, V. S.: Probability Theory
Bridges/Vita: Techniques of Constructive Analysis
Brunt B. van: The Calculus of Variations
Bühlmann, H.; Gisler, A.: A Course in Credibility Theory and its Applications
Carleson, L.; Gamelin, T. W.: Complex Dynamics
Cecil, T. E.: Lie Sphere Geometry: With Applications of Submanifolds
Chae, S. B.: Lebesgue Integration
Chandrasekharan, K.: Classical Fourier Transform
Charlap, L. S.: Bieberbach Groups and Flat Manifolds
Chern, S.: Complex Manifolds without Potential Theory
Chorin, A. J.; Marsden, J. E.: Mathematical Introduction to Fluid Mechanics
Cohn, H.: A Classical Invitation to Algebraic Numbers and Class Fields
Curtis, M. L.: Abstract Linear Algebra
Curtis, M. L.: Matrix Groups
Cyganowski, S.; Kloeden, P.; Ombach, J.: From Elementary Probability to Stochastic Differential Equations with MAPLE
Da Prato, G.: An Introduction to Infinite Dimensional Analysis
Dalen, D. van: Logic and Structure
Das, A.: The Special Theory of Relativity: A Mathematical Exposition

Debarre, O.: Higher-Dimensional Algebraic Geometry

Deitmar, A.: A First Course in Harmonic Analysis. 2^{nd} edition

Demazure, M.: Bifurcations and Catastrophes

Devlin, K. J.: Fundamentals of Contemporary Set Theory

DiBenedetto, E.: Degenerate Parabolic Equations

Diener, F.; Diener, M.(Eds.): Nonstandard Analysis in Practice

Dimca, A.: Sheaves in Topology

Dimca, A.: Singularities and Topology of Hypersurfaces

DoCarmo, M. P.: Differential Forms and Applications

Duistermaat, J. J.; Kolk, J. A. C.: Lie Groups

Dumortier.: Qualitative Theory of Planar Differential Systems

Dundas, B. I.; Levine, M.; Østvaer, P. A.; Röndip, O.; Voevodsky, V.: Motivic Homotopy Theory

Edwards, R. E.: A Formal Background to Higher Mathematics Ia, and Ib

Edwards, R. E.: A Formal Background to Higher Mathematics IIa, and IIb

Emery, M.: Stochastic Calculus in Manifolds

Emmanouil, I.: Idempotent Matrices over Complex Group Algebras

Endler, O.: Valuation Theory

Engel, K.-J.; Nagel, R.: A Short Course on Operator Semigroups

Erez, B.: Galois Modules in Arithmetic

Everest, G.; Ward, T.: Heights of Polynomials and Entropy in Algebraic Dynamics

Farenick, D. R.: Algebras of Linear Transformations

Foulds, L. R.: Graph Theory Applications

Franke, J.; Härdle, W.; Hafner, C. M.: Statistics of Financial Markets: An Introduction

Frauenthal, J. C.: Mathematical Modeling in Epidemiology

Freitag, E.; Busam, R.: Complex Analysis

Friedman, R.: Algebraic Surfaces and Holomorphic Vector Bundles

Fuks, D. B.; Rokhlin, V. A.: Beginner's Course in Topology

Fuhrmann, P. A.: A Polynomial Approach to Linear Algebra

Gallot, S.; Hulin, D.; Lafontaine, J.: Riemannian Geometry

Gardiner, C. F.: A First Course in Group Theory

Gårding, L.; Tambour, T.: Algebra for Computer Science

Godbillon, C.: Dynamical Systems on Surfaces

Godement, R.: Analysis I, and II

Goldblatt, R.: Orthogonality and Spacetime Geometry

Gouvêa, F. Q.: p-Adic Numbers

Gross, M. et al.: Calabi-Yau Manifolds and Related Geometries

Grossman, C.; Roos, H.-G.; Stynes, M.: Numerical Treatment of Partial Differential Equations

Gustafson, K. E.; Rao, D. K. M.: Numerical Range. The Field of Values of Linear Operators and Matrices

Gustafson, S. J.; Sigal, I. M.: Mathematical Concepts of Quantum Mechanics

Hahn, A. J.: Quadratic Algebras, Clifford Algebras, and Arithmetic Witt Groups

Hájek, P.; Havránek, T.: Mechanizing Hypothesis Formation

Heinonen, J.: Lectures on Analysis on Metric Spaces

Hlawka, E.; Schoißengeier, J.; Taschner, R.: Geometric and Analytic Number Theory

Holmgren, R. A.: A First Course in Discrete Dynamical Systems

Howe, R., Tan, E. Ch.: Non-Abelian Harmonic Analysis

Howes, N. R.: Modern Analysis and Topology

Hsieh, P.-F.; Sibuya, Y. (Eds.): Basic Theory of Ordinary Differential Equations

Humi, M., Miller, W.: Second Course in Ordinary Differential Equations for Scientists and Engineers

Hurwitz, A.; Kritikos, N.: Lectures on Number Theory

Huybrechts, D.: Complex Geometry: An Introduction

Isaev, A.: Introduction to Mathematical Methods in Bioinformatics

Istas, J.: Mathematical Modeling for the Life Sciences

Iversen, B.: Cohomology of Sheaves

Jacod, J.; Protter, P.: Probability Essentials

Jennings, G. A.: Modern Geometry with Applications

Jones, A.; Morris, S. A.; Pearson, K. R.: Abstract Algebra and Famous Inpossibilities

Jost, J.: Compact Riemann Surfaces

Jost, J.: Dynamical Systems. Examples of Complex Behaviour

Jost, J.: Postmodern Analysis

Jost, J.: Riemannian Geometry and Geometric Analysis

Kac, V.; Cheung, P.: Quantum Calculus

Kannan, R.; Krueger, C. K.: Advanced Analysis on the Real Line

Kelly, P.; Matthews, G.: The Non-Euclidean Hyperbolic Plane

Kempf, G.: Complex Abelian Varieties and Theta Functions

Kitchens, B. P.: Symbolic Dynamics

Kloeden, P.; Ombach, J.; Cyganowski, S.: From Elementary Probability to Stochastic Differential Equations with MAPLE

Kloeden, P. E.; Platen; E.; Schurz, H.: Numerical Solution of SDE Through Computer Experiments

Koralov, L. B.; Sinai, Ya. G.: Theory of Probability and Random Processes. 2^{nd} edition

Kostrikin, A. I.: Introduction to Algebra

Krasnoselskii, M. A.; Pokrovskii, A. V.: Systems with Hysteresis

Kuo, H.-H.: Introduction to Stochastic Integration

Kurzweil, H.; Stellmacher, B.: The Theory of Finite Groups. An Introduction

Kyprianou, A.E.: Introductory Lectures on Fluctuations of Lévy Processes with Applications

Lang, S.: Introduction to Differentiable Manifolds

Lefebvre, M.: Applied Stochastic Processes

Lorenz, F.: Algebra I: Fields and Galois Theory

Luecking, D. H., Rubel, L. A.: Complex Analysis. A Functional Analysis Approach

Ma, Zhi-Ming; Roeckner, M.: Introduction to the Theory of (non-symmetric) Dirichlet Forms

Mac Lane, S.; Moerdijk, I.: Sheaves in Geometry and Logic

Marcus, D. A.: Number Fields

Martinez, A.: An Introduction to Semiclassical and Microlocal Analysis

Matoušek, J.: Using the Borsuk-Ulam Theorem

Matsuki, K.: Introduction to the Mori Program

Mazzola, G.; Milmeister G.; Weissman J.: Comprehensive Mathematics for Computer Scientists 1

Mazzola, G.; Milmeister G.; Weissman J.: Comprehensive Mathematics for Computer Scientists 2

Mc Carthy, P. J.: Introduction to Arithmetical Functions

McCrimmon, K.: A Taste of Jordan Algebras

Meyer, R. M.: Essential Mathematics for Applied Field

Meyer-Nieberg, P.: Banach Lattices

Mikosch, T.: Non-Life Insurance Mathematics

Mines, R.; Richman, F.; Ruitenburg, W.: A Course in Constructive Algebra

Moise, E. E.: Introductory Problem Courses in Analysis and Topology

Montesinos-Amilibia, J. M.: Classical Tessellations and Three Manifolds

Morris, P.: Introduction to Game Theory

Mortveit, H.; Reidys, C.: An Introduction to Sequential Dynamical Systems

Nicolaescu, L.: An Invitation to Morse Theory

Nikulin, V. V.; Shafarevich, I. R.: Geometries and Groups
Oden, J. J.; Reddy, J. N.: Variational Methods in Theoretical Mechanics
Øksendal, B.: Stochastic Differential Equations
Øksendal, B.; Sulem, A.: Applied Stochastic Control of Jump Diffusions. 2^{nd} edition
Orlik, P.; Welker, V.: Algebraic Combinatorics
Poizat, B.: A Course in Model Theory
Polster, B.: A Geometrical Picture Book
Porter, J. R.; Woods, R. G.: Extensions and Absolutes of Hausdorff Spaces
Procesi, C.: Lie Groups
Radjavi, H.; Rosenthal, P.: Simultaneous Triangularization
Ramsay, A.; Richtmeyer, R. D.: Introduction to Hyperbolic Geometry
Rautenberg, W.: A concise Introduction to Mathematical Logic
Rees, E. G.: Notes on Geometry
Reisel, R. B.: Elementary Theory of Metric Spaces
Rey, W. J. J.: Introduction to Robust and Quasi-Robust Statistical Methods
Ribenboim, P.: Classical Theory of Algebraic Numbers
Rickart, C. E.: Natural Function Algebras
Rotman, J. J.: Galois Theory
Rubel, L. A.: Entire and Meromorphic Functions
Ruiz-Tolosa, J. R.; Castillo E.: From Vectors to Tensors
Runde, V.: A Taste of Topology
Rybakowski, K. P.: The Homotopy Index and Partial Differential Equations
Sagan, H.: Space-Filling Curves
Samelson, H.: Notes on Lie Algebras
Sauvigny, F.: Partial Differential Equations I
Sauvigny, F.: Partial Differential Equations II
Schiff, J. L.: Normal Families
Schirotzek, W.: Nonsmooth Analysis
Sengupta, J. K.: Optimal Decisions under Uncertainty
Séroul, R.: Programming for Mathematicians
Seydel, R.: Tools for Computational Finance
Shafarevich, I. R.: Discourses on Algebra
Shapiro, J. H.: Composition Operators and Classical Function Theory
Simonnet, M.: Measures and Probabilities
Smith, K. E.; Kahanpää, L.; Kekäläinen, P.; Traves, W.: An Invitation to Algebraic Geometry
Smith, K. T.: Power Series from a Computational Point of View
Smorynski: Self-Reference and Modal Logic
Smoryński, C.: Logical Number Theory I. An Introduction
Srivastava: A Course on Mathematical Logic
Stichtenoth, H.: Algebraic Function Fields and Codes
Stillwell, J.: Geometry of Surfaces
Stroock, D. W.: An Introduction to the Theory of Large Deviations
Sunder, V. S.: An Invitation to von Neumann Algebras
Tamme, G.: Introduction to Étale Cohomology
Tondeur, P.: Foliations on Riemannian Manifolds
Toth, G.: Finite Möbius Groups, Minimal Immersions of Spheres, and Moduli
Tu, L. W.: An Introduction to Manifolds
Verhulst, F.: Nonlinear Differential Equations and Dynamical Systems
Weintraub, S. H.: Galois Theory
Wong, M. W.: Weyl Transforms
Xambó-Descamps, S.: Block Error-Correcting Codes
Zaanen, A.C.: Continuity, Integration and Fourier Theory
Zhang, F.: Matrix Theory
Zong, C.: Sphere Packings
Zong, C.: Strange Phenomena in Convex and Discrete Geometry
Zorich, V. A.: Mathematical Analysis I
Zorich, V. A.: Mathematical Analysis II