Critical Point Theory.

Elementary Critical Point Theory

The essence of Morse Theory is a collection of theorems describing the intimate relationship between the topology of a manifold and the critical point structure of real valued functions on the manifold. This body of theorems has over and over again proved itself to be one of the most powerful and far-reaching tools available for advancing our understanding of differential topology and analysis. But a good mathematical theory is more than *just* a collection of theorems; in addition it consists of a tool box of related conceptualizations and techniques that have been gradually built up to help understand some circle of mathematical problems. Morse Theory is no exception, and its basic concepts and constructions have an unusual appeal derived from an underlying geometric naturality, simplicity, and elegance. In these lectures we will cover some of the more important theorems and applications of Morse Theory and, beyond that, try to give a feeling for and an ability to work with these beautiful and powerful techniques.

Preliminaries

We will assume that the reader is familiar with the standard definitions and notational conventions introduced in the Appendix. We begin with some basic assumptions and further notational conventions. In all that follows $f : M \to \boldsymbol{R}$ will denote a smooth real valued function on a smooth finite or infinite dimensional hilbert manifold M. We will make three basic assumptions about M and f:

(a) (Completeness). M is a complete Riemannian manifold.

(b) (Boundedness below) The function f is bounded below on M. We will let B denote the greatest lower bound of f, so our assumption is that $B > -\infty$.

(c) (Condition C) If $\{x_n\}$ is any sequence in M for which $|f(x_n)|$ is bounded and for which $\|df_{x_n}\| \to 0$, it follows that $\{x_n\}$ has a convergent subsequence, $x_{n_k} \to p$.
(By continuity, $\|df_p\| = 0$, so that p is a critical point of f).

Of course if M is compact then with *any* choice of Riemannian metric for M all three conditions are automatically satisfied. In fact we recommend that a reader new to Morse theory develop intuition by always thinking of M as compact, and we will encourage this by using mainly compact surfaces for our examples and diagrams. Nevertheless it is important to realize that in our

formal proofs of theorems only (a), (b), and (c) will be used, and that as we shall see later these conditions do hold in important cases where M is not only non-compact, but even infinite dimensional.

Recall that p in M is called a *critical point* of f if $df_p = 0$. Other points of M are called *regular points* of f. Given a real number c we call $f^{-1}(c)$ the *c-level* of f, and we say it is a *critical level* (and that c is a *critical value* of f) if it contains at least one critical point of f. Other real numbers c (even those for which $f^{-1}(c)$ is empty!) are called *regular values* of f and the corresponding levels $f^{-1}(c)$ are called *regular levels*. We denote by M_c (or by $M_c(f)$ if there is any ambiguity) the "part of M below the level c", i.e., $f^{-1}((-\infty, c])$. It is immediate from the inverse function theorem that for a regular value c, $f^{-1}(c)$ is a (possibly empty) smooth, codimension one submanifold of M, that M_c is a smooth submanifold with boundary, and that $\partial M_c = f^{-1}(c)$. We will denote by $\mathcal{C}$ the set of all critical points of f, and by $\mathcal{C}_c$ the set $\mathcal{C} \cap f^{-1}(c)$ of critical points at the level c. Then we have the following lemma.

9.1.1. Lemma. *The restriction of f to $\mathcal{C}$ is proper. In particular, for any $c \in \boldsymbol{R}$, $\mathcal{C}_c$ is compact.*

PROOF. We must show that $f^{-1}([a, b]) \cap \mathcal{C}$ is compact, i.e. if $\{x_n\}$ is a sequence of critical points with $a \leq f(x_n) \leq b$ then $\{x_n\}$ has a convergent subsequence. But since $\|\nabla f_{x_n}\| = 0$ this is immediate from Condition C. ∎

Since proper maps are closed we have:

9.1.2. Corollary. *The set $f(\mathcal{C})$ of critical values of f is a closed subset of $\boldsymbol{R}$.*

Recall that the gradient of f is the smooth vector field ∇f on M dual to df, i.e., characterized by $Yf = \langle Y, \nabla f \rangle$ for any tangent vector Y to M. Of course if Y is tangent to a level $f^{-1}(c)$ then $Yf = 0$, so at each regular point x it follows that ∇f is orthogonal to the level through x. In fact it follows easily from the Schwarz inequality that, at a regular point, ∇f points in the direction of most rapid increase of f. We will denote by φ_t the maximal flow generated by $-\nabla f$. For each x in M $\varphi_t(x)$ is defined on an interval $\alpha(x) < t < \beta(x)$ and $t \mapsto \varphi_t(x)$ is the maximal solution curve of $-\nabla f$ with initial condition x. Thus $\frac{d}{dt}\varphi_t(x) = -\nabla f_{\varphi_t(x)}$ and so $\frac{d}{dt} f(\varphi_t(x)) = -\nabla f(f) = -\|\nabla f\|^2$, so $f(\varphi_t(x))$ is monotonically decreasing in t. Since f is bounded below by B it follows that $f(\varphi_t(x))$ has a limit as $t \to \beta(x)$.

We shall now prove the important fact that $\{\varphi_t\}$ is a "positive semi-group", that is, for each x in M $\beta(x) = \infty$, so $\varphi_t(x)$ is defined for all $t > 0$.

9.1.3. Lemma. *A C^1 curve $\sigma : (a, b) \to M$ of finite length has*

relatively compact image.

PROOF. Since M is complete it will suffice to show that the image of σ is totally bounded. Since $\int_a^b \|\sigma'(t)\|\,dt < \infty$, given $\epsilon > 0$ there exist $t_0 = a < t_1 < \ldots < t_n < t_{n+1} = b$ such that $\int_{t_i}^{t_{i+1}} \|\sigma'(t)\|\,dt < \epsilon$. Then by the definition of distance in M it is clear that the $x_i = \sigma(t_i)$ are ϵ-dense in the image of σ. ∎

9.1.4. Proposition. *Let X be a smooth vector field on M and let $\sigma : (a, b) \to M$ be a maximal solution curve of X. If $b < \infty$ then $\int_0^b \|X_{\sigma(t)}\|\,dt = \infty$. Similarly if $a > -\infty$ then $\int_a^0 \|X_{\sigma(t)}\|\,dt = \infty$.*

PROOF. Since σ is maximal, if $b < \infty$ then $\sigma(t)$ has no limit point in M as $t \to b$. Thus, by the lemma, $\sigma : [0, b) \to M$ must have infinite length, and since $\sigma'(t) = X_{\sigma(t)}$, $\int_0^b \|X_{\sigma(t)}\|\,dt = \infty$. ∎

9.1.5. Corollary. *A smooth vector field X on M of bounded length, generates a one-parameter group of diffeomorphisms of M.*

PROOF. Suppose $\|X\| \leq K < \infty$. If $b < \infty$ then $\int_0^b \|X_{\sigma(t)}\|\,dt \leq bK < \infty$, contradicting the Proposition. By a similar argument $a > -\infty$ is also impossible. ∎

9.1.6. Theorem. *The flow $\{\varphi_t\}$ generated by $-\nabla f$ is a positive semi-group; that is, for all $t > 0$ φ_t is defined on all of M. Moreover for any x in M $\varphi_t(x)$ has at least one critical point of f as a limit point as $t \to \infty$.*

PROOF. Let $g(t) = f(\varphi_t(x))$ and note that $B \leq g(T) = g(0) + \int_0^T g'(t)\,dt = g(0) - \int_0^T \|\nabla f_{\varphi_t(x)}\|^2\,dt$. Since this holds for all $T < \beta(x)$, by the Schwarz inequality

$$\int_0^{\beta(x)} \|\nabla f_{\varphi_t(x)}\|\,dt \leq \sqrt{\beta(x)} \left(\int_0^{\beta(x)} \|\nabla f_{\varphi_t(x)}\|^2\,dt \right)^{\frac{1}{2}},$$

which is less than or equal to $\sqrt{\beta(x)}(g(0) - B)^{\frac{1}{2}}$, and hence would be finite if $\beta(x)$ were finite. It follows from the preceding proposition that $\beta(x)$ must be infinite and consequently $\|\nabla f_{\varphi_t(x)}\|$ cannot be bounded away from zero as $t \to \infty$, since otherwise $\int_0^\infty \|\nabla f_{\varphi_t(x)}\|^2\,dt$ would be infinite, whereas we know

it is less than $g(0) - B$. Finally, since $f(\varphi_t(x))$ is bounded, it now follows from Condition C that $\varphi_t(x)$ has a critical point of f as a limit point as $t \to \infty$. ∎

9.1.7. Remark. An exactly parallel argument shows that as $t \to \alpha(x)$ either $f(\varphi_t(x)) \to \infty$ or else $\alpha(x)$ must be $-\infty$ and $\varphi_t(x)$ has a critical point of f as a limit point as $t \to -\infty$.

9.1.8. Corollary. *If x in M is not a critical point of f then there is a critical point p of f with $f(p) < f(x)$.*

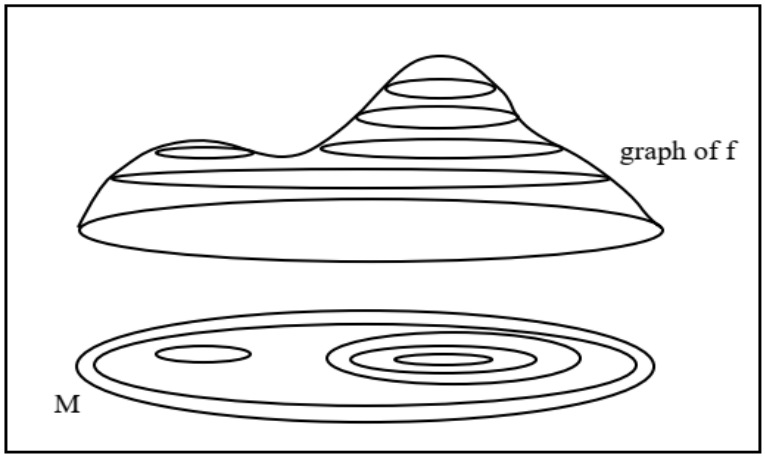
PROOF. Choose any critical point of f that is a limit point of $\varphi_t(x)$ as $t \to \infty$. ∎

9.1.9. Theorem. *The function f attains its infimum B. That is, there is a critical point p of f with $f(p) = B$.*

PROOF. Choose a sequence $\{x_n\}$ with $f(x_n) \to B$. By the preceding corollary we can assume that each x_n is a critical point of f. Then by Condition C a subsequence of $\{x_n\}$ converges to a critical point p of f, and clearly $f(p) = B$. ∎

In order to understand and work effectively with a complex mathematical subject one must get behind its purely logical content and develop some intuitive picture of the key concepts. Normally these intuitions are imprecise and vary considerably from one individual to another, and this often can be a barrier to the easy communication of mathematical ideas. One of the pleasant and special features of Morse Theory is that it has a generally accepted metaphor for visualizing many of its basic concepts. Since much of the terminology and motivation of the theory is based on this metaphor we shall now explain it in some detail.

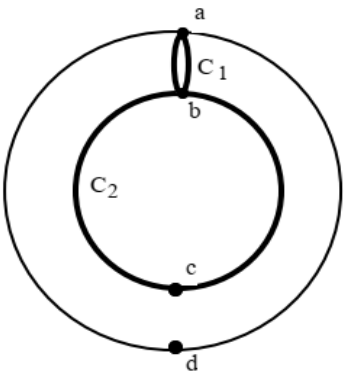
Starting with our smooth function $f : M \to \boldsymbol{R}$ we build a "world" $\mathcal{W} = M \times \boldsymbol{R}$. We now identify M *not* with $M \times \{0\}$, (which we think of as "sea-level") but rather with the graph of f; that is we identify $x \in M$ with $(x, f(x)) \in \mathcal{W}$.



The projection $z : \mathcal{W} \to \boldsymbol{R}$, $(x,t) \mapsto z(x,t) = t$ we think of as "height above sea-level". Since $z(x, f(x)) = f(x)$ this means that our original function f represents altitude in our new realization of M. And this in turn means that the a-level of f becomes just that, it is the intersection of the graph of f with the altitude level-surface $z = a$ in $\mathcal{W}$. The critical points of f are now the valleys, passes, and mountain summits of the graph of f, that is the points where the tangent hyperplane to M is horizontal. We think of the projection of $\mathcal{W}$ onto M as providing us with a "topo map" of our world; projecting the a-level of f in $\mathcal{W}$ into M gives us the old $f^{-1}(a)$ which we now think of as an isocline (surface of constant height) on this topographic map.

We give $\mathcal{W}$ the product Riemannian metric, and recall that the negative gradient vector field $-\nabla f$ represents the direction of "steepest descent" on the graph of f; pointing orthogonal to the level surfaces in the downhill direction. Thus (very roughly speaking) we may think of the flow φ_t we have been using as modelling the way a very syrupy liquid would flow down the graph of f under the influence of gravity. We shall return to this picture many times in the sequel to provide intuition, motivation, and terminology.

There is a particular Morse function that, while not completely trivial, is so intuitive and easy to analyze, that it is is everybody's favorite model, and we will use it frequently to illustrate various concepts and theorems. Informally it is the height above the floor on a tire standing in ready-to-roll position. More precisely, we take M to be the torus obtained by revolving the circle of radius 1 centered at $(0,2)$ in the (x,y)-plane about the y-axis, and define $f : M \to \boldsymbol{R}$ to be orthogonal projection on the z-axis.



This function has four critical points: a maximum $a = (0, 0, 3)$ at the level 3, a minimum $d = (0, 0, -3)$ at the level -3, and two saddle points $b = (0, 0, 1)$, and $c = (0, 0, -1)$, at the levels 1 and -1 respectively. The reader should analyze the asymptotic properties of the flow $\varphi_t(x)$ of $-\nabla f$ in this case. Of course the four critical points are fixed. Other points on the circle $C_1 : x = 0, y^2 + (z-2)^2 = 1$ tend to b, other points on $C_2 : y = 0, x^2 + y^2 = 1$ tend to c, and all remaining points tend to the minimum, d. We shall refer to this function as the "height function on the torus".

The study of the flow $\{\varphi_t\}$ generated by $-\nabla f$ (or more generally of vector fields proportional to it) is one of the most important tools of Morse theory. We have seen a little of its power above and we shall see much more in what follows.

9.2. The First Deformation Theorem

We shall now use the flow $\{\varphi_t\}$ generated by $-\nabla f$ to deform subsets of the manifold M, and see how this leads to a very general method (called "minimaxing") for locating critical points of f. We will then illustrate minimaxing with an introduction to Lusternik-Schnirelman theory.

9.2.1. Lemma. *If O is a neighborhood of the set $\mathcal{C}_c$ of critical points of f at the level c, then there is an $\epsilon > 0$ such that $\|\nabla f\|$ is bounded away from zero on $f^{-1}(c-\epsilon, c+\epsilon) \setminus O$.*

PROOF. Suppose not. Then for each positive integer n we could choose an x_n in $f^{-1}(c - \frac{1}{n}, c + \frac{1}{n}) \setminus O$ such that $\|\nabla f_{x_n}\| < \frac{1}{n}$. By Condition C, a subsequence of $\{x_n\}$ would converge to a critical point p of f with $f(p) = c$, so $p \in \mathcal{C}_c$ and eventually the subsequence must get inside the neighborhood O of $\mathcal{C}_c$, a contradiction. ∎

Since $\mathcal{C}_c$ is compact,

9.2.2. Lemma. *Any neighborhood of $\mathcal{C}_c$ includes the neighborhoods of the form $N_\delta(\mathcal{C}_c) = \{x \in M \mid \rho(x, \mathcal{C}_c) < \delta\}$ provided δ is sufficiently small.*

Now let U be any neighborhood of $\mathcal{C}_c$ in M, and choose a $\delta_1 > 0$ such that $N_{\delta_1}(\mathcal{C}_c) \subseteq U$. Since $\|\nabla f_p\| = 0$ on $\mathcal{C}_c$ we may also assume that $\|\nabla f_p\| \leq 1$ for $p \in N_{\delta_1}(\mathcal{C}_c)$.

If ϵ is small enough then, by 2.1, for any $\delta_2 > 0$ we can choose $\mu > 0$ such that $\|\nabla f_p\| \geq \mu$ for $p \in f^{-1}([c-\epsilon, c+\epsilon])$ and $\rho(p, \mathcal{C}_c) \geq \delta_2$ (i.e., $p \notin N_{\delta_2}(\mathcal{C}_c)$). In particular we can assume $\delta_2 < \delta_1$, so that $N_{\delta_2}(\mathcal{C}_c) \subseteq N_{\delta_1}(\mathcal{C}_c) \subseteq U$.

9.2.3. First Deformation Theorem. *Let U be any neighborhood of $\mathcal{C}_c$ in M. Then for $\epsilon > 0$ sufficiently small $\varphi_1(M_{c+\epsilon} \backslash U) \subseteq M_{c-\epsilon}$.*

PROOF. Let $\epsilon = \min(\frac{1}{2}\mu^2, \frac{1}{2}\mu^2(\delta_1 - \delta_2))$, where δ_1, δ_2, and μ are chosen as above. Let $p \in f^{-1}([c-\epsilon, c+\epsilon]) \backslash U$. We must show that $f(\varphi_1(p)) \leq c - \epsilon$, and since $f(\varphi_t(p))$ is monotonically decreasing we may assume that $\varphi_t(p) \in f^{-1}([c-\epsilon, c+\epsilon])$ for $0 \leq t < 1$. Thus by definition of δ_2 we can also assume that if $\rho(\varphi_t(p), \mathcal{C}_c) \geq \delta_2$ then $\|\nabla f_{\varphi_t(p)}\| \geq \mu$.

Since $\varphi_0(p) = p$ and $\frac{d}{dt} f(\varphi_t(p)) = -\|\nabla f_{\varphi_t(p)}\|^2$ we have:

$$\begin{aligned} f(\varphi_1(p)) &= f(\varphi_0(p)) + \int_0^1 -\|\nabla f_{\varphi_t(p)}\|^2 \, dt \\ &\leq c + \epsilon - \int_0^1 \|\nabla f_{\varphi_t(p)}\|^2 \, dt \,, \end{aligned}$$

so it will suffice to show that

$$\int_0^1 \|\nabla f_{\varphi_t(p)}\|^2 \, dt \geq 2\epsilon = \min(\mu^2, \mu^2(\delta_1 - \delta_2)) \,.$$

We will break the remainder of the proof into two cases.

Case 1. $\rho(\varphi_t(t), \mathcal{C}_c) > \delta_2$ for all $t \in [0, 1]$.

Then $\|\nabla f_{\varphi_t(p)}\| \geq \mu$ for $0 \leq t \leq 1$ and hence
$\int_0^1 \|\nabla f_{\varphi_t(p)}\|^2 \, dt \geq \mu^2 \geq \min(\mu^2, \mu^2(\delta_1 - \delta_2)) \,.$

Case 2. $\rho(\varphi_t(t), \mathcal{C}_c) \leq \delta_2$ for some $t \in [0, 1]$.

Let t_2 be the first such t. Since $p \notin U$, *a fortiori* $p \notin N_{\delta_1}(\mathcal{C}_c)$, i.e., $\rho(\varphi_0(p), \mathcal{C}_c) \geq \delta_1 > \delta_2$, so there is a last $t \in [0, 1]$ less than t_2 such that $\rho(\varphi_0(p), \mathcal{C}_c) \geq \delta_1$. We denote this value of t by t_1, so that $0 < t_1 < t_2 < 1$, and in the interval $[t_1, t_2]$ we have $\delta_1 \geq \rho(\varphi_t(p), \mathcal{C}_c) \geq \delta_2$. Note that $\rho(\varphi_{t_1}(p), \mathcal{C}_c) \geq \delta_1$ while $\rho(\varphi_{t_2}(p), \mathcal{C}_c) \leq \delta_2$ and hence by the triangle inequality $\rho(\varphi_{t_1}(p), \varphi_{t_2}(p)) \geq \delta_1 - \delta_2$. It follows that any curve joining $\varphi_{t_1}(p)$

to $\varphi_{t_2}(p)$ has length greater or equal $\delta_1 - \delta_2$, and in particular this is so for $t \mapsto \varphi_t(p), \quad t_1 \le t \le t_2$. Since $\frac{d}{dt}\varphi_t(p) = -\nabla f_{\varphi_t(p)}$ this means:

$$\int_{t_1}^{t_2} \|\nabla f_{\varphi_t(p)}\|\, dt \ge \delta_1 - \delta_2 \,.$$

By our choice of δ_1, $\|\nabla f_{\varphi_t(p)}\| \le 1$ for t in $[t_1, t_2]$, since $\rho(\varphi_t(p), \mathcal{C}_c) \le \delta_1$ for such t. Thus

$$t_2 - t_1 = \int_{t_1}^{t_2} 1\, dt \ge \int_{t_1}^{t_2} \|\nabla f_{\varphi_t(p)}\|\, dt \ge \delta_1 - \delta_2 \,.$$

On the other hand, by our choice of δ_2, for t in $[t_1, t_2]$ we also have $\|\nabla f_{\varphi_t(p)}\| \ge \mu$, since $\rho(\varphi_t(p), \mathcal{C}_c) \ge \delta_2$ for such t. Thus

$$\begin{aligned}
\int_0^1 \|\nabla f_{\varphi_t(p)}\|^2\, dt &\ge \int_{t_1}^{t_2} \|\nabla f_{\varphi_t(p)}\|^2\, dt \\
&\ge \int_{t_1}^{t_2} \mu^2\, dt = \mu^2(t_2 - t_1) \\
&\ge \mu^2(\delta_2 - \delta_1) \\
&\ge \min(\mu^2, \mu^2(\delta_1 - \delta_2)) \,. \qquad \blacksquare
\end{aligned}$$

9.2.4. Corollary. *If c is a regular value of f then, for some $\epsilon > 0$, $\varphi_1(M_{c+\epsilon}) \subseteq M_{c-\epsilon}$.*

PROOF. Since $\mathcal{C}_c = \emptyset$ we can take $U = \emptyset$. $\blacksquare$

Let $\mathcal{F}$ denote a non-empty family of non-empty compact subsets of M. We define minimax$(f, \mathcal{F})$, the minimax of f over the family $\mathcal{F}$, to be the infimum over all F in $\mathcal{F}$ of the maximum of f on F. Now the maximum value of f on F is just the smallest c such that $F \subseteq M_c$. So minimax$(f, \mathcal{F})$, is the smallest c such that, for any positive ϵ, we can find an F in $\mathcal{F}$ with $F \subseteq M_{c+\epsilon}$. The family $\mathcal{F}$ is said to be invariant under the positive time flow of $-\nabla f$ if whenever $F \in \mathcal{F}$ and $t > 0$ it follows that $\varphi_t(F) \in \mathcal{F}$.

9.2.5. Minimax Principle. *If $\mathcal{F}$ is a family of compact subsets of M invariant under the positive time flow of $-\nabla f$ then minimax$(f, \mathcal{F})$ is a critical value of M.*

PROOF. By definition of minimax we can find an F in $\mathcal{F}$ with $F \subseteq M_{c+\epsilon}$. Suppose c were a regular value of f. Then by the above Corollary

$\varphi_1(M_{c+\epsilon}) \subseteq M_{c-\epsilon}$ and *a fortiori* $\varphi_1(F) \subseteq M_{c-\epsilon}$. But since $\mathcal{F}$ is invariant under the positive time flow of $-\nabla f$, $\varphi_1(F)$ is also in the family $\mathcal{F}$ and it follows that $\text{minimax}(f, \mathcal{F}) \leq c - \epsilon$, a contradiction. ∎

Of course any family $\mathcal{F}$ of compact subsets of M invariant under homotopy is *a fortiori* invariant under the positive time flow of $-\nabla f$. Here are a few important examples:

- If α is a homotopy class of maps of some compact space X into M take $\mathcal{F} = \{\text{im}(f) | f \in \alpha\}$.
- Let α be a homology class of M and let $\mathcal{F}$ be the set of compact subsets F of M such that α is in the image of $i_* : H_*(F) \to H_*(M)$.
- Let α be a cohomology class of M and let $\mathcal{F}$ be the set of compact subsets F of M that support α (i.e., such that α restricted to $M \backslash F$ is zero).

There are a number of related applications of the Minimax Principle that go under the generic name of "Mountain Pass Theorem". Here is a fairly general version.

9.2.6. Definition. Let M be connected. We will call a subset $\mathcal{R}$ of M a *mountain range* relative to f if it separates M and if, on each component of $M \setminus \mathcal{R}$, f assumes a value strictly less that $\inf(f|\mathcal{R})$.

9.2.7. Mountain Pass Theorem. *If M is connected and $\mathcal{R}$ is a mountain range relative to f then f has a critical value $c \geq \inf(f|\mathcal{R})$.*

PROOF. Set $\alpha = \inf(f|\mathcal{R})$ and let M^0 and M^1 be two different components of $M \setminus \mathcal{R}$. Define $M^i_\alpha = \{x \in M^i \mid f(x) < \alpha\}$. By assumption each M^i_α is non-empty, and since M is connected we can find a continuous path $\sigma : I \to M$ such that $\sigma(i) \in M^i_\alpha$. Let Γ denote the set of all such paths σ and let $\mathcal{F} = \{\text{im}(\sigma) \mid \sigma \in \Gamma\}$, so that $\mathcal{F}$ is a non-empty family of compact subsets of M. Since $\sigma(0)$ and $\sigma(1)$ are in different components of $M \setminus \mathcal{R}$ it follows that $\sigma(t_0) \in \mathcal{R}$ for some $t_0 \in I$, so $f(\sigma(t_0)) \geq \alpha$ and hence $\text{minimax}(f, \mathcal{F}) \geq \alpha$. Thus, by the Minimax Principle, it will suffice to show that if $\sigma \in \Gamma$ and $t > 0$ then $\varphi_t \circ \sigma \in \Gamma$, where φ_t is the positive time flow of $-\nabla f$. And for this it will clearly suffice to show that if x is in M^i_α then so is $\varphi_t(x)$. But since $f(\varphi_0(x)) = f(x) < \alpha$, and $f(\varphi_t(x))$ is a non-increasing function of t, it follows that $f(\varphi_t(x)) < \alpha$, so in particular $\varphi_t(x) \in M \setminus \mathcal{R}$, and hence x and $\varphi_t(x)$ are in the same component of $M \setminus \mathcal{R}$. ∎

In recent years Mountain Pass Theorems have had extensive applications in proving existence theorems for solutions to both ordinary and partial differential equations. For further details see [Ra].

We next consider Lusternik-Schnirelman Theory, an early and elegant application of the Minimax Principle. This material will not be used in the remainder of these notes and may be skipped without loss of continuity.

A subset A of a space X is said to be contractible in X if the inclusion map $i : A \to X$ is homotopic to a constant map of A into X. We say that A *has category* m *in* X (and write $\mathrm{cat}(A, X) = m$)) if A can be covered by m (but no fewer) closed subsets of X, each of which is contractible in X. We define $\mathrm{cat}(X) = \mathrm{cat}(X, X)$. Here are some obvious properties of the set function cat that follow immediately from the definition.

(1) $\mathrm{cat}(A, X) = 0$ if and only if $A = \emptyset$.
(2) $\mathrm{cat}(A, X) = 1$ if and only if $\bar{A}$ is contractible in X.
(3) $\mathrm{cat}(A, X) = \mathrm{cat}(\bar{A}, X)$
(4) If A is closed in X then $\mathrm{cat}(A, X) = m$ if and only if A is the union of m (but not fewer) closed sets, each contractible in X.
(5) $\mathrm{cat}(A, X)$ is monotone; i.e., if $A \subseteq B$ then $\mathrm{cat}(A, X) \leq \mathrm{cat}(B, X)$.
(6) $\mathrm{cat}(A, X)$ is subadditive; i.e., $\mathrm{cat}(A \cup B, X) \leq \mathrm{cat}(A, X) + \mathrm{cat}(B, X)$.
(7) If A and B are closed subsets of X and A is deformable into B in X (i.e., the inclusion $i : A \to X$ is homotopic as a map of A into X to a map with image in B), then $\mathrm{cat}(A, X) \leq \mathrm{cat}(B, X)$.
(8) If $h : X \to X$ is a homeomorphism then $\mathrm{cat}(h(A), X) = \mathrm{cat}(A, X)$.

To simplify our discussion of Lusternik-Schnirelman Theory we will temporarily assume that M is compact. For $m \leq \mathrm{cat}(M)$ we define $\mathcal{F}_m$ to be the collection of all compact subsets F of M such that $\mathrm{cat}(F, M) \geq m$. Note that $\mathcal{F}_m$ contains M itself and so is non-empty. We define $c_m(f) = \mathrm{minimax}(f, \mathcal{F}_m)$. By the monotonicity of cat($, M$) we can equally well define $c_m(f)$ by the formula

$$c_m(f) = \inf\{a \in \boldsymbol{R} \mid \mathrm{cat}(M_a(f), M) \geq m\}\,.$$

9.2.8. Proposition. *For $m = 0, 1, \ldots, \mathrm{cat}(M)$, $c_m(f)$ is a critical value of M.*

PROOF. This is immediate from The Minimax Principle, since by (7) above, $\mathcal{F}_m$ is homotopy invariant. ∎

Now $\mathcal{F}_{m+1}$ is clearly a subset of $\mathcal{F}_m$, so $c_m(f) \leq c_{m+1}(f)$. But of course equality *can* occur (for example if f is constant). However as the next result shows, this will be compensated for by having more critical points at this level.

9.2.9. Lusternik-Schnirelman Multiplicity Theorem. *If $c_{n+1}(f) = c_{n+2}(f) = \cdots c_{n+k}(f) = c$ then there are at least k critical points at the level c. Hence if $1 \leq m \leq \mathrm{cat}(M)$ then f has at least m critical points at or below the level $c_m(f)$. In particular every smooth function $f : M \to \boldsymbol{R}$ has at least $\mathrm{cat}(M)$ critical points altogether.*

PROOF. Suppose that there are only a finite number r of critical points $x_1, \ldots, x_r$ at the level c and choose open neighborhoods O_i of the x_i whose

closures are disjoint closed disks (hence in particular contractible). Putting $O = O_1 \cup \ldots \cup O_r$, clearly $\mathrm{cat}(O, M) \leq r$. By the First Deformation Theorem, for some $\epsilon > 0$ $M_{c+\epsilon} \setminus O$ can be deformed into $M_{c-\epsilon}$. Since $c - \epsilon < c = c_{n+1}$, $\mathrm{cat}(M_{c-\epsilon}, M) < n+1$, and so by (7) above $\mathrm{cat}(M_{c+\epsilon} \setminus O, M) \leq n$. Thus, by subadditivity and monotonicity of cat,

$$\mathrm{cat}(M_{c+\epsilon}, M) \leq \mathrm{cat}((M_{c+\epsilon} \setminus O) \cup O, M) \leq n + r$$

and hence

$$c < c + \epsilon < \inf\{a \in \boldsymbol{R} \mid \mathrm{cat}(M_a, M) > n + r + 1\} = c_{n+r+1}(f)\,.$$

Since on the other hand $c = c_{n+k}(f)$, (and $c_m(f) \leq c_{m+1}(f)$) it follows that $n + r + 1 > n + k$, so $r \geq k$. ∎

Taken together the following two propositions make it easy to compute exactly the category of some spaces.

9.2.10. Proposition. *If M is connected, and A is a closed subset of M, then $\mathrm{cat}(A, M) \leq \dim(A) + 1$.*

PROOF. (Cf. [Pa5]). Let $\{O_\alpha\}$ be a cover of A by A-open sets, each contractible in M. Letting $n = \dim(A)$, by a lemma of J. Milnor (cf. [Pa4, Lemma 2.4]), there is a an open cover $\{G_{i\beta}\}$, $i = 0, 1, \ldots, n$, $\beta \in B_i$ of A, refining the covering by the O_α, such that $G_{i\beta} \cap G_{i\beta'} = \emptyset$ for $\beta \neq \beta'$. Since each $G_{i\beta}$ is contractible in M, and M is connected, it follows that $G_i = \bigcup\{G_{i\beta} \mid \beta \in B_i\}$ is contractible in M for $i = 0, 1, \ldots, n$. Let $\{U_{i\beta}\}$, $\beta \in B_i$ be a cover of A by A-open sets with $\overline{U}_{i\beta} \subseteq G_{i\beta}$. Then for $i = 0, 1, \ldots, n$, $A_i \overset{\mathrm{def}}{\equiv} \bigcup\{\overline{U}_{i\beta} \mid \beta \in B_i\}$ is a subset of G_i and hence contractible in M, and $A = \cup A_i$. Finally, since the $\overline{U}_{i\beta}$ are closed in A and locally finite, each A_i is closed in A and hence in M, so $\mathrm{cat}(A, M) \leq n + 1$. ∎

9.2.11. Proposition. $\mathrm{cat}(M) \geq \mathrm{cuplong}(M) + 1$, *provided M is connected.*

PROOF. Cf. [BG].

The topological invariant $\mathrm{cuplong}(M)$ is defined as the largest integer n such that, for some field F, there exist cohomology classes $\gamma_i \in H^{k_i}(M, F)$, $i = 1, \ldots, n$, with positive degrees k_i, such that $\gamma_1 \cup \ldots \cup \gamma_n \neq 0$. Thus

9.2.12. Proposition. *If M is an n-dimensional manifold and for some field F there is a cohomology class $\gamma \in H^1(M, F)$ such that $\gamma^n \neq 0$, then $\mathrm{cat}(M) = n + 1$.*

9.2.13. Corollary. *The n-dimensional torus T^n and the n-dimensional projective space $\boldsymbol{RP}^n$ both have category $n+1$.*

Recall that $\boldsymbol{RP}^n$ is the quotient space obtained by identifying pairs of antipodal points, x and $-x$, of the unit sphere $\boldsymbol{S}^n$ in $\boldsymbol{R}^{n+1}$. Thus a function on $\boldsymbol{RP}^n$ is the same as a function on $\boldsymbol{S}^n$ that is "even", in the sense that it takes the same value at antipodal points x and $-x$.

9.2.14. Proposition. *Any smooth even function on $\boldsymbol{S}^n$ has at least $n+1$ pairs of antipodal critical points.*

An important and interesting application of the latter proposition is an existence theorem for certain so-called "non-linear eigenvalue problems". Let $\Phi : \boldsymbol{R}^n \to \boldsymbol{R}^n$ be a smooth map. If $\lambda \in \boldsymbol{R}$ and $0 \neq x \in \boldsymbol{R}^n$ satisfy $\Phi(x) = \lambda x$, then x is called an eigenvector and λ an eigenvalue of Φ. In applications Φ is often of the form ∇F for some smooth real-valued function $F : \boldsymbol{R}^n \to \boldsymbol{R}$, and moreover F is usually even. For example if A is a self-adjoint linear operator on $\boldsymbol{R}^n$ and we define $F(x) = \frac{1}{2}\langle Ax, x\rangle$, then F is even, $\nabla F = A$, and we are led to the standard linear eigenvalue problem. Usually we look for eigenvectors on $S_r = \{x \in \boldsymbol{R}^n \mid \|x\| = r\}$, $r > 0$.

9.2.15. Proposition. *A point x of S_r is an eigenvector of ∇F if and only if x is a critical point of $F|S_r$. In particular if F is even then each S_r contains at least n pairs of antipodal eigenvectors for ∇F.*

PROOF. Define $G : \boldsymbol{R}^n \to \boldsymbol{R}$ by $G(x) = \frac{1}{2}\|x\|^2$, so $\nabla G_x = x$ and hence all positive real numbers are regular values of G. In particular $S_r = G^{-1}(\frac{1}{2}r^2)$ is a regular level of G. By the Lagrange Multiplier Theorem (cf. Appendix A) x in S_r is a critical point of $F|S_r$ if and only if $\nabla F_x = \lambda \nabla G_x = \lambda x$ for some real λ. ∎

9.3. The Second Deformation Theorem

We will call a closed interval $[a, b]$ of real numbers *non-critical* with respect to f if it contains no critical values of f. Recalling that the set $f(\mathcal{C})$ of critical values of f is closed in $\boldsymbol{R}$ it follows that for some $\epsilon > 0$ the interval $[a - \epsilon, b + \epsilon]$ is also non-critical. If $[a, b]$ is non-critical then the set $\mathcal{N} = f^{-1}([a, b])$ will be called a *non-critical neck* of M with respect to f. We will now prove the important fact that $\mathcal{N}$ has a very simple structure: namely it is diffeomorphic to $\mathcal{W} \times [a, b]$ where $\mathcal{W} = f^{-1}(b)$.

Since $(\nabla f)f = \|\nabla f\|^2$, on the set $M \setminus \mathcal{C}$ of regular points, where $\|\nabla f\| \neq 0$, the smooth vector field $Y = -\frac{1}{\|\nabla f\|^2}\nabla f$ satisfies $Yf = -1$. More generally if $F : \boldsymbol{R} \to \boldsymbol{R}$ is any smooth function vanishing in a neighborhood of $f(\mathcal{C})$, then $X = (F \circ f)Y$ is a smooth vector field on M that vanishes in a neighborhood of $\mathcal{C}$, and $Xf = -(F \circ f)$. We denote by Φ_t the flow on M generated by X. Let us choose $F : \boldsymbol{R} \to \boldsymbol{R}$ to be a smooth, non-negative function that is identically one on a neighborhood of $[a, b]$ and zero outside $[a - \epsilon/2, b + \epsilon/2]$.

9.3.1. Proposition. *With the above choice of F, the vector field X on M has bounded length and hence the flow Φ_t it generates is a one-parameter group of diffeomorphisms of M.*

PROOF. From the definition of Y it is clear that $\|Y\| = \frac{1}{\|\nabla f\|}$ so that $\|X\| = \frac{1}{\|\nabla f\|}|F \circ f|$. Since F has compact support it is bounded, and since $|F \circ f|$ vanishes outside $f^{-1}([a-\epsilon/2, b+\epsilon/2])$, it will suffice to show that $\frac{1}{\|\nabla f\|}$ is bounded on $f^{-1}([a - \epsilon/2, b + \epsilon/2])$, or equivalently that $\|\nabla f\|$ is bounded away from zero on $f^{-1}([a - \epsilon/2, b + \epsilon/2])$. But if not, then by Condition C we could find a sequence $\{x_n\}$ in $f^{-1}([a - \epsilon/2, b + \epsilon/2])$ converging to a critical point p of f. Then $f(p) \in [a - \epsilon/2, b + \epsilon/2]$, contrary to our assumption that the interval $[a - \epsilon, b + \epsilon]$ contains no critical values of f. ∎

Denote by $\gamma(t, c)$ the solution of the ordinary differential equation $\frac{d\gamma}{dt} = -F(\gamma)$ with initial value c. Since $\frac{d}{dt}(f \circ \Phi_t(x)) = X_{\Phi_t(x)}f = -F(f \circ \Phi_t(x))$, it follows that $f(\Phi_t(x)) = \gamma(t, f(x))$, and hence that $\Phi_t(f^{-1}(c)) = f^{-1}(\gamma(t, c))$. In particular *the flow Φ_t permutes the level sets of f.* From the definition of $\gamma(t, c)$ it follows that $\gamma(t, c) = c - t$ for $c \in [a, b]$ and $c - t \geq a$, while $\gamma(t, c) = c$ if $c > b + \epsilon$ or $c < a - \epsilon$. Since $\Phi_t(f^{-1}(c)) = f^{-1}(\gamma(t, c))$, it follows that if we write W for the b level of f, then Φ_{b-c} maps W diffeomorphically onto $f^{-1}(c)$ for all c in $[a, b]$ while, for all t, Φ_t is the identity outside the non-critical neck $f^{-1}([a - \epsilon/2, b + \epsilon/2])$.

In all that follows we shall denote by I the unit interval $[0, 1]$, and if $G : X \times I \to Y$ is any map, then for t in I we shall write $G_t : X \to Y$ for the map $G_t(x) = G(x, t)$. Recall that an *isotopy* of a smooth manifold M is a smooth map $G : M \times I \to M$ such that G_t is a diffeomorphism of M for all t in I and G_0 is the identity map of M. If A and B are subsets of M with $B \subseteq A$ then we say G deforms A onto B if $G_t(A) \subseteq A$ for all t and $G_1(A) = B$. And we say that G *fixes* a subset S of M if $G_t(x) = x$ for all (x, t) in $S \times I$. Finally if $f : M \to \boldsymbol{R}$ then we shall say G *pushes down the levels* of f if for all $c \in \boldsymbol{R}$ and $t \in I$ we have $G_t(f^{-1}(c)) = f^{-1}(c')$, where $c' \leq c$.

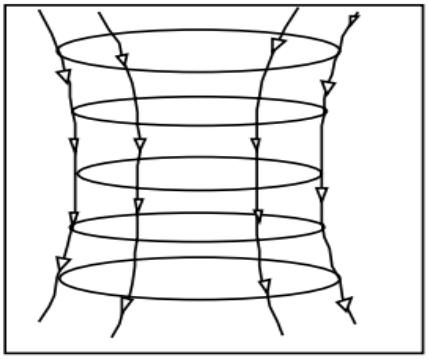
9.3.2. Second Deformation Theorem. *If the interval $[a, b]$ is non-critical for the smooth function $f : M \to \boldsymbol{R}$ then there is a deformation*

G of M that pushes down the levels of f and deforms M_b onto M_a. If $\epsilon > 0$ then we can assume G fixes the complement of $f^{-1}(a-\epsilon, b+\epsilon)$.

PROOF. Using the above notation we can define the deformation G by $G(x,t) = \Phi_{(b-a)t}(x)$. ∎

9.3.3. Non-Critical Neck Principle. *If $[a,b]$ is a non-critical interval of a smooth function $f : M \to \boldsymbol{R}$ and W is the b-level of f, then there is a diffeomorphism of the non-critical neck $\mathcal{N} = f^{-1}([a,b])$ with $W \times [a,b]$, under which the restriction of f to $\mathcal{N}$ corresponds to the projection of $W \times [a,b]$ onto $[a,b]$.*

PROOF. We define the map G of $W \times [a,b]$ into $\mathcal{N}$ by $G(x,t) = \Phi_{(b-t)}(x)$. Since $x \in W$, $f(x) = b$ and hence $f(G(x,t)) = (b-(b-t)) = t$. If $v \in TW_x$ then $DG(v, \frac{\partial}{\partial t}) = D\Phi_t(v) + X$. Now Φ_t maps W diffeomorphically onto $\tilde{W} = f^{-1}(t)$ and $TM_{\Phi_t(x)}$ is clearly spanned by the direct sum of $T\tilde{W}_{\Phi_t(x)}$ and $X_{\Phi_t(x)}$. It now follows easily from the Inverse Function Theorem that G is a diffeomorphism. ∎



A Non-Critical Neck.
The ellipses represent the level surfaces, and the vertical curves represent flowlines of the gradient flow.

The intuitive content of the above results deserves being emphasized. As a ranges over a non-critical interval the diffeomorphism type of the a-level of f, the diffeomorphism type of M_a, and even the diffeomorphism type of the the pair (M, M_a) is *constant*, that is it is independent of a. Now, as we shall see shortly, if we assume that our function f satisfies a certain simple, natural, and generic non-degeneracy assumption (namely, that it is what is called a Morse function) then the set of critical points of f is discrete. For simplicity let us assume for the moment that M is compact. Then the set of critical points is finite and of course the set of critical values of f is then *a fortiori* finite. Let us denote them, in

increasing order, by $c_1, c_2, \ldots, c_k$, and let us choose real numbers $a_0, a_1, \ldots, a_k$ with $a_0 < c_1 < a_1 < c_2 < \ldots < a_{k-1} < c_k < a_k$. Notice that c_1 must be the minimum of f, so that M_{a_0} is empty. And similarly c_k is the maximum of f so that M_{a_k} is all of M. More generally, by the above remark, the diffeomorphism type of M_{a_i} does not depend on the choice of a_i in the interval (c_i, c_{i+1}), so **we can think of a Morse function f as providing us with a specific method for "building up" our manifold M inductively in a finite number of discrete stages**, starting with the empty M_{a_0} and then, step by step, creating $M_{a_{i+1}}$ out of M_{a_i} by some "process" that takes place at the critical level c_{i+1}, finally ending up with M. Moreover the "process" that gives rise to the sudden changes in the topology of $f^{-1}(a)$ and of M_a as a crosses a critical value is not at all mysterious. From the point of view of M_a it is called "adding a handle", while from the point of view of the level $f^{-1}(a)$ it is just a "cobordism". From either point of view it can be analyzed fairly completely and is the basis for almost all classification theorems for manifolds.

9.4. Morse Functions

An elementary corollary of the Implicit Function Theorem is an important local canonical form theorem for a smooth function $f : M \to \boldsymbol{R}$ in the neighborhood of a regular point p; namely $f - f(p)$ is linear in a suitable coordinate chart centered at p. Equivalently, in this chart f coincides near p with its first order Taylor polynomial: $f(p) + df_p$.

But what if p is a critical point of f? Of course f will not necessarily be locally constant near p, but a natural conjecture is that, under some "generic" non-degeneracy assumption, we should again have a local canonical form for f near p, namely in a suitable local chart, (called a Morse Chart), f should coincide with its *second* order Taylor polynomial near p. That such a canonical form does exist generically is called The Morse Lemma and plays a fundamental role in Morse Theory. Before stating it precisely we review some standard linear algebra, adding some necessary infinite dimensional touches.

Let V be the model hilbert space for M, and let $\mathcal{A} : V \times V \to \boldsymbol{R}$ be a continuous, symmetric, bilinear form on V. We denote by $f_{\mathcal{A}} : V \to \boldsymbol{R}$ the associated homogeneous quadratic polynomial; $f_{\mathcal{A}}(x) = \frac{1}{2}\mathcal{A}(x, x)$. Now $\mathcal{A}$ defines a bounded linear map $\hat{\mathcal{A}} : V \to V^*$ by $\hat{\mathcal{A}}(x)(v) = \mathcal{A}(x, v)$. Using the canonical identification of V with V^* we can interpret $\hat{\mathcal{A}}$ as a bounded linear map $A : V \to V$, characterized by $\mathcal{A}(x, v) = \langle Ax, v\rangle$, so that $f_{\mathcal{A}}(x) = \frac{1}{2}\langle Ax, x\rangle$. Since $\mathcal{A}$ is symmetric, A is self-adjoint. The bilinear form $\mathcal{A}$ is called non-degenerate if $\hat{\mathcal{A}} : V \to V^*$ (or $A : V \to V$) is a linear isomorphism, i.e., if 0 does not belong to $\mathrm{Spec}(A)$, the spectrum of A. While we will be

concerned primarily with the non-degenerate case, for now we make a milder restriction. Let $V^0 = \ker(A)$. The dimension of V^0 is called the *nullity* of the quadratic form $f_{\mathcal{A}}$. There is a densely-defined self-adjoint linear map $A^{-1} : (V^0)^\perp \to (V^0)^\perp$. But of course A^{-1} may be unbounded. Since $\|A\| = \sup\{|\lambda| \mid \lambda \in \mathrm{Spec}(A)\}$ and $\mathrm{Spec}(A^{-1}) = (\mathrm{Spec}(A))^{-1}$, equivalently $\mathrm{Spec}(A)$ might have 0 as a limit point. It is *this* that we assume does not happen.

9.4.1. Assumption. *Zero is not a limit point of the Spectrum of A. Equivalently, if A does not have a bounded inverse then $V^0 = \ker(A)$ has positive dimension and A has a bounded inverse on $(V^0)^\perp$.*

(Of course in finite dimensions this is a vacuous assumption).

Choose $\epsilon > 0$ so that $(-\epsilon, \epsilon) \cap \mathrm{Spec}(A)$ contains at most zero. Let $p^+ : \boldsymbol{R} \to \boldsymbol{R}$ be a continuous function such that $p^+(x) = 1$ for $x \geq \epsilon$ and $p^+(x) = 0$ for $x \leq \frac{\epsilon}{2}$. And define $p^- : \boldsymbol{R} \to \boldsymbol{R}$ by $p^-(x) = p^+(-x)$. Finally let $p^0 : \boldsymbol{R} \to \boldsymbol{R}$ be continuous with $p^0(0) = 1$ and $p^0(x) = 0$ for $|x| \geq \frac{\epsilon}{2}$. Then using the functional calculus for self-adjoint operators [La], we can define three commuting orthogonal projections $P^+ = p^+(A)$, $P^0 = p^0(A)$, and $P^- = p^-(A)$ such that $P^+ + P^0 + P^-$ is the identity map of V. Clearly $V^0 = \mathrm{im}(P^0)$ and we define $V^+ = \mathrm{im}(P^+)$ and $V^- = \mathrm{im}(P^-)$, so that V is the orthogonal direct sum $V^+ \oplus V^0 \oplus V^-$. (In the finite dimensional case V^+ and V^- are respectively the direct sums of the positive and of the negative eigenspaces of A). The dimension of V^- is called the *index* of the quadratic form $f_{\mathcal{A}}$ and the dimension of V^+ is called its *coindex*.

Let $\varphi : \boldsymbol{R} \to \boldsymbol{R}$ be a continuous strictly positive function with $\varphi(\lambda) = \sqrt{\frac{2}{|\lambda|}}$ for $|\lambda| \geq \epsilon$, and $\varphi(0) = 1$. Then $\Phi = \varphi(A)$ is a self-adjoint linear diffeomorphism of V with itself. Since $\frac{1}{2}\varphi(\lambda)\lambda\,\varphi(\lambda) = \mathrm{sgn}(\lambda) = p^+(\lambda) - p^-(\lambda)$ for all λ in $\mathrm{Spec}(A)$, it follows that $\frac{1}{2}\Phi A \Phi = P^+ - P^-$, so that

$$\begin{aligned} f_{\mathcal{A}}(\Phi(x)) &= \tfrac{1}{2}\langle A\Phi x, \Phi x\rangle \\ &= \langle \tfrac{1}{2}\Phi A \Phi x, x\rangle \\ &= \langle P^+ x, x\rangle - \langle P^- x, x\rangle \\ &= \|P^+ x\|^2 - \|P^- x\|^2 \,. \end{aligned}$$

9.4.2. Proposition. *Let $A : V \to V$ be a bounded self-adjoint operator and $f_{\mathcal{A}} : V \to \boldsymbol{R}$ the homogeneous quadratic polynomial $f_{\mathcal{A}}(x) = \frac{1}{2}\langle Ax, x\rangle$. If 0 is not a limit point of Spec(A) then V has an orthogonal decomposition $V = V^+ \oplus V^0 \oplus V^-$ (with $V^0 = \ker(A)$) and a self-adjoint linear diffeomorphism $\Phi : V \approx V$ such that*

$$f_{\mathcal{A}}(\Phi(x)) = \|P^+(x)\|^2 - \|P^-(x)\|^2 \,,$$

where P^+ and P^- are the orthogonal projections of V on V^+ and V^- respectively.

We now return to our smooth function $f : M \to \boldsymbol{R}$. (For the moment we do not need the Riemannian structure on M.)

We associate to each pair of smooth vector fields X and Y on M, a smooth real valued function $B(X,Y) = X(Yf)$. We note that $B(X,Y)(p)$ is just the directional derivative of Yf at p in the direction X_p, so in particular its value depends on X only through its value, X_p, at p. Now if p is a critical point of f then $B(X,Y)(p) - B(Y,X)(p) = X_p(Yf) - Y_p(Xf) = [X,Y]_p(f) = df_p([X,Y]) = 0$. It follows that in *this* case $B(X,Y)(p) = B(Y,X)(p)$ also only depends on Y through its value, Y_p, at p. This proves:

9.4.3. Hessian Theorem. *If p is a critical point of a smooth real valued function $f : M \to \boldsymbol{R}$ then there is a uniquely determined symmetric bilinear form $\mathrm{Hess}(f)_p$ on TM_p such that, for any two smooth vector fields X and Y on M, $\mathrm{Hess}(f)_p(X_p, Y_p) = X_p(Yf)$.*

We call $\mathrm{Hess}(f)_p$ the *Hessian bilinear form* associated to f at the critical point p, and we will also denote the related Hessian quadratic form by $\mathrm{Hess}(f)_p$ (i.e., $\mathrm{Hess}(f)_p(v) = \frac{1}{2}\mathrm{Hess}(f)_p(v,v)$). (Given a local coordinate system $x_1, \ldots, x_n$ for M at p, evaluating $\mathrm{Hess}(f)_p(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j})$ we see that the matrix of $\mathrm{Hess}(f)_p$ is just the classical "Hessian matrix" of second partial derivatives of f.)

We shall say that the critical point p is *non-degenerate* if $\mathrm{Hess}(f)_p$ is non-degenerate, and we define the nullity, index, and coindex of p to be respectively the nullity, index, and coindex of $\mathrm{Hess}(f)_p$. Finally, f is called a *Morse Function* if all of its critical points are non-degenerate.

Using the Riemannian structure of M we have a self-adjoint operator $\mathrm{hess}(f)_p$, defined on TM_p, and characterized by $\langle \mathrm{hess}(f)_p(X), Y\rangle = \mathrm{Hess}(f)_p(X,Y)$. Then the nullity of p is the dimension of the kernel of $\mathrm{hess}(f)_p$, p is a non-degenerate critical point of f when $\mathrm{hess}(f)_p$ has a bounded inverse, and, in finite dimensions, the index of p is the sum of the dimensions of eigenspaces of $\mathrm{hess}(f)_p$ corresponding to negative eigenvalues.

Let ∇ denote any connection on TM (*not* necessarily the Levi-Civita connection). Then ∇ induces a family of associated connections on all the tensor bundle over M, characterized by the fact that covariant differentiation commutes with contraction and the "product rule" holds. The latter means that, for example given vector fields X and Y on M,

$$\nabla_X(Y \otimes df) = \nabla_X(Y) \otimes df + Y \otimes \nabla_X(df)\,.$$

Contracting the latter gives:

$$X(Yf) = df(\nabla_X Y) + i_Y i_X(\nabla df)\,.$$

If we define $\mathrm{Hess}^{\nabla}(f)$ to be ∇df then we can rewrite this equation as

$$\mathrm{Hess}^{\nabla}(f)(X,Y) = X(Yf) - df(\nabla_X Y)\,.$$

This has two interesting consequences. First, interchanging X and Y and subtracting gives:

$$\mathrm{Hess}^{\nabla}(f)(X,Y) - \mathrm{Hess}^{\nabla}(f)(Y,X) = df(\tau^{\nabla}(X,Y))\,,$$

where τ^{∇} is the torsion tensor of ∇. Thus if ∇ is a symmetric connection (i.e. $\tau^{\nabla} = 0$), as is the Levi-Civita connection, then $\mathrm{Hess}^{\nabla}(f)$ is a *symmetric* covariant two-tensor field on M. And in any case, at a critical point p of f, where $df_p = 0$, we have:

$$\mathrm{Hess}^{\nabla}(f)(X_p,Y_p) = X_p(Yf) = \mathrm{Hess}(f)_p(X_p,Y_p).$$

9.4.4. Proposition. *If ∇ denotes the Levi-Civita connection for M, then $\mathrm{Hess}^{\nabla}(f) \stackrel{\mathrm{def}}{\equiv} \nabla df$ is a symmetric two-tensor field on M that at each critical point p of f agrees with $\mathrm{Hess}(f)_p$.*

9.4.5. Corollary. *$\mathrm{hess}^{\nabla}(f) \stackrel{\mathrm{def}}{\equiv} \nabla(\nabla f)$ is a field of self adjoint operators on M that at each critical point p of f agrees with $\mathrm{hess}(f)_p$.*

There is yet another interpretation of $\mathrm{Hess}(f)_p$ that is often useful. The differential df of f is a section of T^*M that vanishes at p, so *its* differential, $D(df)_p$, is a linear map of TM_p into $T(T^*M)_{0_p}$ (where 0_p denotes the zero element of T^*M_p). Now $T(T^*M)_{0_p}$ is canonically the direct sum of two subspaces; the "vertical" subspace, tangent to the fiber T^*M_p, which we identify with T^*M_p, and the "horizontal" space, tangent to the zero section, which we identify with TM_p. If we compose $D(df)_p$ with the projection onto the vertical space we get a linear map $TM_p \to T^*M_p$ that, under the natural isomorphism of bilinear maps $V \times V \to \boldsymbol{R}$ with linear maps $V \to V^*$, is easily seen to correspond to $\mathrm{Hess}(f)_p$. With this alternate definition of $\mathrm{Hess}(f)_p$, the condition for p to be non-degenerate is that $\mathrm{Hess}(f)_p$ map TM_p isomorphically onto T^*M_p.

It is clear that at the critical point p of f, $\mathrm{Hess}(f)_p$ determines the second order Taylor polynomial of f at p. But what is less obvious is that, at least in the non-degenerate case, f "looks like" its second order Taylor polynomial near p, a fact known as the Morse Lemma.

Let us put $V = T^*M_p$, $A = \mathrm{hess}(f)_p$, and let V^+, V^0, and V^- be as above, i.e., the maximal subspaces of V on which A is positive, zero, and negative. Recall that a chart for M centered at p is a diffeomorphism Φ of a neighborhood

O of 0 in V onto a neighborhood U of p in M with $\Phi(0) = p$. We call Φ a *Morse chart of the first kind at* p if $f(\Phi(v)) - f(p) = \mathrm{Hess}(v) = \frac{1}{2}\langle Av, v\rangle$. And Φ is called a *Morse chart of the second kind at* p (or simply a *Morse chart at* p) if $f(\Phi(v)) - f(p) = \|P^+v\|^2 - \|P^-v\|^2$, where P^+ and P^- are the orthogonal projections on V^+ and V^-. It is clear that a Morse chart of the second kind *is* a Morse chart of the first kind. Moreover, by the proposition at the beginning of this section, if a Morse chart of the first kind exists at p, then so does a Morse chart of the second kind. In this case we shall say simply that Morse charts exist at p.

9.4.6. Morse Lemma. *If p is a non-degenerate critical point of a smooth function $f : M \to \boldsymbol{R}$ then Morse charts exist at p.*

PROOF. Since the theorem is local we can take M to be V and assume p is the origin 0. Also without loss of generality we can assume $f(0) = 0$. We must show that, after a smooth change of coordinates φ, f has the form $f(x) = \frac{1}{2}\langle Ax, x\rangle$ in a neighborhood $\mathcal{O}$ of 0. Since $df_p = 0$, by Taylor's Theorem with remainder we can write f near 0 in the form $f(x) = \frac{1}{2}\langle A(x)x, x\rangle$, where $x \mapsto A(x)$ is a smooth map of $\mathcal{O}$ into the self-adjoint operators on V. Since $A(0) = A = \mathrm{hess}(f)_0$ is non-singular, $A(x)$ is also non-singular in a neighborhood of 0, which we can assume is $\mathcal{O}$. We define a smooth map B of $\mathcal{O}$ into the group $\boldsymbol{GL}(V)$ of invertible operators on V by $B(x) = A(x)^{-1}A(0)$, and note that $B(0)$ is I, the identity map of V. Now a square root function is defined in the neighborhood of I by a convergent power series with real coefficients, so we can define a smooth map C of $\mathcal{O}$ into $\boldsymbol{GL}(V)$ by $C(x) = \sqrt{B(x)}$. Since $A(0)$ and $A(x)$ are self-adjoint it is immediate from the definition of B that $B(x)^*A(x) = A(x)B(x)$. This same relation then holds if we replace $B(x)$ by any polynomial in $B(x)$, and hence if we replace $B(x)$ by $C(x)$ which is a limit of such polynomials. Thus

$$C(x)^*A(x)C(x) = A(x)C(x)^2 = A(x)B(x) = A(0)$$

or $A(x) = C_1(x)^*AC_1(x)$, where we have put $C_1(x) = C(x)^{-1}$. If we define a smooth map φ of $\mathcal{O}$ into V by $\varphi(x) = C_1(x)x$, then $f(x) = \langle C_1(x)^*AC_1(x)x, x\rangle = \langle A\varphi(x), \varphi(x)\rangle$, so it remains only to check that φ is a valid change of coordinates at 0, i.e., that $D\varphi_0$ is invertible. But $D\varphi_x = C_1(x) + D(C_1)_x(x)$, so in particular $D\varphi_0 = C_1(0) = I$. ∎

9.4.7. Corollary 1. *A non-degenerate critical point of a smooth function $f : M \to \boldsymbol{R}$ is isolated in the set $\mathcal{C}$ of all critical points of f. In particular if f is a Morse function then $\mathcal{C}$ is a discrete subset of M.*

PROOF. Maintaining the assumptions and notations introduced in the proof of the Morse Lemma we have $f(x) = \frac{1}{2}\langle Ax, x\rangle$ in a neighborhood $\mathcal{O}$ of

0, and hence $df_x = Ax$ for x in $\mathcal{O}$. Since A is invertible, df_x does not vanish in $\mathcal{O}$ except at 0. ∎

9.4.8. Corollary 2. *If a Morse function $f : M \to \boldsymbol{R}$ satisfies Condition C then for any finite interval $[a, b]$ of real numbers there are only a finite number of critical points p of f with $f(p) \in [a, b]$. In particular the set $\mathcal{C}$ of critical values of f is a discrete subset of $\boldsymbol{R}$.*

PROOF. We saw earlier that Condition C implies that f restricted to $\mathcal{C}$ is proper, so the set of critical points p of f with $f(p) \in [a, b]$ is compact. But by Corollary 1 it is also discrete. ∎

Since we are going to be focusing our attention on Morse functions, a basic question to answer is, whether they necessarily exist, and if so how rare or common are they. Fortunately, at least in the finite dimensional case this question has an easy and satisfactory answer; Morse functions form an open, dense subspace in the C^2 topology of the space $C^2(M, \boldsymbol{R})$ of all C^2 real valued functions on M. The easiest, but not the most elementary, approach to this problem is through Thom's transversality theory. Let ξ be a smooth vector bundle of fiber dimension m over a smooth n-manifold M . Recall that if s_1 and s_2 are two C^1 sections of ξ with $s_1(p) = s_2(p) = v$, then we say that these sections have *transversal intersection* (or are transversal) at p if, when considered as submanifolds of M, their tangent spaces at v span the entire tangent space to ξ at v. We say s_1 and s_2 are transversal if they have transversal intersection wherever they meet. Since each section has dimension n, and ξ has dimension $m + n$ the condition for transversality is that the intersection of their tangent spaces at v should have dimension $(n+n)-(n+m) = n-m$. So if ξ has fiber dimension n then this intersection should have dimension zero and, since Ds_i maps TM_p isomorphically onto the tangent space to s_i at v, this just means that $Ds_1(u) \neq Ds_2(u)$ for $u \neq 0$ in TM_p. In particular for ξ the cotangent bundle T^*M, a section s vanishing at p is transversal to the zero section at p if and only if $\mathrm{im}(Ds)$ is disjoint from the horizontal space at p, or equivalently if and only if the composition of Ds with projection onto the vertical subspace, T^*M_p is an isomorphism. Recalling our alternate interpretation of $\mathrm{Hess}(f)_p$ above we see:

9.4.9. Lemma. *The critical point p of $f : M \to \boldsymbol{R}$ is non-degenerate if and only if df is transversal to the zero section of T^*M at p. Thus f is a Morse function if and only if df is transversal to the zero section.*

Thom's k-jet transversality theorem [Hi, p.80] states that if s_0 is a C^{k+1} section of a smooth vector bundle ξ over a compact manifold M and $J^k\xi$ is the corresponding bundle of k-jets of sections of ξ, then in the space $C^{k+1}(\xi)$ of

C^{k+1} sections of ξ with the C^{k+1} topology, the set of sections s whose k-jet extension $j_k s$ is transversal to $j_k s_0$ is open and dense. If we take for ξ the trivial bundle $M \times \boldsymbol{R}$ then a section becomes just a real valued function, and we can identify $J^1\xi$ with T^*M so that j_1f is just df. Finally, taking $k = 1$ and letting s_0 be the zero section, Thom's theorem together with the above lemma gives the desired conclusion, that Morse functions are open and dense in $C^2(M, \boldsymbol{R})$.

As a by-product of the section on the Morse Theory of submanifolds of Euclidean space, we will find a much more elementary approach to this question, that gives almost as complete an answer.

9.5. Passing a Critical Level

We now return to our basic problem of Morse Theory; reconstructing the manifold M from knowledge about the critical point structure of the function $f : M \to \boldsymbol{R}$.

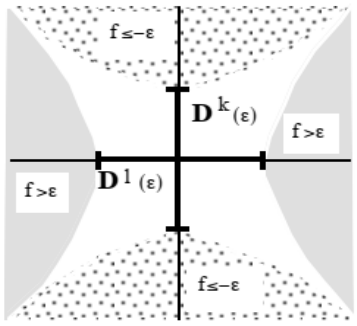
To get a satisfactory theory we will supplement the assumptions (a), (b), and (c) of the Introduction with the following additional assumption:

(d) f is a Morse function.

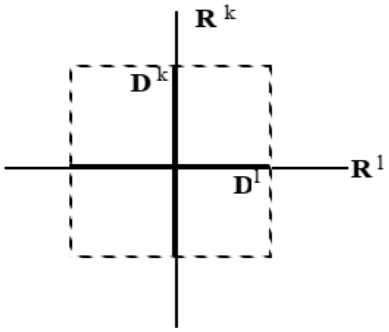
As we saw in the preceding section this implies that for any finite interval $[a, b]$ there are only a finite number of critical points p of f with $f(p)$ in $[a, b]$, and hence only a finite number of critical values of f in $[a, b]$.

Our goal is to describe how M_α changes as α changes from one non-critical value a to another b. Now, by the Second Deformation Theorem, the diffeomorphism type of M_α is constant for α in a non-critical interval of f, hence we can easily reduce our problem to the case that there is a single critical value c in (a, b), and without loss of generality we can assume that $c = 0$. So what we want to see is how to build M_ϵ out of $M_{-\epsilon}$ when 0 is the unique critical value of f in $[-\epsilon, \epsilon]$. In general there could be a finite number of critical points $p_1, \ldots, p_k$ at the level 0, and eventually we shall consider that case explicitly. But the discussion will be greatly simplified (with no essential loss of generality) by assuming at first that there is a *unique* critical point p at the level 0. We will let k and l denote the index and coindex of f at p and $n = k + l$ the dimension of M. If $n = \infty$ then one or both of k and l will also be infinite; nevertheless we shall write $\boldsymbol{R}^k$, and $\boldsymbol{R}^l$ for the Hilbert spaces of dimension k and l, and $\boldsymbol{R}^n = \boldsymbol{R}^l \times \boldsymbol{R}^k$.

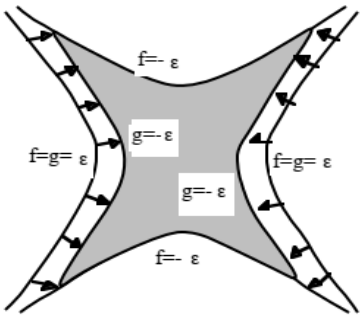
As in all good construction projects we will proceed in stages, and start with some blueprints before filling in the precise mathematical details.



- We will denote by $D^k(\epsilon)$, and $D^l(\epsilon)$ the disks of radius $\sqrt{\epsilon}$ centered at the origin in $\boldsymbol{R}^k$ and $\boldsymbol{R}^l$ respectively. We will write D^k and D^l for the unit disks. The product $D^l \times D^k$, attached in a certain way to $M_{-\epsilon}$ will be called a *handle of index* k.

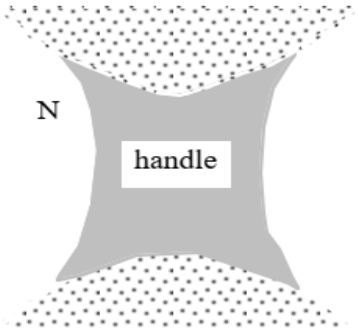


- We will construct a smooth submanifold N of M with $M_{-\epsilon} \subseteq N \subseteq M_{\epsilon}$. Namely, $N = M_{-\epsilon}(g) = \{x \in M | g(x) \leq -\epsilon\}$, where $g : M \to \boldsymbol{R}$ is a certain smooth function that agrees with f where f is greater than ϵ (so that $M_\epsilon = M_\epsilon(f) = M_\epsilon(g)$). Moreover the interval $[-\epsilon, \epsilon]$ is non-critical for g, so by the Second Deformation Theorem there is an isotopy of M that deforms $M_\epsilon = M_\epsilon(g)$ onto $M_{-\epsilon}(g) = N$.



- The manifold N has a second description. Namely, N is an adjunction space that consists of $M_{-\epsilon}$ together with a subset $\mathcal{H}$, (called the "handle")

that is diffeomorphic to the above product of disks and is glued onto $\partial M_{-\epsilon}$, the boundary of $M_{-\epsilon}$, by a diffeomorphism of $\partial D^l \times D^k$ onto $\mathcal{H} \cap \partial M_{-\epsilon}$.



Thus **when we pass a critical level $f^{-1}(c)$ of f that contains a single non-degenerate critical point of index k, $M_{c+\epsilon}$ is obtained from $M_{c-\epsilon}$ by attaching to the latter a handle of index k.**

Now for the details. We identify a neighborhood $\mathcal{O}$ of p in M with a neighborhood of the origin in $\boldsymbol{R}^n = \boldsymbol{R}^l \times \boldsymbol{R}^k$, using a Morse chart (of the second kind). We will regard a point of $\mathcal{O}$ as a pair (x, y), where $x \in \boldsymbol{R}^l$ and $y \in \boldsymbol{R}^k$. We suppose ϵ is chosen small enough that 0 is the only critical level of f in $[-2\epsilon, 2\epsilon]$, or equivalently so that p is the only critical point of f with $|f(p)| \leq 2\epsilon$. We can also assume ϵ so small that the closed disk of radius $2\sqrt{\epsilon}$ in $\boldsymbol{R}^{k+l}$ is included in $\mathcal{O}$. Thus f is given in $\mathcal{O}$ by $f(x, y) = \|x\|^2 - \|y\|^2$. Choose a smooth, non-increasing function $\lambda : \boldsymbol{R} \to \boldsymbol{R}$ that is identically 1 on $t \leq \frac{1}{2}$, positive on $t < 1$, and zero for $t \geq 1$. Then the function g is defined in $\mathcal{O}$ by $g(x, y) = f(x, y) - \frac{3\epsilon}{2}\lambda(\|x\|^2/\epsilon)$.

9.5.1. Lemma. *The function g can be extended to be a smooth function $g : M \to \boldsymbol{R}$ that is everywhere less than f and agrees with f wherever $f \geq \epsilon$ and also, outside $\mathcal{O}$, wherever $f \geq -2\epsilon$. In particular $M_\epsilon(g) = M_\epsilon(f)$.*

PROOF. Suppose (x, y) in $\mathcal{O}$, $f(x, y) \geq -2\epsilon$, and $g(x, y) \neq f(x, y)$. Then $\lambda(\|x\|^2/\epsilon) \neq 0$ and hence $\|x\|^2 < \epsilon$. It follows that $\|x\|^2 + \|y\|^2 = 2\|x\|^2 - f(x, y) < 2\epsilon + 2\epsilon$, i.e., (x, y) is inside the disk of radius $2\sqrt{\epsilon}$. Recalling that the latter disk is interior to $\mathcal{O}$ it follows that if we extend g to the remainder of $f^{-1}([-2\epsilon, \infty))$ by making it equal f outside $\mathcal{O}$, then it will be smooth. Since $g \leq f$ everywhere on the closed set $f^{-1}([-2\epsilon, \infty))$ we can now further extend it to a function $g : M \to \boldsymbol{R}$ satisfying the same inequality on all of M. If $f(q) \geq \epsilon$ then either q is not in $\mathcal{O}$, so $g(q) = f(q)$ by definition of g, or else $q = (x, y)$ is in $\mathcal{O}$, in which case $\|x\|^2 \geq f(x, y) \geq \epsilon$, so $\lambda(\|x\|^2/\epsilon) = 0$, and again $g(q) = f(q)$. ∎

9.5.2. Lemma. *For the function g, extended as above, the interval $[-\epsilon, \epsilon]$ is a non-critical interval. (In fact p is the only critical point of g in $\mathcal{S} = g^{-1}([-2\epsilon, \epsilon])$, and $g(p) = -\frac{3\epsilon}{2}$).*

PROOF. Recalling that $f \geq g$ everywhere, and that, outside $\mathcal{O}$, $f = g$ wherever $f \geq -2\epsilon$, it follows that $f = g$ on $\mathcal{S}\backslash\mathcal{O}$. Thus any critical point of g in $\mathcal{S}\backslash\mathcal{O}$ would also be a critical point of f in $f^{-1}[-2\epsilon, 2\epsilon]$. But by our choice of ϵ, the only such critical point is p, which belongs to $\mathcal{O}$. Thus it will suffice to show that, inside of $\mathcal{O}$, the only critical point of g is $p = (0, 0)$, where $g(x, y) = f(x, y) - \frac{3\epsilon}{2}\lambda(\frac{\|x\|^2}{\epsilon})$ is clearly equal to $-\frac{3\epsilon}{2}\lambda(0) = -\frac{3\epsilon}{2} < -\epsilon$. But in $\mathcal{O}$, $dg = (2 - 3\lambda'(\frac{\|x\|^2}{\epsilon}))x\,dx + 2y\,dy$ and, since λ' is a non-positive function, this vanishes only at the origin. ∎

Now it is time to make the concept of "attaching a handle" mathematically precise.

9.5.3. Definition. Let P and N be smooth manifolds with boundary, having the same dimension $n = k + l$, and with P a smooth submanifold of N. Let α be a homeomorphism of $D^l \times D^k$ onto a closed subset $\mathcal{H}$ of N. We shall say that N *arises from* P *by attaching a handle of index* k *and coindex* l *(or a handle of type* (k, l)*) with attaching map* α if:

(1) $N = P \cup \mathcal{H}$,

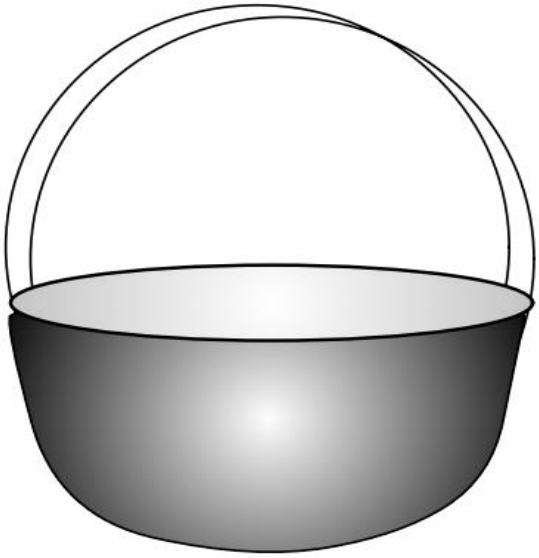
(2) $\alpha|(D^l \times \boldsymbol{S}^{k-1})$ is a diffeomorphism onto $\mathcal{H} \cap \partial P$,

(3) $\alpha|(D^l \times \mathring{D}^k)$ is a diffeomorphism onto $N\backslash P$.

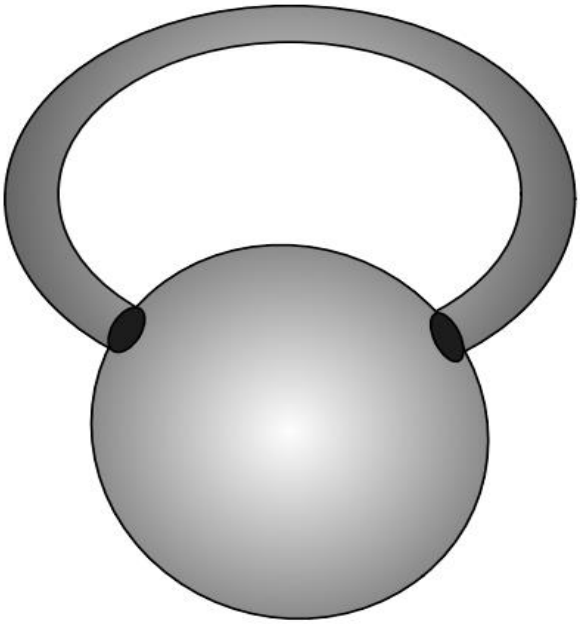
Here $\mathring{D}^k$ denotes the interior of the k-disk. Of course $D^l \times D^k$ is not a *smooth* manifold (it has a "corner" along $\partial D^l \times \partial D^k$), but both $D^l \times \boldsymbol{S}^{k-1}$ and $D^l \times \mathring{D}^k$ *are* smooth manifolds with boundary.

Note that if $k < \infty$, (so, in particular, if $n < \infty$) then $l = n - k$ is determined by k, so in this case it is common to speak simply of attaching a handle of index k.

The following example (with $k = l = 1$) is a good one to keep in mind: P is the lower hemisphere of the standard $\boldsymbol{S}^2$ in $\boldsymbol{R}^3$, (think of it as a basket), and $\mathcal{H}$, the handle of the basket, is a tubular neighborhood of that part of a great circle lying in the upper hemisphere. Of course, where the handle and basket meet, the sharp corner should be smoothed.



Another example that can be easily visualized ($k = 1, l = 2$) is the "solid torus" formed by gluing a 1-handle $D^2 \times D^1$ to the unit disk in $\boldsymbol{R}^3$ (a bowling ball with a carrying handle).



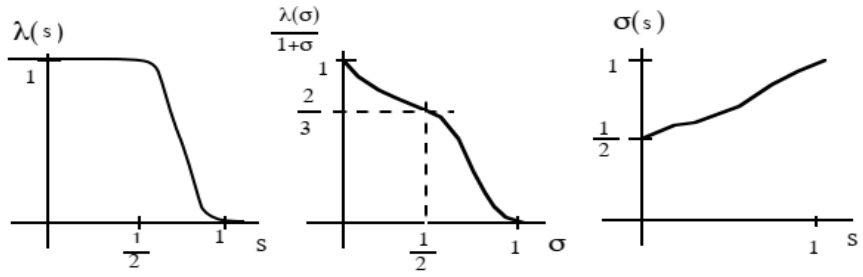
Recall that, in the case of interest to us, $P = M_{-\epsilon}(f)$, $N = M_{-\epsilon}(g)$, and we define the handle $\mathcal{H}$ to be the closure of the set of $(x, y) \in \mathcal{O}$ such that $f(x, y) > -\epsilon$ and $g(x, y) < -\epsilon$. Then recalling that, outside of $\mathcal{O}$, f and g agree where $f \geq -\epsilon$, it follows from the definition of N as $M_{-\epsilon}(g)$ that $N = M_{-\epsilon}(f) \cup \mathcal{H}$. What remains then is to define the homeomorphism α of $D^l \times D^k$ onto $\mathcal{H}$, and prove the properties (2) and (3) of the above definition.

We define α by the explicit formula:

$$\alpha(x,y) = (\epsilon\sigma(\|y\|^2))^{\frac{1}{2}}x + (\epsilon\sigma(\|y\|^2)\|x\|^2 + \epsilon)^{\frac{1}{2}}y$$

where $\sigma : I \to I$ is defined by taking $\sigma(s)$ to be the unique solution of the equation

$$\frac{\lambda(\sigma)}{(1+\sigma)} = \frac{2}{3}(1-s).$$



Clearly $\lambda(\sigma)/(1+\sigma)$ is a smooth function on I with a strictly negative derivative on $[0,1)$. It is then an easy consequence of the Inverse Function Theorem that σ is smooth on $[0,1)$ and strictly increasing on I. Moreover $\sigma(0) = 1/2$ and $\sigma(1) = 1$.

9.5.4. Lemma. *Define real valued functions F and G on $\boldsymbol{R}^2$ by $F(x,y) = x^2 - y^2$ and $G(x,y) = F(x,y) - (\frac{3\epsilon}{2})\lambda(\frac{x^2}{\epsilon})$. (so that, in $\mathcal{O}$, $f(u,v) = F(\|u\|,\|v\|)$ and $g(u,v) = G(\|u\|,\|v\|)$). Then in the region $\mathcal{U}$ that is the closure of the set $\{(x,y) \in \boldsymbol{R}^2 \mid F(x,y) > -\epsilon$ and $G(x,y) < -\epsilon\}$ we have*

$$x^2 \leq \epsilon\sigma\left(\frac{y^2}{\epsilon + x^2}\right).$$

PROOF. We must show that the function $h : \boldsymbol{R}^2 \to \boldsymbol{R}$, defined by $h(x,y) = x^2 - \epsilon\sigma(y^2/(\epsilon + x^2))$, is everywhere non-positive in $\mathcal{U}$. Now for fixed y, h is clearly only critical for $x = 0$, where it has a minimum. Hence h must assume its maximum on the boundary of $\mathcal{U}$ and it will suffice to show that everywhere on this boundary it is less than or equal to zero. But the boundary of $\mathcal{U}$ is the closure of the union of the two curves $\partial_1 = \{(x,y) | F(x,y) = -\epsilon, \quad G(x,y) < -\epsilon\}$ and $\partial_2 = \{(x,y) | F(x,y) > -\epsilon, \quad G(x,y) = -\epsilon\}$ and we will show that $h \leq 0$ both on ∂_1 and on ∂_2.

Indeed on ∂_1, since $G < F$, $(-3\epsilon/2)\lambda(x^2/\epsilon) < 0$ so $\lambda(x^2/\epsilon) > 0$, which implies $x^2/\epsilon < 1$ or $x^2 < \epsilon$. On the other hand, since $x^2 - y^2 = F(x,y) = -\epsilon$, $y^2/(\epsilon + x^2) = 1$ so $\sigma(y^2/(\epsilon + x^2)) = 1$ and hence $h(x,y) = x^2 - \epsilon < 0$.

On ∂_2 we again have $G < F$, so as above $x^2/\epsilon < 1$. The equality $G(x,y) = -\epsilon$ gives

$$\frac{y^2}{\epsilon + x^2} = 1 - \left(\frac{3}{2}\right)\frac{\lambda(x^2/\epsilon)}{(1 + x^2/\epsilon)}.$$

Now $x^2/\epsilon < 1/2$ would imply both $\lambda(x^2/\epsilon) = 1$ and $1 + x^2/\epsilon < \frac{3}{2}$, so the displayed inequality would give the impossible $y^2/(\epsilon + x^2) < 0$. Thus $1/2 \leq x^2/\epsilon < 1$, so x^2/ϵ is in the range of σ, say $x^2/\epsilon = \sigma(\rho)$. Then by definition of σ,

$$\frac{y^2}{\epsilon + x^2} = 1 - \left(\frac{3}{2}\right)\frac{\lambda(\sigma(\rho))}{(1 + \sigma(\rho)} = 1 - \left(\frac{3}{2}\right)\left(\frac{2}{3}\right)(1 - \rho) = \rho,$$

and hence

$$h(x,y) = x^2 - \epsilon\sigma\left(\frac{y^2}{\epsilon + x^2}\right) = \epsilon\sigma(\rho) - \epsilon\sigma(\rho) = 0,$$

so $h \leq 0$ on ∂_2 as well. ∎

The remainder of the proof is now straightforward. We will leave to the reader the easy verifications that if $(u,v) = \alpha(x,y)$ then $f(u,v) \geq -\epsilon$ and $g(u,v) \leq -\epsilon$, so that α maps $D^l \times D^k$ into $\mathcal{H}$.

Conversely, suppose that (u,v) belongs to $\mathcal{H}$. Then $F(\|u\|, \|v\|) = \|u\|^2 - \|v\|^2 \geq -\epsilon$ and $G(\|u\|, \|v\|) \leq -\epsilon$. Thus $\|v\|^2/(\epsilon + \|u\|^2) \leq 1$, so $y = (\epsilon + \|u\|^2)^{-1/2}v \in D^k$. Also $\sigma(\|v\|^2/(\epsilon + \|u\|^2))$ is well defined, and by the preceding Lemma $\|u\|^2/\epsilon\sigma(\|v\|^2/(\epsilon + \|u\|^2)) \leq 1$ so that $x = (\epsilon\sigma(\|v\|^2/(\epsilon + \|u\|^2)))^{-1/2}u \in D^l$. It follows that $\beta(u,v) = (x,y)$ defines a map $\beta : \mathcal{H} \to D^l \times D^k$, and it is elementary to check that α and β are mutually inverse maps, so that α is a homeomorphism of $D^l \times D^k$ onto $\mathcal{H}$. Since σ is smooth and has positive derivative in $[0,1)$ it follows that α is a diffeomorphism on $D^l \times \mathring{D}^k$. On $D^l \times \boldsymbol{S}^{k-1}$ the map α reduces to

$$\alpha(x,y) = \epsilon^{1/2}x + (\epsilon(\|x\|^2 + 1))^{1/2}y$$

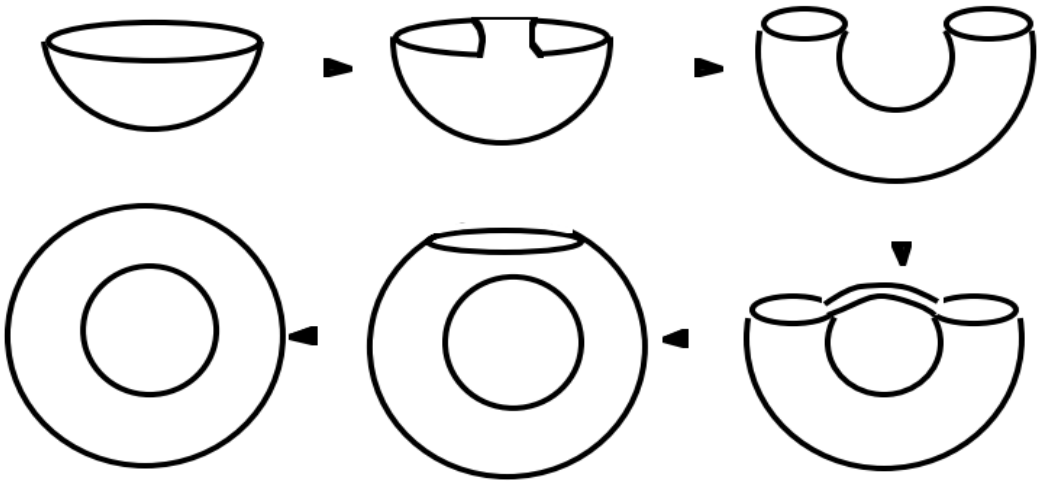
which is clearly a diffeomorphism onto $\mathcal{H} \cap \partial M_{-\epsilon}$. This completes the proof that $M_{c+\epsilon}$ is diffeomorphic to $M_{c-\epsilon}$ with a handle of index k attached.

Finally, let us see what modifications are necessary when we pass a critical level that contains more than one critical point. First note that the whole process of adjoining a handle to $M_{-\epsilon}$ took place in a small neighborhood of p (the domain of a Morse chart at p). Thus if we have several critical points at the same level then we can carry out the same attaching process independently in disjoint neighborhoods of these various critical points.

9.5.5. Definition. Suppose we have a sequence of smooth manifolds $N = N_0, N_1, \ldots, N_s = M$ such that N_{i+1} arises from N_i by attaching a handle of type (k_i, l_i) with attaching map α_i. If the images of the α_i are disjoint then we shall say that *M arises from N by the disjoint attachment of handles of type $((k_1, l_1), \ldots, (k_s, l_s))$ with attaching maps $(\alpha_1, \ldots, \alpha_s)$.*

9.5.6. Theorem. *Let f be a Morse function that is bounded below and satisfies Condition C on a complete Riemannian manifold M. Suppose $c \in (a, b)$ is the only critical value of f in the interval $[a, b]$, and that $p_1, \ldots, p_s$ are all the critical points of f at the level c. Let p_i have index k_i and coindex l_i. Then M_b arises from M_a by the disjoint attachment of handles of type $((k_1, l_1), \ldots, (k_s, l_s))$.*

Let us return to our example of the height function on the torus. That is, we take M to be the surface of revolution in $\boldsymbol{R}^3$, formed by rotating the circle $x^2 + (y-2)^2 = 1$ about the x-axis. The function $f : M \to \boldsymbol{R}$ defined by $f(x, y, z) = z$ is a Morse function with critical points at $(0, 0, -3)$, $(0, 0, -1)$, $(0, 0, 1)$, and $(0, 0, 3)$, and with respective indices 0,1,1,2. Here is a diagram showing the sequence of steps in the gradual building up of this torus, starting with a disk (or 0-handle), adding two consecutive 1-handles, and finally completing the torus with a 2-handle.



9.6. Morse Theory of Submanifolds

As we shall now see, there is a more detailed Morse theory for submanifolds of a Euclidean space. In this section proofs of theorems will often be merely

sketched or omitted entirely, since details can be found in the first two sections of Chapter 4.

We assume in what follows that M is a compact, smooth n-manifold smoothly embedded in $\boldsymbol{R}^N$, and we let k denote the codimension of the embedding. (We recall that, by a classical theorem of H. Whitney, any abstractly given compact (or even second countable) n-manifold can *always* be embedded as a closed submanifold of $\boldsymbol{R}^{2n+1}$, so for $k > n$ we are not assuming anything special about M. We will consider M as a *Riemannian* submanifold of $\boldsymbol{R}^N$, i.e., we give it the Riemannian metric induced from $\boldsymbol{R}^N$.

Let $L(\boldsymbol{R}^N, \boldsymbol{R}^N)$ denote the vector space of linear operators from $\boldsymbol{R}^N$ to itself and $L^s(\boldsymbol{R}^N, \boldsymbol{R}^N)$ the linear subspace of self-adjoint operators. We define a map $P : M \to L^s(\boldsymbol{R}^N, \boldsymbol{R}^N)$, called the *Gauss map* of M by $P_x =$ orthogonal projection of $\boldsymbol{R}^N$ onto TM_x. We denote the kernel of P_x (that is the normal space to M at x) by ν_x. We will write $P_x^\perp$ for the orthogonal projection $I - P_x$ of $\boldsymbol{R}^N$ onto ν_x. Since the Gauss map is a map of M into a vector space, at each point x of M it has a well-defined differential $(DP)_x : TM_x \to L^s(\boldsymbol{R}^N, \boldsymbol{R}^N)$.

9.6.1. Definition. For each normal vector v to M at x we define a linear map $A_v : TM_x \to \boldsymbol{R}^N$, called the *shape operator of M at x in the direction v*, by $A_v(u) = -(DP)_x(u)(v)$.

Since the tangent bundle TM and normal bundle $\nu(M)$ are both subbundles of the trivial bundle $M \times \boldsymbol{R}^N$, the flat connection on the latter induces connections ∇^T and ∇^ν on TM and on $\nu(M)$. Explicitly, given $u \in TM_x$, a smooth curve $\sigma : (-\epsilon, \epsilon) \to M$ with $\sigma'(0) = u$, and a smooth section $s(t)$ of TM (resp. $\nu(M)$) along σ, we define $\nabla^T_u(s)$ (resp. $\nabla^\nu_u(s)$) by $P_x(s'(0))$ (resp. $P_x^\perp(s'(0))$). Clearly ∇^T is just the Levi-Civita connection for M.

The following is an easy computation.

9.6.2. Proposition. *Given u in TM_x and e in $\nu(M)_x$ let $\sigma : (-\epsilon, \epsilon) \to M$ be a smooth curve with $\sigma'(0) = u$ and let $s(t)$ and $v(t)$ be respectively tangent and normal vector fields along σ with $v(0) = e$. Let Pe denote the section $x \mapsto P_x(e)$ of $T(M)$. Then:*

(i) $A_e(u) = -P_x v'(0)$; *hence each A_v maps TM_x to itself,*

(ii) $A_e(u) = \nabla^T_u(Pe)$,

(iii) $\langle A_e(u), s(0)\rangle = \langle e, s'(0)\rangle$.

Suppose $F : \boldsymbol{R}^N \to \boldsymbol{R}$ is a smooth real valued function on $\boldsymbol{R}^N$ and $f = F|M$ is its restriction to M. Since $df = dF|TM_x$, it follows immediately from the definition of the gradient of a function that for x in M we have $\nabla f_x = P_x(\nabla F_x)$, and as a consequence we see that *the critical points of f are just the points of M where ∇F is orthogonal to M*. We will use this fact in what follows without further mention. Also, as we saw in the section on Morse functions, at a critical point x of f $\mathrm{Hess}(f)_x = \nabla^T(\nabla f)$.

We define a smooth map $H : \boldsymbol{S}^{N-1} \times \boldsymbol{R}^N \to \boldsymbol{R}$ by $H(a,x) = \langle a, x\rangle$ and, for each $a \in \boldsymbol{S}^{N-1}$, we define $H_a : \boldsymbol{R}^N \to \boldsymbol{R}$ and $h_a : M \to \boldsymbol{R}$ by $H_a(x) = H(a,x)$ and $h_a = H_a|M$. Each of the functions h_a is called a "height" function. Intuitively, if we think of a as the unit vector in the "vertical" direction, so $\langle a, x\rangle = 0$ defines the sea-level surface, then $h_a(x)$ represents the height of a point $x \in M$ above sea-level. Similarly we define $F : \boldsymbol{R}^N \times M \to \boldsymbol{R}$ by $F(a,x) = \frac{1}{2}\|x-a\|^2$, and for $a \in \boldsymbol{R}^N$ we define $F_a : \boldsymbol{R}^N \to \boldsymbol{R}$ and $f_a : M \to \boldsymbol{R}$ by $F_a(x) = F(a,x)$ and $f_a = F_a|M$. Somewhat illogically we will call each f_a a "distance" function.

For certain purposes the height functions have nicer properties, while for others the distance functions behave better. Fortunately there is one situation when there is almost no difference between the height function h_a and the distance function f_a.

9.6.3. Proposition. *If M is included in some sphere centered at the origin, then h_a and f_{-a} differ by a constant; hence they have the same critical points and the same Hessians at each critical point.*

PROOF. Suppose that M is included in the sphere of radius ρ, i.e., for x in M we have $\|x\|^2 = \rho^2$. Then

$$\begin{aligned} f_{-a}(x) &= \frac{1}{2}\|x+a\|^2 \\ &= \frac{1}{2}(\|x\|^2 + \|a\|^2) + \langle x, a\rangle \\ &= \frac{1}{2}(\rho^2 + \|a\|^2) + h_a(x). \quad \blacksquare \end{aligned}$$

Thus if the particular embedding of M in Euclidean space is not important we can always use stereographic projection to embed M in the unit sphere in one higher dimension and get both the good properties of height functions and of distance functions at the same time.

9.6.4. Proposition. *The gradient of h_a at a point x of M is $P_x a$, the projection of a on TM_x, so the critical points of h_a are just those points x of M where a lies in the space ν_x, normal to M at x. Similarly the gradient of f_a at x is $P_x(x-a)$, so the critical points of f_a are the points x of M where the line segment from a to x meets M orthogonally.*

PROOF. Since H_a is linear, $d(H_a)_x(v) = H_a(v) = \langle a, v\rangle$, so that $(\nabla H_a)_x = a$. Similarly, since F_a is quadratic we compute easily that $d(F_a)_x(v) = \langle x-a, v\rangle$ so $(\nabla F_a)_x = x - a$. $\blacksquare$

By another easy computation we find:

9.6.5. Proposition. *At a critical point x of h_a, $\mathrm{hess}(h_a)_x = A_a$. Similarly at a critical point x of f_a, $\mathrm{hess}(f_a)_x = I + A_{x-a}$.*

Thus, because the hessian of h_v is self-adjoint we see

9.6.6. Corollary. *For each v in $\nu(M)$, A_v is a self-adjoint operator on TM_x.*

We recall that for v in $\nu(M)_x$, the second fundamental form of M at x in the direction v is the quadratic form II_v on TM_x defined by A_v, i.e.,

$$II_v(u_1, u_2) = \langle A_v u_1, u_2 \rangle,$$

and the eigenvalues of A_v are called the principal curvatures of M at x in the normal direction v.

9.6.7. Proposition. *Given e in $\nu(M)_x$, let $v(t) = x+te$. Then for all real t, x is a critical point of $f_{v(t)}$ with hessian $I - tA_e$. Thus the nullity of $f_{v(t)}$ at x is just the multiplicity of t^{-1} as a principal curvature of M at x in the direction e. In particular, x is a degenerate critical point of $f_{v(t)}$ if and only if t^{-1} is a principal curvature of M at x in the direction e. If 1 is not such a principal curvature then x is a non-degenerate critical point of f_{x+e}, and its index is*

$$\sum_{0<t<1} \textit{nullity of } f_{v(t)} \textit{ at } x.$$

PROOF. The first statement follows directly from the above propositions by taking $a = x + te$, and it is then immediate that the nullity of $f_{v(t)}$ is $\mu(t^{-1})$, where $\mu(\lambda)$ denotes the multiplicity of λ as an eigenvalue of A_e. On the other hand, the multiplicity of λ as an eigenvalue of $\mathrm{hess}(f_{x+te})_x = 1 - A_e$ is clearly $\mu(1-\lambda)$. Since $\lambda < 0$ if and only if $1 - \lambda$ equals t^{-1} for some t in $(0, 1)$, the formula for the index of f_{x+e} at x follows. ∎

We will denote by $Y : \nu(M) \to \boldsymbol{R}^N$ the "exponential" or "endpoint" map $(x, v) \mapsto x + v$ of the normal bundle to M into the ambient $\boldsymbol{R}^N$.

9.6.8. Definition. If $a = Y(x, e)$ then a is called *non-focal* for M with respect to x if $DY_{(x,e)}$ is a linear isomorphism. If on the contrary $DY_{(x,e)}$ has a kernel of positive dimension m then a is called a *focal point* of multiplicity m for M with respect to x. A point a of $\boldsymbol{R}^N$ is called a *focal point of M* if, for some $x \in M$, a is focal for M with respect to x.

9.6.9. Proposition. *The point $a = Y(x,e)$ is a focal point of multiplicity m for M with respect to x if and only if x is a degenerate critical point of f_a of nullity m.*

PROOF. Let $\gamma(t) = (\sigma(t), v(t))$ be a smooth normal field to M along a smooth curve $\sigma(t)$, with $\sigma(0) = x$ and $v(0) = e$. Then:

$$\begin{aligned} DY_{(x,e)}(\gamma'(0)) &= \left(\frac{d}{dt}\right)_{t=0} Y(\sigma(t), v(t)) \\ &= \left(\frac{d}{dt}\right)_{t=0} (\sigma(t) + v(t)) \\ &= \sigma'(0) + v'(0) \\ &= \sigma'(0) + P_x v'(0) + P_x^{\perp} v'(0) \\ &= (I - A_e)\sigma'(0) + P_x^{\perp} v'(0). \end{aligned}$$

since by a proposition above $A_e\sigma'(0) = -P_x v'(0)$. Now taking $\sigma(t) \equiv x$ and $v(t) = e + tv$ gives the geometrically obvious fact that $DY_{(x,e)}$ reduces to the identity on the subspace $\nu(M)_x$. It then follows by elementary linear algebra that $\ker(DY_{(x,e)})$ and $\ker(I - A_e)$ have the same dimension. Since we have seen that $\mathrm{hess}(f_a) = I - A_e$ the final statement follows. ∎

9.6.10. Corollary. *If $a \in \boldsymbol{R}^N$ is not a focal point of M then the distance function f_a is a Morse function on M.*

9.6.11. Morse Index Theorem. *If M is a compact, smooth submanifold of $\boldsymbol{R}^N$, $x \in M$, $e \in \nu(M)_x$, and $a = x + e$ is non-focal for M with respect to x, then x is a non-degenerate critical point of the "distance function" $f_a : M \to \boldsymbol{R}$, $v \mapsto \left(\frac{1}{2}\right)\|v - a\|^2$, and the index of x as a critical point of f_a is just equal to the number of focal points for M with respect to x along the segment joining x to a, each counted with its multiplicity.*

PROOF. Immediate from the above. ∎

Next recall Sard's Theorem. Suppose X and Y are smooth, second countable manifolds of the same dimension and $F : X \to Y$ is a C^1 map. A point p of X is called a *regular point* of F if $DF_p : TX_p \to TY_{f(p)}$ is a linear isomorphism, or equivalently if F is a local diffeomorphism at p. A point q of Y is called a *regular value* of F if all points of $F^{-1}(q)$ are regular points of F; other points of N are called *critical values* of F. Then Sard's Theorem [DR, p.10] states that **the set of critical values of F has measure zero**, so

that in particular regular values are dense. Taking $X = \nu(M)$, $Y = \boldsymbol{R}^N$, and $F = Y$, the critical values are those points of $\boldsymbol{R}^N$ which are focal points of M. Thus, by the above Corollary, the distance function f_a is a Morse function for almost all $a \in \boldsymbol{R}^N$. In particular if f_a is not itself a Morse function, that is if a is a focal point of M, we can nevertheless choose a sequence a_n of non-focal points converging to a, and then f_{a_n} will be a sequence of Morse functions converging to f_a in the C^∞ topology.

As an easy application of this fact we can now give a simple proof that **any** smooth real valued function on M, $G : M \to \boldsymbol{R}$, can be approximated in the C^∞ topology by Morse functions. From the above remark it will suffice to show that G can be realized as a distance function, and of course it does no harm to change G by adding a constant. Define an embedding of M in the sphere of radius r in $\boldsymbol{R}^{N+2}$ by $x \mapsto \left(x, G(x), \sqrt{r^2 - \|x\|^2 - G(x)^2}\right)$, where of course r is chosen greater than the maximum of $\sqrt{\|x\|^2 + G(x)^2}$. Then, looked at in $\boldsymbol{R}^{N+2}$, G is clearly the height function h_a, where $a = (0, 1, 0)$. So, by an earlier remark, G differs by a constant from the distance function f_{-a}.