# 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.

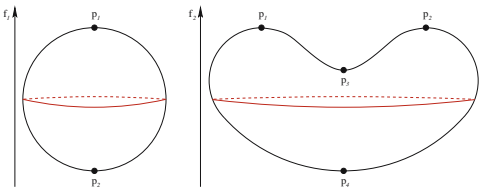


**Fig. 6.1.1.**

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 $d^2f(p)$ 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

$$\begin{aligned} &x : \mathbb{R} \to M \\ &\dot{x}(t) = -\operatorname{grad} f(x(t)) \quad \text{for all } t \in \mathbb{R}. \end{aligned}$$

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 $p$.

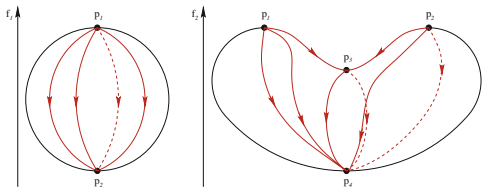


**Fig. 6.1.2.**

In our examples, we have for the Morse indices of the critical points of $f_1$

$$\mu_{f_1}(p_1) = 2\ ,\ \mu_{f_1}(p_2) = 0,$$

and for $f_2$

$$\mu_{f_2}(p_1) = 2\ ,\ \mu_{f_2}(p_2) = 2\ ,\ \mu_{f_2}(p_3) = 1\ ,\ \mu_{f_2}(p_4) = 0\ ,$$

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

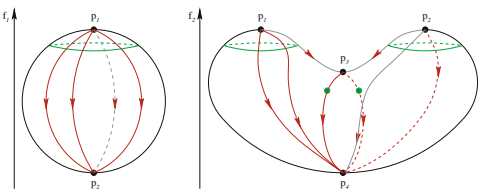


**Fig. 6.1.3.**

pointed sphere ($S^{\mu(p)}$, 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,

$$\chi(X) = \sum_{p \text{ crit. pt. of } f} (-1)^{\mu(p)}\mu(p).$$

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 $\partial$. 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):

$$\partial p = \sum_{\substack{q \text{ crit. pt. of } f \\ \mu(q)=\mu(p)-1}} (\#\{\text{flow lines from } p \text{ to } q\} \bmod 2)q.$$

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 $\partial$ from $C_{k+1}(f, \mathbb{Z}_2)$ is always contained in the kernel of $\partial$ 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.

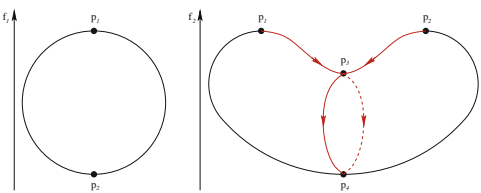


**Fig. 6.1.4.**

For $f_1$, there are no pairs of critical points of index difference 1 at all. Denoting the restriction of $\partial$ to $C_k(f, \mathbb{Z}_2)$ by $\partial_k$, we then have

$$\ker \partial_2 = \{p_1\}$$
$$\ker \partial_0 = \{p_0\},$$

while $\partial_1$ is the trivial operator as $C_1(f_1, \mathbb{Z}_2)$ is 0. All images are likewise trivial, and so

$$H_2(X, f_1, \mathbb{Z}_2) = \mathbb{Z}_2$$
$$H_1(X, f_1, \mathbb{Z}_2) = 0$$
$$H_0(X, f_1, \mathbb{Z}_2) = \mathbb{Z}_2$$

Putting

$$b_k := \dim_{\mathbb{Z}_2} H_k(X, f, \mathbb{Z}_2) \quad \text{(Betti numbers)},$$

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

$$\partial_2 p_1 = \partial_2 p_2 = p_3 \text{ , hence } \partial_2(p_1 + p_2) = 2p_3 = 0$$
$$\partial_1 p_3 = 2p_4 = 0 \text{ (since we are computing mod 2)}$$
$$\partial_0 p_4 = 0$$

Thus

$$H_2(X, f_2, \mathbb{Z}_2) = \ker \partial_2 = \mathbb{Z}_2$$
$$H_1(X, f_2, \mathbb{Z}_2) = \ker \partial_1 / \text{image} \partial_2 = 0$$
$$H_0(X, f_2, \mathbb{Z}_2) = \ker \partial_0 / \text{image} \partial_1 = \mathbb{Z}_2.$$

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.

**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 $0 \le t \le \varepsilon$, for some $\varepsilon > 0$. ($\operatorname{grad} f$ is the gradient of $f$, see (2.1.13); it is the vector field dual to the 1-form $df$.)

*Proof.* Since $x_1$ and $x_2$ are strict relative minima of $f$,
$\exists\, \delta_0 > 0\, \forall\, \delta$ with $0 < \delta \le \delta_0\, \exists\, \varepsilon > 0\ \forall\, x$ with $\|x - x_i\| = \delta$ :

$$f(x) \ge f(x_i) + \varepsilon \quad \text{for } i = 1, 2.$$

Consequently

$$\exists\, \varepsilon_0 > 0\, \forall\, \gamma \in \Gamma\, \exists\, \tau \in (0,1) : f(\gamma(\tau)) \ge \max(f(x_1), f(x_2)) + \varepsilon_0.$$

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

$$\|df(x)\| \ge \alpha \tag{6.2.5}$$

whenever $\kappa - \eta \le f(x) \le \kappa + \eta$.
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 $\max f(\gamma(\tau)) \le \kappa + \eta$ :

$$\begin{aligned} \forall \tau \in [0,1] : \text{either } & f(\gamma(\tau)) \le \kappa - \eta \\ \text{or } & \|df(\gamma(\tau))\| \ge \alpha. \end{aligned} \tag{6.2.7}$$

We let $\varphi(x,t)$ be the solution of (6.2.2) for $0 \le t \le \varepsilon$.
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 $\kappa$. We conclude that there has to exist some $x_3$ with $f(x_3) = \kappa$ and $df(x_3) = 0$. □

The issue of the existence of the negative gradient flow for $f$ will be discussed in the next §. Essentially the same argument as in the proof of Prop. 6.2.1 will be presented once more in Thm. 6.11.3 below.

**Perspectives.** The role of the Palais-Smale condition in the calculus of variations is treated in [142]. A thorough treatment of many further examples can be found in [234] and [39]. 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

$$\phi : X \times \mathbb{R} \to X$$

$$\begin{aligned} \frac{\partial}{\partial t}\phi(x,t) &= -\text{ grad } f(\phi(x,t)) \\ \phi(x,0) &= x. \end{aligned} \tag{6.3.1}$$

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 (2.1.13).

The theorem of Picard-Lindelöf yields the local existence of this flow (see Lemma 1.6.1), 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 $\phi$ 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

$$\phi : X \times \mathbb{R} \to X$$

$$\frac{\partial}{\partial t}\phi = V(\phi(x,t)), \quad \phi(x,0) = x$$

for some vector field $V$ on $X$ which we assume bounded for the present exposition as discussed in 1.6. 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 $\phi$). Therefore, the flow satisfies the group property

$$\phi(x, t_1 + t_2) = \phi(\phi(x, t_1), t_2) \quad \text{for all } t_1, t_2 \in \mathbb{R} \text{ (see Thm. 1.6.1).}$$

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 1.6.1).

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 $\phi$ are defined as

$$W^s(x_0) := \left\{ y \in X : \lim_{t \to +\infty} \phi(y,t) = x_0 \right\}$$
$$W^u(x_0) := \left\{ y \in X : \lim_{t \to -\infty} \phi(y,t) = x_0 \right\}.$$

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

$$\eta(y,t) := tf(y) + \frac{1}{2}(1-t)d^2f(0)(y,y),$$

we have

$$\frac{\partial}{\partial t}\eta(\varphi(y,t),t) = 0, \tag{6.3.5}$$

implying

$$f(\varphi(y,1)) = \eta(\varphi(y,1),1) = \eta(\varphi(y,0),0) = \frac{1}{2}d^2f(0)(y,y)$$

as required. (6.3.5) means

$$\begin{aligned} 0 = f(\varphi(y,t)) + t\,df(\varphi(y,t))\frac{\partial}{\partial t}\varphi(y,t) \qquad (6.3.6)\\ -\frac{1}{2}d^2f(0)(\varphi(y,t),\varphi(y,t)) + (1-t)d^2f(0)(\varphi(y,t),\frac{\partial}{\partial t}\varphi(y,t)). \end{aligned}$$

Now by Taylor expansion, using $df(0) = 0$,

$$\begin{aligned} f(x) &= \int_0^1 (1-\tau)d^2f(\tau x)(x,x)\,d\tau\\ df(x) &= \int_0^1 d^2f(\tau x)x\,d\tau. \end{aligned}$$

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

$$\begin{aligned} T_0(x) &:= -\frac{1}{2}d^2f(0) + \int_0^1 (1-\tau)d^2f(\tau x)\,d\tau\\ T_1(x,t) &:= d^2f(0) + t\int_0^1 (d^2f(\tau x) - d^2f(0))\,d\tau, \end{aligned}$$

(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. $d^2f(0)$ as a linear operator on $B$.

Since we assume that $d^2f(0)$ is nondegenerate, $d^2f(0)$ 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. □

*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

$$f_\lambda := \{y \in X : f(y) \leq \lambda\}$$

as $\lambda$ 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

$$\frac{\partial}{\partial t}\phi(y,t) = -L\phi(y,t)$$
$$\phi(y,0) = y.$$

If $v$ is an eigenvector of $L$ with eigenvalue $\lambda$, 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:

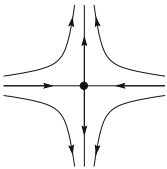

**Fig. 6.3.1.** The horizontal axis is the unstable, the vertical one the stable manifold.

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

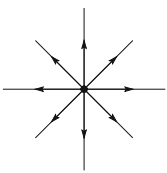

**Fig. 6.3.2.**

If they are different, we may get the following picture, if the one of largest absolute value corresponds to the horizontal direction

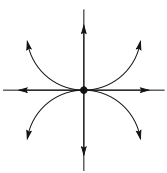

**Fig. 6.3.3.**

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 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 of the linear equation just considered. Namely, we study

$$\begin{aligned} \frac{\partial}{\partial t}\phi(y,t) &= -L\phi(y,t) + \eta(\phi(y,t)) \\ \phi(y,0) &= y, \end{aligned} \tag{6.3.10}$$

in some neighborhood $U$ of 0, where $\eta : H \to H$ satisfies

$$\begin{aligned} \eta(0) &= 0 \\ \|\eta(x) - \eta(y)\| &\leq \delta(\varepsilon)\|x - y\| \end{aligned} \tag{6.3.11}$$

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

$$\begin{aligned} W^u(0,U) = \Big\{ & x \in U : \phi(x,t) \text{ exists and is contained in } U \text{ for all } t \leq 0, \\ & \lim_{t\to-\infty} \phi(x,t) = 0 \Big\} \\ W^s(0,U) = \Big\{ & x \in U : \phi(x,t) \text{ exists and is contained in } U \text{ for all } t \geq 0, \\ & \lim_{t\to+\infty} \phi(x,t) = 0 \Big\}. \end{aligned}$$

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

$$P_\pm : H \to H_\pm, \quad P_+ + P_- = \mathrm{Id}.$$

Since $L$ is bounded, we may find constants $c_0, \gamma > 0$ such that

$$\begin{aligned} \|e^{-Lt}P_+\| &\le c_0 e^{-\gamma t} \qquad \text{for } t \ge 0 \\ \|e^{-Lt}P_-\| &\le c_0 e^{\gamma t} \qquad \text{for } t \le 0. \end{aligned} \tag{6.3.12}$$

Let now $y(t) = \phi(x,t)$ be a solution of (6.3.10) for $t \ge 0$. We have for any $\tau \in [0,\infty)$

$$y(t) = e^{-L(t-\tau)}y(\tau) + \int_\tau^t e^{-L(t-s)}\eta(y(s))\,ds, \tag{6.3.13}$$

hence also

$$P_\pm y(t) = e^{-L(t-\tau)}P_\pm y(\tau) + \int_\tau^t e^{-L(t-s)}P_\pm\eta(y(s))\,ds. \tag{6.3.14}_\pm$$

If we assume that $y(t)$ is bounded for $t \ge 0$, then by (6.3.12)

$$\lim_{\tau\to\infty} e^{-L(t-\tau)}P_- y(\tau) = 0, \tag{6.3.15}$$

and hence such a solution $y(t)$ that is bounded for $t \ge 0$ can be represented as

$$\begin{aligned} y(t) &= P_+y(t) + P_-y(t) \\ &= e^{-Lt}P_+x + \int_0^t e^{-L(t-s)}P_+\eta(y(s))\,ds \\ &\quad - \int_t^\infty e^{-L(t-s)}P_-\eta(y(s))\,ds, \quad \text{with } x = y(0) \end{aligned} \tag{6.3.16}$$

(putting $\tau = 0$ in $(6.3.14)_+$, $\tau = \infty$ in $(6.3.14)_-$). Conversely, any solution of (6.3.16), bounded for $t \ge 0$, satisfies (6.3.13), hence (6.3.10). 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 (6.3.10), with a bounded linear nondegenerate selfadjoint operator $L$ and $\eta$ satisfying (6.3.11). 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 $\eta$ is of class $C^k$ in $U$, so are $W^s(0,U)$ and $W^u(0,U)$.*

*Proof.* We consider, for $x \in P_+H$,

$$T(y,x)(t) := e^{-Lt}x + \int_0^t e^{-L(t-s)} P_+\eta(y(s))\, ds - \int_t^\infty e^{-L(t-s))} P_-\eta(y(s))\, ds. \tag{6.3.17}$$

From (6.3.16) we see that we need to find fixed points of $T$, i.e.

$$y(t) = T(y,x)(t). \tag{6.3.18}$$

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

$$M_\lambda(\varepsilon) := \left\{ y(t) : \|y\|_{\exp,\lambda} := \sup_{t\geq 0} e^{\lambda t}\|y(t)\| \leq \varepsilon \right\}. \tag{6.3.19}$$

$M_\lambda(\varepsilon)$ is a complete normed space. We fix $\lambda$, e.g. $\lambda = \frac{\gamma}{2}$, in the sequel. Because of (6.3.11), (6.3.12), we have for $y \in M_\lambda(\varepsilon)$

$$\begin{aligned}
\|T(y,x)(t)\| &\leq c_0 e^{-\gamma t}\|x\| + c_0\delta(\varepsilon)\left(\int_0^t e^{-\gamma(t-s)}\|y(s)\|\, ds \right.\\
&\qquad \left. + \int_t^\infty e^{\gamma(t-s)}\|y(s)\|\, ds\right)\\
&\leq c_0 e^{-\gamma t}\|x\| + c_0\delta(\varepsilon)\left(\sup_{0\leq s\leq t} e^{\lambda s}\|y(s)\| \int_0^t e^{-\gamma(t-s)}e^{-\lambda s}\, ds \right.\\
&\qquad \left. + \sup_{t\leq s\leq\infty} e^{\lambda s}\|y(s)\| \int_t^\infty e^{\gamma(t-s)}e^{-\lambda s}\, ds\right).
\end{aligned} \tag{6.3.20}$$

Now since

$$\begin{aligned}
\int_0^t e^{-\gamma(t-s)}e^{-\lambda s}\, ds &= e^{-\gamma t}\frac{1}{\gamma-\lambda}\left(e^{(\gamma-\lambda)t} - 1\right) \leq \frac{1}{\gamma-\lambda}e^{-\lambda t},\\
\int_t^\infty e^{\gamma(t-s)}e^{-\lambda s}\, ds &= e^{\gamma t}\frac{1}{\gamma+\lambda}e^{-(\gamma+\lambda)t} = \frac{1}{\gamma+\lambda}e^{-\lambda t},
\end{aligned}$$

(6.3.20) implies

$$\|T(y,x)(t)\| \le c_0 e^{-\gamma t}\|x\| + \frac{2c_0\delta(\varepsilon)}{\gamma-\lambda}e^{-\lambda t}\|y\|_{\exp,\lambda}. \tag{6.3.21}$$

Similarly, for $y_1, y_2 \in M_\lambda(\varepsilon)$

$$\|T(y_1,x)(t) - T(y_2,x)(t)\| \le \frac{4c_0\delta(\varepsilon)}{\gamma-\lambda}e^{-\lambda t}\|y_1 - y_2\|_{\exp,\lambda}. \tag{6.3.22}$$

Because of our assumptions on $\delta(\varepsilon)$ (see (6.3.11), we may choose $\varepsilon$ so small that

$$\frac{4c_0}{\gamma-\lambda}\delta(\varepsilon) \le \frac{1}{2}. \tag{6.3.23}$$

Then from (6.3.22), for $y_1, y_2 \in M_\lambda(\varepsilon)$

$$\|T(y_1,x) - T(y_2,x)\|_{\exp,\lambda} \le \frac{1}{2}\|y_1 - y_2\|_{\exp,\lambda}. \tag{6.3.24}$$

If we assume in addition that

$$\|x\| \le \frac{\varepsilon}{2c_0}, \tag{6.3.25}$$

then for $y \in M_\lambda(\varepsilon)$, by (6.3.21)

$$\|T(y,x)\|_{\exp,\lambda} \le \varepsilon. \tag{6.3.26}$$

Thus, if $\varepsilon$ satisfies (6.3.23), and $\|x\| \le \frac{\varepsilon}{2c_0}$, 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 (6.3.18), for any $x \in P_+H$ with $\|x\| \le \frac{\varepsilon}{2c_0}$.

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 $\|x\| \le \frac{\varepsilon}{2c_0}$)

$$\lim_{t\to\infty} y_x(t) = 0,$$

i.e.

$$y_x(0) \in W^s(0).$$

From (6.3.17), 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 $t \ge 0$. Thus, it also satisfies (6.3.16), 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

$$x = P_+ y_x(0). \tag{6.3.27}$$

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 (6.3.27). 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 (6.3.21), we get assuming (6.3.25),

$$\|y_{x_1}(t) - y_{x_2}(t)\| \leq c_0 e^{-\gamma t}\|x_1 - x_2\| + \frac{1}{2}\|y_{x_1} - y_{x_2}\|_{\exp,\lambda},$$

hence

$$\|y_{x_1}(0) - y_{x_2}(0)\| \leq \|y_{x_1} - y_{x_2}\|_{\exp,\lambda} \leq 2c_0\|x_1 - x_2\|. \tag{6.3.28}$$

We insert the second inequality in (6.328) into the integrals in (6.3.17) and use (6.3.12) as before to get from (6.3.17)

$$\|y_{x_1}(0) - y_{x_2}(0)\| \geq \|x_1 - x_2\| - \frac{4c_0^2\delta(\varepsilon)}{\gamma - \lambda}\|x_1 - x_2\|.$$

If in addition to the above requirement $\frac{1}{\gamma}c_0\delta(\varepsilon) < \frac{1}{4}$ we also impose the condition upon $\varepsilon$ that

$$\frac{4c_0^2\delta(\varepsilon)}{\gamma - \lambda} \leq \frac{1}{2},$$

the above inequality yields

$$\|y_{x_1}(0) - y_{x_2}(0)\| \geq \frac{1}{2}\|x_1 - x_2\|. \tag{6.3.29}$$

Thus, the above map indeed is a bijection between $\{x \in P_+H, \|x\| \leq \frac{\varepsilon}{2c_0}\}$ and its image $W$ in $W^s(0)$. (6.3.28) also shows that our map $x \mapsto y_x(0)$ is Lipschitz, whereas its inverse is Lipschitz by (6.3.29).

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. (6.3.11), (6.3.17) and (6.3.28) (for $x_1 = x, x_2 = 0$, recalling $y_0 = 0$)

$$\begin{aligned}\|P_- y_x(0)\| &= \|\int_0^\infty e^{Ls} P_- \eta(y_x(s))\, ds\| \\ &\le c_0 \int_0^\infty e^{-\gamma s}\delta(\|y_x(s)\|)\|y_x(s)\|\, ds \\ &\le c_0 \int_0^\infty e^{-\gamma s}\delta\left(2c_0 e^{-\lambda s}\|x\|\right) 2c_0 e^{-\lambda s}\|x\| \\ &\le \frac{2c_0^2}{\gamma-\lambda}\delta(2c_0\|x\|)\|x\|.\end{aligned}$$

This implies

$$\frac{\|P_- y_x(0)\|}{\|P_+ y_x(0)\|} = \frac{\|P_- y_x(0)\|}{\|x\|} \to 0 \text{ as } y_x(0) \to 0 \text{ in } W^s(0,U),$$
$$\text{or equivalently } x \to 0 \text{ in } P_+H.$$

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 (6.3.17) depends smoothly on $\eta$. (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 $W^s(0,U)$. □

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 $\phi$ 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 $x$. □

Of course, the corollary holds more generally for the flows of the type (6.3.10) (if we consider only those flow lines $\phi(\cdot,t)$ that exist for all $t \le 0$

resp. $t \geq 0$). (The stable and unstable sets then are as smooth as $\eta$ 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 (6.3.10) 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 $\eta$ is, for example of class $C^k$ if $\eta$ belongs to that class. If*

$$z \in \Lambda^s(z_u) = \varphi_{z_u}(V),$$

*then*

$$\phi(z, t) = \varphi_{\phi(z_u, t)}(P_+\phi(z, t)) \tag{6.3.30}$$

*and*

$$\|\phi(z, t) - \phi(z_u, t)\| \leq c_1 e^{-\lambda t} \tag{6.3.31}$$

*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.*

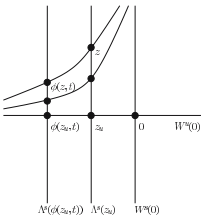


**Fig. 6.3.4.**

**Corollary 6.3.2** *Let $f : X \to \mathbb{R}$ be of class $C^{k+2}$, $k \geq 1$, $x$ 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 Thm. 6.3.2:*

Changing $\eta$ 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 (6.3.11), we may thus assume that the Lipschitz constant of $\eta$ is as small as we like. We apply (6.3.13) to $\phi(z,t)$ and $\phi(z_u,t)$ and get for $\tau \geq 0$, putting $y(t;z,z_u) := \phi(z,t) - \phi(z_u,t)$,

$$y(t;z,z_u) = e^{-L(t-\tau)}y(\tau;z,z_u) + \int_\tau^t e^{-L(t-s)}(\eta(\phi(z,s)) - \eta(\phi(z_u,s)))\,ds. \tag{6.3.32}$$

If this is bounded for $t \to \infty$, then (6.3.12) implies, as in (6.3.15),

$$\lim_{\tau\to\infty} e^{-L(t-\tau)}P_-y(t;z,z_u) = 0. \tag{6.3.33}$$

Consequently, as in (6.3.16) we get,

$$y(t;z,z_u) = e^{-Lt}P_+y(0;z,z_u) + \int_0^t e^{-L(t-s)}P_+(\eta(\phi(z_u,s)+y(s;z,z_u)) - \eta(\phi(z_u,s)))\,ds \tag{6.3.34}$$

$$-\int_t^\infty e^{-L(t-s)}P_-(\eta(\phi(z_u,s)+y(s;z,z_u))-\eta(\phi(z_u,s)))\,ds.$$

As in the proof of Thm. 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

$$\begin{aligned} T(y,z_u,z_+) := {} & e^{-Lt}z_+ \\ & + \int_0^t e^{-L(t-s)}P_+(\eta(\phi(z_u,s)+y(s))-\eta(\phi(z_u,s)))\,ds \\ & - \int_t^\infty e^{-L(t-s)}P_-(\eta(\phi(z_u,s)+y(s))-\eta(\phi(z_u,s)))\,ds, \end{aligned} \tag{6.3.35}$$

for $z_+ \in P_+H$. As in the proof of Thm. 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 Thm. 6.3.1, we get for $0<\lambda<\gamma$ (say $\lambda=\frac{\gamma}{2}$), with $c_0,\gamma$ as in (6.3.12), $\|z_+\|\leq\varepsilon_1$, $y\in M_\lambda(\varepsilon_0)$, i.e. $\|y(t)\|\leq e^{-\lambda t}\varepsilon_0$, and with $[\eta]_{Lip}$ being the Lipschitz constant of $\eta$

$$\|T(y,z_u,z_+)(t)\| \leq c_0\varepsilon_1 e^{-\gamma t} + \frac{2c_0\varepsilon_0}{\gamma-\lambda}[\eta]_{Lip}e^{-\lambda t} \tag{6.3.36}$$

and

$$\|T(y_1,z_u,z_+)(t)-T(y_2,z_u,z_+)(t)\| \leq \frac{4c_0[\eta]_{Lip}}{\gamma-\lambda}e^{-\lambda t}\|y_1-y_2\|_{\exp,\lambda}. \tag{6.3.37}$$

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 (6.3.36) that $T(\cdot, z_u, z_+)$ maps $M_\lambda(\varepsilon_0)$ into itself, and from (6.3.37) 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

$$\begin{aligned} \varphi_{z_u}(z_1) &:= y_{z_u,z_+} \\ z &= y_{z_u,z_+}(0). \end{aligned} \tag{6.3.38}$$

We then have all the required relations:

$$P_+ z = P_+ y_{z_u,z_1}(0) = z_1 \quad \text{from (6.3.35)},$$

and hence $y_{z_u,z_+}$ solves (6.3.34), i.e. is of the form $y(t; z, z_u)$ with $z$ from (6.3.38), and $\phi(z,t) = y(t; z, z_u) + \phi(z_u, t)$ is a flow line. Condition (6.3.30) thus holds at $t = 0$. Since the construction is equivariant w.r.t. time shifts, because of the group property

$$\phi(z, t+\tau) = \phi(\phi(z,t), \tau) \quad \text{for all } t, \tau,$$

(6.3.30) 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 Thm. 6.3.1 that a smoothness property of $\eta$ 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 (6.3.35)) 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. □

**Perspectives.** The theory of stable and unstable manifolds for a dynamical system is classical. Our presentation is based on the one in [49], 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 the 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

$$\begin{aligned} \dot{x}(t) &= -\operatorname{grad} f(x(t)) \quad && \text{for } t \in \mathbb{R} \\ x(0) &= x && \text{for } x \in X. \end{aligned} \tag{6.4.1}$$

We assume that the norms of the first and second derivative of $f$ are bounded. Applying the Picard-Lindelöf theorem (see § 1.6), we then infer that our flow is indeed defined for all $t \in \mathbb{R}$. Also, differentiating (6.4.1), we get

$$\begin{aligned} \ddot{x}(t) (= \nabla_{\frac{d}{dt}} \dot{x}(t)) &= -(\nabla_{\frac{\partial}{\partial x}} \operatorname{grad} f(x(t))) \, \dot{x}(t) \\ &= (\nabla_{\frac{\partial}{\partial x}} \operatorname{grad} f(x(t))) \operatorname{grad} f(x(t)). \end{aligned}$$

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 § 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 $f$ =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

$$\begin{aligned}\langle V, \dot{x}(t)\rangle &= -\langle V,\ \text{grad } f(x(t))\rangle \\ &= -V(f)(x(t)) \qquad \text{by the definition of grad } f, \text{ see (2.1.14)} \\ &= 0 \qquad\qquad\quad \text{since } V \text{ is tangent to a hypersurface on which} \\ &\qquad\qquad\quad\ f \text{ is constant.}\end{aligned}$$

□

We compute

$$\begin{aligned}\frac{d}{dt}f(x(t)) = df(x(t))\dot{x}(t) &= \langle\ \text{grad } f(x(t)), \dot{x}(t)\rangle \quad \text{by (2.1.14)} \\ &= -\|\dot{x}(t)\|^2. \qquad (6.4.2)\end{aligned}$$

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 $t_1, t_2 \in \mathbb{R}$

$$\begin{aligned}f(x(t_1)) - f(x(t_2)) = -\int_{t_1}^{t_2} \frac{d}{dt}f(x(t))\,dt &= \int_{t_1}^{t_2} \|\dot{x}(t)\|^2 \\ &= \int_{t_1}^{t_2} \|\ \text{grad } f(x(t))\|^2\,dt. \qquad (6.4.3)\end{aligned}$$

We also have the estimate

$$
\begin{aligned}
d(x(t_1), x(t_2)) \leq \int_{t_1}^{t_2} \|\dot{x}(t)\|\, dt &\leq (t_2 - t_1)^{\frac{1}{2}} \left( \int_{t_1}^{t_2} \|\dot{x}(t)\|^2\, dt \right)^{\frac{1}{2}} \\
&\quad \text{by Hölder's inequality} \\
&= (t_2 - t_1)^{\frac{1}{2}} (f(x(t_1)) - f(x(t_2)))^{\frac{1}{2}} \\
&\quad \text{by (6.4.3).}
\end{aligned} \tag{6.4.4}
$$

**Lemma 6.4.4** *For any flow line, we have for $t \to \pm\infty$ that 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 $0 \leq t \leq \infty$

$$f_0 := f(x(0)) \geq f(x(t)) \geq f_\infty,$$

and (6.4.3) implies

$$\int_0^\infty \|\dot{x}(t)\|^2 := f_0 - f_\infty < \infty. \tag{6.4.5}$$

Since $\dot{x}(t) = -$ grad $f(x(t))$ is uniformly Lipschitz continuous by Lemma 6.4.1, (6.4.5) implies that

$$\lim_{t\to\infty} \text{grad } f(x(t)) = \lim_{t\to\infty} \dot{x}(t) = 0.$$

□

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 $\phi$ 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 1.6.1, 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, $\dot{x_n}(t)$ are uniformly bounded, and so $x_n$ has to converge in $C^1$. □

*We now assume for the remainder of this § 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 Thm. 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, 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 Thm. 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 Thm. 6.3.1, the convergence is exponential. □

*Remark.* Without assuming that the critical point $x(-\infty)$ is nondegenerate, we still may use (PS) (see Lemma 6.4.8 below) and grad $f(x(t)) \to 0$ for $t \to -\infty$ to see that there exists $t_0 \in \mathbb{R}$ for which $U := \{x(t) : t \le t_0\}$ is precompact and in particular bounded. By Taylor expansion, we have in $U$

$$\| \text{ grad } f(x)\| \le \| \text{ grad } f(x_{-\infty})\| + cd(x, x_{-\infty}) = cd(x, x_{-\infty})$$
$$\text{for some constant } c, \text{ as } \text{ grad } f(x_{-\infty}) = 0.$$

Thus, for $t \le t_n$

$$d(x(t), x_{-\infty}) \le \int_{-\infty}^{t} \|\dot{x}(s)\|\, ds \le c \int_{-\infty}^{t} d(x(s), x_{-\infty})\, ds.$$

The latter integral may be infinite. As soon as it is finite, however, we already get

$$d(x(t), x_{-\infty}) \le c_1 e^{ct} \quad \text{for some constant } c_1,$$

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* $\| \text{ grad } f(x(t))\| \ge \varepsilon$, *for* $t_1 \le t \le t_2$. *Then*

$$d(x(t_1), x(t_2)) \le \frac{1}{\varepsilon}(f(x(t_1)) - f(x(t_2)).$$

*Proof.*

$$\begin{aligned} d(x(t_1), x(t_2)) &\le \int_{t_1}^{t_2} \|\dot{x}(t)\|\, dt \\ &\le \frac{1}{\varepsilon} \int_{t_1}^{t_2} \|\dot{x}(t)\|^2\, dt, \quad \text{since } \|\dot{x}(t)\| = \| \text{ grad } f(x(t))\| \ge \varepsilon \\ &= \frac{1}{\varepsilon}(f(x(t_1)) - f(x(t_2)) \quad \text{by (6.4.3).} \end{aligned}$$

□

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 Thm. 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*

$x_n(-\infty) = p$, $x_n(\infty) = q$.

*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)

$$\| \text{ grad } f(x_n(t_0))\| \to 0,$$

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

$$\| \text{ grad } f(x_n(t_0))\| \geq \varepsilon \quad \text{for all } n \text{ and some } \varepsilon > 0.$$

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

$$\| \text{ grad } f(x_n(t_n))\| = \varepsilon$$
$$\text{and } \| \text{ grad } f(x_n(t))\| \geq \varepsilon \quad \text{for } t_n \leq t \leq t_0.$$

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 Thm. 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 $\infty$ or $-\infty$, say $-\infty$, with

$$\| \text{ grad } f(x_n(t_n))\| \ge \varepsilon \quad \text{for some } \varepsilon > 0. \tag{6.4.6}$$

From (6.4.4), we get the uniform estimate

$$\| \text{ grad } f(x_n(t_1)) - \text{ grad } f(x_n(t_2))\| \le c(t_2 - t_1)^{\frac{1}{2}} \quad \text{for some constant } c. \tag{6.4.7}$$

By (6.4.6), (6.4.7), we may find $\delta > 0$ such that for $t_n - \delta \le t \le t_n$

$$\| \text{ grad } f(x_n(t))\| \ge \frac{\varepsilon}{2},$$

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* $t_{n,i} \in \mathbb{R}$ $(i = 1, \dots, k-1,\ n \in \mathbb{N})$ *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* $y_1 \# y_2 \# \dots \# y_{k-1}$.

*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 Thm. 6.3.1 (see Lemma 6.4.6). □

## 6.5 The Morse-Smale-Floer Condition: Transversality and $\mathbb{Z}_2$-Cohomology

In this §, 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 Thm. 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 Cor. 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 § 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 Def. 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 one.*

*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 Thm. 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 Thm. 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 Thm. 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 $\| \text{ grad } f(x)\| = \varepsilon$, because otherwise we would have a sequence of flow lines $(s_i)_{i\in\mathbb{N}}$ from $p$ to $q$ with $\sup_{x\in s_i} \| \text{ grad } f(x)\| \to 0$ for $i \to \infty$. By (PS) a subsequence would converge to a flow line $s$ (see Lemma 6.4.5) with grad $f(x) \equiv 0$ on $s$. $s$ would thus be constant, in contradiction to Thm. 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 $\| \text{ grad } f(x_i)\| = \varepsilon$, 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 Thm. 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 Thm. 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 $\mathcal{M}^f_{p,q}$.*

*Note.* Let $s_1' \# s_2'$ and $s_1'' \# s_2''$ be broken trajectories contained in the boundary of the same component of $\mathcal{M}_{p,q}^f$. 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.* If a component $\mathcal{M}$ of $\mathcal{M}_{p,q}^f$ 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, Thm. 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 Thm. 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 Thm. 6.3.2, we have the following situation near $p'$: $\mathcal{M}_{p,q}^f$ is a smooth surface containing $s_1$ in its interior. $\mathcal{M}_{p,q}^f$ 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 $\mathcal{M}_{p,q}^f$ 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 $\mathcal{M}_{p,q}^f$. Therefore, in order to compactify $\mathcal{M}_{p,q}^f$ in $W^u(p')$, a single flow line $s_2$ does not suffice.

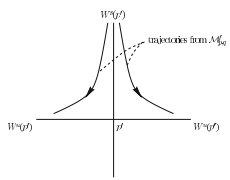


**Fig. 6.5.1.**

Finally, if a broken trajectory through some $p'$ would be a 2−sided limit of $\mathcal{M}_{pq,}^f$, 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

$$\partial p := \sum_{\substack{r \in C_*(f) \\ \mu(p,r)=1}} \left(\#_{\mathbb{Z}_2} \mathcal{M}^f_{p,r}\right) r$$

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 Thm. 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

$$H_k(X, f, \mathbb{Z}_2) := \frac{\ker \partial_k}{\text{image } \partial_{k+1}},$$

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

$$C^k(f, \mathbb{Z}_2) := \text{Hom } (C_k(f, \mathbb{Z}_2), \mathbb{Z}_2)$$

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

$$C_k(f, \mathbb{Z}_2) \to C^k(f, \mathbb{Z}_2)$$
$$p_{j,k} \mapsto p_j^k \text{ with } p_j^k(p_{i,k}) = \delta_{ij} \ (= 1 \text{ for } i = j \text{ and } 0 \text{ otherwise})$$

and

$$\delta^k p_j^k = \sum_{q_{i,k+1} \text{ critical point of } f \text{ of index} k+1} p_{j,k}(\partial q_{i,k+1}) q_i^{k+1}$$

provided that sum is finite, too. Of course, this cohomology theory and the coboundary operator $\delta$ 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

$$y : \mathbb{R} \to X$$
$$\dot{y}(t) = \text{grad } f(y(t)) \text{ for all } t$$

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

$$\dot{x}(t) = -\text{grad } f(x(t))$$

might be well controlled, simply because $f$ is decreasing on such a flow line, and therefore bounded, while along the positive gradient flow

$$\dot{y}(t) = \text{grad } f(y(t)),$$

$f$ 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 §, 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 $\partial$ as in Def. 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 $\pm 1$. 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 $d^2f(p)$, counted with multiplicity. The corresponding eigenvectors span the tangent space $V_p^u \subset T_pX$ 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 $T_yf^{-1}(a)$ 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 $\partial$, 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 §, 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 Thm. 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 Thm. 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 $\mathcal{M}$.

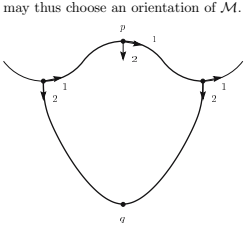


**Fig. 6.6.1.**

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 $-\mathrm{grad}\, f$. 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 $-\mathrm{grad}\, f$, 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)), non-degeneracy of critical points, Morse-Smale-Floer conditions continue to hold, and that the flow is orientable in the sense of Def. 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 Def. 6.5.4.*

*Proof.* We have

$$\begin{aligned}\partial\circ\partial p &= \sum_{\substack{q\in C_*(f)\\ \mu(r,q)=1\\ s_2\in\mathcal{M}^f_{r,q}}}\ \sum_{\substack{r\in C_*(f)\\ \mu(p,r)=1\\ s_1\in\mathcal{M}^f_{r,p}}} n(s_2)n(s_1)q\\ &= \sum_{\substack{q\in C_*(f)\\ \mu(p,q)=2\\ (s_1,s_2)\text{ broken trajectory fromp to }q}} n(s_2)n(s_1)q.\end{aligned}$$

By Thm. 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 Fig. 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. □

In the situation of Thm. 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 §, 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 $\mu$, and a critical point $q$ of $f^2$ of the same Morse index, the unstable manifold of $p$ has dimension $\mu$, 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 $\partial^{f^1}$, $\partial^{f^2}$ 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 $\mu$ the strategy then is to consider the number of flow lines $s(t)$ of (6.7.5) with

$$s(-\infty) = p, \quad s(\infty) = q,$$

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

$$F(x,t) := \begin{cases} f^1(x) & \text{for} \quad -\infty < t \le -1 \\ -t f^1(t) & \text{for} \quad -1 \le t \le 1 \\ -f^1(x) & \text{for} \quad 1 \le t < \infty \end{cases} \tag{6.7.7}$$

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 $\mu^{f^1}$ 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 (6.7.3) 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 $\frac{\partial}{\partial t} + d^2 F(x_0, t)$.

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 $t \le -1$, $\varphi(0) = 0$, we consider the flow

$$\begin{aligned} &\dot{x}(t) = -\varphi(t) \operatorname{grad} f^1(x(t)) \quad \text{for } -\infty < t \le 0, \\ &x(0) = x. \end{aligned}$$

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 $\varphi$ to $(0, \infty)$ by putting

$$\varphi(t) := \varphi(-t) \quad \text{for } t \ge 0,$$

and if we have another Morse function $f^2$ and put

$$\dot{x}(t) = -\varphi(t)\text{grad}\, f^2(x(t)) \quad \text{for } t \geq 0,$$

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 §§, 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 $\dot{x}(t) = -\text{grad}\, f(x(t))$ 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 Thm. 6.5.1. When it come to orientations, however, there is an important difference. Namely, in the situation of Fig. 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 $\mathbb{Z}$.

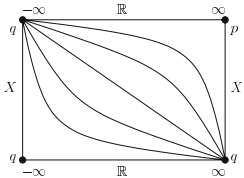


**Fig. 6.7.1.**

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 §, we shall assume that $X$ is a compact, oriented Riemannian manifold. A slight variant of the construction of the preceding § would be the following: Let $f_1, f_2$ be two Morse-Smale-Floer functions, as before. In the preceding §, 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

$$\dot{x}(t) = -\text{grad}\, f_i(x(t)), \quad \text{with } i = 1 \text{ for } t < 0,\ i = 2 \text{ for } t > 2.$$

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 § to equate the homology groups generated by the critical points of $f_1$ and $f_2$ resp.

This construction admits an important generalization:
Let $\Gamma$ 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 $\Gamma$, there is associated a Morse-Smale-Floer function $f_i$ and that the generalized Morse-Smale-Floer condition holds for this collection $f_1, \dots, f_n$.

**Definition 6.8.1** A continuous map $x : \Gamma \to X$ is called a solution of the graph flow for the collection $(f_1, \dots, f_n)$ 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 $\Gamma$ as the flow is automatically smooth in the interior of each edge. If $p_1, \dots, p_{n_1}$ are critical points for the functions $f_1, \dots, f_{n_1}$ resp. corresponding to the edges $e_1, \dots, e_{n_1}$ parametrized on $(-\infty, 0]$, $p_{n_1+1}, \dots, p_{n_1+n_2}$ critical points corresponding to the edges $e_{n_1+1}, \dots, e_{n_1+n_2}$ resp. parametrized on $[0, \infty)$, we let $\mathcal{M}^{\Gamma}_{p_1,\dots,p_{n_1+n_2}}$ be the space of all solutions of (6.8.1) with

$$\lim_{\substack{t \to -\infty \\ t \in e_i}} x(t) = p_i \quad \text{for } i = 1, \dots, n_1$$

$$\lim_{\substack{t \to \infty \\ t \in e_i}} x(t) = p_i \quad \text{for } i = n_1 + 1, \dots, n_1 + n_2,$$

i.e. we assume that on each edge $e_i$, $i = 1, \dots, n_1 + n_2$, $x(t)$ 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 §, the generalized Morse-Smale-Floer condition. Then $\mathcal{M}^{\Gamma}_{p_1,\dots,p_{n_1+n_2}}$ is a smooth manifold, for all tuples $(p_1, \dots, p_{n_1+n_2})$, where $p_i$ is a critical point of $f_i$, with*

$$\dim \mathcal{M}^{\Gamma}_{p_1,\dots,p_{n_1+n_2}} = \sum_{i=1}^{n_1} \mu(p_i) - \sum_{j=n_1+1}^{n_1+n_2} \mu(p_j) - d(n_1 - 1) - d \dim H_1(\Gamma, \mathbb{R}), \tag{6.8.2}$$

*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 $p_i, i = n_1 + 1, \dots, n_1 + n_2$ 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 $\Gamma$, this reduces the dimension by $d$, as the following argument shows:
Let $\Gamma$ 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 $\dot{x}_i(t) = -\text{grad}\, f_i(x_i(t))$, $x_i(0) = x$. 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. □

Again $\mathcal{M}^{\Gamma}_{p_1,\dots,p_{n_1+n_2}}$ is not compact, but can be compactified by flows with broken trajectories on the noncompact edges of $\Gamma$.

The most useful case of Thm. 6.8.1 is the one where the dimension of $\mathcal{M}^{\Gamma}_{p_1,\dots,p_{n_1+n_2}}$ is 0. In that case, $\mathcal{M}^{\Gamma}_{p_1,\dots,p_{n_1+n_2}}$ 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

$$n(\Gamma; p_1, \dots p_{n_1+n_2}).$$

We then define a map

$$q(\Gamma) : \overset{n_1}{\underset{i=1}{\otimes}} C_*(f_i, \mathbb{Z}) \to \overset{n_1+n_2}{\underset{j=n_1+1}{\otimes}} C_*(f_j, \mathbb{Z})$$
$$(p_1 \otimes \dots \otimes p_{n_1}) \mapsto n(\Gamma; p_1, \dots, p_{n_1+n_2})(p_{n_1+1} \otimes \dots \otimes p_{n_1+n_2}).$$

With

$$C^*(f_i, \mathbb{Z}) := Hom(C_*(f_i, \mathbb{Z}), \mathbb{Z}),$$

we may consider $q(\Gamma)$ as an element of

$$\overset{n_1}{\underset{i=1}{\otimes}} C^*(f_i, \mathbb{Z}) \overset{n_1+n_2}{\underset{j=n_1+1}{\otimes}} C_*(f_j, \mathbb{Z}).$$

With the methods of the previous §, one verifies

**Lemma 6.8.1** $\partial q = 0$.

Consequently, we consider $q(\Gamma)$ also as an element of

$$\overset{n_1}{\underset{i=1}{\otimes}} H^*(f_i, \mathbb{Z}) \overset{n_1+n_2}{\underset{j=n_1+1}{\otimes}} H_*(f_j; \mathbb{Z}).$$

Besides the above example where $\Gamma$ 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) $\Gamma$ 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

$$\begin{aligned} q(\Gamma) &\in H^k(X, \mathbb{Z}) \otimes H^{d-k}(X, \mathbb{Z}) \\ &\cong \text{Hom } (H_k(X; \mathbb{Z}); H^{d-k}(X, \mathbb{Z}) \end{aligned}$$

is the so-called Poincaré duality isomorphism.

3) $\Gamma$ 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,

$$\begin{aligned} q(\Gamma) &\in \bigotimes_{j \le k} H^k(K, \mathbb{Z}) \otimes H_j(X, \mathbb{Z}) \otimes H_{k-j}(X, \mathbb{Z}) \\ &\sim \bigotimes_{j \le k} \text{Hom } (H^j(X, \mathbb{Z}) \otimes H^{k-j}(X, \mathbb{Z}), H^k(X, \mathbb{Z}). \end{aligned}$$

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) $\Gamma$ 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

$$\begin{aligned}\nabla_{\frac{d}{ds}} \Big( (\exp_x(t) s\varphi(t))^\bullet + \operatorname{grad} f(\exp_{x(t)} s\varphi(t)) \Big)_{|s=0} \\ = \nabla_{\frac{d}{dt}} \varphi(t) + D_{\varphi(t)} \operatorname{grad} f(x(t)) \text{ with } \nabla_{\frac{d}{dt}} := \nabla_{\dot{x}(t)},\end{aligned}$$

$\nabla$ 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 $x^*\mathrm{End}TX$ 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. J.Jost, X. Li-Jost, Calculus of variations, Cambridge University Press, 1998.

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

$$\|A(t)\| \leq const.,$$

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*

$$\text{ind } L_A = \mu(A(-\infty)) - \mu(A(\infty)).$$

*Proof.* We may find a continuous map $C : \mathbb{R} \to \text{Gl}(n, \mathbb{R})$ and continuous functions $\lambda_1(t), \ldots, \lambda_n(t))$ such that

$$C(t)^{-1} A(t) C(t) = \text{diag } (\lambda_1(t), \ldots, \lambda_n(t)), \quad \lambda_1(t) \leq \lambda_2(t) \leq \ldots \leq \lambda_n(t),$$

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

$$\begin{aligned}\text{ind } L_A &= \dim\ker L_A - \dim\text{coker } L_A\\ &= \dim\ker L_A - \dim\ker L_{-A}\\ &= \mu(A(-\infty)) - \mu(A(\infty)).\end{aligned}$$

□

**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 $\mathrm{End}E_i$, $i = 1, 2$, 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 $\sigma_i(t) = t$ 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*

$$\text{ind } L_A = \text{ind } L_{A_1} + \text{ind } L_{A_2}.$$

*Proof.*

$$\begin{aligned}\text{ind } L_{A_1} + \text{ind } L_{A_2} &= \mu(A_1(-\infty)) - \mu(A_1(\infty)) + \mu(A_2(-\infty)) - \mu(A_2(\infty))\\ &= \mu(A(-\infty)) - \mu(A(\infty))\\ &= \text{ind } L_A, \text{ by Lemma 6.9.2 and construction.}\end{aligned}$$

□

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

$$\text{Det } V := \Lambda^m(V), \quad \text{with } \Lambda^0 V := \mathbb{R}.$$

Then $(\text{Det } V)^* \otimes \text{Det } V$ is canonically isomorphic to $\mathbb{R}$ via $v^* \otimes w \mapsto v^*(w)$. A linear map

$$l : V \to W$$

then induces

$$\det l : \text{Det } V \to \text{Det } W,$$

i.e.

$$\det l \in (\text{Det } V)^* \otimes \text{Det } W.$$

The transformation behavior w.r.t. bases $e_1, \ldots, e_m$ of $V$, $f_1, \ldots, f_m$ of $W$ is given by

$$\det l(e_1 \wedge \ldots \wedge e_m) = le_1 \wedge \ldots \wedge le_m =: \Delta_l f_1 \wedge \ldots \wedge f_m.$$

We may e.g. use the inner product on $W$ to identify the orthogonal complement of $l(V)$ with coker $l$. The exact sequence

$$0 \to \ker l \to V \xrightarrow{l} W \to \text{coker } l \to 0$$

and the multiplicative properties of det allow the identification

$$(\text{Det } V)^* \otimes \text{Det } W \cong (\text{Det } \ker l)^* \otimes \text{Det } (\text{coker } l) =: \text{Det } l.$$

This works as follows:
Put $V_0 = \ker l$, $W_0 = \text{coker } L \ (= l(V)^\perp)$, 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 $e_1, \ldots, e_k$ is a basis of $V_0$, $e_{k+1}, \ldots e_m$ one of $V_1$, $f_1, \ldots f_k$ one of $W_0$, and if we take the basis $le_{k+1}, \ldots, le_m$ of $W_1$, then

$$(e_1 \wedge \ldots e_k \wedge e_{k+1} \wedge \ldots \wedge e_m)^* \otimes (f_1 \wedge \ldots \wedge f_k \wedge le_{k+1} \wedge \ldots \wedge le_m)$$

is identified with

$$(e_1 \wedge \ldots \wedge e_m)^* \otimes (f_1 \wedge \ldots \wedge f_m).$$

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 $(\text{Det } V)^* \otimes \text{Det } W \cong (\text{Det } \ker l)^* \otimes \text{Det coker } l$ 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 $(\text{Det } \ker l)^* \otimes \text{Det coker } l$ can be canonically identified with $\mathbb{R}$, and $\det l$ with $1 \in \mathbb{R}$, by choosing basis $e_1, \ldots, e_m$ of $V$ and the basis $le_1, \ldots le_m$ of $W$, as above.

In a more abstract manner, this may also be derived from the above exact sequence

$$0 \to \ker l \to V \xrightarrow{l} W \to \text{coker } l \to 0$$

on the basis of the following easy algebraic

**Lemma 6.9.3** *Let* $0 \to V_1 \overset{l_1}{\to} V_2 \overset{l_2}{\to} \dots \overset{l_{k-1}}{\to} V_k \to 0$ *be an exact sequence of linear maps between finite dimensional vector spaces. Then there exists a canonical isomorphism*

$$\bigotimes_{i \text{ odd}} \Lambda^{\max} V_i \overset{\sim}{\to} \bigotimes_{i \text{ even}} \Lambda^{\max} V_i.$$

One simply uses this Lemma plus the above canonical identification $(\text{Det } V)^* \otimes \text{Det } V \cong \mathbb{R}$.

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

$$\text{Det } l_y := (\text{Det } \ker l_y)^* \otimes (\text{Det coker } l_y)$$

for each $y$. We intend to show that these lines $(\text{Det } l_y)_{y \in Y}$ constitute a line bundle over $Y$.

$$\begin{aligned} l_y : (\ker l_y)^\perp &\to (\text{coker } l_y)^\perp \\ v &\mapsto l_y v \end{aligned}$$

is an isomorphism, and

$$\text{ind } l_y = \dim \ker l_y - \dim \text{coker } l_y$$

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

$$W'_y := l_y(V'_y) \oplus \text{coker } l_y.$$

Then as above

$$(\text{Det } V'_y)^* \otimes \text{Det } W'_y \cong (\text{Det } \ker l_y)^* \otimes \text{Det coker } l_y.$$

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

$$(\text{Det } V''_y)^* \otimes \text{Det } W''_y \cong (\text{Det } V'_y)^* \otimes \text{Det } W'_y.$$

Once we have verified that property, we can piece the local models $(\text{Det } V'_y)^* \otimes \text{Det } W'_y$ for $\text{Det } l_y$ unambiguously together to get a line bundle with fiber $\text{Det } l_y$ 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

$$\det l_y : \mathrm{Det}\ \bar{V}_y \to \mathrm{Det}\ \bar{W}_y$$

yields a nonvanishing section $\Delta_{l_y}$ of $(\mathrm{Det}\ \bar{V}_y)^* \otimes \mathrm{Det}\ \bar{W}_y$. We then get the isomorphism

$$\begin{aligned}(\mathrm{Det}\ V'_y)^* \otimes \mathrm{Det}\ W'_y &\to (\mathrm{Det}\ V'_y)^* \otimes \mathrm{Det}\ W'_y \otimes (\mathrm{Det}\ \bar{V}_y)^* \otimes \mathrm{Det}\ \bar{W}_y \\ &\cong (\mathrm{Det}\ V''_y)^* \otimes (\mathrm{Det}\ W''_y) \\ s_y &\mapsto s_y \otimes \Delta_{l_y},\end{aligned}$$

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*

$$\mathrm{Det}\ l_y = (\mathrm{Det}\ \ker l_y)^* \otimes (\mathrm{Det}\ \mathrm{coker}\ l_y)$$

*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 $\mathrm{Det}\ l$ 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 Cor. 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 Cor. 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 $\tau$ sufficiently large. Likewise, $A_2{}^{\tau}(t)$ does not depend on $t$ over $(-\infty,1]$ for $\tau$ 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 $\tau$ 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 Cor. 6.9.1 equals

$$\begin{aligned} \operatorname{ind} L_{A_1} + \operatorname{ind} L_{A_2} &= \operatorname{ind} L_A \\ &\leq \dim \ker L_A, \end{aligned}$$

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 $\tau$ 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

$$s_{\tau_n} \to 0 \quad \text{on } [-T, T], \quad \text{for any } T > 0.$$

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. □

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 $x_1(\pm\infty) = x_2(\pm\infty)$, $A_1(\pm\infty) = A_2(\pm\infty)$. Then the determinant lines* $\det L_{(x_1,A_1)}$ *and* $\det L_{(x_2,A_2)}$ *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 $GL(n, \mathbb{R})$ 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, ind $L_{A_1}$ = ind $L_{A_2}$, and 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, 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

$$\text{Det } L_{(x^{-1},A^{-1})} \otimes \text{Det } L_{(x,A)} \equiv \text{Det } L_{(x_0,A_0)}.$$

Finally, for an arbitrary class $L_{(x,A)}$, we find $(x_1, A_1)$ and $(x_2, A_2)$ with

$$\begin{array}{llll} x_1(-\infty) = x_0, & A_1(-\infty) = x_0, & x_1(\infty) = x(-\infty), & A_1(\infty) = A(-\infty), \\ x_2(\infty) = x_0, & A_2(\infty) = x_0, & x_2(-\infty) = x(\infty), & A_2(-\infty) = A(\infty). \end{array}$$

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.

$$\text{Det } L_{(x_1,A_1)} \otimes \text{Det } L_{(x,A)} \otimes \text{Det } L_{(x_2,A_2)} \cong \text{Det } L_{(x_0,A_0)}$$

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 Thm. 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.

$$\dot{x}(t) + \text{grad } f(x(t)) = 0,$$

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 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 Thm. 6.5.1, based on Thm. 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 Fig. 6.6.1

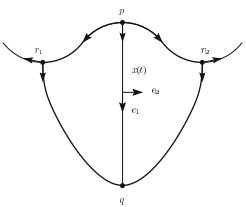


**Fig. 6.9.1.**

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 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 Def. 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, 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 Fig 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.

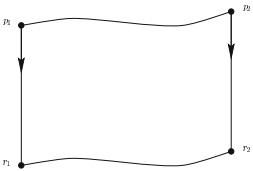


**Fig. 6.9.2.**

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 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 Thm. 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 Fig. 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 Fig. 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 6.9.1** *Under the assumptions of Thm. 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 [74]. It was developed in detail by Schwarz[221], 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 [250]. We also refer the reader to the bibliography of [221] 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[256], 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[75]. This in turn is based on the original work of Quillen[204] 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, 6.9) 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 1.5.1, 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 [221] 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-Steinrod axioms. These axioms are verified for Morse-Floer cohomology in [221]. The background in algebraic topology can be found in [229]. Witten[256] 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 $t = 0$, $\Delta_0$ 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 $\Delta_t$ 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 [182].

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[50], Conley and Zehnder[51]. In another direction, different approaches to Morse theory on singular (stratified) spaces have been developed by Goresky and MacPherson[90] and Ludwig[172].

## 6.10 The Morse Inequalities

The Morse inequalities express relationships between the Morse numbers $\mu_i$, 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 §, 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

$$\dot{x}(t) = -\text{grad}\, f(x(t)) \text{ for } t \in \mathbb{R}$$

and

$$x(t_0) \in A \text{ for some } t_0 \in \mathbb{R} \cup \{-\infty\},$$

then also

$$x(t) \in A \text{ for all } t \geq t_0.$$

We obtain a new boundary operator $\partial^A$ in place of $\partial$ 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

$$\partial^A p := \sum_{\substack{r \in C_*(f)\cap X\setminus A \\ \mu(p,r)=1}} (\#_{\mathbb{Z}_2}\mathcal{M}^f_{p,r})\, r. \tag{6.10.1}$$

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 Thm 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.

$$\dot{x}(t) = -\text{grad } f(x(t)) \text{ for all } t,$$

satisfies

$$x(t_0) \in A \text{ for some } t_0 \in \mathbb{R} \cup \{-\infty\},$$

then there is no $t > t_0$ with $x(t) \in Y \backslash A$.

(ii) If the flow line $x(t)$ satisfies

$$x(t_1) \in Y, x(t_2) \in X \backslash \overset{\circ}{Y}, \text{ with } -\infty \leq t_1 < t_2 \leq \infty,$$

then there exists $t_1 \leq t_0 \leq t_2$ 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 Thm. 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 $T$:

$$Y_p^T := \{x(t) : -\infty \le t \le T, x(\cdot) \text{ flow line in } W^u(p)\} \tag{6.10.8}$$

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 $\infty$, 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 Def. 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 Thm. 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 $\partial$. 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 Thm. 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 $A_j$

$$\ldots A_{i+1} \xrightarrow{f_{i+1}} A_i \xrightarrow{f_i} A_{i-1} \xrightarrow{f_{i-1}} \ldots$$

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 $\pi$ also as an element of $C_k(Y)$, from the inclusion $A \hookrightarrow Y$. If $\pi$ 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, $\gamma$ can be considered as an element of $C_{k+1}(Y)$, and so $\pi$ 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 $\pi$ need not be supported on $A$, but $\partial \pi$ as an element of $H_{k-1}(A)$ does not change if we replace $\pi$ 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 $\pi$ 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** *The sequence*

$$... \, H_k(A) \xrightarrow{i_k} H_k(Y) \xrightarrow{j_k} H_k(Y, A) \xrightarrow{\partial_k} H_{k-1}(A) \longrightarrow \; ...$$

*is exact.*

*Proof.* We denote the homology classes of an element $\gamma$ 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 $A, \pi$ represents an element of $H_{k+1}(Y, A)$, and so $[\gamma] \in$ image $(\partial_{k+1})$. Conversely, for any such $\pi$, $\partial \pi$ 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 $\pi$ 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 $\pi$ 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

$$\begin{array}{ccc} A_2 & \xrightarrow{a} & A_1 \\ f \downarrow & & \downarrow g \\ B_2 & \xrightarrow{b} & B_1 \end{array}$$

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*

$$\begin{array}{ccccccccc} \ldots \longrightarrow H_k(Y_2,Y_3) & \xrightarrow{\partial_k^{2,3}} & H_{k-1}(Y_3) & \xrightarrow{i_k^{2,3}} & H_{k-1}(Y_2) & \xrightarrow{j_k^{2,3}} & H_{k-1}(Y_2,Y_3) & \longrightarrow & \ldots \\ \downarrow i_k^{1,2,3} & & \downarrow i_k^{2,3} & & \downarrow i_k^{1,2} & & \downarrow i_{k-1}^{1,2,3} & & \\ \longrightarrow H_k(Y_1,Y_2) & \xrightarrow{\partial_k^{1,2}} & H_{k-1}(Y_2) & \xrightarrow{i_k^{1,2}} & H_{k-1}(Y_1) & \xrightarrow{j_k^{1,2}} & H_{k-1}(Y_1,Y_2) & \longrightarrow & \ldots, \end{array}$$

*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 $\pi$ 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 $\pi$ 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*

$$\ldots \longrightarrow H_{k+1}(Y_1,Y_2) \xrightarrow{j_{k+1}^{2,3} \circ \partial_{k+1}^{1,2}} H_k(Y_2,Y_3) \xrightarrow{i_k^{1,2}} H_k(Y_1,Y_3) \xrightarrow{j_k^{1,2}} H_k(Y_1,Y_2) \longrightarrow \ldots$$

*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 $\gamma$ as an element of $C_{k+1}(Y_1,Y_2)$ as the class of $\pi$ in $H_k(Y_2,Y_3)$ is not influenced by adding $\partial\omega$ for some $\omega \in C_{k+1}(Y_2)$. Thus $\pi$ 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 $\pi$ 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*
$\ldots \longrightarrow A_3 \xrightarrow{a_3} A_2 \xrightarrow{a_2} A_1 \xrightarrow{a_1} 0$
*be an exact sequence of linear maps between vector spaces.*

*Then for all* $k \in \mathbb{N}$

$$\dim A_1 - \dim A_2 + \dim A_3 - \ldots - (-1)^k \dim A_k + (-1)^k \dim(\ker a_k) = 0. \tag{6.10.14}$$

*Proof.* For any linear map $\ell = V \to W$ between vector spaces,
$\dim V = \dim(\ker \ell) + \dim(\text{image } \ell)$.
Since by exactness

$$\dim(\text{image } a_j) = \dim(\ker a_{j-1})$$

we obtain

$$\dim(A_j) = \dim(\ker a_j) + \dim(\ker a_{j-1}).$$

Since $\dim A_1 = \dim \ker a_1$, we obtain

$$\dim A_1 - \dim A_2 + \dim A_3 - \ldots + (-1)^k \dim(\ker a_k) = 0.$$

□

We now apply Lemma 6.10.4 to the exact sequence of Lemma 6.10.3. With

$$\begin{aligned} b_k(X,Y) &:= \dim(H_k(X,Y)) \\ \nu_k(Y_1,Y_2,Y_3) &= \dim(\ker j_{k+1}^{2,3} \circ \partial_k^{1,2}), \end{aligned}$$

we obtain

$$\begin{aligned} \sum_{i=0}^{k} (-1)^i (b_i(Y_1,Y_2) - b_i(Y_1,Y_3) + b_i(Y_2,Y_3)) \\ - (-1)^k \nu_k(Y_1,Y_2,Y_3) = 0. \end{aligned}$$

Hence

$$\begin{aligned} (-1)^{k-1}\nu_{k-1}(Y_1,Y_2,Y_3) =& (-1)^k \nu_k(Y_1,Y_2,Y_3) - (-1)^k b_k(Y_1,Y_2) \\ &+ (-1)^k b_k(Y_1,Y_3) - (-1)^k b_k(Y_2,Y_3) \end{aligned} \tag{6.10.15}$$

We define the following polynomials in $t$:

$$\begin{aligned} P(t,X,Y) &:= \sum_{k\geq 0} b_k(X,Y) t^k \\ Q(t,Y_1,Y_2,Y_3) &:= \sum_{k\geq 0} \nu_k(Y_1,Y_2,Y_3) t^k \end{aligned}$$

Multiplying the preceding equation by $(-1)^k t^k$ and summing over $k$, we obtain

$$Q(t,Y_1,Y_2,Y_3) = -tQ(t,Y_1,Y_2,Y_3) + P(t,Y_1,Y_2) - P(t,Y_1,Y_3) + P(t,Y_2,Y_3). \tag{6.10.16}$$

We now order the critical points $p_1, ..., p_m$ 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

$$Y_1 := Y_1(i) := \bigcup_{k \geq i} Y(p_k)$$
$$Y_2 := Y_2(i) := \bigcup_{k \geq i+1} Y(p_k)$$
$$Y_3 := \emptyset.$$

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 $\mu_\ell$ be the number of critical points of $f$ of Morse index $\ell$. 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 $(Y_1, Y_2, Y_3)$

$$\begin{aligned} P(t, Y_1(i), Y_2(i)) = {} & P(t, Y_1(i), \emptyset) - P(t, Y_2(i), \emptyset) \\ & + (1+t)(X(t, Y_1(i), Y_2(i), \emptyset), \end{aligned}$$

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

$$\begin{aligned} P(t,X,\phi) &= \sum t^j \dim H_j(X) \qquad (\text{since } H_j(X,\emptyset) = H_j(X)) \\ &= \sum t^j b_j(X) \qquad (\text{see Cor. 6.9.1}) \end{aligned}$$

we conclude

**Theorem 6.10.2** *Let $f$ be a Morse-Smale-Floer function on the compact, finite dimensional orientable Riemannian manifold $X$. Let $\mu_\ell$ be the number of critical points of $f$ of Morse index $\ell$, 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 Thm. 6.10.2*

*(i)* $\mu_k \geq b_k(X)$ *for all* $k$

*(ii)* $\mu_k - \mu_{k-1} + \mu_{k-2} - \ldots \pm \mu_0 \geq b_k(X) - b_{k-1}(X) \ldots \pm b_0(X)$

*(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 $\leq k$. 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

$$b_k(Y_1, Y_3) \leq b_k(Y_1, Y_2) + b_k(Y_2, Y_3). \tag{6.10.22}$$

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 $\partial$. 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 $\partial$ 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 $\mu_\ell$ arise from placing all critical points in different sets and thus gathering only strictly local information while the Betti numbers $b_\ell$ incorporate all the cancellations induced by the boundary operator $\partial$. Thus, the $\mu_\ell$ and the $b_\ell$ only coincide if no cancellations at all take place, as in the example of the function $f_1$ in § 6.1.

**Perspectives.** In this §, we have interpreted the insights of Morse theory, as developed by Thom[242], Smale[228], Milnor[183], Franks[76] 199-215, in the light of Floer's approach. Schwarz[222] 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 $\Lambda_0$ 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

$$(\gamma_1, \gamma_2) := \int_{t\in S^1} \langle D\gamma_1(t), D\gamma_2(t)\rangle dt,$$

where $D\gamma_i(t)$ is the weak first derivative of $\gamma_i$ at $t$, as defined in A.1. This then defines the Riemannian metric of $\Lambda_0$. 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, $\Lambda_0$ assumes the role of the space $X$ in the general context described in the preceding §§.

The Sobolev space $\Lambda_0$ is the natural space on which to define the energy functional

$$E(c) = \frac{1}{2}\int_{S^1} ||Dc(t)||^2 dt$$

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

1) $u_n$ converges uniformly to $u$ $(u_n \rightrightarrows u)$.

2) $E(u_n) \to E(u)$ 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 $S^1 = \bigcup_{\mu=1}^{m} V_\mu$ by open sets such that for sufficiently large $n$

$$u_n(V_\mu), u(V_\mu) \subset U_\mu \text{ for } \mu = 1, \ldots, m.$$

If now $\varphi \in C_0^\infty(V_\mu, \mathbb{R}^n)$ for some $\mu$, then for sufficiently small $|\varepsilon|$

$$f_\mu(u(t) + \varepsilon\varphi(t)) \subset f_\mu(U_\mu) \text{ for all } t \in V_\mu,$$

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 $\varphi$ 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 $f_\mu : U_\mu \to \mathbb{R}^n$ ("$\cdot$" of course denotes a derivative w.r.t. $t \in S^1$.)

$$= \int (g_{ij}(u)\dot{u}^i\dot{\varphi}^j + \frac{1}{2}g_{ij,k}(u)\dot{u}^i\dot{u}^j\varphi^k)dt \quad (\text{using } g_{ij} = g_{ji}) \quad (6.11.1)$$

If $u \in H^{2,2}(S^1, M)$, this is

$$= -\int (g_{ij}(u)\ddot{u}^i\varphi^j + g_{ij,\ell}\dot{u}^\ell\dot{u}^i\varphi^j - \frac{1}{2}g_{ij,k}\dot{u}^i\dot{u}^j\varphi^k)dt \quad (6.11.2)$$

$$= -\int (\ddot{u}^i + \Gamma^i_{k\ell}(u)\dot{u}^k\dot{u}^\ell)g_{ij}(u)\varphi^j dt \quad \text{as in 1.4.}$$

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 (6.11.5) below) so that also the second terms in (6.11.1) and (6.11.2) are integrable.

We may put

$$\|DE(u)\| = \sup\{\, \frac{d}{d\varepsilon}E(u+\varepsilon\varphi)_{|\varepsilon=0} : \qquad (6.11.3)$$
$$\varphi \in H_0^{1,2}(V_\mu, \mathbb{R}^n) \text{ for some } \mu,$$
$$\int g_{ij}(u)\dot\varphi^i\dot\varphi^j dt \le 1\,\}.$$

For second derivatives of $E$, we may either quote the formula of Theorem 4.1.1 or compute directly in local coordinates

$$\begin{aligned}
&\frac{d^2}{d\varepsilon^2}E(u+\varepsilon\varphi)_{|\varepsilon=0}\\
&\quad= \frac{1}{2}\frac{d^2}{d\varepsilon^2}\int g_{ij}(u+\varepsilon\varphi)(\dot u^i+\varepsilon\dot\varphi^i)(\dot u^j+\varepsilon\dot\varphi^j)dt\\
&\quad= \int (g_{ij}(u)\dot\varphi^i\dot\varphi^j + 2g_{ij,k}\dot u^i\dot\varphi^j\varphi^k + g_{ij,k\ell}\dot u^i\dot u^j\varphi^k\varphi^\ell)dt
\end{aligned}$$

which is also bounded for $u$ and $\varphi$ of Sobolev class $H^{1,2}$.

Suppose now that $u \in \Lambda_0$ satisfies

$$DE(u) = 0.$$

This means

$$0 = \int (g_{ij}(u)\dot u^i\dot\varphi^j + \frac{1}{2}g_{ij,k}(u)\dot u^i\dot u^j\varphi^k)dt \text{ for all } \varphi \in H^{1,2}. \qquad (6.11.4)$$

**Lemma 6.11.1** *Any $u \in \Lambda_0$ with $DE(u) = 0$ is a closed geodesic (of class $C^\infty$).*

*Proof.* We have to show that $u$ is smooth. Then (6.11.2) is valid, and Theorem A.1.5 gives

$$\ddot u^i + \Gamma^i_{k\ell}(u)\dot u^k\dot u^\ell = 0 \text{ for } i = 1,\dots,\dim M,$$

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

$$\begin{aligned}
&\int u^i\ddot{\eta}_i dt = -\int \dot{u}^i\dot{\eta}_i dt \text{ which is valid since } \ u \in H^{1,2}\\
&= -\int (g_{ij}(u(t))\dot{u}^i\dot{\varphi}^j + g_{ij,k}\dot{u}^k\dot{u}^i\varphi^j)dt\\
&= \int (\frac{1}{2}g_{ij,k}(u)\dot{u}^i\dot{u}^j\varphi^k - g_{ij,k}(u)\dot{u}^k\dot{u}^i\varphi^j)dt \text{ by (6.11.4)}\\
&= \int (\frac{1}{2}g_{ij,k}g^{k\ell}\dot{u}^i\dot{u}^j - g_{ij,k}\dot{u}^k\dot{u}^i g^{j\ell})\eta_\ell dt\\
&= \int (\frac{1}{2}g^{i\ell}(g_{jk,\ell} - g_{j\ell,k} - g_{k\ell,j})\dot{u}^j\dot{u}^k\eta_i dt, \text{ renaming indices}\\
&= -\int \Gamma^i_{jk}\dot{u}^j\dot{u}^k\eta_i dt. \qquad (6.11.5)
\end{aligned}$$

With $v^i = -\Gamma^i_{jk}\dot{u}^j\dot{u}^k \in L^1$, the desired formula

$$\int u^i\ddot{\eta}_i = \int v^i\eta_i \text{ for } \eta \in C_0^\infty(S^1, M) \text{ with sufficiently small support}$$

then holds, and

$$u \in H^{2,1}.$$

By the Sobolev embedding theorem (Theorem A.1.7) we conclude

$$u \in H^{1,q} \text{ for all } q < \infty.$$

(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.1.7, instead of just $\|D^k f\|_{L^p}$.)

In particular, $u \in H^{1,4}(\Omega)$, hence

$$\Gamma^i_{jk}(u)\dot{u}^j\dot{u}^k \in L^2.$$

(6.11.5) then implies

$$u \in H^{2,2}$$

hence $\dot{u} \in C^0$ by Theorem A.1.2 again.

Now

$$\begin{aligned}
\frac{d}{dt}(\Gamma^i_{jk}(u)\dot{u}^j\dot{u}^k) &= 2\Gamma^i_{jk}\dot{u}^j\ddot{u}^k + \Gamma^i_{jk,\ell}\dot{u}^\ell\dot{u}^j\dot{u}^k \quad (\text{using } \Gamma^i_{jk} = \Gamma^i_{kj})\\
&\in L^2,
\end{aligned}$$

since $\ddot{u} \in L^2, \dot{u} \in L^\infty$. Thus

$$\Gamma^i_{jk}(u)\dot{u}^j\dot{u}^k \in H^{1,2}.$$

Then

$$u \in H^{3,2},$$

by (6.11.5) again.

Iterating this argument, we conclude

$$u \in H^{k,2}$$

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*

$$E(u_n) \leq \text{ const.}$$
$$\|DE(u_n)\| \to 0 \text{ as } n \to 0$$

*contains a strongly convergent subsequence with a closed geodesic as limit.*

*Proof.* First, by Hölder's inequality, for every $v \in \Lambda_0, t_1, t_2 \in S^1$

$$\begin{aligned} d(v(t_1), v(t_2)) &\leq \int_{t_1}^{t_2} (g_{ij}(v)\dot{v}^i\dot{v}^j)^{\frac{1}{2}} dt \leq ((t_2 - t_1) \int_{t_1}^{t_2} g_{ij}(v)\dot{v}^i\dot{v}^j dt)^{\frac{1}{2}} \\ &\leq \sqrt{2}|t_2 - t_1|^{\frac{1}{2}} E(v)^{\frac{1}{2}}. \end{aligned} \tag{6.11.6}$$

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 $E(u_n) \leq$ const. contains a uniformly convergent subsequence. We call the limit $u$. $u$ also has finite energy, actually

$$E(u) \leq \liminf_{n\to\infty} E(u_n).$$

We could just quote Theorem 8.4.2 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

$$E(u_n) - E(u) = \int \sum_{\mu=1}^{m} \eta_\mu (g^\mu_{ij}(u_n)\dot{u}^i_n \dot{u}^j_n - g^\mu_{ij}(u)\dot{u}^i\dot{u}^j)dt \tag{6.11.7}$$

where the superscript $\mu$ 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. (6.11.1))

$$\int (g_{ij}(u_n)\dot{u}^i_n\dot{\varphi}^j + \frac{1}{2}g_{ij,k}(u_n)\dot{u}^i_n\dot{u}^j_n\varphi^k)dt \to 0 \text{ as } n \to \infty$$

for all $\varphi \in H^{1,2}$.

We use

$$\varphi^j = \eta_\mu(u^j_n - u^j)$$

(where, of course, the difference is computed in local coordinates $f_\mu$).

Then

$$\begin{aligned} &\int g_{ij,k}(u_n)\dot{u}^i_n\dot{u}^j_n\eta_\mu(u^k_n - u^k)dt \\ &\leq \text{ const.} \cdot \max_t d(u_n(t), u(t))E(u_n) \\ &\to 0 \end{aligned}$$

as $n \to \infty$ since $E(u_n) \leq$ const. and $u_n \rightrightarrows u$ (after selecting a subsequence).

Consequently from (6.11.1), since $\|DE(u_n)\| \to 0$,

$$\int (g_{ij}(u_n)\dot{u}^i_n(\dot{u}^j_n - \dot{u}^j)\eta_\mu + g_{ij}(u^n)\dot{u}^i_n\dot{\eta}_\mu(u^j_n - u^j))dt \to 0.$$

The second term again goes to zero by uniform convergence.

We conclude

$$\int g_{ij}(u_n)\dot{u}^i_n(\dot{u}^j_n - \dot{u}^j)\eta_\mu \to 0 \text{ as } n \to \infty. \tag{6.11.8}$$

Now

$$\begin{aligned} &\int (g_{ij}(u_n)\dot{u}^i_n\dot{u}^j_n - g_{ij}(u)\dot{u}^i\dot{u}^j)\eta_\mu \\ &= \int \{(g_{ij}(u_n)\ \dot{u}^i_n(\dot{u}^j_n - \dot{u}^j) + (g_{ij}(u_n) - g_{ij}(u))\dot{u}^i_n\dot{u}^j \\ &\quad + g_{ij}(u)(\dot{u}^i_n - \dot{u}^i)\dot{u}^j\}\,\eta_\mu\,. \end{aligned} \tag{6.11.9}$$

The first term goes to zero by (6.11.8). 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 $V_\mu$. This implies that the third term goes to zero as well.

(6.11.9) now implies

$$E(u_n) \to E(u) \text{ as } n \to \infty$$

(cf. (6.11.7).)

$u$ 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 $E$, $\nabla E$, 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 $c$

$$\begin{aligned}&(\nabla E(c), V)_{H^1}\\ &= DE(c)(V) = \int_{S^1} \langle \dot{c}, \dot{V}\rangle dt.\end{aligned} \tag{6.11.10}$$

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 $\Lambda_0$ :

$$\begin{aligned}\frac{d}{dt}\Phi(t) &= -\nabla E(\Phi(t))\\ \Phi(0) &= c_0\end{aligned} \tag{6.11.11}$$

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 (6.11.11). Then*

$$\frac{d}{dt}E(\Phi(t)) \leq 0.$$

*Proof.* By the chain rule,

$$\frac{d}{dt}E(\Phi(t)) = DE(\Phi(t))(\frac{d}{dt}\Phi(t)) \tag{6.11.12}$$
$$= -\|\nabla E(\Phi(t))\|^2_{H^1} \leq 0. \qquad \square$$

**Theorem 6.11.2** *For any $c_0 \in \Lambda_0$, there exists a solution $\Phi : \mathbb{R}^+ \to \Lambda_0$ of*

$$\frac{d}{dt}\Phi(t) = -\nabla E(\Phi(t)) \tag{6.11.13}$$
$$\Phi(0) = c_0.$$

*Proof.* Let

$$A := \{T > 0 : \text{ there exists } \Phi : [0,T] \to \Lambda_0 \text{ solving (6.11.13), with } \Phi(0) = c_0\}.$$

(That $\Phi$ is a solution on $[0,T]$ means that there exists some $\varepsilon > 0$ for which $\Phi$ 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 $\Lambda_0$ and a neighborhood $V$ of 0 in the Hilbert space of $H^1$-vector fields along $c$: For $\xi \in V$

$$\xi(\tau) \mapsto \exp_{c(\tau)} \xi(\tau) \tag{6.11.14}$$

(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 $\Phi$ solves (6.11.13) 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 $\Phi$, since differentiable, in particular is continuous in $t$. Therefore, (6.11.14) 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 $\nabla E$ 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 $\Phi$ solves (6.11.13) 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 (6.11.13), and $0 < t_n < T, t_n \to T$ for $n \to \infty$.

Lemma 6.11.2 implies

$$E(\Phi(t_n)) \leq \text{ const.} \tag{6.11.15}$$

Therefore, the curves $\Phi(t_n)$ are uniformly Hölder continuous (cf. (6.11.6)), 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 (6.11.13), and closedness follows. □

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 Prop. 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*

$$E(c_3) = \kappa := \inf_{\lambda\in\Lambda} \max_{\tau\in[0,1]} E(\lambda(\tau)) > \max\{E(c_1), E(c_2)\} \tag{6.11.16}$$

*with $\Lambda := \Lambda(c_1, c_2) := \{\lambda \in C^0([0, 1], \Lambda_0) : \lambda(0) = c_1, \lambda(1) = c_2\}$, the set of all homotopies between $c_1$ and $c_2$.*

*Proof.* We first claim

$$\begin{aligned} &\exists \delta_0 > 0\, \forall \delta \text{ with } 0 < \delta \leq \delta_0\, \exists \varepsilon > 0\, \forall c \text{ with } d_1(c, c_i) = \delta : \\ &\qquad E(c) \geq E(c_i) + \varepsilon \quad \text{for } i = 1, 2. \end{aligned} \tag{6.11.17}$$

Indeed, otherwise, for $i = 1$ or 2,

$$\begin{aligned} &\forall \delta_0\, \exists 0 < \delta \leq \delta_0\, \forall n\, \exists \gamma_n \text{ with } d_1(\gamma_n, c_i) = \delta, \\ &\qquad E(\gamma_n) < E(c_i) + \frac{1}{n} \end{aligned}$$

If $\|DE(\gamma_n)\| \to 0$, then $(\gamma_n)$ is a Palais-Smale sequence and by Theorem 6.11.1 converges (after selection of a subsequence) to some $\gamma_0$ with $d_1(\gamma_0, c_i) = \delta, E(\gamma_0) = E(c_i)$, contradicting the strict local minimizing property of $c_i$.

If $\|DE(\gamma_n)\| \geq \eta > 0$ for all $n$, then there exists $\rho > 0$ with

$$\|DE(\gamma)\| \geq \frac{\eta}{2} \text{ whenever } d_1(\gamma_n, \gamma) \leq \rho. \tag{6.11.18}$$

This follows, because $\|D^2E\|$ is uniformly bounded on $E$-bounded sets.

(6.11.18) 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 (6.11.17) is correct.

(6.11.17) implies

$$\kappa > \max(E(c_1), E(c_2)). \tag{6.11.19}$$

We let now $K^\kappa$ be the set of all closed geodesics, i.e. curves $c$ in $\Lambda^0$ with $DE(c) = 0, E(c) = \kappa$, homotopic to $c_1$ and $c_2$.

We have to show

$$K^\kappa \neq \emptyset.$$

We assume on the contrary

$$K^\kappa = \emptyset. \tag{6.11.20}$$

We claim that there exists $\eta > 0, \alpha > 0$ with

$$\|DE(c)\| \geq \alpha \tag{6.11.21}$$

whenever $c$ is homotopic to $c_1, c_2$ and satisfies

$$\kappa - \eta \leq E(c) \leq \kappa + \eta. \tag{6.11.22}$$

Namely, otherwise, there exists a sequence $(\gamma_n)_{n\in\mathbb{N}}$ of $H^1$-curves homotopic to $c_1, c_2$, with

$$\lim_{n\to\infty} E(\gamma_n) = \kappa$$
$$\lim_{n\to\infty} DE(\gamma_n) = 0$$

$(\gamma_n)_{n\in\mathbb{N}}$ 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 (6.11.21) has to hold if $\kappa - \eta \leq E(c) \leq \kappa + \eta$.

From Theorem 6.11.2, we know that for any $t > 0$, there is a map

$$\begin{array}{rcl} \Lambda_0 & \to & \Lambda_0 \\ c & \mapsto & \Phi_t(c), \quad \text{where} \quad \begin{array}{rcl} \Phi_t(c) & = & \Phi(t) \text{ solves} \\ \frac{d}{dt}\Phi(t) & = & -\nabla E(\Phi(t)) \\ \Phi(0) & = & c. \end{array} \end{array}$$

With the help of this gradient flow, we may now decrease the energy below the level $\kappa$, contradicting (6.11.20). For that purpose,

let $\lambda \in \Lambda$ satisfy

$$\max_{\tau\in[0,1]} E(\lambda(\tau)) \leq \kappa + \eta. \tag{6.11.22}$$

Then, as in the proof of Lemma 6.11.2

$$\frac{d}{dt}E(\Phi_t(\lambda(\tau))) = -\|\nabla E(\Phi_t(\lambda(\tau)))\|^2 \leq 0. \tag{6.11.23}$$

In particular, for $t > 0$

$$\max E(\Phi_t(\lambda(\tau))) \leq \max E(\lambda(\tau)) \leq \kappa + \eta. \tag{6.11.24}$$

Since $c_1$ and $c_2$ are closed geodesics, i.e. critical points of $E$, $\nabla E(c_i) = 0$ ($i = 1, 2$), hence

$$\Phi_t(c_i) = c_i \quad \text{for all } t \geq 0.$$

Therefore

$$\Phi_t \circ \lambda \in \Lambda \quad \text{for } t \geq 0.$$

(6.11.21), (6.11.23) imply

$$\frac{d}{dt}E(\Phi_t(\lambda(\tau))) \leq -\alpha^2 \quad \text{whenever } E(\Phi_t(\lambda(\tau))) > \kappa - \eta. \tag{6.11.25}$$

(6.11.22), (6.11.25) imply

$$E(\Phi_s(\lambda(\tau))) \leq \kappa - \eta$$

for $s \geq \frac{2\eta}{\alpha^2}$ and all $\tau \in [0, 1]$, contradicting the definition of $\kappa$. Therefore, (6.11.20) cannot hold, and the theorem is proved. □

As the culmination of this §, 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.* □

We now *prove 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 1.4.6. 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 $\partial D^{i-1}$ 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

$$\begin{aligned} \Lambda := \{&\lambda : D^{i-1} \to \Lambda_0, \lambda \text{ homotopic to } H \\ &\text{as described above, in particular mapping } \partial D^{i-1} \text{ to} \\ &\text{point curves}\}, \end{aligned}$$

and put

$$\kappa := \inf_{\lambda \in \Lambda} \max_{z \in D^{i-1}} E(\lambda(z)).$$

As in the proof of Thm. 6.11.3, we see that there exists a closed geodesic $\gamma$ with

$$E(\gamma) = \kappa.$$

It only remains to show that $\kappa > 0$, in order to exclude that $\gamma$ 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 $\varepsilon$. 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 $t$. $H(z,s)(t) := q_{z,t}(1-s)$ then defines a homotopy between $\lambda_\varepsilon$ 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 $\lambda_\varepsilon$ is homotopic to a constant map, hence so are $H$ and $h$, contradicting the choice of $h$. Therefore, $\kappa$ 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[158], 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[77] and Bangert[13]. For an explicit estimate for the growth of the number of closed geodesics of length $\leq \ell$, see Hingston[119] 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[91], Jost[130], as well as Klingenberg[158].

## 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 of 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 $\gamma$ be a smooth convex closed Jordan curve in the plane $\mathbb{R}^2$. Show that there exists a straight line $\ell$ 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 $\gamma$ orthogonally in two points.
(Hint: $\gamma$ bounds a compact set $A$ in $\mathbb{R}^2$ by the Jordan curve theorem. For every line $\ell$ in $\mathbb{R}^2$, put

$$L_A(\ell) := \text{length}(A \cap \ell).$$

Find a nontrivial critical point $\ell_0$ 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 10) as follows:
Let $M$ be diffeomorphic to $S^2$, $\gamma$ a smooth closed Jordan curve in $M$. Show that there exists a nontrivial geodesic arc in $M$ meeting $\gamma$ 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?